

# Újdonságok az OpenOffice.org Lingucomponent moduljában

*Németh László*

Lingucomponent  
projektvezető

2009-11-06



# Összefoglaló

- Mondatellenőrzés: API és bővítmények
- Tezaurusz komponens: tövezés és ragozás
- Elválasztás: kétszintű, angol javítások
- Writer: a kötőjel a szavak része
- Helyesírás-ellenőrzés
  - Bemeneti és kimeneti karakterátalakítás
  - Tövezés és szóalaktani elemzés
  - Kiejtés alapú javaslattevés
  - Speciális összetételek
  - Új eszközök



# Korrektúra, mondatellenőrzés

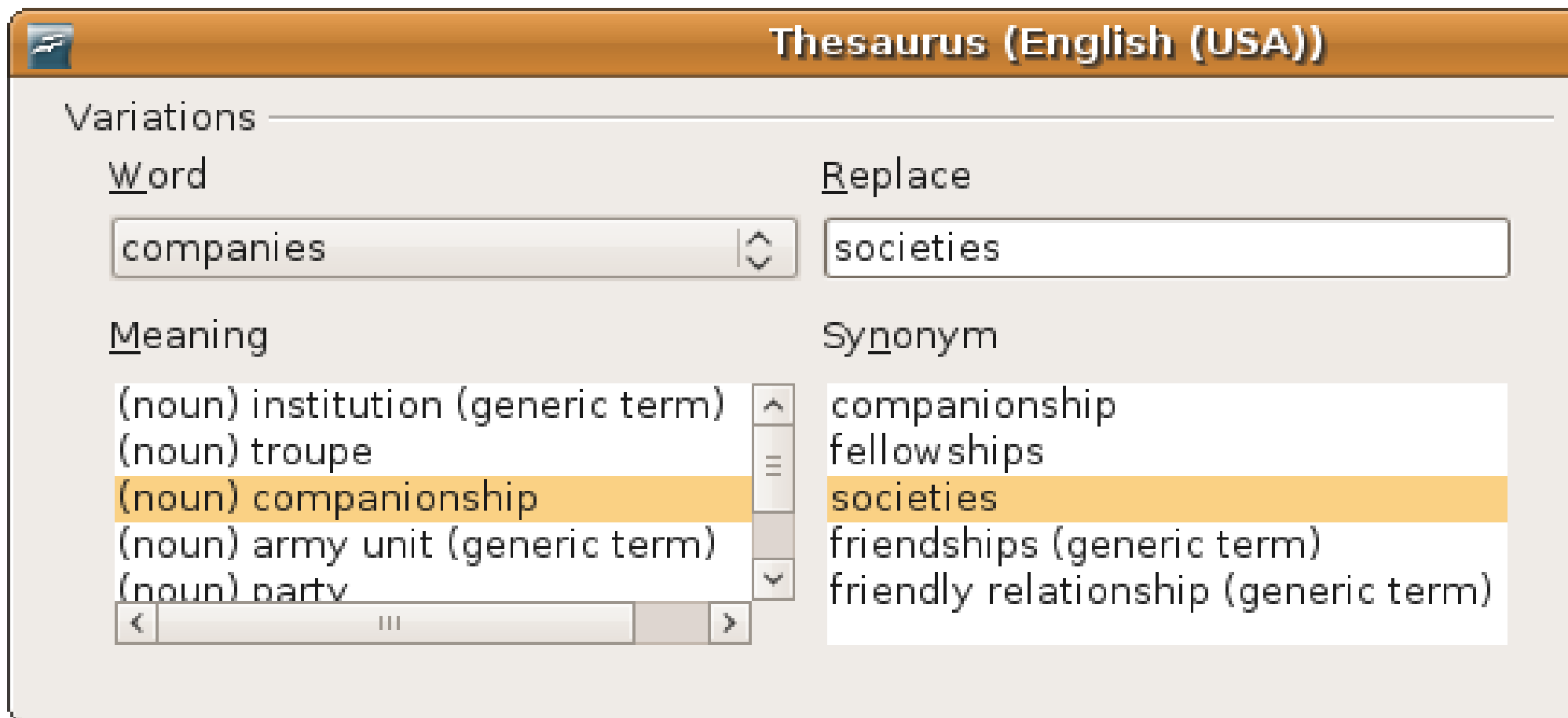
- OOo 3.0.1 (2009-01-28): API (Proofreader)
- CoGrOO (bővítmény a portugál nyelvhez)
- Esperantilo (bővítmény az eszperantóhoz)
- LanguageTool (18 nyelv támogatása)
  - Java + Java programkönyvtárak használata
  - Közösségi tesztelés a Wikipédia szövegein
- Lightproof (pehelysúlyú, magyar)
  - Python + regex + Hunspell
  - 6 kB forráskód



# I. Tezaurusz, szinonimatár modul

- Tövezés: “*companies*” → “*company*”
- Ragozás: “*society*” → “*societies*”

companies



The screenshot shows a software window titled "Thesaurus (English (USA))". It is divided into two main sections: "Variations" and "Synonym".

**Variations**

Word	Replace
companies	societies

**Meaning**

- (noun) institution (generic term)
- (noun) troupe
- (noun) companionship
- (noun) army unit (generic term)
- (noun) party

**Synonym**

- companionship
- fellowships
- societies
- friendships (generic term)
- friendly relationship (generic term)

# II. Tezaurusz

- A tövezés működik a legtöbb szótárral (a Hunspell stem() .dic bejegyzéseket ad)
- Más tő megadása az "st:" .dic mezővel:  
**mice st:mouse**

mice

The screenshot shows a window titled "Thesaurus (English (USA))". It is divided into two main sections: "Variations" and "Synonym".

**Variations**

Word	Replace
mice	somebodies

**Meaning**

(noun) rodent (generic term)	^
(noun) shiner	
(noun) person (generic term)	=
(noun) computer mouse	
(verb) sneak	
(verb) manipulate (generic term)	v

**Synonym**

persons (generic term)	^
individuals (generic term)	
someones (generic term)	=
somebodies (generic term)	
mortals (generic term)	
souls (generic term)	v

# III. Tezaurusz

- A ragozáshoz új adatokat is meg kell adni

- .dic fájl:

**mice st:mouse is:plural**

**mouse al:mice ts:nominative**

**society/S ts:nominative**

- .aff fájl:

**SFX S y ies [^aeiou]y is:plural**



# IV. Tezaurusz

- Toldalékok
  - **ds:** (képző): a tövezés visszaadja a képzőket tartalmazó alakokat is
  - **is:** (rag): a tövezés levágja a ragot vagy ragokat
  - **ts:** (terminális rag): mint az előző, de toldalékolásnál csak akkor számít, ha utolsó elem
- Hivatkozás az alternatív alakokra
  - **al:** (allomorf): egy szótári bejegyzésnek számos “al:” mezője lehet
  - A toldalékolás minden allomorfot végignéz



# V. Tezaurusz

- Tövezési, toldalékolási adatokkal bővített angol és magyar szótárak
- Új eszköz: *analyze*

```
cd /opt/openoffice.org3/share/uno_packages/cache/uno_packages/V6Jkz  
D_/dict-en.oxt
```

```
~/hunspell-1.2.8/src/tools/analyze en_US.aff en_US.dic /dev/stdin  
society
```

```
> society
```

```
analyze(society) = st:society ts:0
```

```
stem(society) = society
```

```
societies
```

```
> societies
```

```
analyze(societies) = st:society ts:0 is:Ns
```

```
stem(societies) = society
```

```
mouse societies [generate from 1st item using the analysis of 2nd item]
```

```
generate(mouse, societies) = mice
```

```
society mice
```

```
generate(society, mice) = societies
```



# I. Elválasztás - új képességek

- LEFTHYPHENMIN, RIGHTHYPHENMIN: elválasztási hely minimális távolsága a szó végétől
  - A TeX elválasztási minták helyes működéséhez
- Kétszintű elválasztás
  - Összetett szavak szintje
  - Szótövek szintje
- COMPOUNDLEFTHYPHENMIN, COMPOUNDRIGHTHYPHENMIN:
  - Az elválasztási hely minimális távolsága a szóhatártól az összetételekben



## II. Elválasztás - jobb angol

- Brit angol elválasztási szótár visszahelyezése (a brit és az amerikai elválasztás nem keverendő)
- RIGHTHYPHENMIN=3 az amerikai angolhoz
- Aposztróf melletti elválasztás tiltása (\**can*'*t*, \**o*'=*clock*, \**o*'*c*=*lock* stb.)
- 's és 't nem számít a HYPHENMIN értékbe
- A szó végi aposztróf sem (pl. *boys*' )



# III. Elválasztás - angol minták

ISO8859-1

LEFTHYPHENMIN 2

RIGHTHYPHENMIN 3 # jav. az amerikaihoz

COMPOUNDLEFTHYPHENMIN 2

COMPOUNDRIGHTHYPHENMIN 3

1' # aposztóf nem számít

1's./' = s,1,2 # trükk: új inaktív elválasztási pont

1't./' = t,1,2 # megadása: túl közel a szóvéghez

**NEXTLEVEL**

4'4

'c4



# IV. Elválasztás - szótárkészítés

- Első szint előállítása
  - Kiindulás: szavak csak összetételi határral jelölve, PatGennel feldolgozva
  - Alternatív minták hozzáadása (ggy/gy=gy)
  - OOo formátum előállítása a substrings.pl-lel
- Következő szint előállítása
  - Az első szinten nem elválasztott szavak (tövek és kimaradt összetételek) és az összetételek tagszavai PatGennel feldolgozva.
  - Alternatív minták
  - Substrings.pl



# V. Elválasztás - szótárkészítés

- Az elválasztás rekurzív az első szinten (a PatGen nem): ellenőrzés és javítás
  - A két mintafájl egybecsomagolása, a NEXTLEVEL kulcsszóval elválasztva:
    - [karakterkódolás]
    - [fejléc] (HYPHENMIN értékek)
    - [összetett szavak szintjének elválasztási mintái]
- NEXTLEVEL**
- [a következő szint mintái]



# A kiskötőjel (-) a szavak része

- a magyarnál eddig is az volt, de pl. az angol “*scot-free*”, “*topsy-turvy*” kezelésére eddig nem volt lehetőség (a “*scot*”, “*topsy*”, “*turvy*” magában hibás)
- A Hunspell tetszőleges kötőjeles szerkezetet felismer (egy-kettő-három)
- Módosítás: BREAK affix fájl paraméter
- Lásd a 64400-as számú OpenOffice.org hibát



# Hunspell

- De facto standard
  - Apple Mac OS X
  - Mozilla Firefox
  - OpenOffice.org, StarOffice
  - Opera
  - SDL Trados
  - Webkit (Google Chrome, Apple Safari)
- Támogatott nyelvek
  - ~100 (arab, baszk, koreai stb.)
  - Héber szótár optimalizálása



# Hunspell - I. i/o átalakítás

- ICONV és OCONV affix fájl paraméterek
- Unicode normalizáció
- Szubsztenderd kódolások támogatása
- Szótagjelek kódolása fonémaszintre





# Hunspell - II. i/o átalakítás

- Koreai Hunspell szótár (Changwoo Ryu)

Hangul szótagjelek (fonémák csoportosítva) → jobb javaslatok fonémaszinten → átalakítás *dzsamó* fonémákra → **각 → 가**

ICONV 1117

ICONV [U+AC00] [U+1100][U+1161]

ICONV [U+AC01] [U+1100][U+1161][U+11A8]

ICONV [U+AC02] [U+1100][U+1161][U+11A9]

...

- Javaslatok visszaalakítása: OCONV

# Hunspell - III. i/o átalakítás

- A joruba ó UTF-8-ban: ó[U+0329], o[U+0329][U+0301] vagy o[U+0301][U+0329]
- Tagolt normálforma kanonikus sorrendje:  
o[U+0329][U+0301] (LATIN SMALL LETTER O,  
COMBINING VERTICAL LINE BELOW, COMBINING  
ACUTE ACCENT)
- Hunspell normalizáció

ICONV 2

ICONV o[U+0301][U+0329] o[U+0329][U+0301]

ICONV ó[U+0329] o[U+0329][U+0301]

# Kiejtés alapú javaslattevés

- PHONE (Aspellből, jelenleg csak ASCII)
- Szótári, a .dic fájl **ph:** mezőjével  
**chihuahua ph:csivava**

# Összetett szavak kezelése

- **Buss + sjåfor** → **bussjåfor** (norvég, svéd)  
**CHECKCOMPOUNDTRIPLE** (hármás betűk tiltása)  
**SIMPLIFIEDTRIPLE** (kettős betűs alak elfogadása)
- Általánosítás (pl. indiai nyelvek sandhi hasonulásának kezelésére)  
**CHECKCOMPOUNDPATTERN** [vég[/kapcsolók]]  
[kezdet[/kapcsolók]] [csere]  
pl. **CHECKCOMPOUNDPATTERN ss s ss**  
(ekvivalens az első példával)

# Új eszközök

- Affixcompress
  - Kiindulás: szólista, eredmény: Hunspell szótár
- Doubleaffixcompress
  - Kiindulás: Hunspell szótár, eredmény: Hunspell szótár kétszintes toldalékolással (szuffixumok)
  - Héber szótár optimalizálása: normális betöltési idő
- Analyze (l. korábban)
- Chmorph:
  - Morfológiai átalakító szövegfájlokhoz



# XML Hunspell UNO API

- Alternatív SPELLML API a tövezéshez és ragozáshoz a spell() és suggest() Hunspell függvényeken keresztül
- Az XspellChecker UNO API isValid() és spell() metódusaival is használható
- Az OpenOffice.org tezaurusz és a Lightproof mondatellenőrző használja



# Köszönöm a figyelmet!

- Lingucomponent projekt
  - <http://lingucomponent.openoffice.org/>
- Hunspell, Hyphen:
  - <http://hunspell.sourceforge.net>

