

Ana Zwitter Vitez in Špela Arhar Holdt:

5 Govor in spletna slovenščina

5.1 IZROČKI – govor in spletna slovenščina



Govor in spletna slovenščina

Ana Zwitter Vitez
FHŠ Univerza na Primorskem, FF Univerza v Ljubljani
Špela Arhar Holdt
FF Univerza v Ljubljani, ZUS Trojina

Ljubljana, 27. 8. 2015





- interdisciplinarno področje
- neraziskani rezultati
- pomembne družbene posledice

O čem bomo govorili

- specifikke govornega jezika
- odnos med govorom, pisno in spletno slovenščino
- kaj lahko besedilo pove o značilnostih avtorja

Od jutra do večera ...







1. stereotip: pišemo standardno, govorimo nestandardno



tinka110
09.02.12
08:16

Spoštovani, knjižni jezik je eno, pogovorni pa drugo.

Knjižni jezik se uporablja v pisani besedi (vam kakšna ura učenja ne bi škodila), pogovorni pa v izreki.

 Odgovori
  Citiraj sporočilo
  Povezava

 Na vrh





gos

☒ Iskanje po pogovornem zapisu
 ☐ Iskanje po standardiziranem zapisu

Uporabljaš [enostavno iskanje](#)
[Napredno iskanje](#)

-obseg:
milijon besed (112 ur posnetkov)

-vzorec:
demografsko reprezentativen vzorec govorcev v najpogostejših govornih situacijah

-sestava:
60% javni diskurz: mediji, predavanja, šolski pouk
40% nejavni diskurz: zasebni pogovori, delovni sestanki

JANES (••	 
<u>a velikokrat ne ostane le pri tem kar smo se lahko prepričali tudi med našim snemanjem</u>	Antigona se je kljub temu odločila, da Polinejka pokoplje.
<u>ma tu si lahko duol čez rit potjgneš</u>	igor-za tekmo bo pa jutri pomoje odloču tomaž hahaha +1 -0 18.08.2015 ob 08:52 Prijavi neprimerno vsebino

JANES (••
 


Pišemo standardno, govorimo pa nestandardno?










www.clipartof.com · 1046439

Pa vendar: specifike govora

Sprotno tvorjenje besedil

[eem eem eee bom eee hitro povedal eee iz eee znam nardit a masažu](#)

Izgovorjava

[tu je blo ko truebla brez konca](#)

Interakcija s sogovorcem

[ja hvala lepa jst eee eee \[1 delno\] se seveda **strinjam** eee v veliki meri](#)

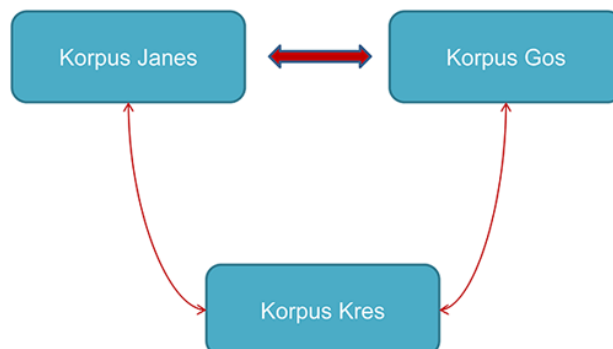
2. stereotip: na spletu pišemo, kot govorimo



Preseganje stereotipov: empirične raziskave



Preseganje stereotipov: empirične raziskave



Kako primerjamo korpuse med seboj?

>> najbolj tipične besede enega korpusa

Ključne besede:
- v opazovanem korpusu so pogoste,
- v primerjanem korpusu ni veliko.

NoSketch Engine

uporabnik: Janes korpus: JANES v0.3.1 Forum podkorpus: Forum L2,3

Iskanje
Seznam
O korpusu
?

Shrani
Spremeni
Nastavitve

Seznam besed

Korpus: JANES v0.3.1 Forum
Podkorpus: Forum L2,3

Referenčni korpus: KRES (uravnoteženi)
[Spremeni cili in referenčni \(pod\)korpus](#)

Stran 1 | Pojdi | [Naslednja >](#)

word	JANES v0.3.1 Forum : Forum L2,3		KRES (uravnoteženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
lp	18.374	749.3	249	2.1	244.6
feltno	4.601	187.7	8	0.1	177.0
lp	9.247	397.5	195	1.6	152.1
ops	3.254	144.9	14	0.1	130.7
LP	15.595	635.9	563	4.7	112.3
com	4.399	179.4	84	0.7	106.3
felne	1.902	77.6	0	0.0	78.6
xenon	1.941	79.2	6	0.0	76.3
sm	31.049	1266.1	2.033	16.9	70.9
oglasnik	1.770	72.2	12	0.1	66.5
tuning	1.250	79.5	32	0.3	63.6
jst	11.729	480.3	802	6.7	62.8
felge	1.512	61.7	2	0.0	61.6
mam	11.604	473.2	829	6.9	60.2
original	5.093	207.7	310	2.6	58.4

Metapodatki v korpusu Gos

gos | Iskanje | Seznam

Iskanje po go | Iskanje | Seznam

Iskanje po pogovornem zapisu | Iskanje po standardiziranem zapisu

Uporabljaj [reket](#) | Najdi

Uporabljaj [iskanje po seznamih](#) | Podrobnosti

Osnovne oblike
reči (1.278)

Tip govora
Neznan zapis
Javni inform
Javni razved
Neznan nez
Vse

Kanal
Osetni srb
Zdravila
Zdravila
Bado (180)
Vse

Standardizirana osnovna oblika
reči (1.278)

Besedna vrsta
glasol (1.278)

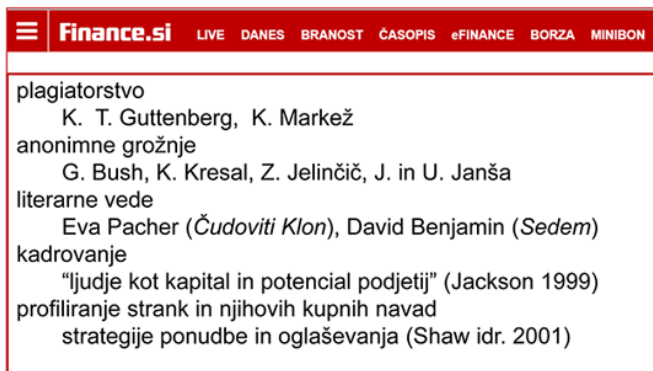
Tip govora
Neznan zapis (576)
Javni informativno-izobraževalni (328)
Javni razvedrilni (213)
Neznan nez (161)
Vse

Prikazujem 1-20 od 30 zadetkov (0.31 sekund)

☐ loči besede glede na oblikoslovne lastnosti

Pogovorni zapis	Standardizirana oblika	Število pojavitev
reku	rekel	885
reko	rekel	192
rekel	rekel	83
rek	rekel	27
reku	rekel	23
riku	rekel	18

3. stereotip: uporabnost jezikoslovnih raziskav



Finance.si LIVE DANES BRANOST ČASOPIS eFINANCE BORZA MINIBON

- plagiatorstvo
K. T. Guttenberg, K. Markež
- anonimne grožnje
G. Bush, K. Kresal, Z. Jelinčič, J. in U. Janša
- literarne vede
Eva Pachter (*Čudoviti Klon*), David Benjamin (*Sedem*)
- kadrovanje
"ljudje kot kapital in potencial podjetij" (Jackson 1999)
- profiliranje strank in njihovih kupnih navad
strategije ponudbe in oglaševanja (Shaw idr. 2001)

Ugotavljanje avtorstva besedil: "Volivci v trenirkah"

uradna spletna stran parlamentarne stranke

psevdonim

vznemirjena javnost:

1. "volivci s tujim naglasom"
2. "volivci v športnih oblačilih (trenirkah)"
3. "s kemičnim svinčnikom napisana številka, ki jo morajo obkrožiti na glasovnici"

Hipoteza in zajem besedil

Če je avtor besedilo anonimno objavil na uradni spletni strani stranke, je verjetno na isti spletni strani objavil še kakšno besedilo pod drugim/pravim imenom.

Kriterij za zajem

- tri mesece pred in tri mesece po objavi spornega besedila

Korpus

- 75 besedil 21 avtorjev
- 55.000 pojavnic (650 do 9000 pojavnic na avtorja).

Priprava besedil

pretvorba v format .txt,
anonimizacija besedil in tvorjenje glav dokumentov
oblikoslovno označevanje (Grčar idr. 2012)

Oznaka: Ppmeid
Kratka oznaka: Pp
Lema: relativen

viča sta dvignila umetno ustvarjen strah in slovenska radodarnost z državljanstvi

Relativni zmagovalec predčasnih volitev Zoran Janković je na volitvah prejel dobrih 300.000 glasov.

Vsi komentatorji so se veliko ukvarjali s strukturo tega volilnega telesa, ob tem pa so se v celoti namerno izognili analizi strukture njegovih volivcev vidika novih državljanov.

Izračun značilke

Leksikalne značilke:

- raznolikost besedišča (*lexical density*)
- Brunetova formula (Brunet 1988)
- hapax legomena (Holmes 1992)
- Honoréjeva statistika (Honoré 1979)

Berljivostne značilke:

- Flesh-Kincaid
- Coleman-Liau
- Automated Readability Index
- Gunning Fog

<http://www.usingenglish.com/members/text-analysis/help/readability.html>

http://nl.ijs.si/isit14/proceedings/isit2014_23.pdf

Rezultati analize

Majer: leksikalne značilke

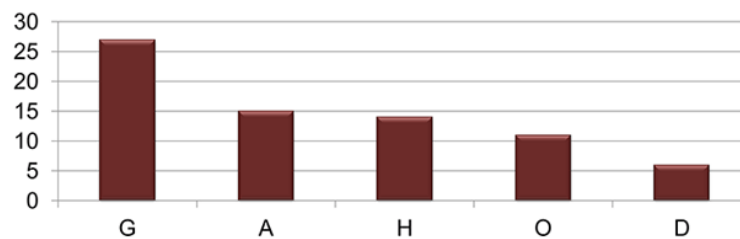
Avtor	DRB	BI	HS	HL	DRL	BI-L	HS-L	HL-L
G	0.02 ± 0.11	0.14 ± 0.61	70.30 ± 372.30	0.02 ± 0.11	0.02 ± 0.12	0.03 ± 0.95	77.01 ± 193.59	0.01 ± 0.11
A	0.05 ± 0.02	0.11 ± 0.04	56.17 ± 230.80	0.05 ± 0.01	0.06 ± 0.02	0.35 ± 0.02	95.76 ± 254.19	0.05 ± 0.00
H	0.06 ± 0.06	0.22 ± 0.27	15.38 ± 74.07	0.06 ± 0.06	0.06 ± 0.06	0.39 ± 0.44	12.16 ± 83.69	0.05 ± 0.05
O	0.06 ± 0.02	0.26 ± 0.18	420.85 ± 315.92	0.07 ± 0.03	0.06 ± 0.03	0.46 ± 0.33	197.93 ± 154.95	0.06 ± 0.03
N	0.10 ± 0.04	0.48 ± 0.13	1.36 ± 58.86	0.10 ± 0.04	0.09 ± 0.05	0.72 ± 0.29	53.15 ± 134.35	0.08 ± 0.04
D	0.07 ± 0.04	0.34 ± 0.18	196.48 ± 175.76	0.07 ± 0.04	0.10 ± 0.04	0.84 ± 0.30	316.35 ± 83.34	0.09 ± 0.04
P	0.10 ± 0.06	0.44 ± 0.30	183.11 ± 43.80	0.10 ± 0.05	0.10 ± 0.04	0.72 ± 0.32	167.55 ± 109.35	0.09 ± 0.03
M	0.10 ± 0.02	0.47 ± 0.05	242.78 ± 160.22	0.10 ± 0.02	0.12 ± 0.03	1.01 ± 0.13	378.17 ± 135.76	0.12 ± 0.02
J	0.12 ± 0.00	0.48 ± 0.00	38.12 ± 0.00	0.12 ± 0.00	0.15 ± 0.00	1.10 ± 0.00	91.45 ± 0.00	0.13 ± 0.00
B	0.13 ± 0.04	0.57 ± 0.30	443.25 ± 231.55	0.10 ± 0.02	0.16 ± 0.03	1.27 ± 0.34	26.28 ± 367.71	0.14 ± 0.00
E	0.17 ± 0.00	0.78 ± 0.00	434.37 ± 0.00	0.17 ± 0.00	0.17 ± 0.00	1.33 ± 0.00	471.62 ± 0.00	0.18 ± 0.00
F	0.15 ± 0.03	0.74 ± 0.31	279.47 ± 94.69	0.14 ± 0.03	0.17 ± 0.04	1.34 ± 0.48	36.19 ± 117.19	0.15 ± 0.04
V	0.17 ± 0.06	0.95 ± 0.63	68.63 ± 263.70	0.16 ± 0.05	0.17 ± 0.07	1.49 ± 0.86	53.65 ± 303.20	0.15 ± 0.05

Razlikovalna moč značilik

Avtor G

Značilika	Utež
Delež različnih besed	0.79 ± 0.27
Delež različnih lem	0.78 ± 0.26
Hapax legomena	0.77 ± 0.26
Hapax legomena (leme)	0.76 ± 0.26
Brunetov indeks (leme)	0.68 ± 0.30
Brunetov indeks	0.54 ± 0.33
Honoréjeva statistika	0.48 ± 0.27
Honoréjeva statistika (leme)	0.47 ± 0.28
Razmerje med št. besed in št. povedi	0.46 ± 0.27
Razmerje med št. znakov in št. besed	0.45 ± 0.26

Podobnost avtorjev z anonimnim besedilom



Avtorji z najmanjšo povprečno absolutno razliko glede na anonimno besedilo.

Omejitve raziskave

križno preverjanje
prisotnost dejanskega avtorja besedila

Odperta vprašanja

dolžnosti raziskovalca
družbene posledice

Prihodnost jezikovnih tehnologij

Povej mi, s kom se družiš, in povem ti, kdo si.
Povej mi, kaj bereš, in povem ti, kdo si.
Povej mi, kaj voziš, in povem ti, kdo si.
Povej mi, kaj ješ, in povem ti, kdo si.
Povej mi, kako se znojiš, in povem ti, kdo si.

Povej mi **kar koli** in povem ti, kdo si.

Kaj bomo delali?

Kako govorimo?

Ali na spletu pišemo, kot govorimo?

Kaj lahko na podlagi podatkov iz korpusov povemo o avtorju besedila?