Lithuanian Parliament Corpus (for Authorship Attribution)

Lithuanian Parliament corpus is designed for authorship attribution task. The corpus consists of transcripts of parliamentarians in Lithuanian Seimas.

Period: March, 1990 – December, 2013
Number of authors: 147
Minimum number of words in a text: 100

Each line in a file contains a different text feature that can be used in the authorship attribution task (Kapočiūtė Dzikienė et al. 2014):

1) average sentence length;
2) average word length;
3) type / token ratio;
4) text;
5) lemmatized text (lemmatized by the tool "Lemuoklis" (Zinkevičius 2000): known words are replaced by their lemmas; lowercased; unknown words are unchanged;
6) part of speech (POS tagger - "Lemuoklis");
7) dependency tags (processed by MaltParser[1], that was trained on the Lithuanian treebank[2]);
8) function words (conjunctions, interjections particles, prepositions, pronouns: that were identified by "Lemuoklis");
9) character 2-grams (within a text document). A text is split into two characters (by using a moving window that moves by a one symbol step). Spaces are replaced by _. For example, “visas tekstas” becomes "vi", "is", "sa", "as", "s_", "_t", "te", "ek", "ks", "st", "ta", "as";
10) character 3-grams (within a text document);
11) character 4-grams (within a text document);
12) character 5-grams (within a text document);
13) character 6-grams (within a text document);
14) character 7-grams (within a text document);
15) concatenated words and part of speech *lexpos*;
16) concatenated lemmas and part of speech *lempos*;
17) concatenated words and morphological information (e.g. “Noun Common Masculine Singular Genitive");
18) concatenated lemmas and morphological information.

References:

Kapočiūtė-Dzikienė, Jurgita, Utka, Andrius, Šarkutė, Ligita. 2014. Feature exploration for authorship attribution of Lithuanian parliamentary speeches. Text, speech and dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014: proceedings, 93-100.

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), 12-21.

Zinkevičius, Vytautas. 2000. Lemuoklis - morfologinei analizei. Gudaitis, L. (ed.) Darbai ir Dienos, 24: 246-273.

[1] MaltParser is taken from http://www.maltparser.org/.

[2] Lithuanian Treebank was compiled during the project "Internet Resources: Annotated Lithuanian Corpus and Tools of Annotation (ALKA 2)" in 2007-2008.

Statistics

No.	Class	Samples	Tokens	Types	Avg. Sample length
1.	ABIŠALA A A	873	181729	19494	208.17
2.	ALBERTYNAS A	229	58229	12842	254.28
3.	ALEKNAITĖ-ABRAMIKIENĖ V	390	90152	16610	231.16
4.	ANDRIUKAITIS V P	3932	926247	54282	235.57
5.	ANTANAVIČIUS K	981	286707	28937	292.26
6.	ASTRAUSKAS V	203	40522	8366	199.62
7.	AUŠTREVIČIUS P	619	125572	19699	202.86
8.	AŽUBALIS A	207	47941	10876	231.60
9.	BALEŽENTIS A	570	95483	13859	167.51
10.	BALTRAITIENĖ V	418	79584	11711	190.39
11.	BALČYTIS Z	528	123344	16233	233.61
12.	BASKAS A	430	101262	15797	235.49
13.	BEINORTAS J	1105	244656	28032	221.41
14.	BEKINTIENĖ D	243	49000	9504	201.65
15.	BENDINSKAS A	238	40153	8436	168.71
16.	BERNATONIS J	1223	227944	20213	186.38
17.	BIČKAUSKAS E	704	128212	16084	182.12
18.	BLINKEVIČIŪTĖ V	774	192638	18070	248.89
19.	BOBELIS K	322	72530	12981	225.25
20.	BOGUŠIS V	275	54305	11865	197.47
21.	BRADAUSKAS B	695	131421	19638	189.09
22.	BRAZAUSKAS A M	826	305606	30658	369.98
23.	BURBIENĖ S	506	118664	15598	234.51
24.	BUTKEVIČIUS A	761	208524	24655	274.01
25.	DAGYS R J	1934	406769	35046	210.33
26.	DAUKŠYS K	796	161541	20576	202.94
27.	DEGUTIENĖ I	2491	484498	29686	194.50
28.	EINORIS V	273	55133	11114	201.95
29.	ENDRIUKAITIS A	543	121442	23024	223.65
30.	ENDZINAS A	281	44689	9811	159.04
31.	GAPŠYS V	229	49108	9902	214.45
32.	GEDVILAS V	1131	213084	16088	188.40
33.	GENTVILAS E	1221	231640	20337	189.71
34.	GLAVECKAS K	764	213584	22730	279.56
35.	GRAKAUSKAS E	308	80593	12061	261.67
36.	GRAUŽINIENĖ L	862	164964	17815	191.37
37.	GRAŽULIS P	1442	287855	30870	199.62
38.	GRICIUS A	485	93202	14435	192.17
39.	GRUBLIAUSKAS V	211	44177	9612	209.37
40.	GYLYS P	320	79762	15396	249.26
41.	JACKŪNAS Ž J	276	81894	12209	296.72
42.	JAKUČIONIS P	389	84861	15535	218.15
43.	JANKAUSKAS D	468	95453	12981	203.96
44.	JARAŠIŪNAS E	277	77521	12204	279.86
45.	JASKELEVIČIUS L K	274	68862	12387	251.32
46.	JUKNEVIČIENĖ R	660	147449	21607	223.41
47.	JUKNEVIČIUS Z	512	106924	15400	208.84
48.	JURŠĖNAS Č	7779	1454762	52595	187.01
49.	KAKTYS S	210	49687	10293	236.60
50.	KAROSAS J	354	77662	14287	219.38
51.	KATKUS J A	253	54071	10503	213.72
52.	KAŠĖTA A	442	80757	13048	182.71
53.	KIRKILAS G	575	164584	21007	286.23
54.	KLUMBYS E	1252	264553	27781	211.30
55.	KUBILIUS A	5093	1076084	53593	211.29
56.	KUNEVIČIENĖ E J	707	172178	19205	243.53
57.	KUNČINAS A	223	44472	8930	199.43
58.	KUPČINSKAS R	244	40825	9129	167.32

59.	LANDSBERGIS V	1732	595325	53521	343.72
60.	LAPINSKAS K	287	103365	13632	360.16
61.	LINKEVIČIUS L A	285	65164	11171	228.65
62.	LIONGINAS J	366	79571	11924	217.41
63.	LISTAVIČIUS J	1402	271738	24894	193.82
64.	MALKEVIČIUS S	344	89696	15839	260.74
65.	MARGEVIČIENĖ V V	284	45360	9043	159.72
66.	MASIULIS E	930	187508	23354	201.62
67.	MASIULIS K	320	68581	14002	214.32
68.	MATULAS A	910	185155	20558	203.47
69.	MATULEVIČIUS A	729	151747	20479	208.16
70.	MAZURONIS A	289	53754	11066	186.00
71.	MAZURONIS V	560	119346	17819	213.12
72.	MIKUTIENĖ D	427	81991	13900	192.02
73.	MILČIUS L	362	96203	16903	265.75
74.	MIŠKINIS P A	314	71143	14422	226.57
75.	MONKEVIČIUS A	253	60799	10857	240.31
76.	MUNTIANAS V	2241	384398	22249	171.53
77.	NAVICKAS V	314	82654	12789	263.23
78.	OLEKAS J	1614	317478	32332	196.70
79.	OZOLAS R	216	58654	12862	271.55
80.	PALAITIS R	347	67820	12554	195.45
81.	PANGONIS J	277	59772	10741	215.78
82.	PAPOVAS P	697	155169	14840	222.62
83.	PATACKAS A V	250	57002	13576	228.01
84.	PAULAUSKAS A	1808	361556	32174	199.98
85.	PAUŽA B	203	41827	8273	206.04
86.	PAVIRŽIS G A	492	124205	18595	252.45
87.	PEKELIŪNAS A	277	40863	5542	147.52
88.	PETRAUSKAS V	448	102260	15435	228.26
89.	PEČELIŪNAS S	2430	458652	35146	188.75
90.	PRAPIESTIS J	484	118672	15444	245.19
91.	PRONCKUS M	524	133542	18518	254.85
92.	PRUNSKIENĖ K D	971	292414	31034	301.15
93.	PUPINIS E	559	97951	14553	175.23
94.	RAMONAS J	201	51177	9427	254.61
95.	RASIMAVIČIUS L N	207	52116	10228	251.77
96.	RAZMA J	1719	316059	32168	183.86
97.	RAŠKINIS A J	287	60240	12012	209.90
98.	RINKEVIČIUS V	279	51047	9921	182.96
99.	RUDYS A	704	189548	22065	269.24
100.	RUDZYS R	316	60426	12165	191.22
101.	RUPEIKA B V	305	62829	14106	206.00
102.	SABATAUSKAS J	840	165027	19158	196.46
103.	SABUTIS L	1142	241574	21457	211.54
104.	SADECKAS A	228	49780	9385	218.33
105.	SAKALAS A	2024	400853	32606	198.05
106.	SALAMAKINAS A	553	93212	14095	168.56
107.	SAULIS V	233	42663	8763	183.10
108.	SKARDŽIUS A	1298	211002	15637	162.56
109.	SMETONA R	366	81533	13586	222.77
110.	STANCIKIENĖ A	270	52548	10632	194.62
111.	STANKEVIČIUS Č V	880	197587	22631	224.53
112.	STARKEVIČIUS K	243	48846	10502	201.01
113.	STASIŠKIS A N	601	137910	19982	229.47
114.	STEPONAVIČIUS G	2539	456598	28934	179.83
115.	STOMA S	200	41199	9294	206.00
116.	STUNDYS V	360	66216	12259	183.93
117.	SYSAS A	2130	451444	34044	211.95
118.	TAMULIS J	241	57118	9916	237.00
119.	TAURANTAS A	846	178287	16803	210.74

120.	TREINYS M	304	109141	17361	359.02
121.	VAGNORIUS G	1237	517318	36131	418.20
122.	VAIŠNORAS A	455	100213	16098	220.25
123.	VALINSKAS A	511	88656	10740	173.50
124.	VARAŠKA M	209	47556	10198	227.54
125.	VESELKA J	3018	672644	51589	222.88
126.	VIDŽIŪNAS A	2654	455756	27287	171.72
127.	VISOKAVIČIENĖ B T	255	65582	9934	257.18
128.	VĖSAITĖ B	503	102037	17538	202.86
129.	ZASČIURINSKAS M	594	114568	15924	192.88
130.	ZIMNICKAS V V	308	58160	9540	188.83
131.	ZINGERIS E	262	63668	13557	243.01
132.	ČAPLIKAS A	779	165450	18501	212.39
133.	ČEKUOLIS J	413	67995	12763	164.64
134.	ČEPAS V	408	81582	16714	199.96
135.	ČIGRIEJIENĖ V M	277	48416	11429	174.79
136.	ŠEDBARAS S	542	111373	14048	205.49
137.	ŠEDŽIUS A	260	50452	9711	194.05
138.	ŠIAULIENĖ I	354	83213	15835	235.06
139.	ŠILGALIS Ž	300	63785	10505	212.62
140.	ŠIMAŠIUS R	332	80244	13210	241.70
141.	ŠIMĖNAS A	474	157179	18532	331.60
142.	ŠIMĖNAS J	541	120778	18069	223.25
143.	ŠLIČYTĖ Z	335	86908	14469	259.43
144.	ŠUKYS R	1106	259971	23437	235.06
145.	ŽAKARIS E	273	49315	8234	180.64
146.	ŽIEMELIS V	601	130369	19100	216.92
147.	ŽUKAUSKAS H	314	52734	9672	167.94
	Total	**110908**	**23908302**	**279494**	**215.57**