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

Tekstyną sudaro parlamentarų pasisakymai Seime
Laikotarpis: 1990 m. kovo mėn. – 2013 m. gruodžio mėn.
Klasės: iki 29; 30-39; 40-49; 50-59; 60-69; virš 70
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 automatiniu 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 automatiniu įrankiu MaltParser[1], prieš tai jį apmokius su lietuviškuoju medžių banku[2] bei didžiausią tikslumą duodančiais parametrais [3])
8) funkciniai žodžiai (jungtukai, jaustukai, dalelytės, prielinksniai, įvardžiai: nustatyti automatiniu 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-Dzikienė, Jurgita; Utka, 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 Lituanistikos mokslinių tyrimų prioriteto įgyvendinimo 2007–2008 metų programos projekto "Internetiniai ištekliai: anotuotas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA 2)" metu.

Statistika

klasė	pavyzdžių kiekis	žodžių kiekis	skirtingų žodžių kiekis	vid. pvz. ilgis žodžiais
iki 29	707	145682	20207	206.06
30-39	16549	3577765	107128	216.19
40-49	28700	6139128	146607	213.91
50-59	42325	9199619	188650	217.36
60-69	17895	3867463	129477	216.12
virš 70	4732	978645	60604	206.81
Visi	**110908**	**23908302**	**279494**	**215.57**