Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres

Béatrice Bouchou, Denis Maurel

Université François Rabelais Tours - Laboratoire d'Informatique
{beatrice.bouchou, denis.maurel}@univ-tours.fr

RÉSUMÉ. *Nous présentons dans cet article l'expression de Prolexbase dans le cadre défini par le projet de norme LMF (Language resource management - Lexical Markup Framework). Prolexbase est une ressource lexicale de noms propres, organisée suivant une approche onomasiologique, regroupant les entrées sous le concept multilingue de nom propre conceptuel, mais qui met également l'accent sur la description morphologique des noms propres. Nous montrons qu'il est possible, sans perte d'information, d'exprimer Prolexbase selon une approche sémasiologique, conforme à LMF.*

ABSTRACT. *We present in this paper how Prolexbase can be expressed within the LMF (Language resource management - Lexical Markup Framework) ISO project framework. Prolexbase is a lexical resource of proper names, organized according to an onomasiological view, with entries based on the concept of conceptual proper name. In the same time, Prolexbase also focuses on morphological description of proper names. We show that it is possible, without loss of information, to represent Prolexbase according to a semasiological view compliant with LMF.*

MOTS-CLÉS : *dictionnaire électronique, norme ISO, LMF, nom propre.*

KEYWORDS: *Electronic dictionary, ISO standard, LMF, Proper name.*

1. Introduction

Le projet Prolex, initié dans les années 90 et piloté par le laboratoire d'informatique de l'université François Rabelais de Tours, a pour but le traitement automatique des noms propres. Une de ses dernières réalisations, dans le cadre de l'appel Technolangue (Maurel *et al.*, 2006), a été celle d'un dictionnaire multilingue de noms propres, Prolexbase (Tran et Maurel, 2006), muni de relations, aujourd'hui disponible sur le site du CNRTL[1].

Dans cet article nous montrons comment Prolexbase peut s'appréhender dans le modèle LMF, afin d'en faire une ressource lexicale standard, mais aussi pour guider la construction de ressources du même type dans des langues de plus en plus nombreuses et favoriser leur interopérabilité.

Par ailleurs, le développement d'une ressource ou d'un modèle de ressource conforme à LMF représente en soi une contribution à la réflexion en cours sur ce standard. En effet, LMF (pour *Lexical Markup Framework*, future norme ISO 24613 (ISO/TC 37/SC 4, 2007)), est encore un projet de norme concernant les lexiques pour le traitement automatique des langues. Ce projet de norme consiste en un modèle abstrait, destiné à servir de cadre à la définition de modèles de données lexicales (comme par exemple les noms propres).

LMF se concentre sur les lexiques (Francopoulo *et al.*, 2006b ; Salmon-Alt *et al.*, 2005), mais plus généralement l'ISO/TC 37 regroupe un ensemble de standards pour la création et l'utilisation des ressources linguistiques, dans le but de faciliter les échanges de données, voire l'intégration de ressources. L'étendue de ces standards est vaste, cela va des ressources monolingues aux multilingues, et concerne les différents niveaux de description (morphologie, syntaxe et sémantique). Les langages considérés ne se limitent pas aux langues européennes et il n'y a pas non plus de limite *a priori* concernant les applications auxquelles sont destinées ces ressources.

L'utilisation de ces normes apporte donc, d'une part, l'avantage de la réutilisation d'un modèle existant, déjà validé, dans la construction de nouvelles ressources lexicales, ainsi que la possibilité d'échanges et d'intégration de ressources. D'autre part, le fait de normaliser ces ressources ne peut que favoriser le développement d'applications utilisatrices, en leur offrant une interface d'accès standard. Ainsi que le montrent les auteurs de (Francopoulo *et al.*, 2006a), les standards ne relèvent pas uniquement de l'ingénierie, mais permettent également à la recherche de se développer, en particulier dans le domaine du TAL.

Pour tendre vers l'universalité, il faut en plus utiliser XML[2], mais dans cet article nous n'abordons pas cette étape, nous restons volontairement au niveau des concepts communs aux ressources lexicales. Dans la suite de cet article nous appellerons ProlexLMF le modèle de Prolexbase conforme à LMF.

1. `http://www.cnrtl.fr/lexiques/prolex/`
2. Ce que préconise LMF et ce que nous faisons également.

Nous rappelons dans un premier temps les caractéristiques essentielles de Prolexbase (section 2), puis les parties principales de LMF et plus particulièrement celles que nous utilisons (section 3), avant de décrire ProlexLMF. La description procède en trois temps : l'architecture générale (section 4), le niveau linguistique (section 5) et le niveau interlingue (section 6). En ce qui concerne le niveau linguistique nous présentons le modèle qui correspond à la partie de Prolexbase publiée sur le site du CNRTL, à savoir le dictionnaire de formes fléchies. Il existe dans Prolexbase une partie où sont décrites les règles de flexion qui permettent de générer les formes fléchies, mais nous ne l'abordons pas ici, afin d'offrir un aperçu relativement complet de notre démarche tout en évitant un « effet catalogue ». Enfin nous discutons (section 7) des différents modèles de Prolexbase et de l'adéquation de LMF à cette ressource lexicale.

2. Présentation de Prolexbase

Prolexbase est d'abord un dictionnaire, avec des lemmes et des formes fléchies. Il contient des noms propres, mais aussi des alias de ces noms propres et des dérivés de nom propre, pour peu que la dérivation soit régulièrement porteuse de sens et que ce sens soit directement en relation avec un nom propre de Prolexbase. Cette relation est appelée *lien morphosémantique* dans WordNet (Fellbaum et Miller, 2003). Donnons quelques exemples :

– *Onu*, *Nations unies* et *Organisation des Nations unies* sont des alias d'un même nom propre ;

– *Parisien* et *Parigot* sont des noms dérivés dont le sens se déduit régulièrement du nom propre *Paris* (*habiter/être né à Paris*, avec un sens familier ou péjoratif pour le second) ;

– *parisianisme* est bien un nom dérivé du nom propre *Paris*, mais son sens est lexicalisé (*ensemble des comportements, défauts et qualités, prêtés aux Parisiens ou des caractéristiques censées être celles de la vie (mondaine) parisienne*, d'après le TLFI) et cette dérivation ne se retrouve pas pour les autres noms de ville. Ce mot ne sera pas placé dans Prolexbase, contrairement aux deux précédents.

L'originalité de Prolexbase est de proposer l'ensemble de ces lemmes sous une seule entrée, appelée « prolexème », avec l'idée que, dans un contexte multilingue, la traduction d'un des mots de cet ensemble peut nécessiter l'utilisation d'un autre mot du prolexème de la langue cible. Si *Parisien* a une traduction en anglais, *Parisian*, il n'en va pas de même pour *Tourangeau* qui se traduira par une glose où sera présent le nom propre *Tours*, par exemple *inhabitant of the city of Tours in France*. Certaines langues ont une morphologie dérivationnelle plus conséquente que le français. Par exemple, il existe en serbe un adjectif possessif construit sur chaque nom à trait humain, y compris sur les noms de personne ou sur les noms d'habitant : *C'est la voiture d'un Parisien* peut se traduire par *To je Parižaninov auto*, où *Parižaninov* est un adjectif possessif (Maurel *et al.*, 2007).

Ainsi, le prolexème français *Paris* sera en fait un ensemble de lemmes, associés à une catégorie[3] ; par commodité, on choisit pour désigner cet ensemble l'un de ses représentants, abusivement appelé aussi prolexème :

Prolexème-fra$_{Paris}$={Paris.N, Parisien.NR, Parigot.NRD, parisien.AR, parigot.ARD}

Chaque lemme est associé à une règle de flexion, éventuellement complexe lorsqu'il s'agit de mots polylexicaux (Savary, 2005). Ces règles permettent la génération de toutes les formes fléchies associées à un prolexème :

Instances-fra$_{Paris}$={Paris.N:ms:fs, Parisien.N:ms, Parisienne.N:fs, Parisiens.N:mp, Parisiennes.N:fp, Parigot.N:ms, Parigote.N:fs, Parigots.N:mp, Parigotes.N:fp, parisien.A:ms, parisienne.A:fs, parisiens.A:mp, parisiennes.A:fp, parigot.A:ms, parigote.A:fs, parigots.A:mp, parigotes.A:fp}

2.1. *Un dictionnaire multilingue*

Ensuite, Prolexbase est un dictionnaire multilingue. Comme cela vient d'être présenté ci-dessus, pour une langue donnée, une entrée linguistique de Prolexbase est un prolexème. Chaque prolexème d'une langue est relié à un et un seul pivot interlingue qui est un identificateur unique. C'est par ce pivot que passe la traduction d'une langue à l'autre. Le pivot correspond à ce qu'on pourrait appeler un « sens », s'il s'agissait d'un nom commun... Ici, il désigne un « point de vue sur le référent ». Nous distinguons ainsi *Paris* de *Ville lumière* qui auront dans Prolexbase deux pivots différents (voir section 2.2).

Par exemple, le pivot correspondant à Paris est 38 558. C'est le pivot de chacun des prolexèmes suivants[4] :

Prolexème-fra$_{Paris}$={Paris.N, Parisien.NR, Parigot.NRD, parisien.AR, parigot.ARD} Prolexème-eng$_{Paris}$={Paris.N, Parisian.NR, Parisian.AR} Prolexème-srp$_{Pariz}$={Pariz.N, Parižanin.NRM, Parižanka.NRF, Parižaninov.APM, Parižankin.APF, parižanski.AR...}

Pour passer d'une langue à une autre, suivant les mots et les catégories, on « traduit » ou on « glose » : le nom français *Paris* donnera *Paris* en anglais et *Pariz* en serbe, mais l'adjectif possessif serbe *Parižaninov* donnera *d'un Parisien* en français.

3. On distingue en particulier le nom propre lui-même (N), le nom relationnel (NR), le nom relationnel diastratique (NRD), l'adjectif relationnel (AR) et l'adjectif relationnel diastratique (ARD).

4. Dans cet exemple s'ajoutent quatre autres catégories : le nom relationnel masculin (NRM) et son adjectif possessif (APM), le nom relationnel féminin (NRF) et son adjectif possessif (APF).

2.2. *Un dictionnaire muni de relations*

Enfin, Prolexbase est un dictionnaire muni de relations. Les différents points de vue représentés par les pivots sont liés les uns aux autres par deux relations paradigmatiques (synonymie et méronymie) et par une relation syntagmatique (accessibilité). Reprenons l'exemple du pivot 38 558 (*Paris*) : il est en relation de synonymie (diaphasique) avec le pivot de *Ville lumière*, en relation de méronymie avec le pivot de *Île-de-France* et en relation d'accessibilité (repérage « capitale ») avec le pivot de *France*.

Pour être plus précis :

1) à la relation de synonymie est associé un indicateur diasystématique (selon (Coseriu, 1998)) : diachronique (variété dans le temps, par exemple *Birmanie* et *Myanmar*), diastratique (variété relative à la stratification socioculturelle, comme *Johnny Hallyday* et *Jean-Philippe Smet*) et diaphasique (variété concernant les finalités de l'emploi, *Paris* et *Ville lumière*)[5] ;

2) la relation de méronymie est très largement étendue et ne concerne pas que l'inclusion spatiale ou temporelle, mais aussi le lien entre une célébrité et un pays, entre une entreprise et une filiale, entre un roi et une dynastie ;

3) comme pour la synonymie, la relation d'accessibilité est associée à des repérages généraux, comme la parenté (*Irène Joliot-Curie* est la fille de *Marie Curie*), la création (*Antonio Lucio Vivaldi* est le compositeur des *Quatre Saisons*), la gouvernance (*Henri IV* est un roi de *France*)...

La figure 1 reprend l'exemple du pivot 38 558 (*Paris*). La version XML de cette entrée est présentée en annexe 6.

Ces trois relations sont complétées par des relations d'hyperonymie, soit vers une typologie qui comporte trente types (personne, association, hydronyme, ville, pays, histoire, catastrophe...) et neuf supertypes, soit vers un paradigme d'existence, qui comporte trois instances (historique, fictif et religieux). Plus de détails sont donnés sur ce sujet dans (Tran et Maurel, 2006).

D'autre part, deux relations syntagmatiques existent aussi au niveau linguistique : l'expansion classifiante[6] (collocations libres) et l'éponymie (lexicalisations et figements), voir (Tran et Maurel, 2006).

Prolexbase est actuellement disponible sur le site du CNRTL ; la version téléchargeable ne contient que la partie française, avec 54 774 entrées, les pivots des prolexèmes français. Ceux-ci incluent 730 alias et 20 614 dérivés, ce qui représente

5. En fait, Coseriu en définit un quatrième, diatopique (variété dans l'espace), mais cette sorte de synonymie est dépendante de la langue et est placée au niveau linguistique, à l'intérieur du prolexème. Par exemple, en français, *Naoned* fait partie du prolexème Prolexème-fra$_{Nantes}$. En breton, ce serait lui le prolexème.

6. C'est cette relation et la relation de méronymie qui permettent par exemple la glose proposée ci-dessus pour traduire *Tourangeau* par *inhabitant of the city of Tours in France*.

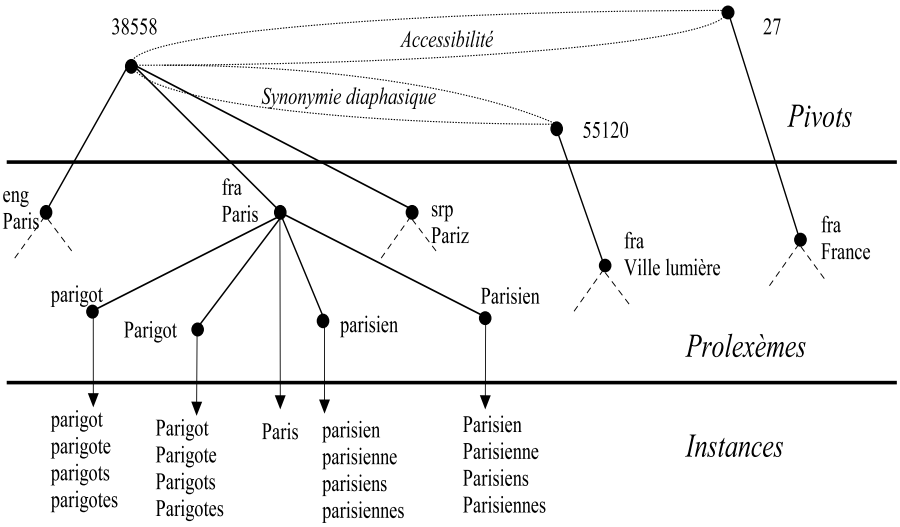

Figure 1. *L'entrée Paris dans Prolexbase, décrite en XML dans l'annexe 6*

75 368 lemmes qui engendrent 123 859 formes fléchies. Ces lemmes se répartissent en 65 805 noms, 10 300 adjectifs et 13 préfixes. Les relations entre pivots sont au nombre de 50 567, correspondant à 2 249 accessibilités, 47 670 méronymies et 648 synonymies.

Prolexbase a permis la génération de dictionnaires utilisés par les logiciels Unitex (Paumier, 2003) et Nooj (Silberztein, 2004). Une autre version de la base est incluse dans la plateforme WS4LR (*Work Station for Lexical Resources*) (Krstev *et al.*, 2006). Faciliter l'utilisation de Prolexbase par d'autres applications est notre principale motivation pour proposer un modèle LMF de Prolexbase.

3. Les bases du projet de norme LMF

LMF s'articule autour de la spécification d'un noyau de descriptions obligatoire, d'une part, et, d'autre part, d'un ensemble de descripteurs élémentaires définis par une autre norme, l'ISO 12620, appelée *catégories de données* (Romary, 2000). Nous présentons d'abord la structure de LMF, puis les principes des catégories de données. Les noms des classes LMF, les noms et les valeurs des catégories de données sont cités en anglais, comme dans la proposition de norme.

3.1. *La structure modulaire de LMF*

Une ressource lexicale, telle que la conçoit LMF, est au moins composée d'un ou plusieurs lexiques (classe *Lexicon*), chacun contenant la description d'une langue particulière. Un lexique contient un ensemble d'entrées lexicales (classe *Lexical Entry*) qui sont composées d'un ensemble de formes (classe *Form*) et, éventuellement, d'un ensemble de sens (classe *Sense*). Les formes peuvent être raffinées en différentes représentations (classe *Form Representation*). Quant aux sens, ils peuvent être multiples, pour représenter la polysémie, pour autant que les informations attachées aux formes soient identiques pour tous les sens.

Prenons l'exemple du mot *tour*. Avec cette structure minimale, si nous souhaitons seulement une liste de mots, nous pouvons définir une entrée lexicale contenant simplement cette forme, éventuellement avec un attribut pour sa catégorie grammaticale. Si notre lexique contient une représentation sémantique, nous pouvons ajouter à cette entrée trois sens, *construction nettement plus haute que large*, *machine-outil servant à façonner des pièces cylindriques tournant sur leur axe* et *boîte, armoire cylindrique tournant sur un pivot* (d'après le TLFI). Si nous complétons encore ce mot par son genre, ce n'est plus une, mais deux entrées lexicales qu'il faudra créer, une féminine associée au premier sens et l'autre masculine, associée aux deux autres sens.

Les lexiques contenant les formes et les sens constituent le cœur de LMF, autrement dit le noyau de description obligatoire. Toutes les autres informations possibles sont réparties dans huit extensions[7], elles aussi normalisées.

Pour ajouter à notre exemple les deux formes fléchies *tour* et *tours*, il nous faudra utiliser l'extension morphologique qui comprend au moins la notion de lemme (classe *Lemma*), mais permet aussi de lister les formes fléchies (classe *Word Form*)[8]. En revanche, si l'information flexionnelle que nous souhaitons voir apparaître consiste en une règle et non en une liste, il faudra à nouveau compléter cette extension par une autre (l'extension paradigmatique).

Remarquons que ces deux extensions sont indépendantes de la partie sémantique de l'entrée lexicale. Celle-ci aussi peut être complétée par l'extension sémantique, avec la possibilité de définir des relations entre sens (classe *Sense Relation*)[9].

Il est important de voir que cette description sémantique est dépendante de la langue. C'est grâce à l'extension multilingue qu'il est possible de relier un sens propre à une langue avec un sens qui peut être partagé par plusieurs langues. On retrouve dans cette dernière extension aussi bien l'idée du pivot de traduction (classe *Sense Axis*) que

7. *Morphology extension*, *Machine Readable Dictionary extension*, *NLP syntax extension*, *NLP semantics extension*, *NLP multilingual notations extension*, *NLP paradigm pattern extension*, *NLP multiword expression patterns extension* et *Constraint expression extension*.
8. Ainsi que les racines (classe *Stem Or Root*) ou les unités des mots polylexicaux (classe *List Of Components*).
9. Tout comme des ensembles de synonymes (classe *Synset*) et des structures prédicativess (classe *Predicative Representation*), éventuellement conjointement à l'extension syntaxique.

celle du transfert (classes *Transfert Axis* et *Example Axis*), notions issues des principaux projets de ressources multilingues pour le traitement automatique des langues. Ces trois classes se rattachent directement à la ressource lexicale, comme les lexiques. Des relations entre ces pivots peuvent être définies *via* la classe *Sense Axis Relation*. De plus, la classe *Interlingual External Ref* permet de lier un pivot à des descriptions externes à la ressource lexicale.

3.2. *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 (Romary, 2000). Celle-ci « 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 définit :

– les catégories de données liées au terme (type d'entrée, grammaire, usage, formation, morphologie, statut...) ;

– l'équivalence (pour l'aspect multilingue) ;

– le domaine ;

– la description (définition, explication, contexte...) ;

– les relations (génériques, partitives, associatives...) ;

– les notions (et les systèmes de notions hiérarchiques) ;

– le langage documentaire et l'information administrative.

La norme ISO 12620 est donc un inventaire des descripteurs élémentaires courants, offrant ainsi un vocabulaire normalisé pour décrire des ressources linguistiques. Elle a été définie à l'origine pour les grands projets sur les bases de données terminologiques (Geneter, MARTIF, etc.).

Cette norme définit la notion de registre de catégories de données (*DCR*). Dans un tel registre, sont précisés le nom et la définition de chaque catégorie de données, les valeurs qu'elle peut prendre, et éventuellement les classes du modèle auxquelles elle se rattache. Comme le souligne Susanne Salmon-Alt (Salmon-Alt, 2006), l'un des avantages majeurs du *DCR* réside dans la maintenance d'un seul référentiel de descripteurs, utilisables à la fois pour le lexique et l'annotation de corpus.

La définition d'un modèle pour les noms propres dans le cadre de LMF revient donc à instancier le noyau obligatoire par les éléments mis en évidence dans Prolexbase, à choisir les extensions nécessaires et à définir l'ensemble des catégories de données[10] qui permettent de décrire les prolexèmes et leurs relations.

10. L'annexe 1 présente cet ensemble, prélude à un DCR pour les noms propres, et l'annexe 2 les adaptations de Prolexbase qui lui sont liées. Le schéma général des classes de ProlexLMF

4. Architecture générale de Prolexbase en LMF

Nous utilisons d'abord le noyau obligatoire de LMF : formes et sens. Pour les formes, Prolexbase contient à la fois des règles de flexion (mono et polylexicale) et l'ensemble des instances. Dans la gestion de la base, un lemme est entré avec ses informations morphologiques et un programme de génération complète automatiquement la table des instances. Pour le lexique de formes fléchies ProlexLMF, la représentation des lemmes et des formes fléchies exige l'instanciation des classes *Lemma* et *Word Form* de l'extension morphologique.

Pour les sens, Prolexbase distingue les relations qui ne dépendent pas de la langue (synonymie, méronymie et accessibilité) et les relations qui en dépendent (aliasisation, dérivation morphosémantique, expansion classifiante, éponymie). Il s'agit de ne pas dupliquer dans chaque langue une même information (inutile de dire à la fois que *Birmanie* est un synonyme diachronique de *Myanmar* et que *Burma* est un synonyme diachronique de *Union of Myanmar* !). C'est pourquoi nous utilisons les relations de l'extension multilingue pour les premières et les relations de l'extension sémantique pour les secondes[11].

Nous venons de justifier notre choix de représentation des relations sémantiques d'un point de vue linguistique, il se justifie aussi par les règles de l'ingénierie des systèmes d'information. En effet, cette solution évite une forte redondance des informations or, la redondance est un facteur évident d'incohérence et d'augmentation des coûts de maintenance. Dans ProlexLMF, l'ajout d'un nouveau langage à la ressource lexicale revient à ajouter une instance de la classe *Lexicon*, dans laquelle les prolexèmes font référence aux instances de la classe *Sense Axis* qui existent déjà dans la ressource lexicale, et à ajouter de nouvelles instances de la classe *Sense Axis* pour les prolexèmes du nouveau langage qui n'auraient de correspondant dans aucune langue déjà représentée.

L'architecture générale de ProlexLMF correspond à l'architecture générale d'une représentation XML de Prolexbase à laquelle nous aboutissions en 2005 (Bouchou *et al.*, 2005) en développant une approche de conception de systèmes d'information orientée XML. Étant donné que notre motivation pour LMF est d'aboutir à un standard de représentation XML, ce constat renforce nos choix, notamment celui d'utiliser la partie multilingue de LMF pour représenter les relations sémantiques (indépendantes de la langue).

Nous présentons dans la section 5 la description des prolexèmes dans une langue et, dans la section 6, celle des pivots. La correspondance entre les termes utilisés dans Prolexbase et des catégories de données est détaillée en annexes 1 et 2. De plus, les

est donné dans l'annexe 3 et les attributs nécessaires à la description des prolexèmes français, dans l'annexe 4.

11. Précisons cependant que, dans la version actuelle, ni l'expansion classifiante, ni l'éponymie ne sont présentes.

annexes 3 et 5 récapitulent les classes de ProlexLMF et un schéma XML pour les représenter (où on retrouve les catégories de données dans les noms d'attributs).

5. Description monolingue

Toutes les descriptions présentées dans cette partie sont dépendantes de la langue. En plus du noyau de description obligatoire du modèle LMF, elles utilisent les extensions morphologique et sémantique, en leur associant un certain nombre de catégories de données (notées ci-dessous entre barres obliques, */data categorie/*). On peut remarquer qu'il existe deux grandes différences entre le modèle de Prolexbase et la présentation dictionairique classique, décrite par LMF.

La première différence est que l'entrée linguistique d'un terme de Prolexbase est un prolexème, c'est-à-dire un ensemble de lemmes éventuellement de catégories grammaticales différentes, alors que celle de LMF correspond à un unique lemme. Pour le nom propre et ses dérivés, nous n'avons d'autre choix que de créer une entrée lexicale LMF pour chacun de ces lemmes. Pour les alias, nous pourrions tous les regrouper sous une même entrée lexicale (en créant différentes représentations du lemme), mais cela ne nous a paru judicieux que pour des alias vraiment très proches (voir section 5.1). Remarquons qu'il faut une entrée lexicale pour chaque expansion classifiante et pour chaque éponyme, mais ces entrées seront considérées comme appartenant à un lexique de noms communs, de termes ou de phrasèmes ; elles ne sont donc pas décrites dans cet article.

La seconde différence importante est que le prolexème se rapporte à un unique pivot, alors qu'une entrée LMF peut regrouper des homonymes. Cela nécessite une réorganisation fort simple qui met l'accent sur les sens puisque, dans LMF, les relations sémantiques propres à une langue et les liens vers un pivot sémantique multilingue portent sur les sens et non sur les entrées lexicales.

5.1. *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[12] :

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

- variante (*/variant/*) - *Pierre Abélard* versus *Pierre Abailard*,
- forme transcrite (*/transcribed form/*) - *Changaï* versus *Shanghai*,
- forme latine (*/romanized form/*) - *Pariz*[13] versus Париз ;

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*.

Conformément au modèle de Prolexbase, le choix du représentant est arbitraire. Nous avons choisi comme représentant du prolexème la forme intégrale (*/full form/*). L'information sur le type d'alias sera, pour les variantes d'écriture, notée dans l'attribut *othographyName* de la représentation et, pour les variantes lexicales, dans l'attribut *termProvenance* du sens (voir section 5.2), car une même entrée lexicale peut correspondre à plusieurs sens.

Pour illustrer notre propos, considérons trois exemples d'entrée lexicale : un représentant de prolexème (*Organisation des Nations unies*, figure 2), un alias (*ONU*, figure 3) et un dérivé (*onusien*, figure 4). Ces trois entrées ont pour attribut leur catégorie grammaticale (*partOfSpeech = "noun"* ou *partOfSpeech = "adjective"*) et sont associées à leur lemme, muni de l'attribut *writtenForm*.

Le prolexème et son alias ont un attribut supplémentaire : leur cooccurrent (*/collocation/*) : *collocation = "l'"*. Les cooccurrents représentent dans Prolexbase une contrainte spécifique, comme, par exemple pour le français, la présence ou non d'un déterminant.

Suivant la morphologie spécifique de la langue décrite, les attributs associés aux formes peuvent varier (le nombre, le genre, le cas, etc.). Sur nos exemples, il y a une forme fléchie pour le prolexème et l'alias (qui en français n'ont pas de varia-

12. La liste complète des catégories d'alias est donnée en annexe 2.
13. Le serbe utilise conjointement l'alphabet latin et l'alphabet cyrillique.

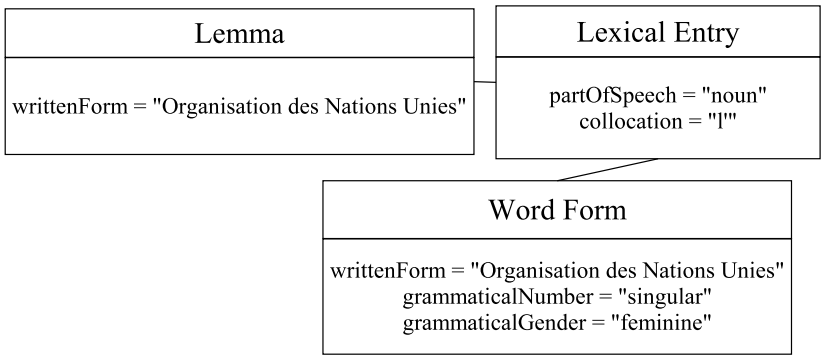

Figure 2. *Représentation du prolexème* Organisation des Nations Unies

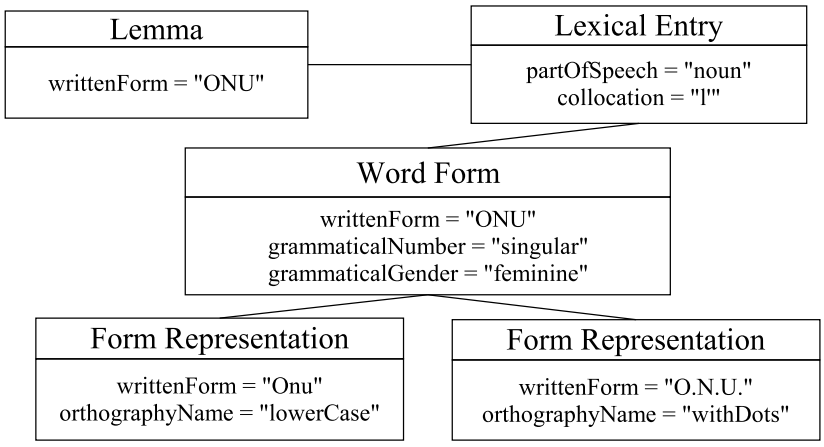

Figure 3. *Représentation de l'alias* ONU

tion flexionnelle) et quatre formes fléchies pour le dérivé : les accords en genre et en nombre d'un adjectif en français.

Enfin, l'écriture de l'alias *ONU* est raffinée en plusieurs représentations : une première avec seulement une majuscule initiale (*orthographyName = "lowerCase"*) et une seconde avec des points (*orthographyName = "withDots"*).

5.2. *Les sens*

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

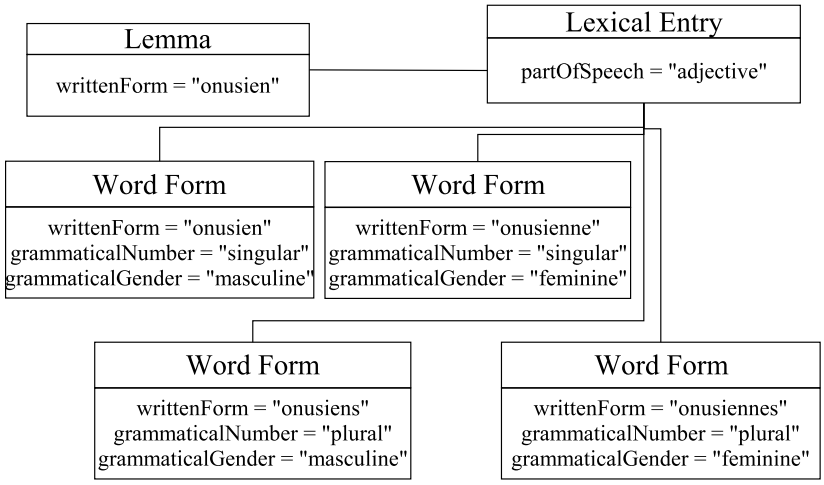

Figure 4. *Représentation de l'adjectif* onusien

– le mode de formation du terme[14] (*/term provenance/*) ;

– l'étymologie (*/etymology/*).

La figure 5 reprend les trois exemples précédents, *Organisation des Nations Unies*, *ONU* et *onusien*. Les deux premiers ont pour mode de formation leur type d'alias (*termProvenance = "fullForm"* et *termProvenance = "initialism"*) et le troisième, sa catégorie de dérivé *termProvenance = "relationalAdjective"*. Dans les trois cas, l'étymologie est l'identificateur unique du pivot correspondant (*etymology = "48 226"*). Par ce partage de la même étymologie, les différents lemmes d'un prolexème peuvent être regroupés.

De plus, les sens de l'alias et du dérivé sont en relation avec celui du représentant du prolexème. Ces relations entre sens comportent comme label le type de relation (*alias* ou *derivative*)[15]. Notons que cette façon de décrire les dérivés permet également de représenter des formes particulières de dérivé (par exemple supplétives), sous la forme d'une arborescence.

Le modèle LMF permet d'avoir une entrée lexicale avec plus d'un sens : par exemple, *Paris* pourra être une entrée lexicale avec une partie *sense* qui fait référence au pivot « capitale de la France » et une autre partie *sense* qui fait référence au pivot « ville du Texas », alors que, dans Prolexbase, les homonymes constituent des entrées

14. Lorsque cet attribut prend la valeur */quasi-synonym/*, nous ajoutons un attribut */usage/*, pour préciser le diasystème.

15. Si des expansions classifiantes ou des éponymes sont présents dans le lexique, nous ajoutons de même des relations entre les sens dont l'attribut *label* prend les valeurs *context* ou *eponymy*.

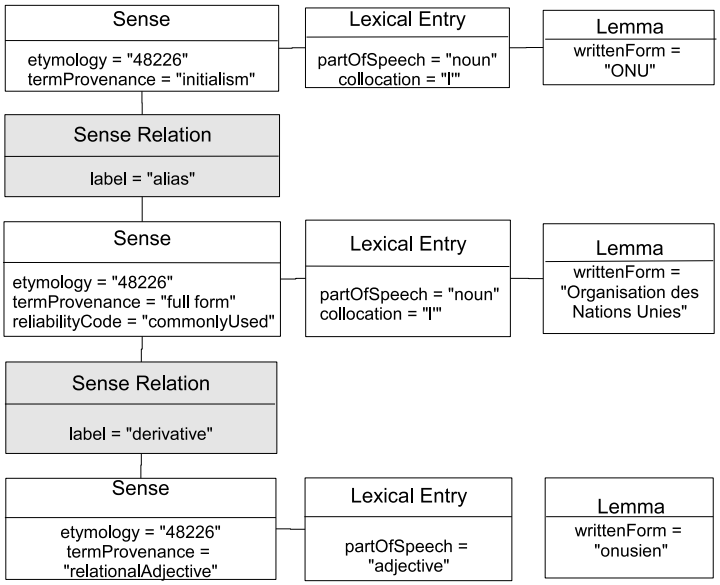

Figure 5. *Représentation des relations entre les sens associés au prolexème* Organisation des Nations Unies, *à l'alias* ONU *et à l'adjectif* onusien

distinctes[16]. Cependant, dans LMF si les différents sens ne sont pas reliés aux mêmes formes il faut créer différentes entrées lexicales. Par exemple, il y a une entrée pour *la France* et une pour *le France*. Nous ajoutons sur le sens du prolexème un attribut de notoriété (*reliabilityCode = "commonlyUsed"*[17]), car celle-ci dépend du sens, en cas d'homonyme (*Paris* est plus connu comme la capitale de la France que comme une ville du Texas...).

Pour résumer, presque tous les alias et tous les dérivés d'un prolexème sont représentés par une entrée lexicale ; ces entrées sont reliées entre elles *via* leurs sens, formant une arborescence dont la racine est le représentant du prolexème. En français, cette arborescence n'a, en général, qu'un niveau sous la racine, mais lorsqu'il y a des alias ou des dérivés de dérivé, ils sont représentés par des niveaux supplémentaires, tous reliés au pivot du prolexème *via* l'attribut d'étymologie. Ainsi, pour traduire *beograna-ninov*, le traducteur pourra savoir qu'il s'agit de l'adjectif possessif du nom relationnel (masculin) du prolexème serbe relié au même pivot que le prolexème français *Belgrade* et suggérer *d'un Belgradois*.

16. Notons que l'homonymie est une relation dépendante de la langue. En anglais, le nom propre *London* correspond à la capitale anglaise, mais aussi à une ville de l'Ontario ; alors qu'en français, ces deux noms ne sont pas homonymes (*Londres* versus *London*).
17. Les valeurs prises par cet attribut sont détaillées en annexe 2.

6. Description multilingue

Nous représentons grâce à l'extension multilingue les pivots, ainsi que leurs relations. Rappelons en effet que, par exemple, le pivot 38558 (*Paris*) est en relation de synonymie (diaphasique) avec le pivot de *Ville lumière*, en relation de méronymie avec le pivot de *Île-de-France* et en relation d'accessibilité (repérage « capitale ») avec le pivot de *France*. Les pivots mettent également en relation les prolexèmes de différentes langues. Par exemple le pivot 38558 relie le prolexème *Paris* en français et le prolexème *Paris* en anglais (figure 1).

La figure 6 illustre ces liens : il faut bien saisir que les classes *Lexicon* et *Sense Axis* se rattachent toutes deux au niveau de la ressource lexicale. Les différentes classes de ProlexLMF sont indiquées en italique sur la gauche de la figure. On y voit les entrées lexicales correspondant au prolexème *Paris*, à son dérivé (nom relationnel) *Parisien* et au prolexème *France*, en français et en anglais. Chaque relation entre sens (ici dérivé de type nom relationnel) est représentée par un point et deux flèches pointant vers les sens liés. De plus, chaque prolexème d'une langue est relié à un pivot, c'est-à-dire un *Sense Axis* en LMF, *via* son étymologie ; inversement chaque *Sense Axis* peut référencer le représentant du prolexème dans différentes langues. Ce lien est représenté dans la figure par une flèche à double sens. La relation d'accessibilité entre *Paris* et *France* est implantée *via* la relation entre les pivots correspondants ; dans la figure 6, cette relation est également représentée par un point et deux flèches pointant vers les pivots.

Les valeurs pour la catégorie de données /*label*/ qui caractérise les relations au niveau des pivots sont les suivantes :

– */quasi-synonym/* pour la synonymie ;

– */partitive relation/* pour la méronymie ;

– */associative relation/* pour l'accessibilité ;

– */generic relation/* pour l'hyperonymie.

La figure 7 montre l'exemple de la relation d'accessibilité que nous avons schématisée par un point et deux flèches vers les pivots 38558 et 27 dans la figure 6. Cette relation associe donc le pivot de *Paris* et le pivot de *France* ; elle a pour label *associativeRelation* et pour domaine (*/subject field/*) *capital*. Les relations de synonymie et de méronymie sont représentées selon le même principe. Pour la synonymie, il est possible de préciser l'indicateur diasystématique *via* la catégorie de données */usage/*.

Il est aussi prévu dans le modèle LMF de faire un lien vers une description extérieure à la ressource, par la classe *Interlingual External Ref*. Nous avons donc conservé tels quels notre typologie et notre paradigme d'existence, en utilisant les catégories de données */external system/* et */external reference/*. Par exemple, la figure 8 montre que le pivot de *Paris* a pour type *City*.

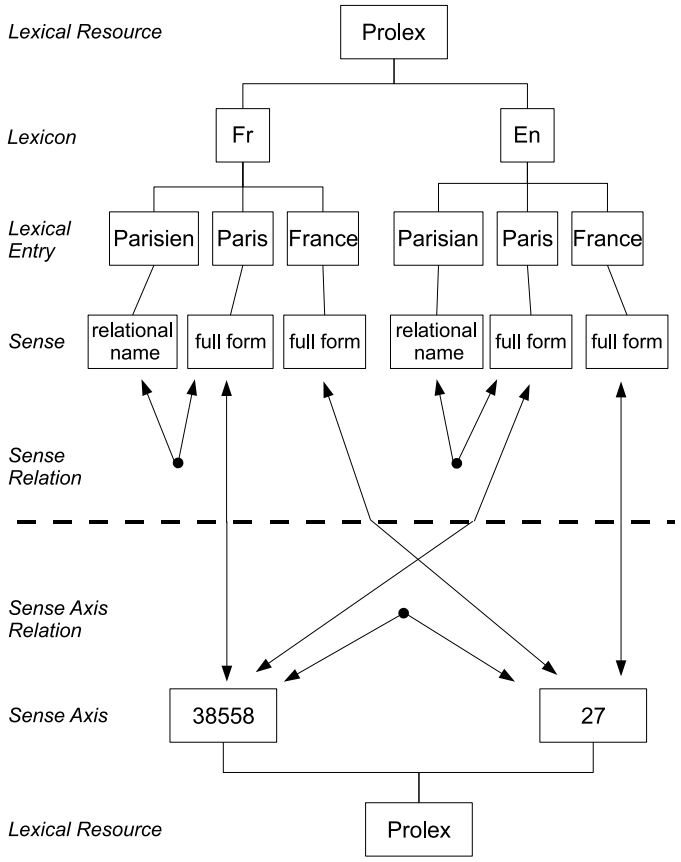

Figure 6. *Vue générale des relations sémantiques et multilingues*

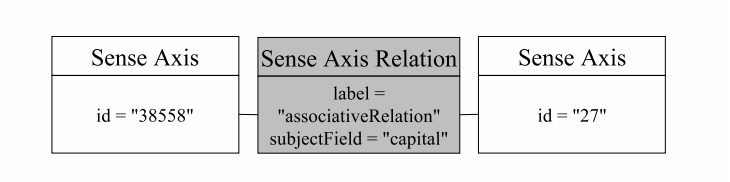

Figure 7. *Une instance de la relation d'accessibilité*

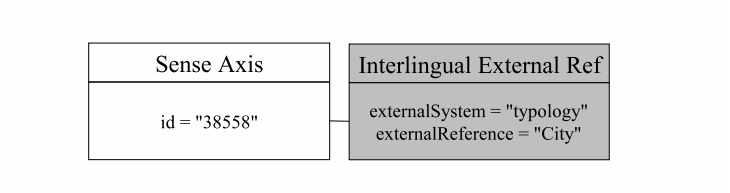

Figure 8. *Une instance du lien vers la typologie*

7. Des modèles de Prolexbase

Comme tout système d'information conçu d'une part pour refléter le plus fidèlement possible ce qu'il représente, et d'autre part pour être exploitable et maintenable, Prolexbase dispose de plusieurs modèles, liés, formant une hiérarchie d'abstraction[18].

Les recherches en modélisation des données ont établi, depuis la seconde moitié des années 60, les avantages, et même les nécessités, pour un système d'information de s'appuyer sur une telle hiérarchie de modèles, qui comporte en général trois niveaux : conceptuel (le plus abstrait), logique (intermédiaire) et physique (le plus proche de la machine) (Elmasri et Navathe, 2007). Ces niveaux correspondent en particulier à différents types d'acteurs interagissant avec le système d'information : les utilisateurs se servent du modèle conceptuel (souvent même de certaines vues particulières de ce modèle), les analystes-développeurs manipulent les concepts du niveau logique et les administrateurs utilisent le modèle physique pour l'optimisation du fonctionnement et de la maintenance du système.

Le modèle conceptuel est celui qui nous permet (nous, êtres humains) de raisonner sur le système d'information, donc de travailler avec cette ressource. Il convient que le modèle conceptuel soit un reflet des plus fidèles du réel qu'il représente : le modèle conceptuel de Prolexbase a été rappelé en section 2. La réflexion qui a abouti à ce modèle conceptuel ne pouvait que relever de la linguistique, puisque le réel est ici la langue naturelle ; ses principes et ses détails ont été publiés dans (Tran et Maurel, 2006). C'est lui que nous exprimons dans le cadre proposé par LMF ; ProlexLMF est un modèle conceptuel.

Le modèle physique sert à optimiser l'organisation, le stockage et la maintenance des données, en particulier la maintenance de la cohérence (une facette importante de la qualité) des données : le modèle physique de Prolexbase est une base de données relationnelle (celui du système de gestion de bases de données MySQL (Dubois *et al.*, 2004)). Les motivations et les caractéristiques techniques de ce modèle ont été décrites dans (Tran, 2006).

18. L'abstraction s'entend ici dans son sens classique en informatique : niveau d'abstraction par rapport au codage binaire.

Par définition, le modèle logique permet aux développeurs d'exploiter les informations. Le modèle logique de Prolexbase est, par défaut, le modèle relationnel (Codd, 1970). Pourtant, le modèle relationnel n'est pas largement utilisé par les développeurs spécialistes du TAL or Prolexbase est une ressource pour le TAL. De plus, notre objectif est d'offrir la ressource en accès libre sur le Web, et le standard de représentation des données dans ce cadre est XML. Les outils de manipulation des données XML (pour exécuter des requêtes en particulier) se servent couramment du schéma de ces données pour optimiser leurs traitements. Ainsi, dans ce cadre, le modèle logique est fourni par le schéma XML. Pour être utile au TAL, Prolexbase doit donc être muni d'un schéma XML aussi naturel que possible pour les chercheurs et développeurs en TAL.

Les initiatives de l'ISO/TC 37 pour définir un ensemble de standards pour la création et l'utilisation des ressources linguistiques vont dans le sens d'une adhésion de la partie la plus large possible de la communauté TAL à un cadre commun de description, et, par suite, d'exploitation de ces ressources. Comme il est courant dans le domaine de la normalisation, la démarche du projet LMF consiste à réaliser une synthèse des meilleures pratiques pour les ressources lexicales. LMF représente clairement un cadre pour l'expression des modèles conceptuels des ressources lexicales. Nous avons exploré des pistes pour exprimer le modèle conceptuel de Prolexbase dans ce cadre, et nous proposons dans cet article une solution qui montre que ce projet LMF est arrivé à un niveau de maturité suffisant pour qu'une ressource comme Prolexbase rentre dans son cadre sans y perdre ses spécificités.

Il ressort de nos analyses que la seule chose qui peut être considérée comme une difficulté de LMF pour Prolexbase est la conservation de la notion même de prolexème, qui regroupe l'ensemble des mots d'une langue liés à un certain point de vue sur un référent (nom propre). Le pivot, quant à lui, trouve un correspondant direct dans la même notion prévue dans LMF. Le fait que, pour Prolexbase, la sémantique se décline essentiellement au niveau interlingue ne pose pas de problème non plus dans LMF, grâce à la possibilité de définir et paramétrer des relations au niveau des pivots. En ce qui concerne le prolexème dans ProlexLMF, au niveau d'un lexique (c'est-à-dire d'une langue donnée) on y accède par n'importe lequel de ses constituants, dont le sens fait référence au pivot (attribut */etymology/*) ; on peut reconstituer l'ensemble du prolexème soit grâce à ce pivot (en sélectionnant toutes les entrées lexicales dont les sens partagent la même valeur pour */etymology/*), soit en suivant les liens qui relient les sens.

Nous nous étions d'abord penchés sur TMF (Terminological Markup Language) (Romary, 2002) du fait de son approche onomasiologique, regroupant les termes sous le concept auquel ils se rattachent[19], qui correspond naturellement à l'or-

19. Les structures des descriptions lexicales peuvent organiser la relation entre les mots et les sens en privilégiant les uns ou les autres. LMF suit une approche sémasiologique en associant l'entrée lexicale aux mots (ou groupes de mots) et en considérant les sens comme des sous-éléments de ces entrées lexicales. TMF permet, quant à lui, une approche onomasiologique, regroupant les mots sous les concepts (ou termes).

ganisation pivot/prolexème de Prolexbase. Cependant TMF n'offre pas la richesse de description de LMF ; or, les noms propres ne sont pas de simples listes de termes, par exemple, ils ont un comportement linguistique qui se rapproche de celui des mots composés lorsqu'il s'agit de les fléchir (même s'il y a plus de phénomènes de figement pour les noms propres). Il est en effet important de rappeler le double aspect du modèle de Prolexbase, qui met l'accent d'une part sur la description ontologique des noms propres et d'autre part sur leur morphologie : en plus des flexions il permet de représenter les dérivations, les variantes (alias), les flexions des dérivations, les flexions des alias, les dérivations des dérivations (et des alias), etc. Prolexbase est une ressource lexicale, qui contient des descriptions morphologiques, syntaxiques et sémantiques, dont le modèle conceptuel nécessite un cadre comparable à celui des autres ressources lexicales.

Rappelons que notre objectif en décrivant le modèle conceptuel de Prolexbase dans un format LMF est d'aboutir à un modèle logique qui facilite les développements autour de cette ressource. Nous remarquons que, bien que se situant au niveau conceptuel, en favorisant l'accès direct aux lemmes (entrées lexicales) un modèle compatible LMF augmente l'efficacité des applications telles que la recherche d'information, l'aide à la traduction, l'alignement, l'extraction terminologique, etc. Dans ProlexLMF, l'accès du pivot vers le prolexème (implanté simplement par l'étymologie) favorise aussi des applications axées sur les concepts (à la TMF[20]).

ProlexLMF est donc une étape essentielle pour rattacher à Prolexbase un schéma XML aussi efficace et aussi naturel que possible pour les chercheurs et développeurs en TAL. Mais passer d'un modèle conceptuel à un modèle logique correspondant est moins évident pour une cible de type schéma XML que pour une cible de type base de données relationnelle (Bouchou *et al.*, 2005). Si un schéma XML accompagne la description de LMF, c'est loin d'être le seul possible, les contributeurs à LMF le rappellent eux-mêmes. En fait, les possibilités offertes par XML sont au moins aussi nombreuses que celles que l'on aurait avec un modèle orienté objet, or Christian Soutou analyse par exemple, dans (Soutou, 2002), pas moins de huit possibilités de représenter dans une base de données orientée objets une association « un à plusieurs » entre deux classes d'un modèle conceptuel... Les choix parmi toutes ces possibilités s'opèrent selon un critère opérationnel, c'est-à-dire en fonction des traitements dont le modèle logique doit être le support. C'est à ce niveau-là (le schéma XML) que doit intervenir la discussion initiée par Éric Laporte dans (Laporte, 2007), concernant l'adéquation des formats standard de lexiques aux applications qui utilisent ces lexiques. Nous présentons en annexe une DTD issue de ProlexLMF, que nous utilisons pour une vue XML de Prolexbase proposée sur le site du CNRTL[21], mais une discussion des avantages et inconvénients de cette DTD échappe au cadre de cet article.

20. À ce propos il nous paraît intéressant de relever le fait que notre représentation de l'organisation pivot/prolexème dans ProlexLMF peut être directement reprise pour adapter une ressource en TMF à un format LMF.

21. Un exemple en XML (Paris) se trouve également en annexe.

8. Conclusion

Après avoir présenté à la fois le dictionnaire de noms propres multilingue muni de relations Prolexbase et le projet de norme LMF (*Language resource management, Lexical markup framework*), nous avons montré qu'il est possible de disposer d'une vision de Prolexbase conforme à LMF. Ceci bien que l'approche entre les deux projets soit *a priori* différente, puisque Prolexbase privilégie l'aspect conceptuel du nom propre plutôt que les entrées lexicales qui lui correspondent dans chaque langue.

Cette réorganisation est possible essentiellement par l'utilisation de l'extension multilingue de LMF, elle-même dépendante de l'extension sémantique. La représentation interne à chaque langue s'appuie sur le noyau, l'extension morphologique et l'extension sémantique. Au passage, cela nous a aussi amenés à remplacer certains noms d'attributs ou certaines valeurs utilisés dans Prolexbase par des catégories de données conformes à la norme ISO 12620. En effet, plus nous avancions dans notre étude de LMF (relativement à Prolexbase) et plus l'importance d'un langage de description standard s'est imposée.

Prolexbase regroupe tous les niveaux de description (morphologie, syntaxe et sémantique) ; son originalité est de permettre de décrire finement à la fois l'ontologie des noms propres et leur morphologie. Dans cet article nous avons axé la présentation sur les formes fléchies, pour autant Prolexbase s'appuie également sur des ensembles de règles de flexion, qui peuvent être exprimées dans l'extension morphologique proposée par LMF (cette expression est en cours de mise au point). L'ajout d'autres caractéristiques dans Prolexbase est encore prévu, comme la description des contextes d'apparition de certains noms propres (grammaires locales) par exemple : là encore, le cadre proposé par LMF pour la description syntaxique sera utile. Ces deux exemples justifient notre motivation pour l'utilisation de LMF, dans le sens où ce projet de norme se révèle offrir un cadre à la fois assez complet et assez souple pour modéliser une ressource lexicale telle que Prolexbase.

Remerciements
Les auteurs remercient Agata Savary pour ses remarques pertinentes, ainsi que les relecteurs anonymes qui ont beaucoup contribué à l'amélioration de cet article.

9. Bibliographie

Bouchou B., Tran M., Maurel D., « Towards an XML Representation of Proper Names and Their Relationships », *NLDB'2005, in Lecture Notes in Computer Science, 3513*, p. 44-55, 2005.

Codd E., « A Relational Model of Data for Large Shared Data Banks », *CACM*, 1970.

Coseriu E., « Le double problème des unités dia-s », *Les Cahiers dia. Etudes sur la diachronie et la variation linguistique, Université de Gent, Belgique*, vol. 1, p. 9-16, 1998.

Dubois P., Hinz S., Pedersen C., *MySQL : Guide officiel*, CampusPress, Paris, 2004.

Elmasri R., Navathe S., *Conception et architecture des bases de données*, Pearson Education, 4e édition, 2007.

Fellbaum C., Miller G. A., « Morphosemantic Links in WordNet », *TAL 44*, vol. 2, p. 69-80, 2003.

Francopoulo G., Declerck T., Monachini M., Romary L., « The relevance of standards to research infrastructure », *LREC*, 2006a.

Francopoulo G., Monte G., Calzolari N., Monachini M., Bel N., Soria C., « Lexical Markup Framework (LMF) », *LREC 2006, Genoa, Italie*, 2006b.

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

Krstev C., Stanković R., Vitas D., Obradović I., « WS4LR - a Workstation for Lexical Resources », *LREC 2006, Genoa, Italie*, p. 1692-1697, 2006.

Laporte E., « Lexicon management and standard formats », *Cornell University Library, http ://arxiv.org/abs/0711.3449v1*, 2007.

Maurel D., Tran M., Friburger N., « Projet Technolangue NomsPropres : Constitution et exploitation d'un dictionnaire relationnel multilingue de noms propres », *TALN 2006, Cahiers du Cental, Louvain, Belgique*, p. 927-936, 2006.

Maurel D., Vitas D., S. S. K., Koeva S., « Prolex : a lexical model for translation of proper names. Application to French, Serbian and Bulgarian », *Bulag*, vol. 32, p. 55-72, 2007.

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

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

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

Salmon-Alt S., « $V^1\Omega a = able$ ou «Normaliser des lexiques TAL est délectable» », *TALN*, 2006.

Salmon-Alt S., Akrout A., Romary L., « Proposals for a normalized representation of Standard Arabic full form lexica », *Second International Conference on Machine Intelligence (ACIDCA-ICMI)*, 2005.

Savary A., « A formalism for the computational morphology of multi-word units », *Archives of Control Sciences, 15(LI), Silesian University of Technology*, 2005.

Silberztein M., « Nooj : A cooperative Object Oriented Architecture for NLP », *Cahiers de la MSH Ledoux, Série Archive, Bases, Corpus*, vol. 1, p. 351-361, 2004.

Soutou C., *De UML à SQL : conception de bases de données*, Eyrolles, Paris, 2002.

Tran M., *Prolexbase, un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne*, thèse de doctorat en informatique de l'Université François Rabelais de Tours, 2006.

Tran M., Maurel D., « Prolexbase : un dictionnaire relationnel multilingue de noms propres », *TAL*, vol. 47, n° 3, p. 115-139, 2006.

Annexe 1. Les termes de Prolexbase et les catégories de données

Le tableau 1 est le résultat de notre analyse des catégories de données. Nous avons recherché celles qui correspondent aux termes que nous utilisons pour représenter dans Prolexbase les noms propres, leurs dérivés et leurs relations. Par exemple, la catégorie de dérivé sera comprise comme le /mode de formation du terme/ et sera notée /term provenance/ dans le schéma XML.

Prolex	**Data Categories**		
	Position	**Français**	**Anglais**
Langue	A10.7	indicatif de la langue	language symbol
Prolexème et alias	A.2.1.1	entrée principale	main entry term
Dérivé	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 (Blark)	A.2.3.4	fréquence	reliability code
Phonétique	A.2.5	prononciation	pronunciation
Expansion classifiante	A.5.3	contexte	context
Détermination (contrainte)	A.2.1.18.1	cooccurent	collocation
Classe	A.2.2.1	catégorie grammaticale	part of speech
Morphologie	A.2.2	morphologie	morphology
Antonomase	A.2.4.1	mode de formation du terme	term provenance
Terminologie	A.2.4.1	mode de formation du terme	term provenance
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

Tableau 1. *Correspondance entre les termes de Prolexbase et les catégories de données*

Annexe 2. Adaptations de Prolexbase aux catégories de données (types d'alias et notoriété)

L'analyse des catégories de données a mené également à certaines adaptations dans Prolexbase. Ainsi, il ressort du tableau 2 qu'il faut ajouter dans Prolexbase la catégorie d'alias *nom usuel* et diversifier les catégories *Abréviation* et *Sigle ou acronyme*.

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	A.2.1.13	quasi-synonyme	quasi synonym
Synonyme diastratique			
glose	A.5.2	explication	explanation

Tableau 2. *Correspondance entre les catégories d'alias de Prolexbase et les type de terme des catégories de données*

De même, les attributs de notoriété de Prolexbase (voir le tableau 3) ont été modifiés et sont passés de quatre degrés à trois, pour suivre les recommandations de l'ISO 12620.

Prolex	**Data Categories**		
	Position	**Français**	**Anglais**
International	A.2.3.4	fréquent	commonly used
Européen			
National		peu fréquent	unfrequently used
Détails		rare	rarely used

Tableau 3. *Correspondance entre les attributs de notoriété de Prolexbase et ceux de fréquence des catégories de données*

Annexe 3. Schéma conceptuel de ProlexLMF (classes UML)

La figure 9 présente l'ensemble des classes de ProlexLMF (la même classe Lexical Resource est placée en haut et en bas de cette figure).

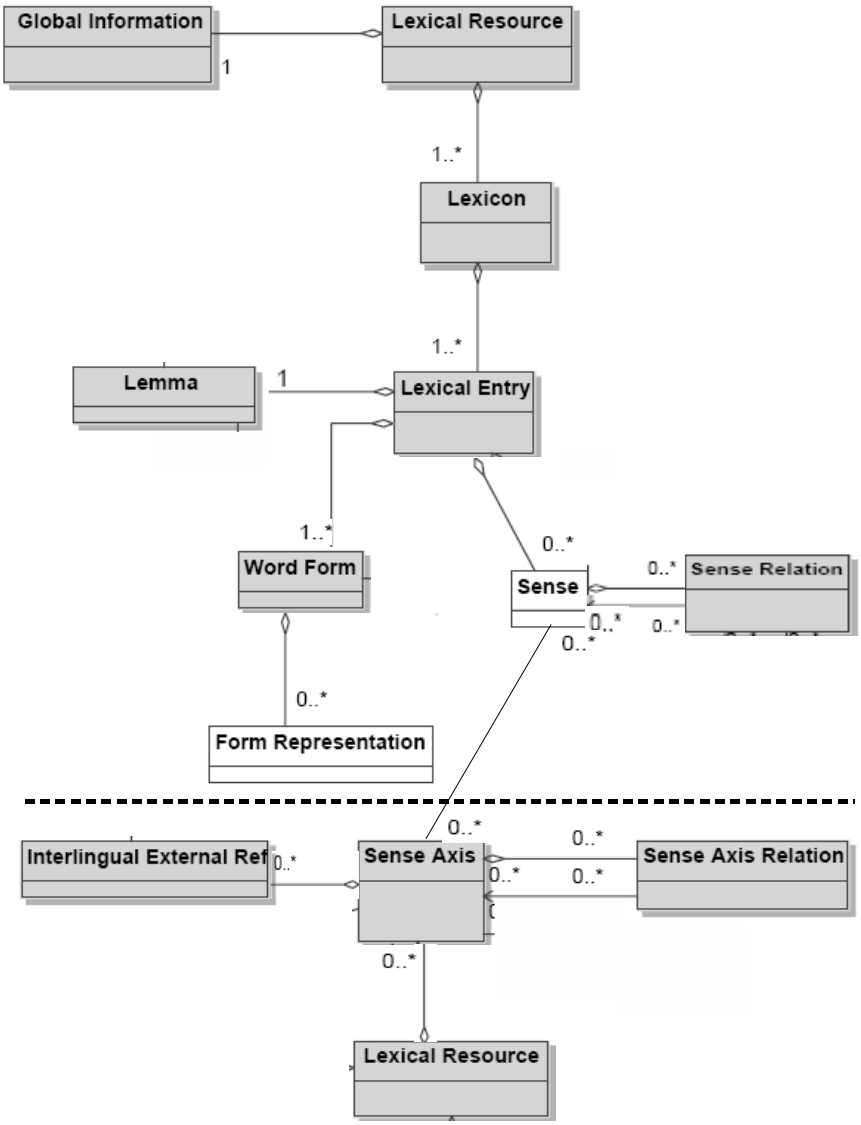

Figure 9. *Les classes LMF utilisées par Prolexbase*

Annexe 4. Exemples d'attributs pour les différentes classes utilisées dans ProlexLMF

À chacune des classes présentées sur la figure 9 de l'annexe 3 sont associés des attributs. Les noms de ces derniers ont été choisis parmi les catégories de données existantes. L'attribut /entrySource/ peut être ajouté à la classe Lexical Entry pour distinguer les entrées lexicales relevant de prolexèmes dans le cas où il y aurait dans le lexique d'autres entrées lexicales que celles provenant de Prolexbase, par exemple en cas de fusion avec une autre ressource lexicale. Le tableau 4 présente les attribut nécessaires à la description du français.

Data Categories	
Classes	**Exemples d'attributs**
Lexical Entry	partOfSpeech
	collocation
	entrySource
Lemma	writtenForm
Word Form	writtenForm
	grammaticalNumber
	grammaticalGender
	grammaticalCase
	grammaticalTense
	grammaticalMood
	person
Form representation	writtenForm
	orthographyName
Sense	etymology
	termProvenance
	usage
	reliabilityCode
Sense Relation	label
Sense Axis	id
Sense Axis Relation	label
	subjectField
	usage
Interlingual External Ref	externalSystem
	externalReference

Tableau 4. *Exemples d'attributs pour les différentes classes utilisées dans Prolexbase*

Annexe 5. Une DTD de ProlexLMF

Contrairement à ce qui est préconisé dans le cadre TMF et LMF[22], nous explicitons les attributs des classes sous la forme d'attributs XML[23].

```
<!DOCTYPE LexicalResource [
    <!ELEMENT LexicalResource (GlobalInformation, Lexicon+, SenseAxis*)>
    <!ELEMENT GlobalInformation ([...])>
    <!ELEMENT Lexicon (LexicalEntry+)>
        <!ATTLIST Lexicon languageSymbol CDATA #REQUIRED>
    <!ELEMENT LexicalEntry (Lemma, WordForm+, Sense+)>
        <!ATTLIST LexicalEntry partOfSpeech CDATA #REQUIRED>
        <!ATTLIST LexicalEntry entrySource CDATA #IMPLIED "prolex">
        <!ATTLIST LexicalEntry collocation CDATA #IMPLIED>
    <!ELEMENT Lemma ()>
        <!ATTLIST Lemma writtenForm CDATA #REQUIRED>
    <!ELEMENT WordForm (FormRepresentation*)>
        <!ATTLIST WordForm writtenForm CDATA #REQUIRED>
        <!ATTLIST WordForm grammaticalNumber CDATA #IMPLIED>
        <!ATTLIST WordForm grammaticalGender CDATA #IMPLIED>
        [...]
    <!ELEMENT FormRepresentation ()>
        <!ATTLIST FormRepresentation writtenForm CDATA #REQUIRED>
        <!ATTLIST FormRepresentation orthographyName CDATA #IMPLIED>
    <!ELEMENT Sense (SenseRelation*)>
        <!ATTLIST Sense id ID #REQUIRED>
        <!ATTLIST Sense etymology CDATA #REQUIRED>
        <!ATTLIST Sense termProvenance CDATA #REQUIRED>
        <!ATTLIST Sense usage CDATA #IMPLIED>
        <!ATTLIST Sense reliabilityCode CDATA #IMPLIED>
    <!ELEMENT SenseRelation ()>
        <!ATTLIST SenseRelation label CDATA #REQUIRED>
        <!ATTLIST SenseRelation targets IDREFS #REQUIRED>
    <!ELEMENT SenseAxis (SenseAxisRelation*, InterlingualExternalRef*)>
        <!ATTLIST SenseAxis id ID #REQUIRED>
    <!ELEMENT SenseAxisRelation ()>
        <!ATTLIST SenseAxisRelation label CDATA #REQUIRED>
        <!ATTLIST SenseAxisRelation targets IDREFS #REQUIRED>
        <!ATTLIST SenseAxisRelation subjectField CDATA #IMPLIED>
        <!ATTLIST SenseAxisRelation usage CDATA #IMPLIED>
    <!ELEMENT InterlingualExternalRef ([...])> ] >
```

22. N'introduire les caractéristiques des classes que sous forme de couples d'attributs (*att*, *val*) d'un élément générique appelé *feat*.

23. Ceci afin de détecter les erreurs d'éléments ou d'attributs à l'aide des validateurs XML.

Annexe 6. L'entrée *Paris* dans le format XML de l'annexe 5

```
< ?xml version="1.0" encoding="utf-8" ?>
<LexicalResource>
    <GlobalInformation type="Prolex"> [...] </GlobalInformation>
    <Lexicon languageSymbol="Fr">
        [...]
        <LexicalEntry partOfSpeech="noun">
            <Lemma writtenForm="Paris"/>
            <WordForm writtenForm="Paris" grammaticalGender="masculine
                feminine" grammaticalNumber="singular"/>
            <Sense id="75000" etymology="38558" termProvenance="fullForm"
                reliabilityCode="commonlyUsed">
                <SenseRelation label="derivative to" targets="75001 [...]"/>
            </Sense>
        </LexicalEntry>
        [...]
        <LexicalEntry partOfSpeech="adjective">
            <Lemma writtenForm="parisien"/>
            <WordForm writtenForm="parisien" grammaticalGender="masculine"
                grammaticalNumber="singular"/>
            <WordForm writtenForm="parisiens" grammaticalGender="masculine"
                grammaticalNumber="plural"/>
            <WordForm writtenForm="parisienne" grammaticalGender="feminine"
                grammaticalNumber="singular"/>
            <WordForm writtenForm="parisiennes" grammaticalGender="feminine"
                grammaticalNumber="plural"/>
            <Sense id="75001" etymology="38558"
                termProvenance="relationalAdjective" reliabilityCode="commonlyUsed">
                <SenseRelation label="derivative from" targets="75000"/>
            </Sense>
        </LexicalEntry>
        [...]
        <LexicalEntry partOfSpeech="noun" collocation="la">
            <Lemma writtenForm="France"/>
            <WordForm writtenForm="France" grammaticalGender="feminine"
                grammaticalNumber="singular"/>
            <Sense id="33" etymology="27" termProvenance="fullForm"
                reliabilityCode="commonlyUsed">
                <SenseRelation label="derivative to" targets="[...]"/>
            </Sense>
        </LexicalEntry>
```

```
        [...]
    </Lexicon>
    <Lexicon languageSymbol="En">
        [...]
        <LexicalEntry partOfSpeech="noun">
            <Lemma writtenForm="Paris"/>
            <WordForm writtenForm="Paris" grammaticalNumber="singular"/>
            <Sense id="75000" etymology="38558" termProvenance="fullForm"
                reliabilityCode="commonlyUsed">
                <SenseRelation label="derivative to" targets="[...]"/>
            </Sense>
        </LexicalEntry>
        [...]
    </Lexicon>
    [...]
    <SenseAxis id="38558">
        <SenseAxisRelation label="AssociativeRelation to" targets="27"
            subjectField="capital"/>
        [...]
        <InterlingualExternalRef externalSystem="typology"
            externalReference="city"/>
    </SenseAxis>
    [...]
</LexicalResource>
```