Prolexbase
Un dictionnaire relationnel multilingue de noms propres

Mickaël Tran, Denis Maurel

Université François Rabelais Tours
Laboratoire d'informatique
EPU-DI, 64, avenue Jean-Portalis, 37200 Tours, France
denis.maurel ad univ-tours.fr

RÉSUMÉ. *Cet article présente la modélisation du domaine des noms propres définie dans le projet Prolex. Celle-ci repose sur deux concepts centraux : le nom propre conceptuel et le prolexème. Le nom propre conceptuel ne représente pas le référent, mais un point de vue sur ce référent. Il possède dans chaque langue un concept spéficique, le prolexème, qui est une famille structurée de lexèmes. Autour d'eux, nous avons défini d'autres concepts et des relations (synonymie, méronymie, accessibilité, éponymie, etc.). Chaque nom propre conceptuel est en relation d'hyperonymie avec un type et une existence au sein d'une ontologie.*

ABSTRACT. *This paper presents the modelling of Proper Name domain defined by the Prolex project. This modelling is based on two main concepts: the Conceptual Proper Name and the Prolexeme. The Conceptual Proper Name do not represents the referent, but a point of view on this referent. It has a specific concept in each language, the Prolexeme, that is a structured family of lexemes. Around them, we have defined other concepts and relations (synonymy, meronymy, accessibility, eponymy...). Each Conceptual Proper Name is an hyponym of a type and an existence within an ontology.*

MOTS-CLÉS : *nom propre, synonymie, alias, dérivé, méronymie, accessibilité, typologie, prolexème.*

KEYWORDS: *proper name, synonymy, alias, derivative, meronymy, accessibility, typology, prolexeme.*

TAL Volume 47 – n°3/2006, pages 115 à 139

1. Introduction

Les ressources linguistiques sont indispensables aux applications du TAL, mais la nature et la taille de ces ressources dépendent largement des méthodes ou logiciels utilisés. Il existe aujourd'hui de nombreuses ressources dictionnairiques ou lexicales de noms communs (comme par exemple les dictionnaires électroniques du LADL (Courtois, 1992), Wordnet (Miller, 1995), Morphalou (Romary *et al.*, 2004), le projet Papillon (Mangeot-Lerebours *et al.*, 2003), etc.) et des ressources terminologiques spécialisées, mais aussi de nombreuses listes de noms propres, souvent multilingues, comme par exemple le dictionnaire CJK[1] avec plus de 150 000 noms propres, EuroGeographics[2], News Explorer[3], etc. Notre travail ne consiste pas en la création de listes supplémentaires mais en celle d'un dictionnaire contenant des informations syntaxiques, morphologiques, sémantique, etc.

Faut-il créer des ressources spécifiques aux noms propres ? Les avis des chercheurs sur cette question sont très divisés. Pour (Mikheev *et al.*, 1999), un système de règles couplé avec une liste de mots liés aux noms propres suffit. D'autres, comme (Ren et Perrault, 1992), considèrent que tout mot inconnu capitalisé peut être classé comme un nom propre. Dans le cadre de nos travaux, le projet Prolex, nous soutenons l'idée que la constitution de ressources lexicales, multilingues et relationnelles de noms propres est nécessaire pour leur traitement automatique. Les recherches présentées ici ont reçu un soutien financier du ministère de l'Industrie dans le cadre d'un projet Technolangue (Maurel *et al.*, 2006) et les ressources créées sont mises gratuitement à la disposition de la communauté TAL, sur le site du CNRTL[4], sous une licence LGPLLR (la *Lesser General Public License For Linguistic Resources*[5]).

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 qui inclut 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 MUC6.

Une simple liste ne suffit pas, il est nécessaire de relier les noms propres entre eux (Maurel *et al.*, 2000) et des les associer à différentes formes morphologiques. Ainsi, *Paris* sera associé à *Parisien*, *parisien*, *Parigot*, etc. et en relation avec *France*, *Ville lumière*, *Lutèce*, etc. Cette remarque peut d'ailleurs s'appliquer aussi aux listes de termes qui possèdent souvent de nombreuses relations entre eux. De ce fait, nous nous

1. http ://www.cjk.org
2. http ://www.eurogeographics.org
3. http ://press.jrc.it/NewsExplorer/home/fr/latest.html
4. http://www.cnrtl.fr/lexiques/prolex/
5. http://igm.univ-mlv.fr/ unitex/lgpllr.html

sommes inspirés, pour l'échange de nos données, de la norme ISO 16642 (Romary, 2002) qui propose un standard de représentation pour des données terminologiques multilingues en XML.

La section 2 présente les concepts clés de notre modélisation des noms propres. La section 3 décrit les relations existant dans notre dictionnaire : la synonymie, la méronymie, l'accessibilité, l'expansion classifiante et l'éponymie. La section 4 est consacrée à notre typologie et la section 5 à notre ontologie. Suivent, section 6, quelques chiffres sur la constitution du dictionnaire à la fin du projet Technolangue.

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

– le niveau méta-conceptuel : la typologie et l'existence ;

– le niveau conceptuel : le nom propre conceptuel et les relations qui ne dépendent pas de la langue ;

– le niveau linguistique : le prolexème, les alias, les dérivés et les relations qui dépendent de la langue ;

– le niveau des instances : l'ensemble des formes fléchies d'une langue.

2. Les concepts

2.1. *Le nom propre conceptuel*

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 Joseph Wojtyla*), en italien (*Giovanni Paolo II* et *Carol Josef Wojtyla*), etc. Ces deux noms correspondent tous les deux à un certain point de vue sur un même et unique référent.

Nous définissons donc le nom propre conceptuel non pas comme le référent, mais plutôt, comme un certain point de vue sur celui-ci. Ainsi les noms propres *Allemagne* en français, *Alemania* en espagnol, *Deutschland* en allemand, etc. seront associés à un même nom propre conceptuel, tandis que les noms propres *République fédérale d'Allemagne* en français, *República Federal de Alemania* en espagnol, *Bundesrepublik Deutschland* en allemand, etc. seront associés à un autre nom propre conceptuel. Ces deux noms propres conceptuels seront en relation de synonymie.

Pour définir ces différents points de vue, nous nous sommes basés sur un marquage diasystématique, qui provient des travaux sur la métalexicographie de (Coseriu, 1998) qui propose un diasystème basé essentiellement sur quatre variétés distinctes (tableau 1).

Diachronique	variété dans le temps
Diaphasique	variété concernant les finalités de l'emploi
Diatopique	variété dans l'espace
Diastratique	variété relative à la stratification socioculturelle

Tableau 1. *Le diasystème de Coseriu*

Le nom propre conceptuel nous servira de pivot entre différentes langues. Il sera représenté dans notre modèle par un numéro d'identité unique (ID), le pivot.

2.2. *Le prolexème*

Dans notre modèle, le prolexème correspond à une projection du nom propre conceptuel dans une langue donnée. Chaque prolexème d'une langue donnée sera donc relié à un seul et unique nom propre conceptuel. 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 de synonymes. 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.

Par exemple, les noms propres *Nations unies*, *Onusien*, *ONU* auront *Organisation des Nations unies* comme prolexème pour la langue française. Les noms propres *Onusian*, *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 nom propre conceptuel.

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 amené à créer trois prolexèmes différents. En revanche, 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 (Piton et Maurel, 2004). Cette information sera rajoutée au niveau des expansions classifiantes du prolexème (voir section 3.4).

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 à une ville du Canada ou à une ville en Angleterre, ce qui n'est pas le cas en français à cause de l'existence d'un exonyme (*Londres*).

Les différents alias provenant d'un même nom propre peuvent être considérés comme des synonymes. Ainsi, les phrases (i), (ii) et (iii) sont sémantiquement identiques, puisque le nom propre *États-unis d'Amérique* et ses deux alias, *États-unis* et *USA*, renvoient à une même et unique entité.

> *(i) De leur côté, en effet, les* ***États-unis*** *considèrent que les frontières doivent être le produit d'un accord. (Libération, 24/5/2006)*
> *(ii) De leur côté, en effet, les* ***États-unis d'Amérique*** *considèrent que les frontières doivent être le produit d'un accord.*
> *(iii) De leur côté, en effet, les* ***USA*** *considèrent que les frontières doivent être le produit d'un accord.*

Comme le note (Gross, 1997), certains dérivés peuvent être considérés comme des *synonymes transformationnels* du nom dont ils dérivent. Ces dérivés seront inclus dans notre base[6]. Ainsi, le remplacement du dérivé *brésilien* dans la phrase (iv) par le groupe prépositionnel contenant le nom propre *Brésil* (v) ne change pas le contenu sémantique de la phrase :

> *(iv) Le président* ***brésilien*** *Luiz Inacio Lula da Silva a de fortes chances d'être réélu en octobre pour un second mandat. (Libération, 23/5/2006)*
> *(v) Le président* ***du Brésil*** *Luiz Inacio Lula da Silva a de fortes chances d'être réélu en octobre pour un second mandat.*

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 existant dans une langue L_1 peut être totalement absente dans une langue L_2. Un système de traduction automatique devra alors être capable de proposer une traduction de l'alias de la langue L_1 en utilisant la traduction du nom propre associé à cet alias dans L_1. De même, pour le cas de la traduction d'un dérivé qui n'existe pas dans une langue donnée. Par exemple, le dérivé *Tourangeau* se traduira en anglais par *inhabitant of Tours*.

2.3. *Les alias*

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 synonymes approximatifs, diatopiques ou diastratiques.

Il est parfois possible de définir des règles basées sur la structure interne (MacDonald, 1996) d'un prolexème afin de générer ses différents alias, ce que nous envisageons d'étudier systématiquement dans de futurs travaux...

6. A contrario, les dérivés lexicalisés, comme *pasteuriser*, n'y figureront pas.

La formation d'un alias résulte quelquefois de la variation d'un ou plusieurs des caractères qui composent le prolexème :

– la hauteur de casse : *Peugeot* ou *PEUGEOT* ;

– l'esperluette : *Science et Vie Junior* ou *Science & Vie Junior*[7] ;

– le remplacement des lettres comportant un signe diacritique : *Épinay-sur-Seine* ou *Epinay-sur-Seine* pour le français, *München* ou *Muenchen* pour l'allemand, *Århus*, *Arhus* ou *Aarhus* pour le danois ;

– le plus, le trait d'union et l'espace : *Canal Plus* ou *Canal +* ;

– etc.

En français, la ligature n'est pas optionnelle : le mot *cœur* ne doit pas s'écrire *coeur*, pour le cas des noms propres la lexicalisation n'est pas courante, dans les textes nous pouvons trouver les deux formes, nous avons décidé de les considérer comme des variantes. Il s'agit d'une variante orthographique provenant d'une erreur sur le diacritique (gluon) porté par le e de *cœur*. Dans le cas des noms propres cette utilisation est moins stricte. Les deux noms *Crèvecœur-en-Brie* et *Crèvecoeur-en-Brie* se rencontrent.

Les alias peuvent être aussi des abréviations du prolexème, c'est-à-dire une *suppression de mots ou de lettres dans une forme plus longue désignant le même concept* [ISO 1087-1:2000] : *Jean-François Delharpe* ou *Jean-François Delaharpe*, *Organisation des Nations unies* ou *Nations unies*.

Pour la plupart des noms de célébrité, il sera possible de modéliser la création de leurs alias. Par exemple, le nom propre *François Mitterrand* pourra être associé à *F. Mitterrand* ou même à *Mitterrand*. Il existe malheureusement des exceptions parmi les noms de célébrité qui ne suivent pas ces règles. S'il est possible à partir de *François-René de Chateaubriand* de générer les alias *de Chateaubriand* et *Chateaubriand*, il ne sera par contre pas possible de créer l'alias *Gaulle* à partir de *Charles de Gaulle*. Il sera aussi peu acceptable d'appliquer ces règles à des noms propres non contemporains, comme *Marco Polo* (* ?*M. Polo*) ou *Jules César* (* ?*J. César*). Nous pouvons aussi créer des règles pour certains noms d'entreprise et pour certains noms de ville.

Parmi ces abréviations, se trouvent aussi les acronymes (*Sofres* pour *Société française d'enquêtes par sondages*) et les sigles (*OCDE* pour *Organisation de coopération et de développement économiques*). Certains sont des sigles communs à de nombreuses langues (*UNESCO* pour *Organisation des Nations unies pour l'éducation*) et d'autres sont spécifiques à certaines langues (*OEA* pour *Organisation des États américains* pour le français et *OAS*pour *Organization of American States* pour l'anglais). Il arrive même que, dans une langue, on utilise un sigle formé à partir

7. Ces deux écritures ont été trouvées sur le site de http ://www.scienceetviejunior.fr consulté le 23/05/06.

d'une autre langue, comme par exemple en français *ESA* pour *European Space Agency* et pour *Agence spatiale européenne*.

En français, les acronymes s'écrivent théoriquement toujours en majuscules ou avec une majuscule initiale (nouvelle écriture) et les sigles tout en majuscules avec des points ou sans point (nouvelle écriture). Certains acronymes et certains sigles sont aussi sujets à des variations sur certains des caractères qui les composent (*ASSEDIC* ou *Assédic* pour *Association pour l'emploi dans l'industrie et le commerce*, *INaLCO* et *INALCO* pour *Institut National des Langues et Civilisations Orientales*).

Nous avons aussi intégré les transcriptions et les translittérations dans les catégories d'alias. Une translittération est une opération qui consiste à transposer signe par signe un ou plusieurs mots écrits dans un système d'écriture vers un autre. Une transcription est souvent basée sur la phonétique.

Les transcriptions ne sont pas identiques d'une langue à l'autre, par exemple, le nom propre russe Владимир Владимирович Маяковский se transcrit :

– *Vladimir Vladimirovitch Mayakovski*, *Maïakovski* ou *Mayakovsky* en français ;

– *Vladimir Vladimirovich Mayakovsky* en anglais ;

– *Vladimir Vladimirovitsj Majakovski* en néerlandais.

Un exemple particulièrement visible est celui des noms propres chinois qui apparaissent souvent dans les textes journalistiques français. Les noms propres *Pékin* et *Mao Tsé-toung* transcrits avec l'EFEO (système mis au point par l'École française d'Extrême-Orient) sont beaucoup plus connus des Français que leur forme pinyin (système de transcription officiel du gouvernement chinois) *Beijing* et *Mao Zedong*.

En serbe, qui utilise deux alphabets, tous les mots écrits en alphabet cyrillique possèdent une transcription en alphabet latin. Par exemple, Организација уједињених нација en alphabet cyrillique se transcrit *Organizacija ujedinjenih nacija* (*Organisation des Nations unies*) en alphabet latin.

Certains alias, comme *Mère Angélique* (pour *Marie Jacqueline Angélique Arnaud*), *le Second Pitt* (pour *William Pitt*) constituent des cas discutables (des synonymes diastratiques ?). Nous avons décidé de les placer dans la partie qui dépend de la langue, car leur formation dépend souvent de la culture qui lui est liée. Enfin, nous avons regroupé dans la variété diatopique les noms propres d'une ou plusieurs langues régionales d'un même pays qui sont en relation de synonymie avec le prolexème (la ville de *Nantes* est appelée par les Bretons *Naoned*, qui doit aussi être considéré comme un nom de la langue française[8]).

8. En breton, *Noaned* serait le prolexème associé au prolexème français *Nantes*.

2.4. *Les dérivés*

Parmi les types de dérivés existant en français, nous avons principalement deux catégories : les noms relationnels, qui débutent normalement par une majuscule (*Parisien*, *Marseillais*...), et les adjectifs relationnels, en général[9] identiques au nom relationnel, à la majuscule près (*parisien*, *marseillais*...).

En français, la formation des dérivés à partir d'un prolexème ou d'un alias résulte de règles morphologiques complexes. Parfois, au lieu de s'appliquer à la base du nom propre, ces règles s'appliquent à une forme supplétive, comme *Bellifontain* pour désigner les habitants de la ville de *Fontainebleau* (Eggert, 2002).

En français, quelques noms de pays produisent des préfixes dérivés (franco, américano, etc.) provenant parfois d'une forme supplétive (hispano, luso, etc.). Il arrive quelquefois aussi qu'un prolexème possède un nom relationnel diastratique comme dérivé. C'est le cas du prolexème *Paris* avec son dérivé *Parigot*. Cette forme est souvent connotée péjorativement.

Dans certaines langues, comme le serbe, les noms relationnels et les adjectifs relationnels ne sont pas systématiquement identiques et leur création se fait par des mécanismes morphologiques totalement différents du français. Le serbe, qui est une langue beaucoup plus riche et complexe sur le plan morphologique, distingue deux catégories de noms relationnels : les noms relationnels féminins et les noms relationnels masculins. À partir de ces deux types de noms relationnels, il devient alors possible de former des adjectifs possessifs et des adjectifs relationnels. La formation des adjectifs ne se fait pas uniquement à partir des noms relationnels dérivés, mais peut aussi se faire à partir du prolexème ou des différents alias.

À partir du prolexème Београд (*Beograd* en serbe latin, *Belgrade* en français) nous pouvons obtenir les dérivés suivants :

– београдски (*beogradski*) : adjectif relationnel (*les rues belgradoises*) ;

– Београдов (*Beogradov*) : adjectif possessif (*l'allure de Belgrade*) ;

– Београђанин (*Beogradanin*) : nom relationnel masculin (*Belgradois*) ;

- београђански (*beogradanski*) : adjectif relationnel (*les habitudes des Belgradois*) ;

- Београђанинов (*Beogradjaninov*) : adjectif possessif (*la maison d'un Belgradois*) ;

– Београђанка (*Beogradanka*) : nom relationnel féminin (*Belgradoise*) ;

9. Une exception bien connue : *suisse* est l'adjectif relationnel féminin et *Suissesse* est le nom relationnel féminin provenant du prolexème *Suisse*.

- београђански (*beogradanski*) : adjectif relationnel (*les habitudes des Belgradoises*) ;

- Београђанкин (*Beogradankin*) : adjectif possessif (*la maison d'une Belgradoise*).

3. Les relations

Une fois que les différents concepts du domaine des noms propres ont été identifiés et définis, il faut rechercher les relations qui lient les noms propres entre eux. Dans cette partie, nous présenterons différentes relations linguistiques dans lesquelles peuvent intervenir des noms propres.

Les relations linguistiques qui peuvent exister entre les unités lexicales d'une langue sont essentiellement divisées en deux catégories distinctes (Polguère, 2003) : les relations paradigmatiques et les relations syntagmatiques. Lorsqu'une unité lexicale peut être substituée à une autre unité lexicale dans un même contexte, on dit alors que ces deux unités lexicales sont reliées par une relation paradigmatique. On distingue les relations paradigmatiques de similarité, comme la synonymie et l'antonymie, et les relations paradigmatiques d'inclusion, comme la méronymie et l'hyperonymie. Les relations syntagmatiques sont des relations qu'entretiennent les unités lexicales entre elles dans une même phrase selon un principe de combinaison (ou collocation).

Dans le cas des relations qui ne dépendent pas de la langue, nous avons retenu trois relations paradigmatiques (méronymie, synonymie, hyperonymie) et une relation syntagmatique (accessibilité). La relation d'hyperonymie sera étudiée en détail dans la section 4, car elle nécessite l'introduction de la typologie des noms propres.

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).

3.1. *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. Par exemple, le nom propre *Molière* est plus connu que son synonyme *Jean-Baptiste Poquelin*.

Nous avons considéré la variation diatopique comme un alias (voir section 2.3, page 119). Il nous reste donc à présenter les trois variations restantes : diachronique, diastratique et diaphasique.

La première variation correspond à un point de vue diachronique, qui permet d'exprimer la notion de variété dans l'espace temporel, que l'on appelle aussi variation

historique. Il s'agit principalement d'entités ou d'objets existants et connus sous un certain nom pendant une période donnée et qui, à cause de diverses raisons (politique, économique, stratégique, etc.), adoptent un nouveau nom et, à partir de cet instant, leur ancien nom cesse d'être utilisé :

– *Zaïre* et *République démocratique du Congo* en français ;

– *Zaire* et *Democratic Republic of the Congo* en anglais ;

– *Zaire* et *República Democrática do Congo* en portugais.

Certaines transformations permettent aussi un passage entre ces synonymes, par exemple :

> « *SUR LES MASSACRES DANS L'**EX-ZAÏRE***
> [...]
> *DEPUIS l'arrivée au pouvoir de Mobutu Sese Seko en 1965, **l'ancienne République démocratique du Congo** (RDC), devenue Zaïre...* »
> *(le Monde diplomatique, 12/1997)*

La seconde variation est diastratique, c'est-à-dire liée à la classe socioculturelle. Pour des raisons diverses (pseudonyme d'auteur, nom religieux, sobriquet, etc.), certains référents correspondent à plusieurs noms propres conceptuels. Nous avons rassemblé ici les variantes familières et savantes : lorsque l'on parle de l'auteur du roman *La mare au diable*, il sera préférable de parler de *George Sand* plutôt que de *Aurore Dupin, baronne de Dudevant* au risque d'être incompris par une majorité de personnes.

Enfin, une variation diaphasique est liée à une différence de finalité d'emploi. Ainsi, par exemple, pour un effet stylistique, on utilisera :

– *Paris* et *Ville lumière* en français ;

– *Paris* et *City of Light* en anglais ;

– *Parigi* et *Città delle luci* en italien.

3.2. *La relation de méronymie*

La relation de méronymie constitue sans doute une des relations importantes que l'on retrouve dans le système WordNet et qui apparaît aussi dans de nombreux autres projets (EuroWordNet, Balkanet, SIMPLE, etc.). On l'appelle également relation partie-tout (*whole-part*), relation partitive ou encore relation d'inclusion. Lorsque deux unités lexicales A et B sont en relation de méronymie, on dit que A est un méronyme de B, et on dit que B est un holonyme de A si et seulement si A est une partie de B.

Cette relation permet d'établir une hiérarchisation sur plusieurs niveaux entre les éléments contenants (holonymes) et les éléments contenus (méronymes). Dans la plupart des cas, elle ne participe pas directement aux différentes étapes de la traduction, mais elle apporte une aide non négligeable dans le domaine de la recherche d'information.

(Winston *et al.*, 1987), dans le cadre de leurs travaux, ont proposé un découpage de la relation de méronymie en six catégories différentes, utilisées par le projet EuroWordNet, tandis que WordNet se base uniquement sur trois catégories. Dans le cadre de nos travaux sur les noms propres, seulement deux catégories s'appliquent, mais il faudrait en ajouter une septième, la relation de méronymie temporelle (Campenhoudt, 1996). Voir le tableau 2.

Type de méronymie	**Exemple**	**Wordnet**	**Eurowordnet**	**Prolex**
composant/objet	anse/tasse	oui	oui	non
membre/collection	arbre/forêt	oui	oui	oui
portion/masse	grain/sel	non	oui	non
matière/objet	bois/table	oui	oui	non
trait/activité	âge/adolescence	non	oui	non
lieu/lieu	oasis/désert	non	oui	oui
période/période	matinée/journée	non	non	oui

Tableau 2. *Taxonomie des relations de méronymie*

Le tableau 3 donne des exemples de relations de méronymie entre les différentes classes de noms propres.

3.3. *La relation d'accessibilité*

Les premiers travaux du projet Prolex portaient uniquement sur les toponymes. Il existait une relation qui permettait de préciser qu'une ville était la capitale d'une région ou d'un pays. Au début de nos études, nous avions étendu cette relation à la relation *Chef*, qui est une fonction lexicale appelée *Cap* du Dictionnaire Explicatif et Combinatoire du français contemporain (Mel'čuk, 1984, 1988, 1992, 1999). La relation *Chef* permettait de préciser si une entité est à la tête d'une autre entité (ou de plusieurs...). En plus de la relation régions-capitales, nous pouvions modéliser les relations entre les anthroponymes et les anthroponymes collectifs. Mais très vite, nous avons voulu introduire une relation entre les auteurs et leurs œuvres, les personnes et leur famille, etc. C'est pour cela que nous avons décidé d'utiliser plutôt la relation d'accessibilité présentée par Jonasson.

Si l'on cherche, par exemple, dans un dictionnaire édité, le nom propre *Tours*, il est impossible à partir de la définition que nous y trouvons de nous représenter la ville de Tours. En revanche, nous pouvons situer l'emplacement de cette ville sur une carte géographique. Cette description relie le nom propre *Tours* à d'autres noms propres : *Indre-et-Loire*, *Loire*, *Paris*. En recherchant dans le dictionnaire le nom propre *Indre-et-Loire*, nous retrouvons une référence au nom propre *Tours*. Cependant, ce lien n'est pas systématique pour tous les noms propres du dictionnaire. Prenons par exemple le cas du nom propre *Aaron* dans le *Petit Larousse* : celui-ci est présenté comme le *Frère aîné de Moïse*, alors qu'en recherchant l'article sur le nom propre *Moïse*, nous n'ob-

Type de méronymie	**Exemple**
Célébrité/Association	François Hollande/le Parti Socialiste
Célébrité/Ensemble	Syd Barrett/les Pink Floyd
Célébrité/Entreprise	Franck Riboud /Danone
Célébrité/Institution	Marguerite Yourcenar/Académie française
Célébrité/Organisation	Shafqat Kakakhel/PNUE
Célébrité/Dynastie	Charlemagne/Carolingien
Célébrité/Œuvre	Lancelot du lac/Cycle du roi Arthur
Célébrité/Pays	Victor Hugo/France
Célébrité/Histoire	Louis XIV/Ancien Régime
Entreprise/Entreprise	Air France /Air France-KLM
Entreprise/Pays	SNCF/France
Entreprise/Supra	EADS/Europe
...	...
Astronyme/Astronyme	Jupiter/le Système solaire
Géonyme/Pays	la Forêt-Noire/Allemagne
Hydronyme /Pays	Seine/France
Hydronyme /Supranational	Danube /Europe
Pays/Supranational	France/Europe
Pays/Organisation	France/ONU
Région/Pays	La Vendée/France
Région/Région	Indre-et-Loire/Région Centre
Ville/Région	Tours/Indre-et-Loire
Ville/Œuvre	Minas Tirith/Le retour du roi
Édifice/Ville	Panthéon/Rome
Édifice/Œuvre	tour de Babel/Bible
Voie/Ville	la place de l'Étoile/Paris
...	...
Objet/Œuvre	Excalibur/cycle du roi Arthur
Œuvre/Œuvre	Le retour du roi/Le Seigneur des Anneaux
Produit/Produit	Mégane/Renault
...	...
Histoire/Histoire	la Prise de la Bastille/la Révolution française
...	...

Tableau 3. *Relation de méronymie entre les noms propres*

tenons aucune référence à *Aaron*. En ironisant un peu, on pourrait affirmer que *Moïse* n'est pas le frère de *Aaron* et que la relation de fraternité entre ces deux personnes n'est pas une relation symétrique. Cela montre bien que l'accès au nom propre *Aaron* se fait par l'intermédiaire du nom propre *Moïse* et non l'inverse.

À partir de ce constat, nous pouvons affirmer qu'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, conçues à partir des expansions classifiantes (voir section 3.4) Mais, nous avons dans notre base de données un grand nombre d'expansions différentes pour un nom propre. Créer autant de relations que d'expansions (par exemple : fils, frère, élève, etc.) risque d'être coûteux et de nuire à la lisibilité du modèle. 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 ne pouvons pas créer dans le niveau interlingue des relations qui ne serviront que pour une seule langue. Nous avons donc décidé de les regrouper dans une seule et unique relation, à laquelle nous avons ajouté des repérages généraux (voir tableau 4). Les informations sur les expansions sont conservées dans la relation d'expansion classifiante. 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. *Marie* est la mère de *Jésus*, *Louis XIII le Juste* est le fils d'*Henri IV* ;

– créateur : les auteurs et les œuvres. *Richard Wagner* est le compositeur de *l'Anneau du Nibelung*, *Victor Hugo* est l'auteur de *Ruy Blas* ;

– capitale : les toponymes et leurs capitales. *La Rochelle* est le chef-lieu de la *Charente-Maritime*, *Bangkok* est la capitale de la *Thaïlande* ;

– dirigeant politique : les hommes politiques et les pays. *Jacques Chirac* est un homme politique *français* ;

– dirigeant non politique : les dirigeants et les entreprises. *Franck Riboud* est le PDG du *groupe Danone*.

– fondateur : les fondateurs d'une association, d'un groupe, d'une entreprise, d'un parti, d'une institution ; *José Maria Escrivá de Balaguer* est le fondateur de l'*Opus Dei*, *Richelieu* est le fondateur de l'*Académie française* ;

– élève : les disciples et leurs maîtres. *Aristote* est le disciple de *Platon* ;

– siège : les entreprises, associations ou organisations et le toponyme correspondant au siège social. *Peugeot* est une firme *sochalienne* ;

– etc.

3.4. *La relation d'expansion classifiante*

Cette relation, que l'on appelle aussi relation de classifieur (Jonasson, 1994) associe à chaque prolexème une expansion. Un nom propre apparaît régulièrement dans les textes journalistiques, quelle que soit la langue, accompagné d'expansions se trouvant

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. Par exemple, le français distingue l'expansion *rivière* et *fleuve* pour le nom d'un cours d'eau alors que l'anglais utilise seulement l'expansion *river*. La traduction des expansions peut parfois poser quelques problèmes. Par exemple, la traduction de *Rechtsanwalt Paul Bischof* (allemand) ne donne pas en français *Avocat Paul Bischof* mais plutôt *Maître Paul Bischof*. 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* deviendrait 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. Le tableau 4 en donne quelques exemples.

Repérage	**Expansions**
Capitale	Capitale, chef-lieu, préfecture, etc.
Créateur	Sculpteur, auteur, peintre, etc.
Dirigeant non politique	Patron, directeur, chef, etc.
Dirigeant politique	Président, roi, empereur
Élève	Disciple, élève, apôtre, etc.
...	

Tableau 4. *Repérages et expansions*

Nous avons prévu d'associer aux expansions classifiantes des liens vers des descriptions syntaxiques (grammaires locales (Gross, 1989)) ou sémantiques (les classes d'objets (Le Pesant et Mathieu-Colas, 1998), EuroWordNet, Framenet (Fillmore *et al.*, 2003)). En français, les toponymes, se construisent dans une phrase avec des prépositions locatives spécifiques, listées, par exemple, par (Constant, 2003). Nous avons aussi envisagé d'intégrer ces informations sous forme de grammaires locales.

3.5. *L'éponymie*

Les noms propres apparaissent parfois dans les textes sous la forme de simples subsantifs. Cette possibilité existe dans un grand nombre de langues, mais, pour un nom propre donné, dépend de la langue considérée. Nous avons appelé relation d'éponymie la relation entre un nom propre et une forme lexicalisée, soit dans le vocabulaire courant, soit dans une expression idiomatique ou terminologique. 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. Par exemple, il ne faudra pas reconnaître *Parkinson* comme nom propre dans *maladie de Parkinson*.

L'antonomase est une figure de rhétorique par laquelle un nom propre est remplacé par un nom commun ou inversement. Nous avons pris en compte uniquement, dans le

cadre de la relation d'antonomase, les antonomases à partir d'un nom propre. Un nom propre employé par antonomase perd la plupart du temps, dans le cas du français, sa majuscule initiale, surtout quand le lien qui l'unit au nom propre originel tend à s'effacer :

un *bic* = un *stylo-bille* ;
un *kleenex* = un *mouchoir en papier*.

Cette figure de style existe dans la plupart des langues :

pampersy pour *couches jetables* en polonais ;
kalodont pour *crème dentifrice* en serbe ;
biro pour *stylo-bille* en anglais ;
ксерокс (xerox) pour photocopieuse en russe.

Certaines antonomases peuvent exister dans une langue et être totalement absentes dans d'autres langues. Le nom propre *Pampers* a donné lieu à une antonomase en polonais, alors que ce n'est pas le cas en français.

Les tournures idiomatiques sont parfois construites à l'aide d'un ou plusieurs noms propres. Certaines tournures idiomatiques comprenant un nom propre dans une langue donnée peuvent se traduire vers une autre langue à l'aide d'une autre tournure idiomatique pouvant ne pas comporter de nom propre ou un nom propre différent. C'est le cas des exemples suivants :

être en tenue d'Adam = to be in one's birthday suit ;
not for all the tea in China = pour rien au monde ;
I don't know him from Adam = je ne le connais ni d'Ève ni d'Adam ;
(Dictionnaire Hachette-Oxford).

Le sens d'une expression figée peut varier d'une langue vers une autre :

zwischen Scylla und Charybdis sein = être entre deux dangers (Duden 11 / Redewendungen) ;
tomber de Charybde en Scylla = quitter un mal pour un autre pire encore (Petit Larousse).

Enfin, on retrouve de nombreux noms propres dans les terminologies scientifiques (le *théorème d'al-Kashi* sur le calcul des longueurs des côtés d'un triangle non rectangle, les *équations de Maxwell* qui caractérisent les interactions entre charges, etc.), juridiques (la *loi Evin* relative à la lutte contre le tabagisme et l'alcoolisme, la *loi de Robien* sur l'investissement locatif, etc.) ou médicales (la *maladie de Creutzfeldt-Jakob*, la *maladie de Parkinson*, etc.).

Nous n'avons pas intégré ces termes dans la classe des noms propres, car ils appartiennent plus à une langue spécialisée qu'à la langue générale. De plus, leur traduction, loin d'être triviale, nécessite parfois l'utilisation d'une expression définie. Par exemple, la *loi Pasqua* ne se traduira pas en allemand par *Pasqua-Gesetz* mais plutôt par *französisches Einwanderungs und Staatsangehörigkeitsgesetz*.

D'une langue à l'autre, les noms propres utilisés dans une terminologie peuvent être sujets à des variations, permutation[10] ou même insertion[11].

4. La typologie

Dans cette section, nous allons nous intéresser au domaine de la typologie des noms propres. Nous nous sommes basés sur les différentes typologies utilisées dans le domaine de la linguistique et sur celles qui ont conduit à des systèmes de reconnaissance de noms propres. Nous avons ensuite établi une liste de types de noms propres (Grass *et al.*, 2002) et nous avons appliqué la méthode de (Noy et McGuinness, 2003) pour définir et hiérarchiser nos différents concepts par une relation d'hyperonymie.

Cette typologie a pour racine le concept de nom propre, pour nœuds des supertypes et pour feuilles des types.

4.1. *Les quatre premiers supertypes*

Situés juste en dessous du concept de nom propre, les quatre premiers supertypes classent les noms propres suivant des traits syntaxico-sémantiques assez généraux. Ces traits peuvent souvent être reconnus par des systèmes d'extraction automatique de noms propres en se basant essentiellement sur le contexte linguistique apparaissant autour d'eux dans le texte (Friburger, 2002).

Dans notre ontologie, nous avons distingué :

– les anthroponymes : trait humain ;

– les ergonymes : trait inanimé ;

– les pragmonymes : trait événement ;

– les toponymes : trait locatif.

Le supertype anthroponyme, comme le supertype toponyme, est un concept largement connu et communément admis dans le domaine de l'onomastique ou de l'étude des noms propres. Le trait humain est sans doute le trait le plus facile à percevoir et à reconnaître chez un nom propre. Les anthroponymes renvoient sur le plan sémantique à la notion de personne. Nous avons partagé le supertype anthroponyme en deux autres supertypes : les anthroponymes individuels (*Lassie*, *George Orwell*, etc.) et les anthroponymes collectifs (*Mérovingiens*, *Organisation mondiale de la santé*, etc.).

Nous avons rassemblé sous le concept de toponyme tous les noms de lieux au sens général. Les toponymes regroupent diverses entités qui possèdent chacune une taille

10. *Maladie de Legg-Perthes-Calvé* vs *Perthes-Legg-Calvé-Krankheit* (Bodenreider et Zweigenbaum, 2000a)
11. *Maladie de Weber-Christian* vs *Pfeifer-Weber-Christian-Krankheit* (Bodenreider et Zweigenbaum, 2000b))

extrêmement variée. Cela peut aller du nom donné à une rue ou à un bâtiment, en passant par le nom d'une vaste zone géographique pouvant regrouper plusieurs pays, jusqu'à s'étendre au nom d'un ensemble contenant des millions de galaxies.

Ergonyme (du grec *ergon* : travail, force) est un mot, emprunté à (Bauer, 1985), qui désigne les fabrications humaines. Sous le type ergonyme, on peut retrouver des noms propres qui possèdent les traits sémantiques inanimé concret (*Coca-Cola*) ou inanimé abstrait (*Alice au pays des merveilles*).

Les pragmonymes peuvent être définis comme des noms d'événements dont l'homme peut être l'auteur, la victime ou les deux à la fois.

4.2. *Les types*

Le type correspond à une classification plus détaillée que le supertype d'un nom propre. Cette classification est destinée principalement à la recherche d'information et à la traduction automatique. Pour associer un type à un nom propre, il faut souvent une intervention humaine. Dans le cadre de nos travaux, nous avons retenu au total trente types que nous allons présenter dans cette partie. Le tableau 5 liste des exemples de noms propres classés en fonction de ces types.

Cependant, certaines distinctions sont difficiles à réaliser et peuvent sembler arbitraires. Nous avons donc décidé de créer deux autres supertypes :

– un supertype hyponyme des anthroponymes collectifs, supertype que nous appellerons « Groupement » et qui rassemble les types correspondant à une entreprise, une association ou une institution (politique, religieuse, culturelle, nationale, internationale, etc.).

– un supertype hyponyme des toponymes, supertype que nous appellerons « Territoire », car il est parfois difficile de faire une distinction entre les pays (au sens *États indépendants*) et les régions incluses ou non dans les pays.

5. L'ontologie

5.1. *La relation d'hyperonymie*

Une relation d'hyperonymie relie les noms propres conceptuels (pivots) de notre ontologie avec les types (ou supertypes), eux-mêmes en relation d'hyperonymie entre eux. Cependant, nous avons souhaité faire une distinction entre une relation plus « naturelle » (hyperonymie primaire) où chaque type est en relation avec un et un seul supertype hyperonyme et des relations « complémentaires » (hyperonymie secondaire). Par exemple, une ville est d'abord vue comme un lieu, avant d'être considérée comme un humain collectif ou une fabrication humaine... Le type « Ville » est donc à la fois en relation d'hyperonymie primaire avec le supertype « Toponyme » et en relation

Types	**Exemples**
Association	*les Restaurants du cœur*, *l'Union chrétienne-démocrate* , etc.
Astronyme	*l'étoile Polaire*, *le Bélier*, *Pluton*, etc.
Catastrophe	*Erika*, *Tchernobyl*, *Katrina*, etc.
Célébrité	*Platon*, *Blanche-Neige*, *Antoine de Saint-Exupéry*, etc.
Dynastie	*Carolingien*, *Michelin*, *Ming*, etc.
Édifice	*le Colisée*, *le Palais Bourbon*, *la Grande Muraille*, etc.
Ensemble	*Les Beatles*, *le cercle de Prague*, *De Stijl*, etc.
Entreprise	*Air France*, *Nestlé*, *DaimlerChrysler*, etc.
Ethnonyme	*Étrusque*, *Aztèque*, *Sabin*, etc.
Fête	*Noël*, *Halloween*, *la Pentecôte*, etc.
Géonyme	*les Alpes*, *le désert de Syrie*, *le Kilimandjaro*, etc.
Histoire	*le Moyen-Âge*, *la bataille d'Austerlitz*, *le traité de Rome*, etc.
Hydronyme	*l'Amazone*, *le lac Léman*, *l'océan Pacifique*, etc.
Institution	*le Collège de France*, *Scotland Yard*, *l'institut Pasteur*, etc.
Manifestation	*le Tour de France*, *le Festival d'Avignon*, *la coupe Davis*, etc.
Météorologie	*l'anticyclone des Açores*, *El Niño*, *la tramontane*, etc.
Objet	*le Saint-Graal*, *Durandal*, *la Toison d'or*, etc.
Œuvre	*l'Avare*, *les Demoiselles d'Avignon*, *la Vénus de Milo*, etc.
Organisation	*la Croix-Rouge*, *l'Organisation mondiale de la santé*, etc.
Patronyme	*Dupont*, *Durant*, etc.
Pays	*le Portugal*, *l'Australie*, *la République de Corée*, etc.
Pensée	*Christianisme*, *Islam*, *Communisme*, etc.
Prénom	*Louis*, *Jean*, *Pierre*, etc.
Produit	*Adidas*, *Ferrari 250 GTO*, *Coca-Cola*, etc.
Pseudo-anthroponyme	*Pégase*, *C-3PO*, *Donald*, etc.
Région	*l'Austrasie*, *le Tartare*, *la Californie*, etc.
Supranational	*les Antilles*, *l'Eurasie*, *les pays Baltes*, etc.
Vaisseau	*le Titanic*, *Apollo 11*, *Enterprise*
Ville	*Marseille*, *Nha Trang*, *Chiang Rai*, etc.
Voie	*la place Rouge*, *les Champs-Élysées*, *l'autoroute du Soleil*, etc.

Tableau 5. *Les types*

d'hyperonymie secondaire avec les supertypes « Anthroponyme collectif » et « Ergonyme ».

Nous avons donc deux relations d'hyperonymie, qui sont définies précisément sur les tableaux 6 (hyperonymie primaire) et 7 (hyperonymie secondaire).

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

Tableau 6. *Hyperonymie primaire*

Types	**Hyperonymes secondaires**
Pays Région Supranational Territoire	Anthroponyme collectif
Ville	Anthroponyme collectif Ergonyme
Édifice Voie	Ergonyme
Fête Histoire Manifestation	
Association Ensemble Entreprise Groupement Institution Organisation	Ergonyme Toponyme
Vaisseau	Toponyme

Tableau 7. *Hyperonymie secondaire.*

5.2. *L'existence*

Enfin, nous définissons une troisième relation d'hyperonymie, que nous avons appelé *existence*, et qui relie les noms propres conceptuels à l'une des trois instances :

– être un nom propre du domaine historique (*Mozart*, *le Danube*, *Paris*...) ;

– être un nom propre du domaine de la croyance (*Zeus*, *Adam...*) ;

– être un nom propre du domaine de la fiction (*Tintin*, *Utopie*, *Atlantis...*).

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, *Blanche-Neige* devient :

– *Sneeuwwitje* en néerlandais ;

– *Biancaneve* en italien ;

– *Schneewitchen* en allemand.

5.3. *Ontolgie et représentation globale*

Chaque nom propre conceptuel (ou pivot) est en relation d'hyperonymie avec un type et une existence. Finalement, notre ontologie des noms propres est constituée des relations qui ne dépendent pas de la langue (section 3, p. 123), de la typologie (section 4, p. 130) et de quatre relations d'hyperonymie (figure 1).

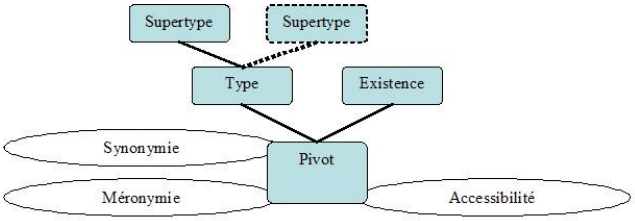

Figure 1. *Ontologie des noms propres*

Nous pouvons aussi représenter les différents concepts du domaine des noms propres sous la forme d'une arborescence (figure 2) qui peut se décomposer en deux niveaux : un niveau qui ne dépend pas des langues (notre ontologie) et un niveau qui dépend de la langue.

Le niveau qui ne dépend pas de la langue est lui-même composé de deux niveaux. Le premier niveau, que l'on appelle le niveau méta-conceptuel, comprend les types et l'existence. Le deuxième niveau est le niveau conceptuel, qui inclut le concept de nom propre conceptuel et les relations de méronymie, de synonymie et d'accessibilité.

Le niveau qui dépend de la langue est aussi subdivisé en deux niveaux différents : le niveau linguistique et le niveau des instances. Le niveau linguistique englobe les concepts de prolexème, d'alias et de dérivé. Chaque langue possédera sa propre arborescence à partir de la forme canonique d'un nom propre, ou prolexème, qui sera relié à un même niveau, qui ne dépend pas de la langue, à travers un ensemble de noms propres conceptuels. En raison des grandes divergences et de la complexité des mécanismes s'appliquant aux noms propres, nous ne pouvons définir une arborescence

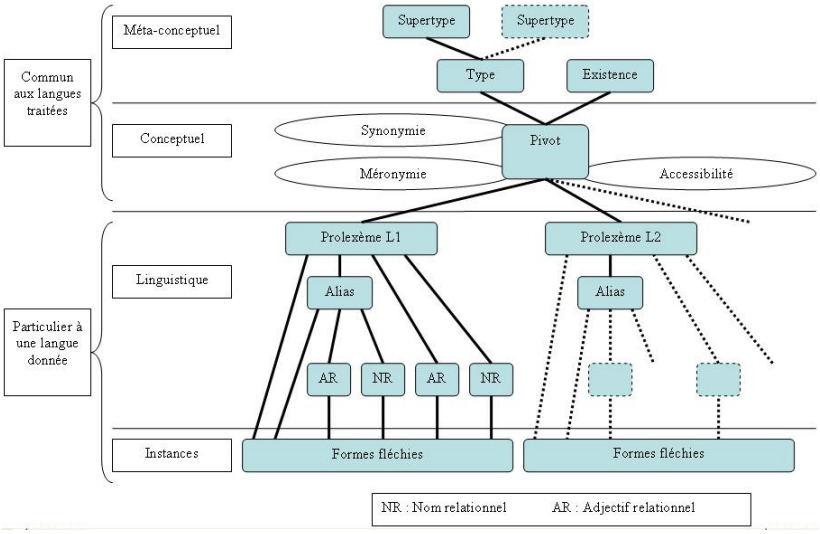

Figure 2. *Représentation arborescente des noms propres*

générale qui pourra s'appliquer pour toutes les langues. Selon la langue, cette arborescence pourra être plus ou moins complexe. Le niveau des instances regroupe toutes les formes fléchies que l'on peut obtenir en appliquant des règles morphologiques, plus ou moins compliquées selon les langues, sur un nom propre.

6. Implémentation

La dernière partie de nos travaux a été consacrée à l'implémentation de notre modèle : création d'une base de données relationnelles et d'une interface de travail collaboratif[12] (Tran *et al.*, 2005) ; création d'une interface de consultation[13], avec des réponses conviviales ou sous un format XML.

Nous avons testé et validé la pertinence de notre modèle en travaillant sur les données toponymiques déjà collectées dans le cadre du projet Prolex et sur les noms propres extraits de dictionnaires de niveau collège, censés représenter un socle de culture commune à l'ensemble des français. À la fin du projet *Technolangue*, Prolexbase contient 54 164 prolexèmes français, associés à 493 alias et à 20 609 dérivés. Le tableau 8 présente le nombre de prolexèmes pour la partie française en fonction de leur type et le tableau 9 le nombre de relations qui ne dépendent pas de la langue.

12. http://tln.li.univ-tours.fr/Tln_prolexbase
13. http://tln.li.univ-tours.fr/tln_prolex/prolex.php

Anthroponyme	4 048
Association	32
Célébrité	3 735
Dynastie	50
Ensemble	14
Entreprise	3
Ethnonyme	134
Institution	57
Organisation	20
Patronyme	0
Prénom	0
Pseudo Anthroponyme	3
Toponyme	49 566
Astronyme	24
Édifice	93
Géonyme	205
Hydronyme	4 348
Pays	398
Région	2 627
Supranational	53
Ville	41 804
Voie	14
Ergonyme	166
Objet	0
Œuvre	76
Pensée	3
Produit	86
Vaisseau	4
Pragmonyme	216
Catastrophe	1
Fête	11
Histoire	200
Manifestation	3
Météorologie	1

Tableau 8. *Nombre de prolexèmes français de Prolexbase.*

Nous avons eu l'occasion de tester notre modèle pour d'autres langues que le français : le serbe (606 prolexèmes) et le coréen (113 prolexèmes). Nous prévoyons de mettre en place des collaborations avec des laboratoires européens pour ajouter d'autres langues.

Relations	47 145
Synonymie	641
Méronymie	44 260
Accessibilité	2 244

Tableau 9. *Nombre de relations qui ne dépendent pas de la langue dans Prolexbase.*

7. Conclusion

La création du dictionnaire relationnel multilingue de noms propres Prolexbase peut se résumer en trois parties, détaillées dans (Tran, 2006) :

– tenter de répondre à diverses questions : qu'est-ce qu'un nom propre ? quelles informations inclure dans un dictionnaire de noms propres ? comment construire un dictionnaire multilingue ?

– modéliser le domaine des noms propres ;

– implémenter le modèle.

Nous avons uniquement présenté dans cet article la modélisation du domaine des noms propres. Nous avons défini deux concepts centraux : le nom propre conceptuel et le prolexème. Le nom propre conceptuel 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, représenté par une forme vedette. Autour d'eux, nous avons défini d'autres concepts (alias, dérivé, etc.) et relations (synonymie, méronymie, accessibilité, éponymie, etc.). Chaque nom propre conceptuel est en relation d'hyperonymie avec un type et une existence au sein d'une ontologie.

Pour le français, nous comptons maintenant travailler la morphologie, en utilisant le rapport du projet *Transweb 2* (Martineau, 2005) et le système *Multiflex* défini par Agata Savary (Savary, 2006). Il nous faut aussi impérativement compléter notre base avec des données « modernes », car les noms de personnalité et d'entreprise sont très présents dans les journaux, mais bien peu dans le dictionnaire[14].

Parallèlement à ce travail, nous envisageons de développer des outils pour le traitement automatique des noms propres dans des textes (pour les applications d'aide à la rédaction et à la traduction, la traduction automatique, la recherche d'information multilingue, l'alignement de textes multilingues, l'indexation des noms propres...).

14. Jacques Chirac et Lionel Jospin sont présents dans les dictionnaires que nous avons utilisés, mais pas Dominique de Villepin, ni Nicolas Sarkozy, pas plus que Ségolène Royal ; quant aux entreprises, elles sont quasiment absentes : nous en avons trouvé trois !

8. Bibliographie

Bauer G., *Namenkunde des Deutschen*, Germanistische Lehrbuchsammlung Band 21, Berlin, 1985.

Bodenreider O., Zweigenbaum P., « Identifying proper names in parallel medical terminologies », *Medical Infobahn for Europe (MIE2000)*, p. 443-447, 2000a.

Bodenreider O., Zweigenbaum P., « Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles », *Traitement automatique des langues*, vol. 41-3, p. 727-757, 2000b.

Campenhoudt M. V., « Recherche d'équivalences et structuration des réseaux notionnels : le cas des relations méronymiques », *Terminology*, vol. 3:2, p. 53-83, 1996.

Constant M., Grammaires locales pour l'analyse automatique de textes : Méthodes de construction et outils de gestion, Thèse de doctorat en informatique, Université de Marne-la-Vallée, 2003.

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

Courtois B., « Dictionnaire électronique des mots simples du français DELAS V07-E1 », *Rapport de recherche n°33 du LADL, Université Paris VII*, 1992.

Eggert E., La dérivation toponymes-gentilés en français : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique, Thèse de doctorat en linguistique, cotutelle des universités de Tours et Münster, 2002.

Fillmore C., Johnson C., Petruck M., « Background to Framenet », *International Journal of Lexicography*, vol. 16, p. 235-250, 2003.

Friburger N., Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques, Thèse de doctorat en informatique, Université François-Rabelais de Tours, décembre, 2002.

Grass T., Maurel D., Piton O., « Description of a multilingual database of proper names », *PorTal, in LNCS 2389*, Faro, Portugal, p. 137-140, juillet, 2002.

Gross M., « The Use of Finite Automata in the Lexical Representation of Natural Language », *Electronic Dictionaries and Automata in Computational Linguistics, LNCS*, vol. 377, p. 34-50, 1989.

Gross M., « Synonymie, morphologie dérivationnelle et transformations », *Langage*, vol. 128, p. 72-90, 1997.

Jonasson K., *Le nom propre. Constructions et interprétations*, Duculot, Paris, 1994.

Le Pesant D., Mathieu-Colas M., « Introduction aux classes d'objets », *Langages*, vol. 131, p. 6-33, 1998.

MacDonald D., « Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names », *Corpus Processing for Lexical Acquisition*, Massachussetts Institute of Technology, p. 21-39, 1996.

Mangeot-Lerebours M., Sérasset G., Lafourcade M., « Construction collaborative d'une base lexicale multilingue - Le projet Papillon », *Traitement Automatiques des Langues (TAL), édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ? (Electronic dictionaries : for humans, machines or both ?) ed. Michael Zock and John Carroll*, vol. 44(2), p. 151-176, 2003.

Martineau C., Outils multilingues de génération de formes fléchies et dérivées, Technical report, Rapport technique, Projet Transweb 2, IGM, Université de Marne-la-Vallée, 2005.

Maurel D., Piton O., Eggert E., « Les relations entre noms propres : lieux et habitants dans le projet Prolex », *t.a.l.*, vol. 41, n° 1, p. 623-641, 2000.

Maurel D., Tran M., Friburger N., « Projet Technolangue NomsPropres : Constitution et exploitation d'un dictionnaire relationnel multilingue de noms propres », *Atelier Les Ressources dans le traitement de la langue écrite, conférence associée à TALN 2006*, Louvain, Belgique, p. 927-936, avril, 2006.

Mel'čuk I., « Dictionnaire explicatif et combinatoire du français contemporain », *Recherches lexico-sémantiques I, II, III, IV, Montréal, Presses de l'Université de Montréal*, 1984, 1988, 1992, 1999.

Mikheev A., Moens M., Grover C., « Named Entity Recognition without Gazetteers », *EACL'99*p. 1-8, 1999.

Miller G., « Wordnet : A lexical database for English », *Communication of the ACM*, vol. 38(11), p. 39-41, 1995.

Noy N. F., McGuinness D. L., « Ontologie pour le Web sémantique », *Web sémantique, rapport final de l'Action spécifique 32 CNRS/STIC*, 2003.

Piton O., Maurel D., « Les noms propres géographiques et le dictionnaire Prolintex », *Cahiers de la MSH Ledoux, Série Archive, Bases, Corpus, n° 1*, p. 53-76, 2004.

Polguère A., *Lexicologie et sémantique lexicale. Notions fondamentales*, Presses de l'Université de Montréal, Montréal, 2003.

Ren X., Perrault F., « The Typology of Unknown Words : an Experimental Study of Two Corpora », *COLING 92, Nantes*, p. 408-414, 1992.

Romary L., « De la sémantique des contenus à la sémantique des structures », *La recherche d'information sur les réseaux, Sciences de l'information, série Études et techniques, ADBS Éditions*, p. 203-230, octobre, 2002.

Romary L., Salmon-Alt S., Francopoulo G., « Standards going concrete : from LMF to Morphalou », *Workshop on Electronic Dictionaries, COLING 2004*, Geneva, p. 22-28, 2004.

Savary A., MULTIFLEX. User's Manuel and Technical Documentation Version 1.0, Technical report, Université François-Rabelais de Tours, IUT de Blois, France, 2006.

Tran M., Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception, implantation et gestion en ligne, Thèse de doctorat en informatique, Université François Rabelais Tours, octobre, 2006.

Tran M., Maurel D., Vitas D., Krstev S., « A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names », *Papillon 2005 workshop on Multilingual Lexical Databases, in Association with the Sixth Symposium on Natural Language Processing (SNLP 2005)*, Chiang Rai, Thailande, p. 2 :67-71, décembre, 2005.

Winston M. E., Chaffin R., Herrmann D., « A Taxonomy of Part-Whole Relations », *Cognitive Science*, vol. 11, p. 417-444, 1987.