# Интеграција српског језика у вишејезични речник Prolexbase

УДК 811.163.41’322.2

Цветана Крстев
cvetana@matf.bg.ac.rs
*Универзитет у Београду*
*Филолошки факултет*
*Србија*

Дени Морел
denis.maurel@univ-tours.fr
*Универзитет у Туру*
*Француска*

Душко Витас
vitas@matf.bg.ac.rs
*Универзитет у Београду*
*Математички факултет*
*Србија*

**САЖЕТАК:** Овај рад представља вишејезични речник властитих имена *Prolexbase*, посебно његов српски део. Представљена је комплексност властитих имена у српском језику, посебно одлике које се тичу њиховог превођења: правопис, деривација, флективне промене и дијалекатске варијације. Описује се модел базе *Prolex*, са посебним освртом на решења неопходна за интеграцију у њу српског језика (коришћење два писма, више нивоа деривације, постојање вишеструких облика). На крају се дају неки бројчани подаци који показују српски језик у бази *Prolex*.

**КЉУЧНЕ РЕЧИ:** властита имена, вишејезична база података, онтологија властитих имена, LMF формат, српски језик, Prolexbase.

**РАД ПРИМЉЕН:** 4. децембар 2018.
**РАД ПРИХВАЋЕН:** 14. децембар 2018.

## 1. Мотивација

Као и друге особености језика (неологизми, полилексемске јединице, идиоми и тако даље), властита имена могу бити узрок невероватних грешака. На пример, да ли би *Bush* требало превести на српски са *грм* (биљка) или *Буш* (лично име)? Да ли се *Casablanca* и *White House* односе на исто место? Влада опште уверење да се властита имена не могу преводити. У суштини, преводиоци користе све врсте преводилачких процеса – адаптацију, дослован превод итд. – када преводе текст са изворног на циљни језик (Lecuit et al., 2011).

Властита имена представљају изазов и за обраду приодних језика, а посебно за задатке везане за *именоване ентитете*[1]. Први задаци везани за именоване ентитете постављени у оквиру конференција о разумевању порука (Message Understanding Conferences MUC-6 и MUC-7) односили су се на попуњавање база података одговорима на питања као што су „кога је терориста напао?“, „где?“, „када?“ или „која фирма је преузела власништво друге фирме?“, „колики је удео?“, „по којој цени?“ и тако даље (Chinchor, 1997). Данас су изазови скоро супротни: ентитете из текста треба повезати са уносима у бази података (Hachey et al., 2013) јер, на пример, властита имена треба да постану једнозначна (видети за пример конференције о анализи текста (Text Analysis Conferences) (McNamee et al., 2010). За овакве задатке се често користи Википедија као и многе друге семантичке базе података: DBpedia (Auer and Lehmann, 2007), GeoNames, YAGO2 (Hoffart et al., 2012), BabelNet (Navigli and Ponzetto, 2012). Ове базе података представљају део система повезаних отворених података (Link Open Data system (LOD)) у коме властита имена заузимају посебно важно место.

Prolexbase је вишејезична релациона база података властитих имена (Maurel, 2008). Циљ Prolex базе је да помогне приликом превођења. Она укључује морфолошке, деривационе и семантичке релације. На пример, ако би требало превести реченицу *Београђанка ми је рекла да је Дунав прелеп* могло би бити корисно да се она прошири на следећи начин: *Женски [флексија] становник града [семантичко проширење] Београда [деривациона релација] у Србији [релација доступности] ми је рекла да је река Дунав [семантичко проширење] прелеп*. Вратићемо се на овај пример на крају овог рада.

Прву верзију Prolex базе која је укључивала осам језика (француски, немачки, енглески, италијански, холандски, пољски, португалски и шпански) је подржао француски пројекат *RNTL-Technolangue Project* (2003-2005). У ствари, у оквиру овог пројекта је конструисан модел базе података, остварена је висока покривеност за француски, док су други језици били слабије заступљени. У исто време је покренут пројекат *Egide Pavle Savic* (2004-2005) чији циљ је био укључивање српског језика у Prolex базу. Укључивање српског тима је било веома важно јер је оно омогућило да се боље разуме сложеност морфологије и деривације у моделу, који је до тада био сувише под утицајем француског

[1] Именовани ентитети се обично дефинишу референцијално или својом јединственошћу.

и енглеског језика. Други проблем је представљало коришћење два писма, ћириличног и латиничног. У овој првој верзији изабрано је незадовољавајуће решење за коришћење два писма: за српски језик су изграђена два блока, један који користи ћирилицу и други који користи латиницу.

Друга верзија Prolex базе коју је подржао пројекат *Hubert Curien Polonium* је донела добру покривеност за енглески и пољски (Savary et al., 2013). Српски део базе је био значајно унапређен у трећој верзији Prolex базе као резултат једномесечне посете проф. Цветане Крстев Универзитету у Туру. Током ове посете побољшана је покривеност српског језика, а раздвојена репрезентација за два писма је спојена у један блок. Такође је припремљен могући опис дијалекатских облика, екавског и ијекавског.

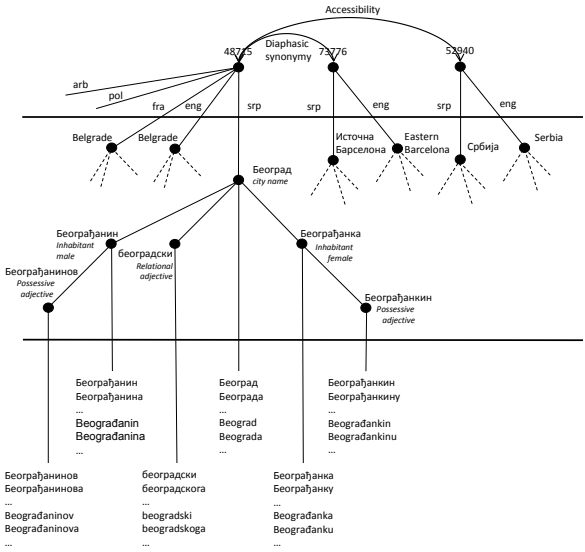
## 2. База Prolex

### 2.1 Модел Prolex базе

Пошто је база Prolex вишејезична база података потребан нам је модел који омогућава повезивање различитих појављивање властитих имена у разним језицима. Одлучили смо да дефинишемо лингвистичку класу властитих имена (и облика изведених деривацијом) као онтологију у смислу датом у (Gruber, 1995): „Концептуализација је апстрактан, поједностављен поглед на свет који желимо да представимо с неким циљем... Онтологија је експлицитна спецификација концептуализације“.

Средиште онтологије базе Prolex је *концептуално властито име*, *стожер*, који представља референта с одређене тачке гледишта. Примери су *Папа Фрањо* и *Хорхе Марио Бергољо* или *Београд* и *Источна Барселона*[2]. Превођење преко стожера (Boitet, 1988) данас није уобичајено иако стожер може да буде профињен у неким језицима више него у другима. На пример, у пројекту Papillon (Mangeot, 2000), стожер за *rice* (*пиринач*) на енглеском има два финија стожера на јапанском, *raw rice* (*сирови пиринач*) и *cooked rice* (*кувани пиринач*). Како за *концептуално властити име* профињавање није потребно овај модел се може користити без проблема. У сваком језику стожер је повезан са јединственим скупом властитих имена, *пролексемом*. Овај скуп садржи властитито име, и по потреби његове псеудониме

[2] На интернету се *Београд* некад реферише као *Источна Барселона*




**Слика 1.** Пример: *Београд* у моделу базе Prolex

(алијасе) као и морфосинтаксички изведене облике (видети 2.3). Стожери представљају концептуални ниво модела док пролексеме представљају његов линвистички ниво. Онтологију допуњавају још два нивоа: метаконцептуални ниво (типови и супертипови) испод кога је ниво примерака (облици власититх имена онако како се јављају у тексту). На слици 1 је модел илустрован на примеру властитог имена *Београд*.

## 2.2 Формат LMF

База Prolex је отворени ресурс који се користи под лиценцом LGPL-LR[3]. Формат за размену је инспирисан форматом за лексичко обележавање (Lexical Markup Format (LMF)) (ISO/TC 37/SC 4, 2007). На слици 2 су приказане LMF класе за репрезентацију модела базе Prolex. Овај формат представља избор класа из језгра LMF модела са деловима додатим из LMF проширења (пакети за морфологију, семантику,

[3] http://www.cnrtl.fr/lexiques/prolex/

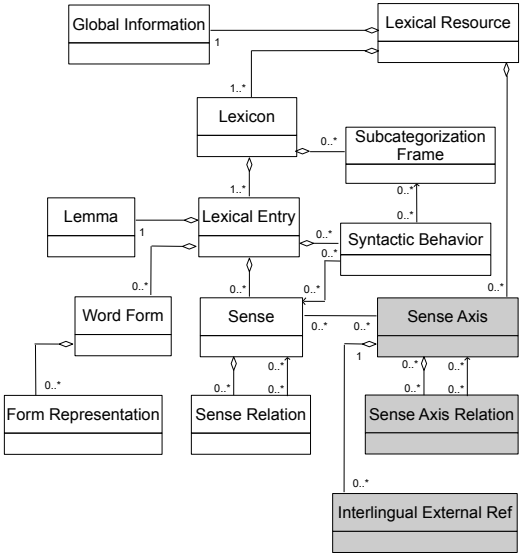
вишејезичну нотацију и синтаксу за потребе обраде природних језика). Вишејезични описи су представљени сивим кућицама. Цео ресурс је представљен класом *Lexical Resource* (*лексички ресурс*) с којом су повезане неке информације – код језика, писма, карактери које ресурс користи (класа *Global Information* (*опште информације*)). Ресурс садржи концептуални ниво (класа *Sense Axis* (*оса значења*)) и лингвистички ниво са више лексикона (класа *Lexicon* (*лексикон*)) који су једнојезични описи. Један од њих је српски лексикон. Лексички уноси су све леме пролексеме (властита имена, алијаси и деривациони облици) са одговарајућим облицима речи (сви примерци):класе *Lexical Entry* (*лексички унос*), *Lemma* (*лема*), *Word Form* (*облик речи*) и *Form Representation* (*репрезентација облика*). Ове леме су повезане са значењима која су стожери којима је додељена категорија, као *описни придев* или *присвојни придев од имена мушког рода* (класе *Sense* (*значење*) и *Sense Relation релација значења*)). Ови стожери су дефинисани у класи *Senses Axis* која спада у вишејезични део ресурса. Класа *Sense Axis Relation* (*Релација осе значења*) представља релацију између концептуалних властитих имена док класа *Interlingual External Ref* (*међујезичка спољашња референца*) представља типологије. Уочавају се и неке информације о класификационом контексту властитих имена (класа *Subcategorization Frame (оквир за субкатегоризацију)* и неке изузетне колокације (класа *Syntactic Behavior* (*синтаксично понашање*)). Ове класе се још увек не користе у српском блоку, а неки примери би били коришћење различитих предлога уз властита имена, на пример, *Србија је **на** Балкану и **у** Европи.*

### 2.3 Релације

Велики део базе Prolex чине релације између стожера (језички независне релације): синонимија, меронимија и доступност.

Релација синонимије, или прецизније, квази-синонимије или блискозначности, је релација између два стожера који се реферишу на исти референт за који постоје различите тачке гледања. Преводилац треба да изабере исправну тачку гледита, што није увек могуће. Разликујемо три различите тачке гледишта које одређују три дијасистематска својства (Coseriu, 1998):

– Дијахроне варијације зависе од времена: *Савезна Република Југославија* према *Државна Заједница Србија и Црна Гора*;




**Слика 2.** LMF шема базе Prolex

– Дијастратске варијације зависе од социокултурне стратификације: *Јосип Броз* према *Тито*;
– Дијафазне варијације зависе од употребе: *Београд* према *Источна Барселона*,

Релација меронимије, или партитивна релација, је релација укључивања. Пример укључивања за географска имена је: *Србија* је на *Балкану* који је део *Европе*, а пример временске релације је: бомбардовање Београда 6. априла 1941. године је део (догодио се током) Другог светског рата. Ову врсту релације проширујемо на друге домене, као што су економија, националности и тако даље.

Релација доступности (Ariel, 1990), или релација асоцијативности, значи да је властито име доступно преко неког другог властитог имена. У речницима, властита имена, за разлику од заједничких именица, немају дефиниције – оне су замењене релацијом са неким познатијим именом. С тога су ове релације ретко симетричне: у речницима се тако може прочитати да је *Арон* брат *Мојсија*, али се *Мојсије* не представља као брат *Арона* већ као предводник Јевреја. Према томе, *Арон* је доступан преко Мојсија, док је *Мојсије* доступан преко одговарајуће јеврејске приче. Разликујемо 12 оваквих релација:

– рођачке везе: *Арон* је брат *Мојсија*;
– главни град: *Београд* је главни град *Србије*;
– вођа: *Тито* је био политички вођа *Југославија*;
– оснивач: *Растко Немањић* је основао *Српску православну цркву*;
– следбеник: *Петар* је ученик *Исуса*;
– креатор: *Бранислав Нушић* је написао комедију *Госпођа министарка*;
– менаџер: *Ранко Жеравица* је био српски кошаркашки тренер који је водио *Југословенску кошаркашку репрезентацију*;
– становник: *Александар Вучић* борави у *Новом двору*;
– наследник: *Кнез Михаило Обреновић* је наследио *Кнеза Милоша Обреновића*;
– седиште: Седиште корпорације *Montinvest Beograd* се налази у *Београду*;
– супарник: *Партизан* је фудбалски ривал *Црвене звезде*;
– друг: *Мирко* је *Славков* најбољи друг и саборац.[4]

Језички зависне релације су фреквенција, је-алијас-од, је-дериват-од, колокација, контекст и епонимија.

Пролексема је скуп свих лема које су семантички повезане с властитим именом у једном језику. На пример, пролексему *Београда* чини: Београд, београдски, Београђанин, Београђанка, Београђанинов, Београђанкин као што се види на слици 1. Три главне релације на језички зависном нивоу су фреквенција, је-алијас-од и је-дериват-од. Фреквенција која указује колико је име познато може да има три вредности: често се користи, не користи се често и ретко се користи. Данас се ова фреквенција може израчунати из повезаних отворених

[4] Главни ликови из веома популарног југословенског стрипа о два партизанска дечака-курира.

података (LOD), углавном из Википедије (Elashter and Maurel, 2016). Алијаси су различите варијације властитог имена: краћи облици, скраћенице, акроними, разлике у писању, алтернативна транскрипција, дијатопски квази-синоними и објашњења. Што се тиче деривације, узимају се убзир само леме изведене морфосемантичким средствима (и леме изведене из њих). На пример, глагол *пастеризовати* који се односи на процес делимичне стерилизације је изведен из имена *Пастер* (Pasteur) али није с њим семантички повезан.

Релације колокације и контекста тичу се локалног коришћења властитих имена. У неким језицима, на пример у француском, испред имена државе често долази члан мушког или женског рода, при чему не постоји посебан разлог за избор рода: тако се каже, на пример, *la* (женски род) *France* и *le* (мушки род) *Montenegro*. Контекст је релација између властитог имена и типичних речи које га окружују. Контекст може да се односи на класификацију или доступност. MacDonald (1990) назива *екстерна структура* властитог имена класификациони контекст који је проширење именичке фразе (главни град, краљ, тренер, итд.). Класификациони контекст може да буде користан приликом превођења. Не пример, *Сава* се преводи на енглески са *Sava River*. Контекст доступности је именичка фраза која уводи релацију доступности између два стожера. Она се може разумети као врста објашњења за властито име његовим повезивањем са неким добро познатим властитим именом. На пример, *Београд* се може превести са *Београд, главни град Србије*.

Релација епонимије се разликује од осталих релација: она говори да се превод не односи на властито име већ на заједничку именицу (антономазија), као у примеру *жилет* који у српском означава све врсте оштрица за бријање, а не само оне које производи фирма Gillette, или на термине, какви су *Паркинсонова болест* или *Питагорина теорема*, или, пак, на идиоматску фразу каква је *све ми је равно до Косова* са значењем *сасвим ми је свеједно*.

### 2.4 Типологија

Метаконцептуални ниво се бави постојањем концепта и типологијом властитих имена.

Постојање концепта дели властита имена у три групе: (а) историјска која постоје или су постојала, као *Београд*; (б) религијска, чије постојање зависи од веровања, на пример *Архангел Михаил*, или (в) измишљена од стране аутора. У општем случају имена која спадају у последње две

наведене групе често захтевају да буду преведена, као што се *Snow White* са енглеског преводи на српски са *Снежана*.

Циљ типологије базе Prolex је класификација властитих имена. Дефинисали смо четири велике класе које одговарају примарном семантичком својству и назвали их *супертиповима*): особе (*антропоними*), локације (*топоними*), конкретне ствари (*ергоними* – артифакти и имена дела) и догађаји (*прагмоними*). Дефинисали смо укупно тридесет типова који су представљени у табели 1. Ову типологију дефинише примарна релација хиперонимије успостављена између стожера и типа. Допунили смо је другом релацијом, секундарно релацијом хиперонимије, која је релација метонимије између типова, као што се види у табели 2.

| властито име | | | | | | |
|---|---|---|---|---|---|---|
| антропоним | | | ергоним | прагмоним | топоним | |
| индивидуално | колективно | | | | | |
| | | група | | | | територија |
| позната лич. | династија | друштво | објекат | несрећа | астроним | држава |
| патроним | етноним | ансамбл | дело | празник | зграда | регион |
| лично име | | фирма | мисао | историја | геоним | наднационално |
| псеудо- | | институција | производ | манифестација | хидроним | |
| антропоним | | организација | пловило | метеорологија | град | |
| | | | | | пут | |

**Табела 1.** Типологија базе Prolex – примарна релација хиперонимије

## 3. Властита имена у српском језику

### 3.1 Писма

У Србији је употреба ћириличног писма прописана законом (Zakon, 2010, члан 1), док је коришћење латиничног писма дозвољено у посебним случајевима (саобраћајни знаци, називи улица, итд). Међутим, латинично писмо је из историјских и других разлога у широкој употреби и у Правопису српског језика (Пешикан et al., 1993, чланови 1–6) је дефинисано као равноправно са ћириличним писмом. Српско писмо, било ћирилично или латинично, има 30 слова, а између ова два скупа

| **типови** | **секундарни хипероними** |
|---|---|
| држава<br>регион<br>наднационалан<br>територија | колективни антропоним |
| град | колективни антропоним<br>ергоним |
| зграда<br>пут | ергоним |
| празник<br>историја<br>манифестација | |
| друштво<br>ансамбл<br>фирма<br>група<br>институција<br>организација | ергоним<br>топоним |
| пловило | топоним |

**Табела 2.** Секундарна релација хиперонимије

постоји 1-1 пресликавање као што је приказано у Табели 3. Редослед слова у ћирилици и латиници је различит; слова у Табели 3 су приказана у ћириличном редолседу. У латиничном писму за српски се не користе нека од 26 слова енглеског алфабета[5] и то Q, W, X и Y, али се нека слова користе са дијакритичким знацима: Č, Ć, Đ, Š и Ž„ а нека су представљена као диграфи, односно, комбинације других слова: Lj, Nj и Dž. Диграфи се у електронским тектовима обично приказују помоћу кодова слова која га чине премда су у Unicode били уведени посебни кодови за ове симболе[6]. Треба имати на уму да велика слова диграфа lj, nj и dž могу бити представљена на два начина: у првом је само прво слово диграфа велико слово – Lj, Nj и Dž – а у другом су оба слова диграфа велика слова – LJ, NJ и DŽ. Овај други начин се користи када

[5] Слова описана ASCII кодом.
[6] Видети кодну страну Unicode Latin Extended-B

је цела реч (или дужи текст) записан само великим словима. Ово је такође изражено у Unicode-у тако што за ове две репрезентације постоје различити кодови.

| | | | | | | |
|---|---|---|---|---|---|---|
| ћирилица | а А | ђ Ђ | ј Ј | н Н | с С | х Х |
| латиница | a A | đ Đ | j J | n N | s S | h H |
| ћирилица | б Б | е Е | к К | њ Њ | т Т | ц Ц |
| латиница | b B | e E | k K | nj Nj | t T | c C |
| ћирилица | в В | ж Ж | л Л | о О | ћ Ћ | ч Ч |
| латиница | v V | ž Ž | l L | o O | ć Ć | č Č |
| ћирилица | г Г | з З | љ Љ | п П | у У | џ Џ |
| латиница | g G | z Z | lj Lj | p P | u U | dž Dž |
| ћирилица | д Д | и И | м М | р Р | ф Ф | ш Ш |
| латиница | d D | i I | m M | r R | f F | š Š |

**Табела 3.** Српско ћирилично и латинично писмо – редослед ћирилице одозго надоле и слева у десно.

### 3.2 Имена страног порекла

Властита имена страног порекла се, по правилу, у српском не пишу у изворном облику него се транскрибују. Ово начело се подједнако примењује како на лична тако и на геополитичка имена. Правопис српског језика (Пешикан et al., 1993, чланови 101–180) допушта употребу изворног облика у текстовима на српском записаним латиничним писмом, али се ово ретко користи у пракси. Један од разлога је што транскрибовање поједностављује објављивање текста на оба писма и омогућава аутоматску конверзију текста из једног у друго писмо.[7].

У српском се праописна или уобичајена транскрипција користи како би се изговор у изворном језику прилагодио српском фонолошком

[7] На пример, сви чланци на српској Википедији се могу прегледати и у ћириличном и у латиничном писму – видети, на пример, Страницу Википедије о *правописној транскрипцији* у српском. Исту могућност конверзије пружају поједини новински портали, као на пример, *Politika*.

систему. Правопис српског језика (Пешикан et al., 1993, articles 101–180) наводи транскрипциона правила за 27 језика укључујући ту и латински, стари и савремени грчки, јапански и кинески. Међутим, велики број властитих имена се не уклапа у ова правила, углавном због тога што је друкчији изговор прихваћен одавно или што друкчији облик више одговара српском језику и његовим морфолошким својствима. Неколико примера за ово је наведено у Правопису за географска имена: *Москва* (уместо *Масква*), *Волгоград* (уместо *Валгаграт*), *Њујорк* (уместо *Њујок*) и *Лајпциг* (уместо *Лајпцих*) као и за лична имена: *Ганди* (уместо *Гандхи*) и *Стриндберг* (umesto *Стриндберј*). За нека страна географска имена, српско име не одгобара ни изворном изговору ни транскрипцији као у примеру *Беч* за *Vienna*.

Вишечлана географска имена се по правилу транскрибују као полилексичка имена, као у примерима *Њу Хемпшир* (*New Hampshire*) и *Солт Лејк Сити* (*Salt Lake City*). Постоје ипак изузеци у односу на ово правило као у случају имена *Порторико* (*Puerto Rico*). Страна вишечлана географска имена које садрже као конституенте једну или више заједничких речи су понекада преведена, понекада делимично, а некада остају непреведена. На пример, *Rocky Mountains* се преводи као *Стеновите планине*, док је *Long Island* транскрибован као *Лонг Ајланд*. Исте заједничке речи су понекада преведене, а понекада само транскрибоване као у случају *Нови Јужни Велс* (*New South Wales*) према *Њу Делхи* (*New Delhi*).

За поједина страна имена у српском понекада постоје вишеструки облици као у примеру *Кот д'Ивоар* што представља транскрибовано име у званичној употреби за *Côte d'Ivoire* док његов преведени облик, *Обала слоноваче*, преовлађује у свакодневној употреби. Ово је често и у случајевима имена места са мешовитим становништвом као *Целовец* и *Клагенфурт* (Klagenfurt) (град у Аустрији) или оних локација чија су имена промењена из политичких разлога као што је то *Град Хо Ши Мин* (*Ho Chi Minh City*) уместо ранијег *Сајгон* (*Saigon*).

Транскрипциона правила није увек једноставно применити тако да су објављени додатни приручници који могу помоћи приликом писања страних имена као што су то *The transcription dictionary of English personal names* (Prćić, 1992) и *The English-Serbian dictionary of geographic names* (Prćić, 2004). Нажалост, информације у овим приручницима су понекада у супротности са Правописом: на пример, у (Prćić, 2004) транскрипција за *Rio de Janeiro* је *Рио де Жанејро* док Правопис

за исто име предлаже *Рио де Жанеиро* које није сасвим сагласно са транскрипционим правилима за португалски језик, али је одомаћено.

Имена организација показују особености у односу на друга властита имена. Она се чешће од других користе у оригиналном облику, а посебно акроними као што је *IBM* или *FBI*. Поред оваквих случајева, имена организација могу бити транскрибована као *Мајкрософт* (Microsoft) или преведена као *Организација за економску сарадњу и развој* (Organization for Economic Cooperation and Development). Штавише, за исту организацију су у употреби и транскрибовано и преведено име као у примеру *Британска телевизијска мрежа* и (ретко) *Бритиш броудкастинг корпорејшн* (British Broadcasting Corporation).Одговарајући акроними могу бити као у оригиналу: *BBC* или према изговору *Би-Би-Си* (Krstev et al., 2015).

### 3.3 Деривација

| | **Име** | **Становник**<br>**Присвојни придев** | **Становница**<br>**Присвојни придев** | **Придев** |
|---|---|---|---|---|
| континент | **Европа** | Европљанин<br>Европљанинов | Европљанка<br>Европљанкин | европски |
| надрегион | **Балкан** | Балканац<br>Балканчев | Балканка<br>Балканкин | балкански |
| земља | **Француска** | Француз<br>Французов | Францускиња<br>Францускињин | француски |
| регион | **Прованса** | Провансалац<br>Провансалчев | Провансалка<br>Провансалкин | провансалски |
| град | **Београд** | Београђанин<br>Београђанинов | Београђанка<br>Београђанкин | београдски |
| део града | **Дорћол** | Дорћолац<br>Дорћолчев | Дорћолка<br>Дорћолкин | дорћолски |

**Табела 4.** Имена становника и придеви изведени из појединих типова топонима у српском језику

Из већине географских властитих имена се могу извести друге именице и придеви[8].

Имена становника или демоними се изводе из различитих географских имена као што су имена за континенте, надрегионе, земље, регионе, градове или делове града као што је приказано у табели 4. За нека имена овог типа није могуће извести име за становника као у случају топонима *Осло* (*Oslo*) и тада се користи израз *становник Осла*. Ако се из топонима може извести име становника, тада се, по правилу, може извести и име становнице, а за оба ова изведена облика се могу извести облици присвојних придева, као и релациони придев који се односи на полазни топоним. За *Осло* се ни такав придев не може извести. С друге стране, од неких имена становника се могу извести и други придеви као у примеру *Парижанин* → *Парижанинов* (*који припада становнику Париза*) → *парижански* (*који се на начин понашања Парижанина*) (насупрот *париски* (*који се односи на Париз*)). У појединим случајевима се могу извести и деминутиви из имена становника као, на пример, *Српче* и *Српчић*, који су деминутивни облици од *Србин*, а обично се односе на децу.

У појединим случајевима могу се извести два или чак три имена за мушког становника што даље даје вишеструка имена становница и облике присвојног и релационог придева. Пример су:

– двоструко име изведено из *Кореја*[9]:
  - *Корејац* (м), *Корејка* (ж), *Корејчев* (м присв.пр.), *Корејкин* (ж присв.пр.), *корејски* (пр.);
  - *Кореанац* (м), *Кореанка* (ж), *Кореанчев* (м присв.пр.), *Кореанкин* (ж присв.пр.), *кореански* (пр.);
– троструко име изведено из*Париз* (*Paris*)[10]:
  - *Парижанин* (м), *Парижанка* (ж), *Парижанинов* (м присв.пр.), *Парижанкин* (ж присв.пр.), *париски* и *паришки* (пр.);
  - *Парижлија* (м), *Парижлијка* (ж), *Парижлијин* (м присв.пр.), *Парижлијкин* (f poss.);
  - *Паризлија* (м), *Паризлијка* (ж), *Паризлијин* (м присв.пр.), *Паризлијкин* (ж присв.пр.).

[8] Овде нећемо посматрати глаголе изведене из географских властитих имена као што је *пофранцузити се* (*понашати се као Француз или Францускиња*) како је објашњено у одељку 2.3.

[9] Ови примери су према (Стијовић, 2016).

[10] Ови примери су према (Стевановић, 1967).

За поједина вишечлана географска имена демоними и придружени придеви се изводе или слагањем чланова или употребом само једног од њих. Резултат је у оба случаја монолекемска реч као што је показано у табели 5[11]. Међутим, за бројна вишечлана географска имена се не могу извести демоними и одговарајући придеви.

| **Српско и изворно име** | **Становник и становница** | **Придев** |
|---|---|---|
| Кабо Верде<br>(Cabo Verde) | Кабоверђанин<br>Кабоверђанка | кабовердски |
| Буркина Фасо<br>(Burkina Faso) | Буркинац<br>Буркинка | буркински |
| Тринидад и Тобаго<br>(Trinidad and Tobago) | становник Тринидада и Тобага<br>становница Тринидада и Тобага | *описно* |
| Нови Сад | Новосађанин<br>Новосађанка | новосадски |
| Бачко Ново Село | становник Бачког Новог Села<br>становница Бачког Новог Села | *описно* |

**Табела 5.** Имена становника и придеви изведени из вишечланих имане топонима на српском

Придеви се изводе и из других типова географских имена као што су хидроними или ороними. Пример за хидрониме су *дунавски* изведен из *Дунав* или *сенски* из *Сена*, а за ороними *алпски* из *Алпи*, *копаонички* из *Копаоник*. За неке хидрониме и ороними се релациони придев не може извести као у примеру *Волга*. Ако се релациони придеви могу извести из вишечланих хидронима или оронима, онда су они монолексемске речи као у примерима *великоморавски* полазећи од *Велика Морава* и *старопланински* од *Стара Планина*.

Присвојни придеви могу бити изведени и из личних имена: имена, презимена и надимака. На пример, присвојни придев изведен из сваког од делова имена *Иво Лола Рибар* би били *Ивов*, *Лолин* и *Рибаров*. Из имена познатих личности се могу извести нове иенице и придеви.

[11] Примери су према (Стијовић, 2016).

На пример, из имена филозофа *Карла Маркса* генеришу се у српском изведенице: *марксизам* за доктрину, *марксиста* и *марксисткиња* за следбенике *марксизма*, *марксологија* за научну дисциплину, *марксолог* и *марксолошкиња* за истрживаче који пручавају *марксологију*, затим придеви *марксистички* и *марксолошки*. Многи од овако изведених придева и именица могу бити префиксирани нпр, префиксима *анти-*, *нео-*, *пост-*, итд. (Vitas and Krstev, 2013). Оваке изведенице нису узете у обзир у бази Prolex, као што је напоменуто у одељку 2.3.

Присвојни придеви се могу извести обично и из монолексемских имена организација. На пример, *Мајкрософтов* је присвојни придев изведен из *Мајкрософт*. Присвојни придеви се употребљавају и код акронима имена организација – у таквим случајевима се деривациони суфикс додаје на акроним полсе цртице, нпр. *IBM-ов*.

### 3.4 Граматичка својства

Властита имена у српском, као и именице и придеви изведени из њих, имају иста флективна својства као и заједничке именице и придеви.

Род географских имена, топонима, оронима и хидронима, може бити мушки, женски или средњи, а имена су променљива по падежима (седам различитих падежа). Број имена је непромељив и може бити или сингулар или плурал. Примери су дати у табели 6.

Географска имена су по правилу неаниматна премда постоји неколико збуњујућих примера: неколико градова у Србији носе имена знаменитих личности као што су *Јаша Томић* и *Алекса Шантић*. Ако се ова имена посматрају као неаниматна, онда би реченица *I travel to Jaša Tomić* била на српском *Путујем у Јаша Томић* која је неприхватљива, јер је Јаша Томић, као особа, обележен као аниматан[12].

Демоними изведени из географских имена су мушког рода (за становнике) и женског рода. Облик придева изведених из географских имена се мења према падежу, броју, роду и аниматности. Треба напоменути да присвојни придеви немају облике компаратива и

[12] Облик акузатива сингулара именица мушког рода зависи од аниматности: за неаниматне именице је он једнак облику номинатива, а за аниматне облику генитива. У овом примеру, предлог *у* захтева акузатив који је за (неаниматно) *Јаша Томић* једнак номинативу, док је за (аниматно) *Јаша Томић* једнак генитиву *Јашу Томића* (упоредити са реченицом *Милица се заљубила у Јашу Томића*.

| тип | изворни или енглески | име | број | род |
|---|---|---|---|---|
| топоним | Belgrade | Београд | сингулар | мушки |
| ороним | Olympos | Олимп | | |
| хидроним | Danube | Дунав | | |
| топоним | Karlovci | Карловци | плурал | |
| ороним | Alpes | Алпи | | |
| хидроним | Dardanelles | Дарданели | | |
| топоним | Athens | Атина | сингулар | женски |
| ороним | Aconcagua | Аконкагва | | |
| хидроним | Seine | Сена | | |
| топоним | Budějovice | Будјеовице | плурал | |
| ороним | Divčibare | Дивчибаре | | |
| хидроним | Plitvice | Плитвице | | |
| топоним | Valjevo | Ваљево | сингулар | средњи |
| ороним | Pohorje | Похорје | | |
| хидроним | Oranjerivier | Орање | | |
| топоним | Kaštela | Каштела | плурал | |

**Табела 6.** Географска имена у српском различитог рода и броја

суперлатива, као ни релациони придеви осим изузетно као у примеру *Војводина је најевропскији део Србије*.

Српска лична имена и надимци могу бити мушког или женског рода, у сингулару су и мењају се по падежима. Српска презимена су мушког рода и, у општем случају се мењају и у броју и про падежима. Нека презимена, углавном страног порекла, се не мењају према броју због морфолошких запрека. Сложена правила слагања се примењују на пуно име у српском чији облик зависи од рода личног имена и редоследа личног имена и презимена – једно правило је да је презиме, када припада женској особи, непроменљиво (Gucul-Milojević, 2010). Женске особе се понекада обележавају обликом присвојног придева презимена у женском роду или именицом женског рода изведеном из придева моцијом рода. Примери су дати у табели 7.

Имена организација се мењају по падежима али се њихов род и број не мењају и зависе, у општем случају, од морфолошких

| **Облик** | **Презиме** | **Женски облик** |
|---|---|---|
| номинатив сингулара | *Петровић* | *Петровићка* |
| | | *Петровићева* |
| генитив сингулара | | |
| | | *Петровићеве* |
| номинатив плурала | *Петровићи* | *Петровићке* |
| | | *Петровићеве* |

| **Облик** | **Пуно мушко име)** | **Пуно женско име** |
|---|---|---|
| номинатицв сингулрара | *Петар Петровић* | *Зорка Петровић* |
| | | *Зорка Петровићка* |
| | | *Зорка Петровићева* |
| генитив сингулара | *Петра Петровића* | *Зорке Петровић* |
| | | *Зорке Петровићке* |
| | | *Зорке Петровићеве* |

**Табела 7.** Мушка и женска лична имена и њихова промена

својстава њиховог имена у случају монолексемских речи, а од свосјтава носеће речи у случају вишечланих речи. На пример, *Мајкрософт* је мушког, док је *Сорбона* женског рода. Вишечлано име организације ***Универзитет*** *у Београду* је мушког, *Београдска аутобуска* ***станица*** женског, а ***Удружење*** *спортских новинара Београда* средњег рода. Имена организација ***Лекари*** *без граница* и *Међународне мировне* ***снаге*** имају само облик множине [13].

### 3.5 Дијалекти

У српском се користе две стандардне варијанте изговора, екавска и ијекавска. Оне се разликују по рефлеку протословенске фонеме *јат*: у екавској варијанти она се замењује најчешће са *е*, док се у ијекавској замењује слоговима *ије* или *је*.

Ове варијанте немају много утицаја на властита имена јер их већина и не садржи рефлекс фонеме *јат*. Када је то ипак случај, име се обично

[13] Носећа реч у овим вишечланим именима организација је наглашена црним слогом.

користи у само једној варијанти. На пример, у именима градова *Ријека* и *Ријека Црнојевића* заједничка именица се користи само у ијекавској варијанти – *ријека*, а не *река*). С друге стране, женско лично име настало из заједничке именице *вера/вјера* – постоји и у екавској *Вера* и у ијекавској варијанти *Вјера*. Међутим, лично име у једној варијанти изговора се неће мењати ако се нађе у тексту записаном у другој варијанти, то јест, оно је непроменљиво.

У полилексемским именима организација појављују се многе заједничке именице које могу бити у једној или другој варијанти изговора. Ове варијанте се онда одражавају и у именима организација у зависности од тога коју варијанту користи текст у коме се појављују, на пример, екавска варијанта *Светска банка* или ијекавска варијанта *Свјетска банка*.

## 4. Остварени резултати

### 4.1 Допринос српског језика моделу базе Prolex

Као што смо рекли у одељку 1., укључивање српског језика довело је до развоја бољег модела базе Prolex. Сарадња између истраживачких група са Универзитета у Туру и Универзитета у Београду је била плодна по много чему, али ми ћемо истаћи две најважније ствари: релација деривације и репрезентација облика.

**Релација деривације.** У одељку 3.3 смо представили сложеност деривационих правила српског језика, као што су могућности квази-систематског извођења из имена људи, као и из геополитичких имена (описни придеви и имена становника). На пример, (видети слику 1), из имена града *Београд* генерише се, као што је случај и у многим другим језицима *Београђанин* (мушки становник *Београда*), док се из *Београђанин* генерише *Београђанинов* (присвојни придев од мушког становника *Београда*). У енглеском и француском постоји само један ниво извођења: *Belgrade/Belgradian* у енглеском и *Belgrade/Belgradois* у француском. Први модел базе података није предвиђао релацију табеле *Derivative* са самом собом. Додали смо ту релацију у касније моделе и онда схватили да таква релација постоји и у француском: на пример, име награде, каква је Нобелова награда, дозвољава да се квази-систематски изведе глагол са значењем *доделити награду*, у овом примеру *nobeliser*,

док се из таквих глагола регуларно стварају други деривати, као *nobelisable* (особа која је могући изабраник Нобеловог комитета), и тако даље.

**Репрезентација облика.** Властита имена смештамо у моделу базе података Prolex у две табеле, *Prolexeme* (најдужи облик имена) и *Alias* (други облици). Међутим, у LMF репрезентацији (видети слику 2) ова разлика нестаје јер су сви алијаси равноправни уноси које повезује значење.

Поставило се питање да ли су име записано ћирилицом и исто име записано латиницом алијаси или не. Могло би изгледати чудно да је реч алјас самој себи! Али, ипак није тако. Прво смо размотрили могућност да дефинишемо две пролексеме за српски језик (за ћирилицу и латиницу), али овакво решење нарушава услов јединствености пројекције стожера за одређени језик. С тога смо прихватили друго решење које дефинише два лексикона, српски ћирилични лексикон и српски латинични лексикон. Коначно, у трећој верзији базе Prolex произведеној у Универзитету у Туру систематски смо додали под *Word Form* једну или више *Form Representations*. На пример, за *Београд* сада имамо:

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>Београд</Lemma>
  <WordForm grammaticalGender="masculine"
  grammaticalNumber="singular"
  grammaticalCase="nominative"
  grammaticalAnimacy="nonAnimate">
    <FormRepresentation script="cyrl">
      Београд
    </FormRepresentation>
    <FormRepresentation script="latn">
      Beograd
    </FormRepresentation>
  </WordForm>
  ...
</LexicalEntry>
```

Пошто је такав избор начињен, додали смо за неке уносе и разликовање између екавског и ијекавског изговора (видети одељак 3.5), за шта смо користили исту LMF репрезентацију:

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>Немачка</Lemma>
  <WordForm grammaticalGender="feminine"
  grammaticalNumber="singular"
  grammaticalCase="nominative"
  grammaticalAnimacy="nonAnimate">
    <FormRepresentation script="cyrl">
      Немачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ekavsk">
      Немачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ijekavsk">
      Њемачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ijekavsk">
      Њемачка
    </FormRepresentation>
    <FormRepresentation script="latn">
      Nemačka
    </FormRepresentation>
  </WordForm>
  ...
</LexicalEntry>
```

Коначно, користили смо исти концепт *form representation* за неке различите варијанте писања, на пример у горњем примеру, *Њемачка* и *Њемачка* (други начин се ретко користи), али и за неке друге случајеве, као што су разлике у транскрипцији, *Рио де Жанејро* и *Рио де Жанеиро* (видети одељак 3.2), или различити облици за исти скуп вредности граматичких категорија – презиме *Чехов* има три варијанте облика датива у сингулару: *Чехову*, *Чеховом* и *Чеховому*. Овај приступ смо применили и на друге језике елиминишући тако категорију алијаса *Variant*.

### 4.2 Имплементација базе Prolex

У табели 8 су дати неки подаци о имплементацији српског језика. Ручно смо унели пролексеме и повезали их са стожером из одабраног скупа француских пролексема, а додали смо и неке алијасе. Потом смо аутоматски генерисали имена изведена деривацијом и, наравно, све флективне облике пролексема, алијаса и изведених имена.

| | |
|---|---|
| Serbian prolexemes | 8 526 |
| Serbian aliases | 21 |
| Serbian derivatives | 920 |
| Serbian instances | 108 325 |
| Serbian pivot relations | 29 567 |

**Табела 8.** Имплементација српског језика

Овим бројевима бисмо могли да додамо и невероватан број примерака имена изведених из *Београд*.[14] Да бисмо употпунили слику 1 са свим примерцима, морали би да додамо 626 облика...

Ако се вратимо на пример из одељка 1.: *Београђанка ми је рекла да је Дунав прелеп* сада добијамо:

Београђанка
    female inhabitant (категорија деривације)
    Belgrade (пролексема)
        city (класификациони контекст)
        Serbia (доступност)
        capital (контекст доступности)
→ The female inhabitant of the city of Belgrade, capital of Serbia
ми је рекла да је
→ has told me that
Дунав
    river (класификациони контекст)
→ the Danube River
прелеп

[14] У поређењу с енглеским, па чак и француским!

→ is splendid

### 4.3 Закључак

Показали смо да је сложеност морфологије српског језика знатно допринела пројекту вишејезичног речника Prolex. Побољшања која су учињена да би се српски језик укључио у базу показала су се корисна и за друге језике из базе Prolex. Ова побољшања се посебно односе на решавање проблема деривације и на репрезентацију вишеструких облика. То је потврђено када су у базу укључени и неевропски језици, попут арапског, јер је интерна структура базе могла да их моделира. Овај рад је такође потврдио колико је важно да се у лингвистичке вишејезичне пројекте укључе не само блиски већ и разноврсни језици.

**Захвалност**

Аутори се захваљују Универзитету у Туру што је омогућио ово истраживање финансирањем једномесечног боравка проф. Цветана Крстев Универзитету. Аутори се такође захваљују фонду *Egide Pavle Savić* који је подржао пројекат којим је покренута сарадња између универзитета у Туру и Београду 2004. Део истраживања је подржало Министарство просвете, науке и технолошког развоја Републике Србије кроз пројекте 138006 и III47003.

## Литература

Ariel, M. *Accessing Noun Phrases Antecedents*, 1990

Auer, S. and J. Lehmann. “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”. У *ESWC 2007*, no. 4519, LNCS, 503–517. 2007

Boitet, C. *Pros and cons of the pivot and transfer approaches to multilingual machine translation*, 93–106. 1988

Chinchor, N. “Muc-7 Named Entity Task Definition”, 1997, URL http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

Coseriu, E. “Le double problème des unitès dia-s”. У *Les Cahiers dia. Etudes sur la diachronie et la variation linguistique, Universitè de Gent, Belgique*, Vol. 1, 9–16. 1998

Elashter, Mouna and Denis Maurel, "Estimer la notoriété d'un nom propre via Wikipedia", У *TALN 2016. Paris*, 2016, URL https://jep-taln2016.limsi.fr/actes/

Gruber, T. R.. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing". *Int. Journal of Human-Computer Studies* Vol. 43 (1995): 907–928

Gucul-Milojević, Sandra. "Personal Names in Information Extraction". *INFOtheca* Vol. 11, no. 1 (2010): 53a–63a

Hachey, B, W Radford, J Nothman, M Honnibal and R Curran, J. "Evaluating entity linking with Wikipedia", У *Artificial Intelligence*, 194, 130–150. 2013

Hoffart, J., F. M. Suchanek, K. Berberich and G. Weikum. "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia". *Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources* (2012)

ISO/TC 37/SC 4. *Language resource management - Lexical markup framework (LMF)*, 2007. http://lirics.loria.fr/documents.html

Krstev, Cvetana, Duško Vitas and Ranka Stanković. "A Lexical Approach to Acronyms and their Definitions". У *Proceedings of 7th Language & Technology Conference, November 27–29, 2015, Poznań, Poland.* 2015

Lecuit, Émeline, Denis Maurel and Duško Vitas. "A tagged and aligned corpus for the study of Proper Names in translation". У *Workshop Annotation and exploitation of parallel corpora, International Conference Recent advance in Natural Language Processing (RANLP 2011),*, 11–18. 2011, URL http://aclweb.org/anthology/W11-43

MacDonald, D. *Internal and external evidence in the identification and semantic categorisation of Proper Names*, 21–39. 1990

Mangeot, M. "Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links". У *7th Workshop on Advanced Information Network and System, Kasetsart University, Bangkok, Thailand.* 2000

Maurel, D. "Prolexbase: A Multilingual relational Lexical Database of Proper Names". У *LREC 2008*, 334–338. 2008

McNamee, P., H. T. Dang, H. Simpson, P. Schone and S. M. Strassel. "An evaluation of technologies for knowledge base population". У *LREC 2010*, 369—372. 2010

Navigli, Roberto and Simone Paolo Ponzetto. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". *Artificial Intelligence* Vol. 193 (2012): 217–250

Prćić, Tvrtko. *Transkripcioni rečnik engleskih ličnih imena [Transcription dictionary of English personal names]*. Nolit, 1992

Prćić, Tvrtko, *Englesko-srpski rečnik geografskih imena [English-Serbian dictionary of geographic names]*. Zmaj, 2004

Savary, A., L. Manicki and M. Baron. “Populating a Multilingual Ontology of Proper Names from Open Sources”. *Journal of Language Modelling* Vol. 1, no. 2 (2013)

Vitas, Duško and Cvetana Krstev. “Derivational Morphology in E-Dictionaries of Serbian”. У *Proceedings of the 32nd International Conference on Lexis and Grammar, September 10–14, 2013, Faro, Portugal.* 2013

Zakon, ур. *Zakon o službenoj upotrebi jezika i pisma [Law on Official Usage of Language and Script]*. Službeni glasnik Republike Srbije, 2010

Пешикан, Митар, Јован Јерковић and Мато Пижурица, . *Правопис српскога језика [The Orthography of Serbian Language]*. Матица српска, 1993

Стевановић, Михаило и др., ур. *Речник српскохрватскога књижевнога језика [Serbo-Croatian literary language dictionary]*. Матица српска, 1967

Стијовић, Рада. “Званични пуни скраћени називи држава на српском и енглеском језику [Official and shorten names of countries in Serbian and English]”, 2016, internal report