LABORATOIRE d'INFORMATIQUE

(UPRES EA n° 2101)

Multilingual Relational Database of Proper Names:

Prolexbase

Documentation

Malgorzata SPEDZIA
Denis MAUREL
Agata SAVARY

Ecole Polytechnique de l'Université de Tours
Département Informatique
64 Av. Jean Portalis
37200 TOURS - FRANCE
Tél. : 02 47.36.14.14. Fax : 02 47.36.14.22.

e-mail : denis.maurel@univ-tours.fr

Avril 2011
Rapport Interne n°297x36p.

Table of contents

1 Introduction ... 1

1.1 Motivations ... 1

1.2 Proper names ... 2

2 Ontology ... 3

2.1 Prolexbase ontology ... 3

2.2 The metaconceptual level ... 4

2.2.1 The existence ... 9

2.3 The conceptual level ... 9

2.3.1 Synonymy ... 10

2.3.2 Meronymy ... 11

2.3.3 Accessibility ... 11

2.3.4 Example of the relations between pivots ... 12

2.4 Linguistic level ... 12

2.4.1 Prolexeme ... 12

2.4.2 Aliases ... 13

2.4.3 Derivatives ... 13

2.4.4 Language dependent relations ... 14

2.5 Instances level ... 16

2.6 Summary of the main points ... 19

2.7 Example ... 20

3 Database ... 21

Metaconceptual level ... 24

Conceptual level ... 24

Linguistic level ... 25

Instances level ... 25

4 Prolexbase and LMF ... 25

4.1 LMF – basic information ... 25

4.1.1 Core package and extensions ... 25

4.1.2 DCR ... 27

4.2 ProlexLMF ... 28

4.2.1 Instances level ... 29

4.2.2 Linguistic level ... 31

4.2.3 Language independent part of ProlexLMF ... 32

4.2.4 Outline of Prolexbase adaptations to LMF 33
5 Summary 34
6 Bibliography of the project Prolex 36
7 References 37

1 Introduction

1.1 Motivations

Natural language processing (NLP) depends largely on the language resources which are adapted to the envisaged applications and used methods. The great majority of applications require a word list furnished with morphological, semantic and syntaxic information, in other words they need an electronic dictionary. Its size varies according to the model of language the system uses: one that combines the stochastic and linguistic information, or one that relies on the linguistic information only. Up to the present day, the scientific community of NLP has concentrated on dictionary resources of common nouns (e.g. System DELA [Courtois, Silberztein, 1990]) as well as on special resources of terminological terms [Sager, 1990]. The problem is that a particular category of nouns, to be more precise: proper names, is scarcely ever present in these dictionaries. And yet not only do proper names constitute a significant part of many texts (they represent more than 10% of all running words in journalistic texts [Coates-Stephens, 1993]) but they can be distinguished by the richness of specific semantic information as well.

From the MUC Conferences and its Named Entity Task, proper names are a challenge for NLP applications. Proper names as well as dates, percentage and monetary amounts constitute together the category of Named Entities [Chinchor, 1997]:

"On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts".

Tasks such as spelling or translation aid, multilingual alignment, lexical anaphora resolution or named entity recognition require an exhaustive description of the linguistic properties of proper names. In other words, they ought to be well defined, categorized and, if possible, linked. This precise description is prominent for two reasons. First of all, different heuristic procedures reduce the extraction of proper names to more or less successful approximations. One of them is an approximation using the rule with the minimal list of proper names since "through a judicious use of internal and external evidence relatively small gazetteers are sufficient to give good Precision and Recall" [Mikheev et al. 1999]. Another one is an approximation that relies on the orthographic practice that proper names are written using the initial capital letters. Taking it into account, a very simple criterion can be deducted that "a proper name is any unknown name that starts with a capital letter". However, this criterion is correct in no more than 50% of cases due to the homography and compounds [Maurel, 2004]. Secondly, proper names have a special linguistic status since they cannot be subjected to the usual defining methods. Because of their rich semantic information, their systematic, exhaustive and explicit description is not the easiest task.

This description becomes even more complex in the environment of multilingual applications. In spite of the fact that proper names represent cognates par excellence, in many cases their graphic variations disable their recognition by the approximative string pattern matching. Let's take the example of the previous pope's name:

English: *John Paul II*
Polish: *Jan Paweł II*

French: *Jean Paul II*
Italian: *Giovanni Paolo II*
Croatian: *Ivan Pavao II*

Even in cases when it is possible to spot an occurrence of a proper name in a monolingual text with simple methods, the identification of the concept that is represented by that name in a multilingual text is almost impossible. It must not be forgotten that proper names share the morphological properties of the language in which they are realized. In addition, the proper name is not always written with the capital letter. It can be illustrated by the example with possessive and relational adjectives:

English: *the Chopin tradition* or *Chopinian tradition* or *Chopin's tradition*
Polish: *tradycja szopenowska*
French: *la tradition chopinienne*
Serbian: *šopenovska tradicija*

The description of proper names in a multilingual application poses a problem. Owing to the complexity of semantic relations that connect proper names, it is impossible to reduce their description to the construction of an ordinary multilingual electronic dictionary. It seems that in a multilingual context it is more appropriate to represent proper names as ontology. The *Prolex project* was initiated in 1990s with a relatively simple goal of producing a database of French inhabitants names and toponyms, with some linguistic information for NLP. Today, the main motivation of the Prolex is to develop a multilingual dictionary of proper names and their relationships. This development was supported by a RNTL-Technolangue project of the French Ministry of Industry in collaboration with two companies, *Systran* (leading provider of the world's most scalable and modular translation architecture) and *Exalead* (editor of Frenchsoftware specialist in web navigation). It aimed to create a multilingual database of proper names, *Prolexbase*, with some linguistic information for natural language processing: machine translation, computer aided translation, information retrieval and spelling dictionaries. From June 2007, this resource is free and available on the CNRS resource website[1] (CNRTL) in XML format [Maurel, 2008].

1.2 Proper names

It is not easy to define precisely a proper name and linguists are not unanimous. One of the first definitions is provided by J.S Mill:

"A proper name signifies nothing but the individual whose name it is; and when we apply it to the individual, we neither affirm nor deny anything concerning him".

For him, no semantic link exists between the proper name and his referent that is why he claims that proper names do not have any meaning. Contrary to his opinion, G. F. Benecke suggests:

"That we cannot make an intelligent use of names without knowing their meanings and that if proper names have a meaning this must apply to them just as much as to general names, is so self-evident that it seems trifling to insist on it.

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

This definition is suitable for the names of human beings, countries, rivers or cities, e.g.:

Louis-Napoléon Bonaparte, China, Vistula, Berlin.

Such proper names are classified as “pure” in opposition to the ones that are “descriptive”. The latter result from the composition of a proper name with a common name (their expansion), e.g.:

Eiffel Tower, Rodin Museum.

Another subcategory of proper names is what we call fixed definite descriptions. These are proper names that are constituted of common nouns only, e.g.:

*Jardin des Plantes, United Nations Organization***.**

In regard to the fact that the two definitions quoted above are not applicable to the descriptive proper names, we thought it justifiable and reasonable to adopt the point of view of [Jonasson, 1994] since her definition includes both: pure and descriptive proper names.

“Every expression associated with a given individual in long-term memory by virtue of the link that is denominative, conventional and stable”.

Before creating our database, we had to define the scope of our project. Our goal was to gather all proper names that one can find in the everyday language. Thus we made a decision to exclude all expressions that are related to medical, scientific or juridical terminology (e.g. *Parkinson's disease*, *the Pythagoras' theorem* or *the law Pasqua*). Equally, names of diplomas and competitive examinations are omitted (e.g. *CAPES*).

2 Ontology

2.1 Prolexbase ontology

An ontology, according to [Temmerman, 2003]:

“*represents an agreed upon conceptualization of the real world*”.

Our ontology aims at modeling the linguistic class of proper names and it seems that in the multilingual context it is more suitable to represent proper names as ontology in the sense of [Gruber, 1995] who writes:

“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”.

That is the reason why we introduced a conceptual proper name: the pivot that is the referent from different points of view. The analysis of the properties of proper names shows that such proper names ontology must be structured in at least four levels (also called layers): two language independent levels: the conceptual (the numerical pivots) and the metaconceptual (types and supertypes) as well as two language dependent ones: the level of instances (the proper names such as they appear in a written text in a specific language) and the linguistic level (the level of so called "prolexemes"). In addition, we define relations between proper names to assure that our database is not only a list of words but a real relational dictionary. The architecture of such ontology is represented in Figure 1.

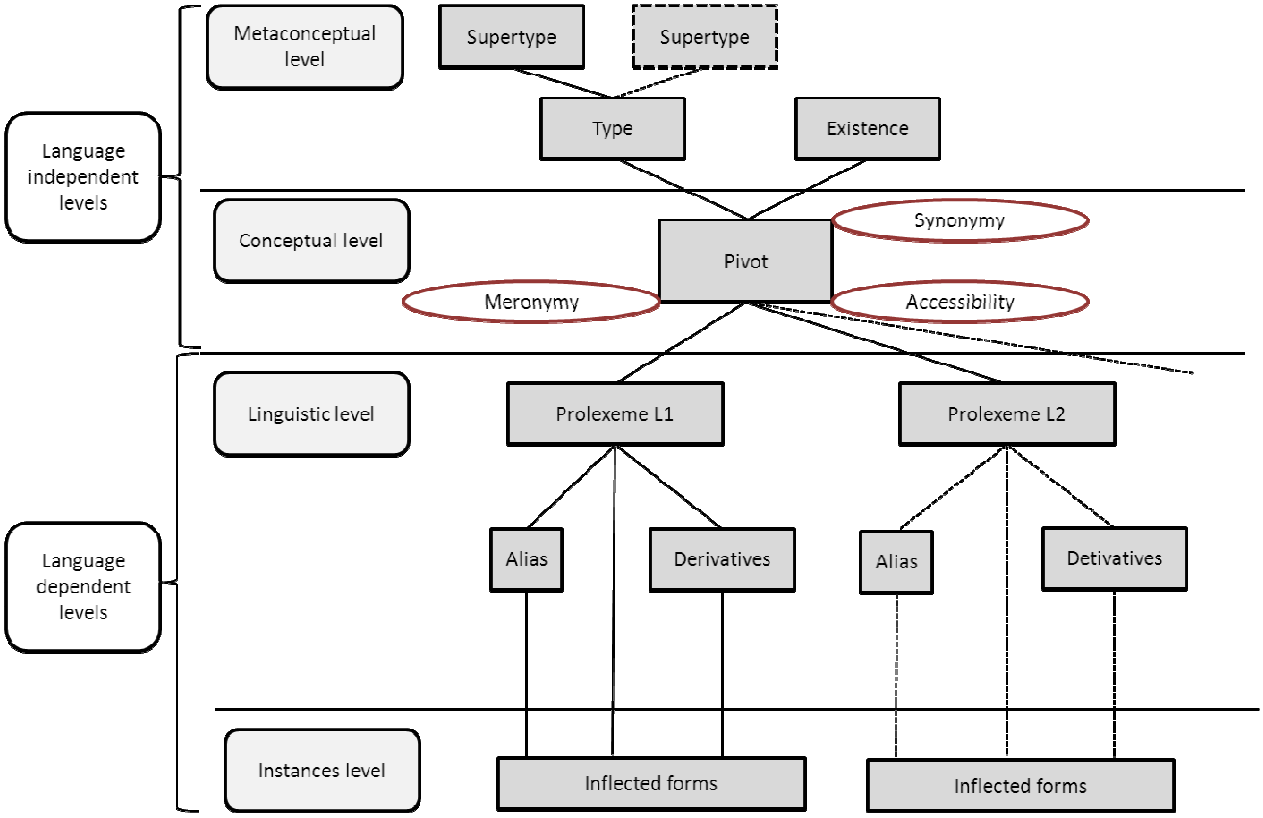


Figure 1 The general structure of the ontology of proper names

2.2 The metaconceptual level

This level enables a homogenous classification of proper names on the bases of supertype and type that are associated to every proper name. A supertype gives minimal information about a proper name and classifies it according to their traditional syntactic and semantic properties. It is generally possible to recognize the supertype from the linguistic context, without any human supervision. We have already developed some automatic procedures based on finite-state transducer cascades with 93,2% recall and 94,4% precision for the recognition of personal, organization and place names in French [Friburger, 2002]. In our ontology we distinguish four sypertypes corresponding to primary semantic features like *human*, *location*, *concrete* and *event*:

Anthroponyms	names of human beings, robots and animals, this supertype is divided into personal and collective names
Toponyms	place names with a subcategory of territory
Ergonyms	artifacts and work names, proper names that designate something concrete or abstract produced by a human being
Pragmonyms	event names

While supertypes provide us with basic information about the proper name, types, being more precise, give a more refined classification. These types are lexical classes determined by homogeneous semantical characteristics and a quite homogenous syntaxical behavior that can be useful for computer aided translations.

Thirty types were defined for the organization of this structured multilingual relational database:

1) Anthroponyms:
a) Individual:

- **Celebrity** - what these names have in common is that they consist of family names which are used without first names. The criterion is the lack of an honorific (*Mr, Me, Dr*) or an introductory (*president*) in relation with a family name in a text. This type includes diverse forms of celebrities names depending on the country and the epoch (*Plato, Victor Hugo, Alexander the Great, l'Abbé Pierre*). Pseudonyms belong to this type as well (Stephen King for Richard Bachman). We rejected the idea of making one particular type for pseudonyms and nicknames because there is mostly no indication in text that shows that a family name is a pseudonym.

  In French: **Célébrité**

- **Patronymic** - family names (*Shakespeare, Hitler*). Generally, we do not translate family names except for the stylistic reasons in literature.

  In French: **Patronyme**.

- **First name** - (*Charles*) appears mostly in combination with a family name. Although a first name can have different forms depending on the language (*Louis, Ludwig, Luis, Luigi*, etc.) or on the epoch (*Johan, Jehan, Jean*), they exist synchronously and in general are not translated.

  In French: **Prénom**

- **Pseudo-anthroponym** - names of animals, robots, machines, etc. (*Laika* - a soviet space dog).

  In French: **Pseudo-anthroponyme**

b) Collective:

- **Dynasty** - these are usually derivatives (*Merovingians).*

  In French: **Dynastie**

- **Ethnonym** - people names (*Angles*) or names of inhabitants of a country, town or region (*Londoners*). Should an ethnonym be a relational name which is associated to a toponym, it will not constitute a separate prolexeme (for example *Polish* will not constitute an independent entry but will be under the prolexeme *Poland*).

  In French: **Ethnonyme**

- **Association** - names of associations or political parties (*Labour Party)*.

  In French: **Association**

- **Ensemble** - artistic ensembles or sporting clubs names that include football teams (*Manchester United*) as well as rock groups (*The Smashing Pumpkins*).

  In French*:* **Ensemble**

- **Firm** – names of firms (*Ikea, General Electric, BASF*).

  In French: **Entreprise**

- **Institution** - names of public or private institutions, hospitals, universities, foundations, etc. (Oxford University),

  In French: **Institution**

- **Organization** - names of international and non-governmental organizations (*Unesco*).

  In French: **Organisation**

2) Toponyms:

a) Territory:

- **Country** - names of countries, states and kingdoms that exist (*Portugal*) or that used to exist (*Soviet Union*).

  In French: **Pays**

- **Region** - this name means a subdivision of a country such as regions, provinces, departments, voivodeships, etc. (Cambridgeshire). Islands (*Maui* - the second-largest of the Hawaiian Islands) will also belong to this type if they are not an independent state (for example the *Bahamas* will be considered as a country, not as a region).

  In French: **Région**

- **Supranational** - name of a group of countries (*the Balkans*).

  In French: **Supranational**

b) Others:

- **Astronym** - celestial bodies names (which are defined by a place and not as a phenomenon) include planets, asteroids, galaxies, etc. (*Jupiter*).

  In French: **Astronyme**

- **Building** - building names include names of parks, gardens, monuments, bridges, theatres, etc. (*Jardin du Luxembourg, Eiffel Tower*).

  In French: **Édifice**

- **City** - city, town, village and quarter names (*Warsaw, Bronx*).

  In French: **Ville**

- **Geonym** - these names are natural geographical sites (*the Sahara Desert, Mont-Blanc*), defined as natural forms of landscape. Geonyms are elements of physical geography like mountains, deserts, glaciers, caves, plains or forests.

  In French: **Géonyme**

- **Hydronym** - these names correspond to water areas, river or stream names (*the Atlantic Ocean, the Vistula*).

  In French: **Hydronyme**

- **Way** - square, road and street names (*RN 7, Fifth Avenue*).

  In French: **Voie**

3) Ergonyms:

- **Object** - fictive or mythical object names (*Excalibur*).
  In French: **Objet**
- **Product** - brand names or trademarks (*Mercedes, Microsoft Word*).

  In French: **Produit**

- **Thought** - names of religions, doctrines and theories (*Catholicism, Marxism*).

  In French: **Pensée**

- **Vessel** - names of vessels which travel on water, in air and in space (*Titanic*).

  In French : **Vaisseau**

- **Work** - names for books, films, theorems, sculptures, tables, etc. (the *Odyssey*).

  In French: **Œuvre**

4) Pragmonyms:

- **Disaster** –names of catastrophes stemming from events such as earthquakes, floods, accidents, fires, explosions, etc. (Chernobyl).

  In French: **Catastrophe**

- **Event** - sporting or cultural event names (*World Cup*).

  In French: **Manifestation**

- **Feast** - feast names with a cyclic character (*Easter*).

  In French: **Fête**

- **History** - historical or political event names (*French Revolution*).

  In French: **Histoire**

- **Meteorology** – names of regular meteorological events (wind *Mistral*).

  In French: **Météorologie**

Types are related to the supertypes by the relation of hyponymy. This means that a supertype is a hypernym of several types. Every type, however, is linked to one supertype only. The same relation is established between types and pivots.

This typology can be represented by Table 1:

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

Table 1 The Prolex typology

We can add to this strict hypernymy a secondary one (Table 2). For instance, a name of a building can be used to designate an artifact (*the Pyramid of Cheops*) or a name of a city can be perceived as a collective anthroponym (*London is waiting for the Manchester supporters*).

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

Table 2 Secondary typology of Prolex

Due to the fact that a pivot (see 2.3) can be directly related to a supertype in an automatic procedure, types and supertypes are listed in the same table. Each proper name is

associated to only one type, otherwise we consider them as homonyms and we duplicate their pivots, e.g.:

Washington as a city /toponym

Washington as a celebrity /anthroponym

Washington as a region /toponym

These three homonyms get three different pivots.

As far as a detailed classification of proper names is considered, there are two requirements: on the one hand, the types must be sufficiently discriminative so that it is possible for a non-specialist to assign one type to each proper name which is found; on the other hand, the types must be sufficiently precise to constitute the beginning of a definition.

2.2.1 The existence

Apart from the supertypes and types, the metalinguistic level possesses also what we call *existence*. Each pivot is linked to one and only one value of existence. This feature is often important information for translation.

We distinguish three types of existence:

1. **Historical**: names of people, events etc. that we know for certain that they have existed.
2. **Fictitious**: proper names are also used by authors of novels, story, play, film, etc.
3. **Religious**: this third feature depends on the faith of people. If *Jesus* and *Mohammed* are historical proper names, it is not the role of the linguist to say if the archangel *Gabriel* really exists or not.

Generally, the names linked to the features *Fictitious* or *Religious* are translated while the names linked to the feature *Historical* remain unchanged. For instance, *Snow White* is translated in French (*Blanche-Neige*), Serbian (*Snežana* and Снежана) and Polish (Królewna Śnieżka).

2.3 The conceptual level

The second language independent level: the conceptual one, is organized around the pivot, which is represented by the unique identification number (ID). The pivot has the role of an interlingual identifier enabling the connection of proper names that represent the same concepts in different languages. 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]…).

A concept is the link between a pivot and a canonical form of a proper name in one language. Pivots represent different points of view of the referent of a proper name (so they do not correspond directly to the language referent). For instance, although *Karol Wojtyła* and *John Paul II* refer to the same person, they get two different pivots since they represent two different points of view.

Conceptual proper names enable a definition of some relations on the conceptual level, such as synonymy, meronymy, accessibility.

2.3.1 Synonymy

As opposed to a commonly accepted idea, a lot of proper names do have synonyms, for instance almost all countries have a short and a long form, e.g.:

In French: *France* and *République française*.

Yet, we have to precise that it is almost impossible to find perfect synonyms in one language. That is why, the proper names like *France* and *République française* will be called quasi-synonyms as it is advised by ISO 12620.

Synonymy is a relation between two pivots designating the same referent. It is also the only relation in our database that can be both language dependent and language independent.

In our description of semantic relations, we use the four diasystematic features of Coseriu [Coseriu, 1998], defined in Table 3. These features correspond to different points of view that we can have about the linguistic referent.

Diachronic	variety depending on time
Diatopic	variety depending on the area
Diastratic	variety depending on sociocultural stratification
Diaphasic	variety depending on the usage purpose

Table 3 The diasystem of Coseriu

On the conceptual level (so in the language independent part of the database), three out of four diasystematic features are used: diachronic, diaphasic and diastratic.

Diachronic: a name has sometimes changed because of the historical reasons, for instance *Petersburg*, *Petrograd* and *Leningrad* in Russia or *Burma* and *Union of Myanmar*.

Diaphasic: it can be used for stylistic reasons, for instance, a tour operator prefers to use the name *City of Light* instead of *Paris.* It can be also used in certain contexts, for example, a political discourse often uses the system of government to speak about a country, such as *Kingdom of Morocco*, versus *Morocco*.

Diastratic: a famous person can have more than one name, but generally not with the same fame, for instance, the American singer-songwriter *Bob Dylan* is well-known, but very few are familiar with his real name: *Robert Allen Zimmerman*.

In the language dependent part, more precisely on the linguistic level, there are two kinds of synonymy: diastratic (so it is used twice in the database) and diatopic (see 2.4.2).

Diastratic: contrary to the diastratic synonymy on the conceptual level, this one depends on a specific knowledge and is not shared by other languages. This point of view shows

differences between standard and informal language, e.g. *Pole* (in the formal register) and *Polack* (in slang).

Diatopic: the existence of different regional languages in one country allows some cities, towns, etc. to have more than one name, e.g. *Nantes* (in French) and *Naoned* (in Breton) to designate the same city in the region of Brittany, France.

2.3.2 Meronymy

The relation of meronymy is well-known in terminological contexts. Meronymy (a partitive relation) is a relation between pivots [Miller et al., 1990]. The tables of meronyms list proper names that are themselves specified by another proper name in a relation of inclusion. It is natural to use it so as to describe the inclusion of toponyms or events, e.g.:

Serbia and *Bulgaria* are in the *Balkans*, which 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 (*The European Aeronautic Defense and Space Company*) is in *Europe*,

St Matthew's Gospel can be found in the *New Testament*,

Novak Djokovic is a citizen of *Serbia*, etc.

The relation of meronymy is frequently used in economical registers, for instance the *European firm* from *EADS*, or in the sport register, for instance the *Serbian tennis man* from *Novak Djokovic*.

2.3.3 Accessibility

Accessibility (associative relation) [Ariel, 1990] means that something/someone is accessible through something/someone else. In explanatory dictionaries, in contrast to common nouns, proper names do not have definitions. Some relations towards different names, generally better known, are usually 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*. It is thanks to *Moses* that we have access to *Aaron*. By contrast, *Moses* will be accessible through *Hebrews'* story.

We precise twelve subject files:

- **relative** (*Aaron* is the brother of *Moses*),

- **capital** (*Paris* is the capital of *France*),

- **leader** (*Angela Merkel* is a German *politician*),

- **founder** (*Henry Dunant* founded *the Red Cross*),

- **follower** (*Peter* is a disciple of *Jesus*),

- **creator** (*The Magic Flute* is an opera of *Wolfgang Amadeus Mozart*),

- **manager** (*Alex Ferguson* is a Scottish football manager currently managing the *Manchester United*),

- **tenant** (*Barack Obama* is the tenant of the *White House*),

- **heir** (*Charles, Prince of Wales* is the heir of *Queen Elisabeth II*),

- **headquarters** (In *Clermont-Ferrand,* there are the corporate headquarters of *Michelin*),

- **rival** (*Quick* is the rival company of *McDonald's*),

- **companion** (*Patroclus* was *Achilles'* beloved comrade and brother-in-arms).

2.3.4 Example of the relations between pivots

Thanks to the pivot 38558 - *Paris* we can have an overview of the existing language independent relations.

First of all, it is in the relation of (diaphasic) synonymy with the pivot 55120 - *City of Light* as they designate the same referent. It is also in the relation of meronymy with the pivot 5 - *Île-de-France* (Paris is in Île-de-Frace) and in the relation of accessibility with the pivot 27 - *France* (Paris is the capital of France).

The following figure presents these three relations around the pivot 38558.

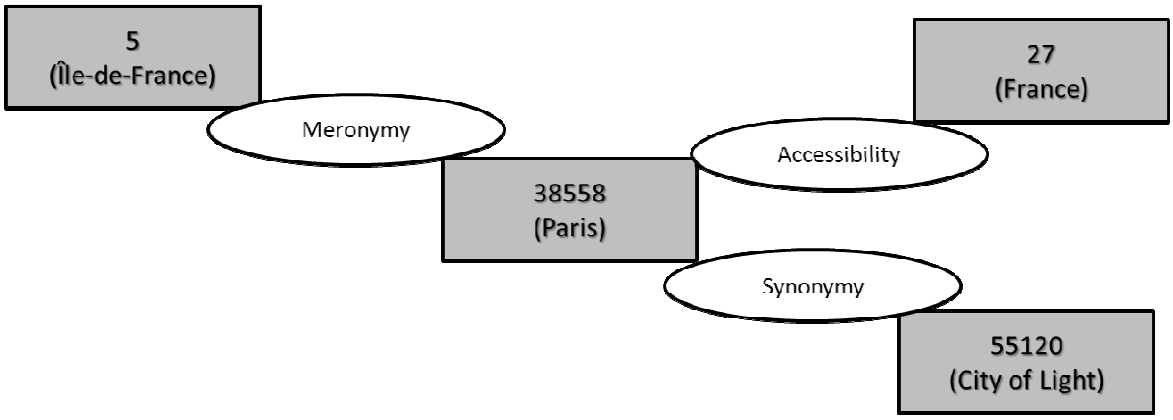


Figure 2 Relations of synonymy, meronymy and accessibility

2.4 Linguistic level

Linguistic level describes the realizations of a proper name in the given language.

2.4.1 Prolexeme

On the linguistic level, the canonical forms (lemmas), also called prolexemes, are defined and connected to the ID (pivot) for the particular language. A prolexeme is the set of all lemmas semantically linked to a proper name in the given language.
For example the prolexeme for an international organization whose stated aims are facilitating international cooperation and achievement of world peace is:

in English: *United Nations Organization*
in French: *Organisation des nation unies*
in Polish: *Organizacja Narodów Zjednoczonych*

These prolexemes in three different languages are linked to the same pivot as they represent the same point of view on the proper name's referent (Figure 3):

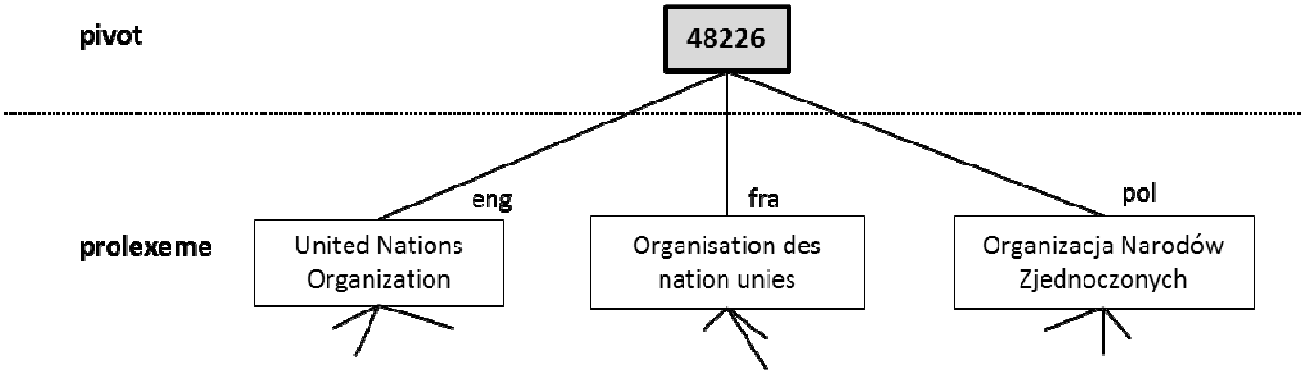


Figure 3 Pivot and prolexemes of the United Nations Organization

The prolexeme can have language dependent variations: aliases and derivatives.

2.4.2 Aliases

Aliases represent variations such as:

- other forms: short forms, abbreviations, acronyms, different orthographies, other transcriptions etc., e.g.:
  aliases for *George Walter Bush* are: *George W. Bush, George Bush, Bush*
  aliases for *United Nations Organization* are: *United Nations, UNO.*
- diatopic synonyms and some diastratic ones - these synonyms are too dependent on the language so as to have a pivot (see 2.3.1).
- explanation, e.g.:
  Caritas USA Organization to explain *Catholic Relief Services*.

Aliases are always connected to the prolexemes. In our database, we distinguish the following types of aliases:

- abbreviation,
- short form,
- acronym,
- explanation,
- transcribed form,
- variant,
- diastratic quasi synonym,
- diatopic quasi synonym,
- translation variant,
- Cyrillic transcribed form,
- Cyrillic abbreviation,
- Cyrillic viariant.

2.4.3 Derivatives

Derivatives are obtained by a morphosemantic derivation with a regular form-meaning sense:

1 **relational adjective**, e.g.:
Polish,

2 **relational name**, e.g.:
Pole,

3 **quasi relational name (diastratic),** e.g.:
Polack,

4 **prefix**, ex:
Anglo-

5 **quasi relational adjective (diastratic),** e.g.:
In French: une vodka *polak*.

According to the language, there can be more categories, e.g.:
progressive adjective in Serb.

We only add to the database the derivatives that are semantically linked to the proper name (we will not include the word *to pasteurize* despite being a derivative of the name *Pasteur*, because this is considered a lexicalized word with a specific definition which is independent of the name *Pasteur*).

2.4.4 Language dependent relations

On the one hand, in the Prolexbase, there are language independent relations between the pivots such as synonymy, meronymy and the relation of accessibility. On the other one, we can also distinguish language dependent relations that concern the prolexemes:

- **Collocation** – indicates the link that can be established between a proper name and a function word such as determiners, prepositions, etc. e.g.:

  Almost all names of countries take an article in French: *la France*, *le Portugal*.

  In French, when we want to say that we live or are going to a particular country, we use two different prepositions depending on the gender of the proper name: *en France* but *au Portugal*.

- **Context** – is a relation between a canonical form of a proper name and typical words appearing with it. It refers to external structure, as defined by [MacDonald, 1996]. We distinguish two types of context which are linked with information about the position and the syntactic structure of the proper name. We use templates where $1 represents the proper name from the first constituent, $2 from the second one, etc.

- the **classifying context** is an expansion of the noun phrase (capital, king, coach…). Let's take a look at the example of the French city *Le Mans*:
  Prolexeme: *Le Mans*
  Classifying context: *la ville du $2*
  This combination will give us: *la ville du Mans*.
  The classifying context is often useful for translation, e.g.:
  In Polish: *Wisła* (the Vistula) → in French: *la rivière Vistule.*

- the **accessibility context** is a noun phrase which can be perceived as a sort of explanation of the proper name. This noun phrase is very often in apposition, e.g.:

38558 (pivot1) is in the relation of accessibility with 27 (pivot2). In English the pivot 38558 corresponds to *Paris* whereas the pivot 27 to *France*.

Prolexeme1: *Paris*
Accessibility context: *the capital of $1*
Prolexeme2: *France*
In this case, we will obtain: *Paris, the capital of France.*
The capital of France is in apposition to *Paris* and at the same time it "explains" the proper name.

- **Eponymy** – this relation, in contrast to the other ones, informs us that the translation does not refer to a proper name but to:
  - a common noun (**antonomasia**): although *Pampers* is a brand of baby products, this word is used to designate the concept of disposable diapers in Polish: *pampersy*.
  - a terminological term (**terminology**): *Alzheimer's disease, Pythagoras' theorem*,
  - an idiomatic phrase (**idiom**): *Not for all the tea in China.*
- **Reliability code** – each prolexeme has a reliability code which indicates whether the proper name is well known or not. There are three features advised by ISO 12620 (Computer applications in terminology – Data categories):
  - commonly used,
  - infrequently used,
  - rarely used.
- **Sort** - a sorting relation gives the information on how to classify multiword proper names. Namely, many dictionaries arrange multiword proper names by inverting their constituent parts [Tran et al., 2005]. For instance, if we want to look up *George Washington* in encyclopedia, we have to search it under letter *W*, not *G*. In this case, it is the second constituent that counts.
- **Language** - each prolexeme is associated to one language. That means that two homographs in two different languages are duplicated. For example, there are two items *France*, one for French and one for English.

  Each language treated in the database appears with its ISO 639 Language Code in the table, e.g. *fra* for *French* or *eng* for *English*. Because of the multilingual dimension, a multibyte character encoding for Unicode: UTF-8 is used.

  In Prolexbase, one can find the following languages:
  French, German, Italian, Portuguese, Spanish, Dutch, English, Serbian, Korean and Polish.
- **Phonetics** - if a proper name does not have a translation in the target language which in addition does not use the same alphabet, thanks to the phonetics we can obtain the transcription.

These language dependent relations are represented by Figure 4.

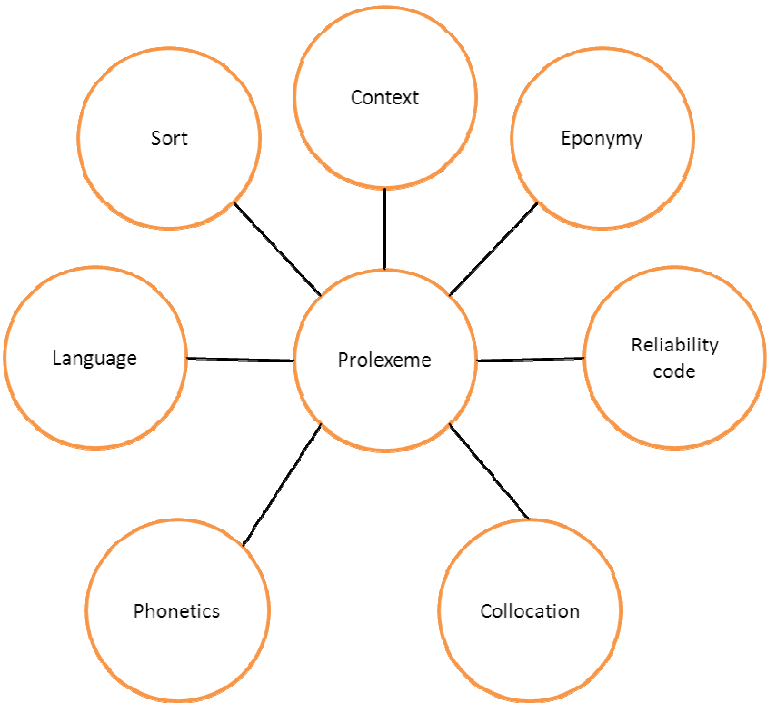


Figure 4 The Prolex language dependent relations

Let's take a look at the example of name translations from French to English:

Un Tourangeau m'a dit que la Loire est magnifique. →
An inhabitant of the city of Tours in France has told me that the Loire River is splendid.

This translation could be deduced from Prolexbase:

• *Tourangeau*	• *Loire*
[Prolexeme] → Tours [Morphosemantic] → Derivative (Relational noun) [Possible meaning] → inhabitant [Classifying context] → city [Partitive relation] → France	[Prolexeme] → Loire [Classifying context] → river.

2.5 Instances level

The level of instances contains sets of all real forms of prolexèmes, their aliases and derivatives. In other words, we can find on this level all inflected forms (instances) that a proper name can have. Each instance is linguistically described: its inflectional properties are given as well as the part of speech to which it corresponds (*noun, adjective, prefix, verb*).

The number of inflected forms on the instances level depends on the morphology of the given language. In general, nouns, pronouns and adjectives inflect for *gender, number* and *case* while verbs can be modified depending on the *person* and *TAM* (*Tense, Aspect, Mood*). To exemplify some differences between languages, let's take a look at the prolexeme *Italy* in English, French and Polish.

Abbreviations:

N = noun,	mp = personal masculine,	pl = plural,	Inst = instrumental,
A = adjective,	nmp = non personal masculine,	Nom = nominative,	Loc = locative,
P = prefix,	sg = singular,	Gen = genitive,	Voc = vocative
m = masculine,		Dat – dative,	
f = feminine,		Acc = accusative,	
n = neuter,			

- English, being a morphologically poor language, has no more than five instances of the prolexeme *Italy*:

Inflexion	**Instances**
N, sg	Italy
N, sg	Italian
N, pl	Italians
A	Italian
P	Italo

- French is slightly richer than English. We can find ten instances of the prolexeme *Italie*:

Inflexion	**Instances**
N, f, sg	Italie
N, m, sg	Italien
N, m, pl	Italiens
N, f, sg	Italienne
N, f, pl	Italiennes
A, m, sg	italien
A, m, pl	italiens
A, f, sg	italienne
A, f, pl	italiennes
P	italo

- Polish, like other Slavic languages, is highly inflected. Polish retains seven cases from the old-Slavic case system (*nominative, genitive, dative, accusative, instrumental, locative* and *vocative*), has three gender classes (masculine, feminine, neuter) and inflects for number. As a result, on the instances level of the prolexeme *Włochy* (Italy), there are up to seventy inflected forms:

Inflexion	Instances						
	Nom	Gen	Dat	Acc	Inst	Loc	Voc
N, nmp, pl	Włochy	Włoch	Włochom	Włochy	Włochami	Włoszech	Włochy!
N, m, sg	Włoch	Włocha	Włochowi	Włocha	Włochem	Włochu	Włochu!
N, mp, pl	Włosi	Włochów	Włochom	Włochów	Włochami	Włochach	Włosi!
N, f, sg	Włoszka	Włoszki	Włoszce	Włoszkę	Włoszką	Włoszce	Włoszko!
N, f, pl	Włoszki	Włoszek	Włoszkom	Włoszki	Włoszkami	Włoszkach	Włoszki!
A, m, sg	włoski	włoskiego	włoskiemu	włoskiego	włoskim	włoskim	włoski!
A, mp, pl	włoscy	włoskich	włoskim	włoskich	włoskimi	włoskich	włoscy!
A, f, sg	włoska	włoskiej	włoskiej	włoską	włoską	włoskiej	włoska!
A, n, sg	włoskie	włoskiego	włoskiemu	włoskie	włoskim	włoskim	włoskie!
A, nmp, pl	włoskie	włoskich	włoskim	włoskie	włoskimi	włoskich	włoskie!

In Prolexbase, there is a possibility to add rules thanks to which we can obtain all inflected forms of a proper name. A specific tool generates all instances by use of finite-state transducers. For the French language, this tool is based on the Unitex software [Paumier, 2003], and the Multiflex system [Savary, 2005] for multiword units.

In the French part of the database, every simple word is associated with an inflectional code (rule), e.g.:
The French word: *Tourangeau* (=an inhabitant of the city of Tours in France) is a relational noun of the prolexeme *Tours*. This derivative is associated with the inflectional code N72 which corresponds to the endings: *au*, *lle*, *aux* et *lles* (Figure 5):

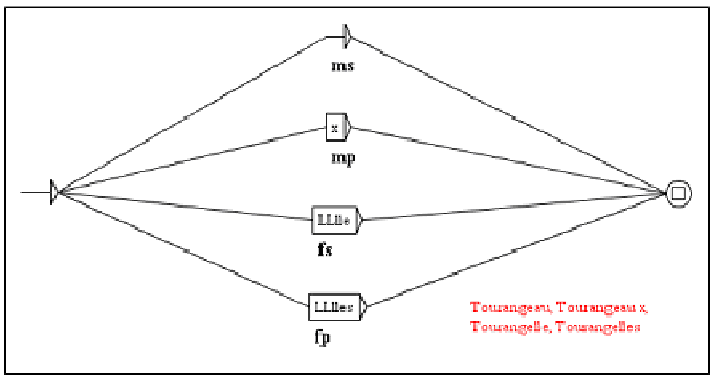


Figure 5 Inflectional graph N72

As far as multiword units are concerned, they need a grammar-based approach in order to be processed since they are linguistic objects on the border between morphology and syntax. We present the following example in French:

Antiguais-et-Barbudien, relational noun of the prolexeme *Antigua-et-Barbuda*, is associated with two units: *Antiguais*.N61:ms and *Barbudien*.N41:ms. In regard to the fact that both of his constituents must be inflected (this is not always the case!), the rule of the multiword inflexion that this derivative requires is NXXXN (Figure 6).

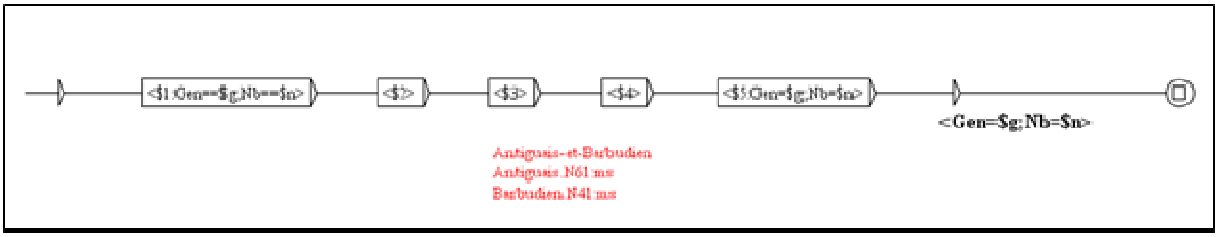


Figure 6 The multiword inflectional graph NXXXN

<$1>, <$2>, <$3>, <$4>, <$5> are arguments (everything counts, all lemmas, hyphens and even spaces) of the French multiword unit: *Antiguais-et-Barbudien*.

2.6 Summary of the main points

As previously mentioned, Prolexbase is a relational database. There are two language independent levels and two language dependent ones. Relations between these levels, pivots and prolexemes are established. The following figure represents the ontological model of Prolexbase:

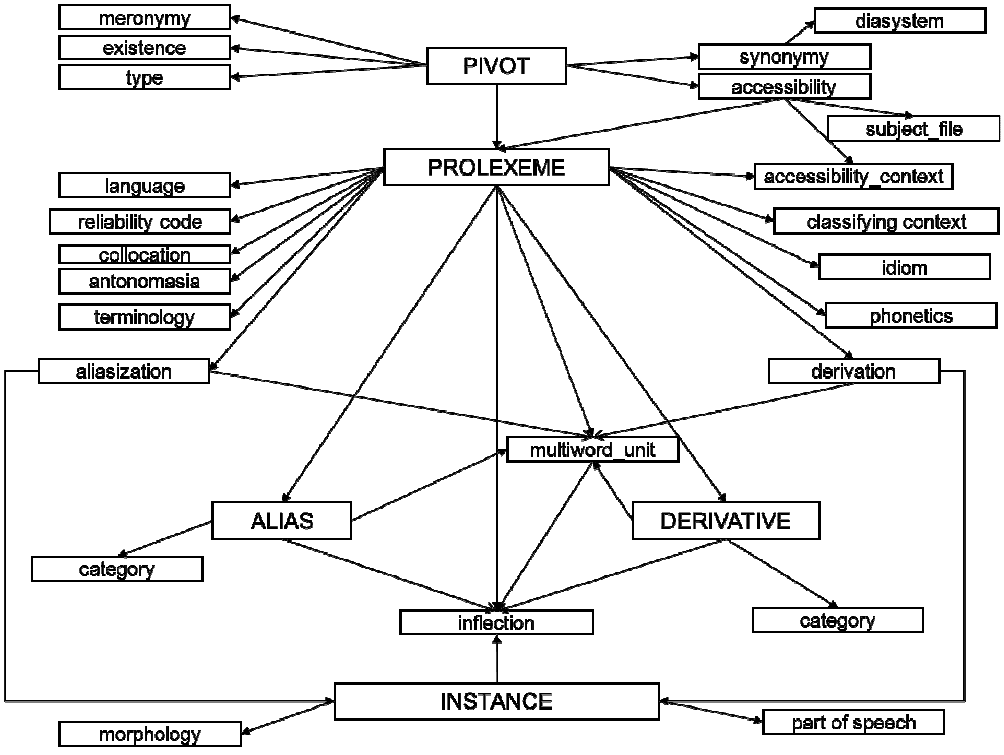


Figure 7 Ontological model of Prolexbase

2.7 Example

As an example of the multilingual ontology of proper names, a simplified model of the implementation of the proper name *Paris* in French and in Serbian is represented in Figure 8. On the conceptual level, the unique pivot 38558 corresponds to the proper name *Paris*. Its type on the metaconceptual level is City with the supertype Toponym. This pivot is in the relation of diaphasic synonymy with the pivot whose prolexeme is *Ville lumière* (City of Light). We know that *Paris* is a real city, that is why we chose historical as its value of existence.

On the linguistic level, the pivot 38558 is realized in French as the prolexeme *Paris*, which does not have any aliases, but has a derivative *Parisien* - an inhabitant of Paris - that has a diastratic synonym *Parigot* - an inhabitant of Paris in French slang. Though synonymous, there are sociolinguistic differences between them as well as differences in the communication situations in which they are used. On the level of instances, only one instance *Paris* marked as a masculine/feminine gender noun (MF) in singular (S) corresponds to the prolexeme *Paris*, while four instances correspond to derivative *Parisien* defined by its inflective paradigm. The prolexeme *Paris* has also the relational adjectives *parisien* and *parigot* as derivatives. This relational adjective possesses its own instances but they are not represented in Figure 8 due to the lack of space.

In Serbian, the prolexeme corresponding to the pivot 38558 is *Pariz*, and its alias is its Cyrillic recording *Париз*. Derivational processes in Serbian are much more complex than in French. Besides the relational adjective *pariski*, and the name for a masculine inhabitant, *Parižanin*, a separate form exists for a feminine inhabitant, *Parižanka*. From inhabitant names, not only a relational adjective can be derived *parižanski* (which is related to the inhabitants of Paris), but also possessive ones *Parižaninov* (belonging to a *Parižanin*) and *Parižankin* (belonging to a *Parižanka*) - these derived forms (and their Cyrillic counterparts) are not represented in Figure 8. On the level of instances, the set of inflected forms corresponds to the prolexeme, its aliases, and all derived forms, the correspondence being established by appropriate regular expression. It should be noted that in Serbian, the derivational level has itself two levels: on the first level are forms derived directly from the prolexeme or its aliases, while on the second level are forms that are being systematically produced from derived forms by the mechanism of structural derivation.

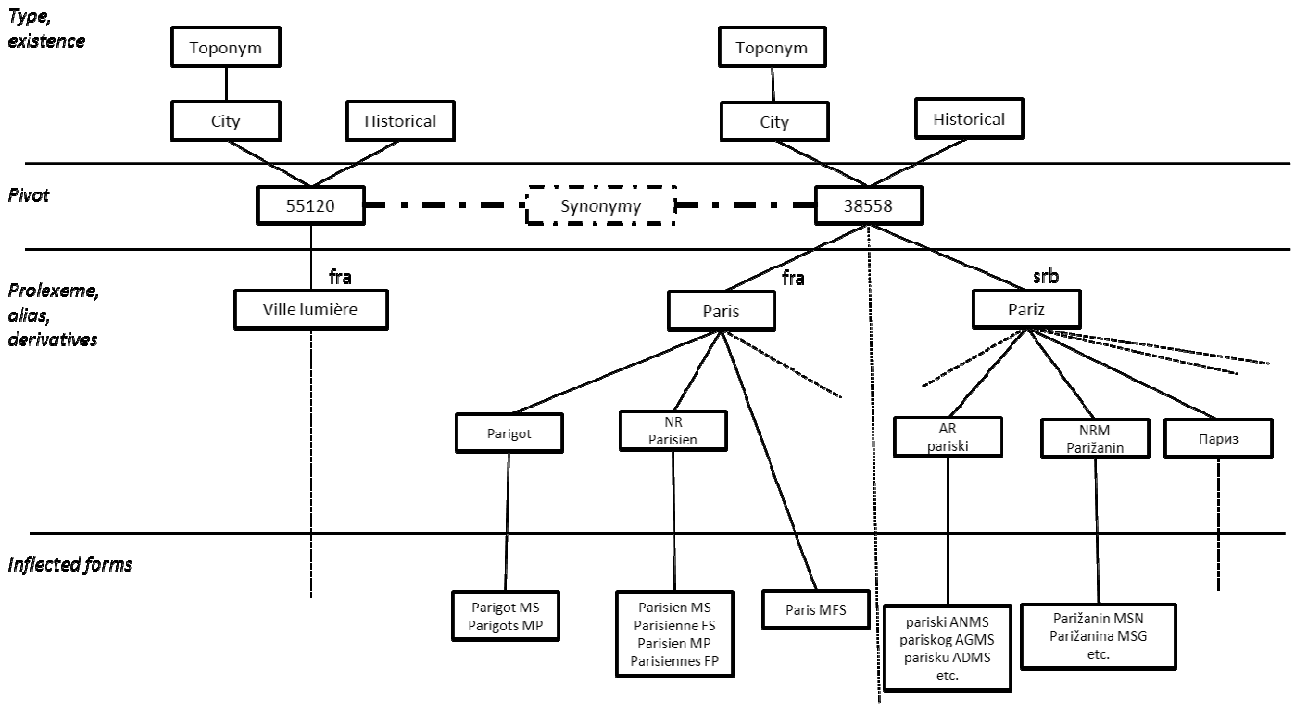


Figure 8 The concept of the proper name *Paris* in French and Serbian

3 Database

As it is based on the idea that NLP requires dual skills, to create the multilingual relational database of proper names Prolexbase, both linguists and computer scientists from different countries were brought together. Today, Prolexbase contains essentially proper names in French, but also some translations in other languages. The French part of the database contains 75 368 lemmas, shared among 65 805 nouns, 10 300 adjectives and 13 prefixes; these lemmas generate 123 859 inflected forms.

We have built this model from the different concepts and relations found in our ontology of proper names which results from the studies on their typology and on their inflectional and derivational mechanisms in different languages. We choose an Entity-Relationship model to represent our data. Figure 9 gives an overview of this model. The relationships are represented by ovals while all the data (entities) by rectangles.

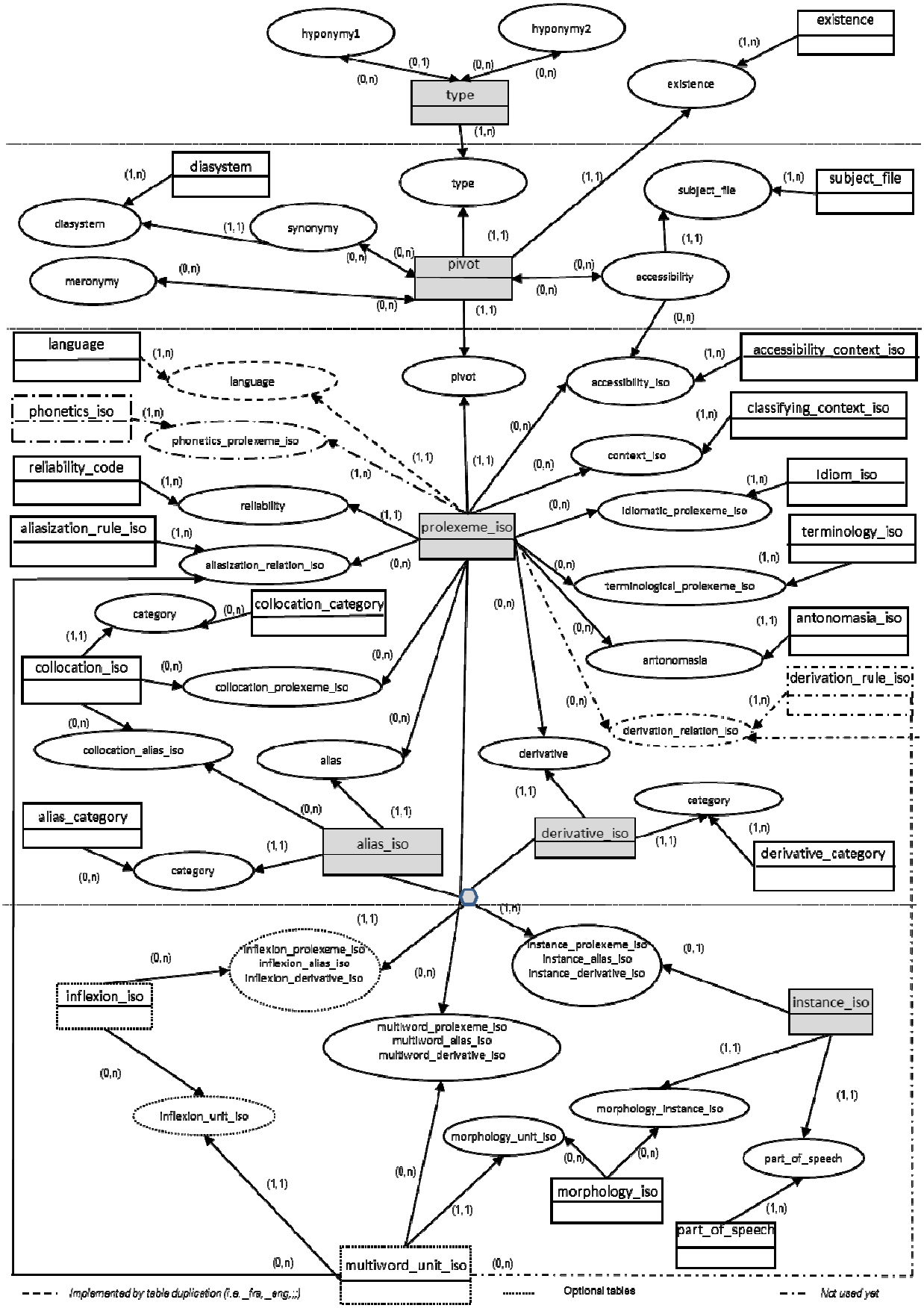
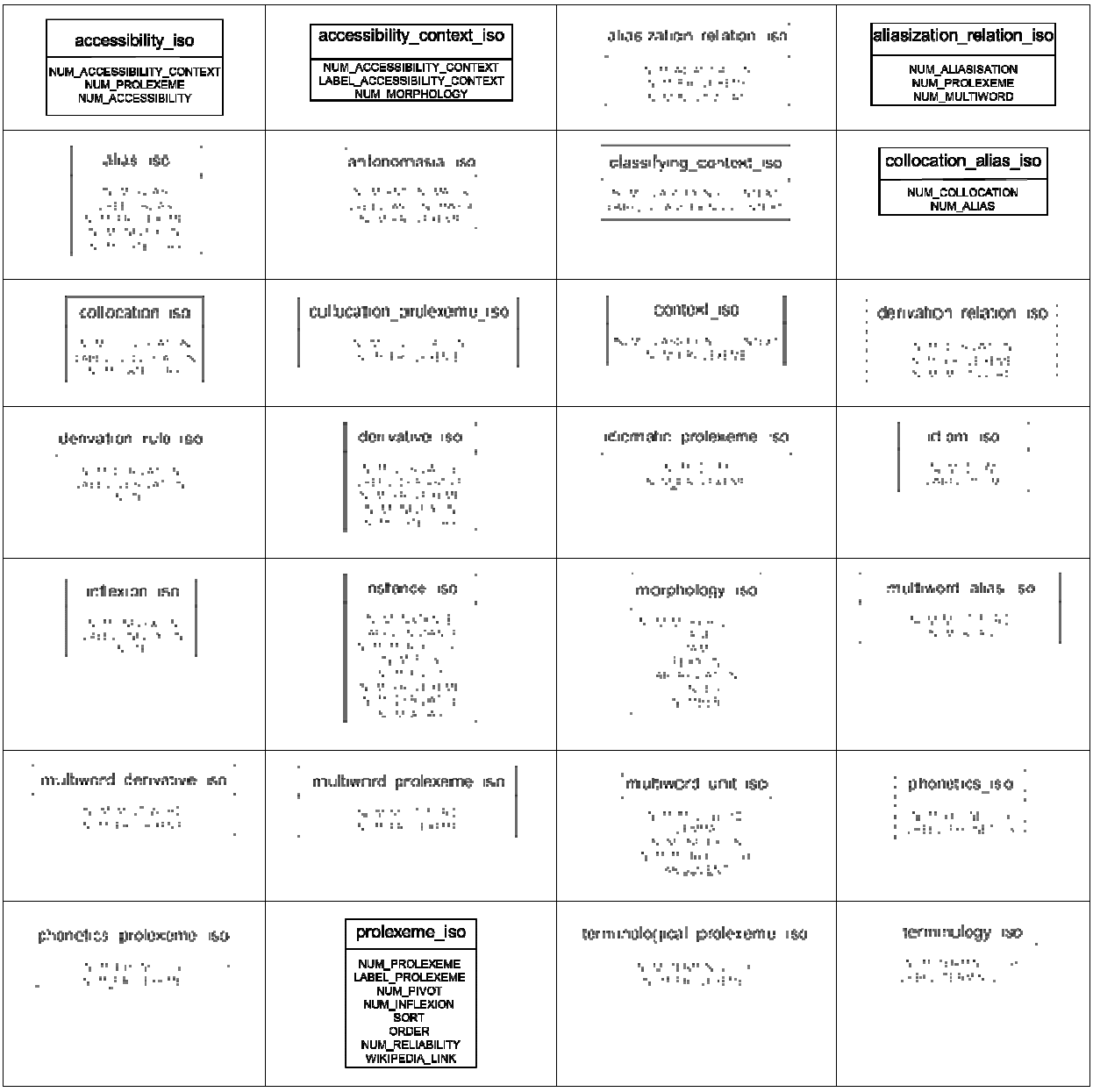


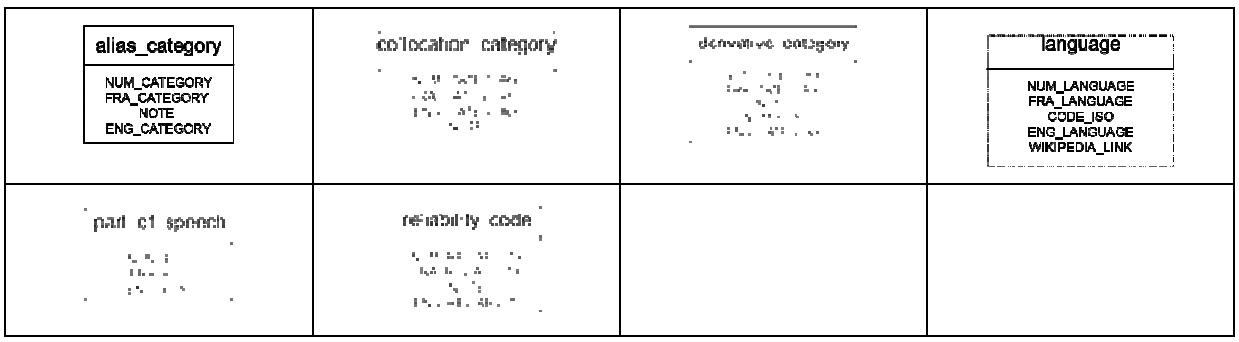
Figure 9 Model of Prolexbase

In Prolexbase, we can find the following tables:

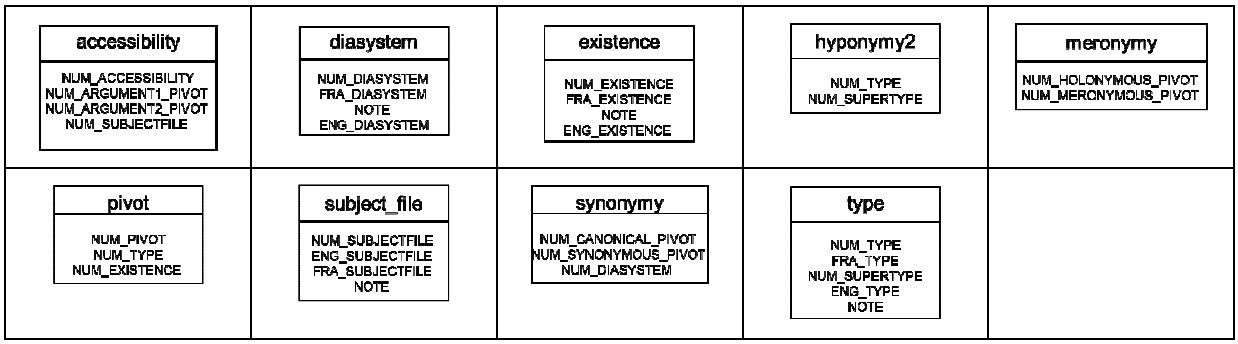
- **thirty four language dependent tables, six of which are common:**



Common language dependent tables:



- **nine language independent tables**



Explanations:

Num_	stands for the number, it is the identifier in the database (attribute),
Label_	symbolizes the entry,
Note	is a commentary or an example,
FRA_	means that the category, language, part of speech, etc. are written in French,
ENG_	means that the category, language, part of speech, etc. are written in English.

In the Prolexbase, we can find the following data categories:

Metaconceptual level

TABLE	COLUMN	VALUES
existence	**ENG_EXISTENCE**	historical, fictitious, religious.
type	**ENG_TYPE**	properName, anthroponym, individual, celebrity, patronymic, firstName, pseudoAnthroponym, collective, dynasty, ethnonym, group, association, ensemble, firm, institution, organization,toponym, territory, region, country, supranational, astronym, building, city, geonym, hydronym, way,ergonym, object, product, thought, vessel, work,pragmonym, disaster, event, feast, history, meteorology.

Conceptual level

TABLE	COLUMN	VALUES
diasystem	**ENG_DIASYSTEM**	diachronic, diaphasic, diastratic.
subject_file	**ENG_SUBJECTFILE**	relative, capital, leader, founder, follower, creator, manager, tenant, heir, headquarters, rival, companion.

Linguistic level

TABLE	COLUMN	VALUES
alias_category	**ENG_CATEGORY**	abbreviation,variant, shortForm, diatopicQuasiSynonym, diastraticQuasiSynonym, transcribedForm, acronym, explanation,translationVariant, cyrillicViariant, cyrillicTranscribedForm, cyrillicAbbreviation.
collocation_category	**ENG_CATEGORY**	determiner, locativePreposition.
derivative_category	**ENG_CATEGORY**	relationalAdjective,possessiveAdjective,quasiRelationalAdjective,cyrillicRelationalAdjective, relationalName, quasiRelationalName,prefix,masculineRelationalName, femininеRelationalName, cyrillicMasculineRelationalName, cyrillicFemininеRelationalName, cyrillicPossessiveAdjective.
language	**ENG_LANGUAGE**	French, German, Italian, Portuguese, Spanish, Dutch, English, Serbian, Korean, Polish
reliability_code	**ENG_RELIABILITY**	commonlyUsed, infrequentlyUsed, rarelyUsed.

Instances level

TABLE	COLUMN	VALUES
part_of_speech	**ENG_POS**	adjective, noun, prefix, verb.
morphology_iso	**CASE**	**NULL**
morphology_iso	**TAM**	**NULL**
morphology_iso	**PERSON**	**NULL**
morphology_iso	**GENDER**	masculine, feminine, masculineFeminine, none.
morphology_iso	**NUMBER**	singular, plural, singularPlural, none.

4 Prolexbase and LMF

4.1 LMF – basic information

LMF stands for *lexical markup framework* and is the ISO standard for natural language processing published on 17 November 2008. Nowadays, in the context of multilingual communication, a standardization of principles and methods relating to language resources is inevitable. LMF offers a universal model not only for the creation and use of lexical resources but also to exchange data between and among these resources. Being a result of five years of work, LMF should be seen as a synthesis of the state of the art in NLP lexicon field. It explains why this ISO standard is able to represent a wide range of lexicons, no matter whether they are small or large, simple or complex, monolingual, bilingual or even multilingual. The considered languages are not limited to European ones either.

In the following part, all given definitions come from ISO 12620 and ISO/TC 37.

4.1.1 Core package and extensions

We can distinguish two components of LMF: the core package (the structural skeleton) and its extensions. The basic hierarchy of information in a lexical entry is described by the core package (Figure 10). In this hierarchy, we can find the following details:

- **Global information** is a class which represents administrative information (at least the language coding but it is also possible to add the script coding and the character coding),
- **Lexical resource** represents a database consisting of one or several lexicons (e.g. Prolexbase),
- **Lexicon** is a class which contains all the lexical entries of a given language within the entire resource,
- **Lexical entry** is a class representing a lexeme in a given language,
- **Form** is an abstract class which represents a lexeme, a morphological variant of a lexeme or a morph,
- **Form Representation** is a class representing one orthographic variant of a *Form*,
- **Sense** is a class representing one meaning of a lexical entry,

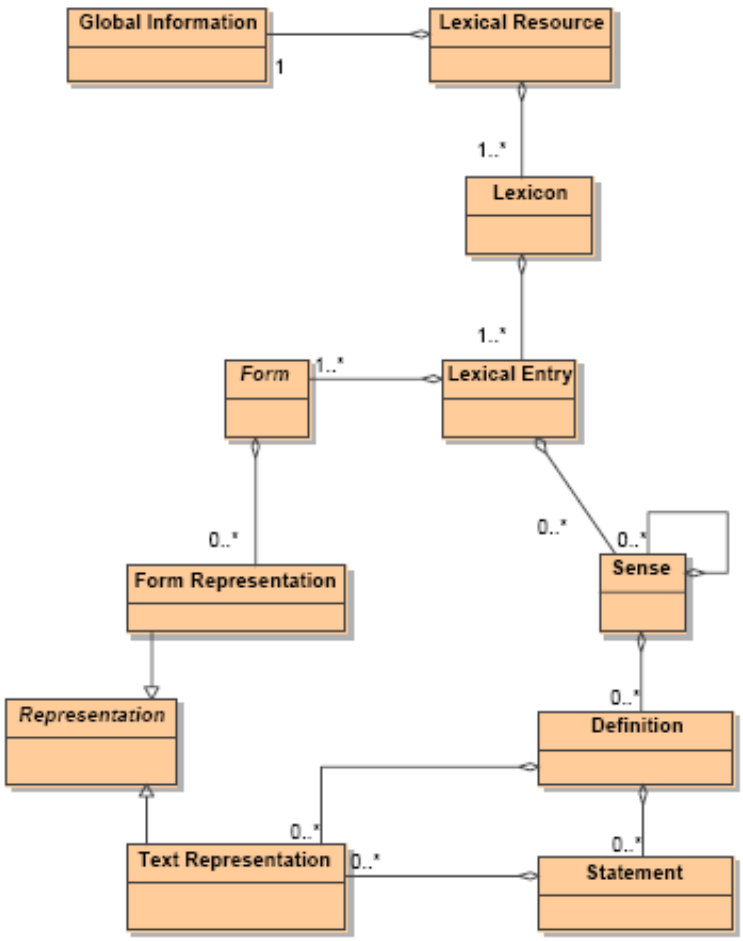


Figure 10 LMF core package

The lexicons containing forms and senses are the heart of the LMF, in other words, they are an obligatory core of the description. All other information can be found in eight extensions of the core package. These extensions are specifically dedicated to **morphology, Machine Readable Dictionary, NLP syntax, NLP semantics, NLP multilingual notations, NLP paradigm patterns, multiword expressions patterns** as well as **constraint expression ones** (Figure 11).

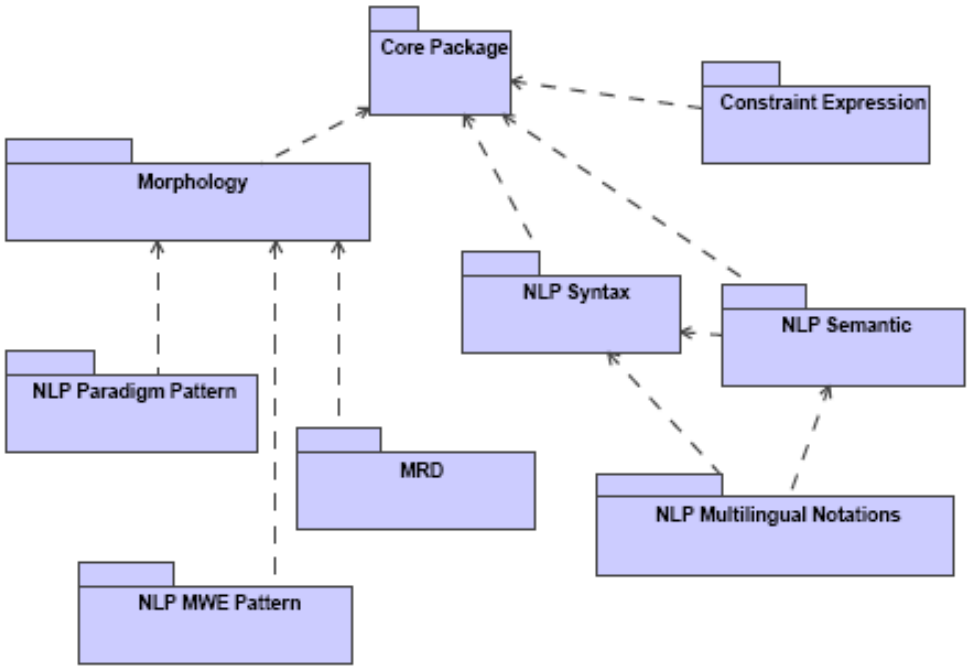


Figure 11 Extensions packages

If we wish to add inflected forms, we have to use the *Morphology extension* with the classes **Lemma** and **Word Form**. On the other hand, if we want to link a sense that is proper to one language with a sense that can be shared by other languages, the thing to do is to use the *Multilingual notation extension*. In this extension, we can find the idea of the pivot (multilingual identifier used for translation): **Sense Axis**, a class representing the relationship between different closely related senses in different languages. This class, just like *Lexicons*, is linked directly to the *Lexical Resource.* Thanks to the class of **Sense Axis Relation**, we can define relationships between pivots whereas the **Interlingual External Ref** class allows the pivot to be linked with external descriptions.

4.1.2 DCR

ISO 12620: "Data items appearing in individual terminological entries are themselves identified according to data category. Differences in approach and individual objectives inevitably lead to variations in data category definition and in the assignment of category names. The use of uniform data category names and definitions, at least at the interchange level, contributes to system coherence and enhances the reusability of data". While classes and their links are standardized by LMF, the attributes that we want to attach to them are in

need of DCR (Data Category Registry). DCR is a set of data category specifications determined by ISO 12620. This standard defines the following data categories:

- term - a designation of a defined concept,
- term-related information - an attribute assigned to a term (term type, usage, term formation, pronunciation, syllabification, hyphenation, morphology, term status, degree of synonymy),
- equivalence - for the multilingual aspect (degree of equivalence, false friend, directionality, reliability code, transfer comment),
- subject field - an area of human knowledge to which a terminological record is assigned (classification system, classification number),
- concept-related description - explanatory material (definition, explanation, context, example, nontextual illustrations, unit, range, characteristic),
- concepts relation - a semantic link between concepts (generic, partitive, sequential, temporal, associative, spatial),
- conceptual structures (concept system, concept position),
- note - additional information,
- documentary language - a formalized language used to characterize data so as to enable their storage and retrieval (thesaurus name, thesaurus descriptor, nondescriptor, keyword, index heading),
- administrative information (terminology management transactions and functions, subset identifier, authorization information, user suggestion, administrative term qualifiers, language symbol, foreign text, collating sequence, entry type, element working status, target database, entry source, concept identifier, entry identifier, record identifier, file identifier, cross-reference, source, source identifier, namespace identifier, originating entity).

In this registry, the name and definition of each category are specified as well as values that every category can take. Every designer of terminology databases ought to make sure that the content of data categories in their systems conforms to the content defined in DCR.

The listing of all data categories can be found on the following site [Romary, 2000]: http//:www.ttt.org/clsframe/datcats.html

4.2 ProlexLMF

If we want to export Prolexbase within the framework of LMF, the following measures must be taken: to instantiate the obligatory core by the elements highlighted in Prolexbase, to choose the necessary extensions and to define all the data categories so as to be able to describe the prolexemes and their relations. Henceforth the Prolexbase model conforming to LMF will be called ProlexLMF. ProlexLMF is presented as an XML-based format.

In ProlexLMF, the obligatory core of LMF is exploited: that is to say forms and senses. Apart from it, the *Morphology* and *Semantic extensions* are used as well. We would like to point out that there are two predominant differences between the Prolexbase model and the classical presentation described by LMF.

The first principal difference concerns the term of lexical entry. In Prolexbase, it corresponds to the prolexeme which is a set of lemmas whereas in LMF, it is a single lemma. The second important difference is that the prolexeme relates to only one pivot while one LMF entry can group homonyms. Because of these differences, ProlexLMF requires certain reorganization. Not only do we have to create a lexical entry for every lemma (prolexeme, derivative, alias – see 4.2.1), we are obliged in some measure to arrange our sense system differently as well.

4.2.1 Instances level

Since the *Morphology extension* provides only one lemma per lexical entry, a lexical entry is created for every derivative and alias.

In our opinion, the full form is the most appropriate lemma to represent the prolexeme. As far as the types of aliases are concerned, the representation attribute *orthographyName* contains the information about the writing variants.

Let's take a look at the following three examples:
Figure 12: (fra) *Organisation des Nations unies* – a representative lemma of the prolexeme,
Figure 13: (fra) *ONU* – an alias,
Figure 14: (fra) *onusien* – a derivative (relational adjective).

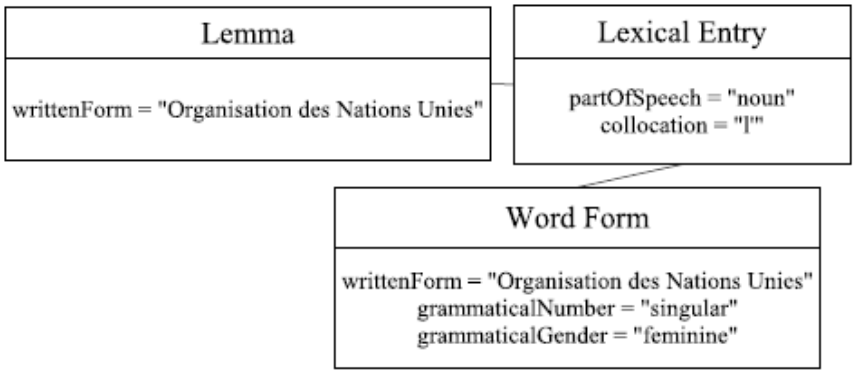


Figure 12 Representative lemma "Organisation des Nations unies"

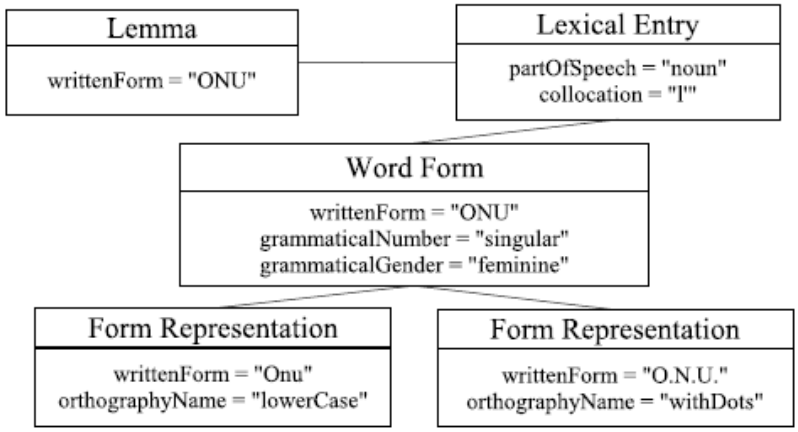


Figure 13 Alias “ONU”

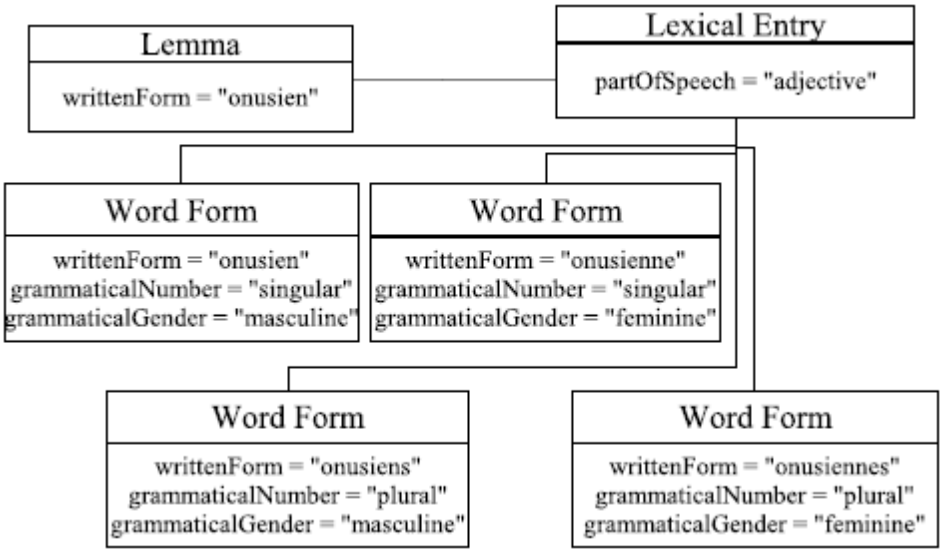


Figure 14 Derivative “onusien”

The grammatical category appears as the attribute of these three lexical entries (*partOfSpeech*) whereas their lemmas contain the attribute *writtenForm*. In addition, we can see that the prolexeme and its alias have another attribute: *collocation*. Depending on the morphology of the given language, the attributes associated with forms can vary (number, gender, case, etc.). Since French is a relatively poor language as far as morphology is concerned, in our examples, there is only one inflected form for the prolexeme and alias (there is no inflectional variation) and four ones for the derivatives (concord in gender and number). Finally, the alias *ONU* can have two other representations (*orthograhyName*): written with only one capital letter (*lower case*) or with dots.

4.2.2 Linguistic level

In general, two data categories are used to describe the semantics of entries of the same prolexeme: term provenance and etymology. Thanks to them, the link between the prolexeme and its aliases and derivatives is kept:

- *term provenance* – classification of a term according to the methodology employed in creating the term,
- *etymology* – information on the origin of a word and the development of its meaning.

Figure 15 takes up the three previous examples: *Organisation des Nations unies, ONU* and *onusien*.

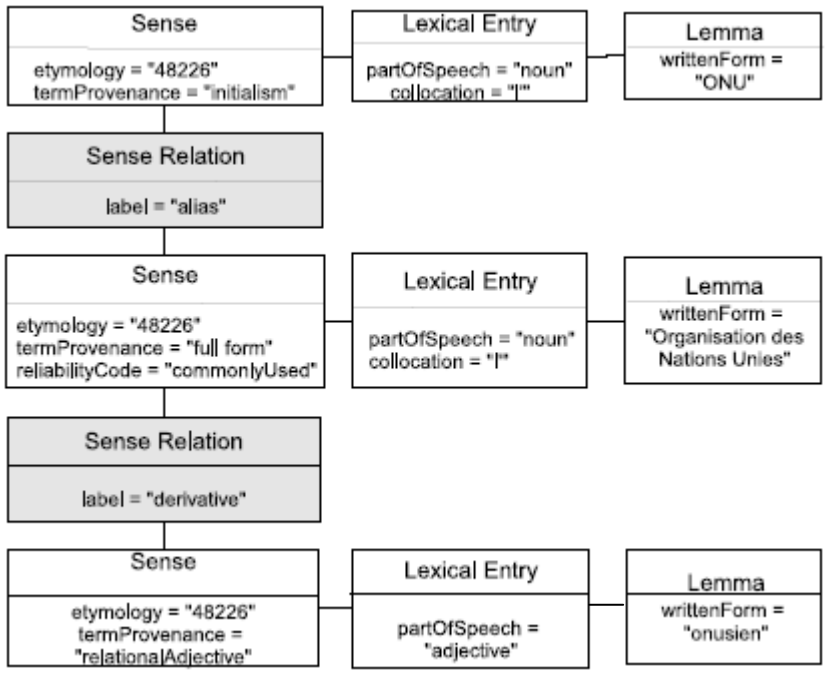


Figure 15 Sense relations for *Organisation des Nations unies, ONU* and *onusien*

As for their term formation, it is the type of alias that is relevant for *Organisation des Nations unies* (*termProvenance = "fullForm"*) and *ONU* (*termProvenance = "initialism"*) while for *onusien* it is the derivative category (*termProvenance = "relationalAdjective"*). In these three cases, etymology represents a unique identifier of the corresponding pivot (*etymology = "48 226"*). The idea of prolexeme is not lost as different lemmas of the same prolexeme can be grouped by sharing the same etymology. We also add the "fame attribute" (*reliabilityCode= "commonlyUsed"*) to the sense of the prolexeme since this information can be helpful in case of homonymy. As it can be observed in Figure 16, all three lexical entries are in relation. The sense relations contain information (*label*) about the type of lexical entry (alias, derivative).

To sum up, all aliases and derivatives are represented by a separate lexical entry. These lexical entries being linked by means of sense relations form an arborescence whose root is the representative lemma of the prolexeme (the full form).

4.2.3 Language independent part of ProlexLMF

Thanks to the *Multilingual notations extension*, we are able to represent the pivots as well as their relationships. For instance, the pivot 38 558 (*Paris*) is in the relation of accessibility (capital) with the pivot 27 (*France)*. Thanks to the pivots, prolexemes in different languages are linked up: the pivot 38 558 links the prolexeme *Paris* in French with the prolexeme *Paris* in English. Figure 16 gives an overview of these multilingual semantic relations.

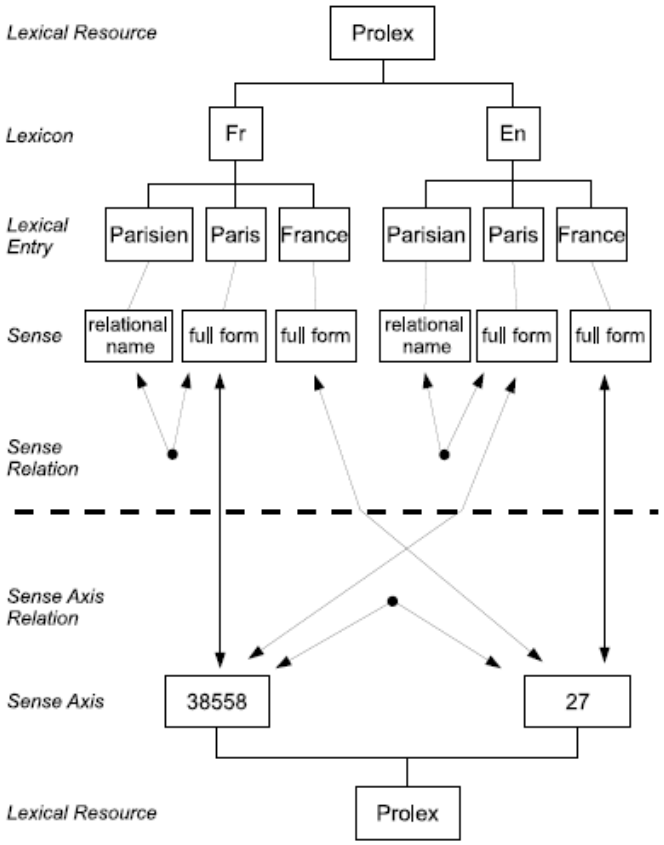


Figure 16 Multilingual relations in ProlexLMF

It is essential to understand that the classes *Lexicon* and *Sense Axis* are linked at the *Lexical Resource* level. We can see the lexical entries that correspond to the prolexeme *Paris*, to its derivative *Parisien* and to the prolexeme *France*, all that in English and in French. Every lemma is linked to the *Sense Axis* via its etymology, and vice-versa, every *Sense Axis* can refer to a lemma in different languages. The relation of accessibility between *Paris* and *France* is established through a relation between corresponding pivots. The relations at the level of the pivots can be characterized by the following data categories:

- quasi-synonym – synonymy,
- partitive relation – meronymy,
- associative relation – accessibility,
- generic relation – hypernymy.

In Figure 17, we can see an example of the relation of accessibility. This relation associates two pivots: *Paris* and *France*, its label is *associativeRealtion* and its subject field is *capital*.

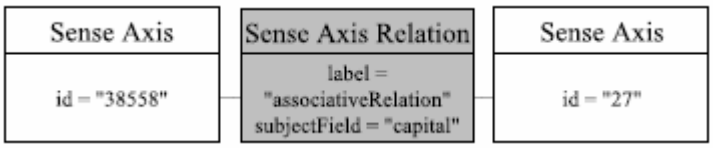


Figure 17 Relation of accessibility in ProlexLMF

The relations of synonymy and of meronymy can be represented in the same way. Furthermore, it is possible for the relation of synonymy to precise the diasystematic indicator via the data category: *usage*.

As previously mentioned, the class *Interlingual External Ref.* allows us to determine a relationship between a *Sense Axis* instance and an external system. We are able to preserve our typology and existence paradigm thanks to the following data categories: *external system* and *external reference*. Figure 18 shows that *City* is the type of the pivot *Paris*.

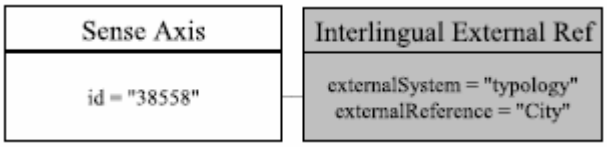


Figure 18 Example of a link towards an external system

4.2.4 Outline of Prolexbase adaptations to LMF

At first sight it can seem difficult to adapt the conceptual model of Prolexbase to LMF because of different concepts concerning lexical entries and the presentation of homonymy. Nevertheless, in spite of these differences, it is totally possible to adjust Prolexbase to the LMF standard. Simplistically, we can draw the following parallel between Prolexbase and LMF:

Prolexbase		**LMF**
Prolexeme	corresponds to	Sense
Pivot		Sense Axis
Type, existence		Interlingual External Ref.
Relation		Sense Axis Relation

Figure 19 represents the set of classes used in ProlexLMF:

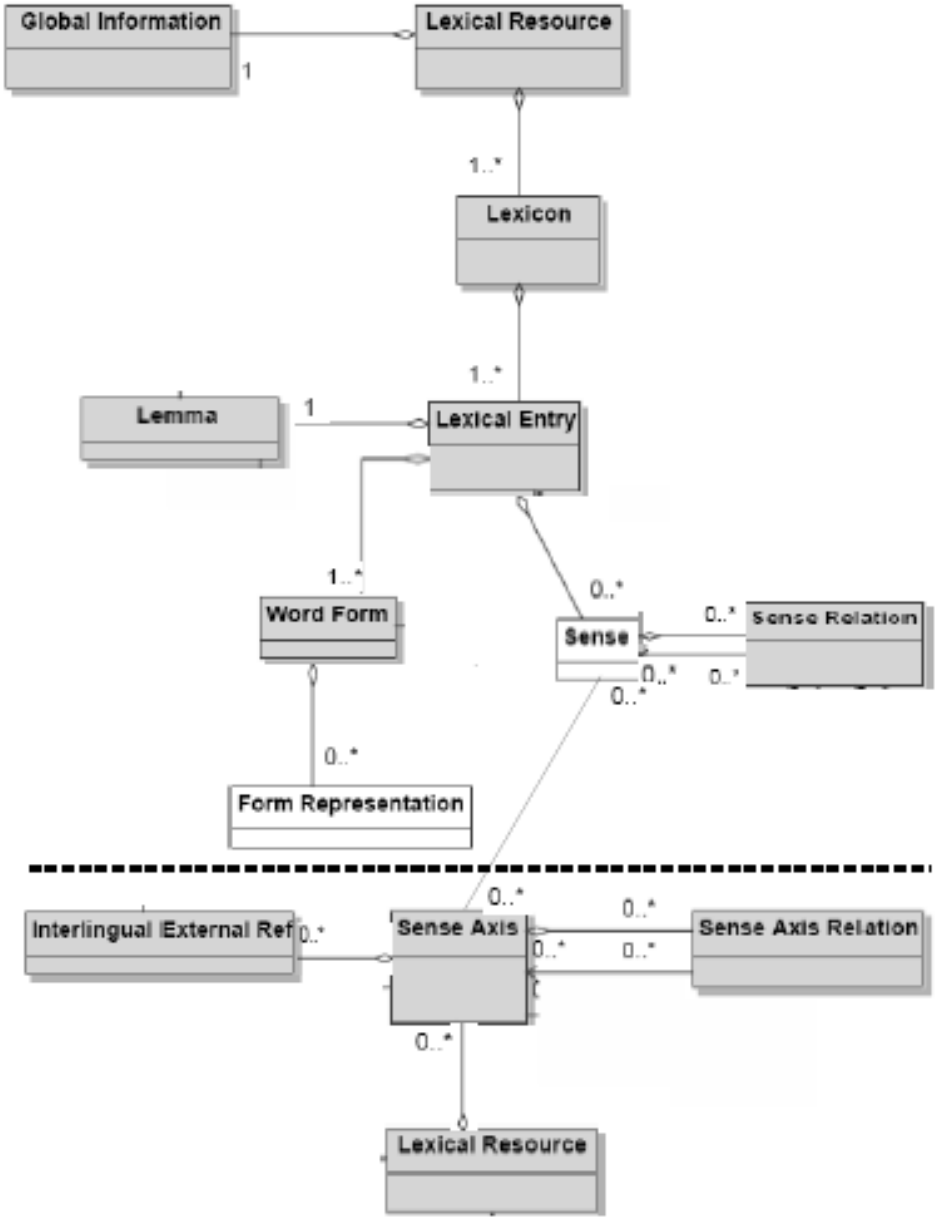


Figure 19 LMF classes used by Prolexbase

5 Summary

We have presented the project Prolex whose aim is to process automatically proper names in different languages. This project resulted, inter alia, in the creation of a multilingual dictionary of proper names, Prolexbase.

First of all, we showed the originality of Prolexbase which allows us to describe not only ontology and relations of proper names but also their morphology. The four-level ontology (metaconceptual, conceptual, linguistic and instances levels) is based on two main concepts:

the language independent pivot which represents a point of view about a referent, and the language dependent prolexeme which is a set of lemmas: name, its aliases and derivatives. The typology, language independent and language dependent relations as well as morphology proper to each language complete the description.

Secondly, we presented the structure (entities and relations) of our database. This relational database consists of nine language independent tables and thirty two language dependent ones.

Finally, we described the standard LMF whose goal is to provide a common model for lexical resources. We also pointed out that it is possible to have a version of Prolexbase that would be in keeping with LMF. According to the classical database design, we built a conceptual model, conforming to the LMF standard, which, in turn, has been translated into a logical model in order to efficiently store, maintain and use the dictionary of proper names. While a relational database should be used for the model implementation, an XML schema will come in useful for data exchange.

When completed, the database should cover most of the European languages. Such a multilingual relational database of proper names will be extremely useful and valuable in majority of NLP applications.

6 Bibliography of the project Prolex

Agafonov C., Grass T., Maurel D., Rossi-Gensane N., Savary A. (2006), La traduction multilingue des noms propres dans PROLEX, META, vol. 51/4, pp. 622-636.

Bouchou B., Maurel D. (2008), Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres, TAL, vol. 49 (1), pp. 61-88.

Bouchou B., Tran M., Maurel D. (2005), Towards an XML Representation of Proper Names and Their Relationships, *NLDB2005, in Lecture Notes in Computer Science*, 3513, pp. 44-55.

Friburger N. (2002), *Reconnaissance automatique des noms propres; application à la classification automatique de textes journalistiques*, thèse de doctorat en informatique, Université de Tours.

Grass T., Maurel D., Piton O. (2002), Description of a Multilingual Database of Proper Names, PorTal 2002, Faro, Portugal, 23-26 July, in *Lecture Notes in Computer Science*, pp. 2389: 137-140.

Krstev S., Vitas D., Maurel D., Tran M. (2005), Multilingual Ontology of Proper Names, Second Language & Technology Conference, Poznań, Poland, pp. 116-119.

Maurel D. (2008), Prolexbase: A Multilingual relational Lexical Database of Proper Names, *LREC2008*, pp. 334-338.

Maurel D. (2010), Dénomination et anaphore lexicale – Le réseau sémantique de Prolexbase, in Osu S., Col G., Garric N., Toupin F., Construction d'identité et processus d'identification, Peter Lang SA Edition scientifiques internationales.

Maurel D., Tran M., Grass T. (2004), An Ontology for Multilingual Treatment of Proper Names, in *OntoLex* 2004, *in Association with LREC2004*, pp. 75-78.

Maurel, D., Tran, M., Vitas, D., Grass, Th., Savary, A. (2006) Prolex: Implantation d'une ontologie multilingue des noms propres. Tours: Université François Rabelais, Rapport Interne n°279x50p.

Maurel D, Vitas D, Koeva S. (2007), Prolex: A Lexical Model for Translation of Proper Names. Application to French, Serbian and Bulgarian, *Bulag* 32, pp. 55-72.

Savary A. (2005), A Formalism for the Computational Morphology of Multiword Units, *Archives of Control Sciences, 15 (LI), Silesian University of Technology*.

Tran M., Maurel D. (2006), Prolexbase: Un dictionnaire relationnel multilingue de noms propres, *TAL*, 47(3), pp. 115-139.

Tran M., Maurel D., Savary A. (2005), Implantation d'un tri lexical respectant la particularité des noms propres, *Linguisticae Investigationes*.

Tran M., Maurel D., Vitas D., Krstev C. (2005), A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names, Papillon 2005 workshop on Multilingual Lexical Databases, in Association with the Sixth Symposium on Natural Language Processing, Thailand, pp. 2: 67-71.

Vaxelaire J.-L. (2005), *Les noms propres: une analyse lexicologique et linguistique*, H. Champion, Paris.

7 References

Ariel M. (1990), *Accessing Noun Phrases Antecedents*, Routledge, London.

Chinchor N. (1997), *Muc-7 Named Entity Task Definition*, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendice.

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, *Les Cahiers dia. Etudes sur la diachronie et la variation linguistique*, Belgium, vol. 1, pp. 9-16.

Courtois B., Silberztein M, (1990), Dictionnaires électroniques des mots simples du français, *Langue française* 87, pp. 11-22.

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

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

Jonasson K. (1994), *Le nom propre. Constructions et interprétations*, Ducelot, Paris.

Maurel D. (2004), Les mots inconnus sont-ils des noms propres?, *JADT2004*, Louvain-la-Neuve, Belgium, pp. 776-784.

Mikheev A., Moens M., Grover C. (1999), Named Entity Recognition without Gazetteers, *EACL99*, pp. 1-8.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. (1990), Introduction to WordNet : an on-line lexical database, *International Journal of Lexicography*, n°3, p. 235-244.

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

Romary L, (2000), *CLS Framework: Listing of ISO 12620 Data Categories*, http://www.ttt.org/clsframe/datcats.html.

Romary L. (2002), *The ISO 16642 document (draft), Version ISO/TC 37/SC 3*, http://www.loria.fr/projets/TMF/tmf.html.

Sager J.C. (1990), *A Practical Course in Terminology Processing*, John Benjamins, Amsterdam.

Savary A. (2005), MULTIFLEX: Users Manual and Technical Documentation, version 1.0, Rapport interne du Laboratoire d'Informatique de l'Université de Tours, n°285.

Temmerman R. (2003), The Ontology Shift in Terminography, Seminar on Multilingual Terminography: *Towards Intelligent Dictionaries*?