# Prolexbase: A multilingual relational lexical database of proper names

**Denis MAUREL**
Université François Rabelais Tours
Laboratoire d'informatique
France
E-mail: denis.maurel@univ-tours.fr

**Abstract**

This paper deals with a multilingual relational lexical database of proper name, Prolexbase, a free resource available on the CNRTL website.
The Prolex model is based on two main concepts: firstly, a language independent pivot and, secondly, the prolexeme (the projection of the pivot onto particular language), that is a set of lemmas (names and derivatives). These two concepts model the variations of proper name: firstly, independent of language and, secondly, language dependent by morphology or knowledge. Variation processing is very important for NLP: the same proper name can be written in different instances, maybe in different parts of speech, and it can also be replaced by another one, a lexical anaphora (that reveals semantic link).
The pivot represents different referent's points of view, i.e. language independent variations of name. Pivots are linked by three semantic relations (quasi-synonymy, partitive relation and associative relation).
The prolexeme is a set of variants (aliases), quasi-synonyms and morphosemantic derivatives. Prolexemes are linked to classifying contexts and reliability code.

## 1. The Prolex project

From the MUC Conferences and its Named Entity Task, proper names are a challenge for NLP applications. If the use of lexical database is not advised for recognizing proper names by (Mikheev et al., 1999), it is not almost the case for other tasks, as spelling or translation aid, multilingual alignment, lexical anaphora resolution…
The *Prolex project* was initiated in 1990s, in order to process proper names, first with the study of toponyms in French and second with the development of a Serbian version. Then, a relational multilingual dictionary of Proper Names, *Prolexbase*, in the form of relational database, was designed and constructed (Krstev et al., 2005; Tran & Maurel, 2006). From June 2007, this resource is free and available on the CNRS resource website [1] (CNRTL) in XML format (Maurel, 2008). Finally, we are working on a new version of this resource, using the TMF[2] and the LMF[3] ISO standard. The TMF version will present the database as it is described in this paper (ordered by pivots); the LMF version will propose a dictionary of proper names and derivatives (ordered by lemmas).
Today, Prolexbase contains essentially proper names in French, but also some translations in other languages, almost for Serbian. 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 mainly selected these entries from junior high school dictionaries, during a project supported by the French Ministry of Industry[4].

[1] http://www.cnrtl.fr/lexiques/prolex/.
[2] ISO 16642:2003, Computer applications in terminology - Terminological Markup Framework (TMF), http://www.loria.fr/projets/TMF/.
[3] ISO/TC 37/SC 4, Language resource management - Lexical markup framework (LMF), http ://lirics.loria.fr/documents.html, 2007.

## 2. The Prolex model of proper names

The Prolex model is based on two main concepts: firstly, a language independent *pivot* and, secondly, 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. These two concepts model the variations of proper name: firstly, independent of language and, secondly, language dependent by morphology or knowledge. Variation processing is very important for NLP: the same proper name can be written in different instances, maybe in different parts of speech, and it can also be replaced by another one, a lexical anaphora (that reveals semantic link).

### 2.1 The pivot definition

To define the *pivot*, we use the quasi-synonymy relation completed by diasystematic features of Coseriu (1998). A *pivot* is a diachronic, diastratic or diaphasic referent's point of view, i.e. a language independent variation of a name:

1. Diachronic: a name has sometimes changed because of the history of the country, for instance *Petersburg*, *Petrograd* and *Leningrad* in Russia or *Burma* and *Union of Myanmar*.
2. Diastratic: a famous person can have more than one name, but generally not with the same fame, for instance, 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*. And if the American singer-songwriter *Bob Dylan* is well-known, how many people know his real name, *Robert Allen Zimmerman*?
3. Diaphasic: for instance, a tour operator prefers use the

[4] http://www.technolangue.net/article.php3?id_article=155.

name *Town of Light* instead of *Paris* and a political discourse often uses the system of government to speak about a country, such as *Kingdom of Morocco*, versus *Morocco*.

## 2.2 The prolexeme definition

To define the *prolexeme*, we distinguish between three types of language dependent variation:

1. The name and its written form aliases:
   full form (*United Nations Organization*)
   short form (*United Nations*)
   initialism (*UNO*)
   acronym
   orthographic variant
   transcribed form (*Marat Safin*)
   transliterated form
   Romanized form (in Serbian, *Belgrade* is written Београд or *Beograd*)…
2. Quasi-synonyms:
   diastratic, by a specific knowledge not shared by foreign countries
   diatopic, by a local language
   explanation (the *Caritas USA Organization* to explain *Catholic Relief Services*);
3. Derivatives, obtained by a morphosemantic derivation with a regular form-meaning sense, link to the proper name, as (Fellbaum & Miller, 2003):
   relational name (*Onusian.N*)
   relational adjective (*Onusian.A*)
   etc. (depend of the language).

We call the set of these language dependent variations, the *prolexeme*. And we associate to each lemma inflectional rule to generate all its inflected forms. These are also put in the instance table of the database.

For instance, the Figure 1 presents the pivot 48 226:

Prolexeme-eng$_{UNO}$={United Nations Organization, United Nations, UNO, Onusian.N, Onusian.A}

Instances-eng$_{UNO}$={United Nations Organization, United Nations, UNO, Onusian.N, Onusians, Onusian.A}

There are only six instances, because the English morphology is very poor… In French, it is eleven and in Serbian, more than fifty!

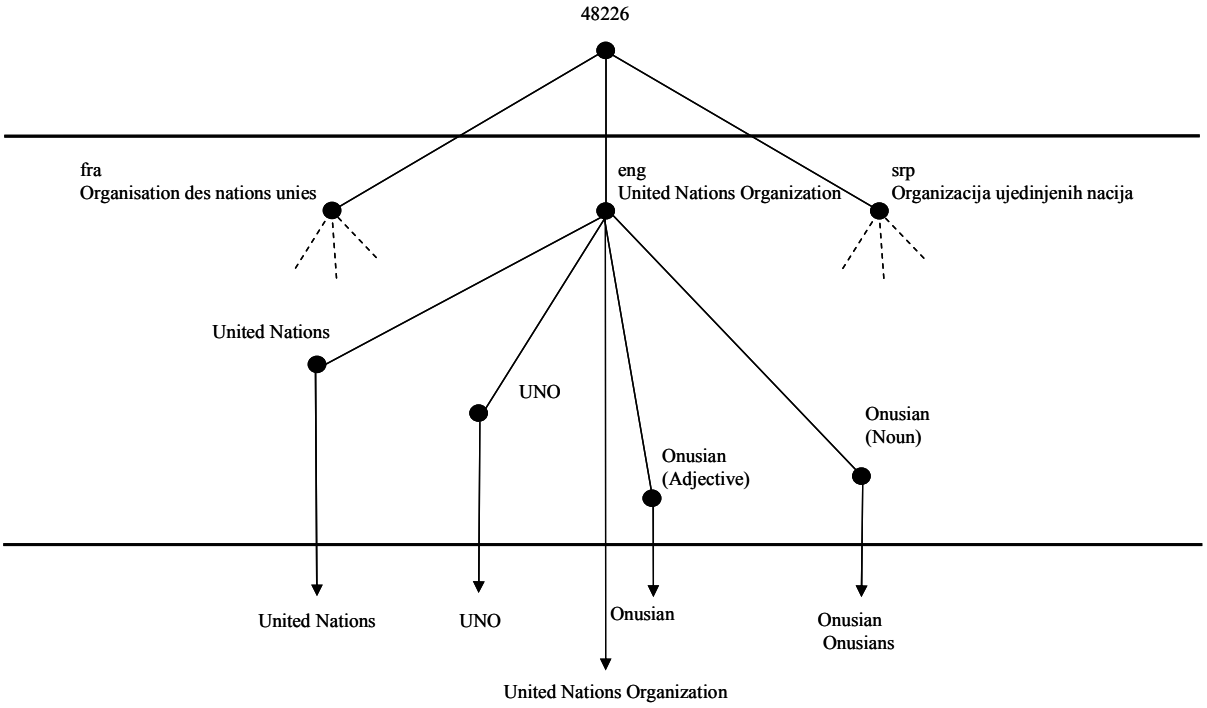


Figure 1: Pivot, prolexemes and instances from *United Nations Organization*

## 3. On the language independent variations

As we have said before, we simply use a pivot, as in many lexical databases: *EuroWordnet* (Vossen, 1998) and *Balkanet* (Tufiş et al., 2004), *Papillon* (Mangeot-Lerebours et al., 2003)…

Three semantic relations between these pivots make possible anaphora:

1. The quasi-synonymy (see, section 2.1);
2. The partitive relation:
   Firstly, the meronymy of toponyms or events:
   *Morocco⊂ Maghreb ⊂ North Africa*
   *Operation Torch ⊂ Second World War*
   Secondly, its extension to other contexts:
   *EADS ⊂ Europe*
   *Psalms ⊂ Bible*
   *Al Gore ⊂ USA*…
3. The associative relation:
   Relative: *Irène Joliot-Curie* is the daughter of *Marie Curie*
   Capital: *Rabat* is the capital of *Morocco*
   Politician: *Gordon Brown* is an *English* politician
   Creator: *The Heroic Symphony* is an opera of *Beethoven*
   etc.

The associative relation replaces in dictionaries the definition of common nouns and allows the accessibility of the name (Ariel, 1990).

We define also two generic relations to tag each pivot with two features from:

1. A limited typology of thirty types and nine super types (see Table 1). The first level is shared between four super types: Anthoponym (human feature), Toponym (locative), Ergonym (artifact) and Pragmonym (event feature). Types and super types are also in relation of hyperonymy. For instance, *UNO* is an *Organization*, is a *Group*, is a *Collective anthroponym*, is an *Anthroponym*, is a *Proper name*. And, secondarily, it is also a *Toponym* and an *Ergonym*.

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

Table 1: The Prolex typology

2. Three values of existence (historical, fictitious and religious). Religious feature does not tag religious man or event, but name from the religious belief, for instance the *archangel Gabriel*… These features help the translation: Fictitious and religious names are often translated, as *Snow White* in English and *Blanche-Neige* in French.

Figure 2 presents these relations between pivots.

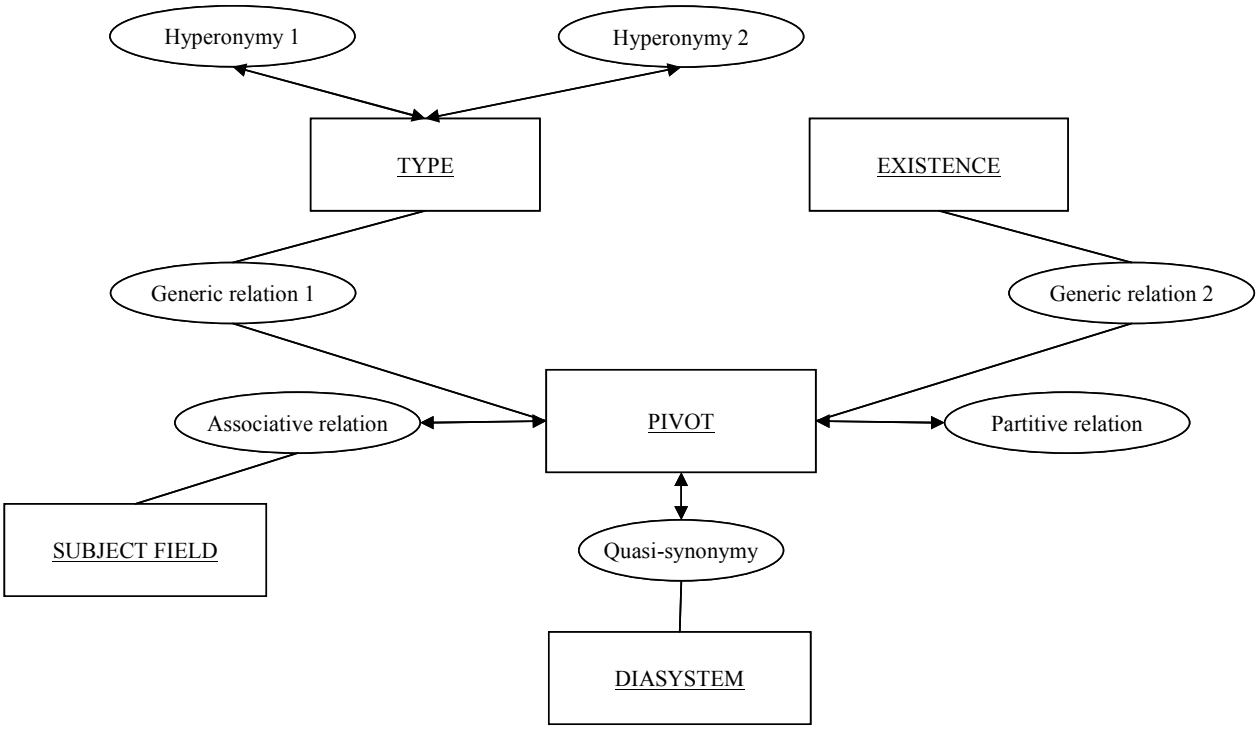


Figure 2: The Prolex language independent relations

## 4. On the language dependent variations

As we have seen section 2.2, the prolexeme is the set of the language dependent variations. The lemmas of the prolexeme are linked to inflectional paradigms and we use finite-state transducers to generate instances: more precisely, we use the Unitex system (Paumier, 2003) and the Multiflex system (Savary, 2005).

Some anaphora result of the classifying context (*capital*, *king*, *river*…) that we also note here and which is often useful for translation: for instance, the name *Loire* in French is translated *Loire River* in English.

At the opposite direction of preceding relations, the relation of eponymy points out that translation does not refer to proper name but to common noun (antonomasia: *a Rembrandt* for an artist…), terminological terms (*Parkinson's disease*...) or idiom (*not for all the tea in China*…).

We have added also to each prolexeme a reliability code with the three features advised by ISO 12620 (Computer applications in terminology - Data categories): *commonly used*, *infrequently used* and *rarely used*. And we indicate if the name is or not constructed with a determiner.

Figure 3 presents these relations.

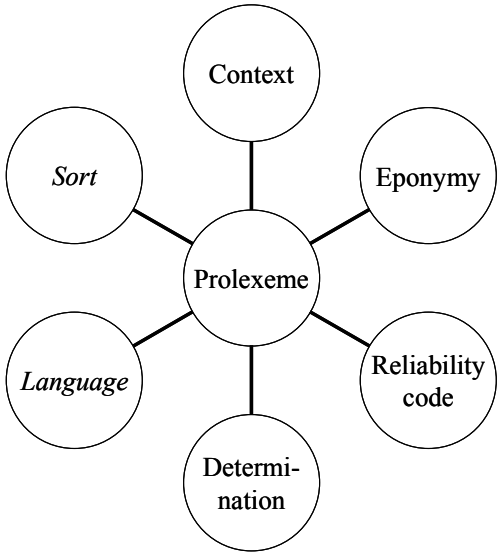


Figure 3: The Prolex language dependent relations

## 5. Conclusion

We have presented a lexical database of proper names (and derivatives). This database is multilingual and relational. We can now complete the number of entries given section 1: Today, the French part of Prolexbase contains 54 774 proper names, 730 aliases and 20 614 derivatives. It contains also 50 567 relations: 2 249 associative relations, 47 670 partitive relations and 648 quasi-synonymies.

Let us show on a last example of name translations from French to English, with the sentence:

*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*
  [Prolexeme] ⇨ *Tours*
  [Morphosemantic] ⇨ Derivative (Relational noun)
  [Possible mining] ⇨ inhabitant
  [Classifying context] ⇨ city
  [Partitive relation] ⇨ *France*
- *Loire*
  [Prolexeme] ⇨ *Loire*
  [Classifying context] ⇨ river

In prospect of this work, we will increase coverage of Prolexbase with new entries and new languages. And, as we have said section 1, we are working in ISO format.

We are also working on a named entity tagger for French transcribed speech that uses the CasSys system, a transducer cascade (Friburger & Maurel, 2004).

We project to build a proper name processing platform using the database and the Prolex model, first to compare and align multilingual texts.

## 6. References


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

Coseriu E. (1998), Le double problème des unités dia-s, *Les Cahiers δια. Etudes sur la diachronie et la variation linguistique* 1, pp. 9-16.

Fellbaum C., Miller G. A. (2003), Morphosemantic Links in WordNet, *TAL*, 44(2) , pp. 69–80.

Friburger N., Maurel D. (2004), Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

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

Mangeot-Lerebours M., Sérasset G., Lafourcade M. (2003), Construction collaborative d'une base lexicale multilingue, le projet Papillon, TAL, 44(2) , pp. 151-176.

Maurel D. (2008), Prolexbase : Une base de données lexicale de noms propres pour le Tal, Colloque *Lexicographie et informatique : bilan et perspectives*, (Actes p. 137-144), Nancy, 23-25 janvier.

Mikheev A., Moens M., Grover C. (1999), Named entity Recognition without Gazetteers, *EACL'99*, pp. 1-8.

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

Savary A. (2005), Towards a Formalism for the Computational Morphology of Multi-Word Units, Second Language & Technology Conference, pp. 305-309, Poznan, Poland.

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

Tufiş D., Cristea D., Stamou S. (2004), BalkaNet: Aims, Methods, Results and Perspectives. A General Overview, *Romanian journal of Information science and technology*, 7-1-2, pp. 9-44.

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