Elektronski in spletni naslovi v časopisu DELO, 1998-2000

Primož Jakopin

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 4, 1000 Ljubljana, Slovenija
primoz.jakopin@uni-lj.si

Abstract

In the paper a statistical analysis of e-mail addresses and web page addresses (URLs), as they appear in the electronic edition of the leading Slovenian daily newspaper, from January 1998 to June 2000, is given. The DELO text collection is part of the BESEDA corpus at the Fran Ramovš Institute of Slovenian Language ZRC SAZU (http://bos.zrc-sazu.si) and contains 42 million words (8/2000) among which 2.500 such units were identified (1.940 different ones).

1. Uvod

Namen prispevka je kvantitativno osvetliti elektronske in spletne naslove v enem od množičnih občil. DELO je namenjeno najširšemu krogu bralcev in je zato v njegovem besedilu vsaj implicitno skrit odgovor, koliko je povprečen prebivalec naše dežele izpostavljen internetnim podatkom, med katerimi so najbolj otipljivi naslovi spletnih strani in elektronski poštni naslovi posameznikov in ustanov ter podjetij.

Teh naslovov v časopisu ni tako veliko, kot bi domneval kdo, ki se ukvarja z jezikovnimi tehnologijami in se pri svojem delu vsak dan sreča vsaj z nekaj deset takimi naslovi, še vedno pa toliko, da utegnejo biti rezultati obdelave zanimivi.

2. Gradivo

Prispevek se nanaša na gradivo iz dnevnika DELO (od ponedeljka do sobote), del katerega, kak dan več, kak dan manj, navadno približno dve tretjini, je v elektronski obliki po elektronski pošti dostopen predvsem slepim in slabovidnim, poleg tega pa še nekaterim raziskovalcem besedilnih korpusov, med drugim tudi piscu teh vrstic.

To časopisno gradivo je del besedilnega internetnega korpusa BESEDA, ki je v obliki konkordančnika in slovarja besednih oblik s frekvencami prosto dostopen na Inštitutu za slovenski jezik, preko spletne strani: http://bos.zrc-sazu.si (Jakopin, 2000). Količina v prispevku zajetega gradiva je navedena v tabeli 1.

Za leto 2000 je zajeto gradivo od januarja do konca junija, letnika 1998 in 1999 pa sta zajeta v celoti. Števila povedi in besed so približna (stanje 22. 7. 2000): vsaka nova obdelava gradiva odkrije kaj takega, kar je treba popraviti in se potem spremenijo tudi sumarni podatki, vendar ne bistveno. Leto 2000 je po količini skromnejše od prvih polovic prejšnjih let predvsem zaradi tega, ker je letos žal izpadlo precej sobotnih prilog.

3. Identifikacija elektronskih in spletnih naslovov

Kako obravnavani skupini naslovov poiskati v časopisnem besedilu je le na prvi pogled videti preprosto; naslovi spletnih strani npr. imajo v splošnem obliko **način_dostopa://[uporabnik@]računalnik [:vrata]/[imenik[/imenik]/datoteka]** (Peterlin, 2000). Dejansko se naslovi spletnih strani v opazovanem gradivu skoraj vedno začnejo z nizom: *http://* (le v enem primeru s *https://* ali *gopher:*, z *ftp:* tudi v le treh primerih), elektronski poštni naslovi pa vsebujejo *at-znak* (*@*, pogovorno tudi *afna*, *viseča opica*, *a v srajčki* in podobno). Žal navedeno velja le za zadnje naslove, spletni naslovi pa so dostikrat navedeni samo delno, brez niza *http://* spredaj ali, kadar jih je v povedi več, samo del, ki je različen od zadnjega prej zapisanega naslova. Tako je treba, če naj bo raziskava kolikor toliko verodostojna, najprej zajeti vse nize znakov, ki vsebujejo zaporedji "<x>/<x>" oziroma "<x>.<x>" , kjer je <x> črka angleške abecede ali številka ali tilda (~ za /). Nizi smejo vsebovati poleg črk angleške abecede in številk tudi nekatera ločila in posebni znaki: . / _ - + ~ % = ? # @, spredaj in zadaj pa morajo biti omejeni z ločilom, ki ni v zadnjem seznamu (končno piko za začetek tudi še pustimo pri miru).

Izkaže se, da je takih nizov, potencialnih elektronskih in spletnih naslovov, v obravnavanem gradivu veliko, kar 12.902. Univerzalnega algoritma, s katerim bi bilo mogoče med njimi iskane naslove strojno poiskati, avtor žal ni našel. Seznam je premetal ročno, pri čemer si je bilo za določitev opazovanega niza dostikrat treba pomagati s kontekstom. V abecedno urejenem seznamu so sicer res bile večje sklenjene skupine naslovov, npr. vseh, ki se začnejo na *http://* ali *www*, in nizov, ki naslovi očitno niso (najpogostejši so navedeni v tabeli 2), bilo je pa tudi veliko napak (predvsem nepotrebnih presledkov), ki so bile popravljene.

Na koncu sta iz prvotnega nastala dva seznama: prvi z 2.509 ustreznimi nizi in drugi z 10.393 napačnimi, predvsem kraticami. 10 najpogostejših neustreznih nizov s frekvencami in razlago je navedenih v tabeli 2.

Tabela 1: Obseg obdelanega vzorca

	1998	1999	2000	Skupaj
dni	298	292	149	739
strani	4249	4205	1848	10302
povedi	924,633	882,694	392,168	2.199,495
besed	16.692,625	15.939,021	7.130,162	39.761,808

Tabela 2: Najpogostejši nepravi nizi, ki vsebujejo znaka / ali . s frekvencami

1.	t.m.	2210	tega meseca
2.	km/h	618	kilometrov na uro
3.	d.o.o.	377	družba z omejeno odgovornostjo
4.	d.d.	292	delniška družba
5.	STA/AFP	236	Slovenska tisk. agencija/Agence France Presse
6.	t.i.	215	tako imenovani
7.	STA/DPA	203	Slovenska tisk. agencija/Deutsche Presse Agentur
8.	in/ali	147	-
9.	m/a	122	metrov z avtoštartom (pri konjskih dirkah)
10.	STA/AP	117	Slovenska tiskovna agencija/Associated Press

V prvem seznamu prevladujejo pravi spletni naslovi (http://, 1286) ali nebistveno skrajšani (ki se začnejo na www, 500) in elektronski poštni naslovi (96), je pa avtor vanj vključil tudi precej drugih imen, v nadaljnjem bodo skupaj z naslovi označena kot pikasta imena, ki imajo zelo jasno povezavo z internetom. Na zanimiv način označujejo naš čas in bi jih bilo škoda odvreči. Lepša oznaka bi bila morda imena s pikami, a je dolga in okorna, izraz pikčast pa v nasprotju s pikast pomeni nekoga ali nekaj, kar je bolj posejano s pikami kot pikast (SSKJ, 1994).

4. Pikasta imena

Skupaj jih je bilo 2.509, od tega 1.944 različnih.V tabeli 3 in na sliki 1 je navedena njihova porazdelitev po letih in mesecih. Kot je bilo že omenjeno v uvodu, je gradiva v letu 2000 po mesecih nekoliko manj kot v prejšnjih letih in zato tudi trend naraščanja števila teh imen s časom ni jasno razpoznaven. Očiten pa je vrh ob koncu poletja in zgodaj jeseni lanskega leta.

Tabela 3: Porazdelitev pikastih imen po letih in mesecih

	1998	1999	2000
januar	70	105	94
februar	94	92	58
marec	89	111	80
april	100	72	96
maj	78	71	98
junij	48	114	81
julij	65	53	-
avgust	86	123	-
september	53	126	-
oktober	45	110	-
november	62	100	-
december	52	83	-
Skupaj	842	1160	507

Slika 1: Histogram porazdelitve po mesecih, 1998-2000, 100% je celota

Oglejmo si še pogostejša pikasta imena. Najdemo jih v tabeli 4: v njej so navedene njihove leme; marsikatero tako ime, še posebej pa najpogostejše, Si.mobil, je bilo poleg samostalniške vloge deležno že tudi pridevniške (npr. *V zaključnem delu Si.mobil lige v Tivoliju bodo igrali ...*). Imena s piko ali at-znakom poleg podjetij, kot je recimo *Amazon.com*, označujejo tudi dogodke, npr. *offline@online, festival sodobne elektronske umetnosti*, projekte, npr. *b.ALT.ica* Moderne galerije ali osebe, npr. *španski umetnik Marcel.li (Antunez Roca)*.

V tabeli 5 so navedeni najpogostejši sestavni deli pikastih imen, vsi ki so dosegli frekvenco vsaj 10. Pri izdelavi tabele so bile velike črke spremenjene v male, odvrženi pa vsi posebni znaki razen podčrtaja in pomišljaja.

Tabela 4: Pogostejših 23 pikastih imen s frekvencami

Si.mobil	145	S5.net	7
Amazon.com	68	MP3.com	6
malar@delo.si	43	SI.CERT	6
http://www.mzt.si/mzt/novo.html	19	Slowwwenia.com	6
http://www.yahoo.com	14	Amis.net	5
Beograd.com	13	EM.TV	5
b.ALT.ica	11	http://mail.yahoo.com	5
http://www.eyebeam.org	11	http://www.altavista.com	5
Marcel.li	8	http://www.ijp.si/DMFA/Kolokviji	5
http://www.amazon.com	7	http://www.mzt.si	5
K2.net	7	www.ris.org	5
SRC.SI	7		

Tabela 5: Pogostejših 63 sestavnih delov pikastih imen s frekvencami

www	1492	edu	35	beograd	14
http	1276	sigov	33	news	14
com	1035	geocities	29	at	13
si	754	uk	29	k2	13
html	324	novo	27	mss	13
org	189	slo	27	slowwwenia	13
htm	146	home	26	altavista	12
net	140	co	25	aol	12
mobil	131	ljudmila	24	b	12
amazon	83	doc	21	kiss	12
index	71	members	19	ris	12
mzt	66	mp3	19	amis	11
delo	51	microsoft	17	eyebeam	11
uni-lj	48	telekom	17	it	11
arnes	45	asp	16	shtml	11
www2	45	alt	15	cgi-bin	10
ijs	43	edus	15	eu	10
malar	43	mobitel	15	go	10
de	38	siol	15	si21	10
gov	36	s5	15	to	10
yahoo	36	zdnet	15	zrc-sazu	10

Tabela 6: Končnice prvega dela spletnih naslovov s frekvencami

com	815	it	8	dk	2	ok	1
si	492	au	7	fr	2	ot	1
org	171	nl	6	il	2	ru	1
net	101	hr	5	is	2	sg	1
edu	34	lu	5	jp	2	sh	1
de	26	nu	5	pk	2	sk	1
uk	25	be	4	se	2	tc	1
gov	20	ca	4	cz	1	tp	1
at	13	hu	4	ee	1	us	1
to	10	id	3	ie	1	va	1
ch	8	za	3	ke	1		
int	8	co	2	no	1		

V zadnji tabeli, tabeli 6, so prikazane vse končnice (46) prvih delov spletnih naslovov - kadar ima jedro naslova vsaj eno poševno črto, je to zadnje ime pred njo, kadar je nima, pa zadnje ime v naslovu.

5. Sklep

Analiza imen, ki jih vsak dan srečujemo vsi, ki se tako ali drugače ukvarjamo z računalniki, je odgovorila na nekaj vprašanj in odprla nekaj novih. Nedvomno bi bili zanimivi tudi rezultati raziskave, ki bi zajela vsa tovrstna imena, ki v določenem časovnem obdobju, npr. mesecu dni, potujejo preko večjega vozliščnega računalnika na slovenskem delu svetovnega spleta. V prihodnosti pa bi bila nujna tudi izdelava algoritma za prepoznavanje spletnih naslovov, elektronskih poštnih naslovov in drugih pikastih imen. Pri slednjih utegne biti naloga, kot pri osebnih imenih, vsako leto težje rešljiva.

6. Viri

Jakopin, Primož, 2000. BESEDA - a text corpus of Slovenian. *Digital Resources for the Humanities*, Conference Abstracts, University of Sheffield, Sept. 2000, str. 70-72.

Peterlin, Primož, 2000. Re: Pocitniski pozdrav in vabilo. Elektronsko sporočilo v okviru konference *Slovenska literarna veda <slovlit.ijs.si>*, 28. avgusta 2000, ob 10.41.16.

SSKJ, 1994. Slovar slovenskega knjižnega jezika. DZS, Ljubljana 1994.