

Tekstynas **GROŽ_AMŽIUS_PROF** skirtas autorių profilių (pagal amžių) nustatymo tyrimams.

Tekstyną sudaro grožinių kūrinių autorių tekstai

Klasės: iki 29; 30-39; 40-49; 50-59; 60-69; virš 70

Minimalus teksto ilgis: 2000 simbolių (įskaitant tarpus)

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) simbolių 2-gramos (žodžio ribose). Tekstas suskaidomas į elementus po du simbolius (pritaikius slenkančio lango metodą per vieną simbolį). Pvz.: "visas_tekstas": "vi", "is", "sa", "as", "te", "ek", "ks", "st", "ta", "as".
- 16) simbolių 3-gramos (žodžio ribose)
- 17) simbolių 4-gramos (žodžio ribose)
- 18) simbolių 5-gramos (žodžio ribose)
- 19) simbolių 6-gramos (žodžio ribose)
- 20) simbolių 7-gramos (žodžio ribose)
- 21) tekstas junginiais: žodžiais ir jų kalbos dalimis
- 22) tekstas junginiais: lemomis ir žodžių kalbos dalimis
- 23) tekstas junginiais: žodžiais, jų kalbos dalimis bei kita morfologine informacija (morfologinė informacija, pvz. linksnis, skaičius, laipsnis, laikas ir kt. nustatyta "Lemuoklio" pagalba)
- 24) tekstas junginiais: lemomis, žodžių kalbos dalimis bei kita morfologine informacija
- 25) tekstas junginiais: kalbos dalimis kartu su morfologine informacija

Nuorodos:

[1] Kapočiūtė-Dzikienė, 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-Dzikienė, 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 ištekliai: anototas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA 2)" metu.

Statistika

klasė	pavyzdžių kiekis	žodžių kiekis	skirtingų žodžių kiekis	vid. pvz. ilgis žodžiais
iki 29	2264	658348	97176	290.79
30-39	5727	1659564	186572	289.78
40-49	6465	1843627	191583	285.17
50-59	6523	1889891	205488	289.73
60-69	5937	1751719	183320	295.05
virš 70	6724	1958928	215670	291.33
Visi	33640	9762077	504156	290.19