

Prolexbase 1.1

Sommaire

1 Introduction	1
2 Contributions et remerciements	2
3 Bibliographie	2
4 La modélisation des noms propres dans Prolexbase	3
4.1 Les concepts.....	3
4.1.1 Le pivot.....	3
4.1.2 Le prolexème.....	4
4.1.3 Les alias et dérivés.....	4
4.1.4 Les instances.....	5
4.2 Les relations.....	6
4.2.1 La relation de synonymie.....	6
4.2.2 La relation de méronymie.....	7
4.2.3 La relation d'accessibilité.....	7
4.2.4 La relation d'expansion classifiante.....	8
4.2.5 L'éponymie.....	8
4.3 Les relations d'hyponymie.....	8
4.3.1 La typologie des noms propres de Prolexbase.....	8
4.3.2 La typologie d'existence de Prolexbase.....	9
5 La version LMF de Prolexbase	9
5.1 Les catégories de données.....	9
5.2 Les classes LMF utilisées.....	11
5.3 Les formes.....	12
5.4 Les sens.....	15
5.5 L'extension multilingue.....	17
5.6 Représentation de Prolexbase par LMF.....	21
6 Couverture linguistique de Prolexbase 1.1	23

1 Introduction

Prolexbase est un lexique relationnel multilingue de noms propres. La version 1 comprend essentiellement du français, mais aussi quelques traductions, [consultables en ligne](#). Cependant la version téléchargeable est uniquement française pour l'instant.

La modélisation du domaine des noms propres définie dans le projet Prolex repose sur deux concepts centraux : le *pivot* et le *prolexème*. Le pivot ne représente pas le référent, mais un point de vue sur ce référent. Il possède dans chaque langue un concept spécifique, le prolexème, qui est une famille structurée de lexèmes. Autour d'eux, sont définis d'autres concepts et des relations (synonymie, méronymie, accessibilité, éponymie, etc.). Chaque pivot est en relation d'hyponymie avec un type et un paradigme d'existence.

Il n'est pas évident de définir la notion de nom propre. La plupart des définitions insistent sur le caractère unique de son référent et sur une sémantique et une syntaxe qui lui est propre.

Nous avons choisi d'adopter le point de vue de (Jonasson, 1994) qui propose une définition plus large incluant ce qu'elle appelle les noms propres purs (noms de personne et noms de lieu) et les noms propres descriptifs qui résultent souvent de la composition d'un nom propre avec une expansion (Tour Eiffel, musée Rodin, etc.). Un nom propre descriptif peut être considéré comme une expression définie figée ou en cours de figement (Jardin des Plantes, Médecins sans frontières, etc.). Cette définition est assez proche de celle utilisée dans le domaine du Tal depuis la conférence MUC6.

Alors que la version 1.0 se présentait sous une forme plutôt "terminologique", puisque les entrées correspondaient aux pivots, la version 1.1 adopte les recommandations de la norme [LMF \(ISO-24613:2008\)](#) et se présente donc sous une forme dictionnaire où les entrées sont des lemmes.

Prolexbase 1.1 est en accès libre sous acceptation de la licence LGPL-LR. La maintenance et la mise à jour du lexique sont assurées par le [Laboratoire d'informatique](#) (LI) de l'[université François Rabelais Tours](#), l'hébergement, par le CNRTL (ATILF).

2 Contributions et remerciements

Prolexbase 1.1 est [un projet Tal](#) du LI, en collaboration avec :

- Le groupe de recherche *Langues et Représentation* (L&R) de l'université François-Rabelais.
- L'université de Belgrade.

La conception de la base est principalement due à Denis Maurel (LI), Mickaël Tran (thèse au LI), Thierry Grass (L&R), Duško Vitas (université de Belgrade), Agata Savary (LI) et Béatrice Bouchou (LI) ; son implémentation à Denis Maurel, Mickaël Tran, Marie Ndiaye et Coralie Villes ; son installation sur le site du CNRTL à Etienne Petitjean (ATILF). La version 1.1 a été conçue par Béatrice Bouchou et Denis Maurel, avec l'aide d'Estelle Leton pour son implémentation.

Prolexbase 1.1 a bénéficié de l'aide et des contributions de [tous les participants](#) au projet Prolex.

Ce projet a reçu le soutien :

- De l'action [Technolangue](#) du Ministère de l'Industrie.
- Du programme d'action intégré [Egide](#) Pavle-Savic du Ministère des Affaires étrangères.

3 Bibliographie

Principalement :

Tran M. (2006), [Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne](#), Thèse de doctorat d'informatique, Université François Rabelais Tours.

Tran M., Maurel D. (2006), [Prolexbase : Un dictionnaire relationnel multilingue de noms propres](#), Traitement automatique des langues, Vol. 47(3):115-139.

Bouchou B., Maurel D. (2008), [Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres](#), Traitement automatique des langues, 49(1):61-88.

Voir aussi la [bibliographie complète du projet Prolex](#).

4 La modélisation des noms propres dans Prolexbase

L'ensemble de notre modèle se présente sous la forme d'une arborescence qui peut se décomposer en deux niveaux distincts :

- Le niveau qui ne dépend pas de la langue : le pivot et certaines relations ; la typologie des noms propres et la typologie d'existence.
- Le niveau qui dépend de la langue : le prolexème, les alias, les dérivés et certaines relations ; les instances (ensemble des formes fléchies d'une langue).

Après avoir présenté les concepts clés de notre modélisation des noms propres, nous décrirons les relations existant dans notre dictionnaire, puis nos typologies.

4.1 Les concepts

4.1.1 Le pivot

Pour une langue donnée, des noms propres totalement différents sur le plan graphique peuvent renvoyer à un même et unique référent et ce phénomène se retrouve généralement d'une langue à l'autre. Par exemple, les noms propres *Jean-Paul II* et *Karol Jozef Wojtyla* en français correspondent tous les deux à un certain point de vue sur un même et unique référent et il en est de même en anglais (*John Paul II* et *Karol Jozef Wojtyla*), en italien (*Giovanni Paolo II* et *Karol Jozef Wojtyla*), etc.

Nous définissons donc le pivot non pas comme le référent, mais, plutôt, comme un certain point de vue sur celui-ci. Il sera représenté dans la base par un numéro identifiant unique. Ainsi les noms propres *Pologne* en français, *Polonia* en espagnol, *Polen* en allemand, etc. seront associés à un même pivot, tandis que les noms propres *République de Pologne* en français, *República Polaca* en espagnol, *Republik Polen* en allemand, etc. seront associés à un autre pivot. Ces deux pivots seront en relation de synonymie.

Pour définir ces différents points de vue, nous nous sommes basés sur un marquage diasystématique, provenant des travaux sur la métalexigraphie d'Eugenio Coseriu qui propose un diasystème basé essentiellement sur quatre variétés distinctes : diachronique (variété dans le temps), diaphasique (variété concernant les finalités de l'emploi), diatopique (variété dans l'espace) et diastratique (variété relative à la stratification socio-culturelle).

4.1.2 Le prolexème

Dans notre modèle, le prolexème correspond à une projection du pivot dans une langue donnée. Chaque prolexème d'une langue donnée sera donc relié à un seul et unique pivot. C'est en se basant sur cette relation que l'on va pouvoir traduire les prolexèmes d'une langue vers une autre. Le concept de prolexème peut aussi se définir comme une classe d'équivalence.

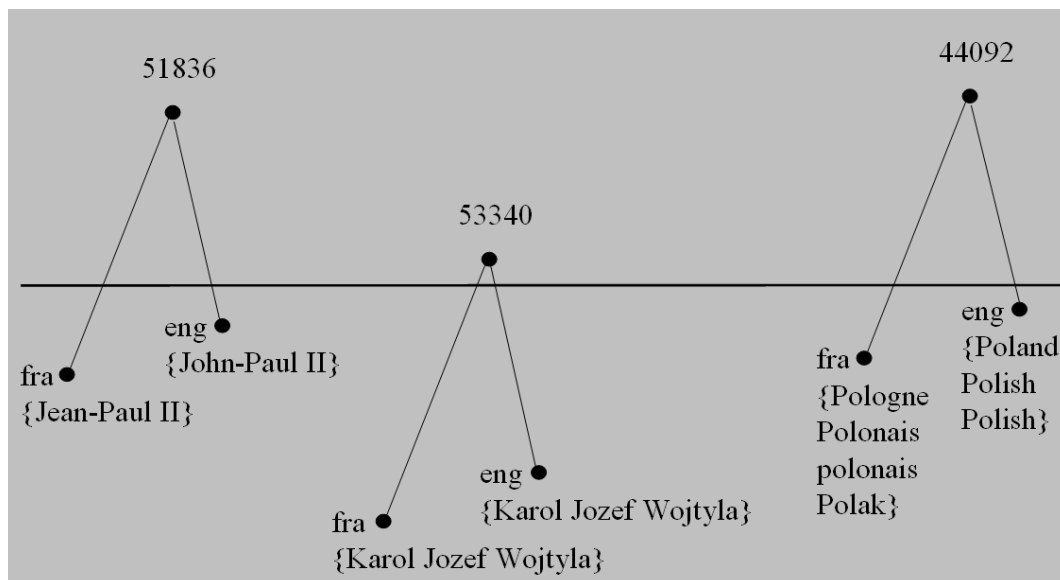


Figure 1 : Pivots et prolexèmes

Pour simplifier, nous considérons aussi le prolexème comme le lemme associé aux différentes formes d'un nom propre qui apparaissent dans les différents textes d'une langue donnée. Il peut ainsi être considéré comme la forme vedette d'un ensemble de dérivés et d'alias.

Les noms propres polysèmes, qui sont classés sous des catégories différentes, seront reliés à des prolexèmes différents. Par exemple, *Verdun* est à la fois connu comme étant une célèbre bataille durant la Première Guerre Mondiale, comme un traité entre les trois fils de l'empereur Louis le Pieux pour partager son Empire et, enfin, comme le chef-lieu de la Meuse. Pour ce cas-là, nous serons amenés à créer trois prolexèmes différents. Par contre, dans le cas de toponymes correspondant à la fois à un lieu et une entité administrative (comme par exemple *Paris* qui est à la fois une ville et un département), nous avons décidé de ne pas dupliquer les prolexèmes pour éviter l'abondance d'homographes. Cette information sera rajoutée au niveau des expansions classifiantes du prolexème.

Les noms propres homographes seront aussi associés à des prolexèmes différents. En recherchant le nom propre *Sydney* dans un dictionnaire, on trouvera deux entrées distinctes : une qui correspondra à une ville en Australie et l'autre à une ville située au Canada. Il est à noter que l'homonymie dépend de la langue. Par exemple, en anglais, le nom propre *London* correspond au chef-lieu de l'île Kiritimati ou à la capitale de l'Angleterre, ce qui n'est pas le cas en français à cause de l'existence d'un exonyme (*Londres*).

4.1.3 Les alias et dérivés

Nous définissons les alias comme des synonymes qui dépendent de la langue. Nous avons regroupé dans le terme d'alias, d'une part, des synonymes exacts (les variantes d'écriture

-caractères, abréviations, acronymes et sigles, transcriptions-, les variantes orthographiques et, d'autre part, des quasi-synonymes, diatopiques ou diastratiques.

Par exemple, les noms *Nations Unies*, *Onusien*, *ONU* auront *Organisation des Nations Unies* comme prolexème pour la langue française. Les noms *United Nations* et *UNO* auront pour prolexème *United Nations Organization* pour la langue anglaise. Le prolexème français *Organisation des Nations Unies* et le prolexème anglais *United Nations Organization* seront reliés à un même pivot (48226).

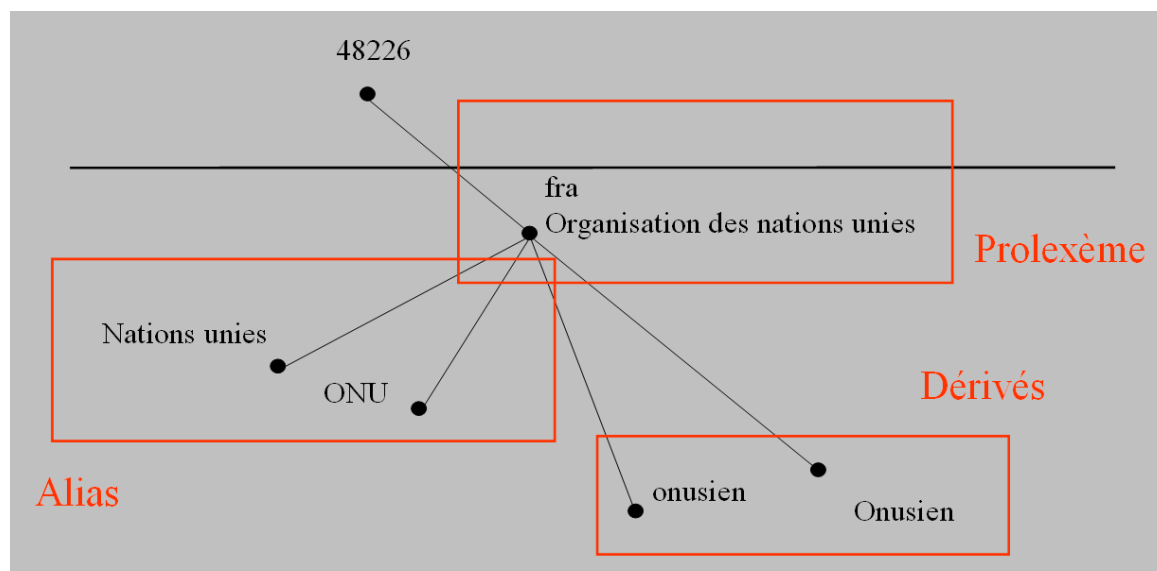


Figure 2 : Alias et dérivés

Les dérivés présents dans Prolexbase sont des dérivés morphosémantiquement liés aux prolexèmes. Ainsi *polonais* y figure (*le président de Pologne* versus *le président polonais*), mais pas *pasteuriser* (qui est lexicalisé dans un sens très précis et qui ne peut se paraphraser à partir du prolexème : **faire comme Pasteur faisait*).

Le classement des alias et des dérivés dans la partie qui dépend de la langue s'explique notamment par la raison que la créativité lexicale est propre à chaque langue. Une variante d'écriture ou un dérivé existant dans une langue L_1 peut être totalement absent dans une langue L_2 . Par exemple, le dérivé *Tourangeau* se traduira en anglais par *inhabitant of Tours*.

4.1.4 Les instances

Les instances sont toutes les formes fléchies que l'on peut obtenir à partir d'un nom propre, d'un de ses alias ou d'un de ses dérivés. La version téléchargeable de Prolexbase 1.1 ne contient que du français, des noms, adjectifs ou préfixes ; chaque forme nominale ou adjectivale sera donc accompagnée de son genre et de son nombre.

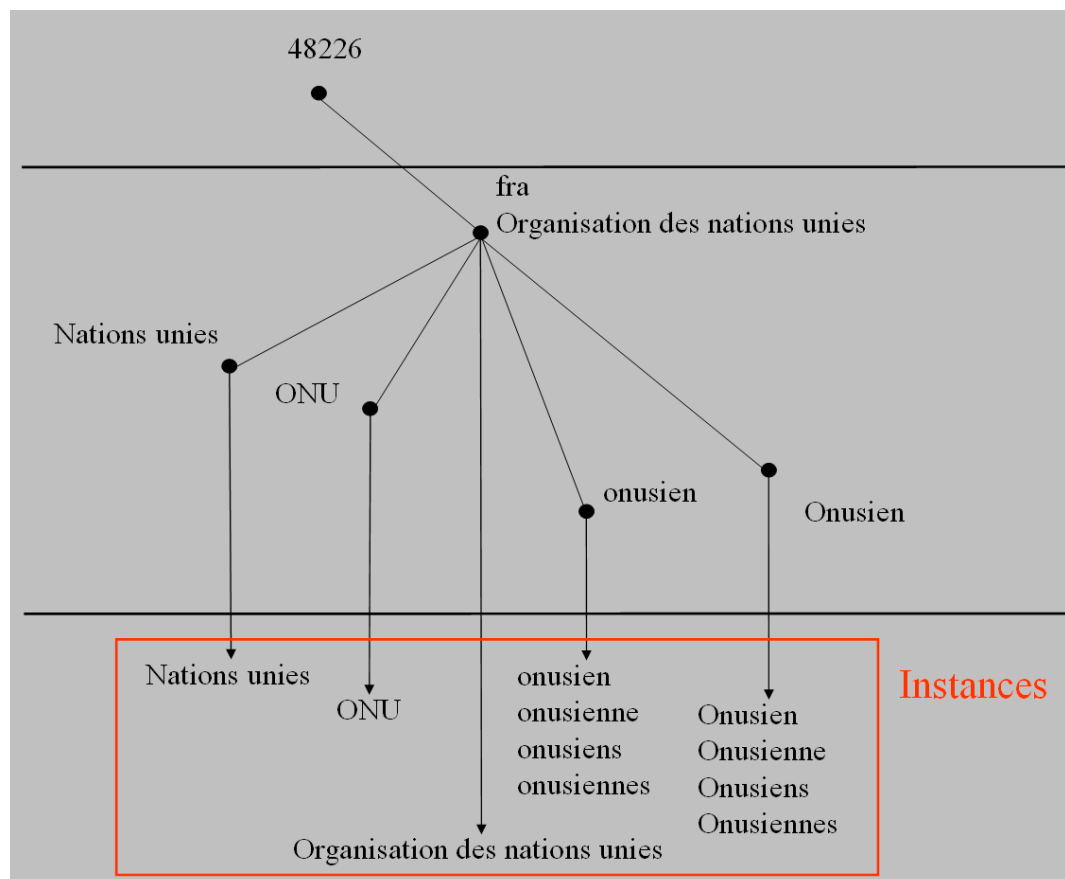


Figure 3 : Instances

4.2 Les relations

Dans le cas des relations qui ne dépendent pas de la langue, nous avons retenu pour Prolexbase trois relations paradigmatiques (méronymie, synonymie, hyperonymie) et une relation syntagmatique (accessibilité). La relation d'hyperonymie correspond, dans Prolexbase, à une typologie des noms propres et à un paradigme d'existence que nous présenterons plus loin.

Dans le cas des relations qui dépendent de la langue, nous avons retenu deux relations syntagmatiques : l'expansion classifiante (collocation libre) et l'éponymie (collocation figée). Mais celles-ci ne sont [consultables qu'en ligne](#).

4.2.1 La relation de synonymie

Dans une synonymie, l'un des termes est souvent préférable à l'autre. On appellera le premier la forme canonique et l'autre la forme synonyme. Cette forme canonique en général correspond à la forme la plus connue. Nous avons considéré la variation diatopique comme un alias (voir plus haut). Il reste donc, au niveau indépendant de la langue, les trois variations : diachronique (*Zaïre* et *République démocratique du Congo*), diastratique (*Molière* et *Jean-Baptiste Poquelin*) et diaphasique (*Paris* et *Ville Lumière*).

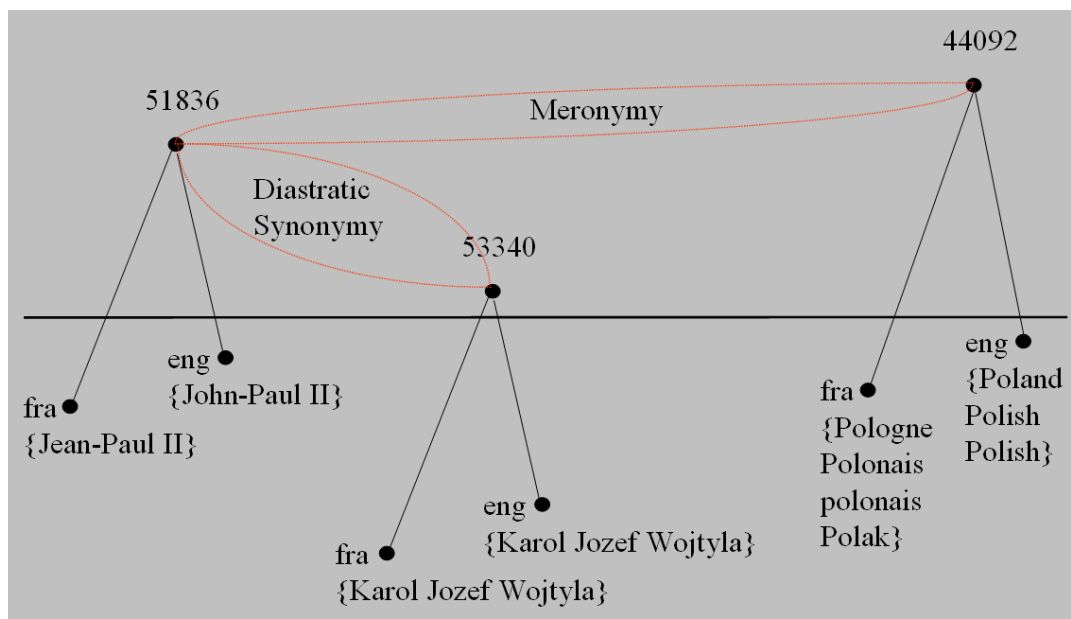


Figure 4 : Relations qui ne dépendent pas de la langue

4.2.2 La relation de méronymie

La relation de méronymie permet d'établir une hiérarchisation sur plusieurs niveaux entre les éléments contenant (holonymes) et les éléments contenus (méronymes). Lorsque deux unités lexicales A et B sont en relation de méronymie, on dit que A est un méronyme de B (et que B est un holonyme de A) si A est une partie de B.

Nous entendons cette relation au sens large, par exemple une célébrité et son pays, un roi et sa dynastie, un personnage et une œuvre, des toponymes entre eux, des évènements entre eux, etc.

4.2.3 La relation d'accessibilité

Un nom propre dans un dictionnaire n'est pas associé à une définition classique, mais à une description encyclopédique, faite avec d'autres noms propres sur lesquels se base son accessibilité. Cette relation pourrait en fait correspondre à une multitude de relations, mais cela risquerait d'être coûteux et de nuire à la lisibilité du modèle, d'autant que certaines expansions existent dans une langue et sont absentes dans d'autres langues. Par exemple, en France, nous faisons la distinction entre un chef-lieu, une préfecture, une capitale, etc. Nous avons donc décidé de les regrouper dans une seule et unique relation, à laquelle nous avons ajouté des repérages généraux. Les informations spécifiques sont conservées dans la relation d'expansion classifiante (voir ci-dessous). Cette relation d'accessibilité n'est pas une modélisation idéale, mais correspond à une solution économique, suffisante pour la plupart des applications de Tal.

Donnons quelques exemples de repérages qui peuvent apparaître dans le cadre d'une relation d'accessibilité :

- Parent : les personnes et les membres de leur famille. *Joseph* est la mère de *Jésus*, *Henri II* est le fils de de *François I^{er}*, etc.

- Créateur : les auteurs et les œuvres. *Molière* est l'auteur de *Tartuffe*, etc.
- Capitale : les toponymes et leurs capitales. *Tours* est le chef-lieu de l'*Indre et Loire*, *Bangkok* est la capitale de la *Thaïlande*, etc.
- Dirigeant politique : les hommes politiques et les pays. *Jacques Chirac* est un homme politique *français*, etc.
- Siège : les entreprises, associations ou organisations et le toponyme correspondant au siège social. *Peugeot* est une firme *sochaliennne*, etc.
- Etc.

4.2.4 La relation d'expansion classifiante

Cette relation, qui n'est [accessible qu'en ligne](#), associe à un prolexème une expansion qui peut apparaître soit à sa gauche, soit à sa droite. Toutes les expansions qui existent dans une langue ne se retrouvent pas forcément dans une autre langue. Si l'expansion d'un nom propre est omise dans un texte, il est parfois nécessaire de la rétablir lors de la traduction de celui-ci, afin d'apporter un complément d'information au lecteur. Ainsi, le nom propre *la Loire* devient en anglais *the Loire River*. Comme il a été dit ci-dessus, certaines expansions complètent la notion de repérage associée à la relation d'accessibilité entre deux noms propres.

4.2.5 L'éponymie

Les noms propres apparaissent parfois dans les textes sous la forme de simples substantifs. Cette possibilité existe dans un grand nombre de langues, mais, pour un nom propre donné, dépend de la langue considérée. La relation d'éponymie, qui n'est [accessible qu'en ligne](#), est donc la relation entre un nom propre et une forme lexicalisée, soit dans le vocabulaire courant (*un tartuffe*), soit dans une expression idiomatique (*tous les chemins mènent à Rome*) ou terminologique (*maladie de Parkinson*). Contrairement aux autres relations, l'objectif de la prise en compte de la relation d'éponymie est d'empêcher une reconnaissance abusive des noms propres dans des textes.

4.3 Les relations d'hyponymie

4.3.1 La typologie des noms propres de Prolexbase

La typologie des noms propres de Prolexbase est hiérarchisée par une relation d'hyponymie qui a pour racine le concept de nom propre, pour nœuds des supertypes et pour feuilles des types.

Situés juste en dessous du concept de nom propre, les quatre premiers supertypes classent les noms propres suivant des traits syntaxico-sémantiques assez généraux : les *anthroponymes* (trait humain), les *ergonymes* (les fabrications humaines - trait inanimé), les *pragmonymes* (trait événement) et les *toponymes* (trait locatif).

Nous avons partagé le supertype anthroponyme en deux autres supertypes : les *anthroponymes individuels* et les *anthroponymes collectifs*.

Les types correspondent à une classification plus détaillée, mais volontairement limitée. Dans Prolexbase 1.1, nous avons retenu au total trente types (voir Figure 5). Comme certaines distinctions sont difficiles à réaliser et peuvent sembler arbitraires, nous avons créé deux autres supertypes : un supertype hyponyme des *anthroponymes collectifs*, les *groupements*, et un supertype hyponyme des *toponymes*, les *territoires*.

Nom propre							
Anthroponyme			Toponyme		Ergonyme	Pragmonyme	
Individuel	Collectif			Territoire			
Personne Patronyme Prénom Pseudo Anthroponyme	Dynastie Ethnonyme	Groupement		Astronyme Edifice Géonyme Hydronyme Ville Voie			Pays
		Association Ensemble Entreprise Institution Organisation					Région Supranational
				Objet Œuvre Pensée Produit Vaisseau	Catastrophe Fête Histoire Manifestation Météorologie		

Figure 5 : Les types de Prolexbase

Dans Prolexbase 1.1, chaque pivot est en relation d'hyponymie avec un type. Donc seuls ceux-ci apparaissent.

4.3.2 La typologie d'existence de Prolexbase

L'hyponymie d'*existence* relie les pivots à l'une des trois instances :

- Historique : Etre un nom propre du domaine historique (*Mozart, le Danube, Paris*, etc.).
- Religieux : Etre un nom propre du domaine de la croyance (*Zeus, Adam, Gabriel*, etc.).
- Fictif : Etre un nom propre du domaine de la fiction (*Don Quichotte de la Manche, Eldorado*, etc.).

La distinction entre les noms propres historiques et les autres s'avère utile pour la traduction, car, dans la majorité des cas, ces derniers possèdent des traductions distinctes d'une langue à l'autre. Par exemple, *Don Quijote de la Mancha* en espagnol devient *Don Quichotte de la Manche* en français et *Don Quixote of La Mancha* en anglais.

5 La version LMF de Prolexbase

5.1 Les catégories de données

Si les classes et leurs liens sont normalisés par LMF, il n'en est pas de même pour les attributs que l'on souhaite leur attacher. Cependant, il est recommandé de suivre autant que faire se peut la norme [ISO 12620](#) qui « spécifie les catégories de données utilisées pour

l'enregistrement de l'information terminologique [...] ainsi que pour l'échange et la recherche d'information terminologique ». Cette norme est complétée par [Francopoulo et al., 2008] et par quelques ajouts de notre part. Le tableau de la Figure 6 présente la correspondance entre les termes de Prolexbase et les catégories de données.

Prolex	Data Categories		
	Position	Français	Anglais
Langue	A10.7	indicatif de la langue	language symbol
Prolexème, alias et dérivés	A.2.1.1	entrée principale	main entry term
	A.2.1.1	entrée principale	main entry term
Catégorie de dérivé	A.2.4.1	mode de formation du terme	term provenance
Relation de dérivation	A.2.4.2	étymologie	etymology
Notoriété	A.2.3.4	fréquence	reliability code
Expansion classifiante	A.5.3	contexte	context
Détermination	A.2.1.18.1	cooccurrent	collocation
Classe	A.2.2.1	catégorie grammaticale	part of speech
Morphologie	A.2.2	morphologie	morphology
Antonomase, terminologie et figement	A.2.4.1	mode de formation du terme	term provenance
Relation d'éponymie	A.2.4.2	étymologie	etymology
Pivot	A3	équivalence	equivalence
Relation	A6	relation internotion	concept relation
Hyperonymie	A6.1	relation générique	generic relation
Méronymie	A6.2	relation partitive	partitive relation
Accessibilité	A6.4	relation associative	associative relation
Synonymie	A.2.1.13	quasi-synonyme	quasi-synonym
Repérage	A4	domaine	subject field
Diasystème	A.2.3.4	usage	usage

Figure 6 : Correspondance entre les termes de Prolexbase et les catégories de données

Le même travail a été réalisé pour les catégories d'alias (Figure 7).

Prolex	Data Categories		
	Position	Français	Anglais
Prolexème	A.2.1.7	forme intégrale	full form
0	A.2.1.5	nom usuel	common name
variante	A.2.1.9	variante	variant
abréviation	A.2.1.8. 1	abréviation	abbreviation
	A.2.1.8. 2	forme courte	short form
sigle ou acronyme	A.2.1.8. 3	sigle	initialism
	A.2.1.8. 4	acronyme	acronym
0	A.2.1.10	forme translittérée	transliterated form
transcription	A.2.1.11	forme transcrite	transcribed form
latin	A.2.1.12	forme romanisée	romanized form
synonyme diatopique ou diastratique	A.2.1.13	quasi-synonyme	quasi-synonym
glose	A.5.2	explication	explanation

Figure 7 : Correspondance entre les catégories d'alias de Prolexbase et les catégories de données

5.2 Les classes LMF utilisées

Nous utilisons d'abord le noyau obligatoire de LMF, avec la langue et, à l'intérieur d'une langue, les formes et les sens d'une entrée.

Les formes correspondent aux lemmes (prolexèmes, alias et dérivés) de Prolexbase, ainsi qu'à leurs instances. Nous utilisons donc pour cela les classes *Lemma* et *Word Form* de l'extension morphologique de LMF.

Pour les sens, Prolexbase distingue les relations qui ne dépendent pas de la langue (synonymie, méronymie et accessibilité) et les relations qui en dépendent (aliasation, dérivation morphosémantique, expansion classifiante, éponymie). Il s'agit de ne pas dupliquer dans chaque langue une même information. C'est pourquoi nous utilisons les relations de l'extension multilingue pour les premières et les relations de l'extension sémantique pour les secondes.

Nous détaillons ci-dessous les formes, les sens et l'extension multilingue.

5.3 Les formes

L'extension morphologique ne prévoit qu'un unique lemme par entrée lexicale. Dès lors, pour représenter les différents alias d'un nom propre, deux solutions s'offraient à nous :

- Créer une seule entrée lexicale par nom propre en choisissant comme lemme un représentant et en le décomposant en plusieurs représentations (une par alias), puis faire de même pour chaque forme.
- Créer plusieurs entrées lexicales en associant un lemme à chaque alias.

La première solution, quoique acceptable pour le français, serait limitative pour des langues dont la morphologie est plus complexe. Même en français, nous pouvons noter par exemple qu'*Organisation des Nations unies* est un singulier, alors que *Nations unies* est un pluriel... En revanche, la deuxième solution apparaît trop verbeuse, voire redondante, lorsqu'il s'agit d'alias très proches, comme *O.N.U.*, *ONU* et *Onu*.

Nous avons donc choisi une solution intermédiaire en partageant en deux sous-ensembles les alias d'un nom propre :

1. les variantes d'écriture, regroupées sous un même lemme :

- variante (/variant/) - *Pierre Abélard* versus *Pierre Abailard*,
- forme transcrite (/transcribed form/) - *Changai* versus *Shanghai*,
- forme latine (/romanized form/) - *Pariz* versus *Παρις* ;

2. les variantes lexicales, correspondant à des lemmes différents :

- forme intégrale (/full form/) - *Organisation des Nations unies*,
- abréviation (/abbreviation/) - *Microsoft corp.*,
- forme courte (/short form/) - *Nations unies*,
- sigle (/initialism/) - *ONU*,
- acronyme (/acronym/) - *Inalco*,
- quasi-synonyme (/quasi-synonym/) - *Naoned* ;
- explication (/explanation/) - *le Secours Catholique allemand* versus *Caritas Allemagne*.

Le choix de la forme vedette, le prolexème, est arbitraire ; nous avons choisi la forme intégrale (/full form/). L'information sur le type d'alias sera, pour les variantes d'écriture, notée dans l'attribut *orthographyName* de la représentation et, pour les variantes lexicales, dans l'attribut *termProvenance* du sens (voir section 5.3), car une même entrée lexicale peut correspondre à plusieurs sens.

Donnons trois exemples d'entrée lexicale : *Organisation des Nations unies* (Figure 8), *ONU* (Figure 9) et *onusien* (Figure 10). Ces trois entrées ont pour attribut leur catégorie grammaticale (*partOfSpeech*) et sont associées à leur lemme, muni de l'attribut *writtenForm*. Le prolexème et son alias ont un attribut supplémentaire : leur cooccurent (*/collocation/*). Les cooccurents représentent dans Prolexbase une contrainte spécifique, comme, par exemple pour le français, la présence ou non d'un déterminant.

Chaque entrée est accompagnée d'une ou plusieurs formes (les flexions avec l'attribut *writtenForm* et les attributs flexionnels, par exemple *grammaticalGender* et *grammaticalNumber*). Les formes sont éventuellement complétées de différentes représentations.

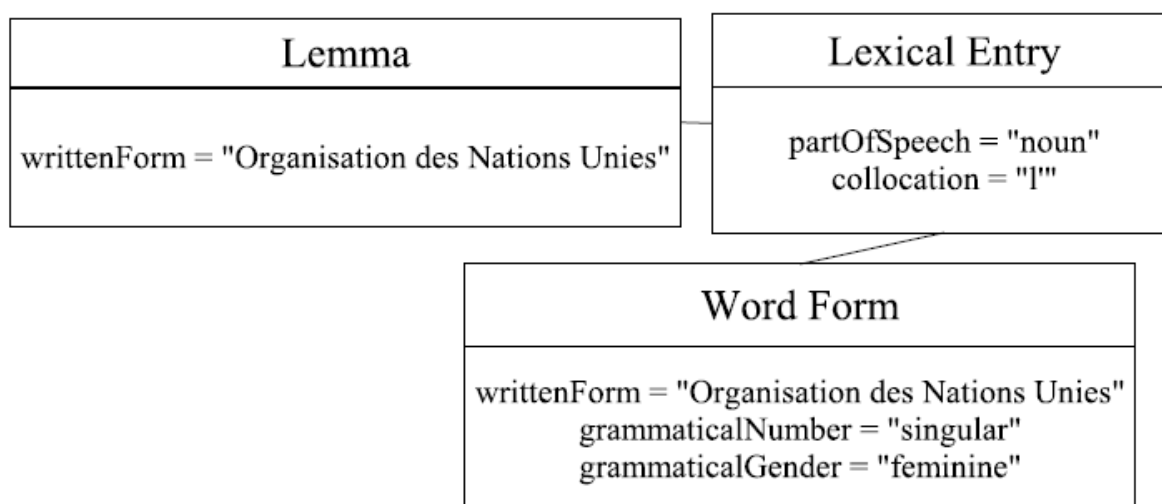


Figure 8 : L'entrée lexicale *Organisation des Nations unies*

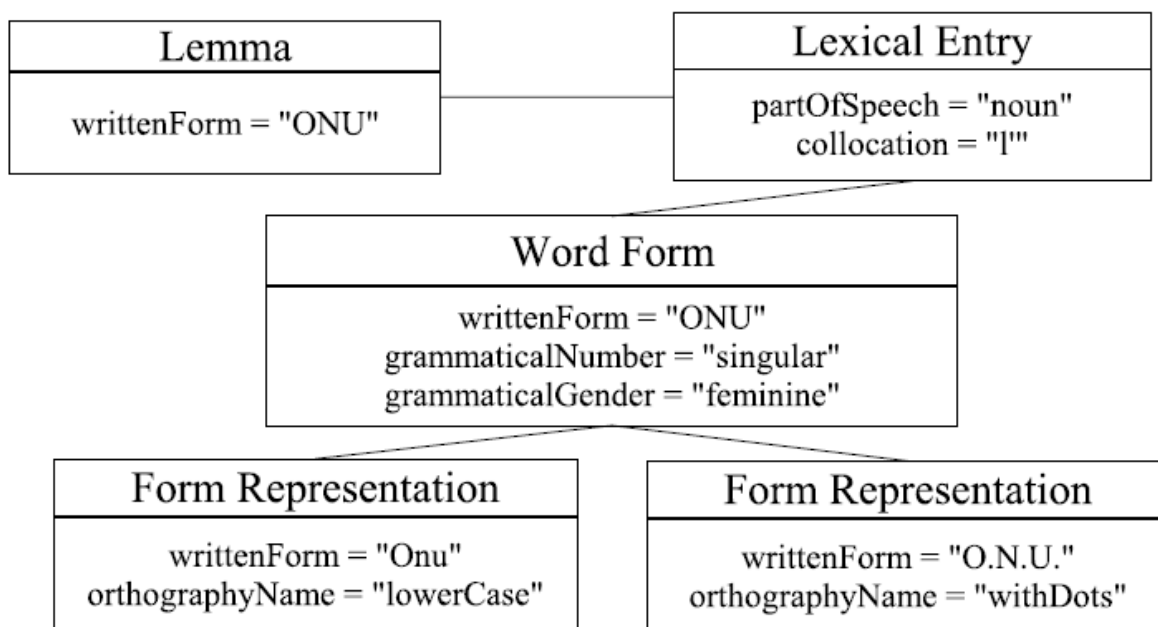
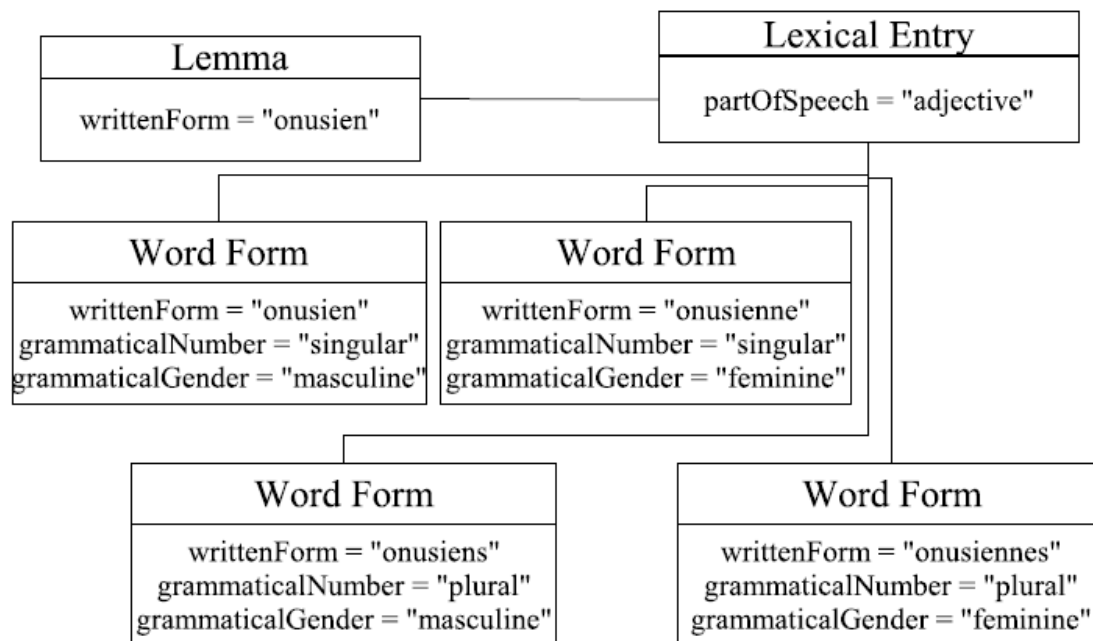


Figure 9 : L'entrée lexicale *ONU*

Figure 10 : L'entrée lexicale *onusien*

5.4 Les sens

Pour décrire la sémantique des entrées d'un même prolexème, nous utilisons deux catégories de données comme attributs :

- le mode de formation du terme (*/term provenance/*) ;
- l'étymologie (*/etymology/*) qui pointe sur le pivot.

La Figure 11 reprend les trois exemples ci-dessus et la Figure 12 présente les entrées *Paris*, *Parisien* et *parisien* :

Remarquons que les homonymes (au niveau des formes) doivent, dans le modèle LMF être regroupés sous une même entrée, contrairement à ce qui est implanté dans la base. Ainsi, les différentes villes nommées *London* devraient être juste distinguées par leur sens.

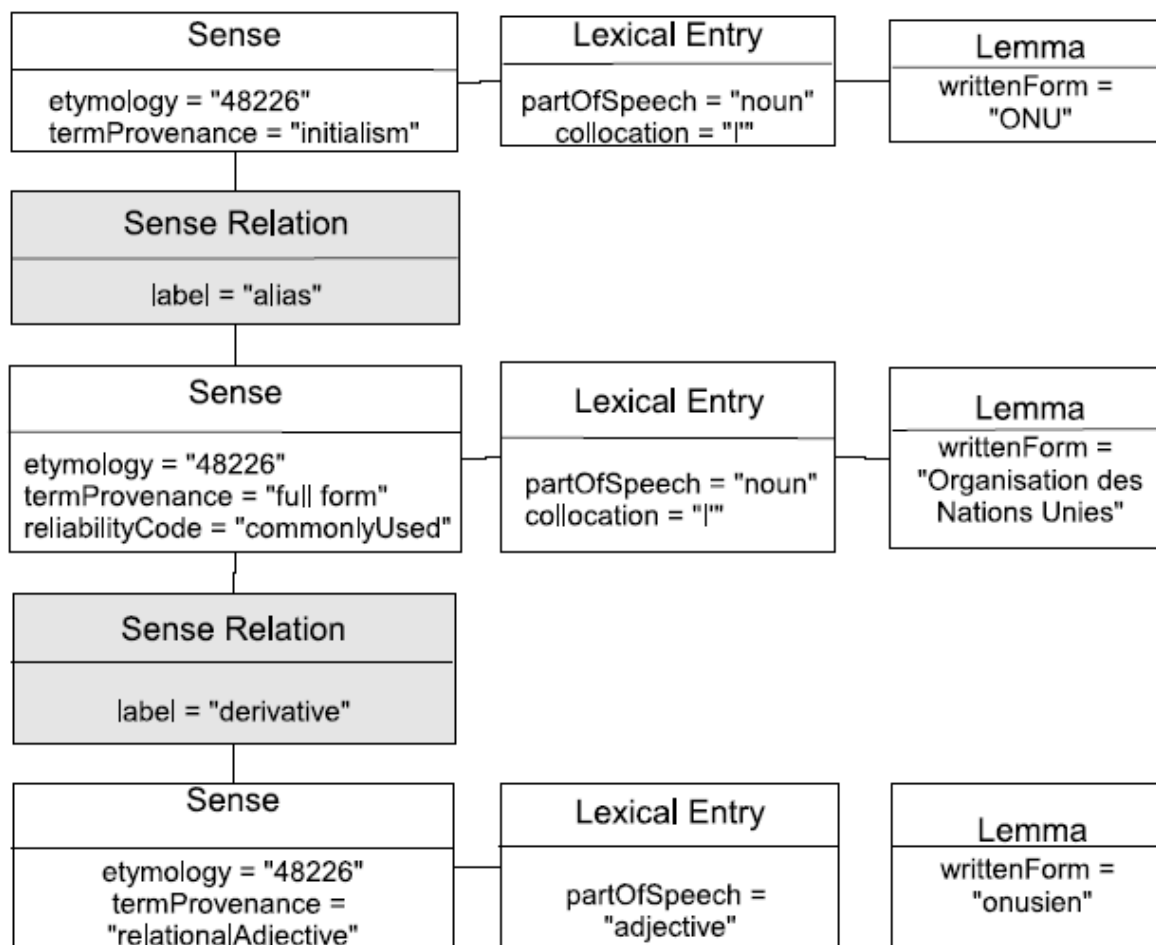


Figure 11 : Liens sémantiques entre les entrées lexicales
Organisation des Nations unies, ONU et onusien

```

<LexicalEntry partOfSpeech="noun">
  <Collocation determiner="zero"/>
  <Lemma writtenForm="Paris"/>
  <WordForm writtenForm="Paris"
grammaticalGender="masculineFeminine" grammaticalNumber="singular"/>
  <Sense id="38558" etymology="38558" termProvenance="fullForm"
reliabilityCode="commonlyUsed"/>
</LexicalEntry>
<LexicalEntry partOfSpeech="noun">
  <Lemma writtenForm="Parisien"/>
  <WordForm writtenForm="Parisiennes" grammaticalGender="feminine"
grammaticalNumber="plural"/>
  <WordForm writtenForm="Parisiens" grammaticalGender="masculine"
grammaticalNumber="plural"/>
  <WordForm writtenForm="Parisienne" grammaticalGender="feminine"
grammaticalNumber="singular"/>
  <WordForm writtenForm="Parisien" grammaticalGender="masculine"
grammaticalNumber="singular"/>
  <Sense id="18666" etymology="38558" termProvenance="relationalName"
reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558"/>
  </Sense>
</LexicalEntry>
<LexicalEntry partOfSpeech="adjective">
  <Lemma writtenForm="parisien"/>
  <WordForm writtenForm="parisiennes" grammaticalGender="feminine"
grammaticalNumber="plural"/>
  <WordForm writtenForm="parisiens" grammaticalGender="masculine"
grammaticalNumber="plural"/>
  <WordForm writtenForm="parisienne" grammaticalGender="feminine"
grammaticalNumber="singular"/>
  <WordForm writtenForm="parisien" grammaticalGender="masculine"
grammaticalNumber="singular"/>
  <Sense id="18667" etymology="38558"
termProvenance="relationalAdjective" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558"/>
  </Sense>
</LexicalEntry>

```

Figure 12 : Les entrées *Paris*, *Parisien* et *parisien*

5.5 L'extension multilingue

Les pivots et leurs relations sont représentés grâce à l'extension multilingue. Nous utilisons, pour les pivots, la classe *Sense Axis*, qui est directement rattachée à la ressource lexicale, comme les lexiques, et, pour les relations qui ne dépendent pas de la langue, la classe *Sense Axis Relation*, avec pour attribut */label/* (Figure 13).

Synonymie	<i>/quasi-synonym/</i>
Méronymie	<i>/partitive relation/</i>
Accessibilité	<i>/associative relation/</i>
Hyperonymie	<i>/generic relation/</i>

Figure 13 : Correspondance entre les relations de Prolexbase et les catégories de données

Par exemple, la Figure 14 représente l'accessibilité de Paris comme capitale de la France, et la Figure 15 son type.

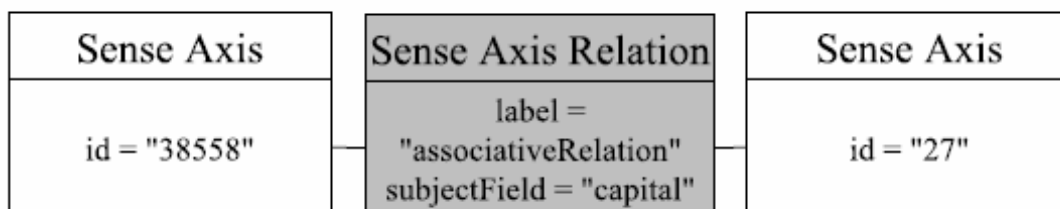


Figure 14 : Accessibilité de *Paris* comme *capitale* de la *France*

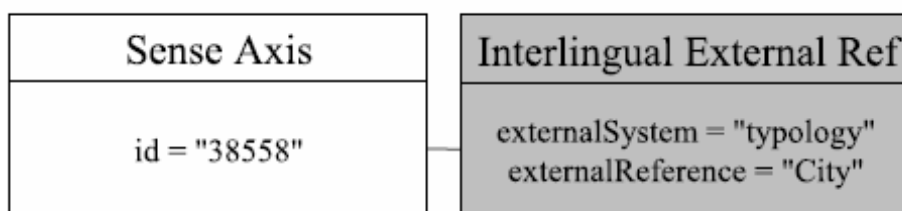


Figure 15 : Le type associé à *Paris*

Finalement les relations de Prolexbase sont représentées, soit par la classe *Sense Relation*, soit par la classe *Sense Axis Relation*, comme sur la Figure 16, où le pivot 38558 (*Paris*) est en relation d'accessibilité (repérage *capitale*) avec le pivot 27 (*France*). Les pivots mettent également en relation les prolexèmes de différentes langues.

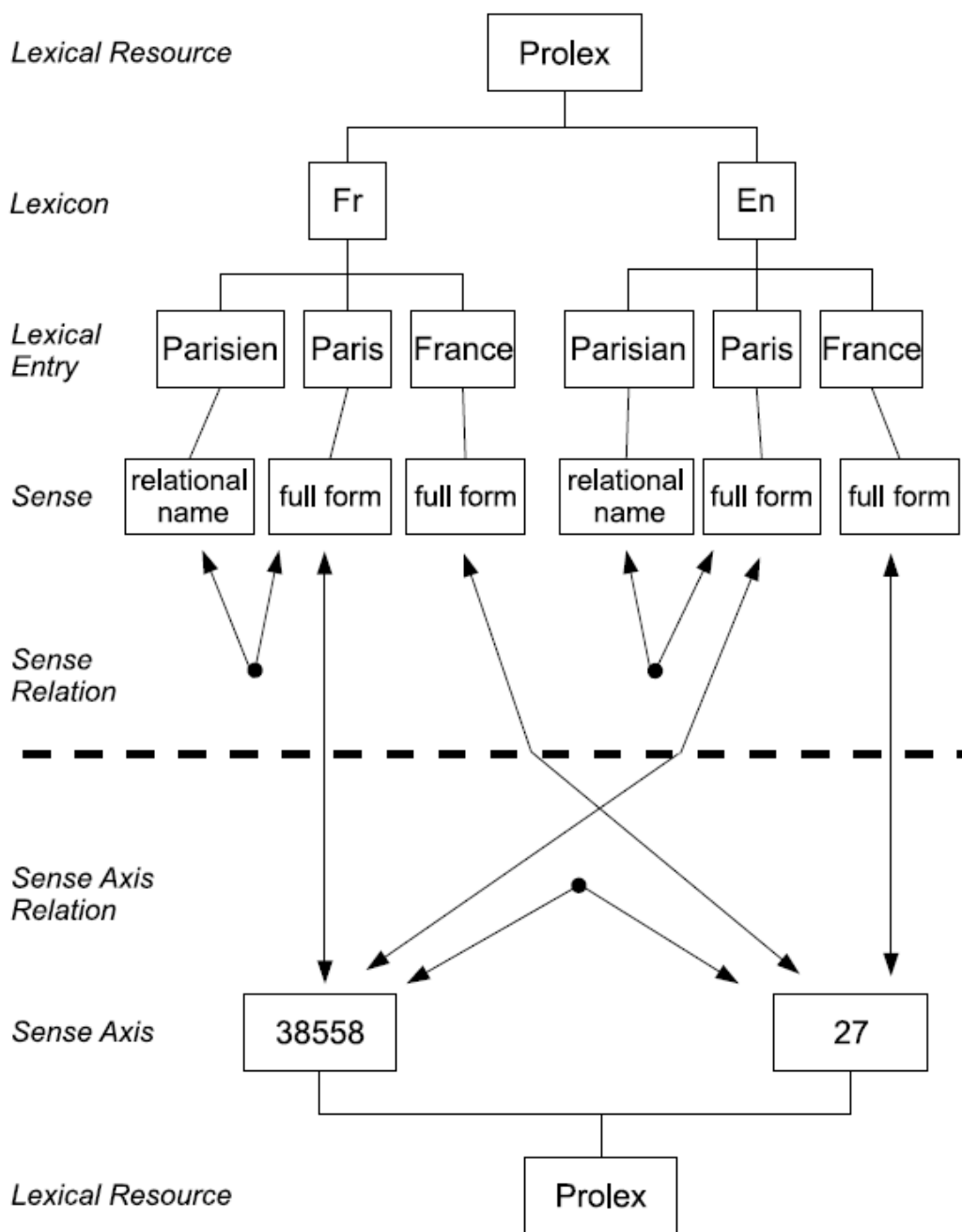


Figure 16 : Les relations de Prolexbase

La Figure 17 présente l'exemple du pivot 38558 (*Paris*), en relation par exemple avec les pivots 55120 (*Ville lumière*), 53865 (*parc de la Villette*), 53687 (*palais des Tuileries*)..., 48149 (*Perche*), 27 (*France*), 5 (*Île-de-France*), 54453 (*Air liquide*) et 54914 (*Arkema*).

```

<SenseAxis id="38558">
  <InterlingualExternalRef externalSystem="typology"
externalReference="city"/>
  <InterlingualExternalRef externalSystem="existence"
externalReference="historical"/>
  <SenseAxisRelation label="quasiSynonym" targets="55120"
subjectField="diaphasic"/>
  <SenseAxisRelation label="partitiveRelation" targets="53865"/>
  <SenseAxisRelation label="partitiveRelation" targets="53687"/>
  ...
  <SenseAxisRelation label="partitiveRelation" targets="48149"/>
  <SenseAxisRelation label="associativeRelation" targets="27"
subjectField="capital"/>
  <SenseAxisRelation label="associativeRelation" targets="5"
subjectField="capital"/>
  <SenseAxisRelation label="associativeRelation" targets="54453"
subjectField="headquarters"/>
  ...
  <SenseAxisRelation label="associativeRelation" targets="54914"
subjectField="headquarters"/>
</SenseAxis>

```

Figure 17 : L'entrée 38558

5.6 Représentation de Prolexbase par LMF

Les classes LMF utilisées par Prolexbase sont présentées à la Figure 18 et les attributs de ces différentes classes, Figure 19.

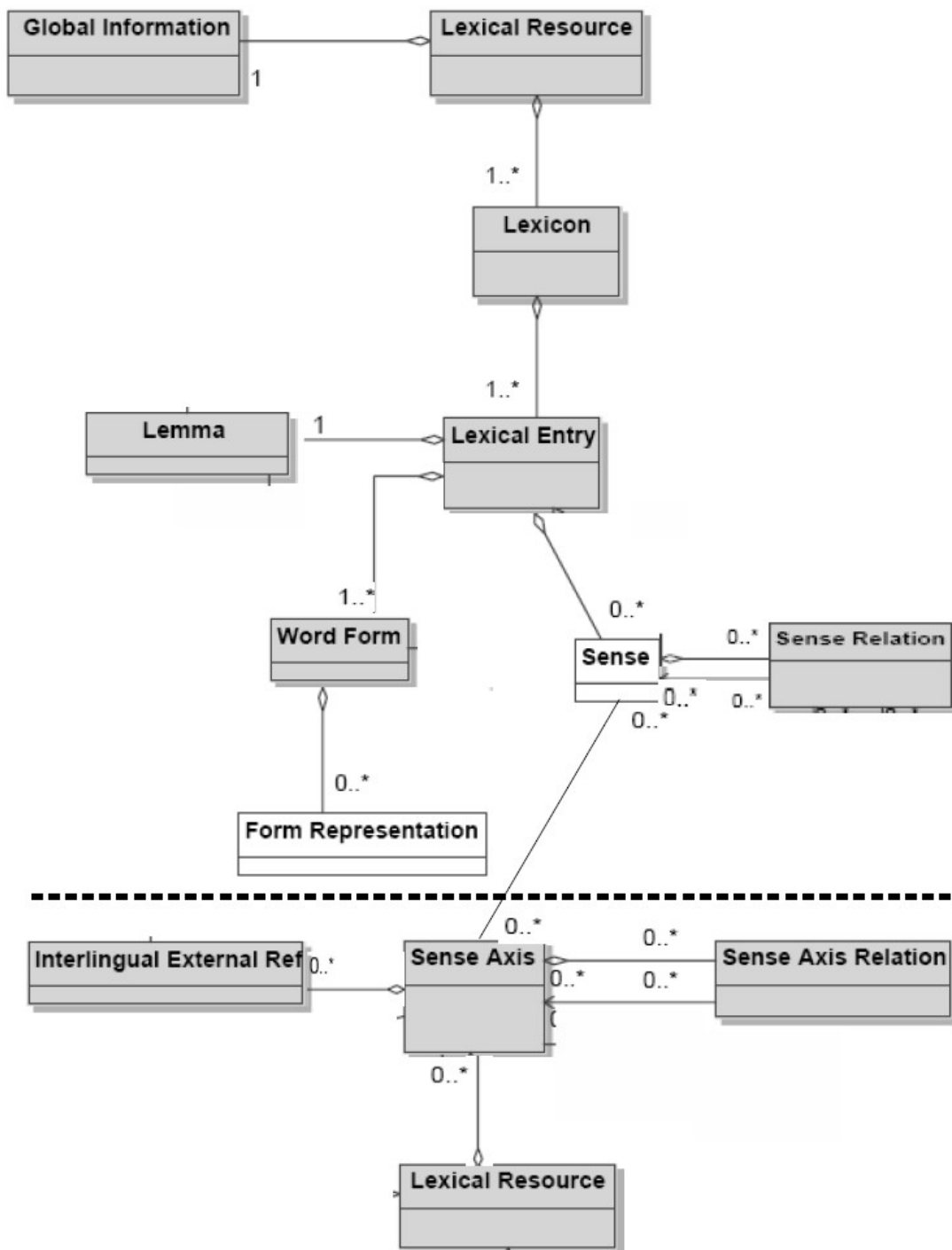


Figure 18 : Les classes LMF utilisées par Prolexbase

Classes	Attributs
Lexical Entry	writtenForm partOfSpeech reliabilityCode
Form	writtenForm variantType collocation
Form Representation	writtenForm orthographyName
Lemma	writtenForm
Word Form	writtenForm grammaticalNumber grammaticalGender grammaticalCase grammaticalTense grammaticalMood person
Sense	etymology termProvenance
Sense Relation	label
Sense Axis	id
Sense Axis Relation	label domain usage
Interlingual External Ref	externalSystem externalReference

Figure 19 : Les attributs des classes LMF utilisées par Prolexbase

6 Couverture linguistique de Prolexbase 1.1

Prolexbase 1.1 contient 54 774 prolexèmes français, 730 alias et 20 614 dérivés.

Les prolexèmes sont inégalement répartis entre les différents supertypes (4 588 anthroponymes, 49 789 toponymes, 175 ergonymes et 222 pragmonymes). La Figure 20 précise la répartition par type.

Nom propre		54 774
Anthroponyme		4 588
<i>Individuel</i>		3 767
	Personne	3 764
	Patronyme	0
	Prénom	0
	Pseudo Anthroponyme	3
<i>Collectif</i>		821
	Dynastie	50
	Ethnonyme	140
<i>Groupement</i>		631
	Association	33
	Ensemble	14
	Entreprise	506
	Institution	57
	Organisation	21
Toponyme		49 789
	Astronyme	27
	Edifice	86
	Géonyme	207
	Hydronyme	4 368
	Ville	41 941
	Voie	15
<i>Territoire</i>		3 145
	Pays	436
	Région	2 644
	Supranational	65
Ergonyme		176
	Objet	1
	Œuvre	81
	Pensée	3
	Produit	87
	Vaisseau	4
Pragmonyme		221
	Catastrophe	1
	Météorologie	1
	Fête	10
	Histoire	206
	Manifestation	3

Figure 20 : La répartition par type dans Prolexbase

Les relations entre pivots sont au nombre de 50 567, correspondant à 2 249 accessibilités, 47 670 méronymies et 648 synonymies.

Enfin, Prolexbase 1.1 contient 75 368 lemmes qui engendrent 123 859 formes fléchies. Ces lemmes se répartissent en 65 805 noms, 10 300 adjectifs et 13 préfixes.