Pregledni znanstveni rad
UDK 008(497.5):004

MOGUĆNOST OZNAČITELJSKIH ALATA UNUTAR MREŽNOG OKVIRA ZA ISTRAŽIVANJE HRVATSKE KULTURNE BAŠTINE

Mario Essert
(Fakultet strojarstva i brodogradnje Sveučilišta u Zagrebu)

U radu je predstavljen mrežni okvir za čuvanje i obradbu dokumenata iz različitih kategorija hrvatske kulturne baštine (književnost, slikarstvo, arhitektura i dr.) pohranjenih u različitim medijima (digitalizirani rukopisi, tekst, slike, zvučni atlasi, filmovi...). Okvir omogućuje postavljanje kategoriziranog digitalnog zapisa s različitim obilježjima u vremensko-prostorne koordinate i pretraživanje po različitim kriterijima. Tekstni dokumenti, osim klasičnih (bibliotekarskih) podataka, mogu se pretraživati i po riječima koje oni sadrže te prikazivati u vremenskim trajektorijima, što omogućuje praćenje željenih riječi kroz stoljeća, od njihova nastanka do (eventualnog) iščeznuća i novih pojavaka. Za sintaktično-semantička označivanja izgrađen je vizualni editor TEIMark, a za označivanja slika (npr. digitaliziranih rukopisa) načinjen je program DocMark. Oba editora omogućuju postavljanje vizualnih oznaka (tagova) iznad informacije (teksta ili slike) u nizu slojeva, koji se mogu po želji sakriti, prikazati ili spremiti u XML/TEI-zapisu. Svaki dokument može imati svoj skup trojaca (*triplets*), koji se onda preko baze Virtuoso triplestore može pretraživati naredbama SparQL. Mrežni okvir prati i dodatni razvojni sustav za lingvističku obradu teksta, kao i program koji iz rečenica teksta izvlači s-p-o-informaciju prema korisnički definiranim uzorcima. Razvojni sustav omogućuje poluautomatsko stvaranje abecedarija i rječnika te njihovo povezivanje preko *linked data* unutar definicija i natuknica na postojeće *online*-rječnike. To je temelj budućeg ontologijskog povezivanja ovakvih podatka (LOD) u globalni mrežni oblak.

Ključne riječi: alati za vizualno označivanje, semantički okviri, izvlačenje informacije, jezikoslovni i kulturološki povezani podaci

UVOD

Hrvatska, iako teritorijem nije velika, iznimno je bogata raznolikom kulturnom baštinom. Tijekom seobe Slavena, prema materijalnim dokazima, Hrvati stižu na ove prostore već u 7. stoljeću poslije Krista. Dugotrajna nazočnost Otomanskog Carstva i vladavina Austro-Ugarske te brojni drugi utjecaji ostalih naroda koji su se ispreplitali na ovom prostoru (Slaveni, Romani i Germani) stvorili su osebujnu cjelinu na jedinstven i neponovljiv način. Da bi se kulturna baština sačuvala barem u virtualnom svijetu, provodi se, pogotovo u zadnjem desetljeću, digitalizacija materijalne i nematerijalne baštine (rukopisi, književna djela, dijalekti, folklor, umijeća…). Digitalizirani uzorci pohranjuju se u brojnim digitalnim repozitorijima (npr. Digitalni akademski repozitorij[1] Sveučilišta u Zagrebu, Digitalni akademski arhivi i repozitoriji[2]) i dokumentacijskim zbirkama (Digitalna zbirka HAZU[3], Virovitička dokumentacijska zbirka[4] i sl.). Nedostatak je dosadašnjeg pristupa u međusobnoj nepovezanosti spremljenih digitaliziranih objekata i njihova nedohvatljivost – zatvorenost prema istraživanjima različitih struka. Za nestručne pak korisnike velika je poteškoća pretraga po naslovima djela ili imenima autora (koji su im najčešće nepoznati), jer ne postoji dohvatljivost informacije preko vremensko-prostornih koordinata.

Ovaj rad opisuje usmjeren, dobro razrađen višegodišnji pokušaj stvaranja mrežnih okvira za spremanje, dohvaćanje i obradbu digitaliziranih uzoraka hrvatske kulturne baštine s obzirom na vrijeme i prostor nastanka, kao i povezivanja spremljene informacije s drugim repozitorijima, *online*-enciklopedijama i zbirkama objavljenima na mreži.

Da bi se to postiglo, bilo je potrebno projektirati i izvesti uz pomoć suvremenih tehnologija tri velika računalna mrežna sustava:

1. MHKB (Mrežna hrvatska kulturna baština, engl. *Croatian web heritage*) za spremanje, traženje i vizualizaciju digitaliziranih uzoraka različitih medija iz svih područja kulture

---

[1] http://dar.nsk.hr/ Internet. 30. siječnja 2016.

[2] https://dabar.srce.hr/ Internet. 30. siječnja 2016.

[3] http://dizbi.hazu.hr/ Internet. 30. siječnja 2016.

[4] www.muzejvirovitica.hr Internet. 30. siječnja 2016.

2. MHJ (Mrežno hrvatsko jezikoslovlje, engl. *Croatian web linguistics*) kao popratni, istraživački sustav, za temeljnu jezikoslovnu obradbu i pripravu podataka za MHKB
3. CroLOD (Hrvatski povezani otvoreni podaci, engl. *Croatian linked open data*) za povezivanje hrvatskih otvorenih (jezikoslovnih i kulturoloških) podataka u svjetski oblak (engl. *LOD cloud*)[5]

Svi su mrežni sustavi modularni, realizirani u tzv. MVC-tehnologiji[6] (engl. *model-view-control*), a obuhvaćaju više integriranih ili vanjskih modula, od kojih će samo temeljni moduli biti kratko opisani.

Sustav MHKB izveden je u tehnologiji PHP/MySQL[7], koja se zbog obradbe LOD-podataka povezuje s bazom Virtuoso triplestore[8], a preko API-a (JSON[9]) sa sustavom MHJ. Sustav MHJ izveden je u tehnologiji Web2py[10] s modulima Python[11] (poglavito NLTK[12] i SciKit-learn[13]). CroLOD za razvitak ontologije koristi Protege[14], a za pohranu trojaca i upite SPARQL[15] koristi se server Virtuoso. Sve su tehnologije *open-source*, dakle besplatne za razvoj u akademskom okruženju. Svi sustavi za svoju vizualizaciju koriste tehnologije jQuery/bootstrap/C3/D3[16] Javascript s dodatnim modulima (*timeline*, *fancytree*, webGL i dr.).

U ovom radu središnje mjesto pripast će sustavu MHKB, a ostali će sustavi biti prikazani u onoj mjeri koja tumači međusobnu zavisnost i osigurava koherentnost općeg projekta. Njihova poveznica bit će preslikavanje – matematička funkcija koja je u temelju svih označivanja.

---

5 http://lod-cloud.net/ Internet. 30. siječnja 2016.

6 https://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller Internet. 30. siječnja 2016.

7 https://secure.php.net/ Internet. 30. siječnja 2016.

8 http://virtuoso.openlinksw.com/ Internet. 30. siječnja 2016.

9 http://www.json.org/ Internet. 30. siječnja 2016.

10 http://www.web2py.com/ Internet. 30. siječnja 2016.

11 https://www.python.org/ Internet. 30. siječnja 2016.

12 http://www.nltk.org/ Internet. 30. siječnja 2016.

13 http://scikit-learn.org Internet. 30. siječnja 2016.

14 http://protege.stanford.edu/ Internet. 30. siječnja 2016.

15 http://sparql.org/ Internet. 30. siječnja 2016.

16 https://jquery.com/, http://getbootstrap.com/, http://c3js.org/, http://d3js.org/ Internet. 30. siječnja 2016.

POSTOJEĆI RADOVI

Navedeno područje naših istraživanja i izvedbe računalnih sustava, s naglaskom na označiteljske alate, preširoko je područje da bismo obuhvatili sve postojeće svjetske radove koji su utjecali na njihov nastanak. Kompromis je navesti samo glavna područja i najvažnije autore, odnosno djela, s kojima je ovaj rad najviše povezan:

- Digitalizirana kulturna baština – CULTURESAMPO[17]: infrastruktura za finsku nacionalnu ontologiju (muzeja, knjižnica, arhiva i sl.) s naprednim semantičkim označivanjem (Hyvönen, Makela 2009) i vizualizacijama (Ioannides 2014)
- Vizualno označivanje – DIRT (Digital research tool)[18]: povezuje tisuće rješenja i projekata za vizualizaciju
- Leksikologija – generativni rječnici (Pustejovsky 1998), sinonimija (Šarić 2008) i leksička semantika (Lieber 2009)
- Semantička analiza – MTT (Meaning-Text Theory) u izvlačenju informacija (Melčuk 2015) i leksičke funkcije (Gelbukh, Kolesnikova 2014)
- Označiteljski mrežni okviri – LMF (Lexical Markup Framework) u temelju globalnog *wordneta*[19] (Francopoulo 2013)
- Otvoreni povezani podaci – LOD (Linked Open Data): temelj su općih mrežnih ontologija[20], a također i jezikoslovnih (Chiarcos, Nordhoff, Hellmann 2012)

Na internetu dosad nisu bili prisutni označiteljski alati koji bi objedinjavali sva ova područja.

---

[17] www.kulttuurisampo.fi Internet. 30. siječnja 2016.

[18] http://dirtdirectory.org/ Internet. 30. siječnja 2016.

[19] http://globalwordnet.org/ Internet. 30. siječnja 2016.

[20] http://protege.stanford.edu/ Internet. 30. siječnja 2016.

OZNAČIVANJE – MATEMATIČKA FUNKCIJA

Označivanje je matematička funkcija koja elementu iz jednog skupa (domene) pridružuje element drugog skupa (kodomene). S obzirom na takvo jednoznačno pridruživanje, označivanje se često zove i *preslikavanje*. Dva ili više elemenata iz domene mogu se preslikati u jedan element kodomene, ali ne i obratno: nijedan element domene ne smije se preslikati u više elemenata kodomene. To naprimjer znači da za skupove X={1,2,3} i Y={A,B,C,D}, ako je X domena, a Y kodomena, onda preslikavanje f={(1,D),(2,C),(3,C)} jest funkcija, a preslikavanje h={(1,D),(1,C),(3,C)} to nije (jer se element 1 iz skupa X preslikao i u element D i u element C u skupu Y). Kako se vidi funkcija je skup uređenih (poredanih) parova u kojima je prvi element član domene, a drugi kodomene. Pojedinačno preslikavanje elementa domene u neki element kodomene obično se prikazuje izrazom *f(element_iz_X)=element_iz_Y* ili $f(x \varepsilon X) = y \varepsilon Y$, na primjer f(1)=D ili f(2)=C i sl. Element iz skupa X, element_iz_X u ovakvoj notaciji zove se još i argument funkcije.

Funkcije se mogu *komponirati* tako da se prvo primijeni funkcija *f* na argument *x*, a potom primijeni funkcija *g* na rezultat prethodne primjene. Na taj se način dobije funkcija $g \circ f$ (*g* komponirano sa *f*), definirana sa $(g \circ f)(x) = g(f(x))$, za svaki element *x* u skupu *X*. U slučaju više nezavisnih domena, naprimjer X1 i X2, moguća je i kompozicija funkcije za više varijabli $f(x1, x2) = g(x1) \circ h(x2)$ ili drugačije, s dva argumenta napisano: *f(g(x1), h(x2))*, gdje su: *x1* element iz skupa X1, a *x2* element iz skupa X2.

Označivanje u jezikoslovlju također je matematička funkcija, a izborom različitih skupova i njihovih pripadnih elemenata postižemo različite vrste označivanja. U ovom radu pokazat će se označivanja jezikoslovnih dokumenata kad su domene na razini slova, npr. rukopisnog zapisa (paleografske, kodikološke, tipografske i druge domene), zatim morfosintaktičkih ili semantičkih obilježja diskurzivnih oblika i rečenica, kao i preslikavanja čitavih tekstnih dokumenata ili fonoloških zapisa u dijakronijske vremensko-prostorne koordinate. Štoviše, po uzoru na različite konstruktivističke pristupe, pretpostavlja se (po Saussureu) da je i sam jezik proizvoljan sustav označivanja; dakako, sa svojom strukturom koja nastaje i mijenja se uporabom. Sva nabrojena označivanja nismo samo teorijski osmislili nego i računalno izveli te potom iskoristili u nekoliko javnih jezikoslovnih projekata. Neka programska rješenja još čekaju na svoj prikladni pojavak.

Ovo označivanje i analiza digitaliziranog dokumenta ostvareno je računalnim sustavom *DocMark* koji koristi tehnologiju WebGL, JavaScript i PHP/MySQL, pa se poziva preko interneta. Korisniku se omogućuje rad u jednom ili više slojeva nad slikom dokumenta. U svakom sloju, nakon izbora markera, korisnik klikom miša postavlja marker (vizualnu oznaku) na željeno mjesto slike dokumenta.

DocMark – MREŽNI PROGRAM ZA INTERAKTIVNO OZNAČIVANJE S POMOĆU VIZUALNIH OZNAKA

U ovom je desetljeću u Hrvatskoj digitaliziran velik broj starih rukopisa (pisanih glagoljicom, latinicom ili ćirilicom) i spremljen u digitalnom formatu slike (npr. formatu jpg ili png). Za njihovo istraživanje bilo bi potrebno dugotrajno i mučno transkribiranje i/ili često transliteriranje (npr. glagoljičkih tekstova). Problem je taj što se automatsko, strojno prepoznavanje ili prevođenje rukopisa, pogotovo starih pisama, teško može načiniti, a prepisivanje u digitalni, tekstni oblik može se povjeriti samo nekolicini stručnjaka. S druge pak strane različiti istraživači svoju pozornost usmjeruju na različite aspekte takvih dokumenata: kodikolozima će dokument ili knjiga biti važni u smislu obilježja koja imaju kao fizički objekti (materijal od kojeg je dokument načinjen, vrsta tinte ili boje kojim je napisan/nacrtan i sl.), paleografima će biti zanimljive margine, komentari, dekoracije i dr., jezikoslovcima pak riječi i njihova sintaksa (Tomić, Glumac, Essert 2003).

Rješenje je pronađeno preslikavanjem (lijepljenjem) vizualnih oznaka u slojevima nad pojedinom slikom dokumenta, ne mijenjajući pritom samu sliku. Svakom sloju moguće je pridružiti drugu domenu, npr. domenu stranič nog postava s pripadnim oznakama (širine margina, razmaka između redaka, odjeljaka i stupaca, inicijala i istaknutih slova, spacioniranja, uporaba velikih slova, oznaka za paragrafe i sl.).

Koristeći oznake iz Slike 1 – skromni presjek među 387 definiranih kodikoloških, grafetičkih i aproprijacijskih oznaka doktorskog rada Marijane Tomić (2013) – lako je uočiti da označiteljska funkcija za glagoljički broj folije unutar arka daje vizualnu oznaku plave sklopkice (S7_23, kao indeks oznake u računalnoj bazi). To se rješenje dobije jednostavnom kompozicijom funkcija $g(x1 \ \varepsilon \ X1) \circ h(x2 \ \varepsilon \ X2)$, gdje je skup X1 domena boja X1 = {narančasta, modra, zelena, smeđa}, s pripadnom kodomenom Y1={arak,

Arak						
	Arapski broj		S7_11	9	010	orange
	Rimski broj		S7_12		041	orange
	Glagoljički broj		S7_13		165	orange
	Latinica		S7_14	A	179	orange
Folija unuar arka						
	Arapski broj		S7_21	9	010	blue
	Rimski broj		S7_22		041	blue
	Glagoljički broj		S7_23		165	blue
	Latinica		S7_24	A	179	blue
Folijacija						
	Arapski broj		S7_31	9	010	green
	Rimski broj		S7_32		041	green
	Glagoljički broj		S7_33		165	green
	Latinica		S7_34	A	179	green
Paginacija						
	Arapski broj		S7_41	9	010	red
	Rimski broj		S7_42		041	red

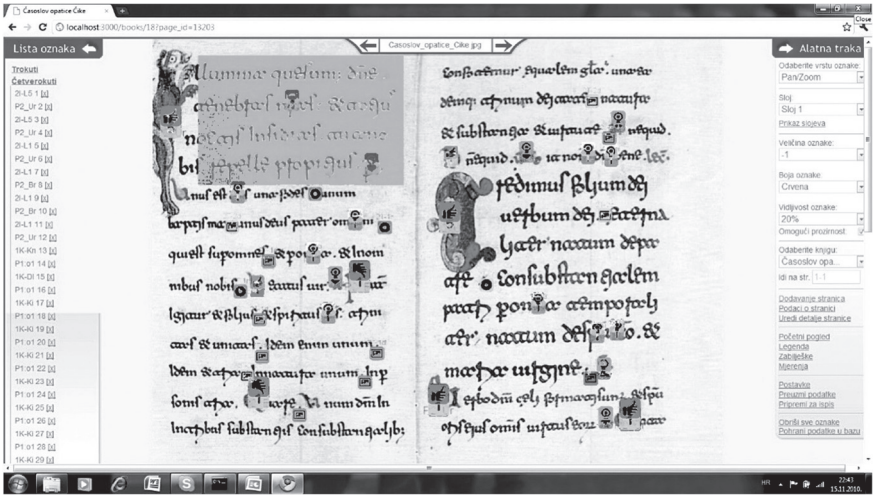
Slika 1. Oznake/markeri programa *DocMark*

folija unutar arka, folijacija, paginacija}, dok je skup X2 domena brojeva X2={arapski, rimski, glagoljički, latinica}, a kodomena Y2={broj 9, plamen, sklopkica, A}, odnosno skup brojeva {010, 041,165,179} neobojanih vizualnih oznaka spremljenih u odgovarajućoj datoteci. Tek će se djelovanjem, kompozicijom druge funkcije svaka oznaka prikladno obojiti. Funkciju *g()* stoga možemo zvati *bojanje()*, a funkciju *h()* zvati *brojni_zapis()*. Uobičajeno je da se funkcija označuje imenom iza kojeg slijede otvorena i zatvorena uglasta zagrada, u koju se za konkretan slučaj navode još i argumenti.

Slijedom navedenih općih matematičkih izraza za ovaj konkretan slučaj možemo napisati: *brojni_zapis (glagoljički broj) = sklopkica* ili *brojni_zapis (glagoljički broj) = 165* te *bojanje (modra) = folija_unutar_arka*, pa će funkcija *označivanje (bojanje (modra), brojni_zapis (glagoljički broj)) = brojni_zapis (glagoljički broj)◦ bojanje (modra)* dati željenu oznaku programa *DocMark*: *označivanje (bojanje (modra), brojni_zapis (glagoljički broj)) = modra sklopkica*.

Na isti način spremaju se i dohvaćaju u posebnom sloju grafetička sredstva (ligature, kratice, interpunkcija, punktuacija), aproprijacije (zapisi u knjigama, ispravci koje je unosio pisar ili tiskar) ili bilo koja druga svojstva sabrana u nekoj domeni.

Korisnik izabire elemente domene predočene u vrstama oznaka, tipovima vizualnih markera (npr. trokuta, četverokuta, strelice, križića, ikonica),



Slika 2. Označena stranica *Časoslova opatice Čike*, latinično pismo, Zadar, 11. st.

različitih boja i prozirnosti. Svakom tipu markera autor prije označivanja pridružuje neko svojstvo/kategoriju (matematičku domenu).

Mrežni program[21] načinjen je tako da korisnik klikom miša izabire i postavlja neku oznaku/marker na željeno mjesto slike otvorenog dokumenta izabirući prije toga sloj u kojem će se postaviti. Kod pregleda označenih dokumenata moguće je prikazivati pojedinačne slojeve ili skupno, bilo koju njihovu kombinaciju. S obzirom na mrežnu programsku realizaciju, više istraživača može raditi na istom dokumentu, u istim ili različitim slojevima, prema dopuštenjima administratora dokumenta. I to bez transkripcije ili transliteracije što značajno ubrzava vrijeme priprave i analize označenih dokumenata. Postupak je isti za bilo koje digitalizirane dokumente, bilo koja pisma ili slikovne prikazbe (vidi Sliku 2).

Za mjesta u dokumentu gdje je bitna mjerna veličina istraživačkog objekta (npr. istaknuta glagoljička slova, inicijali), margine, razmaci između stupaca i druge bjeline na slici, sustav nudi alat za precizno mjerenje (do stotinke milimetra). Među oznakama/markerima posebno mjesto zauzimaju

[21] http://docmark.cingel.hr/session/new (korisničko ime: test1 ili test2 ili… test20; zaporka: 123456)

linije koje korisnik može povlačiti od neke pozicije do druge te rastezljiva polja kojima označuje zanimljiva područja dokumenta (obično s visokom prozirnošću kako bi se originalni dokument ispod njih mogao vidjeti). Svi markeri, sva mjerenja i svi opisi automatski se spremaju u bazu, a na poziv automatski dohvaćaju, što omogućuje naknadno editiranje dokumenta.

Analiza označenog dokumenta *DocMark*om daje jednostavno i brzo prebrojavanje markera/tagova po vrstama, slojevima i stranicama/slikama dokumenta, računa udaljenosti željenih markera različitih tipova, obrađuje zadana područja dokumenata (npr. koliko markera nekog tipa ima u određenim poljima/područjima) te brzo pozicioniranje i prikaz dokumenta prema odabranom (pojedinačnom) markeru, uz mogućnost njegova automatskog zumiranja na središte stranice. U slučaju transliteracije teksta markeri se s lakoćom mogu pretvoriti u bilo koje XML/TEI-tagove i obrađivati u LOD-oblaku.

TEIMark

Po uzoru na program *DocMark* načinili smo program *TEIMark* koji omogućuje označivanje teksta nekog digitaliziranog dokumenta. Dok se označivanje u programu *DocMark* izvršavalo nad slikom dokumenta, u *TEIMark*u se to provodi nad prepoznatim ili utipkanim riječima, rečenicama ili odlomcima. Na taj je način moguće označiti stvarni sadržaj napisanog teksta, ne više vizualnog izgleda dokumentacijske informacije nego riječi i njihove skladbe (sintakse) što je više orijentirano jezikoslovnim istraživanjima. Dakako, takvo označivanje može se s lakoćom povezati i sa značenjem (semantikom) promatranog teksta, samo će u tom slučaju domena biti drugačija. Možemo naprimjer za domenu X uzeti skup boja koje se preslikavaju u kodomenu Y1, koja je skup vrste riječi (POS – Part Of Speech) ili se preslikavaju u Y2 kodomenu koja sadrži elemente sentiment-analize (od jako negativnih do jako pozitivnih osjećaja). Onda možemo u općem slučaju funkciju POS definirati kao *POS (x ε boje) = y ε vrste_riječi*, gdje simbol *ε* znači ‘element iz’, a konkretno naprimjer: *POS (crveno) = glagol* ili *POS (modro) = imenica*.

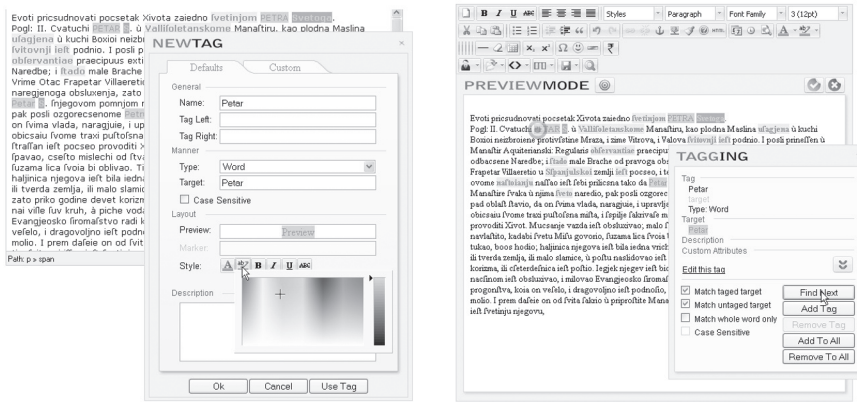
Ako bismo funkcijom *SMT(crveno)=stidljivost* označili skup riječi koje se koriste u opisu stidljivosti, imali bismo pri označivanju već spomenuti problem (ne bismo znali je li označena riječ glagol ili je vezana uz senti-

ment). Matematički gledano ne bismo imali preslikavanje ‘1 na 1’, što je uvjet svake funkcije. Problem se rješava na dva načina: stvara se nova domena ili se funkcija kategorizira, razdvaja u slojeve. U prvom slučaju, domena boja može se razdvojiti, pa prvi dio spektra služi za POS-oznake, a drugi za SMT-oznake ili se u oba slučaja koriste sve boje za označivanje riječi ili izraza, ali u posebnim slojevima. Slojevi su u *TEIMark*u uređeni hijerarhijski do bilo koje dubine, a također i skupovi oznaka. Ručno upravljanje oznakama, slojevima i prikazbom označenog teksta dokumenta načinjeno je vrlo intuitivno. Tako postoji oznaka ‘oka’ na čiji se klik neki sloj ili oznake prikazuju ili sakrivaju.

Funkcionalno gledajući oznake mogu biti “vezane” ili “slobodne”, ovisno o tomu pridružujemo li pri njihovu stvaranju vizualnoj kategoriji (boji, precrtavanju, podcrtavanju, fontu, ukošenosti, pojačanosti i sl.) značenjsko obilježje riječi ili ne. Postupak označivanja moguć je klasičnim, ručnim označivanjem/markiranjem (povlačenjem miša uz pritisnutu lijevu tipku miša) ili poluautomatskim (strojnim) označivanjem prepoznatih riječi ili izraza. Pritom se vizualna oznaka “lijepi” na riječ, točku između dva znaka neke riječi i na izraze (fraze). Poluautomatsko, tj. interaktivno ili automatizirano označivanje fraza riješeno je u smislu naprednog pretraživanja, pa će fraza biti označena bez obzira na to je li njen niz znakova cjelovit ili razlomljen (Slika 3). Naprimjer fraza *zaboravljeni počeci* bit će označena i u nizu *zaboravljeni davni počeci* ili *zaboravljeni pa obnovljeni počeci*.

Budući da se pretraživanje provodi nad nizom znakova spremljenih kao ASCII-kodovi, tj. brojevi, ovim alatom nije moguće detektirati fraze u kojima je došlo do morfoloških promjena, npr. fraza *zaboravljen početak* ili *zaboravljeni počeci* neće biti označeni. Taj problem pokušali smo riješiti razvitkom novih modula koji vode računa o oblikoslovnim, ali i semantičkim obilježjima, pa otkrivaju i fraze koje uključuju istoznačnice.

Tekst koji se označi uz pomoć *TEIMarka* može se izvesti (spremiti) u različitim formatima, kao što su XML i njegova TEI-varijanta, JSON i slično.



Slika 3. Stvaranje oznake i automatsko *online*-označivanje s definiranom vezanom oznakom

HRVATSKA KULTURNA BAŠTINA UNUTAR VREMENSKO-PROSTORNIH KOORDINATA

Označiteljska funkcija ne mora imati samo po jednu ili dvije domene/kodomene, kako smo dosad u označiteljskim alatima u članku pokazivali, već po volji mnogo. Mrežni okvir za hrvatsku kulturnu baštinu zamišljen je i realiziran tako da povezuje mnogobrojne funkcijske skupove u koherentnu cjelinu omogućujući na taj način najbolju svezu do sada nepovezanih područja.

Ako pomoću T označimo skup godina, npr. bilo koji vremenski trenutak od stoljeća sedmog do danas, pomoću G skup zemljovidnih koordinata, pomoću A skup autora, a pomoću K skup kulturoloških područja, onda će funkcija

$$HKB\ (T, G, A, K) \qquad (1)$$

označiti djelo nekog autora iz nekog kulturološkog područja u vremensko-prostornoj koordinati po želji precizne koordinate.

Osnovno grafičko sučelje programskog rješenja Mrežnog okvira hrvatske kulturne baštine (MHKB) zamišljeno je i realizirano kao spoj triju velikih cjelina: vremenske osi, geografske Googleove karte i tražilice sa spomenutim A- i K-obilježjima (domenama), ali i po kategorijama medija

Slika 4. Korisničko sučelje sustava MHKB

u kojem se digitalizirano nacionalno blago čuva, npr. slici, tekstnom dokumentu, zvučnom ili videozapisu i sl.

Mrežni okvir posjeduje javni (dostupan svima) i privatni (istraživački) dio koji je dostupan samo korisnicima s dopuštenjem (s korisničkim imenom i zaporkom). U javnom dijelu zasad su samo dokumenti više desetaka hrvatskih listina od 13. do 18. stoljeća, koje priređuje i održava[22] dr. sc. Jurica Budja s Instituta za hrvatski jezik i jezikoslovlje, u sklopu projekta *Arealno-dijakronijskog korpusa hrvatskoga jezika* (Slika 4).

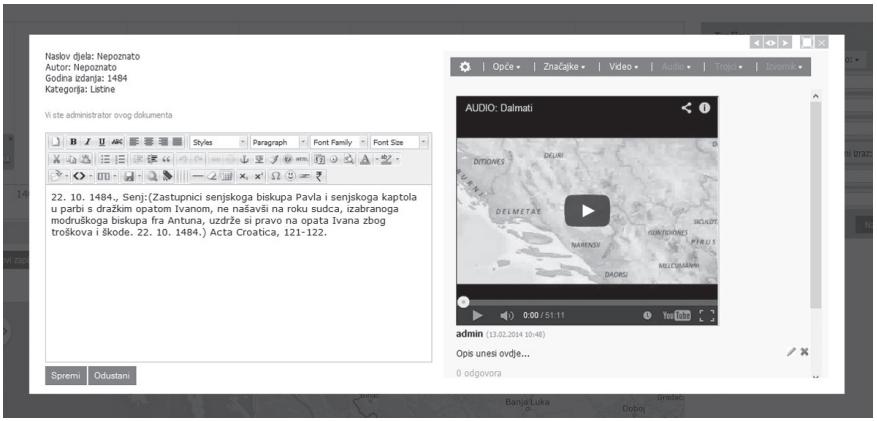
Vremenska os omogućuje zumiranje i pomak, pa se vidljivi dokumenti na njoj mogu prikladno vizualizirati na gruba (stoljeća i desetljeća) ili finija (godine, mjeseci, dani) vremenska razdoblja. Budući da kulturnu baštinu čine dokumenti iz različitih kulturoloških područja, oznake dokumenata na vremenskoj osi prikazane su u različitim bojama (npr. za kategoriju književnosti – žuta, arhitekture – bijela, glazbe – zelena, listine – plava itd.). Isto tako u svakoj kategoriji informacija se može spremiti u različitom mediju, naprimjer rukopis se može spremiti kao digitalizirana slika ili kao prepoznati (OCR) tekst, a može se spremiti i kao zvuk ako je pročitan i zapisan u kojem od zvučnih formata. Zato je ispod osnovne informacije dokumenta (naslova, autora, godine i mjesta), u desnom kutu svake oznake, ikonicom prikazana vrsta informacije (medij) u kojoj je dokument pohranjen. Kod

[22] http://bozoou.com/timeline/ Internet. 30. siječnja 2016.

spremanja (prebacivanja) dokumenta u sustav, vlasnik dokumenta definira osnovnu informaciju po kojoj je kasnije moguće pretraživanje.

OBILJEŽJA DOKUMENATA

Otvaranje i prikaz pronađenog/odabranog dokumenta na lijevoj strani prozora prikazuje sadržaj dokumenta (bilo kojeg medija), a na desnoj pridružena obilježja (npr. popis riječi u tekstu dokumenta s frekvencijom pojavaka; opis dokumenta preko video- ili audiozapisa; komentare drugih korisnika i sl.). Svaka se strana (osnovne i dodatne informacije) može ugasiti (sakriti), odnosno proširiti na cijeli zaslon (Slika 5).



Slika 5. Temeljni (lijevo) i popratni sadržaj (desno)

Ako je dokument otvorio vlasnik, on će imati mogućnost njegova uređivanja standardnim mrežnim uređivačem (Wordu sličnim editorom), koji je proširen mogućnošću spomenutog vizualnog označivanja/tagiranja (*TEIMark*) i dodatnih funkcija (npr. gumba – alata za upis bilo kojeg znaka UTF-8, čime se obuhvaćaju svi poznati znakovi, naglasci, glagoljička i ćirilična slova te posebni znakovi).

Slika 6. Praćenje hrvatskih riječi kroz stoljeća

PRAĆENJE HRVATSKE RIJEČI KROZ VRIJEME

Kodomena funkcije *HKB()* prema izrazu (1), osim digitaliziranog dokumenta dohvaćenog iz riznice kulturnog blaga, može također biti i obična riječ koja se pojavljuje u jednom ili više dokumenata. Takvo označivanje pa pretraživanje daje informaciju o životu neke riječi, njena nastanka i/ili nestanka, kao što se vidi na Slici 6.

Ispod trajektorija pojavaka riječi ispisuje se popis rečenica u kojima se prikazane riječi nalaze, a postoji i mogućnost otvaranja dokumenta (klikom na ikonicu knjige) u kojoj se ta rečenica nalazi.

DIJALEKTOLOŠKO BLAGO HRVATSKOGA JEZIKA

Ako se kao domena uzme skup rečenica s nekim jezikoslovnim obilježjima, a kodomena zemljovid, onda se s pomoću istoga alata (unutar okvira MHKB-a) dobije zvučni atlas[23] kakav su priredili i postojano ga upotpunjuju prof. dr. sc. Velimir Piškorec i njegova suradnica dr. sc. Ivana Kurtović Budja. Cilj je takvog atlasa da korisnik može za istu rečenicu ili izraz čuti izgovor u hrvatskim idiomima svih triju narječja. Uz vrlo intuitivno korisničko sučelje

[23] http://hrvatski-zvucni-atlas.com/interaktivni-zemljovid Internet. 30. siječnja 2016.

razrađeno je i administratorsko sučelje u kojem autor tonskog zapisa na jednostavan način objavljuje zapis na Googleovoj karti, manipulira zvukom (pojačava ili smanjuje jakost), dodaje opis i slično. Razvijena je čak i mogućnost da se zvuk automatski snima iz udaljene lokacije putem interneta, pa je za skupljanje tonskih zapisa dovoljno računalo s mikrofonom i spojem na internet. I, dakako, trud ljubitelja hrvatske riječi koji će okupiti ljude koji će se angažirati oko tog mrežnog okvira i trajno sačuvati svoj govor i kulturu.

MREŽNO HRVATSKO JEZIKOSLOVLJE

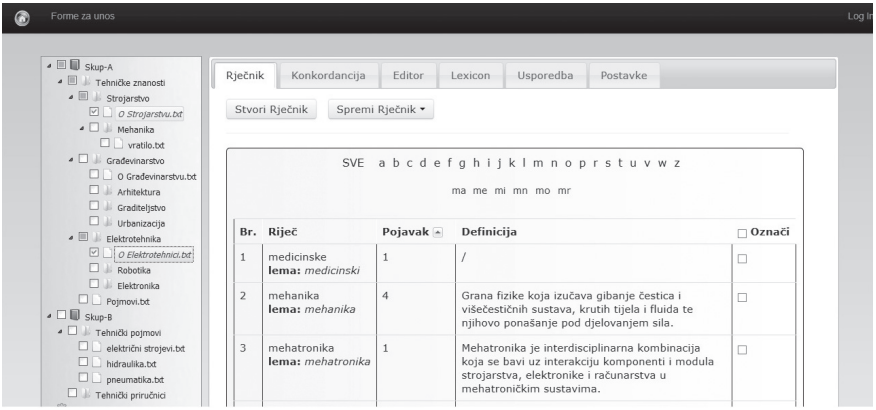
Dosad opisani MHKB-ov okvir koristi više jezikoslovnih obilježja koja su stvorena izvan njega, naprimjer čestotni rječnik nekog dokumenta, izvučena sintaktička ili semantička obilježja, koristi poveznice s mrežnim pojmovima drugih baza i digitalnih repozitorija i sl. Iako sastavljen od više modula središnje mjesto pripada generatoru rječnika koji iz bilo kojeg digitaliziranog teksta izvlači riječi i slaže ih po abecednom redu ili prema broju pojavaka riječi, uzlaznog ili silaznog smjera. Dakako, i tu se radi o označivanju, ali ovog puta strojnim metodama gdje se tekst prvo treba raščlaniti u riječi (problemi interpunkcijskih znakova i kratica), a potom razvrstati različite riječi i složiti ih po nekom ključu. Kako se lako vidi, i rječnik je samo kodomena, skup raznovrsnih riječi, koji se još može preslikavati u sažetiji skup, npr. skup lema. Za lematizaciju različitih prirodnih jezika postoje mnogi algoritmi koji najčešće rade po statističkim zakonima, pa stoga nisu posve precizni. Mi smo izabrali puno teži, inverzni put: načinili smo program koji iz leme bilo koje promjenljive vrste riječi stvara sve njene gramatičke oblike. Jednom generirana i spremljena riječ u svim svojim pojavnostima ne mora se više nikad obrađivati. Načinjeni program studenta J. Markučiča[24] uključuje sve algoritme glasovnih promjena i tvorbe riječi hrvatskoga jezika, pa je postao idealan alat za stvaranje e-rječnika.

To su postigli i studenti Juraj Benić i Jakov Topić mrežnim programom[25] (Slika 7.) koji stvara takve rječnike i povezuje ih s mrežnim rječnicima drugih izvora, npr. s HJP-om, LZMK-om, Strunom[26] i dr. Iako

[24] https://jmarkucic.pythonanywhere.com/morf/default/imenice Internet. 30. siječnja 2016.

[25] https://jt195996.pythonanywhere.com/Test_FancyTree/default/index Internet. 30. siječnja 2016.

[26] http://hjp.novi-liber.hr/, http://hol.lzmk.hr/, http://struna.ihjj.hr/ Internet. 30. siječnja 2016.



Slika 7. Sustav MHJ za jezikoslovna istraživanja

proširiv na sva područja, program je ograničen samo na tehniku i opisan u nagrađenom radu "Mrežni program za tvorbu i obradbu tehničkih rječnika"[27].

S funkcionalnog stajališta taj rad proširuje mogućnosti postojećih mrežnih rječnika (Anić, Jojić, Matasović 2003) i repozitorija u Hrvatskoj: dodaje gramatičke oblike riječi, otkriva stručne riječi u nekom tekstu (napisanom ili povučenom s mreže), ispisuje definicije i druge atribute stručnih riječi iz teksta koje je pronašao u nekom od mrežnih leksikona, omogućuje obradbe teksta (čestotnost, konkordancija, statistika...), daje mogućnost unosa riječi kojih u tehničkom leksikonu ili enciklopediji nema, a nalaze se u stručnom tekstu i slično.

U njemu je također organiziran potpun sustav mrežnih korisnika, tj. izvedena su hijerarhijska dopuštenja za više korisničkih skupina: administratori koji mogu izvršavati sve operacije i dodjeljivati dopuštenja korisnicima, skupine korisnika koji imaju pravo unosa novih riječi ili promjena u leksikonu te korisnici koji imaju samo dopuštenja gledanja, ali ne i editiranja informacije.

Ista načela mrežnog okvira koristili smo i u repozitoriju metafora CMR[28] (HRZZ-ov projekt UIP-11-2013, voditeljica dr. sc. Kristina Štrkalj

[27] Rektorova nagrada, Sveučilište u Zagrebu, 2015.

[28] https://metafora.ihjj.hr:8443/Metafore/ Internet. 30. siječnja 2016.

Despot) na Institutu za hrvatski jezik i jezikoslovlje, samo su domene i kodomene označiteljskih funkcija ovog puta uzete iz semantičkih skupova apstraktnih, izvorišnih (engl. *source*) i realnih, odredišnih (engl. *destination*) značenjskih domena.

STVARANJE TROJACA POVEZANIH PODATAKA

Razvijen je i zaseban program kojim izvlačimo željenu informaciju iz bilo kojeg teksta stvarajući 's-p-o' (subjekt-objekt-predikat) trojce (engl. *triplets*), iako navedeni pojmovi ne moraju stvarno predstavljati službu riječi u rečenici, već su više povezani sa sintaktičko-semantičkim kategorijama.

Izvlačenje informacije iz teksta riješeno je u dva prolaza preko uzoraka i izraza temeljenih na regularnim izrazima. Prvo se definiraju uzorci koji opisuju neku riječ preko stringa ili gramatičkog svojstva koje riječ ima (i nalazi se u našem označenom rječniku s oko 1,5 milijuna riječi, uključujući i sve gramatičke oblike). Nakon toga se definiraju 'extract_izrazi' koji koriste zadane uzorke, tj. s regularnim izrazima nalaze međusobne veze uzoraka. Programska funkcija:

CUPAJ (reg(rijeci), reg(gram_oblika), AND_OR, 'dodatni argumenti') (2)

obradit će bilo koju rečenicu nekog teksta na taj način da će u njoj pronaći sve one uzorke koji su zadani po regularnom izrazu stringa neke riječi ili njezina gramatičkog oblika, vodeći računa o oba (engl. *AND*) ili pojedinom (engl. *OR*) zahtjevu te uvažavajući pritom i dodatne argumente.

Dodatni argumenti odnose se na poziciju drugih uzoraka u tekstu u odnosu na traženi uzorak.

Na taj način iz rečenice (kao niza riječi – stringova) u prvom prolazu načini se jedan jedini string koji na mjestima pronađenih uzoraka ima njihovu oznaku, a na mjestu riječi koje ne odgovaraju uzorku ima znak '-'. U drugom prolazu "napada" se samo taj string ponovo regularnim izrazima, pa ako uvjet bude zadovoljen, dobiva se željena informacija koja se potom lako može spremiti u neki format RDF (Chiarcos 2012). Na opisani način spremaju se do sada svi trojci kao LOD-literali dokumenata ili repozitorija, npr. CMR-a. Ontologije koje će s njima, tj. nad njima raditi, još se razvijaju; nažalost, dosad to u državnim institucijama ide jako sporo.

Druga je važna pretpostavka ta da računalni okvir bude podudaran, kompatibilan s ostalima. Iako od ranih 80-ih godina prošlog stoljeća do da-

nas imamo vrlo mnogo razvijenih mrežnih okvira (npr. Acquilex, Multilex, Genelex, Eagles, Isle, Mile i dr.), danas se teži postići standard poznat kao LMF – Lexical Markup Framework (Francopoulo 2013). On omogućuje povezivanje čak i semantičke informacije (Fontenelle 2012), poput Wordneta, u globalnu mrežu[29]. Takvo rješenje potaknulo bi kvalitetniji razvitak hrvatskog Wordneta s obzirom na ovaj koji trenutačno imamo (Šojat 2009; Raffaelli, Katunar 2012).

BUDUĆI RADOVI

Konstantno potiranje sinonimije, po mišljenju stručnjaka[30] (Šarić, Wittchen 2008), te svođenje gramatičkih oznaka na univerzalne matrice (npr. MULTEXT-EAST[31]) osiromašuje hrvatski jezik i svrstava ga u skupine "minijezika"[32] umjesto da se potiču njegove vrijednosti i divljenje njegovoj čarobnosti. U želji da se to promijeni radi se na novom hrvatskom tezaurusu (Orešković, Essert 2016) koji, umjesto dosadašnjeg označivanja, uvodi raščlanjena gramatičko-semantička obilježja (tzv. WOS/SOW T-strukture) kojim se elegantno postižu značenja kompleksnih riječi (Lieber 2009) kao što su izvedenice, složenice ili riječi podvrgnute različitim vrstama pretvorbi, a sve uz pomoć temeljnih morfemskih značenja (vidi Sliku 8).

Tezaurus povezuje svu relevantnu mrežnu informaciju (Fribley 2012) koju u Hrvatskoj imamo, vodi računa o višeznačnostima riječi, njihovim naglascima, terminologiji, jezičnim savjetima i slično (Geeraerts 2010).

ZAKLJUČAK

Nacionalno kulturno i jezikoslovno blago nije dovoljno čuvati u izoliranim računalnim sustavima, već je nužno potrebno uključiti ga u globalne, svjetske okvire, kako bi Hrvatska što prije izašla iz "geta" koji joj je desetljećima bio nametnut.

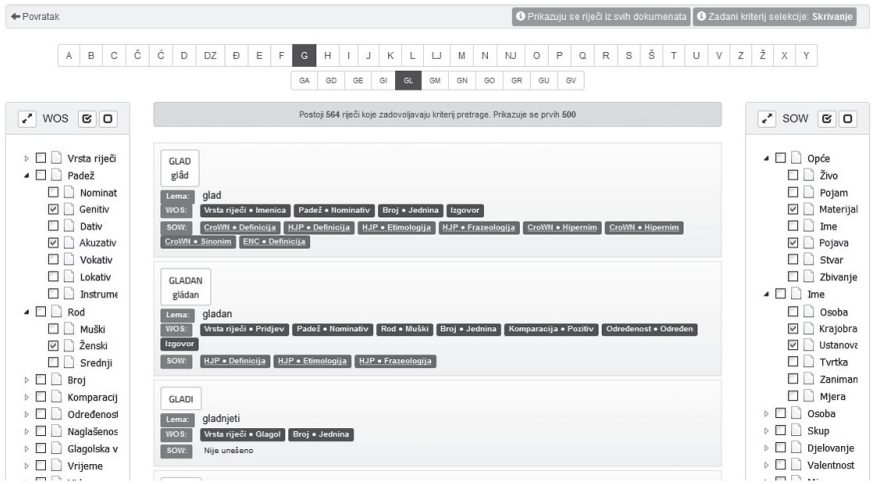
---

[29] http://globalwordnet.org/ Internet. 30. siječnja 2016.

[30] http://www.hkv.hr/kultura/jezik/4382-potiranje-sinonimije-znai-siromaenje-jezika.html Internet. 23. travnja 2016.

[31] http://nlp.ffzg.hr/data/tagging/msd-hr.html Internet. 30. siječnja 2016.

[32] Izraz zabilježen na jednoj *Jezikoslovnoj raspravi* u IHJJ-u.



Slika 8. Mrežni hrvatski tezaurus s obilježjima WOS/SOW

Ovaj rad sa svojim računalnim rješenjima na tragu je te težnje i svojom sveobuhvatnošću daje nadu u mogući uspjeh. Tri velika računalna sustava na jednostavan, intuitivan način rješavaju potrebe običnog korisnika, ali i zahtjevnog istraživača. Među računalnim modulima ističu se alati za vizualno označivanje (*DocMark* i *TEIMark*), repozitorij svih vrsta dokumenata kulturne baštine (po kategorijama i mediju) smještenih u vremensko-prostorne koordinate (arealno-dijakronijskog korpusa i zvučnog atlasa hrvatskih govora), praćenje riječi kroz vrijeme, morfološki generator promjenljivih vrsta riječi, generator naprednog e-rječnika (iz bilo kojeg digitalnog dokumenta), priprema i stvaranje LOD-podataka (s repozitorijem metafora kao prvim primjerom uporabe) te na koncu hrvatski tezaurus s novom strukturom obilježja riječi.

LITERATURA

Anić, V., Jojić, Lj., Matasović, R. 2003. *Hrvatski enciklopedijski rječnik*. Novi liber, Zagreb.

Chiarcos, C., Nordhoff, S., Hellmann, S. 2012. *Linked data in linguistics representing and connecting language data and language metadata*. Springer, Berlin.

Francopoulo, G. 2013. *LMF Lexical Markup Framework*, John Wiley & Sons, ISTE.

Fribley, K. 2012. *Find the right words with thesauruses*. Ann Arbor, Mich., Cherry Lake Pub.

Fontenelle, T. 2012. Wordnet, Framenet and Other Semantic Netorks in the International Journal of Lexicography – the Net Result? International Journal of Lexicography, Vol. 25, str. 437–449.

Geeraerts, D. 2010. *Theories of lexical semantics*. Oxford; New York: Oxford University Press.

Gelbukh, A., Kolesnikova, O. 2014. *Semantic. Analysis of Verbal Collocations with Lexical Functions,* ISBN 978-3-642-28771-8, Springer-Verlag Berlin Heidelberg.

Hyvönen, E., Makela, E. et al. 2009. *CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0*, L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, str. 851–856.

Ioannides, M. et al., 2014. *Digital Heritage – Progress in Cultural Heritage: Documentation, Preservation, and Protection*, 5th Int. Conference, EuroMed, November 3–8, Cyprus.

Jackendoff, R. 2002. *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press, New York.

Lieber, R. 2009. *Morphology and lexical semantics*. Cambridge University Press.

Mel'Cuk, I. 2015. *Semantics*., ISBN 978-90-272-5933-2, John Benjamins.

Orešković, M., Essert, M. 2016. *Croatian Lexicon inside Syntax & Semantic Framework*, International Congress "Word format and lexical combinations", Roma.

Orešković, M., Juraj Benić, J., Essert, M. 2016. *Network integrator Croatian lexicographical resources*, prihvaćeno za XVII EURALEX Int. Congress 6–10 September, Tbilisi, Georgia

Pustejovsky, J. 1998. *The generative lexicon*. MIT Press, Cambridge, Massachusetts.

Raffaelli, I., Katunar, D. 2012. *Lexical-Semantic Structures in Croatian WordNet*. Filologija, Vol. No. 59.

Saint-Dizier, P., Viegas, E. 1995. *Computational lexical semantics*. New York: Cambridge University Press.

Šarić, Lj., Wittschen, W. 2008. *Rječnik sinonima hrvatskoga jezika*. Naklada Jesenski i Turk. Čakovec.

Šojat, K. 2009. *Morphosyntactic annotation in the Croatian Wordnet.* Suvremena lingvistika. Vol. 35, No. 68.

Tomić, M. 2013. *Organizacija i aproprijacija tekstova hrvatskoglagoljskih brevijara na razmeđu rukopisne i tiskane tradicije*, doktorski rad, Zagreb.

Tomić, M., Glumac, N., Essert, M. 2010. *Označivanje i analiza digitaliziranog dokumenta*, AKM 14, str. 2, Poreč.

SUMMARY

NETWORK FRAMEWORK FOR RESEARCH OF CROATIAN CULTURAL HERITAGE PREPARED FOR LINKED DATA

This paper presents a framework for archiving and analysing documents from various categories of Croatian cultural heritage, e.g. literature, painting and architecture, stored in various different media, e.g. digitalised handwriting, text, images, sound recordings and movies. The framework enables easy categorisation in temporal and spatial coordinates of digital recordings with various properties and contains a search functionality with many different search criteria. Apart from classical bibliographic search, words can be tracked in time, meaning their evolution can be tracked through centuries, from when they first appeared, to when they vanished and reappeared again. A visual editor called TEIMark was developed for marking of syntactic and semantic data, while another editor called DocMark was developed for the marking of visual data, e.g. digitalised handwriting. Both editors support visual markings, or tags above the information, e.g. words or images which are organised in a series of layers that can be hidden, shown or saved in the XML/TEI format. Each document can contain its own set of triplets which can be searched with the Virtuoso triple store database using SparQL commands. The presented network framework also contains a development system for linguistic text analysis, and a program called IExtract, which extracts the s-p-o information from a set of sentences using user-defined patterns. The development system enables semi-automatic creation of alphabets and dictionaries, which can then be connected using the linked data paradigm to existing on-line dictionaries. This forms a foundation for future ontology systems that connect such data (LOD) in a global network cloud.

Keywords: tools for visual annotations, semantical framework, information extraction, linguistics and heritages open linked data