Serbian Language Integration in Prolexbase Multilingual Dictionary

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2018.18.2.2

ABSTRACT: In this paper we present the multilingual dictionary of proper names *Prolexbase*, particularly the Serbian volume. We also present the complexity of proper names in the Serbian language, especially those related to their translation: their orthography, derivation, inflection and dialect variations. We describe the model of the *Prolexbase*, with the emphasize on the solutions that had to be made to include the Serbian language in the database (the use of two alphabets, several levels of derivations, the existence of multiple forms). At the end, we give some figures that corroborate the presence of the Serbian language in the *Prolexbase*.

KEYWORDS: proper names, multilingual database, ontology of proper names, LMF format, Serbian language, Prolexbase.

PAPER SUBMITTED: 4 December 2018
PAPER ACCEPTED: 14 December 2018

Cvetana Krstev
cvetana@matf.bg.ac.rs
University of Belgrade
Faculty of Philology, Serbia

Denis Maurel
denis.maurel@univ-tours.fr
University of Tours, France

Duško Vitas
vitas@matf.bg.ac.rs
University of Belgrade
Faculty of Mathematics, Serbia

1 Motivation

As other particularities of language (neologisms, multiwords, idioms and so on), proper names can be responsible of amazing errors. For instance, how should *Bush* be translated to Serbian: as *грм* (plant) or *Буш* (personal name)? Are *Casablanca* and *White House* the same location? It is a common belief that proper names cannot be translated. In fact, all the sorts of translation processes (adaptation, layer, literal translation and so on) are used by translators when they transfer them from a source-language text to a target-language text (Lecuit et al., 2011).

Proper names are also a challenge for Natural Language Processing (NLP) and, more generally, for *Named Entity* tasks[1]. First tasks related to

[1] Named Entities are usually defined by a referent or a kind of uniqueness.

named entities were to complete data bases in the Message Understanding Conferences MUC-6 and MUC-7 conferences with answers to the questions, such as “who did a terrorist attack”, “where?”, “when?” or “what firm took holdings in another one?”, “at what height?”, “for how much dollars?” and so on (Chinchor, 1997). Today, the challenge is almost the opposite: entities in a text have to be linked with database entries (Hachey et al., 2013), i.e. proper names have to be disambiguated (see for instance the Text Analysis Conferences (McNamee et al., 2010)). One often uses for these tasks Wikipedia and a number of other semantic data bases, as DBpedia (Auer and Lehmann, 2007), GeoNames, YAGO2 (Hoffart et al., 2012), BabelNet (Navigli and Ponzetto, 2012). These databases constitute a part of the Link Open Data system (LOD) where proper names have a particularly important place.

Prolexbase is a Multilingual Relational Database of Proper Names (Maurel, 2008). The aim of Prolexbase is to assist in their translation. It merges morphology, derivation and semantic relations. For instance, if a sentence in Serbian *Београдска жена ми је рекла да је Дунав прелеп* has to be translated, it may be helpful to expand it: *The female [inflection] inhabitant of the city [semantic expansion] of Belgrade [derivative relation] in Serbia [accessibility relation] has told me that the Danube River [semantic expansion] is splendid.* We will return to this example at the end of this paper.

The first version of Prolexbase (covering eight languages, French, German, English, Italian, Dutch, Polish, Portuguese and Spanish) has been supported by the French *RNTL-Technolangue Project* (2003-2005). Actually, the model of the database was constructed and its coverage for French was good while for other languages it was weak. However, at the same time, an *Egide Pavle Savic* project (2004-2005) has initiated with the aim to add the Serbian language to Prolexbase. The involvement of the Serbian team was very important as it helped to realize the complexity of the morphology and the derivation in the model, that was too French/English centered. Another problem was the presence of two scripts, Cyrillic and Latin. In this first version, a not satisfactory solution was chosen for the problem of two scripts: two volumes were built for the Serbian language, one using the Cyrillic alphabet and the other using the Latin alphabet.

The second version of Prolexbase has been supported by the *Hubert Curien Polonium* project which brought a good coverage for Polish and English Savary et al. (2013). The Serbian part has been significantly improved in the third version of Prolexbase, as a result of a one month visit of Professor Cvetana Krstev which was sponsored by The University of Tours. We improved the coverage of Serbian and we mostly mixed the two alphabets representations in only one volume. We also prepared a possible description of dialect forms, as *Ekavian* and *Ijekavian*.

2 Prolexbase

2.1 The Prolexbase model

Since Prolexbase is a multilingual databases we need a model enabling the linking of different occurrences of proper names in different languages. We choose to define the linguistic class of proper names (and their derivations) as an ontology in the sense of (Gruber, 1995): "A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose... An ontology is an explicit specification of a conceptualization".

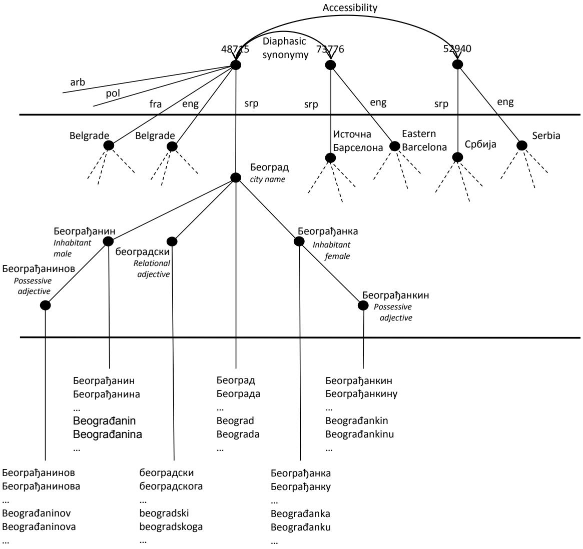

Figure 1. The example of *Београд* in the Prolexbase model

The center of the Prolexbase ontology is the *conceptual proper name*, the *pivot*, which represents the referent from a particular point of view. For instance, *Pope Francis* and *Jorge Mario Bergoglio* or *Београд* (*Belgrade*) and *Источна Барселона* (*Eastern Barcelona*). The translation by pivots (Boitet, 1988) is not very usual today, although a pivot can be refined for some language and not for others. For instance, in the Papillon project (Mangeot, 2000), the pivot for *rice* in English corresponds to two refined pivots in Japanese, *raw rice* and *cooked rice*. For the *conceptual proper name*, no refinement is needed, so we can use this model without any problems. For each language, the pivot is linked to an unique set of proper names, the *prolexeme*. This set contains the proper name and, eventually, its aliases and its morphosyntactic derivatives (see 2.3). The pivots constitute the conceptual level of the model and the prolexemes its linguistic level. The ontology is completed by two other levels, one at the top, the metaconceptual level (types and supertypes), and the other one at the bottom, the instance level (forms of proper names – as they appear in a written text). Figure 1 illustrates the model with the proper name *Београд*.

2.2 The LMF format

Prolexbase is a free and open source resource, under LGPL-LR license[2]. The exchange format is inspired by the Lexical Markup Format (LMF) (ISO/TC 37/SC 4, 2007). Figure 2 shows LMF classes for representing the Prolexbase model. It represents a selection of classes of the LMF core model with additional parts from LMF extensions (packages Morphology, NLP semantics, NLP multilingual notations and NLP syntax). Multilingual descriptions are represented with grey boxes. The whole resource is represented by the class *Lexical Resource* to which some information is linked, such as the language codes, scripts, characters used in the whole resource (class *Global Information*). The resource contains the conceptual level (class *Sense Axis*) and the linguistic level, with several lexicons (class *Lexicon*) that are monolingual descriptions. One of them is the Serbian lexicon. The lexical entries are all lemmas of a prolexeme (proper names, aliases and derivatives) with their word forms (all the instances): classes *Lexical Entry*, *Lemma*, *Word Form* and *Form Representation*.

These lemmas are linked with the senses that are pivots with category assigned, as *relational adjective* or *male possessive adjective* (classes *Sense* and *Sense Relation*). These pivots are defined in the class *Senses Axis* belonging to the multilingual part of the resource. The class *Sense Axis Relation*

[2] Prolexbase (on-line)

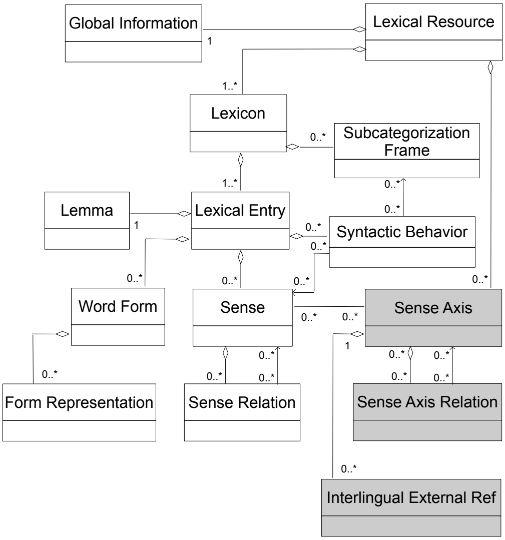

Figure 2. The LMF schema of Prolexbase

represents relations between conceptual proper names while the class *Interlingual External Ref* represents typologies. We also note some information about classifying contexts of proper names (class *Subcategorization Frame* and some idiosyncratic collocations (class *Syntactic Behavior*). These classes are not yet used in the Serbian volume and some examples would be the use of different prepositions for proper names, like in *Србија је* ***на*** *Балкану и* ***у*** *Европи* (*Serbia is in the Balkans and in Europe*).

2.3 Relations

The large part of Prolexbase consists of relations between pivots (language independent relations): synonymy, meronymy and accessibility.

The synonymy relation, or more precisely, the quasi-synonymy relation, is a relation between two pivots referring to the same referent for which different points of view exist. The translator has to choose the same point of view, which is not always possible. We distinguish between three different points of view, designated by three diasystematic features of (Coseriu, 1998):

- Diachronic: variations depending on time. *Савезна Република Југославија* (*Federal Republic of Yugoslavia*) versus *Државна Заједница Србија и Црна Гора* (*the State Union of Serbia and Montenegro*);
- Diastratic: variations depending on sociocultural stratification. *Јосип Броз* (*Josip Broz*) versus *Тито* (*Tito*);
- Diaphasic: variations depending on the usage purpose. *Београд* (*Belgrade*) versus *Источна Барселона* (*Eastern Barcelona*).

The meronymy relation, a partitive relation, is a relation of inclusion. The examples are geographical inclusion: Serbia is in the Balkans, which is a part of Europe, and temporal relation: The bombing of Belgrade on April 6, 1941 is a part of (happened during) the Second World War. We extend this relation to other domains, such as economy, nationality and so on.

The accessibility relation (Ariel, 1990), an associative relation, means that a proper name is accessible through some other proper names. In dictionaries, proper names, contrary to common nouns, do not have definitions – they are rather replaced by a relation to some more known name. Thus, this relation is rarely symmetric: in dictionaries, one can read that Aaron is the brother of Moses, but Moses is not presented as the brother of Aaron, but as the leader of the Hebrews. Consequently, Aaron is accessible through Moses and Moses is accessible through the Hebrews’ story. We distinguish 12 such relations:

- Relative: *Арон* (Aaron) is the brother of *Мојсије* (Moses);
- Capital: *Београд* (Belgrade) is the capital of *Србија* (Serbia);
- Leader: *Тито* (Tito) is a political leader of *Југославија* (Yugoslavia);
- Founder: *Растко Немањић* (Rastko Nemanjić) founded the *Српска православна црква* (Serbian Orthodox Church);
- Follower: *Петар* (Peter) is a disciple of *Исус* (Jesus);
- Creator: *Госпођа министарка* (The Cabinet Minister’s Wife) is a comedy by *Бранислав Нушић* (Branislav Nušić);
- Manager: *Ранко Жеравица* (Ranko Žeravica) was a Serbian basketball coach who used to manage the *Југословенска кошаркашка репрезентација* (Yugoslav national basketball team);

- Tenant: *Александар Вучић* (Aleksandar Vučić) is the tenant of the *Нови двор* (Novi dvor);
- Heir: *Кнез Михаило Обреновић* (Prince Mihailo Obrenović) was the heir of *Кнез Милош Обреновић* (Prince Miloš Obrenović);
- Headquarters: In *Београд* (Belgrade) are the corporate headquarters of *Montinvest Beograd*;
- Rival: *Партизан* (Partizan) is the football rival of *Црвена звезда* (Red Star);
- Companion: *Мирко* (Mirko) was *Славко*'s (Slavko) beloved comrade and brother-in-arms.[3]

The language dependent relations are frequency, is-an-alias, is-a-derivative, collocation, context and eponymy.

A prolexeme is the set of all lemmas semantically linked to a proper name in one particular language. For instance, the prolexeme *Београд* (*Belgrade*) consists of: Београд, београдски, Београђанин, Београђанка, Београђанинов, Београђанкин as shown in Figure 1. The three main relations at the language dependent level are frequency, is-an-alias and is-a-derivative. The frequency which indicates whether the proper name is well known can have three possible values: commonly used, infrequently used and rarely used. Today, this frequency can be calculated from LOD, mainly from Wikipedia (Elashter and Maurel, 2016). Aliases are different variations of a proper name: short forms, abbreviations, acronyms, different orthographies, alternate transcriptions, diatopic quasi-synonyms and explanations. Only the morphosemantic derivatives of a prolexemes are considered (and their derivatives). For instance, *пастеризовати* (to pasteurize), referring to the process of partial sterilization, is a derivative of the name *Пастер* (Pasteur), but it is not semantically linked to it.

Collocation and context relations concern the local usage of proper names. In some languages, such as French, country names are often preceded by an article, masculine or feminine without any particular reason which one should be used: for instance, one says *la* (feminine) *France* and *le* (masculine) *Montenegro*. The context is a relation between a proper name and typical words appearing with it. The context can be classified as a classifying or an accessibility context. The classifying context is an expansion of a noun phrase (capital, king, coach, etc.), called by MacDonald (1990) the *external structure* of a proper name. The classifying context is can be

[3] Two main characters of a very popular Yugoslav comic book series about two Partisan couriers.

useful in translation. For instance, *Сава* (*Sava*) is translated in English as the *Sava River*. The accessibility context is a noun phrase that implements the accessibility relation between two pivots. It can be regarded as a sort of the explanation of a proper name by a link to a well-known proper name. For instance, one can translate *Београд* (*Beograd*) as *Belgrade, the capital of Serbia*.

The eponymy relations differs from other relations: it tells us that the translation does not refer to a proper name but to a common noun (antonomasia), as *жилет* (*žilet*) in Serbian that designates all razor blades and not only the Gillette ones, or to a terminological term, as *Parkinson's disease* or *Pythagoras' theorem*, or again to an idiomatic phrase, as *све ми је равно до Косова* (*sve mi je ravno do Kosova*) that says literally *It's all straight to me up to Kosovo* and idiomatically *It's all equal to me* or *I don't care at all.*

2.4 Typologies

The meta-conceptual level deals with the concept existence and the typology of proper names.

The existence concept divides proper names into three groups: the historical ones that exist or existed, as *Београд* (*Belgrade*); the religious ones whose existence depends on one's beliefs, as *Архангел Михаил* (*Michael the Archangel*), or fictional ones invented by authors. Generally, name belonging to two later categories need to be translated, as *Snow White* that is translated in Serbian as *Снежана.*

The aim of Prolexbase typology is to classify proper names. We defined four big classes (named *supertypes*) corresponding to the primary semantic features: the human (*Anthroponyms*), the location (*Toponyms*), the concrete (*Ergonyms* – artifacts and work names) and the event (*Pragmonyms*). We defined thirty types in total, presented in Table 1. This typology defines the primary hypernymy relation between a pivot and type. We completed it with another relation, the secondary hypernymy relation, which is a metonymy relation between types, as seen in Table 2.

3 Proper names in the Serbian language

3.1 Alphabets

In Serbia, the use of Cyrillic alphabet is prescribed by law (Zakon, 2010, article 1), while the use of Latin alphabet is permitted in special situations

Proper Name						
Anthroponym			Ergonym	Pragmonym	Toponym	
Individual	Collective					
		Group				Territory
Celebrity	Dynasty	Association	Object	Disaster	Astronym	Country
Patronymic	Ethnonym	Ensemble	Work	Feast	Building	Region
First name		Firm	Thought	History	Geonym	Supranational
Pseudo-anthroponym		Institution	Product	Manifestation	Hydronym	
		Organization	Vessel	Meteorology	City	
					Way	

Table 1. The Prolexbase typology – the primary hypernymy

(traffic signs, street names, etc.). However, due to historical and other reasons Latin alphabet is widely used and it is defined as equal to Cyrillic in the Serbian orthography (Пешикан et al., 1993, articles 1–6). The Serbian alphabet, both Cyrillic and Latin, has 30 letters; 1-1 correspondence is established between these two sets, as presented in Table 3. The order of letters in Cyrillic and Latin alphabet is different; letters in Table 3 are presented in the Cyrillic order. The Serbian Latin alphabet does not use some of the 26 letters of English alphabet[4] – Q, W, X and Y. It uses some letters with diacritics – Č, Ć, Đ, Š and Ž – while some are represented as combinations of existing letters as digraphs – Lj, Nj and Dž. Digraphs are in electronic texts traditionally represented with two codes of consisting letters, although Unicode has introduced specific codes for these symbols[5]. One should note that capital letters of digraphs lj, nj and dž can be represented in two ways, with only the first composing capital letter – Lj, Nj and Dž – and with both capital letters – LJ, NJ and DŽ. The later case is used when the whole word (or a longer text) is written in capital letters. This is reflected in Unicode as well that has separate codes for these representations.

3.2 Names of foreign origin

Proper names of foreign origin, as a rule, are not written in Serbian using the original script and spelling, they are rather transcribed. It is applied

[4] Letters represented in ASCII code

[5] See code page Unicode Latin Extended-B

Types	**Secondary hypernymy**
Country Region Supranational Territory	Collective anthroponym
City	Collective anthroponym Ergonym
Buiding Way	Ergonym
Feast History Manifestation	
Association Ensemble Firm Group Institution Organization	Ergonym Toponym
Vessel	Toponym

Table 2. The secondary hypernymy relation

to personal names as well as geopolitical names. The Orthography manual (Пешикан et al., 1993, articles 101–180) permits the use of the original script and spelling for Serbian texts written in Latin; however, in practice it is rarely used. One of the reasons is that the use of transcription for both scripts facilitates publications in both of them, as well as switching between them on Web pages[6].

In Serbian, orthographic, or practical, transcription is used to customize sounds from the original language to the standard Serbian spelling system. The Orthography manual (Пешикан et al., 1993, articles 101–180) lists transcription rules for 27 languages, including Latin, Ancient and Modern Greek, Japanese and Chinese. However, there are number of proper names that do

[6] For instance, all articles in Serbian Wikipedia can be viewed in Cyrillic and Latin alphabet – see, for instance, the Wikipedia page about *Orthographic transcription* in Serbian. The same possibility is offered by some newspaper portals, for instance *Politika*.

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

Table 3. The Serbian alphabet – Cyrillic and Latin; the order of Cyrillic alphabet is present top-down, from left to right.

not conform to these rules, mostly because they are used as such for a very long time, or because they better suit Serbian language and its morphological properties. Some examples listed in the Orthography manual for geographic names are *Москва* (Moscow) (instead of *Масква*), *Волгоград* (*Volgograd*) (instead of *Валгаграт*), *Њујорк* (*New-York*) (instead of *Њујок*) and *Лајпциг* (*Leipzig*) (instead of *Лајпцих*) and for personal names *Ганди* (*Gandhi*) (instead of *Гандхи*) and *Strindberg* (instead of *Стриндберј*). For some foreign geographic names the Serbian name is neither the original nor its transcription, for instance *Беч* for *Vienna.*

Multi-unit geographic names are, as a rule, transcribed into multiword names, for instance, *Њу Хемпшир* (*New Hampshire*) and *Солт Лејк Сити* (*Salt Lake City*). There are exceptions to this rule as well, for instance *Порторико* (*Puerto Rico*). The foreign geographic multi-unit names that have as constituents one or more common words are sometimes translated, partially translated or not translated at all. For instance, *Rocky Mountains* is translated as *Стеновите планине*, while *Long Island* is transcribed as *Лонг Ајланд.* The same common words are sometimes translated and sometimes transcribed, for instance *Нови Јужни Велс* (*New South Wales*) vs. *Њу Делхи* (*New Delhi*).

Sometimes multiple variants exist in Serbian for a single foreign proper name. For instance, *Кот д'Ивоар* is the transcribed name in official use for *Côte d'Ivoire* while its translated name *Обала слоноваче* prevails in everyday use. It is more often the case for names of location with mixed inhabitants, like *Целовец* and *Клагенфурт* (Klagenfurt) (a city in Austria), or for names of locations that changed names due to political reasons, for instance *Град Хо Ши Мин* (*Ho Chi Minh City*), former *Сајгон* (*Saigon*).

Transcription rules are not always easy to master, so additional manuals are published that can help in writing proper names of foreign origin, for instance *The transcription dictionary of English personal names* (Prćić, 1992) and *The English-Serbian dictionary of geographic names* (Prćić, 2004). It is unfortunate that information in these manuals sometimes contradicts the Orthography manual: for instance, in (Prćić, 2004) the transcription for *Rio de Janeiro* is *Рио де Жанејро* while the Orthography manual for the same name suggests *Рио де Жанеиро* that does not conform completely to the transcription rules for Portuguese but is established as a name.

Organization names are specific as compared to other proper names. They are more often than others used in original, especially acronyms such as *IBM* or *FBI*. Besides that, organization names can be transcribed *Мајкрософт* (Microsoft) or translated *Организација за економску сарадњу и развој* (Organization for Economic Cooperation and Development). Moreover, for some organizations both transcribed and translated names are used, for instance *Британска телевизијска мрежа* and (rare) *Бритиш броудкастинг корпорејшн* (British Broadcasting Corporation). The corresponding acronyms can be either in original *BBC* or in spelling *Би-Би-Си* (Krstev et al., 2015).

3.3 The derivation

Nouns and adjectives can be derived from most geographic proper nouns[7].

Names of inhabitants, or demonyms, are derived from various geographic proper names: continents, super-regions, countries, regions, cities, and city districts as represented in Table 4[8]. For some names of these types it is not possible to derive a name for an inhabitant, e.g. *Осло* (*Oslo*) and a phrase is used instead *становник Осла* (an *inhabitant of Oslo*). If a name of a male inhabitant can be derived, than, as a rule, a name for a female inhabitant can also be derived, and for both of them possessive adjectives can be

[7] We will not consider verbs derived from geographic proper nouns, such as *пофранцузити се* (*become as a Frenchman/Frenchwoman*) as explained in Subsection 2.3.

[8] Дорћол (Dorćol) is a central district of Belgrade.

	Name	**Inhabitant (m.)** **Possessive adj.**	**Inhabitant (f.)** **Possesive adj.**	**Adjective**
continent Europe	**Европа**	Европљанин Европљанинов	Европљанка Европљанкин	европски
super-region Balkan	**Балкан**	Балканац Балканчев	Балканка Балканкин	балкански
country France	**Француска**	Француз Француозов	Францускиња Францускињин	француски
region Provence	**Прованса**	Провансалац Провансалчев	Провансалка Провансалкин	провансалски
city Belgrade	**Београд**	Београђанин Београђанинов	Београђанка Београђанкин	београдски
city district Dorćol	**Дорћол**	Дорћолац Дорћолчев	Дорћолка Дорћолкин	дорћолски

Table 4. Names of inhabitants and adjectives derived from certain types of toponyms in Serbian

derived, as well as an adjective describing something as related to the initial toponym. For *Осло* such an adjective cannot be derived either. On the other hand, for some names of inhabitants other adjectives can be derived, for instance, *Парижанин* (a *male inhabitant of Paris*) → *Парижанинов* (*belonging to a male inhabitant of Paris*) → *парижански* (*referring to, in a style of inhabitants of Paris*) (opposed to *париски* (*referring to Paris*)). Occasionally, diminutives can be derived from names of inhabitants, e.g. *Српче* and *Српчић*, diminutives of *Србин* (an inhabitant of Serbia), sometimes referring to children.

In certain cases, double or even triple names of male inhabitants can be derived, leading to multiple names of female inhabitants, possessive and descriptive adjectives. The examples are:

- double names derived from *Кореја* (*Korea*)[9]:
    - *Корејац* (m), *Корејка* (f), *Корејчев* (m poss.), *Корејкин* (f poss.), *корејски* (adj.);
    - *Кореанац* (m), *Кореанка* (f), *Кореанчев* (m poss.), *Кореанкин* (f poss.), *кореански* (adj.);

[9] These examples are corroborated in (Стијовић, 2016).

- triple names derived from *Париз* (*Paris*)[10]:
    - *Парижанин* (m), *Парижанка* (f), *Парижанинов* (m poss.), *Парижанкин* (f poss.), *париски* and *паришки* (adj.);
    - *Парижлија* (m), *Парижлијка* (f), *Парижлијин* (m poss.), *Парижлијкин*(f poss.);
    - *Паризлија* (m), *Паризлијка* (f), *Паризлијин* (m poss.), *Паризлијкин* (f poss.).

For certain multi-unit geographic names demonyms and corresponding adjectives are derived either by composing its constituents or by using just one of them. In either case as a result simple words are derived as illustrated in Table 5[11]. However, for a number of multiword geographic names demonyms and corresponding adjectives cannot be derived.

Name – Serbian original	**Inhabitant – male female**	**Adjective**
Кабо Верде (Cabo Verde)	Кабоверђанин Кабоверђанка	кабовердски
Буркина Фасо (Burkina Faso)	Буркинац Буркинка	буркински
Тринидад и Тобаго (Trinidad and Tobago)	становник Тринидада и Тобага становница Тринидада и Тобага	*descriptive*
Нови Сад	Новосађанин Новосађанка	новосадски
Бачко Ново Село	становник Бачког Новог Села становница Бачког Новог Села	*descriptive*

Table 5. Names of inhabitants and adjectives derived from multi-unit names of toponyms in Serbian

Adjectives are derived from other types of geographic names, hydronyms and oronyms, as well. Examples are, hydronyms *дунавски* derived from *Дунав* (*Danube*), *сенски* derived from *Сена* (*La Seine*) and oronyms *алпски*

[10] These examples are corroborated in (Стевановић, 1967).

[11] These examples are corroborated in (Стијовић, 2016).

derived from *Алпи* (*Alps*), *копаонички* derived from *Копаоник* (a mountain in Serbia). For some hydronyms and oronyms adjectives cannot be derived, for instance *Волга* (Volga River). Adjectives derived from multiword hydronyms and oronyms, if they exist, are simple words, for instance, *великоморавски* derived from *Велика Морава* (a river in Serbia) and *старопланински* derived from *Стара Планина* (a mountain in Serbia).

Possessive adjectives can be derived from personal names: first names, surnames and nicknames. For instance, possessive adjectives derived from all parts of the name *Иво Лола Рибар*[12] would be *Ивов*, *Лолин* and *Рибаров*. Various nouns and adjectives can be derived from names of famous persons. For instance, the name of the philosopher *Карл Маркс* (*Karl Marx*) yields in Serbian: *марксизам* for a doctrine, *марксиста* and *марксисткиња* for supporters of *марксизам*, *марксологија* for a scientific discipline, *марксолог* and *марксолошкиња* for scientists studying *марксологија*, the adjectives *марксистички* (*relating to Marxists*) and *марксолошки* (*relating to Marxologs*). Many of these derived adjectives and nouns can be prefixed, eg. with *анти-*, *нео-*, *пост-*, etc. (Vitas and Krstev, 2013). These derivatives are not considered in the Prolexbase, as explained in Section 2.3.

Possessive adjectives can be derived from some, mostly simple-word, organization names. For instance, *Мајкрософтов* is a possessive adjective derived from *Мајкрософт* (*Microsoft*). Possessive adjectives are also used for acronyms of organization names – in such cases, derivational suffixes are added to an acronym after a hyphen, e.g. *IBM-ов* (*belonging to IBM*).

3.4 Grammatical features

Proper names in Serbian, as well as nouns and adjectives derived from them, share inflectional properties with common nounds and adjectives.

The gender of geographic names, toponyms, oronyms and hydronyms, can be masculine, feminine or neuter while the names inflect in cases (seven different cases). They do not inflect in number which can be either singular or plural. The examples are given in Table 6.

Geographic names are, as a rule, inanimate although there are some confusing examples: there are a few city names in Serbia named after some famous people, for instance *Јаша Томић* and *Алекса Шантић*[13]. If they

[12] Ivo Lola Ribar (1916-1943), Yugoslav national hero.

[13] Jaša Tomić (1856 – 1922) was a politician, Aleksa Šantić (1868 – 1924) was a poet.

are considered inanimate, the sentence *I travel to Jaša Tomić* would be in Serbian *Путујем у Јаша Томић* that looks incorrect, since Јаша Томић, as a person is animate[14].

type	**original or English**	**name**	**number**	**gender**
toponym oronym hydronym	Belgrade Olympos Danube	Београд Олимп Дунав	singular	masculine
toponym oronym hydronym	Karlovci Alpes Dardanelles	Карловци Алпи Дарданели	plural	
toponym oronym hydronym	Athens Aconcagua Seine	Атина Аконкагва Сена	singular	feminine
toponym oronym hydronym	Budějovice Divčibare Plitvice	Будјеовице Дивчибаре Плитвице	plural	
toponym oronym hydronym	Valjevo Pohorje Oranjerivier	Ваљево Похорје Орање	singular	neuter
toponym	Kaštela	Каштела	plural	

Table 6. Geograhic names in Serbian with different gender and number

Demonyms derived from geographic names have masculine gender (for male inhabitants) and feminine gender (for female inhabitants) and they inflect in case and number. Adjectives derived from geographic names inflect

[14] In Serbian, the form of the accusative case singular for the masculine gender nouns depends on the animatness: for the inanimate nouns it is equal to the nominative case while for the animate nouns it is equal to the genitive case. In this example, the preposition *u* invokes the accusative case which for the (inanimate) *Јаша Томић* is the same as the nominative case, while for the (animate) *Јаша Томић* is *Јашу Томића* (for example in the sentence, *Милица се заљубила у Јашу Томића* (Milica fell in love with Jaša Tomić)).

in case, number, gender and animatness. It should be noted that possessive adjectives do not have comparative and superlative forms, and neither do descriptive adjectives except occasionally, e.g. *Војводина је најевропскији део Србије* (*Vojvodina is the most European part of Serbia*).

Serbian first names and nick names can have the masculine or the feminine gender, they are in singular and they inflect in case. Serbian surnames have masculine gender and they, in general, inflect both in number and in case. Some surnames, mostly of foreign origin, do not inflect in number because of morphological restrictions. Complex agreement rules apply to Serbian full names that depend on the gender of a first name and the order a first and a surname in a full name – one rule is that surnames do not inflect in case for female personal names (Gucul-Milojević, 2010). Women are sometimes referred by possessive adjectives of a surname in the feminine gender or by a feminine gender noun derived from a surname by gender motion. Some examples are given in Table 7.

Form	**Surname**	**Feminine forms**
nominative singular	*Петровић*	*Петровићка*
		Петровићева
genitive singular	*Петровића*	*Петровићке*
		Петровићеве
nominative plural	*Петровићи*	*Петровићке*
		Петровићеве

Form	**Full name (male)**	**Full name (female)**
nominative singular	*Петар Петровић*	*Зорка Петровић*
		Зорка Петровићка
		Зорка Петровићева
genitive singular	*Петра Петровића*	*Зорке Петровић*
		Зорке Петровићке
		Зорке Петровићеве

Table 7. Male and female personal names and their inflection

Organization names inflect in case while their number and gender do not change and, in general, depend on organization name form, if a single-word, or on the number and the gender of its head word, if a multiword. For instance, *Мајкрософт* (*Microsoft*) has the masculine gender, while *Сорбона* (*Sorbonne*) has the feminine gender. Among multiword organization names ***Универзитет*** *у Београду* (*University of Belgrade*) has the masculine gender, *Београдска аутобуска* **станица** (*Belgrade bus station*) has the feminine gender, while ***Удружење*** *спортских новинара Београда* (*Association of sport journalists of Belgrade*) has the neuter gender. Organization names ***Лекари*** *без граница* (*Doctors Without Borders*) and *Међународне мировне* **снаге** (*International peacekeeping forces*) have the plural number[15].

3.5 Dialects

In Serbian, two standard variants of pronunciation are in use, Ekavian and Ijekavian. They differ in the reflection of the old Proto-Slavic phoneme (*jat*): in Ekavian variant it is replaced predominantly by *e*, while in the Ijekavian variant its is replaced by syllables *ije/je*.

These variants do not have big influence on proper names, because most of proper names do not contain the reflection of the phoneme (*jat*). In cases when they do, the name is usually used in one of dialects only. For instance, in two city names *Ријека* (in Croatia) and *Ријека Црнојевића* (in Montenegro) the common noun is used only in Ijekavian dialect – *ријека* (and not *река*) (*river*). On the other hand, the feminine first name derived from the common noun *вера/вјера* (*faith*) – has both the Ekavian *Вера* and the Ijekavian variant *Вјера*. However, such a name in one variant would not change if it appears in a text written in another variant, that is, it is unchangeable.

In organization multiword names various common words appear that can be in either of variants. These variants are then reflected in organization names as well depending on the variant a text in which they appear uses, for instance, Ekavian variant *Светска банка* vs. Ijekavian variant *Свјетска банка* (*World Bank*).

4 Achived Results

4.1 The Serbian language contribution to the Prolexbase model

As we said in section 1, the inclusion of the Serbian language led to the development of a better Prolexbase model. The collaboration between research

[15] Head nouns in these multiword organization names are in bold.

groups from the University of Tours and the University of Belgrade was very fruitful in many respects, but we will emphasize two issues that we consider the most important: the derivation relation and the form representation.

The derivation relation. In Section 3.3 we presented the complexity of derivation rules of the Serbian language, such as the quasi-systematic possibility to create derivatives from human names, and thus from derivatives of topological names as well (the relational or inhabitant names). For instance, (see Figure 1), the city name *Београд* (*Belgrade*) generates (as in many other languages) a derivative *Београђанин* (a male inhabitant of *Belgrade*), while *Београђанин* generates in its turn *Београђанинов* (a possessive adjective of a male inhabitant of *Belgrade*). In English and French only one level of derivation exists: *Belgrade*/*Belgradian* in English and *Belgrade*/*Belgradois* in French. The first database model did not include relation from the table *Derivative* to itself. We added this relation to later models, and then we realized that this relation exists in French as well: for instance, the name of a prize, like Nobel prize, quasi-systematically allows the creation of a verb meaning *to give the prize*, for instance, *nobeliser*, while from such verbs it is possible to regularly create other derivatives, like *nobelisable* (a person who is likely to be chosen by the Nobel Prize Committee), and so on.

The form representation. In the Prolexbase database model, we place proper names in two tables, *Prolexeme* (the longest form of the name) and *Alias* (others forms). However, in the LMF representation (see Figure 2) this distinction disappears, because all aliases are equivalent entries, linked by the sense.

The question was whether a name written in the Cyrillic alphabet and the same name written in the Latin alphabet are aliases or not? However, it would look amazing that the same word can be alias of itself! Sure not. We considered first the possibility to define two prolexemes in Serbian language (Cyrillic and Latin), but we abandoned this idea since this solution violates the constraint of the uniqueness of the pivot projection in a particular language. For that reason we adopted a second solution that defines two lexicons, the Serbian Cyrillic lexicon and the Serbian Latin lexicon. Finally, the third version of Prolexbase was produced at the University of Tours in which we added systematically under the *Word Form* one or more *Form Representations*. For instance, for *Београд* (*Belgrade*) we now have:

```
<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>
```

After this choice was done, we added for some entries the distinction between the Ekavian and the Ijekavian dialect (see Section 3.5), for which we used the same LMF representation:

```
<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">
      Нјемачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ijekavsk">
      Њемачка
    </FormRepresentation>
    <FormRepresentation script="latn">
      Nemačka
    </FormRepresentation>
  </WordForm>
  ...
</LexicalEntry>
```

Finally, we used the concept of *form representation* to some variants of writing, as, for instance, in the example above, *Њемачка* and *Нјемачка* (the later rarely used), but also in some other cases, like the differences in transcription, as *Рио де Жанејро* and *Рио де Жанеиро* (*Rio de Janeiro*) (see Section 3.2), or for different forms for the same set of values of grammatical categories – the surname *Чехов* (*Chekhov*) has three variant singular dative forms: *Чехову*, *Чеховом* and *Чеховому*. We enhanced this approach to other languages as well, eliminating thus the alias category *Variant*.

4.2 The Prolexbase implementation

The Table 8 gives some numbers about the Serbian language implementation. We manually introduced the prolexemes with their link to the pivot from a selection of the French ones, and we also added a few aliases. Then we automatically generated the derivatives, and of course, all the inflections of the prolexemes, aliases and derivatives.

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

Table 8. Serbian language implementation

We can add to these numbers the amazing[16] number of the instances derived from *Београд*. If we complete the Figure 1 with all instances, we obtain 626 forms...

[16] Compared to English or even French!

Coming back to the example from Section 1: *Београдска жена ми је рекла да је Дунав прелеп* we now obtain:

београдска
  female inhabitant (derivative category)
  Belgrade (prolexeme)
    city (classifying context)
    Serbia (accessibility)
    capital (accessibility context)
→ The female inhabitant of the city of Belgrade, capital of Serbia
жена ми је рекла да је
→ has told me that
Дунав
  river (classifying context)
→ the Danube River
прелеп
→ is splendid

4.3 Conclusion

We showed that the complexity of the Serbian language morphology led to the essential contribution to the Prolex multilingual dictionary project. The necessary improvements that were introduced in order to accommodate Serbian proved to be useful for other languages in the Prolexbase. These improvements particularly concern the treatment of derivations and the representation of multiple forms. This was proved during our work on inclusion of non-European languages, such as Arabic, into the database since its internal structure was able to model them. This work also showed how important it is to include in linguistic multilingual project the variety of languages, not only the proximate ones.

Acknowledgment

The authors thank the University of Tours for enabling this research by funding the Prof. Cvetana Krstev one-month visit to the University. The authors also thank the *Egide Pavle Savic* fund that sponsored the project that initialized the collaboration between universities of Tours and Belgrade in 2004. This research was partly supported by the Ministry of Education, Science and Technological Development through projects 138006 and III47003.

References

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

Auer, S. and J. Lehmann. “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”. In *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”. In *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”, In *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”, In *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”. In *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”. In *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". In *7^{th} Workshop on Advanced Information Network and System, Kasetsart University, Bangkok, Thailand.* 2000

Maurel, D. "Prolexbase: A Multilingual relational Lexical Database of Proper Names". In *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". In *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". In *Proceedings of the 32^{nd} International Conference on Lexis and Grammar, September 10–14, 2013, Faro, Portugal.* 2013

Zakon, eds.. *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, ed.. *Правопис српскога језика [The Orthography of Serbian Language].* Матица српска, 1993

Стевановић, Михаило и др., eds.. *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