PROLEX: A LEXICAL MODEL FOR TRANSLATION OF PROPER NAMES APPLICATION TO FRENCH, SERBIAN AND BULGARIAN

Denis Maurel, Université François Rabelais Tours, France
Duško Vitas and Cvetana Krstev, University of Belgrade, Serbia
Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria

Résumé

Nous présentons ici le modèle lexical multilingue et relationnel du projet *Prolex*, pour le traitement automatique des noms propres, et son application à la traduction entre le français, le serbe et le bulgare. Ce modèle repose sur deux concepts principaux : le *nom propre conceptuel*, indépendant de la langue, qui représente un point de vue sur le référent ; et le *prolexème*, sa projection sur une langue donnée, qui correspond à un ensemble de lemmes (nom, alias et certains dérivés). Des relations et une typologie complètent la description.

Mots clefs

Nom propre ; TAL ; Multilingue ; Traduction ; Serbe ; Bulgare.

Abstract

We present in this paper the multilingual and relational lexical model developed in the framework of the *Prolex* project, for Proper Name processing, and its application to translation between French, Serbian and Bulgarian. This model is based on two main concepts: the *conceptual proper name*, language independent, representing the point of view about the referent; and the *prolexeme*, language dependent, namely its projection onto a given language that is a set of lemmas (name, aliases and some derivatives). Relations and typology complete the description.

Key-words

Proper name; NLP; multilingual; translation; Serbian; Bulgarian.

1. Motivation

In this paper, we deal with some problems of the interrelation (translation, transliteration, transcription) of Proper Names between French and two Slavic languages, Serbian and Bulgarian. In French linguistic studies (as well as in Serbian and Bulgarian), the Proper Names are not usually described in details regarding both their inflexion and derivation properties. On the other hand, the correct translation of Proper Names is not a trivial issue and the example of Slavic languages and French is very rewarding in this domain. Just like common nouns, Slavic Proper Names undergo declension but their inflectional patterns are more complicated. Rich inflection, derivation and a specific word order (shown in Serbian Personal names) make processing of Proper Names in Slavic languages more difficult than for other languages. There are at least four points that illustrate the main difficulties in the correspondence between French and Slavic Proper Names.

1. First, there are transcription problems between the two alphabets, Latin and Cyrillic: The Cyrillic alphabet is used for Serbian and Bulgarian, the Latin alphabet, for Serbian and French.

For instance, the surname of the Japanese Prime Minister, *Shinzo Abe*[1], is written *Šinzo Abe* and *Шинзо Абе* in Serbian and *Шиндзо Абе* in Bulgarian. The official Japanese transcription of this surname in Latin is *Abe Shinzō*, but the diacritic *ō* does exist neither in French and English, nor in Serbian and Bulgarian and it is omitted.

The name of the Russian President, *Vladimir Putin*, is written *Vladimir Putin* and *Владимир Путин* in Serbian, *Владимир Путин* in Bulgarian and *Vladimir Poutine* in French (the French transcription is not the same as the English one!). Note that the official Russian surname is written with diacritics, Влади́мир Пу́тин, and often omitted.

[1] To illustrate our motivation, we use names of the heads of the states and states from the G8 in June 2007.

2. Second, Bulgarian and Serbian are South Slavic languages with rich nominal inflection: nouns and adjectives are inflected for case, number and gender.

The attributes of the Proper Names lemma and their realized values define the paradigm in the respective languages, i.e. if a Proper Name is Human, it might have vocative form in Bulgarian, if a proper Name is Singularia tantum, it does not have plural forms, etc. Serbian and Bulgarian Proper Names are divided into grammatical subclasses with respect to their Number (Singularia tantum, Pluralia tantum), Gender, Definiteness (only for Bulgarian) and Animateness. The category Gender with Bulgarian Proper Names is a lexical-semantic category (an exception are Family names), which means that a given noun does not possess different word forms expressing masculine, feminine and neuter, although the noun lemmas can be grammatically classified into the three classes: *Даниел* (*Daniel*, masculine), *София* (*Sofia*, feminine) and *Черно море* (*Black see*, neuter). The Definiteness is also a lemma property regarding the Bulgarian Proper Names which are single words. In Serbian the difference between the grammatical and natural gender and number is marked. The Proper Names lemmas in Bulgarian, Serbian and French are characterized by the following categories (Figure 1).

	Serbian	Bulgarian	French
Gender	masculine, feminine, neuter	masculine, feminine, neuter	masculine, feminine
Number	singular, plural	singularia tantum, pluralia tantum	singular, plural
Definiteness		definite, indefinite	
Animateness	human, non human	human, non human	

Figure 1: Lemma attributes of Proper Names

Slavic languages as Serbian (with an exception of Bulgarian and Macedonian) are highly inflected languages regarding to the category of noun case. Thus not only the nominative form of a Proper Name, but also its genitive, accusative, etc. has to be taken into consideration. Seven cases (nominative, genitive, dative, accusative, vocative, instrumental and locative) are used in Serbian. Bulgarian Proper Names may also express the category of case but with the values of nominative and vocative only (some other restriction are shown concerning the vocative inflection of foreign names). Figure 2 below presents all instances of the name *Vladimir* in Serbian, Bulgarian and French

<table>
<tr><td></td><td colspan="2">Serbian</td><td>Bulgarian</td><td>French</td></tr>
<tr><td>Nominative</td><td>Vladimir</td><td>Владимир</td><td rowspan="7">Владимир</td><td rowspan="8">Vladimir</td></tr>
<tr><td>Genitive</td><td>Vladimira</td><td>Владимира</td></tr>
<tr><td>Dative</td><td>Vladimiru</td><td>Владимиру</td></tr>
<tr><td>Accusative</td><td>Vladimira</td><td>Владимира</td></tr>
<tr><td>Instrumental</td><td>Vladimirom</td><td>Владимиром</td></tr>
<tr><td>Locative</td><td>Vladimiru</td><td>Владимиру</td></tr>
<tr><td>Vocative</td><td>Vladimire</td><td>Владимире</td><td>Владимире</td></tr>
</table>

Figure 2: The case instances of the name *Vladimir*

The category Gender is an inflectional category in Serbian: for instance *papa* (pope) is masculine, but its plural form *pape* is feminine. Besides two main categories for number, singular and plural, Serbian nouns also have the so called "paucal" form which represents a synthetic category of number and gender that is used with small numbers (two, three

four): *jedan lepi Zec* (one pretty Zec[2]), *dva lepa Zeca* (two pretty Zec), *pet lepih Zečeva* (five pretty Zec). Animateness is an inflectional category for masculine gender nouns in Serbian; the form of the accusative case is equal to the genitive case for the animate nouns and to the nominative case for the inanimate nouns. In Bulgarian, the category Gender operates at the word form level only for family names, the categories Number and Definiteness - only for Proper names that are Multiword units.

3. Third, Serbian and Bulgarian (but not all other Slavic languages) have a very productive derivative morphology, in particular for nouns with a Human feature, and then for Proper Names (Persons, but also Organizations, and Toponyms with the inhabitants names).

Slavic languages (such as Bulgarian and Serbian) show similar tendencies towards the derivation of Proper Names. A relational noun and a relational adjective from the surname *Putin*, both in French and Slavic languages can be created:

Relational noun: *Poutinien* in French, *putinovac*, masc. and *putinovka*, fem. and путиновац, masc. and путоновка, fem. in Serbian, ?*путинец*[3] masc. and ?*путинка* fem. in Bulgarian.

Relative adjective: *poutinien* in French, *putinski* and путински in Serbian, ?*путинов*, ?*путински* in Bulgarian.

However, there is a possessive adjective in Serbian and Bulgarian, but not in French: *Putinov* and *Путинов* in Serbian, *Путинов* in Bulgarian. Two more derivatives exist in Serbian, *putinovčev*, a possessive adjective built from the relational noun *putinovac* (something belongs to a supporter of Putin), or the relative adjective built from the same noun, *putinovački*, (in the way of a supporter of Putin). Of course, these nouns and adjectives have different instances for the categories of case, gender, number and definiteness (see Figure 1). Thus the relative adjective *putinien* has four instances in French, but sixty-three instances in Serbian and nine instances in Bulgarian.

In Bulgarian collective nouns can be built by Person names; for instance: *Даниел даниеловци* (all persons with the name *Daniel*) *Дон Кихот донкихотовци* (all persons that seem to *Don Quixote*, masculine), *Мария марии* (all persons with the name *Maria*, feminine), *Мими мимета* (all persons with the name *Mimi*, neuter).

The word derivation is widely observed with toponyms, which can also express the Human feature by means of the metonymy (*Russia refuses to accept American missile in Czech*). Only two French derivatives from the name *Russie* (*Russe* and *russe*) can be built, comparing to at least ten possible forms in Serbian and seven in Bulgarian (see Figure 3), counting only the results of regular derivation [Vitas, Krstev 2005].

[2] Familly name *Zec* means rabbit.

[3] The question mark indicates the optional and understandable forms that are not in common use in Bulgarian.

	Serbian		Bulgarian	French
Name	*Rusija*	Русија	Русия	*Russie*
Relative adjective (from the Name)	*ruski*	руски	руски	*Russe*
Possessive adjective (from the Name as human)	*Rusijin*	Русијин	руски	
Relational noun masc. (Male inhabitant)	*Rus*	Рус	руснак	*Russe*
Diminutive from the relational noun	*Ruščе*[4] *Ruščić*	Рушче Рушчић	русначе	
Relational noun fem. (Female inhabitant)	*Ruskinja*	Рускиња	рускиня	
Diminutive from the relational noun	*Ruskinjica*	Рускињица	рускинче	
Collective noun	*Rusi*	Руси	руси	
Possessive adjective (from an inhabitant)	*Rusov* *Ruskinjin*	Русов Рускињин	-	

Figure 3: The derivatives of the toponym *Russia*

A regular tendency both in Serbian and Bulgarian is that derivatives from Multiword units always are single graphical words with complex structure.

4. Fourth, Proper Names (but also a lot of common nouns) are often Multiword units that need specific morphological treatment.

For instance, the Multiword unit *Vladimir Putin* does not have the same morphological treatment in Serbian if the first name and the surname are permutated. While the two parts of the name in nominative are the same, *Vladimir Putin* and *Putin Vladimir*, the genitive is *Vladimira Putina* and *Putin Vladimira* and so on.

It should be pointed out that the Bulgarian and Serbian multiword proper names have their own inflective rules. The part of speech of the head word determines the clustering into grammatical classes; the significant grammatical categories inherent to the lemma of the head word (such as gender for nouns), the number and part of speech of the remaining constituents and the options for inserting some words (such as particles) all show the grouping of Multiword unit grammatical subclasses. For instance, the Bulgarian proper name *Руска федерация* (Russian Federation) is a noun phrase; the members of its paradigm are determined by the head feminine noun: Руска федерация (singular, indefinite), Руската федерация (singular, definite), ?Руски федерации (plural, indefinite), ?Руските федерации (plural, definite).

The inflection type is determined by the inflectional alternations of each member (the adjective and the noun), there are agreement dependencies between adjective and head noun and no other word interventions or word order changes are allowed[5].

The Proper Names constitute a significant part of natural language texts. There are several reports claiming that the Proper Names estimate to about 10% in the newspaper articles [Coates-Stephens 1993]. The statistics over the Bulgarian National Radio corpus show that there are 32,669 Bulgarian Proper Names detected at the total amount of 325,788 words which represents 10.02%. It can be expected that the proportion of the Proper Names will be also relatively high in some thematic domains like history, geography, politics, sport, etc. The similar statistics over the corpus containing

[4] This form denotes rather a child who is a Russian.

[5] To inflect these Multiword units, we use the Multiflex system presented in another paper in the same number of this review (paper by Agata Savary, Cvetana Krstev and Duško Vitas).

Bulgarian political texts show 199,204 Proper Names among 2,224,789 words which is equal to 8.95%. These calculations were made by means of a tokenizer, thus the Multiword units (including Proper Names) are not taken into account. Another statistical data item shows that there are 12,636 Multiword literals out of 51,584 in Bulgarian wordnet (24.49%) and 45,769 compound literals out of 203,147 existing (22.5 %) in Princeton wordnet. The proportion in other European wordnets is similar. Considering the fact that the Multiword units are also a significant part of the human lexis, we can conclude that the relative percentage of Proper Names is more than one tenth.

2. The Prolex project

In order to process Proper Names for NLP applications, the *Prolex project* was initiated in 1990s with the study of toponyms in French [Piton, Maurel, 2000]. This work has been pursued by development of a Serbian version. Finally, a relational multilingual dictionary of Proper Names, *Prolexbase*, in the form of a relational database, was designed and constructed [Krstev et al. 2005] [Tran, Maurel 2006].

This model is based on two main concepts: the *pivot* (that represents the *conceptual proper name*) at language independent level and the *prolexeme* (the projection of the pivot onto particular language), that is a set of lemmas that includes the name, but also its aliases and some of its derivatives.

The definition of these concepts is a representation of the variations of proper name. This variation may be conceptual (and, then, independent of language) or linked to a particular language by morphology or knowledge. These variations are very important for NLP, because the same proper name can be written in different instances, sometimes in different parts of speech, and also, it can be replaced by another proper name, linked by a semantic relation (an anaphora).

We use in our description of semantic relations the four diasystematic features of Coseriu [Coseriu 1998], defined at Figure 4 .

Diachronic	variety to time
Diatopic	variety to area
Diastratic	variety to sociocultural stratification
Diaphasic	variety to usage purpose

Figure 4: The diasystem of Coseriu

When the variation is language independent, it defines a specific point of view about the referent, a concept that we call the *conceptual proper name*. Being language dependent, the variation may be:

1. Cultural (for instance, by a specific knowledge not shared by foreign countries).
2. Based on the written form (variation in orthography, abbreviated forms, acronyms, alphabet, etc.).
3. Obtained by a derivation with a particular and well-known sense, that always refers to the name (see [Gross, 1997]).

We call the set of these variations the *prolexeme*.

The language independent level with its three semantic relations and specific features (typology and existence) is presented below, followed by the language dependent level with the prolexeme and its property.

2.1. The language independent level

We simply use a pivot (a unique number) to represent a specific point of view about the referent, i.e. a *conceptual proper name*. This representation by pivot is common in many lexical databases (*EuroWordnet* [Vossen, 1998] and *Balkanet* [Tufiş et al., 2004], *Papillon* [Mangeot-Lerebours et al., 2003]…).

2.1.1. The relation of synonymy

The relation of synonymy, when it is language independent, concerns variation of a name from the diachronic, diastratic or diaphasic point of view:

1. Diachronic: a new name is sometimes implied by the history of the respective country. This is the case for many toponyms, due to the Communist period in Eastern Europe, for instance, *Petersburg*, *Petrograd* and *Leningrad* in Russia; or due to a change in the political system, for instance, *Zaire* and *Democratic Republic of the Congo*.

2. Diastratic: a name is well-known because of its fame. Some years ago, many people knew the religious name of the pope, *John Paul II*, but only a few knew his surname, *Karol Jozef Wojtyla*. Other examples are the pseudonym that artists frequently use. The cartoonist *Georges Rémi*, the creator of *Tintin*, is well-known with the pseudonym *Hergé*, but not with his surname.

3. Diaphasic: for instance, in an official register, a synonym of a country name includes the system of government, such as *Republic of Bulgaria*, versus *Bulgaria*; in emphatic register (news, tourism, etc.) the *Town of Light* may be used instead of *Paris.*

Sofia, the name of the Bulgarian capital is encountered 485 times in 312 files taken from the Bulgarian National Radio corpus. Although the corresponding anaphora expression (*the Bulgarian capital*) occurs only once, the two named entities express equal sense.

2.1.2. The relation of meronymy

The relation of meronymy is well-known in terminological contexts. It is natural to use it for names to describe inclusion of toponyms or events. *Serbia* and *Bulgaria* are in the *Balkans*, that are in *Europe*; the *Normandy landings* is a particular event of the *Second World War*.

This notion can be extended to other contexts, such as *EADS* $\subset$ *Europe, St Matthew's Gospel* $\subset$ *New Testament*, *Novak Djokovic* $\subset$ *Serbia*, etc. The meronymy relation is frequently used in economical registers, for instance *the European firm* from *EADS*, or in sport register, for instance *the Serbian tennis man* from *Novak Djokovic*.

2.1.3. The relation of accessibility

In explanatory dictionaries, Proper Names do not have definitions, in contrast to common nouns, but usually some relations towards different names, generally better known, are given. For instance, the name *Aaron* is situated with the name of *Moses* (*Aaron* is presented as the brother of *Moses*). If we search for *Moses* in the dictionary, we might not have the symmetrical information (*Moses* is the brother of *Aaron*), but rather *Moses* will be represented as the chief of *Hebrews*.

There are many possible relations and we do not have the aim to model them in our project. So we adopted a unique relation for all of them, that we call accessibility [Ariel 1990]. For instance, we will say that the name *Aaron* is accessible from the name *Moses* that is accessible from *Hebrews' story*, etc. However, we precise large registers as relative (*Aaron* and *Moses*), capital (*Paris* is the capital of *France*), politician (*Angela Merkel* is a *German* politician), founder (*Henry Dunant* has founded the *Red Cross*), follower (*Peter* is a disciple of *Jesus*), creator (*The Magic Flute* is an opera of *Wolfgang Amadeus Mozart*).

2.1.4. The existence

The pivot is specified by a feature of existence. Each pivot is linked to one and only one *existence*. This feature often is important information about translation.

The existence is just made up of three features:

1. Historical: Most of the proper names refer to historical period; we know for certain that they have existed.

2. Fictitious: Proper names are also used by the authors of novels, story, play, film, etc.

3. Religious: This third feature depends of the belief of people. If *Jesus* and *Mohammed* are historical proper names, it is not the role of linguist to say if the archangel *Gabriel* really exists or not…

Generally, the names linked to the features *Fictitious* or *Religious* are translated and not the names linked to the feature *Historical*[6]. For instance *Snow White* is translated in French (*Blanche-Neige*), Serbian (*Snežana* and Снежана) and Bulgarian (Снежанка).

2.1.5. The typology

To again help translation, we use a typology of proper names, inspired by different onomastic, economic or NLP typologies, compiled by Grass [Grass, 2000]. As is done for the existence, each pivot is linked to one and only one type. As is done for the accessibility relation, we have chosen to define only a few types (exactly thirty), obviously general; but we have completed this first level by a hyperonymy of supertypes.

Of course, the whole entries of the database have the feature *Proper name* that is the hyperonym of all other types or supertypes. This supertype is divided into the four classical linguistic features:

1. Human feature: The supertype *Anthroponym* is divided also into individual and collective anthroponyms. Individual anthroponyms concern persons (*Celebrity*, *First Name*, *Patronymic*), but also names of animal (*Laika*, the first dog of the space) or machine (*HAL*, the robot of *2001: A Space Odyssey*, the film directed by Stanley Kubrick), that are linked to the type *Pseudo-anthroponym*. Collective anthroponyms concern *Dynasty*, *Ethnonym* or *Group*, a supertype that specifies the different organizations, economic (*Firm*), politic (*Institution*), religious or associative (*Association*), cultural (*Ensemble*) and international (*Organization*).

2. Location: The supertype *Toponym* concerns natural areas (*Astronym*, *Geonym* and *Hydronym*) as well as man-made ones (*Building*, *City* and *Way*) and also human areas (*Territory*, shared between three types: *Country* for an independent country, *Region* for a region in a country and *Supra-national* for a region spanning countries).

3. Inanimate: The supertype *Ergonym* concerns different human fabrications (except toponyms). We naturally have brands or products (*Product*), novels, sculptures, paintings, films, operas… (*Work*), but also intellectual constructions (*Thought*); we have added names of *Vessel* (the *Pinta,* one of the three ships used by *Christopher Columbus* in his first voyage) and some rare names of *Object*, often mythical (the *Grail*).

4. Event: the supertype *Pragmonym* concerns historical periods or events (*History*), but also cultural (*Event*), as the *Football World Cup,* or religious ones (*Feast*)… And, meteorological phenomena (*Meteorology*), as winds, or, sadly, disasters (*Disaster*), as *Chernobyl disaster*.

The complete typology is presented at Figure 5.

[6] In fact, it is not just as straightforward, because the translation of the names depends also on their type…

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

Figure 5 : The Prolex typology

We have visited the Wikipedia site on June, 13, 2007, with the theme *French revolution*; Figure 6 presents some type examples in our three languages.

Type	*French*	*Serbian*	*Bulgarian*
Celebrity	Louis XVI	Luj XVI, Луј XVI	Луи XVI
	Maximilien Marie Isidore de Robespierre	Maksimilijan Robespjer, Максимилијан Робеспјер	Максимилиан Мари Изидор де Робеспиер
Association	Jacobins, pl.	Jakobinac, Јакобинац, sg.	Якобинци, pl.
City	Paris	Pariz, Париз	Париж
	Versailles	Versaj, Версај	Версай
Building	Tuileries	Tiljerije, Тиљерије	Тюйлери
Country	France	Francuske, Француска	Франция
Supra-national	Europe	Evropa, Европа	Европа
Work	Déclaration des Droits de l'Homme et du Citoyen	Deklaracija o pravima čoveka i građanina, Декларација о правима човека и грађанина	Декларация "Правата на човека и гражданина"
History	Prise de la Bastille	Pad Bastilje, Пад Бастиље	Щурмуването на Бастилията

Figure 6: Some type examples from Wikipedia

We can add to this strict hyperonymy a secondary one (Figure 7). For instance, a name of a territory or a city can be used as human, a building or a way are human fabrications, etc.

Types	*Secondary hyperonym*
Territory	Collective anthroponym
City	Collective anthroponym Ergonym
Building Way	Ergonym
Event Feast History	
Group	Ergonym Toponym
Vessel	Collective anthroponym Toponym

Figure 7: The secondary hyperonymy

2.1.6. An ontology of proper names

Finally, Figure 8 presents the language independent level as ontology of proper names (*Hyperonymy1* is the links between types, *Hyperonymy2*, the type of a prolexeme, *Hyperonymy3*, the existence and *Hyperonymy4*, the secondary hyperonymy).

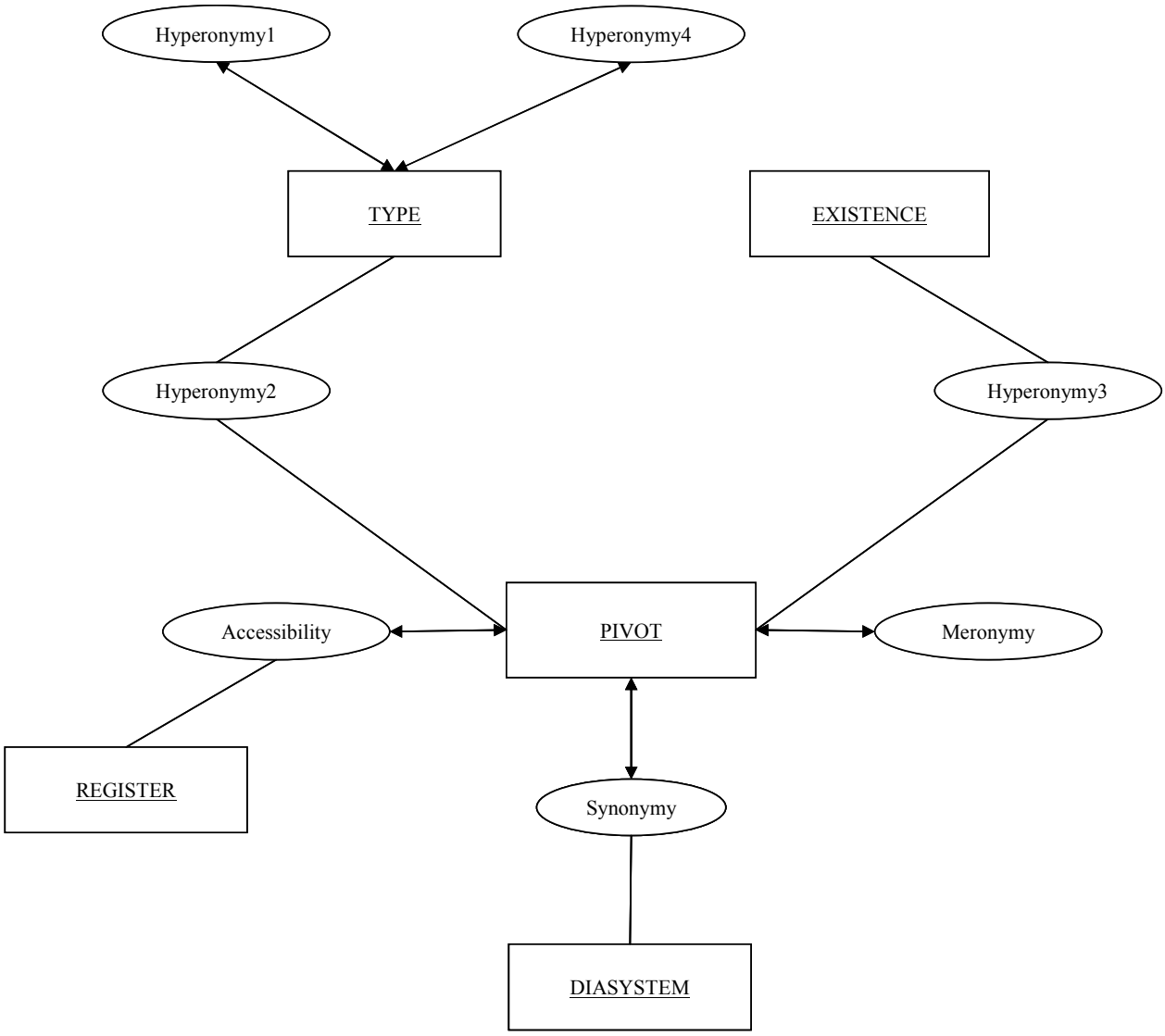

Figure 8: An ontology of proper names

2.2 The language dependent level

The prolexeme is the set of all lemmas semantically linked to a proper name in the observed language. For instance (see Figure 3), the pivot of *Russia* is 45161 and the prolexemes in French, Serbian and Bulgarian are:

French: {*Russie*, *Russe*, *russe*...}

Serbian: {*Rusija*, *ruski*, *Rusijin*, *Rus*, *Rušče*, *Ruskinja*, *Ruskinjica*, *Rusi*, *Rusov*, *Ruskinjin*...}

Bulgarian: {Русия, руски, руски, руснак, русначе, рускиня, рускинче, руси…}

To simplify our database, by misuse of language, we have not implemented the term *French prolexeme* by the pair (45161, fr), but by the name *Russie*. So, we have three tables of lemmas in *Prolexbase*:

1. Prolexemes: Arbitrarily, the longest form of names.

2. Alias: Other forms (short forms, abbreviations, acronyms, different orthographies, other transcriptions…), but also diatopic synonymies and some diastratic ones (that are too dependent of the language to have a pivot).

3. Derivatives: We only add to the database the derivatives that are semantically linked to the Proper Name (*to pasteurize* is a derivative of the name *Pasteur*, but it is a lexicalized word, with a specific definition, independent of the name *Pasteur*).

We note also at this level the classifying context (*capital*, *king*, *coach*…) which is often useful for translation.

We add to the database the relation of eponymy: antonomasia (*this politician is a Cicero* - i.e. is a good orator), terminological terms (*Alzheimer's disease*, *Pythagoras' theorem*...) or idiomatic phrases (*I don't know him from Adam*…). This relation, in the opposite to the other ones, informs that translation does not refer to proper name but to common noun, terminology or idiom.

At this level, each lemma is linked to an inflectional paradigm. A specific tool generates all its instances by use of finite-state transducers. This tool is based on the Unitex software [Paumier, 2003] and the Multiflex system. Figure 9 presents the example of *United States of America*.

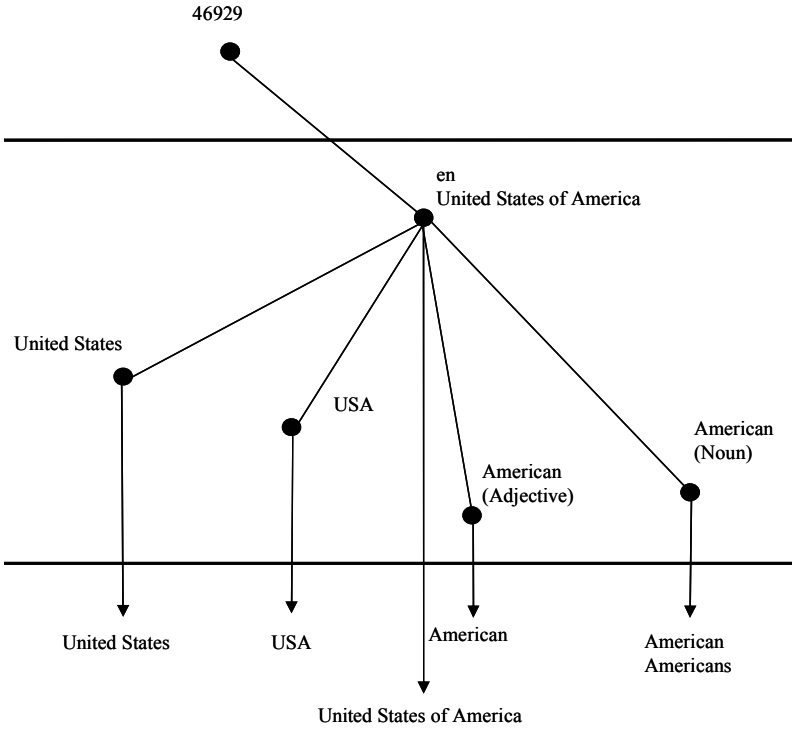

Figure 9: Pivot, prolexeme, aliases, derivatives and instances from *United States of America*

2.3. The inter lingual links

The pivot represents a point of view about a referent. It is linked, in one language, with one, and only one, prolexeme. This prolexeme is linked to all the instances of the name, its aliases or derivatives. This description of languages allows inter lingual links that are not word to word links, but prolexeme to prolexeme links.

For instance, *the car of a supporter of Vladimir Putine* is translated in French by *la voiture d'un Poutinien*, in Serbian by *putinovčev auto* and in Bulgarian by the same phrase as in English (колата на поддръжника на Владимир Путин).

3. Corpus examples

We now present some results from the aligned version of Jules Verne's novel *Le tour du monde en quatre-vingts jours*. This novel has been translated (and recently sentence to sentence aligned) in many Slavic languages. The occurrences of the proper name *Passepartout* (Figure 10) and the toponym *Angleterre* (Figure 11) in Verne's novel are used in order to illustrate consequences and differences between the three languages.

	POS	*French*	*Serbian*	*Bulgarian*
Passepartout	Name	437 times *Passepartout* (one form)	430 times *Paspartu* 366 *Paspartua* 37 *Paspartuom* 3 *Paspartuu* 20	438 times Паспарту (one form)
	Possessive adjective		9 times *Paspartuov* 4 *Paspartuova* 1 *Paspartuovih* 1 *Paspartuovim* 1 *Paspartuovo* 1 *Paspartuovu* 1	

Figure 10: The name *Passepartout* in the Verne's novel

	POS	*French*	*Serbian*	*Bulgarian*
England	Name	37 times *Angleterre*	39 times *Engleska* 9 *Engleske* 8 *Engleskoj* 15 *Englesku* 7	37 times Англия
	Relational adjective	59 times *anglais* 30 *anglaise* 23 *anglaises* 6	58 times *engleska* 6 *engleske* 8 *engleski* 15 *engleskih* 5 *engleskim* 5 *englesko* 2 *engleskog* 8 *engleskoga* 2 *engleskoj* 2 *engleskom* 4 *englesku* 1	62 times английски 14 английския 6 английският 1 английска 7 английската 9 английско 7 английското 5 английски 7 английските 6
	Relational noun	19 times *Anglais* +1 in English *Englishman*	21 times *Englez* 6 *Engleza* 10 *Englezi* 3 *Englezu* 2	20 times англичанин 7 англичанинът 1 англичани 8 англичаните 4
Great Britain (meronymy)	Name	2 times *Grande Bretagne*	2 times *Velike Britanije*	5 times Великобритания
	Relational adjective	4 times *britanique*	6 times *britanska* 2 *britanskog* 1 *britansko* 1 *britanskoj* 1 *britanskom* 1	2 times британски 1 британските 1
United Kingdom (synonymy)	Name	8 times *Royaume-Uni*	8 times *Ujedinjenog Kraljevstva* 5 *Ujedinjenih Kraljevstva* 1 (plural) *Ujedinjeno Kraljevstvo* 1 *Ujedinjenom Kraljevstvu* 1	1 time Обединено кралство

Figure 11: The name *England* in the Verne's novel

For instance, we find *engleska prestonica* from *capitale de l'Angleterre* (*capital of England*) and *francuske i engleske pesmice* from *refrains de France et d'Angleterre* (*chorus from France and England*). The name England is also a part of the multiword proper name *Bank of England* with *Engleska banka* from *Banque d'Angleterre* and *English Indies* with *Engleska Indija* from *Inde anglaise* (see [Maurel, 2004]). We find also terminology (*royal British sauce*).

4. Conclusion

We have shown in this paper that the *Prolex* model is well adapted to translation of proper names, particularly between French and Slavic languages, as Serbian or Bulgarian, due to the importance of morphology (different cases, but also derivatives, etc.). The existing relations between proper names have to be considered also and sometimes the translator use is to replace a name by another one.

The French database is available at the url: http://www.cnrtl.fr/lexiques/prolex/. This model can be used also for Information retrieval, particularly in Slavic language, as is done at the url: http://hlt.rgf.bg.ac.yu/WS4QE/Default.aspx.

References

ARIEL M. (1990), *Memory and context for language interpretation*, Cambridge University Press.

COATES-STEPHENS S. (1993), *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*, Kluwer Academic Publishers, Hingham, MA.

COSERIU E. (1998), Le double problème des unités dia-s, in : *Les Cahiers δια. Etudes sur la diachronie et la variation linguistique* 1:9-16, Université de Gent, Belgique.

GRASS T. (2000), Typologie et traductibilité des noms propres de l'allemand vers le français, in : *TAL*, 41-3, 643-669, Hermès-Lavoisier, Paris.

GROSS, M. (1997) Synonymie, morphologie dérivationnelle et transformations, in : *Langages* 128, Paris, Larousse, p. 72-90.

KRSTEV S., VITAS D., MAUREL D., TRAN M. (2005), Multilingual Ontology of Proper Names, Second Language & Technology Conference, 116-119, Poznań, Poland, 21-23 avril.

MANGEOT-LEREBOURS M., SÉRASSET G., LAFOURCADE M. (2003), Construction collaborative d'une base lexicale multilingue, le projet Papillon,in : TAL, 44-2:151-176, Hermès-Lavoisier, Paris.

MAUREL D. (2004), Les mots inconnus sont-ils des noms propres ?, *Septièmes Journées internationales d'Analyse statistique des Données Textuelles* (*JADT 2004*), Louvain-la-Neuve, Belgique, 10-12 mars, 776-784.

PAUMIER S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

PITON O., MAUREL D. (2000), Beijing frowns and Washington takes notice: Computer Processing of Relations between Geographical Proper Names in Foreign Affairs, *Fourth International Workshop on Applications of Natural Language to Data Bases (NLDB'00),* Versailles, 28-30 juin (Actes p. 66-78).

TUFIŞ D., CRISTEA D., STAMOU S. (2004), BalkaNet: Aims, Methods, Results and Perspectives. A General Overview, *Romanian journal of Information science and technology*, Vol. 7, n°1-2:9-44, Romanian Academy, Bucharest, Romania.

TRAN M., MAUREL D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, Traitement automatique des langues, Vol. 47-3, (à paraître), publication électronique (http://www.atala.org).

VITAS D., KRSTEV C. (2005), Regular derivation and synonymy in an e-dictionary of Serbian, in: *Archives of Control Sciences*, 15-3:469-480, Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland

VOSSEN P. (1998), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.