

Tekstynas **LYTIS_PROF** skirtas autorių profilių (pagal lytį) nustatymo tyrimams.

Tekstyną sudaro parlamentarų pasisakymai Seime
Laikotarpis: 1990 m. kovo mėn. – 2013 m. gruodžio mėn.
Klasės: mot ir vyr
Minimalus žodžių kiekis tekste: 100 žodžių

Eilutės kiekviename iš tekstų atitinka teksto savybes, naudojamas autorystės nustatymo tyrimuose (daugiau apie savybes skaitykite [1]):

- 1) vidutinis sakinio ilgis tekste
- 2) vidutinis žodžio ilgis tekste
- 3) skirtingų žodžių ir visų žodžių tekste santykis
- 4) tekstas
- 5) lemuotas tekstas (lemuota automatinio įrankiu "Lemuoklis"[2]: atpažįstamas žodis verčiamas lema, bendrinis žodis taip pat verčiamas mažosiomis raidėmis; jeigu žodis neatpažįstamas – paliekama neliesta jo forma)
- 6) kalbos dalys (nustatytos "Lemuoklio" pagalba)
- 7) priklausomybių gramatikos sintaksinės žymos kiekvienam iš žodžių (apdorota automatinio įrankiu MaltParser¹, prieš tai jį apmokius su lietuviškuoju medžių banku² bei didžiausią tikslumą duodančiais parametrais [3])
- 8) funkciniai žodžiai (jungtukai, jaustukai, dalelytės, prielinksniai, įvardžiai: nustatyti automatinio lemavimo įrankiu "Lemuoklis")
- 9) simbolių 2-gramos (tekstinio dokumento mastu). Tekstas suskaidomas į elementus po du simbolius (pritaikius slenkančio lango metodą per vieną simbolį), tarpo simboliai prieš tai keičiami "_". Pvz.: "visas_tekstas": "vi", "is", "sa", "as", "s_", "_t", "te" ir t.t.
- 10) simbolių 3-gramos (tekstinio dokumento mastu)
- 11) simbolių 4-gramos (tekstinio dokumento mastu)
- 12) simbolių 5-gramos (tekstinio dokumento mastu)
- 13) simbolių 6-gramos (tekstinio dokumento mastu)
- 14) simbolių 7-gramos (tekstinio dokumento mastu)
- 15) tekstas junginiais: žodžiais ir jų kalbos dalimis
- 16) tekstas junginiais: lemomis ir žodžių kalbos dalimis
- 17) tekstas junginiais: žodžiais, jų kalbos dalimis bei kita morfologine informacija (morfologinė informacija, pvz. linksnis, skaičius, laipsnis, laikas ir kt. nustatyta "Lemuoklio" pagalba)
- 18) tekstas junginiais: lemomis, žodžių kalbos dalimis bei kita morfologine informacija

Nuorodos:

[1] Kapočiūtė-Dzikiėnė, Jurgita; Utkā, Andrius; Šarkutė, Ligita. 2014. Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. *Kalbotyra*, 66: 27–45.

[2] Zinkevičius, Vytautas. 2000. Lemuoklis – morfologinei analizei. *Gudaitis, L. (red.) Darbai ir Dienos*, 24: 246–273.

[3] Kapočiūtė-Dzikiėnė, Jurgita; Nivre, Joakim; Krupavičius, Algis. 2013. Lithuanian Dependency Parsing with Rich Morphological Features. *Empirical Methods in Natural Language Processing – 4th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL'2013)*, psl. 12–21.

¹ Įrankis MaltParser parsisiųstas iš <http://www.maltparser.org/>.

² Medžių bankas sukurtas VDU vykdyto Lietuvos valstybinio mokslo ir studijų fondo Lituaniškos mokslinių tyrimų prioriteto įgyvendinimo 2007–2008 metų programos projekto "Internetiniai ištekļiai: anototas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA 2)" metu.

Statistika

klasė	pavyzdžių kiekis	žodžių kiekis	skirtingų žodžių kiekis	vid. pvz. ilgis žodžiais
mot	10727	2357596	93611	219.78
vyr	100181	21550706	268030	215.12
Visi	110908	23908302	279494	215.57