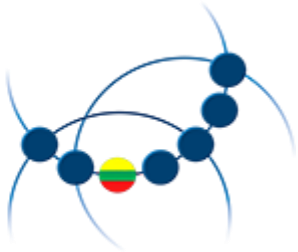
CLARIN-LT



Lietuvių kalbos morfologijos duomenų bazės tobulinimo rezultatai

Vilniaus universitetas

Virginijus Dadurkevičius, Arūnas Samuilis, Danielius Ralys, Jonas Vaičiulis, Franciška Ralienė

Kaunas, 2016 m. gruodžio 9 d.

Įvadas

- **Laikotarpis**. 2015–2016 metai
- **Tikslas**. Išplėsti VDU/KTU „Semantikos“ projekto metu sukurtą lietuvių kalbos morfologijos bazę naujomis lemomis, geriau atspindint šiuolaikinį Lietuvos interneto turinį ir kitus skaitmeniškai prieinamus šaltinius.
- **Vykdytojai**. Vilniaus universiteto Taikomųjų mokslų instituto 5 darbuotojai, iš viso – 2 etatai.
- **Siekiamas rezultatas**. Bent 10000 lemų išplėsta lietuvių kalbos morfologijos duomenų `.aff` ir `.dic` failai, užrašyti naudojant atvirojo kodo ***Hunspell*** platformos formalizmą. Šie failai turi būti perkelti ir naudojami VDU morfologijos analizatoriuje.

Šaltiniai

- **VDU BIT** tekstyno fragmentas (~1 mln. žodžių, 115 tūkst. unikalių),
- **LRS** tekstynas (~400 mln. žodžių, 1 mln. unikalių),
- **VU** mašininio vertimo tekstynas (~800 mln. žodžių, 4 mln. unikalių)

Kuo išplėsta morfologija?

- **Pavardės**. Virš 4000 naujų lemų. Maždaug vienodai vyriškų, moteriškų ir mergaičių.
- **Veiksmų vardai** (-ymas). Apie 800 lemų.
- **Neigiami būdvardžiai** (ne-, nebe-). Apie 1500 lemų.
- **Neigiami veiksmažodžiai** (ne-, nebe-). Apie 7000 lemų.
- **Sangrąžiniai veiksmažodžiai** (-tis). Apie 700 lemų.
- **Neigiami sangrąžiniai veiksmažodžiai** (nesi-, nebesi-). Apie 700 lemų.
- **Veiksmažodžiai su te-, tebe-**. Apie 3000 lemų.
- **Kiti žodžiai**. Apie 450 įvairių kalbos dalių lemų.

- **Iš viso** pridėta 18254 naujos lemos.

Morfologijos duomenų bazės struktūra

- Darybos taisyklių rinkinys (.aff failas)
  - 17781 atskira taisyklė
  - 4865 taisyklių grupės
- Klasifikuotų lemų rinkinys (.dic failas)
  - 41674 bendriniai daiktavardžiai
  - 73596 tikriniai daiktavardžiai
  - 14499 būdvardžiai
  - 34614 veiksmažodžiai
  - 3839 prieveiksmiai
  - 53 įvardžiai
  - 152 skaitvardžiai
  - 2272 kitos kalbos dalys, sutrumpinimai ir pan.
  - Iš viso **170699** lemos

Pavyzdžiai

- .aff failo fragmentas

```
AM is:Indic_PastFreq_Sg_II_short
AM is:Indic_PastFreq_Sg_III_short
AM is:Obli_PastFreq_Masc_Sg
AM is:Obli_PastFreq_Masc_Pl
AM is:Obli_PastFreq_Fem_Sg
...
SFX 350 Y 2
SFX 350 tis čiai . 55
SFX 350 tys čiai . 55

SFX 352 Y 2
SFX 352 tis čiuosna . 53
SFX 352 tys čiuosna . 53

SFX 364 Y 5
SFX 364 uo uo . 37
SFX 364 uo uo . 80
SFX 364 uo esio . 38
SFX 364 uo esiai . 39
SFX 364 uo esių . 40
```

- .dic failo fragmentas

```
terlius/52              5
terma/33                5
termas/6                5
Termentas/28            7
termikas/6              5
Terminaitė/3            6
terminalas/6            5
```

Įdomybės

- Negalima automatiškai, be žmogaus patikrinimo, pridėti lemų iš tekstynuose neatpažintų žodžių sąrašo. Pora pavyzdžių iš VU mašininio vertimo tekstyno su nurodytais dažniais:
  - **neriekti** ?← neriekia (79), neriek (10), neriektu (5), nerieks (5) , nerieki (1) , neriekiat (1)
  - **buriuotis** ?← buriavosi (221), buriuojasi (170), buriuotis (53), buriuosis (12) , buriuodavosi (8) , buriavęsi (2) , buriuotųsi (1) , buriuodamiesi (1)
- Dabartinis VU mašininio vertimo tekstyno „padengiamumas“ –
  - Išplėstu ***Hunspell*** variantu – 92,7%
  - Vytauto Zinkevičiaus ***Lemuokliu*** – 89,2%

Klausimai?