Socialinis skaitmeninis diskursas: nauji iššūkiai kompiuteriniams lingvistams ir kalbos technologams

Darius Amilevičius

Vytauto Didžiojo universitetas

VYTAUTO DIDŽIOJO UNIVERSITETAS

Socialinis Skaitmeninis Diskursas

Socialinis diskursas – tarpusavyje susijusių asmenų asmeninio pobūdžio diskusija filosofinėmis, politinėmis, religinėmis, literatūrinėmis ar pan. temomis.

Skaitmeninių informacijos perdavimo technologijų plėtra atvėrė iš esmės naujas globalaus bendravimo ir sąveikos galimybes – sukūrė pasaulio komunikacinę erdvę be įprastų laiko ir atstumo suvaržymų.

2010 m. Internetas buvo nominuotas Nobelio taikos premijai už dialogo, diskusijų ir bendro sutarimo skatinimą per bendravimą. Nominacijos dokumente buvo teigiama, kad Internetas yra "*a tool for peace. Anyone who uses it can sow the seeds of non-violence. And that's why the next Nobel Peace Prize should go to the Net.* ***A Nobel for each and every one of us***"

Analizė (Social Media Analytics – interdisciplinary approach) – teisėta ir etiška ?

„Gražuolės ir pabaisos" fenomenas (pvz. „taika" vs. „diskriminacija")?

VYTAUTO DIDŽIOJO UNIVERSITETAS

CLARIN-LT

Socialinių tekstų tekstynas LITIS v.1

(sudarė: D. Amilevičius, M. Petkevičius)

182 000 LRytas komentarų / 6 212 854 žodžių/ 507 471 formų

17 909 Delfi komentarų / 430 545 žodžių/ 92 667 formų

Iš viso: 199 909 komentarų / 6 643 399 žodžių/ 456 576 formų

---

1 komentaras = 1 skaitmeninė byla
Skaitmeninės bylos turinys:

1. Thu Apr 10 16:47:00 EEST 2014 (data)
2. Noriu pasitikslinti.... (komentoriaus slapyvardis/ įžanginė frazė?)
3. http://www.delfi.lt/news/daily/lithuania/... (straipsnio adresas)
4. Ministrui dėl V. Vonžutaitės vyro svyla padai (straipsnio pavadinimas)
5. "Вонджутайте" - čia Fedotovos rinkiminis slapyvardis? (komentaro tekstas)

DLKT / BIT / LITIS v.1

(10 dažniausių formų su „stop“ žodžiais)

DLKT (grožinė) (“Norminė kalba”)	**BIT2014_04 (“Kalba be keiksmažodžių)**	**BIT 2015_10 (“Kalba be keiksmažodžių)**	**LITIS v.1 („Patogi kalba“)**
17 mln. (12,3%)	57 mln.	108 mln.	6,6 mln.
599702 ir	ir 1763767	ir 3377079	229477 ir
219630 į	kad 674265	kad 1430009	88417 kad
187648 kad	su 366901	yra 755082	78532 o
144363 iš	yra 366004	iš 737328	74352 tai
130444 o	iš 360662	su 718464	49141 kaip
129112 jis	tai 332909	tai 677487	44346 su
119381 kaip	buvo 279752	buvo 579927	41544 ne
110846 su	kaip 258002	kaip 516578	40016 tik
109475 buvo	ar 236564	ar 469419	36385 bet
102991 tai	savo 226362	bet 422054	36325 ar

DLKT / BIT / LITIS v.1

(10 dažniausių formų be „stop“ žodžių)

DLKT (grožinė) (“Norminė kalba”)		**BIT2014_04 (“Kalba be keiksmažodžių)**		**BIT2015_10 (“Kalba be keiksmažodžių)**		**LITIS v.1 („Patogi kalba“)**	
17 mln. (12,3%)		57 mln.		108 mln.		6,6 mln.	
109475 buvo	9	yra 366004	4	yra 755082	3	32275 yra	12
33451 yra	45	lietuvos 172157	17	buvo 579927	7	20706 buvo	23
24715 būtų	59	metų 138750	26	metų 251221	27	20469 lietuvos	24
16133 būti	89	metu 98417	39	lietuvos 246599	28	13671 reikia	46
15909 vieną	93	būti 97946	40	metu 199059	37	13043 bus	52
15298 gerai	95	turi 89825	46	būti 190101	40	12574 gali	53
15003 kartą	96	daug 74775	59	bus 188923	41	12007 turi	56
14390 reikia	99	galima 73083	60	daugiau 188177	42	11497 lietuva	58
14216 bus	101	sakė 70183	63	sakė 143013	60	9291 lietuvoje	75
14117 viskas	102	lietuvoje 69183	65	galima 141707	62	9238 papildyta	76

Iššūkiai kompiuteriniams lingvistams ir kalbos technologams

„jezau jezau..tigi teip negalim daugiau gyvent! “ (iš komentaro)

Socialinio skaitmeninio diskurso automatinio duomenų parengimo etapo iššūkiai:

1) žodžių segmentavimas (pvz.: „mamamala“, „m.a.m.a“)
2) normalizavimas (pvz.: „is“ ir „iis“ turi tapti „iš“)
3) morfologija (žr. skaidres toliau)
4) sintaksė (automatinis skaidymas frazėmis; sudieu įprastinei sintaksei)
5) Sentimentų analizė (pvz.: jaustukai, necenzūriniai žodžiai – emocijos, socialinės tapatybės statusas... ?)*
6) Įvardytų esybių atpažinimas (pvz.: iš anksto sudaryti sąrašai neatitinka realijų)
7) Tekstų klasifikatorius

---

* Čekuolytė A., “He blet nachui was in a shop”: Swearing Practices and Attitudes to Swearing among Vilnius Adolescents. Taikomoji kalbotyra, 6(2014).

Morfologijos iššūkiais:

Saitažymė angl. ***Hashtag*** (metainformacija)

Socialiniuose tinkluose ir tviteryje sutartiniu ženklu, dažniausiai grotelėmis (#), žymimas žodis ar žodžių junginys kaip kokios nors temos nuoroda (*http://naujazodziai.lki.lt/Index.asp?zodis=saitažodis*)

Saitažymė = Saitažodis = Grotžymė = Grotažymė

… tam tikros temos pranešimams identifikuoti sukuriant saitą į juos. (*Aiškinamasis kompiuterijos žodynas - rastija.lt*)

#TyrėjųNaktis2016MRU

#vgtu4u

#ManNeDzin

#SpreadPatinka

Morfologijos iššūkiais:
Saitažodis angl. *Hashtag* (morfologijos reiškinys)

Tokiu būdu galima generuoti begalę naujų žodžių, kurie pradeda gyventi savarankišką gyvenimą sakinyje.

- Kauniečiai – susitikime ketvirtadienį #VDU apskritojo stalo diskusijai. Tema - „Vytauto Didžiojo universitetas Kauno regione“. *(soc. Tinklai)*
- Vilniaus Simono Daukanto progimnazijos moksleiviai siunčia jums sveikinimą iš Specialiųjų tyrimų tarnybos (#STT). #Sveikinimosidiena prisiminkite savo artimuosius ☺ *(soc. Tinklai)*
- Po pasirodymo suksiu mėgstamiausias #Cumbia ir #Salsa plokšteles. Liko vienas tik vienas kitas bilietas kam reiktu #Pasipatinkinti *(soc. Tinklai)*

- Kai kurie „Twitter“ vartotojai pastebi, kad šis iššūkis turi akivaizdžių trūkumų, ir pažymi, kad nors juo siekiama sustabdyti pajuokas iš nepatrauklios išvaizdos, #DontJudgeChallenge tik dar labiau sustiprina egzistuojančius grožio stereotipus. (*iš Delfi straipsnio*)

Vytauto Didžiojo universitetas

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII