Kontrolor pravopisu pro Esperanto

Bc. Marek Blahuš <xblah@fi.muni.cz>
Fakulta informatiky MU Brno

RASLAN, Karlova Studánka, 2009-12-04

Motivace

- snadnost publikování → mnoho textů → nedostatečná kontrola → nízká kvalita
  - zvláště pro malé jazyky jako esperanto
- automatická kontrola textů:
  - kontrolor pravopisu
    - Pokrovskij: ISpell, slabá morfologie (jediný kmen)
  - kontrolor gramatiky
    - Lingvohelpilo (E@I, 2009): Constraint Grammar

Nový kontrolor pravopisu

- A Spell Checker for Esperanto (bc. práce)
  - prototyp kontroloru v nástroji Hunspell
  - práce obhájena 26. 6. 2008
- MUNI33/212008 (studentský VV-projekt)
  - ladění, rozšiřování, testování, OpenOffice.org
  - ukončení projektu listopad 2009
- vedoucí práce / garant projektu:
  - doc. RNDr. Petr Sojka, Ph.D.

Morfologie esperanta

- aglutinatiční jazyk
  - *mal·bon·a* = špatný
  - *kaf·o·muel·il·o* = mlýnek na kávu
  - *mal·sam·ras·an·oj* = příslušníci jiné rasy
- struktura slova
  - kořeny, lexikální afixy (LA), gramatické afixy (GA)
  - *(LA*·Kořen·LA*·[aeio]?)*·LA*·Kořen·LA*·GA?*
  - kmen = kořen s LA měnícími jeho význam
  - kmeny se skládají do slov, typicky zakončených GA

Charakter kořenů

- volnost v tvorbě slovních tvarů z kořenů
  - *rapid·a* = rychlý, *rapid·e* = rychle,
    rapid·o = rychlost, *rapid·i* = spěchat
- přesto každý kořen základní význam
  - člověk, nástroj, činnost, vlastnost (*rapid*), ...
- afixy se pojí jen s kořeny určitých typů
  - *bo·patr·o* = tchán, *bo·frat·o* = švagr
- celkem 10 prefixů a 31 sufixů

Klasifikace kořenů

- Plena Ilustrita Vortaro (PIV): 16.780 kořenů
- dle chování 10 prefixů a 31 sufixů odvozeno 15 potřebných sémantických tříd pro kořeny
  - vlastnosti, činnosti, objekty
  - osoby, zvířata, rostliny
  - mužský rod, ženský rod, společný rod
  - místa, rodinné vztahy, čísla
  - tranzitivita, funkční slova, schopný utvářet antonyma
- automatická klasifikace
  - odborné slovníky, uzavřené seznamy, korpusy

Povolené struktury slov

- Witkam 2008: ESPSOF – 33.000 slov s vyznačenou morfologickou strukturou
- struktura slova = kombinace kořenů a afixů
- celkem 632 různých struktur, nejčastější:
  - *Kořen·GA*: 39 %
  - *Kořen·Kořen·GA*: 18 %
  - *Kořen·o·Kořen·GA*: 4 %
  - struktury obsahující *LA*: 38 %
    - nejčastější *Root·aĵ·GA*

Afixová pravidla

- odvozena ručně pro každý nalezený vzor identifikací použitelných tříd kořenů
- např. u *Kořen·id·GA* smí být kořen z tříd:
  - zvíře: *kat·o* → *kat·id·o* (kočka → kotě)
  - rostlina: *kverk·o* → *kverk·id·o* (dub → doubek)
  - osoba: *reĝ·o* → *reĝ·id·o* (král → princ)
- výsledkem regulární výraz pro Hunspell
  - (zvíře+rostina+osoba)·*id·o*

Implementace

- Hunspell
  - OS nástroj pro kontrolu pravopisu
  - až dvojitá afixace, ale návrhy oprav jen z jednoduché
  - částečná podpora regulárních výrazů
    - jen * a ?, ne + → nutnost rozepsání + do jednotlivých výrazů
  - celkem navrženo 37.155 regulárních výrazů (pravidel)
- Mozilla Firefox, OpenOffice.org
  - doplněk (extension) – slovník, licence a metainformace
  - bug tracking v OpenOffice.org Issue Tracker
  - součást esperantské lokalizace OpenOffice.org (1/2010)

Závěr

- testování na přepisu rukopisu povídky
- falešná pozitiva
  - Pokrovskij: 3823 ~ 7 % (1565 ~ 15 %)
  - Blahuš: 2140 ~ 4 % (546 ~ 5 %)
- detekce:
  - Pokrovskij: 206 (148) chyb
  - Blahuš: 167 (113) chyb (tj. o 19 % méně)
- nutnost vyvážení obou ukazatelů

Děkuji za pozornost.

Bc. Marek Blahuš