Analiza diskurza in rudarjenje podatkov

Uporaba pri odkrivanju ideoloških razlik v medijskem poročanju

Senja Pollak, Roel Coesemans, Walter Daelemans, Nada Lavrač:

Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining

Pragmatics 21:4.647-683 (2011), International Pragmatics Association

Članek na osnovi magisterskega dela Senje Pollak na Univerzi v Antwerpnu, mentor W. Daelemans

Program P6-0215 – Slovenski jezik - bazične, kontrastivne in aplikativne raziskave (PI: Vojko Gorjanc)

Kontakt: Senja.Pollak@ff.uni-lj.si

Glavni doprinosi k znanosti

- Metodološki:

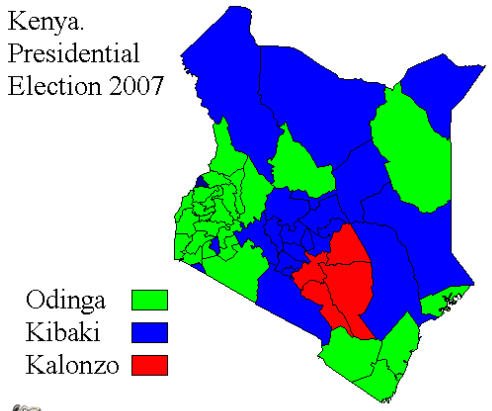
  Raziskava prispeva k povezovanju dveh tradicionalno ločenih disciplin, jezikovne pragmatike in rudarjenja podatkov, ki skupaj doprineseta k popolnejši analizi diskurza v medijih, natančneje k analizi razlik pri poročanju o istih vsebinah

- Vsebinski:

  Raziskava doprinese k analizi ideoloških razlik v člankih o kenijskih predsedniških in parlamentarnih volitvah leta 2007, objavljenih v kenijskih in zahodnih medijih

Kenijske volitve

- Predsedniške in parlamentarne volitve 27. decembra 2007
- Kenija: več kot 40 etničnih skupin, Kikuyu, Luo, Kalenjin, …
- Glavna kandidata:
  - Mwai KIBAKI, pripadnik Kikuyu, stranka PNU - predsednik od 2002
  - Raila ODINGA, pripadnik Luo, stranka ODM - kandidat opozicije
- Parlamentarne volitve: zmaga ODM
- Predsedniške volitve:
  - rezultati oznanjeni s trodnevno zamudo
  - zmaga Kibaki (?) in zapriseže
- Napetost narašča, sledi povolilna kriza
- Mediacija (K. Annan) in sporazum o delitvi moči



Predstavitev problema

- Medijske novice: izbor aktualnih dogodkov, selektivno poročanje, reprezentacija realnosti z določene ideološke pozicije
- Analiza medijskih novic
  - Pragmatika: študij kognitivnih, družbenih in kulturnih aspektov uporabe jezika, leksikalne in diskurzivne izbire ustvarjajo pomene in lahko služijo za ohranjanje in širjenje dominantnih ideologij (npr. Verschueren 1999)
- Analiza medijskih novic z rudarjenjem podatkov:
  - Avtomatsko ugotavljanje razlik in vzorcev v poročanju kenijskih in zahodnih medijev

Predstavitev korpusa

- Časopisni članki različnih medijskih hiš (v angleščini):
  - 4 zahodni časopisi (britanski in ameriški): **WE**

    The Independent, the New York Times, The Times, The Washington Post
  - 2 lokalna (kenijska) časopisa: **LO**

    Daily Nation, The Standard
- Analiza obravnava 464 člankov (232 zahodnih in 232 lokalnih)
- Ista tema: kenijske volitve in povolilna kriza
- Isto obdobje: konec decembra 2007 – konec februarja 2008

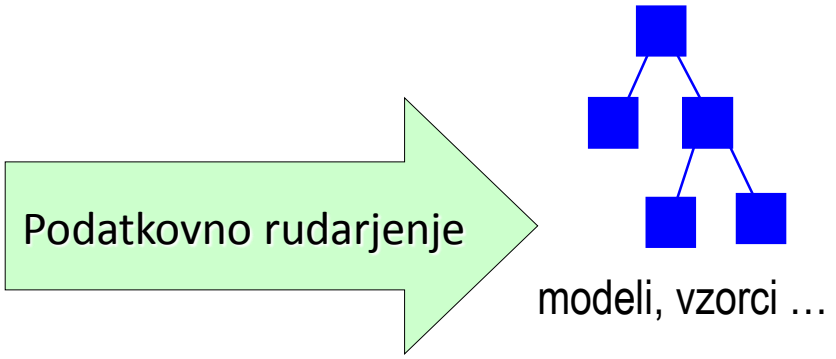
Korpusna analiza

Analiza z orodjem WordSmith

	Local (LO)					Western (WE)				
N	Keyword	Keyness	Freq.	%	Disp.	Keyword	Keyness	Freq.	%	Disp.
1	ODM	10252,93	795	0,51	0,92	KIBAKI	10478,96	827	0,50	0,90
2	KIBAKI	8295,84	651	0,42	0,89	KENYA	9500,88	943	0,57	0,94
3	KENYA	5732,61	602	0,38	0,93	ODINGA	9362,28	755	0,46	0,89
4	RAILA	5409,67	420	0,27	0,79	KENYA'S	6091,10	513	0,31	0,90
5	ANNAN	4696,75	402	0,26	0,79	KIKUYU	5542,06	451	0,27	0,83
6	MR	4256,45	1319	0,84	0,96	ELECTION	4300,16	757	0,46	0,91
7	ODINGA	3733,38	308	0,20	0,87	KENYAN	3672,08	341	0,21	0,95
8	PRESIDENT	3661,46	754	0,48	0,89	OPPOSITION	3298,01	608	0,37	0,87
9	KENYANS	3353,50	273	0,17	0,89	MR	3213,25	1128	0,68	0,80
10	PNU	3215,23	249	0,16	0,77	VIOLENCE	3190,29	527	0,32	0,89
11	VIOLENCE	3179,73	519	0,33	0,86	KIKUYUS	3125,44	244	0,15	0,83
12	ELECTION	2498,46	500	0,32	0,90	RAILA	3086,84	242	0,15	0,95
13	NAIROBI	2495,77	247	0,16	0,91	NAIROBI	3067,93	299	0,18	0,92
14	SAID	2399,71	1510	0,96	0,96	KIBAKI'S	2907,66	227	0,14	0,86
15	MEDIATION	2264,43	232	0,15	0,84	ETHNIC	2794,59	393	0,24	0,78
16	PRESIDENTIAL	2075,46	301	0,19	0,91	KENYANS	2644,00	219	0,13	0,87
17	TALKS	2041,88	390	0,25	0,77	MWAI	2438,22	194	0,12	0,91
18	CRISIS	1844,72	350	0,22	0,84	PRESIDENT	2404,46	569	0,35	0,92
19	ECK	1701,01	146	0,09	0,67	ANNAN	2377,22	216	0,13	0,82
20	KOFI	1674,52	134	0,09	0,83	SAID	2168,71	1475	0,89	0,93
21	GOVERNMENT	1567,30	676	0,43	0,87	LUO	2125,23	176	0,11	0,80
22	LEADERS	1494,84	318	0,20	0,90	ODINGA'S	2042,20	164	0,10	0,77
23	KENYAN	1345,02	139	0,09	0,88	KALENJIN	1580,49	125	0,08	0,80
24	KIVUITU	1291,16	100	0,06	0,55	TRIBE	1543,66	193	0,12	0,83
25	POLITICAL	1273,01	463	0,30	0,89	TRIBAL	1431,85	182	0,11	0,89
26	YESTERDAY	1271,49	383	0,24	0,92	RIFT	1393,17	164	0,10	0,77
27	POLICE	1194,51	427	0,27	0,80	LUOS	1320,77	104	0,06	0,75
28	KISUMU	1157,63	94	0,06	0,68	KOFI	1229,14	100	0,06	0,80
29	MPS	1119,01	195	0,12	0,82	LEADERS	1146,89	268	0,16	0,80
30	ELDORET	1116,15	89	0,06	0,88	VOTE	1071,38	254	0,15	0,82

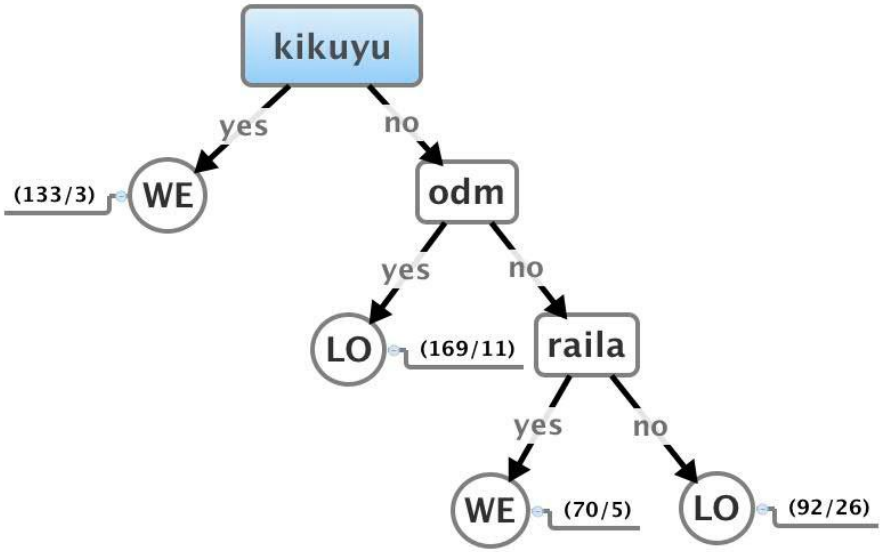
Rudarjenje podatkov

Oseba	Starost	Dioptrija	Astigmat.	Solzenje	Leče
O1	mlad	kratko	ne	zmanjšano	NE
O2	mlad	kratko	ne	normalno	MEHKE
O3	mlad	kratko	da	zmanjšano	NE
O4	mlad	kratko	da	normalno	TRDE
O5	mlad	daleko	ne	zmanjšano	NE
O6-O13	...	...	...	...	...
O14	pr_st_dal	daleko	ne	normalno	MEHKE
O15	pr_st_dal	daleko	da	zmanjšano	NE
O16	pr_st_dal	daleko	da	normalno	NE
O17	st_daleko	kratko	ne	zmanjšano	NE
O18	st_daleko	kratko	ne	normalno	NE
O19-O23	...	...	...	...	...
O24	st_daleko	daleko	da	normalno	NE



Podatki: tabela podatkov (primeri, atributi, razred)
Rezultat: klasifikacijski modeli (odločitvena drevesa, pravila, …), nabor zanimivih vzorcev
Uporaba zgrajenih modelov: klasificiranje novih primerov, odkrivanje novega znanja v podatkih, razumevanje domene, etc.

Identifikacija razlik z učenjem klasifikacijskih modelov in iskanjem kontrastnih besed



	Klasifik. točnost (stand. dev.)		
Preds.	**J48**	**Jrip**	**PRISM**
W1	89.00 (4.64)	89.22 (3.95)	83.61 (6.34)
W2	89.46 (4.17)	90.53 (3.81)	82.53 (4.41)
W3	89.67 (6.03)	90.96 (6.34)	83.86 (5.58)

Uporaba orodja za rudarjenje podatkov Weka

LOCAL	odm, mp, team, mr, pnu, odm_leader, president_kibaki, dr, media, statement
WESTERN	kikuyu, mr_kibaki, opposition, mr_odinga, luo, tribe, tribalism, opposition leader, odinga, ethnic

Uporaba podpornih vektorjev (SVM) v orodju OntoGen

Vsebinski zaključki

- **Zahodni mediji:** interpretacija dogodkov iz etnične perspektive
  - eksplicitno navajanje etničnih pripadnosti akterjev
  - pogosta uporaba ideološko zaznamovanih besed *pleme, plemenski, etnični, etnična skupina,* … v negativnih besedilnih kontekstih (*tribal violence, ethnical conflict, violence, fighting, tensions*)
- **Kenijski mediji:** socio-politični okvir interpretacije, izogibanje etničnim vidikom
  - politične interpretacije (povolilna kriza, povezovanje akterjev s strankami)
  - npr. nasilneži: ali brez specifikacije (*youths, gangs, mobs, protesters or criminals*) ali povezani s političnimi strankami

Metodološki zaključki in nadaljnje delo

- Rudarjenje besedil je omejeno z nerazumevanjem konteksta, jezikovna pragmatika omogoča razumevanje pomena odkritih razlik v medijskem poročanju
- Pragmatična analiza s tekstovnim rudarjenjem:
  - omogoči hiter vpogled v velike korpuse
  - omogoči postavitev hipotez na podlagi zgrajenih modelov, avtomatsko identificiranih kontrastnih besed in vzorcev
- Združitev kvantitativnih (rudarjenje podatkov) in kvalitativnih (analiza diskurza) pristopov omogoča popolnejšo analizo uporabe jezika v medijih
- V nadaljnjem delu smo uporabili metode za iskanje izjem v podatkih za odkrivanje nenavadnih člankov (gostujoči avtorji, netipične vsebine člankov, LREC 2012)