Digitalne biblioteke kao potencijalni lingvistički resurs - stanje u Srbiji

Cvetana Krstev

Digitalne biblioteke u Srbiji

Prema (Arms 2000), neformalna definicija digitalne biblioteke je kontrolisana, tj. sistematski organizovana kolekcija informacija sa pridruženim servisima, pri čemu su informacije uskladištene u digitalnom formatu i može im se pristupiti preko mreže.

Može se postaviti pitanje otkuda tolika popularnost digitalnih biblioteka i koja je njihova suštinska prednost i novina u odnosu na tradicionalne biblioteke. Prema istom autoru, opšte je verovanje da digitalne biblioteke bolje obavljaju one funkcije biblioteke koje se odnose na isporuku informacija, a to se postiže na sledeći način:

- Digitalne biblioteke dovode biblioteku korisniku, dok, da bi koristio tradicionalnu biblioteku korisnik mora da ide u nju.
- Informacije u digitalnim bibliotekama mogu da se dele. Naime, mnoge informacije u bibliotekama i arhivama su jedinstvene, a jednom kad se postave na mrežu one postaju dostupne svima.
- Informacije se lakše osvežavaju, što je veoma važno za one vrste informacija koje su podložne stalnim promenama.
- Informacije su uvek dostupne, jer se vrata digitalne biblioteke zahvaljujući pristupu preko mreže nikad ne zatvaraju.
- U digitalne biblioteke se pohranjuju novi oblici informacija koje nije bilo moguće skladištiti u tradicionalne biblioteke, ili to nije bilo uobičajeno. Na primer, statistički podaci o popisu stanovništva se mogu smestiti u baze podataka čime oni postaju dostuni širokom krugu ljudi.
- Informacije se mogu pregledati i pretraživati. Uprkos mnoštvu sekundarnih i referensnih alata koji su razvijeni za pristup tradicionalnim dokumentima, pronalaženje željene informacije u njima je pravi izazov.

Premda ima autora koji nastanak digitalnih biblioteka vezuju za početke automatizacije biblioteka nastankom mašinski čitljivih kataloga (Neavill 2004), prava suština digitalnih biblioteka je u samim informacijama, a ne u meta-informacijama ili referensnim informacijama, koje su u digitalnom obliku. Od trenutka ekspanzije digitalnih biblioteka, krajem 80-tih i početkom 90-tih godina prošlog veka (Lesk 1997), za šta je najzaslužniji nastanak veba, pokrenuto je mnogo projekata izgradnje digitalnih biblioteka, ili digitalnih kolekcija koje pretenduju na taj naziv. Neki od značajnijih projekata pomenuti su u (Krstev 2002).

Svakako najznačajniji projekat razvoja digitalne biblioteke pokrenut u Evropi je projekat TEL - Evropska biblioteka (The European Library)[1] koji je tokom perioda 2001-2004. vodila i realizovala Konferencija direktora nacionalnih biblioteka Evrope CENL (Conference of European National Libraries). U ovom periodu projekat je realizovan kao prateći program u okviru Petog okvirnog programa za razvoj informacionog društva i tehnologija Evropske unije pa ga je, u skladu s tim, finansirala Evropska komisija, ali i same članice CENL-a. Počev od 2005. godine projekat finansiraju isključivo članice CENL-a. Osnovni ciljevi ovog projekta su izgradnja zajedničkog elektronskog kataloga evropskih nacionalnih biblioteka, postavljanje i dostupnost svih njihovih digitalnih zbirki, organizovanje virtualne izložbe koja bi predstavljala evropsku kulturnu baštinu kao i izgradnja odgovarajućeg portala preko koga bi sve ove informacije i funkcije bile dostupne.

Narodna biblioteka Srbije je postala članica CENL-a 2003. godine, a počev od jula 2005. godine ona postaje i punopravni partner u TEL projektu (Injac 2005), što je bio znak da je ispunila niz tehničkih i organizacionih uslova. Od trenutka priključenja ovom projektu, Narodna biblioteka Srbije je postavila na portal TEL deset kolekcija od kojih su pet pretraživih, i to direktno preko portala TEL-a, dok se ostalih pet mogu pregledati preko veza koje direktno vode na mrežnu stanicu Narodne biblioteke Srbije. Jedna od pet pretraživih kolekcija je zapravo uzajamni katalog COBISS.SR koji trenutno ima preko 1,9 milion zapisa. Ostale četiri pretražive kolekcije su *Srpska dečja digitalna biblioteka* koja predstavlja jedinstvenu digitalnu kolekciju 130 dečijih knjiga, *DOI Serbia* koja sadrži bazu podataka sa člancima punog teksta više naučnih časopisa iz Srbije uglavnom iz oblasti prirodnih i biomedicinskih nauka[2], *Svetogorska grafika* koja predstavlja zbirku digitalizovanih kopija grafika iz svetogorskih manastira i *Zbirka pozorišnih plakata* koja predstavlja digitalne slike pozorišnih plakata štampanih na teritoriji Srbije i Kraljevine Jugoslavije od sredine XIX veka do 1945. godine. Sve ove kolekcije su bibliografski obrađene u Narodnoj biblioteci Srbije i OAI su kompatibilne[3].

Digitalne kolekcije uključene u TEL koje su dostupne za pregledanje su: *Politika (1904-1941)* koja sadrži digitalne slike stranica svih brojeva dnevnog lista Politika od prvog broja koji je izašao 12. januara 1904. godine do poslednjeg predratnog broja 6. aprila 1941. godine, *Tekuća bibliografija Srbije* koja sadrži bibliografski popis svih vrsta građe i publikacija koje se tokom godine izdaju u Republici Srbiji, *Katalog knjiga na jezicima jugoslovenskih*

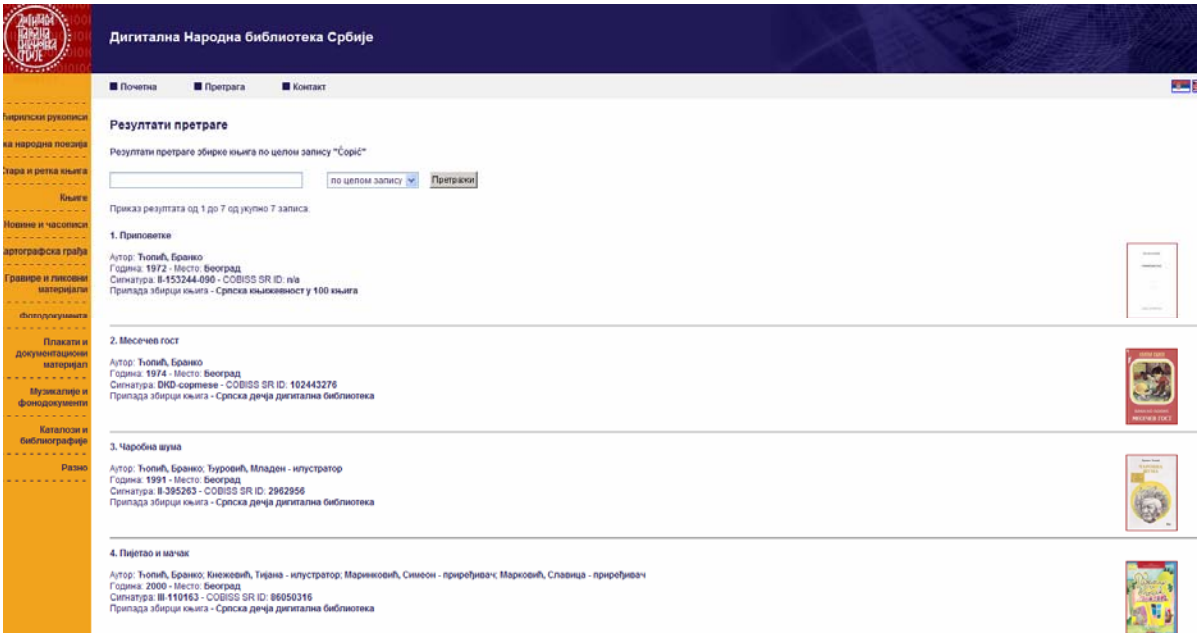
[1] TEL – The European Library (www.theeuropeanlibrary.org)

[2] Naziv ove kolekcije potiče od primene standarda DOI – *Digital Object Identifier* koji predstavlja sistem dodele međunarodnih standardnih brojeva za digitalne dokumente.

[3] OAI – Open Archive Initiative se odnosi na protokol koji definiše mehanizme za ubiranje iz repozitorijuma meta-podataka koji su u XML formatu (http://www.openarchives.org).

naroda (1868-1972) i *Katalog knjiga na jezicima jugoslovenskih naroda (1519-1867)* koji sadrže digitalne kopije odgovarajućih štampanih kataloga i predstavljaju zapravno "slikovni katalog" i, konačno, *Srpska bibliografija. Knjige. 1868-1944* koja sadrži digitalne kopije srpske retrospektivne bibliografije za knjige.

Treba napomenuti da je samo deo kolekcija Digitalne biblioteke Narodne biblioteke Srbije koja danas sadrži oko 600.000 dokumenata dostupan preko TEL portala. Celokupnoj kolekciji se može pristupiti direktno preko portala Digitalne biblioteke[4]. Ova kolekcija sadrži još niz interesantnih zbirki, kao što su *Digitalna zbirka ploča na 78 obrtaja*, *Beograd na starim razglednicama*, *Beograd na starim mapama*, i mnoge druge. Gotovo sve ove kolekcije su pretražive, čak i one za koje ta mogućnost nije ponuđena na TEL portalu. Pretraživost znači da se željena stavka može pronaći u kolekciji preko svog katalošког opisa i, što je najvažnije, njoj se direktno može pristupiti i ona se može pregledati koristeći vrlo dobro dizajnirano sučelje. Na slici 1 su prikazani rezultati pretraživanja kolekcije knjiga ključnom reči „Ćopić". Radi se zapravo o pretraživanju svih polja katalošких zapisa kolekcije zadatom ključnom reči, što kao rezultat daje sedam stavki. Svaki ponuđeni odgovor je propraćen skraćenim katalošким opisom i slikom naslovne strane koja predstavlja početak hipertekstualne veze ka digitalnoj kopiji dela koja se može dalje pregledati.



Slika 1. Pretraživanje kolekcije knjiga u Digitalnoj biblioteci Narodne biblioteke Srbije

Ako uporedimo Digitalnu biblioteku Narodne biblioteke Srbije sa osnovnih sedam funkcija ili karakteristika digitalnih biblioteka koje su date na početku

[4] http://digital.nb.rs/

ovog rada vidimo da ih ona uglavnom sve poseduje i da, s toga, sa punim pravom nosi taj naziv. Jedinu nedoumicu izaziva poslednja od navedenih stavki koja se odnosi na pregledanje i pronalaženje. Autor (Arms 2001) ukazujući na karakteristike koje digitalne biblioteke razlikuju od tradicionalnih posebno ističe značajno unapređeno i olakšano pronalaženje željenih informacija. Međutim, u slučaju kada je pronalaženje u digitalnoj biblioteci svedeno na pretraživanje meta-informacija, a ne samih informacija, postavlja se pitanje da li je pronalaženje unapređeno u onoj meri u kojoj je to potrebno i, u krajnjoj liniji, tehnološki moguće. Kada se govori o pretraživanju samih informacija obično se misli na mogućnost pronalaženja u punom tekstu (engl. full-text retrieval), ali napredne metode pretraživanja se danas fokusiraju i na pronalaženje drugih vrsta informacija: pronalaženje u multimedijalnim dokumentima (Baeze-Yates & Ribeiro-Neto 1999, poglavlja 11 i 12) , pronalaženje u govornom materijalu (Coden et al. 2002) ili pronalaženje u muzičkom materijalu (Downie 2006). Razlike između ove dve vrste pretraživanja ilustrovaćemo na primeru digitalizovane kolekcije *Politika (1904-1941)*.

Pretraživanje digitalnih kolekcija – digitalizovana Politika

Kolekcija *Politika (1904-1941)* sadrži digitalne slika stranica svih brojeva dnevnog lista Politika od prvog broja koji je izašao 12. januara 1904. godine do poslednjeg predratnog broja 6. aprila 1941. godine. Prema (Injac 2006) "Politika je jedini dnevni list na Balkanu koji izlazi u kontinuitetu od 1904. do danas, izuzev u periodima Prvog i Drugog svetskog rata, i predstavlja nezaobilaznu arhivsku građu za proučavanje istorije i kulture Srbije i Balkana u 20. veku". Sa ovom konstatacijom bi se svako, ko je makar kratko vreme živeo u Srbiji i bivšoj Jugoslaviji, veoma lako saglasio. Ako je značaj Politike danas iz raznoraznih razloga ponešto opao, još ne tako davno je u Srbiji vladalo uverenje da ono o čemu Politika nije izvestila se, verovatno, nije ni dogodilo.

Pretraživanje ove zbirke je moguće po godinama, mesecima i datumima izlaženja. To znači da korisnik može, koristeći zajedničko sučelje digitalne kolekcije da pregleda željeni broj novina ili, pak, da redom lista sve brojeve iz određenog perioda u potrazi za traženom informacijom. Pokazaćemo na jednom primeru da to nije uvek dovoljno.

Nedavno su dve školovane i iskusne bibliotekarke-informatičarke dobile zadatak da pronađu sliku izvesnog dr Edvarda Rajana koji je u okviru misije Crvenog krsta Amerike boravio u Srbiji tokom Prvog svetskog rata[5]. One su

[5] Zahvaljujem se svojim bivšim i sadašnjim studentkinjama Biljani Kalezić i Dušici Rajčević koje su obavile ovo pretraživanje što su mi ukazale na ovo zanimljivo iskustvo i dozvolile da ga pomenem u ovom radu.

prvo pokušale da pronađu njegovu sliku u arhivama raznih organizacija (Crveni krst Amerike, Crveni krst Srbije, Vojni arhiv Srbije) ali bez uspeha. Istovremeno su, prirodno, pokušale da pronađu neke relevantne informacije i na Google-u. Pretraživanje ključnim rečima „dr Edvard Rajan site:.yu“ ne daje nikakve rezultate, to jest, ne identifikuje željenu osobu. Međutim, pretraživanje ključnim rečima „dr Edward Ryan Belgrade“ već kao prvi odgovor nudi članak iz arhive Njujork Tajmsa[6] koji je u ovom časopisu objavljen 31. januara 1915. godine, na strani SM3 (u arhivu takođe stoji da članak ima 1675 reči). Naslov ovog članka je „Američki doktor spasao Beograd – supruga srpskog podsekretara za inostrane poslove govori o herojskom radu Crvenog krsta u toj zemlji“ (videti sliku 2). Kada se jednom pristupi arhivi Njujork Tajmsa i pokrene pretraživanje arhive istim ključnim rečima, vidi se da se o dr Rajanu i njegovoj misiji u Beogradu još mnogo pisalo: 31. januara 1914. („London Tajms hvali Amerikanca zbog njegovog rada tokom austrijske okupacije“), 15. januar 1915. („Bolničarka u Beogradu dok šrapneli padaju...“), 30. januar 1915. („Govori o ratnim užasima koje je preživela u Srbiji...“[7]), 17. jun 1915. („Srpska zaraza savladana...“[8]) i tako dalje.

AMERICAN DOCTOR SAVED BELGRADE

Wife of Servian Under Secretary of Foreign Affairs Tells of Heroic Work of Red Cross in That Country.

SEEDS for sowing the devastated fields of Servia are the quest of Mme. Slavko Grouitch, wife of the Servian Under Secretary of Foreign Affairs, who arrived here the other day. Others have come over from the various war-stricken countries to ask for money for various purposes, or for sympathy. Mme. Grouitch wants seeds—wheat, corn, oats, barley, vegetable—anything that will produce food or feed.

An American by birth, having been before her marriage Miss Mabel Dunlop of West Virginia, Mme. Grouitch has become best known as the wife of the Servian Minister to London, her husband having held that post for many years before his promotion early last year. Mme. Grouitch was in Switzerland, on her way to join him in Belgrade, when the war broke out.

It was she who organized and took from England the first Red Cross mission to leave for Servia after the beginning of the war. Then it was she who succeeded in having the American Red Cross unit that went over in charge of Dr. Edward Ryan sent to what has proved the most important post in all of Servia, the hospital at Belgrade, and she believes that if she hadn't persuaded the Servian military authorities that the

ing. An appeal has been made for plows, hoes, and rakes, and it is a similar appeal over here. The farming all has to be done by the women. In one area that had been devastated by a recent battle I saw women cutting corn by moonlight.

"Naturally, when the war came I was anxious to proceed to Belgrade as quickly as possible, but a telegram from my husband had warned me not to attempt to cross Austria.

"I could have gone to Belgrade by way of Italy, but I hurried to England and engaged as many nurses as I could with the limited resources at my command. I took the ten out as Servian Red Cross nurses. The Servian Red Cross Society was the fourth Red Cross society to be organized, and for giving immediate aid it used to be one of the best equipped. In all the three wars Servia has had of late it has taken care of 20,000 wounded a month before outside aid arrived. The nurses are women of all ranks of life, led by women of the intellectual class. In the first two wars there were almost no septic cases, but this war took the Servian Red Cross as well as the Servian Nation absolutely unprepared.

"There were no uniforms for the army. There were no muskets. There was no

Mme. Slavko Grouitch in Prizren Costume.

Slika 2. Članak o doktoru Edvardu Rajanu pronađen u arhivi Njujork Tajmsa

Pošto su na osnovu podataka iz arhive Njujork Tajmsa zaključile da je doktor Rajan zaista značajan za našu istoriju dve bibliotekarke su prionule na pronalaženje informacija u digitalizovanom arhivu dnevnog lista Politika. Na

[6] The New York Times Archives (query.nytimes.com)

[7] U ovom članku se govori da je srpski Princ izrazio javnu zahvalnost dr Edvardu Rajanu zbog njegovog humanitarnog rada.

[8] Dokor Edvard Rajan se i sam razboleo od tifusa od koga se uspešno oporavio i nastavio svoju misiju u Estoniji.

žalost, pronalaženje u ovom arhivu ne ide tako lako. Svi članici u arhivu Njujork Tajmsa su u PDF formatu koji prepoznaje tekst i omogućava njegovo pretraživanje, što nije slučaj sa arhivom Politike. Korist od digitalizovane arhive svakako postoji – može se pretraživati od kuće i ne moraju se listati stari prašnjavi brojevi. Dve bibliotekarke su u svakom slučaju savesno prelistale sve digitalizovane brojeve Politike iz perioda koji je izgledao relevantan za obavljano istraživanje, ali nisu pronašle ni jedan tekst koji bi govorio o misiji dr Rajana. Naravno, ostala je sumnja da je on možda ipak pomenut u nekom tekstu koji je izmakao njihovoj pažnji.

Bibliotekarke, ipak, nisu odustajale i nastavile se da tragaju u drugim izvorima. Njihov trud se konačno isplatio pronalaskom knjige Žarka Vukovića „Da ne zaboravimo : Savezničke medicinske misije u Srbiji, 1915.“[9]. U ovoj knjizi postoje svi podaci o dr Rajanu i njegovom delovanju u Srbiji, pa čak i njegova slika. Iz ove knjige se, štaviše, saznaje da podaci o dr Rajanu potiču iz feljtona „Savezničke medicinske misije u Srbiji 1914-1918. godine“ koji je Politika objavljivala avgusta 1984. godine. Posleratne Politike još uvek nisu u digitalizovanom arhivu, ali čak i da jesu u arhivu ovakvom kakav je sada, kako bi se pregledanjem, bez mogućnosti pretraživanja punog teksta, pronašli podaci o događajima iz 1914. i 1915. godine u brojevima iz 1984?

Posle pronalaženja ovih informacija, bibliotekarske su se vratile digitalizovanoj Politici i u drugom pokušaju pronašle dva relevantna članka: prvi od njih se pojavljuje u broju od 11. decembra 1914. i pod naslovom „Amerikanci i Austrija. Optužba d-r. Rajana, da Austrijanci gaze sve obzire i sve obaveze“ govori o delovanju dr Rajana u okupiranom Beogradu, dok drugi, mnogo kraći, koji se pojavljuje u broju od 14. januara 1915. govori o poseti „amerikanske lekarske misije“ Kraljevom dvoru i ne pominje dr Rajana po imenu.

Ovaj primer jasno pokazuje od kolikog je značaja je izbor pravog formata digitalizovane kolekcije, posebno one za koju se smatra da je od izuzetnog značaja za proučavanje političke i kulturne istorije jednog naroda, a to digitalizovana kolekcija Politike svakako jeste. Ne treba, međutim, misliti da stručnjaci Narodne biblioteke Srbije zaduženi za izgradnju digitalne biblioteke to ne znaju. Izgradnja potpuno pretražive digitalne kolekcije je veoma zahtevan posao sa stanovišta novca, vremena i ljudskog znanja, te su svi ti faktori često ograničavajući prilikom odlučivanja.

Digitalizovana kolekcija Politike, međutim, nije zanimljiva samo kao arhivska građa, već može poslužiti kao izuzetna osnova za raznovrsna lingvistička istraživanja. Ne treba posebno naglašavati da se u obliku u kome je sada digitalizovana Politika ne može koristiti za takva istraživanja, a čak ni

[9] Da ne zaboravimo : savezničke medicinske misije u Srbiji, 1915. / Žarko Vuković. - Beograd : Plato, 2004.

PDF format ne bi bio dovoljan jer postojeći lingvistički softver najčešće traži da ulazni tekst bude u takozvanom tekstualnom formatu (sirovi tekst, ili XML ili HTML obeležen, ponekad i Word format). Neobično zainteresovani za utvrđivanje kakva bi se sve lingvistička i jezička istaživanja mogla obaviti nad digitalizovanom Politikom u tekstualnom formatu – koja bi se onda najpre zvala korpus Politike – pristupili smo izradi manjeg uzorka.

Korpus Politike

Izgradnja korpusa Politike sastoji se od pretvaranja sadašnjeg slikovnog formata u čisti tekstualni format. Naša prva ideja sastojala se u tome da pokušamo da upotrebimo neki softver za optičko prepoznavanje karaktera (optical character recognition – OCR) koji bi čitao već skenirane stranice. Pokazalo se da se to može uraditi, ali da se u uslovima nepostojanja projekta koji bi podržao ovako obiman posao, realno može formirati u nekom razumnom vremenskom periodu samo jedan manji uzorak[10]. Naime, kao što pokazuje slika 3, stari brojevi Politike štampani su starim tehnikama štampe, primerci koji su skenirani često sadrže pečate ili neke rukom pisane zabeleške, a sve to veoma otežava i usporava prevođenje u tekstualni format.

Slika 3 Prva stranica broja Politike od 6. aprila 1904. godine gledana kroz sučelje digitalne biblioteke Narodne biblioteke Srbije

[10] Formiranje uzorka korpusa Politika je uradila Branka Rašić, bibliotekar Gradske biblioteke u Požarevcu, u okviru svojih ispitnih obavza na postdiplomskim studijama na Grupi za bibliotekarstvo i informatiku Filološkog fakulteta u Beogradu.

Zbog svega toga je prevođenje skeniranih brojeva Politike u tekstualni format korišćenjem OCR softvera daleko zathevnije nego da se radi o novijim tekstovima. U nekim slučajevima, stari brojevi su tako zamrljani da nikako nije moguće, čak ni iz konteksta, odgonetnuti šta je na papiru zaista pisalo. Tako je, na primer, deo teksta iz članka „Kraljev put“ sa druge strane Politike od 6. aprila 1904. godine ostao do kraja nedešifrovan:

> ...На неколико места Краљ се силазио (*замрљано*) и разговарао са искупљеним (*замрљано*).
>
> У Црнуће је стигао у 10½ часова (*замрљано*) поздравио га је (*замрљано*), свештеник, а народ, нарочито омладина, бацала је цвеће Краљу и по путу којим је ишао.

Rezultat ovih prvih obavljenih eksperimenata bila je odluka da se za početak formira samo uzorak korpusa koji bi sadržao izbor članaka iz svih brojeva Politike objavljenih 6. aprila[11]. Uzorak bi sadržao tekstove iz predratnih brojeva koji se nalaze u digitalizovanoj arhivi, ali radi potpunosti korpusa i tekstove iz posleratnih brojeva koji nose isti datum. Da bi se ovaj drugi deo posla obavio, brojevi Politke koji postoje u Gradskoj biblioteci u Požarevcu su prvo skanirani, a zatim prevođeni u tekstualni format i korigovani.

Izbor tekstova je vršen nasumično. U principu, za svaki odabarni broj uzorak sadrži jedan tekst koji se odnosi na spoljnu politiku, jedan na unutrašnju politiku, jedan ili više tekstova iz društvenog i kulturnog života, i više kratkih i ličnih vesti.

Do sada prikupljen uzorak sadrži tekstove iz 15 brojeva Politike koji nose datum 6. april: 1904, 1911, 1921, 1924, 1926, 1931, 1941, 1946, 1951, 1956, 1961, 1966, 1971, 1976 i 1981. Uzorak sadrži 185 članaka i nešto više od 55.000 reči. Bez sumnje, ovaj uzorak je veoma mali za bilo kakva ozbiljna istraživanja. Podsetimo se samo da je čuveni Brown korpus engleskog jezika nastao 1961. godine imao jedan milion reči, a danas se uglavnom pominje samo iz istorijskih razloga[12]. Korpus tekstova iz Politike iz 2000. godine, a koji ponekad nazivamo i „Izborna kriza“ jer sadrži tekstove od 9. septembra do 23. septembra 2000. godine i koji je korišćen u više istraživanja (Krstev 2008), sadrži više od 580.000 reči. Iako ni ovaj korpus nije dobijen „pritiskom na dugme“, jer je tekstove trebalo, pre svega pribaviti sa mrežne stanice Politike,

[11] Ovaj datum je izabran zato što je tog dana izašao poslednji broj 1941. godine i tim brojem se digitalizovana arhiva za sada završava.

[12] The Brown Corpus (http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html)

zatim izdvojiti one relevantne (isključiti, na primer, rezultate sportskih prognoza, i slično), prilagoditi kodove i očistiti od nepotrebnih elemenata, kao što su HTML etikete, ovaj korpus je ipak mnogo lakše formiran nego uzorak o kome govorimo. Cilj našeg rada, s toga, nije da prikažemo neke konkretne rezultate, već više da ukažemo kako bi se adekvatno formiran korpus Politike, ili uostalom nekih drugih novina, uz korišćenje određenih softverskih alata i jezičkih resursa mogao koristiti za zanimljiva istraživanja.

Obrada korpusa „Stara Politika“ – neki rezultati

Analizom jezika novinskih tekstova bavili su se mnogi autori (Fowler, R. 1991), ali ova vrsta analize je dobila poseban zamah razvojem korpusne lingvistike (Teubert 2005). Razvoju ove grane lingvistike je naročito doprineo nastanak veba sredinom devedestih godina prošlog veka i dostupnost na njemu ogromne količine tekstova u digitalnom obliku, a pre svega novinskih tekstova, čime je omogućena relativno laka izgradnja korpusa velikih dimenzija. Tako su razvijeni korpusi novinskih tekstova za mnoge jezike, a u mnogim slučajevima je i neopravdano velika količina ove vrste tekstova ušla u opšte korpuse. Pomenuti korpus New York Times-a je, na primer, korišćen za različite istraživanja u mnogim domenima obrade prirodnih jezika. Primeri su automatska akvizicija znanja iz korpusa, kao što je izgradnja specijalizovanih sintaksičkih rečnika (Manning, 1993) ili evaluacija raznovrsnih metoda u obradi prirodnih jezika (Tetreault, 2001). Zanimljivo je pomenuti i konferencije posvećene razumevanju poruka i ekstrakciji informacija MUC (Message Understanding Conferences) koje su počele da se održavaju osamdesetih godina prošlog veka, a koje su se razlikovale od uobičajenih konferencija po tome što su se na njima učesnici takmičili sa svojim programskim sistemima za razumevanje poruka i ekstrakciju informacija poštujući unapred definisana precizna pravila. Korpus agencijskih vesti New York Times-a (period 1995-96, više od 150,000 članaka) je korišćen kao korpus za evaluaciju sistema učesnika sedme i poslednje konferencije MUC-7[13] (Hasegawa et al. 2004). Osim ovakvih ili sličnih primena koje spadaju u domen obrade prirodnih jezika ili veštačke inteligencije, na korpusima su obavljana i mnoga lingvistička istraživanja, bilo da je u njima jezik novina u fokusu interesovanja, ili je, pak, relativno lako pribavljen veliki korpus novinskih tekstova poslužio kao uzorak za opšte jezičke zaključke. Dobar pregled korpusno zasnovanih istraživanja jezika novina i medija uopšte je dat u zborniku (Ungerer 2000), gde je dato više radova koji su posvećeni dijahronim istraživanjima, na primer na korpusu *The Times*-a.

[13] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

U ovom radu ćemo samo naznačiti mogućnosti i metode obrade korpusa koje se zasnivaju na korišćenju elektronskih rečnika i lokalnih gramatika. Podršku ovoj obradi pruža programsko okruženje za obradu korpusa Unitex[14] koje se zasniva na tehnologiji konačnih automata (Paumier 2008). Čitalac se detaljnije može upoznati sa leksičkim resursima razvijenim za srpski jezik koji podržavaju ovu vrstu obrade u (Vitas 2009). Dalje ćemo se u radu zadržati na samo dva aspekta obrade korpusa starih novinskih tekstova. Pokušaćemo da pokažemo u čemu se njihova obrada, posebno korišćenjem usvojenih metoda, razlikuje od obrade savremenih tekstova, a zatim ćemo pokazati kako se može obaviti jedno malo istraživanje.

Čak i mali korpus koji smo sačinili pokazuje da se obrada starih tekstova značajno razlikuje od obrade savremenih tekstova, a razlozi za to su višestruki. Po pravilu, stari tekstovi sadrže više štamparskih grešaka. Korišćenje interpunkcijskih znakova, a posebno zareza, je često nepravilno ili nedosledno, kao i korišćenje velikih slova. Na primer, u istom pasusu jednog članka iz godine 1904. titula „kralj“ je zapisana jednom velikim, a drugi put malim slovom:

> ...свуда, где год се **Краљ Петар** појави народ га дочекује тако одушевљено, тако искрено и тако топло како још ниједног краља пре тога није дочекивао. Ја сам прећашњих година имао прилике да присуствујем дочеку **краља Александра** у једном месту у унутрашњости.

Pravopisna pravila su se menjala više puta u toku vremenskog perioda koji korpus pokriva. To se ogleda, na primer, u korišćenju tačke kao interpunkcijskog znaka. U najstarijim brojevima Politike svi naslovi članaka su se završavali tačkom, što je danas, naravno, neuobičajeno. Naslov i podnaslov iz broja od 6. aprila 1911. godine to ilustruju:

Дневне вести.

Престолонаследник у инспекцији.

Redni brojevi koji se zapisuju ciframa se danas obavezno završavaju tačkom, ali ni to pravilo nije u stabilnoj upotrebi. Odlomak iz jednog teksta iz broja od 6. apila 1941. godine to najbolje pokazuje:

> Онима који су онако мушки и родољубиво извршили ово велико дело, припашће достојно и часно место у историји овога народа, а

[14] Matična stranica Unitex-a http://igm.univ-mlv.fr/~unitex/

> тај дан, 27 март 1941 године, биће исписан у нашој историји као велики национални дан.

Stabilna upotreba tačke i drugih interpunkcijskih znakova je veoma značajna sa stanovišta automatska obrade teksta, a posebno automatske segmentacije na rečenice. U proteklih sto godina menjao se odnos prema beleženju suglasnika *j* u pisanom tekstu: u starim tekstovima se, na primer, nailazi na *audiencija* i *istoriski*, umesto današnjeg pisanja *audijencija* i *istorijski*. Neke fonološke alternacije se ranije nisu beležile u pisanim tekstovima, pa se u najstarijim brojevima može naići na oblik *gubitci* umesto današnjeg *gubici*. Zbog ovakvih pravopisnih razlika broj neprepoznatih reči u leksičkoj analizi elektronskim rečnicima je u proseku veći za stare nego za savremene novinske tekstove.

U starim tekstovima su se razne vrste skraćenica koristile mnogo češće nego danas, verovatno zato što je ušteda novinskog prostora bila veoma važna – sve do 1921. godine svi brojevi Politike su imali samo četiri stranice. Neke od skraćenica su predvidive i veoma često su se koristile, na primer, skraćenica *ov.* je označavala razne oblike zamenice *ovaj* i često se koristila u konstrukcijama nalik na ovoj preuzetoj iz broja od 6. aprila 1926. godine:

> На Благовести 7 ов. м. у 11 часова пре подне четврти симфонијски концерт београдске филхармоније.

U našem malom korpusu ova skraćenica je uz varijantu *o.m.* korišćena trinaest puta, dok se u pomenutom korpusu savremenih novinskih tekstova „Izborna kriza“ koji je više nego 10 puta veći ista skraćenica uopše ne javlja. Javljaju se i mnoge druge skraćenica koje, verovatno, nisu bile opšteprihvaćene već su korišćene za potrebe trenutka. Veridba objavljena u broju od 6. aprila 1931. sadrži čak devet takvih skraćenica:

> Г-ца **Милица Даничићева,** чин. Држ. хип. банке, кћи **Даринке** и **Станимира Даничића,** прокуристе фабрике кожа Бели Орао и г. **Љубомир Д. Марковић,** секретар Држ. савета, син **Савке** и пок. **Душана Марковића,** бивш. полиц. писара, верени.

U starim tekstovima poseban problem predstavljaju skraćenice koje su nastale spajanjem početnih slova nekih višesloženih naziva, jer je posle 50 ili 100 godina ponekad teško odgonetnuti na šta se one odnose. Jedan primer iz broja od 6. aprila 1926. godine koji to pokazuje istovremeno ilustruje i ono o čemu je već bilo reči, a to su razlike u korišćenje tačke kao interpunkcijskog znaka, jer se danas ovakva vrsta skraćenica piše bez tačke.

Данас треба да дође у Београд председник клуба Х. С. С.[15] г. К. Ковачевић.

Da se ne radi samo o skraćenicama koje su bile u upotrebi pre osamdeset i više godina ilustruje primer iz broja od 6. aprila 1981. godine koji mlađi ljudi verovatno više ne razumeju:

...а у Бојнику, центру устаничке Пусте Реке, пред више хиљада грађана, говорио је Веселин Станојевић, председник ОК ССРН[16].

Prilikom obrade starih tekstova poseban problem u prepoznavanju predstavljaju geografski nazivi koji su se često menjali u proteklih sto godina, uglavnom iz političkih razloga. Grad koji se sada naziva *Sankt Petersburg*, se pre toga zvao *Lenjingrad*, a pre toga *Petrograd.* Sličnih promena naziva je bilo i u Srbiji i celoj bivšoj Jugoslaviji: *Jagodina* se jedno vreme zvala *Svetozarevo*, *Zrenjanin* je bio *Stari Bečkerek* (a još ranije *Petrovgrad*), *Podgorica* se zvala *Titograd* itd. Ranije su se, za nazive mesta u susednim zemljama češće koristili slovenski nazivi, dok danas preovlađuju nazivi koji su u upotrebi u tim zemljama. Primer iz broja od 6. aprila 1946. godine govori o *Videmu* i *Trbižu* dok su danas uobičajeni nazivi *Udine* i *Tarvizio*.

У Трбиж, у видемском крају, стигла је Међународна комисија коју је словеначки народ дочекао са великим задовољством.

Mnoga geografska imena su se promenila u postkolonijalnoj eri. Interesantan je primer članka iz broja od 6. aprila 1961. godine koji govori o kongoanskoj krizi.

Леополдвил, 5. априла (Танјуг)

Представник мисије УН у Елизабетвилу потписао је са Чомбеом споразум којим му се пружају тражене гарантије да на тле Катанге, осим у базу Камина, неће ступити ни један индијски војник.

Leopoldvil i *Elizabetvil* su stari kolonijalni nazivi za *Kinšasu* i *Lumbaši*, oblast *Katanga* se od 1971. do 1997. zvanično zvala *Šaba*, dok je ime Moiza Čombea, predsednika secesionističke države Katanga i zaštitnika kolonijalnih interesa Belgije koje je jedno vreme punilo štampu širom sveta danas uglavnom zaboravljeno. Pretraga srpskih mrežnih stanica na internetu pokazuje da je

[15] Ova skraćenica se verovatno odnosi na Hrvatsku seljačku stranku.
[16] Odnosi se na Opštinski komitet Socijalističkog saveza radnog naroda

Čombe čest nadimak mnogih „loših momaka“ sa naših prostora, što je jedni ostatak njegove stare „slave“.

Leksički repertoar starih novinskih tekstova se unekoliko razlikuje od savremenih tekstova. To se najbolje vidi kod imena profesija od kojih su mnoge nestale u proteklom stoleću, pa su i njihova imena zaboravljena, tako da neka od njih ne beleži ni rečnik Matice srpske ni naš elektronski rečnik koji je razvijen na podacima iz više tradicionalnih rečnika i na obradi mnogih, većinom savremenih, tekstova. Neki brojčani podaci mogu ovo da ilustruju. Posle obrade malog korpusa *Stara Politika* uvećali smo naš morfološki elektronski rečnik sa 345 novih ulaza, od kojih je 241 imenica. Među dodatim imenicama su 141 vlastito ime (31 geografsko i 110 ličnih imena), dok od preostalih 100 imenica dodatih u rečnik čak 28 označavaju profesije: asvalter, bonbondžija, elektromehaničar, farbar, firmopisac, kartonažer, kazandžija, mašinbravar, metalostrugar, metlar, pečatorezac, pegler, staklobrusač, tašner, teracer, i tako dalje. Mnoge od njih se odnose na stare, napuštene profesije, kao *kaldrmdžija*, dok su nazivi nekih profesija potpuno zaboravljeni, te jedino konsultovanje više rečnika može da rasvetli o kojim se profesijama radi, na primer *pucer* i *kordaš*[17]. Ni kontekst, na žalost, ne daje nikakvo pojašnjenje:

> Као што се наши читаоци сећају, Милева се 8 ов. м. тровала. Она је попила пуну чашу растворене соде због тога, што се посвађала са својим вереником, једним кројачким радником, и братом **пуцером**.

> Пореска комисија при Народном одбору Општине Палилуле одржаће јавне расправе за утврђивање основица промета и дохотка за 1955 годину од самосталних занимања са територије Народног одбора Општине Палилуле и то по следећем распореду: 16 априла 1956 године: кројачице, израда мидера и прслука, ткачи штопери,

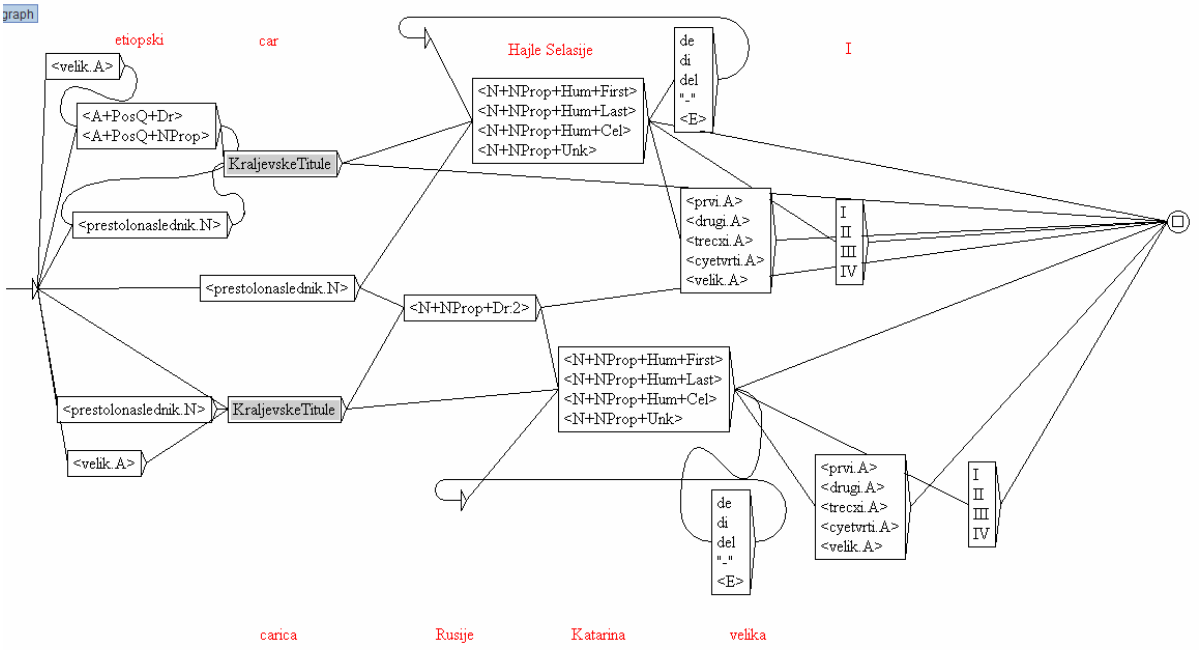
[17] Rečnik (Klajn i Šipka, 2006) beleži odrednicu *pucer* sa značenjem „onaj koji čisti, čistač“, prema nemačkom *Putzer*. Ni jedan od konsultovanih rečnika ne beleži odrednicu *kordaš*. U rečniku (Klajn i Šipka, 2006) se javlja odrednica *korda* (od italijanskog *corda*) sa dva značenja „uže ili konopac“ i „fitilj za paljenje“. Rečnik (RMS 1967) uz ovu odrednicu ima i odrednicu *korda* koja upućuje na *ćorda* (sablja). Rečnik sinonima (Lalević 2004) upućuje sa *korda* na *nož*, što ima potvrdu i u rečniku (SANU 1978) koji osim odrednice *korda* sa zanačenjima uže ili fitilj, navodi i odrednicu *korda* koja potiče iz turskog *gŏrda* sa značenjem sablja ili iz turskog *kürde* sa značenjem (vrsta) noža. Rečnik turcizama (Škaljić 1989) ne navodi ni jedno od ovih značenja. Rečnici (Vujaklija 1961), (Anić & Goldstein 2004) i (Skok 1972) ne nude ništa novo. U rečniku (SANU 1978) se, međutim, javlja i treća odrednica *korda* (od grčkog *chordē*) čije je četvrto značenje koje se koristi u građevinarstvu „vrsta kolica za prevoz zemlje koja vuče konj“. S obzirom da se u primeru iz Politike zanimanje *kordaš* javlja zajedno sa *kočijaš* ovo poslednje deluje i kao najverovatnije značenje. Konsultovanje rečnika je obavila Ljiljana Macura, bibliotekar-informatičar Narodne biblioteke.

обућари, сарачи, ташнери, крзнари и рукавичари,..., маја 1956 године: такси-шофери и власници камиона. 7 маја 1956 године: кочијаши и **кордаши**.... (Politika, 6. april 1956)

Korišćenje svih raspoloživih leksičkih resursa, morfoloških elektronskih rečnika i lokalnih gramatika dolazi do punog izražaja u narednom primeru u kojem iz korpusa pokušavamo da izdvojimo sve one segmente koji govore o pripadnicima kraljevskih ili aristokratskih porodica. Njihovo prepoznavanje se neće zasnivati na poznavanju imena pojedinačnih pripadnika ovih porodica, jer bi takvo prepoznavanje bilo nužno manjkavo zbog nemogućnosti da se prikupe imena svih pripadnika ovakvih porodica u svetu, nekada i sada. Prepoznavanje njihovih imena, međutim, olakšava činjenica da se o njima retko piše neformalno, te su njihova imena gotovo uvek propraćena odgovarajućom titulom: *kralj*, *prestolonaslednik*, *kneginja*, i tako dalje. Za razliku od imena, broj ovih titula je ograničen i one se mogu pobrojati.

Na slici 4. je prikazana lokalna gramatika u formi konačnog automata za prepoznavanje punog imena nekog pripadnika kraljevske ili aristokratske porodice. Prepoznavanje delova teksta konačnim automatom podrazumeva da program može da stigne od početnog čvora konačnog automata (označen je trouglom) do završanog stanja (označen je krugom u kome je upisan kvadrat), sravnjujući sve čvorove koji se na tom putu nalaze redom sa rečima teksta. Centralno mesto u konačnom automatu zauzimaju osenčeni čvorovi koji predstavljaju poziv podređenog konačnog automata (podgrafa) – KraljevskeTitule - koji prepoznaju naziv titule. Delovi automata se izdvajaju u podautomate, pre svega, zbog preglednosti celog automata.

Sravnjivanje čvorova može da bude raznovrsno, a najjednostavniji je slučaj doslovno sravnjivanje niski karaktera, kao što je u našem automatu, na primer, niska „I“ u imenu *etiopski car Haile Selasije I* (Politika, 12. IX 2000). Složenije je sravnjivanje čvorova koje zavisi od sadržaja rečnika jer ono, u principu, reč iz teksta pokušava da sravni sa skupom reči, na primer, svim oblicima jedne leme iz elektronskog rečnika. Takav je, na primer, čvor <prestolonaslednik.N> koji će sravniti sve oblike imeničke leme *prestolonaslenik* i tako prepoznati, na primer, segment *Prestolonaslednika Aleksandra* (Politika, 6. IV 1911). Još složeniji su čvorovi koji takođe zavise od rečnika a koji sravnjuju reči date gramatičke kategorije i drugih zadatih karakteristika. Primer je čvor <A+PosQ+Dr> koji sravnjuje sve reči u tekstu koje predstavljaju oblik relacionog prideva izvedenog iz naziva države. Na osnovu toga bi bio prepoznat prvi član u segmentu *iranskog šaha Mohameda Reze Pahlavija* (Politika, 6. IV 1956).



Slika 4. Konačni automat *Kralj* koji prepoznaje u tekstu segemnte koji referišu pripadnike kraljevskih ili aristokratskih porodica.

Uopšteno govoreći konačni automat sa slike 4. ima dva glavna puta: gornji, u kome tituli i imenu osobe prethodi zemlja kojoj ona pripada naznačena relacionim pridevom, i donji, u kome naznaka zemlje sledi iza titule i imena u vidu naziva države u genitivu (čvor <N+NProp+Dr:2>). Ime osobe se može sastojati od više imena što je u automatu označeno petljom, to jest, povratkom u isti čvor. Imena mogu biti reči prepoznate rečnikom kao lično ime (<N+NProp+Hum+First>), prezime (<N+NProp+Hum+Last>), ime poznate osobe (<N+NProp+Hum+Cel>), a u ovom slučaju to mogu biti i potencijalne reči koje su zapisane početnim velikim slovom, a nema ih u rečniku (<N+NProp+Unk>). Tako bi segment *grofa Lajoša Vermeša* (Politika, 11. IX 2000) bio prepoznat kao sekvencija: titula (čvor *KraljevskaTitula*), lično ime (<N+NProp+Hum+First>) i nepoznato vlastito ime (<N+NProp+Unk>).

Primena ove lokalne gramatike na korpus „Stara Politika“ pronalazi 63 pojavljivanja od kojih se 54 direktno odnose na osobe iz kraljevskih ili plemićkih porodica, aktuelnih ili iz prošlosti. Neka od tih pojavljivanja u kontekstu su:

```
Istraga koja je vođena o odlasku cara Karla iz Pranžena dovršena je.
zvanično preda svoju ostavku britanskoj kraljici Elizabeti Drugoj.
montirani proces princezi Botum Bofa i princu Norodomu Naradipu, deci
razgovarao je s velikim Knezom Vladimirom o propasti "Petropavlovska".
```

U osam slučajeva, radi se o korišćenju imena osoba iz kraljevskih porodica za različita imenovanja. Neki od ovih slučajeva su:

Ovo je, posle lanca ordena Cara Lazara, najveće odlikovanje u Srbiji.
Na uglu Pariske i Cara Dušana ulice, u palati Srpske banke stanuje
vojnom gađanju XVIII pešadijskog puka "Kraljevića Đorđa".

Jedan prepoznati segment je pogrešan. Ova ista lokalna gramatika primenjena na korpus „Izborna kriza“ koji je deset puta duži (po broju reči), ekstrahuje 75 segmenata, od čega se 46 odnosi na osobe iz kraljevskih ili plemićih porodica:

araocu u Evropi, jordanska princeza Vijdan Ali o umetnicama arapskog
Kraljica Elizabeta II je predsedavala njenim Savetničkim komitetom
belgijski prestolonaslednik princ Filip i njegova supruga princeza
13. septembra 1918. godine vojvoda Živojin Mišić završio je rečima:

U poslednjem primeru, *vojvoda* nije plemićka titula već vojnički čin, ali lokalna gramatika ovu razliku u značenju ne može da obuhvati. U 25 slučajeva imena osoba se koriste za različite vrste nazivanja:

Urbing, Subotica, Trg cara Jovana Nenada 15, Subotica, telefoni
Osmi "Sabor vojvode Stojana Čupića - Zmaja od Noćaja" odvijao se
Jelisaveta" je između dva svetska rata nosio naziv "Kraljica Marija"
"Poljaci", prema delu "Kralj Ibi" A. Žarija

Četiri segmenta su pogrešno prepoznata. Iz ovih podataka se vidi da je pojavljivanje ličnosti iz kraljevskih porodica u savremenim novinskim tekstovima srazmeno znatno manje nego što je bilo u prošlosti. Ako se pri tome napravi analiza pojavljivanja aktuelnih pripadnika kraljevskih ili plemićkih porodica, to postaje još uočljivije. U korpusu “Stara Politika” samo jedno od 54 pojavljivanja se ne odnosi na aktuelne nosioce titula. U više navrata se, doduše, govori o “bivšim kraljevima”, ali se radi o živim ljudima koji su izgubili titulu. Na primer, Politika od 6. IV 1921. piše o Karlu Habzburškom, između ostalog:

> Мађарска влада известила је представнике Мале Антанте у Пешти, да ће бивши цар и краљ Карло напустити мађарску територију пре среде.

U korpusu “Izborna kriza” situacije je drukčija jer se od 46 pojavljivanja, manje od pola, tačnije 19, odnosi na aktuelne nosioce titula.

Ostaje da razjasnimo odakle potiču pogrešna pronalaženja. Naša lokalna gramatika *KraljevskaTitula* je prilično komotno napisana, što znači da ne vodi računa o potrebi da se članovi prepoznatih segmenata slažu po određenim gramatičkim kategorijama. Iako ovo deluje kao veliki nedostatak, relativno mali broj pogrešno prepoznatih segmenata pokazuje da se, u ovom slučaju, sam tekst stara o tome da pogrešnih prepoznavanja ne bude. Pogrešno prepoznate sekvencije su:

```
je dan poveći dopis iz Pešte o poseti Kralja Petra Franji Josifu.
tri poena više od drugoplasiranog tima Novosadskog šah-kluba, davao
Predsednik Komisije FIDE za problemski šah Bedrin Formanek i predsed
Dr Vezir Bajrami, lekar u OVK bolnici, međutim, kako naglašava
mog prijatelja primio doktor po imenu Vezir Bajrami, koji je iz Štml
```

Primeri pogrešnog prepoznavanja pokazuju da bi se samo prvi i drugi primer mogli eliminisati iz prepoznavanja produbljivanjem lokalne gramatike uvođenjem uslova slaganja. Četvrti primer bi se mogao isključiti istovremenim korišćenjem lokalnih gramatika za prepoznavanje ličnih imena sa titulama koje bi u ovom slučaju prepoznale duži segment *Dr Vezir Bajrami*, a prepoznavanje dužih segmenata je najčešći princip prepoznavanja. Vrlo detaljne i precizne lokalne gramatike za prepoznavanje ličnih imena, sa i bez titula, zanimanja i funkcija koje su razvijene za srpski su opisane u (Krstev 2008) i (Gucul-Milojević 2008). Treći primer bi rešilo uključivanje višečlanog izraza *problemski šah* u elektronski rečnik kompozita[18], dok bi preostali, peti primer na nivou prepoznavanja samo lokalnim gramatikama bilo teško rešiti.

Zaključak

Digitalne biblioteke danas predstavljaju repozitorijume ogromne količine informacija od kojih je većina tekstualnog ili hibridnog oblika. I pored početnog zaostajanja Srbija danas intenzivno razvija digitalne biblioteke koje se uključuju u mreže svetskih i evropskih digitalnih biblioteka. Ove digitalne biblioteke sadrže ogromno jezičko blago na srpskom jeziku. Primeri digitalnih biblioteka na engleskom jeziku pokazuju da se izborom pogodnog digitalnog formata material iz digitalnih biblioteka može ne samo uspešno pretraživati već i koristiti za izgradnju jezičkih korpusa velikih dimenzija. U ovom radu smo pokazali, da usled raznih ograničenja finansijske, vremenske i pravne prirode, sadržaj najveće digitalne biblioteke u Srbiji za čiji razvoj je zadužena Narodna biblioteka Srbije, nažalost, nije takav da bi pretraživanje po punom tekstu bilo moguće. Rezultat toga je da se one ne mogu direktno koristiti ni za izgradnju korpusa. Nekoliko navedenih primera bi trebalo da pokaže koliko bi dobila srpska naučna populacija transformacijom postojećih digitalnih biblioteka u adekvatniji format.

Literatura

Arms, W. Y. 2000. *Digital libraries*, The MIT Press, Cambridge, Massachusetts, London, England, 2000.

[18] Ovaj izraz (RMS 1967) ne beleži, ali je njegov broj pojavljivanja na Google-u dosta velik (samo u obliku „problemski šah“ 13,600 puta, 13. II 2009). Radi se o posebnoj šahovskoj disciplini koja se sastoji od rešavanja nekog probelma, bez protivnika.

Baeza-Yates, R. / Ribeiro-Neto, B. 1999. *Modern Information Retrieval*, ACM Press & Addison-Wesely.

Coden, A. R., Brown, E. W., Srinivasan, S. (eds.). 2002. *Information Retrieval Techniques for Speech Applications*, Springer.

Downie, J. S. 2006. The Music Information Retrieval Evaluation eXchange (MIREX), *D-Lib Magazine*, 12 (12).

Fowler, R. 1991. *Language in the News –Discourse and Ideology in the Press*, Routledge.

Gucul-Milojević, S. 2008. *Prepoznavanje jedne klase imenovanih entiteta u elektronskim tekstovima na srpskom jeziku*, Magistarska teza, Filološki fakultet.

Hasegawa, T., Sekine, S., and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association For Computational Linguistics* (Barcelona, Spain, July 21 - 26, 2004). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 415.

Injac, V. 2005. "Narodna biblioteka Srbije – punopravni partner u projektu 'Evropska Biblioteka'", Pregled Nacionalnog centra za digitalizaciju, Vol. 7, pp. 49-54.

Injac, V. 2006. "Narodna biblioteka Srbije u projektu 'Evropska biblioteka'", *Glasnik Narodne biblioteke Srbije*, No. 1.

Krstev, C. 2002. "Digitalne biblioteke : razgraničenje pojmova, " u *INFOteka : časopis za informatiku i bibliotekarstvo*, vol. 3, No.1, pp. 3-14, Beograd.

Krstev, C. 2008. *Processing of Serbian : Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade.

Manning, C. D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting on Association For Computational Linguistics* (Columbus, Ohio, June 22 - 26, 1993). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 235-242.

Lesk, M. 1997. *Practical Digital Libraries – Books, Bytes & Bucks,* Morgan Kaufmann, San Francisco.

Neavill, G. 2004. "Emergence of digital libraries" u *INFOteka : časopis za informatiku i bibliotekarstvo*, vol. 5, No.1-2, pp. 25-33, Beograd.

Paumier, S. 2008. *Manuel d'utilisation du logiciel Unitex,* http://www-igm.univ-mlv.fr/~unitex/UnitexManual2.0.pdf,

Tetreault, J. R. 2001. "A Corpus-Based Evaluation of Centering and Pronoun Resolution",*Computational Linguistics*, December 2001, Vol. 27, No. 4, Pages 507-520.

Teubert, W. 2005. "My version of corpus linguistics", *International Journal of Corpus Linguistics*, 10:1, pp. 1-13.

Ungerer, F. (ed.) 2000. *English Media Texts, Past and Present: Language and Textual Structure*, John Benjamins Publishing Company

Vitas, D. 2009. "Resursi i metode za obradu srpskog – stanje i perspektive" u istom zborniku

Rečnici

Anić, V. / Goldstein, I. 2004. *Rječnik stranih riječi*, Zagreb : Novi Liber.

Klajn, I. / Šipka, M. 2006. *Veliki rečnik stranih reči i izraza*, Novi Sad : Prometej.

Lalević, M. S. 2004. *Sinonimi i srodne reči srpskohrvatskog jezika*, Beograd : Nolit (fototipsko izdanje 1974).

RMS 1967. *Rečnik srpskohrvatskoga književnog jezika*, Matica srpska – Matica hrvatska : Novi Sad – Zagreb, knjiga druga, ž – kosište.

SANU 2000. *Rečnik srpskohrvatskog narodnog i književnog jezika*, Beograd : SANU, Knj. 10 : koliti – kukutica (fototipsko izdanje 1978).

Skok, P. 1972. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*, Zagreb : Jugoslovenska akademija znanosti i umjetnosti.
Škaljić, A. 1989. *Turcizmi u srpskohrvatskom jeziku*, Sarajevo : Svjetlost.
Vujaklija, M. 1961. *Leksikon stranih reči i izraza*, Beograd : Prosveta.

Ključne reči: digitalne biblioteke, pronalaženje informacija, formati teksta, korpus, obrada srpskog, elektronski rečnik, lokalne gramatike

Digital Libraries as a Potential Linguistic Resource – Situation in Serbia

Summary

In this paper we present the actual situation in development of digital libraries in Serbia, in particular the efforts invested in this field by the leading library institution National library of Serbia. We are particularly interested in the searching options offered to the users of these libraries. Due to many financial, time and copyright constraints the format of most of included material is picture-oriented rather then text-oriented limiting search to meta-data instead of full-text. This is also valid for the very important collection of all printed issues of the daily newspaper *Politika*, from its first number that appeared in 1904 to the last number before war, dated April 6. 1941. Although web interface to this collection offers new comfort to the users, its inappropriate format severely limits its usage. We show that transformation from picture to text format of this collection is by no means straightforward – however, should it be done, the benefits to many users would be enormous. Among them would be linguists that would obtain large-scale diachronic corpora which they could use for different purposes, using various methods, some of which are presented as a case study.