УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЗОФСКИ ФАКУЛТЕТ
ДС/СС 05/4-02 бр. 397/2-XVII/17
08.03.2016. године

ВЕЋЕ НАУЧНИХ ОБЛАСТИ
ДРУШТВЕНО-ХУМАНИСТИЧКИХ НАУКА

Наставно-научно веће Филозофског факултета у Београду је на својој III редовној седници, одржаној 08.03.2016. године – на основу чл. 202. став 1. алинеја 14. и 15. Статута Факултета, прихватило Извештај Комисије за докторске студије с предлогом теме за докторску дисертацију: ИМЕНОВАНИ ЕНТИТЕТИ У ЛАТИНСКОМ ЈЕЗИКУ, докторанда Марине Голуб.

За ментора је одређен проф. др Војин Недељковић.

ПРЕДСЕДНИК ВЕЋА

Проф. др Војислав Јелић

Доставити:
1x Универзитету у Београду
1x Стручном сараднику за докторске дисертације
1x Шефу Одсека за правне послове
1x Архиви

Факултет Филозофски 04/1-2 бр. 6/14 (број захтева) 8.03.2016. (датум)	УНИВЕРЗИТЕТ У БЕОГРАДУ Веће научних области друштвено-хуманистичких наука (Назив већа научних области коме се захтев упућује)

З А Х Т Е В
за давање сагласности на предлог теме докторске дисертације

Молимо да, сходно члану 46. ст. 5. тач. 3. Статута Универзитета у Београду («Гласник Универзитета», бр. 131/06), дате сагласност на предлог теме докторске дисертације:

Именовани ентитети у латинском језику

(пун назив предложене теме докторске дисертације)

НАУЧНА ОБЛАСТ класична филологија

ПОДАЦИ О КАНДИДАТУ:

Име, име једног од родитеља и презиме кандидата:

Марина, Марин, Голуб

Назив и седиште факултета на коме је стекао високо образовање: Филозофски факултет, Универзитет у Београду

Година дипломирања: 2012.

Назив мастер рада кандидата: Трагови латинске синтаксе у грчком преводу Рес гестае диви Аугусти

Назив факултета на коме је мастер рад одбрањен: Филозофски факултет, Универзитет у Београду

Година одбране мастер рада: 2013.

Обавештавамо вас да је Наставно-научно веће

на седници одржаној 8.03.2016.

размотрило предложену тему и закључило да је тема подобна за израду докторске дисертације.

	ДЕКАН ФАКУЛТЕТА Проф. др Војислав Јелић

Додатак уз Образац 1

ПОДАЦИ О МЕНТОРУ

за кандидата Марину Голуб

Име и презиме ментора: Војин Недељковић

Звање: ванр. професор

Списак радова који квалификују ментора за вођење докторске дисертације:

1. *Увод у проучавање вулгарног латинитета. I: Феноменологија и извори*, Београд 2012.
2. *Нацрт класичног латинског вокализма и прозодије*, Нови Сад—Сремски Карловци 2008.
3. »From Funerary Poetry to Vulgar Prose? The Case of IMS 4.39«, *Illinois Classical Studies* 31 (2006), 114–127.
4. »Приповедачи и проповедници. Скица за једну статистичку студију латинског језичког варирања« *Lucida intervalla* 42 (2013), 79–108.
5. »Натписи позноантичког Ниша као предмет вулгарнолатинских студија« *ЗРВИ* 50 (2013), 45–63.

Заокружити одговарајућу опцију (А, Б, В или Г):

А) У случају менторства дисертације на докторским студијама у групацији техничко-технолошких, природно-математичких и медицинских наука ментор треба да има најмање три рада са SCI, SSCI,
AHCI или SCIe листе, као и Math-Net.Ru листе.

Б) У случају менторства дисертације на докторским студијама у групацији друштвено-хуманистичких наука ментор треба да има најмање три рада са релевантне листе научних часописа (Релевантна листа научних часописа обухвата SCI, SSCI, AHCI и SCIe листе, као и ERIH листу, листу часописа које је Министарство за науку класификовало као М24 и додатну листу часописа коју ће, на предлог универзитета, донети Национални савет за високо образовање. Посебно се вреднују и монографије које Министарство науке
класификује као М11, М12, М13, М14, М41 и М51.)

В) У случају израде докторске дисертације према ранијим прописима за кандидате који су стекли академски назив магистра наука ментор треба да има пет радова (референци) које га, по оцени Већа научних области, квалификују за ментора односне дисертације.
Г) У случају да у ужој научној области нема квалификованих наставника, приложити одлуку Већа докторских студија о именовању редовног професора за ментора.

ДЕКАН ФАКУЛТЕТА

Датум _____ М.П. **проф. др Војислав Јелић**

OBRAZLOŽENJE PREDLOGA TEME DOKTORSKE DISERTACIJE

Imenovani entiteti u latinskom jeziku

Predmet istraživanja

Predmet istraživanja predložene doktorske disertacije jesu imenovani entiteti u latinskom jeziku. Termin imenovani entiteti koji se koristi u obradi prirodnih jezika (*Natural Language Processing – NLP*) nastao je relativno nedavno, 1995. godine za potrebe Šeste konferencije o razumevanje poruka – *MUC*-6 (*Message Understanding Conference*). Ove konferencije je osamdesetih godina 20. veka inicirao *NOSC (the Naval Ocean System Center)* radi analize vojnih poruka koje sadrže tekstualne informacije.

Zadatak učesnika bila je ekstrakcija informacija (*Information Extraction – IE*) iz slobodnog, nestrukturiranog, teksta, tj. identifikacija specifičnih podataka u tekstovima prirodnog jezika (npr. u novinskim člancima), kao i kasnija semantička klasifikacija i strukturiranje tih podataka, radi njihove efikasnije obrade.

Na *MUC* konferencijama organizovanim sredinom devedesetih godina prošlog veka, uz vodeću podršku *DARPA*-e (*the Defense Advanced Research Projects Agency*), primarni zadatak je bila ekstrakcija informacija vezanih za aktivnosti različitih kompanija i organizacija, kao i onih u vezi sa pitanjima bezbednosti i odbrane. Recimo, na Četvrtoj konferenciji o razumevanju poruka *MUC*-4, održanoj 1992. godine, učesnici su kao zadatak imali da iz novinskih tekstova "izvuku", tj. da pronađu, obeleže i klasifikuju specifične informacije vezane za terorističke napade u Južnoj Americi.

Prilikom razvijanja sistema za *IE* učesnici su primetili da je veoma važno prepoznavanje informacija koje predstavljaju lična imena i numeričke podatke, tj. onih informacija koje jasno identifikuju jedan elmenat iz seta drugih elemenata koji imaju slične atribute. Prepoznavanje i ekstrakcija ovih entiteta su tada postali jedan od važnih podzadataka unutar *IE*.

Klasifikacija imenovanih entiteta (*NE*)

Isprva, imenovani entiteti obuhvatali su 3 kategorije, odnosno 7 potkategorija – **lična imena**, u koja spadaju imena osoba, lokacija i organizacija, **temporalni izrazi** (vreme i datum) i **numerički podaci** (novčane i procentualne vrednosti). Ipak, ispostavilo se da ovih 7 tipova nije dovoljno da pokrije sve *NE* i postepeno je broj tipova rastao tako da se od 7 došlo do 200. Ovde ćemo navesti samo neke od tih

naknadno dodatih tipova: reka, jezero, muzej, aerodrom, adresa, e-mail adresa, broj telefona, naziv knjige itd.

Ciljevi istraživanja

Cilj našeg istraživanja je da utvrdimo u koliko je meri pitanje ekstrakcije informacija, preciznije imenovanih entiteta relevantno i izvodljivo u latinskom, budući da je trenutno, barem koliko je nama poznato, ova oblast računarske lingvistike u latinskom poprilično nerazvijena.

Prepoznavanje i klasifikacija imenovanih entiteta – *NER(C)* (*Named Entity Recognition and (Classification)*) su u najvećoj meri razvijeni za engleski. Osim u engleskom, na ovom pitanju se mnogo radilo i u nemačkom, španskom, japanskom, kineskom, grčkom i italijanskom jeziku. Takođe, u srpskom, bugarskom, rumunskom, korejskom, turskom, švedskom, portugalskom itd. ovom problemu je posvećena značajna pažnja.

Za korpusne jezike poput latinskog i grčkog, situacija je vrlo specifična – već krajem 19. veka, na Mommsenovu inicijativu, nastao je leksikon ličnih imena *Prosopographia Imperii Romani* (*PIR*), leksikon imena osoba koji se bavi prosopografijom Rimskog carstva, od kraja I veka pre Hr. pa sve do kraja III veka nove ere, trenutno najstariji i najduže vođeni prosopografski poduhvat u Evropi. Dugotrajnost ovog projekta je jasan pokazatelj važnosti entiteta poput ličnih imena zahvaljujući kojima je moguće utvrditi socijalni kontekst u kom je određena osoba živela, njeno poreklo (etničko ili regionalno), karijeru i porodične veze. Osim toga, njegova dugotrajnost nam, naravno, ukazuje i na činjenicu da je proces prikupljanja ovakvih podataka rukom, proces koji iziskuje mnogo vremena i truda.

Sa početkom digitalne i *ICT* (*Information and Communication Technology* – informacijska i komunikacijska tehnologija) revolucije šezdesetih godina 20. veka javila se ideja o mogućnosti digitalizacije prosopografskog materijala u kompjuterske, lako pretražive baze. Međutim, već sredinom sedamdesetih godina, upravo na primeru *PIR*, ispostavilo se da mogućnosti tadašnje tehnologije još uvek

nisu dorasle željama istraživača. I pored napretka tehnologije u naredne dve ili tri decenije, digitalne prosopografske baze ipak nisu doživele očekivani razvoj.

Tek 2008. ili 2009. godine, koliko je nama poznato, istraživači su uvideli da bi automatski sistemi za prepoznavanje *NE* mogli biti odličan metod za prikupljanje ovakvh vrsta podataka u grčkom jeziku, budući da je ručno prikupljanje izuzetno dugotrajno. U latinskom, prema našim saznanjima, primena sistema za *NER* počinje tek 2014.

Nešto veoma slično se može reći i za topografske podatke – odavno se sakupljaju i beleže, prednosti digitalnog doba su takođe vrlo rano počele da se koriste i da se sprovodi digitalizacija ovakvih baza podataka, ali, kako se čini, ovaj proces se odvija paralelno sa procesom razvoja automatskih sistema za prepoznavanje *NE*, umesto da dolazi do njihovog ukrštanja.

Vidimo, s jedne strane da su istraživači koji se bave latinskim jezikom odavno uvideli značaj nekih tipova imenovanih entiteta ali da su, s druge strane, automatski sistemi za njihovo prepoznavanje u latinskom u velikom zaostatku u odnosu na moderne jezike.

Namera nam je da ovom tezom doprinesemo razvoju metoda za *NER* u latinskom jeziku, budući da je odavno prepoznata centralna uloga koju lična imena, toponimi, datumi i slični entiteti imaju za samo razumevanje teksta, ali i za bolje razumevanje političke i socijalne istorije. Takođe, želimo da pokažemo važnost digitalne tehnologije za klasične filologe i ukažemo na prednosti koje nudi u odnosu na tradicionalne metode istraživanja.

Osnovne hipoteze

- Rad počiva na pretpostavci da je prepoznavanje imenovanih entiteta veoma značajan zadatak u oblasti obrade prirodnih jezika, ali i da je nedovoljno razvijeno za latinski jezik. Smatramo da bi razvijanje metoda za prepoznavanje *NE* u latinskom jezikom uveliko olakšalo posao istraživača kada su u pitanju brzina, preciznost i preglednost dobijenih rezultata. Takođe, pretpostavljamo da bi dalja obrada ovako dobijenih rezultata bila efikasnija i

da bi se tako različiti fenomeni, kako istorijski i socijalni, tako i jezički, mogli posmatrati iz drugačije perspektive.

- Verujemo da bi ovo istraživanje moglo pobliže uputiti i zainteresovati istraživače, ne samo za problem imenovanih entiteta, već i za neka druga pitanja vezana za elektronsku obradu jezika, budući da, barem koliko se nama iz ličnog iskustva čini, naša naučna zajednica ne koristi u dovoljnoj meri sve prednosti digitalnog doba.
- Očekujemo da će i pored specifičnosti po kojima se latinski razlikuje od modernih jezika, npr. po osobitom načinu računanja datuma, metod koji smo izabrali biti efikasan.

Metod istraživanja

U našem radu biće zastupljeni i teorijski i praktični aspekti ispitivanja. Svaki tip imenovanih entiteta koji budemo obrađivali prvo ćemo uporediti sa istim tim tipom u nekom od modernih jezika u kom je ovaj problem već obrađivan da bismo videli koje su sličnosti i razlike, kao i potencijalni problem u prepoznavanju ovih entiteta u latinskom. Zatim ćemo na odgovarajućem tipu nestrukturiranih tekstova na praktičnim primerima pokazati kako izgleda to prepoznavanje, na koji način se ono vrši, kao i u čemu se ogleda značaj tog prepoznavanja i gde sve ono može naći primenu.

Kada kažemo nestrukturirani tekstovi ne mislimo na tekstove u kojima su podaci strukturno nekoherentni već prosto na sve pisane i govorne podatke, slike, audio i video forme koje su enkodirane tako da ih računar teško interepretira, za razliku od strukturiranih podataka koje karakteriše lako unošenje i, za računar, jednostavna analiza, npr. tabelarni prikazi i slične baze podataka. Dakle, tekstovi poput književnih, istorijskih, novinskih bi, po toj kategorizaciji, svakako spadali u nestrukurirane.

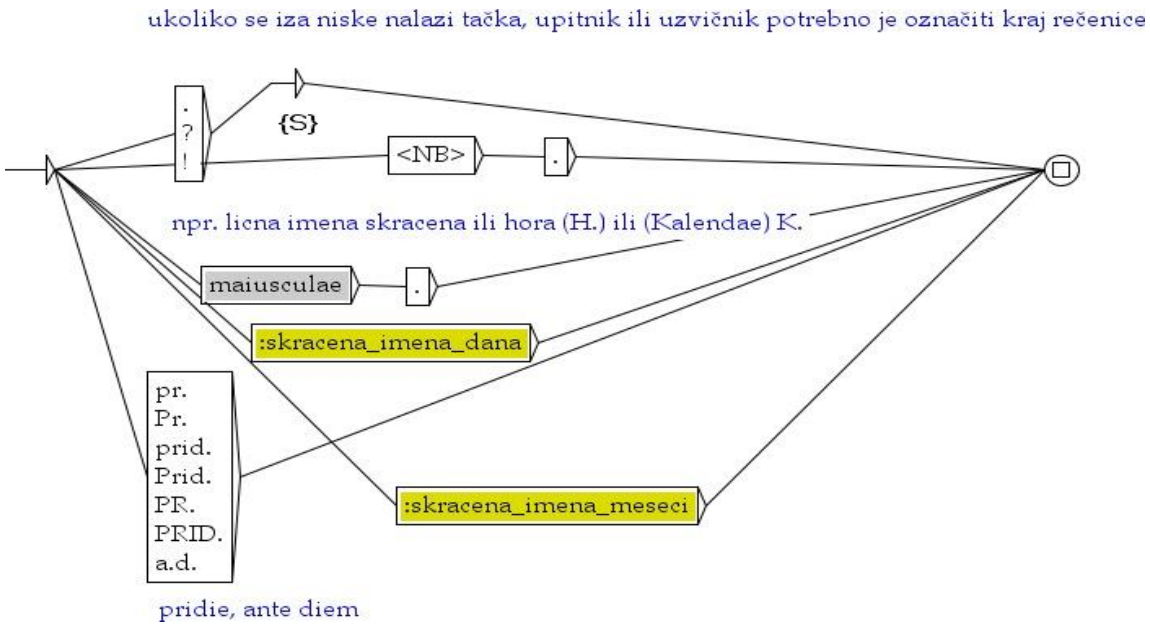
Za praktični deo rada biće nam potreban neki sistem za prepoznavanje imenovanih entiteta. Automatski sistemi za prepoznavanje i anotaciju *NE* u nestrukturiranom tekstu bazirani su ili na elektronskim rečnicima i ručno unapred,

definisanim pravilima (*Handcrafted Rules*), ili na tehnikama nadziranog učenja (*Supervised Learning Techniques*) koje podrazumevaju detaljno ispitivanje odlika imenovanih entiteta na velikom broju uzoraka i kreiranje pravila zasnovanih na tim primerima. Ovakvi sistemi na osnovu primera „uče" da prepoznaju ali i da predvide *NE*, ali su za obuku mašine na ovaj način potrebni veoma veliki korpusi anotiranih tekstova, a i pokazalo se da nisu lako prilagodljivi za razliku od sistema zasnovanih na ručno, unapred definisanim pravilima i elektronskim rečnicima. Neretko se javljaju i neki uspešni hibridni pristupi koji kombinuju prednosti oba metoda.

Sistemi sa unapred definisanim pravilima skoro u potpunosti zavise od jezika koji se obrađuje, jer su bazirani na elektronskim rečnicima, a sama izgradnja ovakvih leksičkih resursa je dugotrajan i zahtevan proces. Prednost ovih sistema se ogleda u tome što korisnik lako može promeniti i modifikovati pravila i na taj način poboljšati učinkovitost programa.

Mi ćemo se ovde služiti programom koji je zasnovan na korišćenju elektronskih rečnika i unapred definisanim pravilima, konkretno Unitexom. Unitex je softver koji se koristi za obradu prirodnih jezika i može se definisati kao skup programa razvijenih za analizu prirodnih jezika uz pomoć jezičkih resursa, npr. elektronskih rečnika i gramatika.

Elektronski rečnici su neophodan leksički izvor u različitim fazama automatske obrade teksta, a gramatike koje se kreiraju i koriste u Unitexu jesu prosto grafički prikazi jezičkih pojava koje korisnik može vrlo lako praviti i modifikovati. Kao primer dajemo nedovršeni grafa kojim se utvrđuje kraj rečenice u latinskom:



Značaj istraživanja

Kao što smo već nagovestili, smatramo da bi naše istraživanje moglo biti od nemale važnosti za razvoj sistema za prepoznavanje imenovanih entiteta, pošto je trenutno ovaj problem, kada se radi o latinskom jeziku, vrlo zapostavljen. Verujemo da bi naš rad mogao biti od značaja i za upućivanje naučne zajednice u ovo pitanje. Cilj nam je da je zainteresujemo i podstaknemo na slična istraživanja ili da joj ponudimo efikasniji, brži i precizniji metod prikupljanja podataka, budući da i sami znamo koliko ovaj zadatak ume da bude težak, i koliko bi bolje poznavanje dostignuća do kojih se došlo u elektronskoj obradi prirodnih jezika moglo umnogome olakšati ovaj posao.

Struktura rada

Rad bi se sastojao iz 10 delova: sadržaja, spiska skraćenica, uvoda, 4 poglavlja, zaključka, rezimea i bibliografije.

Nakon **Sadržaja** i **Spiska skraćenica** koji se nalaze na samom početku sledi **Uvod** koji postavlja okvire teme, daje pojašnjenje metodologije istraživanja, i ispituje trenutna dostignuća u oblasti koja je predmet našeg istraživanja, uz kritički osvrt na postojeću literaturu.

Uvodni deo sledi glavni deo rada – 4 poglavlja od kojih se prva tri bave različitim tipovima imenovanih entiteta u latinskom, a četvrto ovim problemom u epigrafskim spomenicima. U prvom od ova tri poglavlja razmatraćemo entitete koji predstavljaju imena osoba, u drugom topografske, a u trećem numeričke podatke.

Nakon toga sledi **Zaključak** u kome ćemo prikazati rezultate istraživanja i ukazati na pitanja i probleme o kojima bi moglo biti više reči u budućnosti. Na kraju se nalaze **Rezime,** na engleskom jeziku, u kom će biti sumirani rezultati do kojih se došlo tokom istraživanja, i **Bibliografija**, koja predstavlja detaljan popis korišćene literature.

Struktura rada

Sadržaj

Spisak skraćenica

Uvod

- Predstavljanje teme
- Opis metodologije rada
- Kritički osvrt na postojeću literaturu, istorijat i dostignuća dosadašnjih istraživanja

Poglavlja

- Imena osoba
- Toponimi
- Numerički podaci
- Imenovani entiteti u epigrfskim spomenicima

Zaključak

Rezime

Bibliografija

Bibliografija

1. OMAR ALONSO, JANNIK STRÖTGEN, RICARDO BAEZA-YATES, MICHAEL GERTZ, Temporal Information Retrieval: Challenges and Opportunities. *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW)*, 1–8, 2011.
2. AURÉLIEN ARENA, JEAN-PIERRE DESCLÉS, A Formal Ontology for a Computational Approach of Time and Aspect. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, August 16-18 2010, Reykjavik, Iceland, 45–56, 2010.
3. DAVID BAMMAN, GREGORY CRANE, Building a Dynamic Lexicon from a Digital Library, *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 11–20, 2008.
4. TIMOTHY BALDWIN&SU NAM KIM, Multiword Expressions. *Handbook of Natural Language Processing*, Second Edition, Nitin Indurkhya, and Fred J. Damerau (eds.), CRC Press, Taylor and Francis Group, Boca Raton, FL, 267–292, 2010.
5. T. D. BARNES, Prosopography Modern and Ancient. *Prosopography Approaches and Applications: A Handbook*, (ed.) K. S. B. Keats-Rohan, Oxford: Occasional Publications UPR, 71–82, 2007a.
6. T. D. BARNES, Prosopography and Roman History. *Prosopography Approaches and Applications: A Handbook*, (ed.) K. S. B. Keats-Rohan, Oxford: Occasional Publications UPR, 83–93, 2007b.
7. ECKHARD BICK, A Named Entity Recognizer for Danish. *Proc. Conference on Language Resources and Evaluation*, LREC 2004, Lisbon, Portugal, 305–308, 2004.

8. JOACHIM BINGEL, THOMAS HAIDER, Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 26–31 May, Reykjavik, Iceland, ELRA, 2578–2583, 2014.
9. WILLIAM J BLACK, FABIO RINALDI AND DAVID MOWATT, Facile: Description of the NE System Used for MUC-7. *Proc. Message Understanding Conference*, 1998. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/facile_muc7.pdf.
10. MATTHEW BROOK O'DONNELL, STANLEY E. PORTER AND JEFFREY T. REED, OpenText.org: the Problems and Prospects of Working with Ancient Discourse. *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster University (UK), 29 March–2 April 2001, 413–422, 2001.
11. BROUX 2014, YANNE BROUX: Double names in Roman Egypt: A Prosopography Version 1.0, (ed.) W. Clarysse, M. Depauw, *KU Leuven Trismegistos Online Publications*, 2014, http://www.trismegistos.org/top.php.
12. MAYA CARRILLO, ESAÚ VILLATORO-TELLO, AURELIO LÓPEZ-LÓPEZ, CHRIS ELIASMITH, LUIS VILLASEÑOR-PINEDA, MANUEL MONTES-Y-GÓMEZ, Concept Based Representations for Ranking in Geographic Information Retrieval. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, August 16–18 2010, Reykjavik, Iceland, 85–96, 2010.
13. NANCY CHINCHOR, ELAINE MARSH, Appendix D: MUC-7 Information Extraction Task Definition (version 5.1). *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference,* Fairfax, Virginia, April 29–May 1, 1998.
14. PHILIPP CIMIANO, JOHANNA VÖLKER, PAUL BUITELAAR, Ontology Construction. *Handbook of Natural Language Processing*, Second Edition, Nitin Indurkhya, Fred J. Damerau (eds.), CRC Press, Taylor and Francis Group, Boca Raton, FL, 577–604, 2010.
15. WILLIAM W. COHEN, SUNITA SARAWAGI, Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data

Integration Methods. *Proc. Conference on Knowledge Discovery in Data,* 89–98, 2004.

16. JIM COWIE, YORICK WILKS, Information Extraction. *Information Extraction Handbook of Natural Language Processing*, (ed.) Robert Dale, Harold Somers, Hermann Moisl, CRC Press, 241–260, 2000.
17. MICHAEL H. CRAWFORD, *Roman Republican Coinage*, 2 volumes, Cambridge UP, 1974.
18. TOM ELLIOTT, SEAN GILLIES, Digital Geography and Classics, *Digital Humanities Quarterly 3.1,* http://digitalhumanities.org/dhq/vol/3/1/000031/000031.html#elliott2008 2009.
19. JOAQUIM F. FERREIRA DA SILVA, ZORNITSA KOZAREVA, JOSÉ GABRIEL PEREIRA LOPES, Cluster Analysis and Classification of Named Entities. *Proc. Conference on Language Resources and Evaluation*, LREC 2004, Lisbon, Portugal, 321–324, 2004.
20. MARK DEPAUW, TOM GHELDOF, Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops: LCPD 2013, SUEDL 2013, DataCur 2013*, Valletta, Malta, September 22–26, 2013, Revised Selected Papers, 40–52, 2014.
21. MARK DEPAUW, BART VAN BEEK, People in Greek Documentary Papyri: First Results of a Research Project. *The Journal of Juristic Papyrology* 39, 31–47, 2009.
22. LEON DERCZYNSKI, JANNIK STRÖTGEN, RICARDO CAMPOS, Time and Information Retrieval: Introduction to the Special Issue. *Information Processing and Management*, 2015, http://derczynski.com/sheffield/papers/time-ir-ipm.pdf, u štampi.
23. LISA FERRO, LAURIE GERBER, INDERJEET MANI, BETH SUNDHEIM, GEORGE WILSON, *TIDES 2005 Standard for the Annotation of Temporal Expressions*, The MITRE Corporation, 2005.
24. NATHALIE FRIBURGER, DENIS MAUREL, Finite-state Transducer Cascades to Extract Named Entities in Texts. *Theoretical Computer Science*, Volume 313, Issue 1, 93–104, 2004.

25. SOFÍA N. GALICIA-HARO, ALEXANDER GELBUKH, Complex Named Entities in Spanish Texts: Structures and Properties. *Named Entities: Recognition, Classification and Use*, Amsterdam/Philadelphia, 71–96, 2009.
26. RALPH GRISHMAN, BETH SUNDHEIM, Message Understanding Conference MUC-6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING),* I, Kopenhagen, 466–471, 1996.
27. MAURICE GROSS, A Bootstrap Method for Constructing Local Grammars. Bokan, Neda (Ed.) *Proceedings of the Symposium "Contemporary Mathematics"*, Faculty of Mathematics, University of Belgrade, 229–250, 2000.
28. MENA B. HABIB, MAURICE VAN KEULEN, Named Entity Extraction and Disambiguation: The Reinforcement Effect, *In Proc. of MUD 2011*, 9–16, 2011.
29. MENA B. HABIB, MAURICE VAN KEULEN, Improving Toponym Disambiguation by Iteratively Enhancing Certainty of Extraction. *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2012*, 4–7 Oct 2012, Barcelona, Spain, 399–410, 2012.
30. TAKAAKI HASEGAWA, SATOSHI SEKINE, RALPH GRISHMAN, Discovering Relations among Named Entities from Large Corpora. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, 415–422, 2004.
31. JERRY R. HOBBS, ELLEN RILOFF, Information Extraction, *Handbook of Natural Language Processing*, Second Edition, Nitin Indurkhya, Fred J. Damerau (eds.), CRC Press, Taylor and Francis Group, Boca Raton, FL, 511–532, 2010.
32. MARIETTA HORSTER, The Prosopographia Imperii Romani (PIR) and New Trends and Projects in Roman Prospography. *Prosopography Approaches and Applications: A Handbook*, (ed.) K. S. B. Keats-Rohan, Oxford: Occasional Publications UPR, 231–240, 2007.
33. WILLIAM L. HOSCH (ED.), *The Britannica Guide to Numbers and Measurement*, New York, NY: Britannica Educational Publications, 2010.
34. JELENA JAĆIMOVIĆ, Recognition and Normalization of Temporal Expressions in Serbian Texts. *Proceedings of the 5th Balkan Conference in Informatics*, 97–100, 2012.

35. JELENA JAĆIMOVIĆ, Automatic Processing of Temporal Expressions in Serbian. *Proceedings of the Conference 35th Anniversary of Computational Linguistics in Serbia,* 2013, *http://jerteh.rs/wp-content/uploads/2015/05/Jacimovic.pdf*.
36. JELENA JAĆIMOVIĆ, CVETANA KRSTEV, DRAGO JELOVAC, A Rule-Based System for Automatic De-identification of Medical Narrative Texts, *Informatica*, Vol. 39, No. 1, The Slovene Society Informatika, Ljubljana, 45–53, 2015.
37. LUDOVIC JEAN-LOUIS, ROMARIC BESANÇON, OLIVIER FERRET 2010: Using Temporal Cues for Segmenting Texts into Events. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, August 16-18 2010, Reykjavik, Iceland, 150–161, 2010.
38. PANAGIOTA KARANASOU, LORI LAMEL, Comparing SMT Methods for Automatic Generation of Pronunciation Variants. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, August 16-18 2010, Reykjavik, Iceland, 167–178, 2010.
39. SVETLA KOEVA, CVETANA KRSTEV, IVAN OBRADOVIĆ, DUŠKO VITAS, Resources for Processing Bulgarian and Serbian — a Brief Overview of Completeness, Compatibility and Similarities, S. Piperidis, E. Paskaleva (eds.) *Workshop on Language and Speech Infrastructure for Information Access in the Balkanic Countries*, 25 September 2005, Borovets, Bulgaria, 31–38, 2005.
40. SVETLA KOEVA, CVETANA KRSTEV, DUŠKO VITAS, Morpho-semantic Relations in Wordnet – a Case Study for two Slavic Languages. *Proceedings of the 4th Global WordNet Conference 2008*, (eds.) Attila Tanacs *et al*, University of Szeged, Department of Informatics, 239–253, 2008.
41. CVETANA KRSTEV, *Processing of Serbian – Automata, Texts and Electronic dictionaries*, Faculty of Philology, University of Belgrade, Beograd, 2008.
42. KRSTEV 2014: CVETANA KRSTEV, Akronimi u automatskoj obradi tekstova na srpskom jeziku. *Naučni sastanak slavista u Vukove dane - Srpski jezik i njegovi resursi: teorija, opis i primene*, Vol. 43/3, Međunarodni slavistički centar, Beograd, 155-174, 2014.
43. CVETANA KRSTEV, DUŠKO VITAS, SANDRA GUCUL, Recognition of Personal Names in Serbian Texts. G. Angelova (ed.) *Proc. of the International Conference*

Recent Advances in Natural Language Processing, 21–23 September 2005, Borovets, Bulgaria, 288–292, 2005.

44. CVETANA KRSTEV, DUŠKO VITAS, AGATA SAVARY, Prerequisites for a Comprehensive Dictionary of Serbian Compounds. *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL*, (eds.) T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala, Turku, Finland, 23–25 August 2006, Volume 4139, 2006, LNCS (LNAI), Springer, Heidelberg, 552–563, 2006.
45. CVETANA KRSTEV, RANKA STANKOVIĆ, DUŠKO VITAS, IVAN OBRADOVIĆ, WS4LR: A Workstation for Lexical Resources. *Proceedings of LREC'06*, 22–28 May 2006, Genoa, Italy, ELRA, 1692–1697, 2006.
46. CVETANA KRSTEV, RANKA STANKOVIĆ, DUŠKO VITAS, A Description of Morphological Features of Serbian: a Revision Using Feature System Declaration. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, 17–23 May 2010, Valletta, Malta, ELRA, 816–819, 2010.
47. CVETANA KRSTEV, RANKA STANKOVIC, IVAN OBRADOVIC, DUŠKO VITAS, MILOŠ UTVIĆ, Automatic Construction of a Morphological Dictionary of Multi-Word Units. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, August 16-18 2010, Reykjavik, Iceland, 226–237, 2010.
48. CVETANA KRSTEV, DUŠKO VITAS, IVAN OBRADOVIĆ, MILOŠ UTVIĆ, E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, 12–15 July 2011, Blois, France, 48–56, 2011.
49. CVETANA KRSTEV, IVAN OBRADOVIĆ, RANKA STANKOVIĆ, DUŠKO VITAS, An Approach to Efficient Processing of Multi-word Units. *Computational Linguistics Studies in Computational Intelligence*, Volume 458, 109–129, 2013.
50. CVETANA KRSTEV, ANĐELKA ZEČEVIĆ, DUŠKO VITAS, TITA KYRIACOPOULOU, NERosetta – an Insight into Named Entity Tagging. *Proceedings of 6th Language & Technology Conference*, December 7–9, 2013, Poznań, Poland, (ed.) Zygmunt Vetulani, Hans Uszkoreit, 168–172, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, 2013.

51. CVETANA KRSTEV, IVAN OBRADOVIĆ, MILOŠ UTVIĆ, DUŠKO VITAS, A System for Named Entity Recognition Based on Local Grammars. *Journal of Logic and Computation* 24 (2), 473-489, 2014.
52. CVETANA KRSTEV, STAŠA VUJIČIĆ STANKOVIĆ, DUŠKO VITAS, Approximate Measures in the Culinary Domain: Ontology and Lexical Resources. *Proceedings of the 9th Language Technologies Conference IS-LT 2014*, Ljubljana, Slovenia, 9-10 October 2014, (eds.) Tomaž Erjavec, Jerneja Žganec Gros, Institut "Jožef Stefan", 38–43, 2014.
53. CVETANA KRSTEV DUŠKO VITAS, Corpus and Lexicon - Mutual Incompleteness. *Proceedings of the Corpus Linguistics Conference Series 1(1)*, Birmingham University, http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx, 2005.
54. CVETANA KRSTEV, DUŠKO VITAS, Finite State Transducers for Recognition and Generation of Compound Words. IS-LTC 2006, Tomaž Erjavec, Jerneja Zganec Gros (eds.), Ljubljana, Slovenia: Institut "Jožef Stefan", 192–197, 2006.
55. CVETANA KRSTEV, DUŠKO VITAS, The Treatment of Numerals in Text Processing. *Proceedings of Third Language & Technology Conference*, October 5-7, 2007, Poznań, Poland, ed. Zygmunt Vetulani, IMPRESJA Widawnictwa Elektroniczne S.A., Poznań, 418–422, 2007.
56. CVETANA KRSTEV, DUŠKO VITAS, An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds. *Proceedings of Lexis and Grammar Conference*, Bergen, 204–212, 2009.
57. EMELINE LECUIT, DENIS MAUREL, DUŠKO VITAS, CVETANA KRSTEV, Temporal Expressions: Comparisons in a Multilingual Corpus. *Proceedings of 4th Language & Technology Conference*, November 6–8, 2009, Poznań, Poland, (ed.) Zygmunt Vetulani, IMPRESJA Widawnictwa Elektroniczne S.A., Poznań, http://poincare.matf.bg.ac.rs/~cvetana/biblio/ltc-035-lecuit.pdf, 2009.
58. SIMON MAHONY, GABRIEL BODARD: Introduction. *Digital Research in the Arts and Humanities* (eds.) Gabriel Bodard, Simon Mahony, 1–11, 2010.

59. INDERJEET MANI, Recent Developments in Temporal Information Extraction, *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, Amsterdam/Philadephia, 45–60, 2004.
60. GIDEON S. MANN, DAVID YAROWSKY, Unsupervised Personal Name Disambiguation. *Proceedings of CoNLL*, 33–40, 2003.
61. RALPH W. MATHISEN, Where are all the PDBs?: The Creation of Prosopographical Databases for the Ancient and Medieval Worlds. *Prosopography Approaches and Applications: A Handbook*, (ed.) K. S. B. Keats-Rohan, Oxford: Occasional Publications UPR, 95–126, 2007.
62. DENIS MAUREL, Les mots inconnus sont-ils des noms propres?. *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles,* 776–784, 2004.
63. DENIS MAUREL, DUŠKO VITAS, CVETANA KRSTEV, SVETLA KOEVA, Prolex: A Lexical Model For Translation Of Proper Names Application To French, Serbian And Bulgarian. *Les langues slaves et le français : approches formelles dans les études contrastives Bulag*, 32, 55–72, 2007.
64. DENIS MAUREL, Prolexbase. A Multilingual Relational Lexical Database of Proper Names. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, 26 May–1 June 2008, Marrakech, Marocco, ELRA, 334–338, 2008.
65. DENIS MAUREL, BÉATRICE BOUCHOU-MARKHOFF, Prolmf: A Multilingual Dictionary of Proper Names and Their Relations. *LMF Lexical Markup Framework*, Gil Francopoulo (ed.), Wiley, 67–82, 2013.
66. DENIS MAUREL, NATHALIE FRIBURGER, IRIS ESHKOL-TARAVELLA, Enrichment of Renaissance Texts with Proper Names. *Infotheca* Vol. 15, No. 1, Belgrade, 29a–41a, 2014.
67. ELISABETE MARQUES RANCHHOD, Using Corpora to Increase Portuguese MWE Dictionaries: Tagging MWE in a Portuguese Corpus. *Proceedings of the Corpus Linguistics Conference Series 1(1)*, Birmingham University, UK, July 14–17 2005, http://infolingu.univ-mlv.fr/english/Bibliographie/Articles/RanchhodCorpora.pdf, 2005.

68. PAWEL MAZUR, ROBERT DALE, Handling Conjunctions in Named Entities. *Named Entities: Recognition, Classification and Use*, Amsterdam/Philadelphia, 51–70, 2009.
69. ANDREI MIKHEEV, A Knowledge-free Method for Capitalized Word Disambiguation. *Proc. Conference of Association for Computational Linguistics*, 159–166, 1999.
70. MARIE-FRANCINE MOENS, *Information Extraction, Algorithms and Prospects in a Retrieval Context*, Springer, 2006.
71. DIEGO MOLLÁ-ALIOD, JOSÉ-LUIS VICEDO, Question Answering. *Handbook of Natural Language Processing*, Second Edition, Nitin Indurkhya, Fred J. Damerau (eds.), CRC Press, Taylor and Francis Group, Boca Raton, FL, 485–510, 2010.
72. DAVID NADEAU, SATOSHI SEKINE, A Survey of Named Entity Recognition and Classification. *Named Entities: Recognition, Classification and Use*, Amsterdam/Philadelphia, 3–28, 2009.
73. MATTEO NEGRI, LUCA MARSEGLIA, Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Technical Report, http://www.cs.upc.edu/~nlp/meaning/documentation/3rdYear/WP3.6.pdf , 2004.
74. MATTEO NEGRI, ESTELA SAQUETE, PATRICIO MARTÍNEZ-BARCO, RAFAEL MUÑOZ, Evaluating Knowledge-based Approaches to the Multilingual Extension of a Temporal Expression Normalizer. *Proceedings of the Coling/ACL 2006 Workshop on Annotating and Reasoning about Time and Events*, Sydney, Australia, 30–37, 2006.
75. DAMIEN NOUVEL, *Reconnaissance des entités nommées par exploration de règles d'annotation: Interpréter les marqueurs d'annotation comme instructions de structuration locale*, Thèse de doctorat, Machine Learning [cs.LG], Université François-Rabelais, Tours, 2012.
76. DAMIEN NOUVEL, JEAN-YVES ANTOINE, NATHALIE FRIBURGER, DENIS MAUREL, An Analysis of the Performances of the CasEN Named Entities Recognition System in the Ester2 Evaluation Campaign. *Proceedings of the Seventh*

International Conference on Language Resources and Evaluation (LREC'10), 17–23 May 2010, Valletta, Malta, ELRA, 523–529, 2010.

77. DAMIEN NOUVEL, JEAN-YVES ANTOINE, NATHALIE FRIBURGER, ARNAUD SOULET, Recognizing Named Entities using Automatically Extracted Transduction Rules. *Proceedings of the 5th Language and Technology Conference*, Nov 2011, Poznan, Poland, 136–140, 2011.
78. NOUVEL *ET AL*. 2012: DAMIEN NOUVEL, JEAN-YVES ANTOINE, NATHALIE FRIBURGER, ARNAUD SOULET, Coupling Knowledge-Based and Data-Driven Systems for Named Entity Recognition. *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, ACL, 69–77, 2012.
79. DAMIEN NOUVEL, JEAN-YVES ANTOINE, NATHALIE FRIBURGER, Pattern Mining for Named Entity Recognition. *Lecture Notes in Computer Science (LNCS) series / Lecture Notes in Artificial Intelligence (LNAI) subseries*, 226–237, 2014.
80. PARUL PATEL, DR. S. V. PATEL, Approaches for Temporal Information Extraction: A Comparative Study. *International Journal of Engineering Research & Technology*, Vol. 3 - Issue 2, http://www.ijert.org/view-pdf/8071/approaches-for-temporal-information-extraction-a-comparative-study, 2014.
81. SÉBASTIEN PAUMIER, Unitex 3.1.beta User Manual. http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf, 2015.
82. THIERRY POIBEAU, LEILA KOSSEIM, Proper Name Extraction from Non-Journalistic Texts. *Computational Linguistics in the Netherlands*, 144–157, 2001.
83. STANLEY E. PORTER, MATTHEW BROOK O'DONNELL, Theoretical Issues for Corpus Linguistics Raised by the Study of Ancient Languages. *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster University (UK), 29 March–2 April 2001, 483, 2001.
84. ESTELA SAQUETE, PATRICIO MARTÍNEZ-BARCO, RAFAEL MUÑOZ, MATTEO NEGRI, MANUELA SPERANZA, RACHELE SPRUGNOLI, Multilingual Extension of a Temporal Expression Normalizer using Annotated Corpora. *Proceedings of the EACL Workshop on Cross-Language Knowledge Induction*, Trento, Italy,

http://rua.ua.es/dspace/bitstream/10045/22482/1/2006_Saquete_EACL.pdf , 2006.

85. AGATA SAVARY, A Formalism for the Computational Morphology of Multi-Word Units. *Archives of Control Sciences*, 15(LI), 437–449, 2005.
86. AGATA SAVARY, Computational Inflection of Multi-Word Units: A Contrastive Study of Lexical Approaches. *Linguistic Issues in Language Technology – LiLT* volume 1, issue 2, 1–53, 2008.
87. AGATA SAVARY, LESZEK MANICKI, MAŁGORZATA BARON, Populating a Multilingual Ontology of Proper Names from Open Sources. *Journal of Language Modelling*, Vol 1, No 2 (2013), 189–225, 2013.
88. NATHAN SCHNEIDER, SPENCER ONUFFER, NORA KAZOUR, EMILY DANCHIK, MICHAEL T. MORDOWANEC, HENRIETTA CONRAD, NOAH A. SMITH, Comprehensive Annotation of Multiword Expressions in a SocialWeb Corpus. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 26–31 May, Reykjavik, Iceland, ELRA, 455–461, 2014.
89. DAVID R. SEAR, *Roman Coins and Their Values Volume I: The Republic and the Twelve Caesars, 280 BC - AD 96*, London, 2000.
90. DAVID R. SEAR, *Roman Coins and Their Values Volume II From the accession of Nerva to the overthrow of the Severan Dynasty AD 96–AD 235*, London, 2002.
91. SATOSHI SEKINE, NYU: Description of the Japanese NE System used for MET-2. *Proceedings of the 7th Message Understanding Conference, https://aclweb.org/anthology/M/M98/M98-1019.pdf*, 1998.
92. SATOSHI SEKINE, Named Entity: History and Future. http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf, 2004.
93. SATOSHI SEKINE, On-demand information extraction. *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, 731–738, 2006.
94. SATOSHI SEKINE, KIYOSHI SUDO, CHIKASHI NOBATA, Extended Named Entity Hierarchy. *Proceedings of the Third International Conference on Language Resources and Evaluation,* LREC-02, 29–31 May 2002, Las Palmas, Canary Islands, Spain, 1818–1824, 2002.

95. SATOSHI SEKINE, CHIKASHI NOBATA, Definition, dictionaries and tagger for Extended Named Entity Hierarchy. *Proceedings of Conference on Language Resources and Evaluation*, LREC 2004, Lisbon, Portugal, 1977–1980, 2004.

96. YUSUKE SHINYAMA, SATOSHI SEKINE, Named entity Discovery Using Comparable News Articles. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, ACL, 848–853, 2004.

97. DAVID A. SMITH, GREGORY CRANE, Disambiguating Geographic Names in a Historical Digital Library?. *Research and Advanced Technology for Digital Libraries*, volume 2163 of LNCS, 127–136, 2001.

98. DAVID A. SMITH, GIDEON S. MANN, Bootstrapping Toponym Classifiers. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 45–49, 2003.

99. MALGORZATA SPEDIZIA, DENIS MAUREL, AGATA SAVARY, *Multilingual Relational Database of Proper Names: Prolexbase Documentation*, Ecole Polytechnique de l'Université de Tours Département Informatique, Tours, 2011.

100. RALF STEINBERG, BRUNO POULIQUEN, Cross-lingual Named Entity Recognition. *Named Entities: Recognition, Classification and Use*, Amsterdam/Philadelphia, 137–164, 2009.

101. JANNIK STRÖTGEN, MICHAEL GERTZ, HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. *Proceedings of the Workshop on Semantic Evaluation*, ACL, 321–324, 2010.

102. SERHAN TATAR, ILYAS CICEKLI, Automatic Rule Learning Exploiting Morphological Features for Named Entity Recognition in Turkish. *Journal of Information Science*, vol. 37, no 2, 137–151, 2011.

103. MELISSA TERRAS, The Digital Classicist: Disciplinary Focus and Interdisciplinary Vision. *Digital Research in the Arts and Humanities* (eds.) Gabriel Bodard, Simon Mahony, 171–189, 2010.

104. MICKAËL TRAN, DENIS MAUREL, DUŠKO VITAS, CVETANA KRSTEV, A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names. *Proceedings of the 6th Workshop on Multilingual Lexical Databases (PAPILLON'05)*, Chiang Rai, Thailand, 67–71, 2005.

105. HERBERT VERRETH, A Survey of Toponyms in Egypt in the Graeco-Roman Period, Version 2.0, (ed.) Willy Clarysse, Mark Depauw, Heinz Josef Thissen, *KU Leuven Trismegistos Online Publications*, http://www.trismegistos.org/top.php, 2013.

106. DUŠKO VITAS, CVETANA KRSTEV, GORDANA PAVLOVIĆ-LAŽETIĆ, GORAN NENADIĆ, Recent Results in Serbian Computational Lexicography. *Proceedings of the Symposium "Contemporary Mathematics*, Belgrade, (ed.) Neda Bokan, Faculty of Mathematics, Belgrade, 113–130, 2000.

107. DUŠKO VITAS, CVETANA KRSTEV, IVAN OBRADOVIĆ LJUBOMIR POPOVIĆ, GORDANA PAVLOVIĆ-LAŽETIĆ, Processing Serbian Written Texts: An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts. *Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece, 97–104, 2003.

108. DUŠKO VITAS, CVETANA KRSTEV, DENIS MAUREL, A Note on the Semantic and MorphologicalPproperties of Proper Names in Prolex Project. *Named Entities: Recognition, Classification and Use*, Amsterdam/Philadelphia, 117–136, 2009.

109. DUŠKO VITAS, CVETANA KRSTEV, Derivational Morphology in an E-Dictionary of Serbia. *Proceedings of 2nd Language & Technology Conference*, Poznań, Poland, (ed.) Zygmunt Vetulani, 139–143, 2005.

110. DUŠKO VITAS, CVETANA KRSTEV, Structural Derivation and Meaning Extraction: A Comparative Study on French-Serbo-Coratian Parallel Texts. *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, (eds.) Geoff Barnbrook, Pernilla Danielsson, Michaela Mahlberg, Birmingham: The University of Birmingham Press, 166–178, 2005.

111. DUŠKO VITAS, CVETANA KRSTEV, Regular Derivation and Synonymy in an E-Dictionary of Serbian, *Archives of Control Sciences* Volume 15(LI), 2005 No.3, 251–263, 2006.

112. DUŠKO VITAS, CVETANA KRSTEV, Processing of Corpora of Serbian Using Electronic Dictionaries. Prace Filologiczne, vol. LXIII, 279–292, Warszawa, 2012.

113. STAŠA VUJIČIĆ STANKOVIĆ, CVETANA KRSTEV, DUŠKO VITAS, Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain. *The Proceedings of Seventh Global WordNet Conference 2014*, (eds.) Heili Orav, Christiane Fellbaume, Piek Vossan, University of Tartu, Tartu, Estonia, January 25–29, 127–132, 2014.

114. ROSA STERN, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Thèse de doctorat, Computation and Language [cs.CL]. Universit´e Paris-Diderot – Paris VII, 2013.

ВЕЋУ ОДЕЉЕЊА ЗА КЛАСИЧНЕ НАУКЕ, КОМИСИЈИ ЗА ДОКТОРСКЕ СТУДИЈЕ
И
НАСТАВНО-НАУЧНОМ ВЕЋУ ФИЛОЗОФСКОГ ФАКУЛТЕТА У БЕОГРАДУ

Изабрани у комисију за разматрање предлога теме за докторску дисертацију који је Одељењу за класичне науке презентовала **Марина Голуб**, подносимо

РЕФЕРАТ

о квалификованости кандидата и подобности
теме предложене за докторску дисертацију

О кандидату

Марина Голуб отпочела је основне студије Класичних наука на Филозофском факултету у
Београду 2008. године; 2012. је одбранила дипломски рад на тему *Латинска и српска будућа времена: поређење на примерима из Теренцијевог Формиона*, с оценом 10; њена просечна оцена на основним студијама била је 9,33; школске 2011/12. награђена је као студент генерације на Одељењу за класичне науке.
По основним студијама уписала се и на мастер студије на нашем факултету, те је 2013. одбранила мастер рад на тему *Утицај латинске синтаксе на грчки превод Res gestae divi*
Augusti, с оценом 10; њена просечна оцена на мастер студијама била је 10.
Исте године уписала се код нас и на докторске студије; тренутно је на другој години студија,
с досадашњом просечном оценом 10. Поднела је предлог теме за докторску дисертацију
под насловом *Именовани ентитети у латинском језику.*
Од 2009. до 2013. примала је стипендију Министарства просвете, науке и технолошког
развоја РС за студенте основних и мастер академских студија. Од маја 2014. стипендист је
истог министарства и за докторске студије. У том својству сарађује на пројекту Језици и
културе у времену и простору, финансираном од Министарства просвете, науке и технолошког развоја РС (ОИ 178002).
Од октобра 2014. на Одељењу за класичне науке помаже у припреми и извођењу наставе на
курсевима Латински језик 1 и Латински језик 2 за студенте историје.
На научним скуповима досад је учествовала са следећим саопштењима: »Антериорна употреба будућих времена у латинском и њихови еквиваленти у српском језику«, на Четвртој
међународној конференцији Језици и културе у времену и простору, Нови Сад, 22. новембра 2014. (тај рад је примљен за објављивање у предстојећем зборнику Језици и

културе у времену и простору 4); »Position of Genitives in Greek an Latin DPs«, на Првој
регионалној конференцији студената класичних наука GLAS (Graecae Latinaeque antiquitatis
studentes), Београд, 11–12. јула 2015; »Possessive Pronominal Arguments in Greek and Latin
DPs«, Twelfth International Conference on Greek Linguistics, Берлин, 16–19. септембра 2015.

Предмет и циљ дисертације

Предмет предложеног истраживања били би именовани ентитети у латинском језику.
Термин именовани ентитети (Named Entities – NE) користи се у тзв. обради природних језика
(Natural Language Processing – NLP), а односи се на један од битних концепата у области тзв.
извлачења информација (Information Extraction – IE), у склопу појмовно-методске апаратуре
развијене на мултидисциплинарним конференцијама које су у последњој деценији прошлог
века постале познате под акронимом MUC (Message Understanding Conference). Ове конференције примарно су се тицале једног практичног задатка: како из текстова аутоматизовати екстракцију информација везаних за активности различитих компанија и
организација и за питања безбедности и одбране.
С тим у вези брзо се очитовала важност препознавања информација које, базиране на властитим именима и нумеричким подацима, унутар скупова елемената са сличним атрибутима јасно идентификују поједине елементе. На тај начин, препознавање и извлачење
тих ентитета постали су један од важних задатака при извлачењу информација, па се у
протеклих двадесет година много радило на овом проблему у модерним језицима.
Препознавање и класификација именованих ентитета (Named Entity Recognition and Classification – NERC) највише су развијени за енглески, али много се радило и на немачком,
шпанском, јапанском, кинеском, грчком и италијанском језику, а значајна пажња овом проблему посвећена је и у српском, бугарском, румунском, корејском, турском, шведском,
португалском. У латинском, напротив, развијање система за аутоматско препознавање
именованих ентитета сасвим је занемарено.
За корпусне језике попут латинског и грчког, ситуација је врло специфична. Још крајем 19.
века, на Момзенову иницијативу, покренута је велика Prosopographia Imperii Romani (PIR). Taj
први и најдужи просопографски подухват у Европи сабира личности у Римском царству од
краја I в. пре Хр. до краја III в. по Хр. С једне стране, то је убедљив пример важности ентитета
као што су лична имена, са свим подацима која она носе или имплицирају, за изучавање

свих аспеката класичне историје и културе; с друге стране, дуготрајност пројекта PIR најбоље показује колико је времена и труда потребно за мануално прикупљање оваквих података.

Слично се може рећи и о топографским подацима: они се давно прикупљају, па и њихова дигитализација почела је рано, али се тај процес, по свему судећи, одвија мимо развоја аутоматских система за препознавање именованих ентитета.

С обзиром на то, кандидаткињина је намера да својом докторском дисертацијом допринесе развоју метода за препознавање именованих ентитета у латинском језику, са ширим циљем да демонстрира нешто од предности које дигиталија, наспрам традиционалних метода истраживања, нуде класичној филологији данас.

Структура дисертације

Кандидаткиња је предложила следећу структуру:

• Резиме (српски/енглески).

• Садржај.

• Списак скраћеница.

• Увод. – Овде би се поставио оквир теме, описала и оправдала метода, и приказала досадашња достигнућа у области која је предмет истраживања, уз критички осврт на постојећу литературу.

• Средишња поглавља. – Овај главни део рада састојао би се од четири поглавља, од којих би прва три била посвећена различитим типовима именованих ентитета у латинском: просопографским подацима и антропонимима; топографским подацима и топонимима; нумеричким подацима и језичким средствима њиховог исказивања. Четврто поглавље било би посвећено именованим ентитетима у епиграфским споменицима као проблему специјалне природе.

• Закључак – Овде би се сабрали резултати истраживања, размотриле могућности њиховог тумачења у ширем контексту и путеви којима би у будућности требало поћи.

• Библиографија.

Полазне хипотезе

Од почетака до данас, класификација именованих ентитета у језицима развила се од почетних 7 типова до двестотинак о којима се данас говори. У том послу, међутим, латински језик није досад узиман у обзир, и кандидаткиња претпоставља да ће, с тим у виду, категорије именованих ентитета морати у извесном степену да се модификују. Већ на први поглед јасно је да за неким типовима именованих ентитета (попут телефонских бројева) у случају латинског неће бити потребе. Деликатнија питања постављају се у вези с подацима који су у модерно доба нумерички а у римска времена то нису били: на пример, године се код Римљана нису бројале већ означавале именима конзула; на извесне дане у месецу такође се референсало именима а не нумерички. При раду с антропонимима главна тешкоћа

ће се вероватно лоцирати у области апелатива који (карактеристично пре свега за владаре
и сл.) у исти мах представљају титулу и лично име; ово не само због класификације већ и
због идентификације различитих лица на која се може односити један те исти апелатив.
Етноними и топоними такође ће представљати искушење, пре свега зато што модерна,
наизглед саморазумљива ситација у којој нације, државе и територије носе свака своје устаљено име заправо не постоји у старим цивилизацијама.
Утилитарно гледано, рад би почивао на претпоставци да развијање методе за препознавање
именованих ентитата у латинском језику, које данас још нема, може значајно да унапреди
филолошка и сродна истраживања кад су у питању брзина, прецизност и прегледност добијених резултата, и побуди живљи интерес и за друга питања која се посредно или непосредно тичу електронске обраде језика.

Методе које ће се применити

Рад би имао како теоријски тако и практични аспект. Типови именованих ентитета у латинском поредили би се с аналогним типовима у језицима за које је проблем већ обрађиван, да би се кроз сличности и разлике антиципирали проблеми у препознавању који
би могли настати на латинској страни. Затим би се на неструктурираним текстовима
одговарајућег типа показало на чему се заснива и како тече препознавање именованих ентитета, какав му се значај може приписати и где му се може наћи примена.
За практични део рада био би неопходан систем за препознавање именованих ентитета.
Кандидаткиња планира да се послужи Унитексом, који се већ добро показао у препознавању
именованих ентитета у неким модерним језицима. (Унитекс је софтвер који се користи за
обраду природних језика и може се дефинисати као скуп програма развијених за анализу
природних језика уз помоћ језичких ресурса, нпр. електронских речника и граматика.)
Електронски речници су неопходан лексички извор у различитим фазама аутоматске обраде
текста; граматике пак које се у Унитексу креирају, користе, и врло лако модификују, графички
су прикази језичких појава – попут овог недовршеног графа којим се утврђује крај реченице у
латинском:

Очекивани резултати и научни допринос

Допринос овог рада би, као и метода, требало да буде двојак, теоријски и практични. Кандидаткиња би преиспитала досадашњу класификацију именованих ентитета засновану
на модерним језицима и модификовала је тамо где се то покаже неопходним за обраду латинских текстова. Самим тим, највећи практични допринос тицао би се оних типова

именованих ентитета који се у највећој мери разликују од типова познатих из модерних језика. Осим успешног препознавања датума по римском систему, кандидаткиња предвиђа аутоматизацију њихових еквивалената по савременом систему датирања (уз Idibus Martiis, 15. март; уз C. Pansa A. Hirtio consulibus, 43. пре Хр., и томе слично).

Закључак

Досадашњи успеси Марине Голуб на академским студијама, високе способности које је она код себе развила, и склоност истраживачком раду коју је досад исказала, омогућују нам да с поуздањем подржимо њене планове везане за докторску дисертацију. Научни посао који кандидаткиња предлаже одиста је нов у области латинистике, до те мере да се његове импликације, практичне, али и теоретске, не могу сасвим ни разабрати а приори. Кандидаткиња, међутим, прилази својој теми не само с тачно дефинисаним циљевима и јасном представом како их достићи, већ и са импресивним увидом у досадашње послове те врсте у рачунарској лингвистици и, на другој страни, с изврсном спремом у домену класичне филологије и нарочито латинске лингвистике. Ми стога процењујемо да ће она моћи и умети да оствари задатак који је себи поставила.

Са особитим задовољством предлажемо да се Марини Голуб одобри израда докторске дисертације на тему и под насловом који је предложила: Именовани ентитети у латинском језику.

у Београду, 22. октобра 2015.

_____

проф. др Цветана Крстев

_____

доц. др Борис Пендељ

_____

доц. др Драгана Димитријевић

_____

*проф. др Војин Недељковић, ментор*__