Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception, implémentation et mise en ligne

Mickaël Tran

Université François-Rabelais de Tours

École Doctorale : Santé, Sciences et Technologies
Année Universitaire : 2005-2006

THÈSE PRÉSENTÉE POUR OBTENIR LE GRADE DE
DOCTEUR DE L'UNIVERSITÉ DE TOURS

Discipline : Informatique

présentée et soutenue publiquement

par

Mickaël Tran

le 20 octobre 2006

Prolexbase
Un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne

---

Sous la direction de Denis Maurel
et sous la codirection d'Agata Savary

Membres du jury

Christian Boitet	*Professeur*	Université de Grenoble 1	*(Rapporteur)*
Denis Lepesant	*Professeur*	Université de Lille 3	*(Examinateur)*
Denis Maurel	*Professeur*	Université de Tours	*(Directeur de thèse)*
Jean-Marie Pierrel	*Professeur*	Université de Nancy 1	*(Rapporteur)*
Agata Savary	*Maître de Conférences*	Université de Tours	*(Codirecteur de thèse)*
Pierre Zweigenbaum	*Directeur de recherche*	CNRS	*(Examinateur)*

Đa tình tự cổ nan di hận
Chỉ hận miên miên vô tuyệt kỳ

Tây Du Ký

Mauruuru Mulțumesc Gracias

吾该 Mersi

благодарам

ขอบคุณครับ

Dankie

thugs rje gnang

Дзякую Zahvaljujem תודה

Danke

Hvala Dziękuję

Faleminderit Dankon

Tak Paldies

I ni cɛ Milesker Merci

Blagodaria

Mercès 谢谢

Wliwin

Thanks

благодаря

شكرا Kiitos

Tänan

ありがとう

Ačiu Grazas Mèrci

Ευχαριστώ

Diolch

ขอบคุณคะ lời cảm ơn

спасибо

S'il est possible de comparer une thèse à une autre activité, je pense sans hésiter au domaine du bâtiment. Construire sa maison est sans doute une activité pénible, de longue haleine et qui nécessite d'avoir énormément de patience, de courage, du sens de la rigueur et de dépenser sans compter dans de longues et parfois interminables heures de travail sous le froid de l'hiver et sous la chaleur de l'été. Cette belle maison qui s'élève devant vous aujourd'hui est le résultat de longues années de sueurs. Mais c'est sans doute grâce à une telle enrichissante expérience que l'on pourra mieux affronter nos futurs autres grands chantiers.

C'est probablement grâce au plan de l'architecte (*Denis Maurel*), qui a toujours su veiller à la qualité, à corriger les défauts et à fournir de précieux conseils dans la réalisation de l'œuvre, que l'on doit cette magnifique demeure.

On ne peut se lancer dans un tel chantier sans l'aide de l'ingénieure (*Agata Savary*), toujours disponible, qui a su guider cet ouvrier dans la bonne direction et lui apporter plus de méthode et plus de rigueur dans son travail.

Je remercie aussi ces autres grands architectes (*Christian Boitet*, *Denis Lepesant*, *Jean-Marie Pierrel* et *Pierre Zweigenbaum*), pour avoir accordé à cet ouvrier un peu de leur précieux temps pour veiller à ce que son travail respecte les normes. Je tiens à remercier mes rapporteurs pour leurs remarques et commentaires, et tout particulièrement *Christian Boitet* pour ses longues heures passées à corriger les nombreux défauts de cette maison et à qui ce manuscrit doit sa forme actuelle.

Ce travail n'a pu se réaliser que grâce à l'équipe du bureau d'étude (*Claire Agafanov*, *Béatrice Bouchou*, *Nathalie Friburger*, *Thierry Grass*, *Cvetana Krstev*, *Nathalie Rossi* et *Dusko Vitas*) qui ont toujours été là pour m'aider et me conseiller dans les moments de difficultés.

A ces nombreux autres ouvriers que l'on croise sur les chantiers voisins, partageant tous un unique rêve : *Ali*, *Cédric*, *Hassina*, *Karima*, *Lamia*, *Moustafa*... Et à ces compagnons de route que l'on croise le soir à la table de la taverne du village : *Abdalah*, *Anthony*, *Arnaud*, *Cyril*, *Dimitri*, *Johnathan*, *Mathieu*, *Nicolas*, *Samir* et ce cher filleul (bien sûr que je me rappelle plus de ton nom *Vincent*)...

Un grand remerciement à 218 (*Elise*) pour avoir passé du temps à corriger mes fautes d'orthographe et j'espère que ça lui donnera l'envie d'entreprendre la construction de sa propre maison.

Je tiens particulièrement à remercier mes actionnaires (mes parents, *Marianne* et *Raphaël*) qui ont su soutenir, encourager et sponsoriser ce projet tout au long de ces trois années, à qui sans doute cette maison doit son existence.

Il se peut que j'en ai oublié d'autres, je leur prie de bien vouloir m'excuser, car il n'est pas évident d'écrire ses remerciements dans les profondeurs de la nuit sous les lumières de la lune.

Et comme enfin, vient le moment ou il faut conclure par quelques rimes :

Ô toi seule ma Muse, par ton regard, a su m'écouter et m'inspirer
Dans ces durs et interminables moments de labeur très acharné
Par tes si douces et éphémères apparitions durant mes quelques nuits
Si près dans mes pensées et pourtant loin de mon cœur tu t'enfuis

Table des matières

Introduction **13**

I État de l'art **15**

1 Qu'est-ce qu'un nom propre ? **17**

Introduction **17**
1.1 Définitions . . . 17
1.1.1 Définitions des dictionnaires . . . 17
1.1.2 Définitions des linguistes . . . 18
1.2 Statut linguistique du nom propre . . . 19
1.2.1 Critère de la majuscule . . . 19
1.2.2 Critères morphologiques . . . 19
1.2.3 Critères syntaxiques . . . 20
1.2.4 Critères sémantiques . . . 21
1.2.5 Critères pragmatiques . . . 22
1.2.6 Discussion . . . 22
1.3 Typologies des noms propres . . . 22
1.3.1 Typologies linguistiques . . . 22
1.3.2 Typologies issues du TAL . . . 23

Conclusion **27**

2 Les ressources dictionnairiques **31**

Introduction **31**
2.1 Travaux du LADL . . . 32
2.2 EuroWordNet . . . 35
2.2.1 WordNet . . . 35
2.2.2 EuroWordNet . . . 37
2.2.3 Balkanet : une extension d'EuroWordnet . . . 39
2.2.4 Discussion . . . 40
2.3 Le Trésor de la Langue Française informatisé . . . 42
2.4 Dictionnaire Explicatif et Combinatoire . . . 43
2.5 Papillon . . . 46

Conclusion **48**

II Modélisation des noms propres **51**

3 Autour du nom propre conceptuel **53**

Introduction **53**
3.1 Les deux principaux concepts . . . 54
3.1.1 Le nom propre conceptuel . . . 54
3.1.2 Le prolexème . . . 55
3.2 Au niveau du prolexème . . . 56
3.2.1 Les alias . . . 57
3.2.2 Les dérivés . . . 61
3.3 Les relations . . . 63
3.3.1 La relation de synonymie . . . 63
3.3.2 La relation de méronymie . . . 65
3.3.3 La relation d'accessibilité . . . 66
3.3.4 La relation d'expansion classifiante . . . 69
3.3.5 L'éponymie . . . 69
3.4 Représentation sous forme d'un schéma . . . 72

Conclusion **73**

4 Ontologie des noms propres **77**

Introduction **77**
4.1 Ontologie . . . 77
4.1.1 Définition d'une ontologie . . . 77
4.1.2 Méthodologie de construction d'ontologie . . . 78
4.2 Typologie des noms propres . . . 79
4.2.1 Les quatre premiers supertypes . . . 80
4.2.2 Type . . . 81
4.3 Ontologie des noms propres . . . 86
4.3.1 Existence . . . 87
4.3.2 La relation d'hyperonymie . . . 87

Conclusion **88**

III Implémentation de Prolexbase **89**

5 La base de données **91**

Introduction **91**
5.1 La méthode Merise . . . 91
5.1.1 Le modèle conceptuel de données . . . 91
5.1.2 Le modèle relationnel de données . . . 92
5.2 Modèle conceptuel de données . . . 93
5.2.1 Le niveau conceptuel . . . 94
5.2.2 Le niveau méta-conceptuel . . . 96
5.2.3 L'éponymie . . . 96
5.2.4 Les règles . . . 96
5.2.5 Les autres informations . . . 97

5.2.6 Les flexions . . . 99
5.2.7 Les instances . . . 99
5.3 Modèle logique de données . . . 101

Conclusion **101**

6 Exportation des données **105**

Introduction **105**
6.1 État de l'art . . . 105
6.1.1 TEI . . . 105
6.1.2 TMF . . . 107
6.1.3 Discussion . . . 107
6.2 Le modèle XML de Prolexbase . . . 108
6.2.1 Fichier de requête XML . . . 108
6.2.2 Fichier d'exportation XML . . . 109
6.3 Implémentation . . . 111
6.4 Une contribution effective . . . 111

Conclusion **115**

7 Interface web **117**

Introduction **117**
7.1 Cahier des charges . . . 117
7.2 Site de consultation de Prolexbase . . . 118
7.2.1 Le menu *Recherche* . . . 118
7.2.2 Le menu *Texte* . . . 119
7.2.3 Autres menus . . . 119
7.3 Interface de travail . . . 119
7.3.1 Les menus simples . . . 121
7.3.2 Le menu fichier . . . 127
7.3.3 Les menus d'administration . . . 130
7.3.4 Suppression et fusion . . . 131
7.4 Calcul de complexité . . . 131

IV Synthèse **135**

8 Évaluation **137**
8.1 Le modèle . . . 137
8.1.1 Prolexème et forme vedette . . . 137
8.1.2 Date . . . 137
8.1.3 Synonymie et forme canonique . . . 139
8.1.4 Les relations d'hyperonymie et de méronymie . . . 139
8.1.5 La relation d'accessiblité . . . 140
8.2 L'interface de travail de Prolexbase . . . 141
8.3 Analyse quantitative . . . 141
8.4 Le contenu de Prolexbase . . . 144

Conclusion **147**

Liste de publications 149

A MCD et MLD 151

B Codes flexionnels du DELA 155

C Exemple XML : le prolexème *États-Unis d'Amérique* 157

Bibliographie 163

Table des figures

1.1 Typologie de Paik, Liddy, Yu et McKenna 24
1.2 Typologie de Sekine 26
1.3 Format du corpus de CoNLL 27
1.4 Typologie utilisée par la campagne ESTER. 28

2.1 Codes grammaticaux du DELAS. 33
2.2 Traits du DELAS. 33
2.3 Code flexionnel NC_XXN. 34
2.4 Recherche du nom propre *Paris* dans WordNet. 35
2.5 Nombre de mots et de concepts dans WordNet 2.1. 36
2.6 Exemple de relation d'hyperonymie dans WordNet. 37
2.7 Relations sémantiques dans WordNet. 37
2.8 Architecture d'EuroWordNet 39
2.9 EuroWordNet Top-Ontology 40
2.10 Relations internes d'une langue entre les synsets dans EuroWordNet. 41
2.11 Extrait de l'article *RÉMITTENT* du TLFi. 43
2.12 Dico 45
2.13 Macrostructure de Papillon 46
2.14 Schéma XML des lexies. 47
2.15 Différentes architectures 48

3.1 Le diasystème de Coseriu. 54
3.2 Le diasystème de Blanco. 55
3.3 Les variantes graphiques. 59
3.4 Statistique des suffixes de gentilés. 62
3.5 Taxonomie des relations de méronymie. 66
3.6 Relation de méronymie entre les noms propres. 67
3.7 Extrait de la Frame *Text_creation*. 70
3.8 Exemple d'une grammaire locale. 70
3.9 Hiérarchie de *writer* dans EuroWordNet. 71
3.10 Les quatre niveaux. 72
3.11 Le prolexème français. 73
3.12 Le prolexème serbe. 73
3.13 La partie cyrillique du prolexème serbe *Belgrade*. 74
3.14 Les prolexèmes français *Suisse* et *Confédération helvétique*. 75

4.1 Les supertypes. 80
4.2 Les types 82
4.3 La hiérarchie des types. 83
4.4 Hyperonymie secondaire. 87

4.5 Ontologie des noms propres. 88

5.1 Représentation d'une entité. 91
5.2 Représentation des attributs. 92
5.3 Représentation d'une association. 92
5.4 Schéma relationnel. 93
5.5 Le modèle conceptuel de données. 93
5.6 Le niveau conceptuel. 94
5.7 Les repérages. 95
5.8 Le nom propre *Paris* dans EuroWordNet. 95
5.9 Le niveau méta-conceptuel. 96
5.10 L'éponymie. 96
5.11 Les règles. 97
5.12 Les informations. 97
5.13 Les indicateurs actuellement utilisés pour le BLARK. 98
5.14 La flexion. 99
5.15 Le graphe de flexion d'*Antillais-et-Barbudien*. 100
5.16 Les instances. 100
5.17 Le modèle relationnel de données : partie commune. 102
5.18 Le modèle relationnel de données : partie spécifique. 103

6.1 Exemple de balise pour les dictionnaires. 106
6.2 Exemple avec la balise *<persName>*. 106
6.3 Le méta-modèle du TMF. 107
6.4 Exemple de représentation avec le Generic Mapping Tool. 108
6.5 Requête XML. 109
6.6 Le modèle de Prolexbase. 110
6.7 Résultat d'une requête XML. 112
6.8 Exemple de relations. 113
6.9 Architecture d'exportation de Prolexbase. 114
6.10 Traits du dictionnaire Prolex-Toponymes. 114

7.1 Page d'accueil de l'interface de consultation. 118
7.2 Recherche de noms propres. 119
7.3 Recherche avancée. 120
7.4 Menu *Texte*. 120
7.5 Informations sur le mot *luxembourgeois*. 121
7.6 Choix de la langue de l'interface. 121
7.7 Login et mot de passe. 122
7.8 Choix de la langue de travail. 122
7.9 Onglet *Consultation*. 123
7.10 Onglet *Ajout*. 124
7.11 Onglet *Pivot*. 124
7.12 Onglet *Alias*. 125
7.13 Onglet *Dérivé*. 125
7.14 Onglet *Eponymie*. 125
7.15 Onglet *Source*. 125
7.16 Onglet *Modification*. 126
7.17 Onglet *Modification d'une liste*. 126
7.18 Exemple de fichier. 127

7.19 Ajout d'un fichier. 127
7.20 Relation de méronymie. 128
7.21 Exemple de fichier multilingue. 128
7.22 Traduction de prolexèmes dans une autre langue. 129
7.23 Les erreurs. 129
7.24 Onglet *Attributs et Notes*. 130
7.25 Onglet *Compte*. 131

8.1 Exemple de cycle pour la relation de synonymie diachronique. 138
8.2 Exemple de cycle pour la relation de synonymie. 139
8.3 Repérages et expansions. 140
8.4 Extrait d'une partie du fichier de travail sur le Larousse Collège. 142
8.5 Nombre de prolexèmes extraits du Larousse Collège. 143
8.6 Le type Pensée. 144
8.7 Nombre de prolexèmes français de Prolexbase. 146

A.1 Le MLD serbe. 151
A.2 Le MLD français. 152
A.3 Le MCD des noms propres. 153

Introduction

Motivations

Personne ne peut nier l'omniprésence des noms propres dans notre quotidien. Pour le vérifier, il suffit juste par exemple de mettre la radio, d'allumer la télévision, d'écouter nos voisins de table à la terrasse d'un café, de tourner les pages d'un journal ou d'un magazine, de parcourir une œuvre littéraire, ou même de lire une thèse pour se rendre compte de leur importance au sein de notre communication. En répétant cette expérience dans d'autres langues, nous en tirerions sûrement la même conclusion.

Selon une étude de [Coates-Stephens, 1993], les noms propres représentent à eux seuls plus de 10% des textes journalistiques. Sans les noms propres, les textes perdraient sans doute beaucoup de leur richesse sémantique. De ces simples constats, nous ne pouvons négliger l'importance des noms propres et par conséquent admettre qu'ils méritent qu'on leur consacre de nombreux travaux.

Depuis longtemps, des linguistes, philosophes, informaticiens, etc. se sont intéressés aux noms propres. Ils ont essayé à travers leurs différents travaux de définir précisément leur statut et de présenter les mécanismes et relations qui les concernent.

Les noms propres constituent aussi depuis quelques années un problème étudié en traitement automatique des langues (TAL), où ils sont désormais connus comme faisant partie des *entités nommées*. La reconnaissance et le traitement des noms propres sont des thèmes que l'on retrouve dans plusieurs domaines du TAL, tels que la recherche d'information, les systèmes de réponse à des questions, la traduction automatique, l'indexation de documents, etc.

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. Certains logiciels, basés sur des méthodes statistiques, nécessitent par exemple des corpus d'apprentissage. D'autres applications ont besoin d'une liste de mots avec des informations linguistiques, telles que la morphologie, la syntaxe, etc., c'est-à-dire un dictionnaire électronique. Selon le système utilisé, la taille de ces dictionnaires électroniques varie énormément.

Dans le domaine du TAL, il existe aujourd'hui de nombreuses ressources dictionnairiques de noms communs (comme par exemple les Dictionnaires Électroniques du LADL (DELA) [Courtois, 1992], WordNet [Miller, 1995], Morphalou [Romary et al., 2004], le projet Papillon [Mangeot-Lerebours et al., 2003], etc.) et des ressources terminologiques spécialisées. De nombreuses listes de noms propres existent sur Internet, 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

[1] `http://www.cjk.org`

[2] `http://www.eurogeographics.org/`

[3] `http://press.jrc.it/NewsExplorer/home/fr/latest.html`

mais à 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 couplés avec une petite liste de mots liés aux noms propres suffit :

> *The collection of gazetteers need not be a bottleneck : through a judicious use of internal and external evidence relatively small gazetteers are sufficient to give good Precision and Recall.*

D'autres, comme [Ren and Perrault, 1992], considèrent que tout mot inconnu capitalisé peut être classé comme un nom propre. Cela est loin d'être vrai selon [Maurel, 2004], en raison de mots composés et d'homographes : prenons l'exemple du nom propre *Banco Real* (Le Monde du 15 janvier 1999), *Banco* n'est pas un mot inconnu (il a un homographe en français) et *Real* n'est pas un nom propre ; c'est le mot polylexical *Banco Real* qu'il est intéressant de catégoriser.

Dans le cadre de nos travaux, nous soutenons l'idée que la constitution de ressources lexicales de noms propres est nécessaire pour traitement des noms propres pour le TAL.

La recherche d'information, l'extraction d'information ou l'aide à la traduction nécessitent de délimiter précisément les noms propres, de les catégoriser et même, parfois, de les relier entre eux.

Cette thèse s'inscrit dans le cadre du projet Prolex [Maurel et al., 1996], dirigé par le professeur Denis Maurel au sein du Laboratoire d'Informatique de l'Université François-Rabelais de Tours (LI). Les recherches présentées ici ont reçu un soutien financier du ministère de l'Industrie dans le cadre du projet Technolangue NomsPropres. Nous allons décrire, dans cette thèse, les étapes de la conception et de l'implémentation de Prolexbase, un dictionnaire relationnel multilingue de noms propres destiné aux processus automatiques.

Plan de la thèse

Le chapitre 1 présente l'état de l'art sur les définitions, les caractéristiques et les typologies de noms propres à travers les différents travaux de linguistes, d'informaticiens, etc.

Le chapitre 2 est consacré à quelques grands projets de bases lexicales, tels que DELA, Wordnet, EuroWordnet, le DEC, le projet Papillon, etc.

Les chapitres 3 et 4 définissent les différents concepts du domaine des noms propres, les relations entre ces concepts, notre typologie et notre ontologie.

Le chapitre 5 décrit les entités et les associations de la base de données que nous avons créée, à partir des concepts et relations des chapitres 3 et 4, et le chapitre 6, les deux formats XML (un format de requête et un format d'exportation) pour l'échange de nos données.

Le chapitre 7 commente l'interface de consultation et l'interface de travail collaboratif.

Le dernier chapitre fait une évaluation du projet.

Première partie

État de l'art

Chapitre 1

Qu'est-ce qu'un nom propre ?

Introduction

Avant d'étudier les différents modèles de dictionnaires, nous allons dans un premier temps nous intéresser à la notion de nom propre. Qu'est-ce qu'un nom propre ? Quelles sont ses caractéristiques ou propriétés ? Les réponses à ces questions sont essentielles pour la bonne modélisation des noms propres. C'est pourquoi nous analysons les différents travaux réalisés à ce sujet dans le domaine de la linguistique et dans le domaine du TAL.

1.1 Définitions

Prenons par exemple ce texte :

> *C'est par un matin d'automne, juste un peu avant que les premiers rayons de lumière n'effleurent les rideaux de la chambre d'une maison vieillie par quelques hivers, que Diana décide d'oublier son dernier baiser échangé au milieu de la petite forêt de Grandmont avec cet admirable poète sans grand talent et aux rimes encore trop maladroites.*
>
> *Mickaël Tran, le 10 mai 2006 vers 1h du matin*

Tout lecteur humain reconnaîtra sans difficulté les noms propres suivants de cet extrait : *Diana*, *Grandmont*.

Chacun possède une définition assez intuitive de la notion de nom propre. Pourtant, il nous est difficile de donner une définition précise et complète du nom propre.

1.1.1 Définitions des dictionnaires

En cherchant la définition du nom propre dans les dictionnaires, nous obtenons les résultats suivants :

Définition 1 ***Le Larousse***
Nom propre, qui désigne un être ou un objet considéré comme unique (par oppos. à noms communs).

Définition 2 ***Hachette :***
le nom propre est une sous-classe particulière de la classe du nom. Il présente plusieurs spécificités, tant du point de vue sémantique que du point de vue syntaxique.

Définition 3 ***Le Robert***
I. Signe du langage (mot ou groupe de mots) servant à désigner un individu ou une classe d'individus et à les distinguer des êtres de la même espèce.
...
NOM PROPRE : nom désignant un individu et ne correspondant pas à un concept, à une notion (-> ci-dessus, I.).

En résumé, la plupart de ces définitions insistent sur le caractère unique du référent du nom propre. Celui-ci s'oppose au nom commun et possède une sémantique et une syntaxe qui lui est propre.

1.1.2 Définitions des linguistes

Il n'est pas aussi évident de définir précisément ce qu'est un nom propre. Plusieurs linguistes se sont intéressés à cette question :

Définition 4 ***[Molino, 1982]***
Les noms propres renvoient aux trois dimensions de la deixis, la personne, l'espace et le temps.

Définition 5 ***[M. Grevisse, 1986]***
Le nom propre n'a pas de signification véritable, de définition ; il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière.

Définition 6 ***[Gary-Prieur, 1994]***
Alors que l'interprétation d'un nom commun ne met en jeu que la compétence lexicale, celle du nom propre requiert presque toujours une mise en relation avec le référent initial, qui mobilise des connaissances discursives.

Définition 7 ***[Jonasson, 1994]***
Toute expression associée dans la mémoire à long terme à un particulier en vertu d'un lien dénominatif conventionnel stable.

Discussion

La définition de [Molino, 1982] appelle noms propres les noms de personnes, les noms de lieu et les noms d'événements. Que dire des noms d'entreprises, de marques, d'objets, de groupes, etc.[1] ?

[M. Grevisse, 1986] donne une définition des noms propres qui s'applique à des noms bien identifiés (des personnes et des lieux). Pourrait-on inclure dans cette définition des noms propres comme le *Jardin des Plantes* ?

Quand à [Gary-Prieur, 1994], elle insiste sur les relations sémantiques que les noms propres entretiennent entre eux et qui le plus souvent contribuent à leur interprétation en contexte[2]. Dans une entrée du dictionnaire, le nom propre est lui aussi mis en relation avec d'autres noms propres pour permettre son appropriation par le lecteur.

[Jonasson, 1994] 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

[1] voir la liste des types (section 4.2 chapitre 4 page 79)

[2] nous en tiendrons compte principalement par notre relation d'accessibilité (section 3.3.3 chapitre 3 page 66)

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 des entités nommées, qui est très largement utilisée dans le domaine du TAL depuis la conférence MUC de 1995 (voir section 1.3.2). C'est essentiellement à cette définition que nous nous référerons à partir de maintenant.

1.2 Statut linguistique du nom propre

De nombreux linguistes, philosophes, informaticiens, logiciens, anthropologues, etc. se sont intéressés à la question du nom propre. Dans cette partie, nous allons résumer les différentes caractéristiques du nom propre au niveau graphique, au niveau morphologique, au niveau syntaxique, au niveau sémantique et au niveau pragmatique.

1.2.1 Critère de la majuscule

Depuis le primaire et le collège, tous les livres de grammaire nous apprennent qu'un nom propre en français dans un texte commence par une majuscule :

> *Les pêcheurs maintenaient dimanche 30 avril le blocus du port d'**Arcachon**, en **Gironde**, entamé vendredi pour protester contre la hausse du prix du gazole.*
> *(Le Monde, 30/04/06)*

Cependant, le critère de la majuscule est loin d'être suffisant et ne s'applique pas pour toutes les langues. C'est le cas, par exemple, pour l'allemand où tous les noms commencent par une majuscule :

> *In völliger **Dunkelheit** und kalten **Strömungen** gedeihen **Korallen**, die kein **Schnorchler** und kein **Hobbytaucher** je zu **Gesicht** bekommt.*
> *(DER SPIEGEL, 02/04/2006)*

Historiquement, en France, la majuscule au début d'un nom propre est seulement apparue dans les textes avec le développement de l'imprimerie.

Il est impossible de distinguer à l'oral si un nom propre possède ou non une majuscule, pourtant cela empêche rarement sa bonne reconnaissance. Cela signifie donc que le critère de majuscule s'applique seulement au domaine de l'écrit.

L'emploi de la majuscule n'est pas limité aux noms propres, il concerne aussi quelques noms communs. Certains auteurs de la littérature française, afin d'accentuer la personnification de certains concepts, attribuent une majuscule à des noms communs comme *la Mort*, *le Destin*, *la Nature*, etc. La majuscule initiale est utilisée dans les textes pour les mots se trouvant en début de phrase, dans les titres, etc.

Dans le cas de noms propres composés, la majuscule n'apparaît pas pour chaque élément : *Organisation mondiale de la santé*, *la tour Eiffel*, *le quai d'Orsay*.

1.2.2 Critères morphologiques

En morphologie du français, les noms propres sont considérés comme une classe de mots invariables en genre et en nombre. Cependant, il existe des exceptions qui ne respectent pas cette règle. Notons que dans d'autres langues, la morphologie du nom propre varie suivant le cas, le genre, le nombre, etc.

Le genre des noms de personnes correspond le plus souvent au sexe de cette personne. Dans le domaine des noms de pays, il semblerait que les pays finissant par un *e* adoptent

souvent le genre féminin, comme *La Roumanie*, *L'Allemagne*, *La Suisse*, etc., alors que d'autres le genre masculin, comme *Le Congo*, *Le Nigeria*, *Le Canada*, etc. Quelques noms géographiques n'existent qu'au pluriel, comme *Les Pyrénées*, *Les Antilles*, *Les États-Unis*, etc.

Certains noms de famille royale français ou francisés prennent la marque du pluriel, comme *Les Bourbons*, *Les Stuarts*, *Les Césars*, etc., alors que les autres restent au singulier, comme *Les Dupont*, *Les Habsbourg*, *Les Durand*, etc. Les noms de journaux et les noms de marques ne prennent jamais la marque du pluriel (*deux Figaro*, *deux Peugeot*). Il existe quelques incertitudes : faut-il écrire *les deux Corée* ou bien *les deux Corées* ?

1.2.3 Critères syntaxiques

Contrairement à un nom commun, un nom propre apparaissant dans un texte peut être accompagné ou non d'un déterminant :

> *Le ministre n'a pas toujours une vision très claire de la géopolitique. Il s'est laissé plusieurs fois surprendre à confondre* ***Taïwan*** *et* ***la Thaïlande***, ***la Croatie*** *et* ***le Kosovo***.
> *(Le Monde, 27/04/06)*

En général, les noms de célébrité et de ville n'ont pas de déterminant. Il existe bien sûr des exceptions qu'il faut prendre en compte dans la création d'un dictionnaire électronique. Certaines catégories de noms propres, telles que les noms de pays (*la France*, *l'Allemagne*, etc.), les noms de bateaux (*le France*, *le Clémenceau*, etc.), les noms de restaurants ou d'hôtels (*le George V*, *le Vauban*, etc.), etc., possèdent en général un déterminant.

D'après [Jonasson, 1994], les noms propres sont dits non modifiés, s'il sont sans déterminant[3] et ne comportent pas d'adjectif, pas de complément de nom, etc. Un nom propre, tout comme un nom commun, peut être associé à des adjectifs (*Adj*), des noms communs (*Nc*), des déterminants (*Det*) pour former un groupe nominal plus ou moins complexe. [Jonasson, 1994] a listé au total onze constructions possibles autour d'un nom propre (Npr) non modifié :

0-Npr-0	*Paul danse*
0-Npr-Adj	*aux côtés de Reagan malade*
Adj-Npr-0	*Éblouissante Marlène*
Det-Adj-Npr	*Le/ce/mon vieux Théodule*
Det-Nc-Npr	*le bâtonnier Bernier*
Npr-le-Adj	*Marylin la blonde*
Npr-le-Nc	*Mailer le romancier*
Det-Npr-Adj	*l'Henri Brûlard stendhalien*
Det-Npr-de-N	*la Dora de Freud*
Ce-Nc-de-Npr	*cette andouillette de Maguy*
Ce-Npr-de-Nc	*ce Séguin de ministre*

Un nom propre, dans une phrase, peut occuper diverses fonctions comme les noms communs :

- sujet : *Mickaël est parti sur le chemin.*
- objet direct : *Il a vu Diana hier matin.*
- objet indirect : *Et il pense toujours à Diana dans le train.*

[3] exception faite du déterminant dont il est question ci-dessus.

– complément de nom : *En relisant les messages de Diana plein de chagrin.*

Ainsi, il apparaît que les noms propres ne sont pas fondamentalement différents des noms communs sur le plan de la syntaxe [Gary-Prieur, 1994]. Cependant pour [Leroy, 1994] :

> *Il n'en reste pas moins que, dans les cas les plus habituels, l'équivalence entre nom propre et nom commun, si communément admise, est en partie une illusion, construite par le discours grammatical sur la base d'une confusion entre nom (commun) et syntagme nominal.*

Nous ne rentrerons pas dans les détails de la sémantique apportée par leurs constructions syntaxiques [Garrigues, 1993].

1.2.4 Critères sémantiques

La signification des noms propres est un débat qui divise les linguistes. Il existe trois grandes théories qui tentent d'expliquer le sens du nom propre.

Certains linguistes, tels que S. Mill, K. Kripke, J. Molino, M. Noailly, K. Jonasson, etc., défendent la thèse que le nom propre n'a pas de sens et sert seulement d'"*étiquette référentielle*". Il sert seulement à désigner une personne ou un individu sans même le décrire. Selon [Kleiber, 1996], cette théorie présente à la fois un avantage et un inconvénient :

> *Son principal avantage est de rendre compte du fait que le nom propre ne décrit aucun attribut du porteur du nom. Son inconvénient majeur est de ne pouvoir expliquer comment se fait la référence. Elle ne peut, par exemple, montrer la différence entre les énoncés :*
>
> *Napoléon est mort à Sainte-Hélène*
> *Wellington est mort à Sainte-Hélène*
>
> *puisque absence de sens signifiant identité de sens, elle a pour résultat indésirable de rendre les deux énoncés identiques.*

Pour d'autres les noms propres possèdent un sens descriptif. Cette thèse est divisée en deux camps. D'un côté certains linguistes, tels que E. Buyssens, F. Kiefer, M. Gross, etc., soutiennent une version faible où le sens du nom propre se réduit à des spécifications ("ville", "fleuve", etc.) ou à des traits sémantiques généraux (+/- humain, +/- mâle, etc.). De l'autre, on trouve les partisans d'une version forte "*qui assigne aux noms propres un sens identifiant constitué d'une description (ou de descriptions) qui identifie univoquement le référent*" [Kleiber, 1996].

La dernière théorie considère le nom propre comme un prédicat de dénomination, que [Kleiber, 1996] définit de la façon suivante :

> *Nous avons fait l'hypothèse que le nom propre correspondait à un prédicat de dénomination* ***être appelé / N /*** *et qu'un nom propre non articulé représentait l'abréviation d'une description dénominative du type* ***le x appelé / N /*** *(cf.* ***Romulus*** *=* ***le x appelé / Romulus /****).*

Suite à de nombreuses critiques d'autres linguistes, [Kleiber, 1996] a abandonné cette hypothèse et a maintenu l'hypothèse d'un sens de dénomination du nom propre :

> *Ce sens dénominatif n'est alors plus conçu comme une propriété ou description du référent, mais comme l'instruction de chercher et de trouver dans la mémoire stable le référent qui porte le nom en question.*

1.2.5 Critères pragmatiques

Contrairement au nom commun, qui renvoie souvent à un ensemble d'individus, le nom propre se distingue habituellement par son unicité référentielle. Ainsi, le nom commun *arbre* désigne la classe de tous les arbres, que se soit par exemple un pommier, un cerisier, un peuplier, etc. Le nom propre *Romulus* désigne uniquement la personne ayant tué son frère jumeau *Remus* et ayant fondé *Rome*.

Il peut arriver quelque fois qu'un même nom propre renvoie à plusieurs individus différents, phénomène appelé homonymie. C'est le cas par exemple du nom propre *Paris* qui peut soit désigner une ville en France, soit un département, soit d'autres villes à travers le monde (aux États-Unis, au Gabon, à Haïti, au Togo, aux Philippines, à Sao Tomé et au Canada). Un nom de marque tel que *Kleenex* renvoie à un ensemble d'objets. Un nom de famille renvoie à tous les individus portant ce nom, etc.

Inversement, certains noms comme par exemple le *soleil* ou la *lune*, bien qu'ils renvoient à des entités uniques, ne sont pas considérés comme des noms propres mais comme des noms communs. Cela montre bien que le trait d'unicité référentielle ne suffit pas pour caractériser tous les noms propres.

1.2.6 Discussion

Dans cette partie, nous avons présenté les cinq traits linguistiques du nom propre. Ces différents traits, pris séparément ou ensemble, ne suffisent pas pour caractériser un nom propre. La plupart de ces traits permettent de ne caractériser qu'une partie spécifique des noms propres. Dans le cadre d'un travail multilingue, ces traits varient beaucoup d'une langue à l'autre. Construire un dictionnaire multilingue de noms propres nécessite obligatoirement d'inclure des informations graphiques[4] (précisant l'écriture d'un nom propre), syntaxiques (indiquant si celui-ci possède ou non un déterminant[5] et ses constructions possibles dans une phrase[6]) et flexionnelles[7]. Nous ne pouvons oublier d'inclure des informations sémantiques et pragmatiques qui peuvent apparaître sous forme de relations entre les noms propres.

1.3 Typologies des noms propres

Dans cette partie nous allons présenter différentes typologies des noms propres. Nous distinguons deux catégories de typologie, celles utilisées dans le domaine de la linguistique et celles qui ont conduit à des systèmes de reconnaissance des noms propres.

1.3.1 Typologies linguistiques

Typologie de Zabeeh

[Zabeeh, 1968] présente une typologie des noms propres en cinq classes :
- noms de personnes
- périodes de temps ou historiques
- artefacts (produits, arts, etc.)
- noms de lieux

[4] voir notre définition des alias (section 3.2.1 chapitre 3 page 57)
[5] voir détermination (section 5.2.5 chapitre 6 page 97)
[6] voir les grammaires associées aux expansions classifiantes (section 3.3.4 chapitre 3 page 69)
[7] voir les instances (section 3.4 chapitre 3 page 72)

– noms d'institutions politiques ou économiques

Typologie de Bauer

[Bauer, 1985] a proposé une classification pragmatique des noms propres en cinq classes :
– *Anthroponymes* : personnes individuelles ou groupes.
– *Toponymes* : noms de lieux.
– *Ergonymes* : objets ou produits manufacturés.
– *Praxonymes* : faits historiques, maladies ou événements.
– *Phénonymes* : phénomènes météorologiques, astres, etc.

Typologie de Grass

Inspiré des travaux de Bauer, [Grass, 1999] propose une typologie des noms propres définie sur deux niveaux :
– *Anthroponymes* : patronymes, prénoms, pseudonymes, gentilés[8], hypocoristiques, ethnonymes, groupes musicaux modernes, ensembles artistiques et orchestres classiques, partis et organisation, clubs sportifs, noms donnés aux animaux familiers (zoonymes).
– *Toponymes* : pays, villes, microtoponymes, hydronymes, oronymes, installations militaires, monuments.
– *Ergonymes* : marques, entreprises, établissements d'enseignement et de recherche, titres de livres, de films, de publications et d'œuvres d'art, objets mythiques.
– *Praxonymes* : faits historiques, maladies, événements culturels.
– *Phénonymes* : ouragans, zones de haute et de basse pression, astres et comètes.

1.3.2 Typologies issues du TAL

Typologie de Coates-Stephens

La typologie de [Coates-Stephens, 1993] comporte huit classes :
– noms de personnes
– noms de lieux
– noms d'organisations
– noms d'origines ou de gentilés
– noms de législations
– noms de sources d'informations (média, journaux, etc.)
– noms d'événements (guerres, révolutions, catastrophes, etc.)
– noms d'objets (artefacts, produits, etc.)

MUC

Les conférences MUC [MUC-6, 1995] (Message Understanding Conference), pendant plusieurs années, se sont intéressées à l'extraction d'information. Il s'agit d'extraire à partir d'un ensemble de textes journalistiques en anglais des informations sur un thème précis (actes de terrorisme en Amérique latine, rachats d'entreprise, etc...) et de comparer les performances des systèmes développés par chaque participant.

Avec MUC-6, les noms propres, ainsi que les dates et les unités chiffrées sont regroupées sous le terme d'entités nommées [Chinchor, 1997] :

[8] *Nom que portent les habitants d'une ville, d'une région, d'un pays, etc.* (définition du dictionnaire Hachette encyclopédique édition 2000).

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

Les entités nommées de MUC sont réparties en trois catégories :

- ENAMEX :
  - PERSON : noms de personnes ou de familles.
  - ORGANISATION : noms de sociétés, de gouvernements, ou d'autres organisations.
  - LOCATION : villes, provinces, pays, régions internationales, hydronymes, montagnes, etc...
- TIMEX
  - DATE : expressions de date.
  - TIME : expressions de temps.
- NUMEX
  - MONEY : expressions monétaires.
  - PERCENT : pourcentages.

Voici un exemple de texte balisé selon le format de MUC-6 :

<ENAMEX TYPE = "LOCATION"> Washington </ENAMEX>, <TIMEX TYPE = "DATE"> March 7 </TIMEX> (<ENAMEX TYPE = "ORGANIZATION"> Bloomberg </ENAMEX>) – <ENAMEX TYPE = "ORGANIZATION"> MCI Communications Corp. </ENAMEX> and <ENAMEX TYPE = "ORGANIZATION"> News Corp. </ENAMEX> said they will pay <ENAMEX TYPE = "ORGANIZATION">Loral Corp.</ENAMEX> more than <NUMEX TYPE = "MONEY"> $400 million</NUMEX> to build two satellites for a direct television broadcasting venture.

Typologie de Paik, Liddy, Yu et McKenna

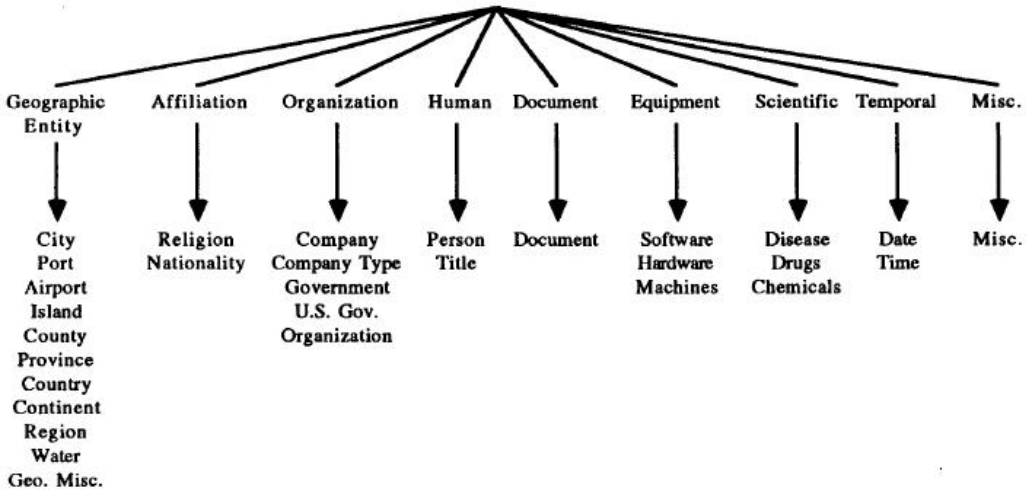

FIG. 1.1 – Typologie de Paik, Liddy, Yu et McKenna

[Paik et al., 1996] ont établi une classification des entités nommées à partir d'un corpus du *Wall Street Journal*. Cette classification (figure 1.1) comprend neuf classes et trente catégories :

– *Geographic Entity* : ville, port, aéroport, île, comté, province, pays, continent, région, hydronyme, nom géographique divers.
– *Affiliation* : religion, nationalité.
– *Organization* : entreprise, type d'entreprise, gouvernement, état américain, organisation.
– *Human* : personne, titre.
– *Document* : document.
– *Equipment* : logiciel, matériel, machine.
– *Scientific* : maladie, médicament, produit chimique.
– *Temporal* : date, heure.
– *Misc* : divers.

Leur système a réussi à répartir 89 % des 588 entités nommées extraites de ce corpus sur 29 catégories et le reste a été classé dans la catégorie **Misc** (divers).

Typologie de Sekine, Sudo et Nobata

Certaines applications du traitement automatique des langues, telles que l'extraction d'information, système de question/réponse, etc., nécessitent le balisage des entités nommées. Pour cela, [Sekine et al., 2002] ont développé une hiérarchie d'entités nommées comprenant 150 types différents (figure 1.2).

La première phase de la construction de leur typologie passe par une classification manuelle de 3 500 candidats pouvant prétendre être des entités nommées sur un corpus de journaux anglais (*Wall Street Journal*, *New York Times* et *Los Angeles Times*).

La deuxième phase comprend une étude et une comparaison de différents systèmes d'extraction d'information et d'autres travaux afin d'étendre la hiérarchie obtenue lors de la première phase.

Enfin, la dernière phase consiste à consulter des thésaurus comme Wordnet et Roget Thesaurus dans le but de trouver des catégories d'entités nommées ayant été oubliées lors des deux précédentes phases.

CoNLL

CoNLL-2002 (Conference on Natural Language Learning) [Sang, 2002] est dédiée à l'extraction d'entités nommées dans des textes, autres que l'anglais. Au cours de cette conférence, les participants ont pu tester leurs systèmes sur des textes en espagnol et en hollandais.

CoNLL divise les entités nommées en 4 catégories :
– PER : noms de personnes.
– ORG : organisations.
– LOC : noms de lieux.
– MISC : noms divers.

Voici un exemple de texte balisé par CoNLL :

> *[PER Wolff], currently a journalist in [LOC Argentina], played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid].*

Un exemple de format du corpus d'entraînement fournit par CoNLL est donné dans la Figure 1.3.

L'étique B-XXX (Begin) indique le premier mot d'une entité nommée. I-XXX (Inside) indique que le mot repéré est interne à une entité nommée. O indique que le mot ne fait

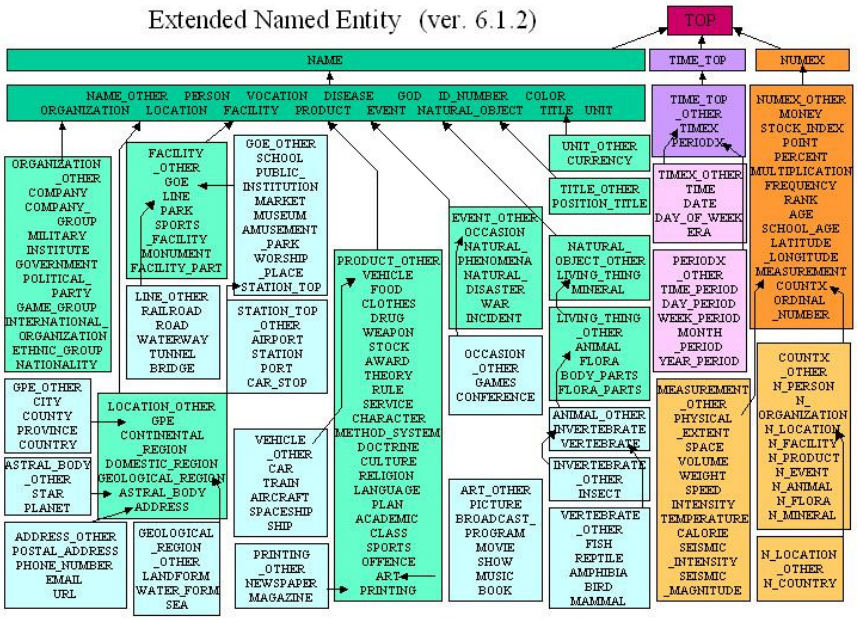

FIG. 1.2 – Typologie de Sekine

pas partie d'une entité nommée.

ESTER

La campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques (ESTER) [Le Meur et al., 2004] [Gravier et al., 2004] s'intéresse à l'indexation et à la transcription d'enregistrements audio de journaux télévisés et de journaux radiophoniques français.

Pour pouvoir annoter manuellement les corpus de transcription, un typage des entités nommées (voir figure 1.4) a été mis en place. Cette typologie comprend huit classes et vingt-neuf sous-classes d'entités nommées :

- Les personnes (pers) : personnes ou animaux.
- Les organisations (org) : organisations politiques, religieuses, culturelles...
- Les lieux (loc) : lieux naturels, bâtis par des humains, adresses postales, téléphoniques, électroniques et fax.
- Les constructions humaines (fac).
- Les productions humaines (prod) : moyens de transport, œuvres, récompenses.
- Les dates et heures (time).
- Les quantifiables (amount) : montants, unités de mesure...
- Incertain (unk) : cas n'appartenant à aucune des classes précédentes.

Discussion

Il n'est pas évident à partir des définitions de noms propres données par les auteurs de dictionnaires ou les linguistes de décider si un nom appartient ou n'appartient pas à la classe des noms propres. L'utilisation d'une typologie de noms propres permet souvent d'aider dans cette prise de décision.

Wolff	B-PER
,	O
currently	O
a	O
journalist	O
in	O
Argentina	B-LOC
,	O
played	O
with	O
Del	B-PER
Bosque	I-PER

FIG. 1.3 – Format du corpus de CoNLL

On distingue deux types de typologies : les typologies destinées à des systèmes automatiques et les typologies linguistiques. La plupart du temps, les typologies automatiques possèdent des classes assez restreintes (par exemple 4 pour CoNLL) et la reconnaissance d'un nom propre se base souvent sur des repères internes ou externes. Les noms inclassables suivant ces critères sont réunis dans une classe spécifique. Un contre-exemple est la typologie de [Sekine et al., 2002], très détaillée, avec 150 types. La typologie linguistique développée par Thierry Grass, à partir des travaux de [Zabeeh, 1968] et de [Bauer, 1985], basée sur deux niveaux (4 supertypes et 34 types) s'avère judicieuse de ce point de vue. Tout d'abord, les supertypes peuvent être utilisés par des systèmes de reconnaissance automatique de noms propres, puis les types, en nombre facilement mémorisable, peuvent être attribués par un expert humain. Nous avons au cours de notre thèse retravaillé sur la typologie de Thierry Grass et nous l'avons complétée en nous inspirant des typologies présentées précédemment (voir chapitre 3 page 53). La typologie à laquelle nous avons abouti a été testée sur un numéro du journal Le Monde [Maurel, 2004] et nous a fourni un cadre stable pour un classement manuel de plus de 7 000 noms propres extraits du Larousse Collège.

Conclusion

Au cours de ce premier chapitre, nous avons montré qu'il est difficile de définir ce qu'est précisément un nom propre. Sa définition est un sujet controversé et divise la plupart des linguistes. Parmi les différentes définitions proposées, nous avons adopté la définition de K. Jonasson, car celle-ci nous a paru avoir une meilleure couverture que les autres définitions.

Nous avons aussi présenté le statut des noms propres en linguistique notamment à travers leurs différents traits, comme leurs traits sémantiques, syntaxiques, morphologiques et pragmatiques. De cette étude nous pouvons conclure qu'aucun de ces traits, pris séparément ou réunis, ne suffit pour définir précisément la notion de nom propre et situer la frontière entre les noms propres et les noms communs.

L'étude des noms propres dans le domaine du TAL nécessite la présence d'une typologie. Selon différents projets, le nombre de classes utilisées dans la typologie varie énormément (entre 4 et 150 classes). Dans le cadre de nos travaux, nous avons choisi d'adapter la typologie proposée par T. Grass, en la complétant d'une étude manuelle de tous les noms propres d'un numéro du journal Le Monde [Maurel, 2004]. Notre typologie s'est inspirée des typologies linguistiques et des typologies du TAL. Nous sommes partis des propositions de Thierry

Classe	Sous-classe	Exemple
pers	pers.hum	[**Richard Virenque**] a gagné...
	pers.anim	mon chien [**Médor**]...
	pers.imag	[**Superman**]...
org	org.pol	Le [**parti communiste français**]...
	org.edu	L'[**ENSAM**]...
	org.non-profit	l'[**ANPE**]...
	org.com	[**France Inter**]...
gsp	gsp.pers	les habitants du nord et le reste de la [**France**]...
	gsp.org	la [**France**] a signé un accord...
	gsp.loc	Ils se sont retrouvés en [**France**]...
loc	loc.geo - loc.geo.line	Le [**Mont-Blanc**]... l'[**autoroute A4**]...
	loc.addr - loc.addr.post - loc.addr.tel - loc.addr.elec	 au [**3 avenue de Matignon 75008 Paris**]... c'est le [**01 45 65 90 90**]... [**www.telecharger.com**]...
fac		...se transfère au [**palais de Chaillot**]...
prod	prod.vehicule	la fusée [**Ariane**]...
	prod.award	le [**prix nobel de la paix**]...
	prod.art prod.printing	la comédie musicale [**Roméo et Juliette**]... [**Les Fleurs du Mal**]...
time	time.date - time.date.abs - time.date.rel	 le [**11 septembre 2001**]... c'était [**hier**]...
	time.hour	sur le coup de [**4 heures du matin**]...
amount	 amount.phy - amount.phy.age - amount.phy.dur - amount.phy.temp - amount.phy.len - amount.phy.wei - amount.phy.spd	 ce monsieur a [**87 ans**]... pendant [**3 semaines**]... il faut [**moins 20 degrés**]... il mesure [**1 mètre 85**] et pèse [**70 kilos**]... le mistral approchera [**60**] [**80 Km/h**]...
	amount.cur	[**7,5 euros**]...
unk		le [**trône d'Angleterre**]...

FIG. 1.4 – Typologie utilisée par la campagne ESTER.

Grass et de Denis Maurel pour définir une typologie "moyenne" en nombre de classes. Nous l'avons, dans le cadre de cette thèse, intégrée à une ontologie.

Chapitre 2

Les ressources dictionnairiques

Introduction

Le dictionnaire, outil dans la diffusion du savoir de notre société, est le résultat d'un long processus de développement et de représentation de notre connaissance des langues.

Les premiers dictionnaires sont apparus dans l'Antiquité sous forme de listes de mots, comme les listes bilingues akkadien-sumérien (vers 2400 av. J.-C.), les listes de mots de la Grèce antique, dont par exemple celle de Protagoras d'Abdère contenant des mots difficiles extraits des poèmes d'Homère (vers le V^e siècle av. J.-C.), ou encore les dictionnaires chinois (II^e siècle av. J.-C.).

C'est seulement vers 1502 qu'Ambrogio Calepino[1] va publier le Dictionarium (dictionnaire bilingue latin-italien), qui au fil de ses éditions va devenir le tout premier dictionnaire multilingue avec onze langues (latin, grec, italien, espagnol, français, allemand, hébreu, flamand, anglais, polonais et hongrois) en 1588.

Aujourd'hui, les dictionnaires papier tels que le Larousse, le Robert ou bien d'autres encore font partie intégrante de notre vie quotidienne. Avec le développement de l'informatique, la plupart des dictionnaires existant sur support papier ont été mis sur support électronique et commercialisés sur CD-ROM, sur DVD-ROM ou bien sont accessibles sur Internet. Il s'agit d'un nouveau type de dictionnaire, que nous allons appeler dictionnaire informatisé.

Depuis une vingtaine d'années, de nombreux chercheurs ont développé un grand nombre de modèles de bases de données lexicales ou dictionnaires électroniques formalisés, que nous appellerons dictionnaires électroniques. Les dictionnaires électroniques comportent des données spécifiques destinées à l'analyse automatique des langues. Nous pouvons distinguer deux types d'usage d'un dictionnaire électronique : usage humain ou usage automatique. Un dictionnaire électronique à usage humain contient souvent des informations implicites qui nécessitent une connaissance de la part du lecteur et qui ne sont pas adaptées aux machines. Un dictionnaire électronique servant de données pour des programmes de TAL a besoin d'informations explicites et non ambiguës.

Dans cette partie, nous allons présenter uniquement les projets qui nous ont inspiré dans la construction de notre dictionnaire. Nous avons utilisé les travaux sur les codes flexionnels des dictionnaires DELA. Les projets EuroWordNet et Papillon montrent la nécessité d'utiliser une approche par pivot dans la structure d'un dictionnaire multilingue. Nous découvrirons une stratégie de peuplement de base lexicale à travers le projet Papillon. Nous avons étudié les relations sémantiques de WordNet et du DEC pour définir les relations

[1] voir l'article *dictionnaire* de l'Encyclopédie Hachette Multimédia

spécifiques aux noms propres.

2.1 Travaux du LADL

Sous la direction de Maurice Gross, le Laboratoire d'Automatique Documentaire et Linguistique (LADL) de l'université de Paris VII a développé plusieurs dictionnaires électroniques, qui peuvent être regroupés en deux catégories. La première catégorie comporte les dictionnaires de formes non fléchies : le DELAS [Courtois, 1992] pour les mots monolexicaux, le DELAP [Laporte, 1990] pour la phonémisation des mots monolexicaux et le DELAC [Silberztein, 1990] pour les mots polylexicaux. La seconde catégorie regroupe les dictionnaires de formes fléchies : le DELAF, le DELAPF et le DELACF.

[Courtois, 1992] définit ainsi l'objectif des dictionnaires du LADL :

> *Un objectif des dictionnaires électroniques est de construire des structures où sont répertoriées les unités de la langue, avec un certain nombre de propriétés nécessaires au traitement automatique.*

Le DELAS, ou Dictionnaire Électronique du LADL de formes simples, pour le français comporte environ 80 000 entrées de mots monolexicaux, c'est-à-dire des séquences de lettres. Une entrée du DELA se présente sous la forme suivante :

abacule,N1+z3

abajoue,N21+z3

cheval,N4+Anl+z1

où le mot *abacule* correspond à la forme canonique. Le code *N1* indique que ce mot est un nom qui suit la classe morphologique numéro 1 : (0,-,s,-) (voir Annexe B) ; *z3* est un code sémantique permettant de préciser que le mot *abacule* appartient à un langage spécialisé, contrairement au mot *cheval*. La figure 2.1 et la figure 2.2 présentent les codes grammaticaux et les codes sémantiques du DELAS [Paumier, 2006].

En appliquant les règles de flexion sur le DELAS, nous obtenons le Dictionnaire Électronique du LADL de formes fléchies ou DELAF, constitué d'environ 900.000 formes fléchies. Une entrée du DELAF se présente sous la forme suivante :

mercantiles,mercantile.A+z1:mp:fp

glace,.N+z1:fs

où *mercantiles* correspond à la forme fléchie et *mercantile* à la forme canonique (ou lemme). *A+z1* précise que ce mot est un adjectif appartenant au langage courant. *mp* et *fp* indiquent que *mercantiles* est la forme du masculin pluriel et aussi la forme du féminin pluriel de la forme canonique *mercantile*.

La structure du DELAC (Dictionnaire Électronique du LADL de mots composés) et celle du DELACF (Dictionnaires Électronique du LADL de mots composés fléchis) sont identiques aux deux dictionnaires précédents. Le DELACF est constitué de plus de 100 000 mots composés (90 000 noms, 15 000 constructions être Prép N, 8 000 adverbes et 500 conjonctions).

Dans le dictionnaire DELAP (Dictionnaire phonémique) et DELAPF (Dictionnaire phonémique de formes fléchies), chaque entrée comporte en plus une représentation phonémique de sa prononciation. Le DELAPF contient environ 620 000 entrées. Voici un exemple du DELAP :

phonémique,fonemik,.A31

Code	Signification
A	adjectif
ADV	adverbe
CONJC	conjonction de coordination
CONJS	conjonction de subordination
DET	déterminant
INTJ	interjection
N	nom
PREP	préposition
PRO	pronom
V	verbe

FIG. 2.1 – Codes grammaticaux du DELAS.

Code	Signification
z1	mot courant
z2	mot rare
z3	mot technique
Abst	abstrait
Anl	animal
AnlColl	animal collectif
Conc	concret
ConcColl	concret collectif
Hum	humain
HumColl	humain collectif
t	verbe transitif
i	verbe intransitif
en	particule pré-verbale (PPV) obligatoire
se	verbe pronominal
ne	verbe à négation obligatoire

FIG. 2.2 – Traits du DELAS.

Il existe aussi des dictionnaires du LADL pour l'allemand, l'anglais, le coréen, l'espagnol, le grec, l'italien, le norvégien, le portugais, le serbe et le thaïlandais.

Multiflex [Savary, 2006] est un programme qui permet de fléchir des mots polylexicaux [Savary, 2000] à partir de leur lemme. Pour cela, un formalisme [Savary, 2005] permettant de décrire la création des formes fléchies a été mis en place.

Des données spécifiques pour chaque langue sont nécessaires. Voici un exemple pour le polonais :

Polish
<CATEGORIES>
Nb : sing, pl
Case : Nom, Gen, Dat, Acc, Inst, Loc, Voc
Gen : masc_pers, masc_anim, masc_inanim, fem, neu
<CLASSES>
noun : (Nb,<var>),(Case,<var>),(Gen,<fixed>)

adj :(Nb,<var>),(Case,<var>),(Gen,<var>)
adv :

La première partie de ce fichier décrit les catégories grammaticales (nombre, cas, genre) qui existent en polonais. La deuxième partie précise pour chaque classe grammaticale si celle-ci varie suivant le nombre, le cas ou le genre. En lisant ce fichier, on constate que le genre des noms polonais est toujours fixe et qu'ils varient suivant le nombre et le cas, tandis que les adverbes sont invariables.

Les mots polylexicaux sont découpés en unités et chaque unité est associée à une variable ($1, $2...). Par exemple, le mot *Athens '04* est décomposé en cinq unités :

$1=Athens
$2=<espace>
$3='
$4=0
$5=4

Chaque unité est associée à un code flexionnel, sauf si celle-ci est invariable. Par exemple :

avant-garde(garde.N21:fs)

Dans cet exemple, le code *N21* indique que l'unité *garde* suit la règle morphologique : (-,0,-,s). Le code *fs* signifie féminin singulier.

On attribue à chaque mot polylexical un code flexionnel :

avant-garde(garde.N21:fs),NC_XXN

Les codes flexionnels de Multiflex peuvent être représentés sous la forme d'un graphe (figure 2.3). Pour notre exemple, on aura donc dans la variable $1 le mot *avant*, dans $2 le trait d'union et dans $3 le mot *garde*. L'expression *Gen==$g* signifie que le genre est fixe et qu'il correspond au genre de la troisième unité, c'est-à-dire féminin. L'expression *Nb=$n* indique que le nombre peut être variable et prendre toutes les valeurs de sa catégorie, à savoir singulier et pluriel. *<Gen=$g ;Nb=$n>* précise le genre et le nombre du résultat qui sont déterminés par l'unification et qu'ils s'accordent avec le genre et le nombre de la 3ème unité. En appliquant le programme Multiflex, on obtiendra le résultat suivant :

- *avant-garde,avant-garde.NC_XXN:fs*
- *avant-gardes,avant-garde.NC_XXN:fp*

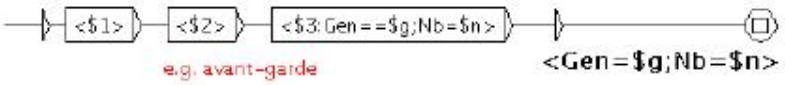

FIG. 2.3 – Code flexionnel NC_XXN.

Discussion

A partir des travaux du LADL et de Multiflex, nous retiendrons la nécessité d'utiliser des codes flexionnels qu'il faudra associer à chaque lemme afin de générer automatiquement toutes les formes fléchies d'un nom propre (voir sections 5.2.6 et 5.2.7 du chapitre 5 page 99). Le but de notre thèse n'étant pas de développer un autre système de codes flexionnels, nous utilisons, pour le cas du français et du serbe, les codes du DELAS pour les noms

propres monolexicaux (voir Annexe B) et envisageons d'utiliser les codes de Multiflex pour les noms propres polylexicaux.

De plus, nous avons prévu, dans le cadre de travaux futurs, de développer un système identique pour la génération d'alias et de dérivés de noms propres (voir sections 3.2.1 et 3.2.2 du chapitre 3 pages 57 et 61).

2.2 EuroWordNet

Avant de présenter la base de données lexicale multilingue EuroWordNet, il nous paraît indispensable de commencer par une description du projet WordNet, qui constitue sans doute une référence indispensable à connaître dans le monde des dictionnaires électroniques et qui sert de point de départ à EuroWordNet.

2.2.1 WordNet

Développé en 1985 par des linguistes du Laboratoire des Sciences Cognitives de l'Université de Princeton, sous la direction de G. A. Miller, WordNet [Miller, 1995] est une base de données lexicales anglaises dont la conception a été inspirée des théories psycholinguistiques et informatiques sur la mémoire lexicale humaine. L'objectif de ce projet est de lister, de classifier et d'établir des relations entre le contenu lexical et le contenu sémantique de la langue anglaise. La version actuelle de WordNet (2.1), consultable sur le site www.cogsci.princeton.edu, comporte plus de 150 000 mots.

WordNet est un réseau lexical où chaque nœud correspond à un synset et chaque arc est formé par les relations entre synsets. Le synset (ou *synonym set*) est défini comme un ensemble de mots interchangeables, représentant un sens particulier. Par exemple, le nom propre anglais *Paris* (figure 2.4) appartient à quatre synsets différents.

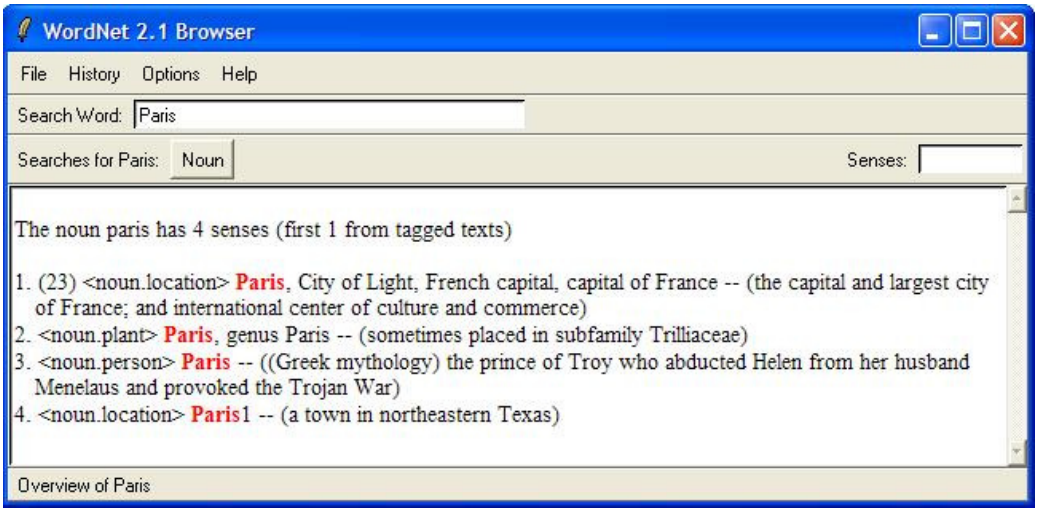

FIG. 2.4 – Recherche du nom propre *Paris* dans WordNet.

Dans WordNet, le lexique est partitionné en quatre catégories syntaxiques : nom, verbe, adjectif, adverbe (figure 2.5). Ce découpage est basé sur une hypothèse cognitive, selon laquelle les mots dans notre mental sont classés en fonction de leur catégorie syntaxique. Chaque catégorie syntaxique possède sa propre hiérarchie de classes sémantiques et ses

POS	Unique Strings	Synsets	Word-Sense Pairs
Noun	117 097	81 426	145 104
Verb	11 488	13 650	24 890
Adjective	22 141	18 877	31 302
Adverb	4 601	3 644	5 720
Totals	155 327	117 597	207 016

FIG. 2.5 – Nombre de mots et de concepts dans WordNet 2.1.

propres relations sémantiques. Il n'existe aucune relation entre des unités lexicales de catégories syntaxiques différentes.

Les noms sont regroupés dans vingt-cinq classes :

- act, action, activity
- attribute, property
- quantity, amount
- natural object
- plant, flora
- event, happening
- animal, fauna
- body, corpus
- relation
- natural phenomenon
- possession
- food
- artifact
- process
- group, collection
- person, human being
- communication
- substance
- location, place
- time
- motive
- shape
- state, condition
- cognition, knowledge
- feeling, emotion

Les verbes sont regroupés en quinze familles :

- body : verbs of grooming, dressing and bodily care.
- change : verbs of change of size, temperature, intensity, etc.
- cognition : verbs of thinking, judging, analyzing, doubting, etc.
- communication : verbs of telling, asking, ordering, singing, etc.
- competition : verbs of fighting, athletic activities, etc.
- consumption : verbs of eating and drinking.
- contact : verbs of touching, hitting, tying, digging, etc.
- creation : verbs of sewing, baking, painting, performing, etc.
- emotion : verbs of feeling.
- motion : verbs of walking, flying, swimming, etc.
- perception : verbs of seeing, hearing, feeling, etc.
- possession : verbs of buying, selling, owning, and transfer.
- social : verbs of political and social activities and events.
- stative : verbs of being, having, spatial relations.
- weather : verbs of raining, snowing, thawing, thundering, etc.

Les adjectifs sont divisés en deux classes :

- adjectifs descriptifs (*big*, *interesting*)
- adjectifs relationnels, qui sont des dérivés de noms (*fraternal*, *presidential*)

Les adverbes ne possèdent aucune structure hiérarchique dans WordNet.

WordNet est construit autour de deux relations principales : la synonymie, qui est modélisée à travers le concept de synset, et l'hyperonymie (figure 2.6), une relation transitive permettant de construire une hiérarchie entre les synsets.

Autour des synsets, WordNet a défini d'autres relations sémantiques (figure 2.7). La méronymie, relation inverse de l'holonymie, permet de spécifier si un synset est une partie

=> entity, something
=> object, physical object
=> artifact, artefact
=> instrumentality, instrumentation
=> conveyance, transport
=> vehicle
=> motor vehicle, automotive vehicle
car, auto, automobile, machine, motorcar

FIG. 2.6 – Exemple de relation d'hyperonymie dans WordNet.

d'un autre synset. L'antonymie exprime les sens opposés entre les synsets. La relation d'implication (*entailment*) s'applique uniquement pour les verbes.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless
Hyponymy (subordinate)	N	sugar maple, maple maple, tree
Meronymy (part)	N	brim, hat gin, martini
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

FIG. 2.7 – Relations sémantiques dans WordNet.

2.2.2 EuroWordNet

Le projet européen EuroWordNet [Vossen, 1998], coordonné par P. Vossen de l'université d'Amsterdam, a été lancé en 1996. L'objectif d'EuroWordNet est de construire une base de données lexicales multilingue contenant plusieurs langues européennes. Comportant au départ seulement quatre langues (néerlandais, italien, espagnol et anglais), EuroWordNet s'est achevé pendant l'été 1999 avec quatre langues de plus (allemand, français, estonien et tchèque).

Selon [Vossen et al., 1997], il existe plusieurs manières de développer une base de données multilingue :

– La première solution, sans doute la plus coûteuse, consiste à créer des liens par paire de langues. Pour une base de données multilingue contenant quatre langues, il faudrait 12 liens interlingues différents (néerlandais → italien, italien → néerlandais, néerlandais → espagnol, espagnol → néerlandais, néerlandais → anglais, anglais → néerlandais, italien → espagnol, espagnol → italien, italien → anglais, anglais → italien, espagnol → anglais, anglais → espagnol). L'ajout d'une nouvelle langue peut s'avérer très compliqué. La complexité du problème augmente avec le nombre de langues.

– Une deuxième solution consiste à créer une langue artificielle structurée qui va servir d'interlangue. La mise en place d'une langue artificielle nécessite de résoudre plusieurs difficultés. Le lexique doit être précis et assez large pour pouvoir englober les lexiques des différentes langues. L'ajout d'une nouvelle entrée dans une langue peut parfois amener à revoir et améliorer la langue artificielle.
– Une autre solution serait de prendre une des langues comme pivot. Mais cela rend le modèle dépendant de la structure de la langue servant de pivot. Si un sens donné d'un mot est absent dans la langue pivot alors qu'il existe dans une autre langue, cela peut aussi être gênant pour le modèle.
– Une quatrième solution, celle qui a été adoptée par les concepteurs d'EuroWordNet, envisage d'utiliser un ensemble de concepts non structurés, qui servent de liens interlingues. L'avantage d'une telle solution est que cette liste d'index non structurée ne doit respecter aucune théorie linguistique ou cognitive, car elle contiendra simplement des numéros d'identité uniques et ne possédera pas de structure interne. De plus l'ajout d'une nouvelle langue ne remettra pas en cause la totalité de l'index ou les relations que les wordnets entretiennent déjà avec l'index, mais seulement une petite partie de celui-ci.

L'architecture globale d'EuroWordNet [Vossen, 1999] [Jansen, 2004] (figure 2.8) est formée de trois niveaux. Le premier niveau comprend les différentes bases de données lexicales monolingues, qui ont été développées suivant le modèle de WordNet 1.5. Le deuxième niveau, indépendant des langues, comprend un *Inter-Lingual-Index* (ILI). Les synsets de wordnets monolingues ayant été reliés à un même élément de l'ILI (*enregistrement-ILI*) seront considérés comme des concepts équivalents. L'ensemble des synsets de WordNet 1.5 a servi de point de départ à l'ILI d'EuroWordNet. Le dernier niveau contient une ontologie de domaine (*Domain Ontology*) et une ontologie supérieure (*Top Ontology*) (figure 2.9) [Vossen et al., 1998]. L'ontologie supérieure fournit une hiérarchie sémantique des différents enregistrements-ILI et l'ontologie de domaine permet de répartir les enregistrements-ILI selon des thèmes (sport, hôpital, restaurant, trafic aérien, etc.).

L'ontologie supérieure se décompose en trois parties :
– Entité du premier ordre (*1stOrderEntity*) : entité concrète de notre environnement. Par exemple : *Comestible (Function)*, *Living (Natural, Origin)*, etc.
– Entité du deuxième ordre (*2ndOrderEntity*) : situation statique ou dynamique. Par exemple : *length (Property)*, *day (Time)*, etc.
– Entité du troisième ordre (*3rdOrderEntity*) : entité non observable. Par exemple : *idea*, *thought*, *information*, *theory*, *plan*, etc.

Contrairement à WordNet, EuroWordNet autorise des relations entre les différentes catégories syntaxiques. Dans le projet EuroWordNet, il existe deux types de relations : les relations internes d'une langue entre les synsets (figure 2.10) et les relations entre les synsets et les enregistrements-ILI.

Voici les relations les plus importantes entre les enregistrements-ILI et les synsets d'EuroWordNet :
– EQ_SYNONYM : si le synset correspond à un seul et unique enregistrement-ILI (synset : diventare IT / enregistrement-ILI : to become).
– EQ_NEAR_SYNONYM : si un synset correspond à plusieurs ILI-records, si plusieurs synsets correspondent à un même enregistrement-ILI, ou encore s'il y a des doutes sur le choix de l'enregistrement-ILI.
– EQ_HAS_HYPERONYM : si un synset est plus spécifique que les enregistrements-ILI disponibles (synset : kunstproduct NL (artifact substance) / enregistrements-ILI : artifact ; product).

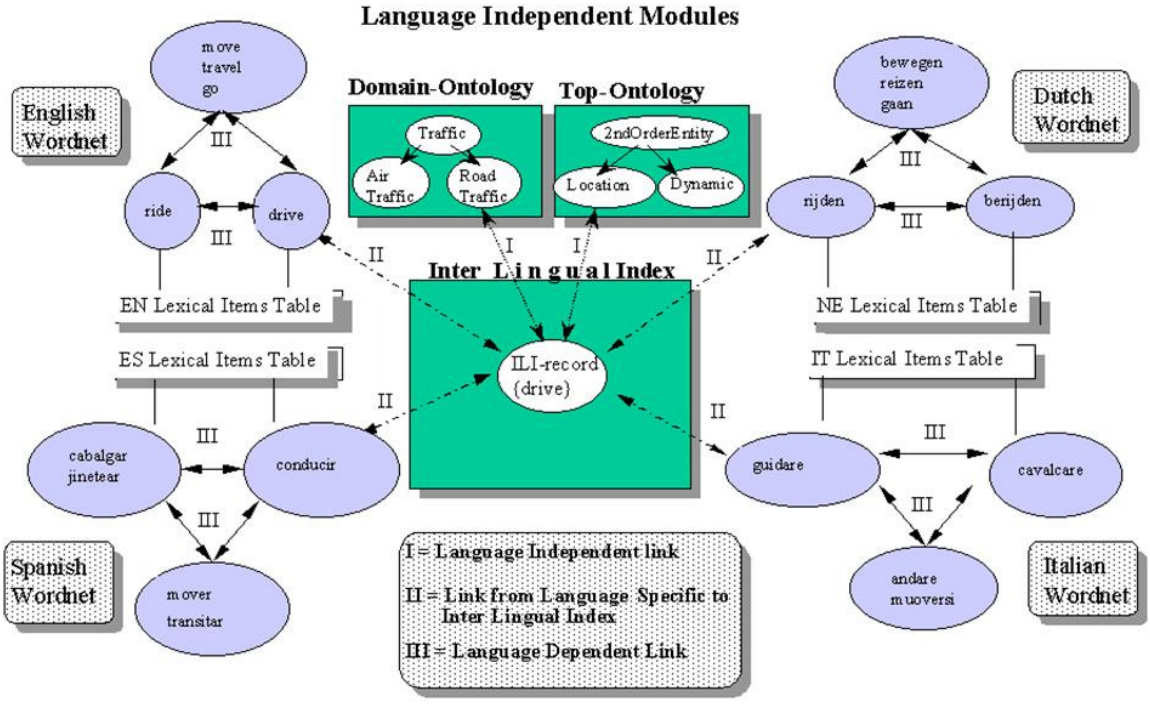

FIG. 2.8 – Architecture d'EuroWordNet

– EQ_HAS_HYPONYM : si un synset peut être associé à plusieurs enregistrements-ILI (synset : dedo ES (a finger or toe) / enregistrements-ILI : toe ; finger).

2.2.3 Balkanet : une extension d'EuroWordnet

Le projet Balkanet [Tufis et al., 2004] [Krstev et al., 2004] est une extension d'EuroWordNet appliquée aux langues des Balkans et à quelques autres langues européennes : bulgare, grec, roumain, serbe, turc et tchèque. Ce projet a débuté en septembre 2001 et s'est achevé en août 2004.

L'objectif du projet Balkanet est de traduire un ensemble de 8 000 concepts dans les six langues du projet pour produire des ressources lexicales et des outils pour le TAL qui soient assez flexibles et réutilisables par différentes applications. L'ILI de Balkanet (BILI) est le résultat de deux sélections sur l'ILI d'EuroWordNet. Le premier ensemble de concepts sélectionné par Balkanet correspond aux concepts de base (Base Concept) d'EuroWordNet, qui regroupent les concepts les plus utilisés et qui constituent une référence pour d'autres concepts. Le dernier ensemble est formé des enregistrements-ILI qui ont été utilisés par la plupart des langues d'EuroWordNet.

Les concepteurs de Balkanet se sont vite aperçus que l'utilisation des enregistrements-ILI d'EuroWordNet posait certains problèmes :

– la difficulté à trouver une traduction juste pour chaque enregistrement-ILI.
– le manque d'information sur les enregistrements-ILI pour pouvoir affecter les liens entre chaque synset des WordNets de Balkanet et les enregistrements-ILI.
– la non structuration des ILI, qui peut nuire à l'application du projet dans des systèmes d'extraction d'information.

En raison de ces problèmes, ils ont décidé de remplacer l'ensemble des enregistrements-ILI d'EuroWordNet par WordNet 1.7 et de considérer l'anglais comme langue pivot.

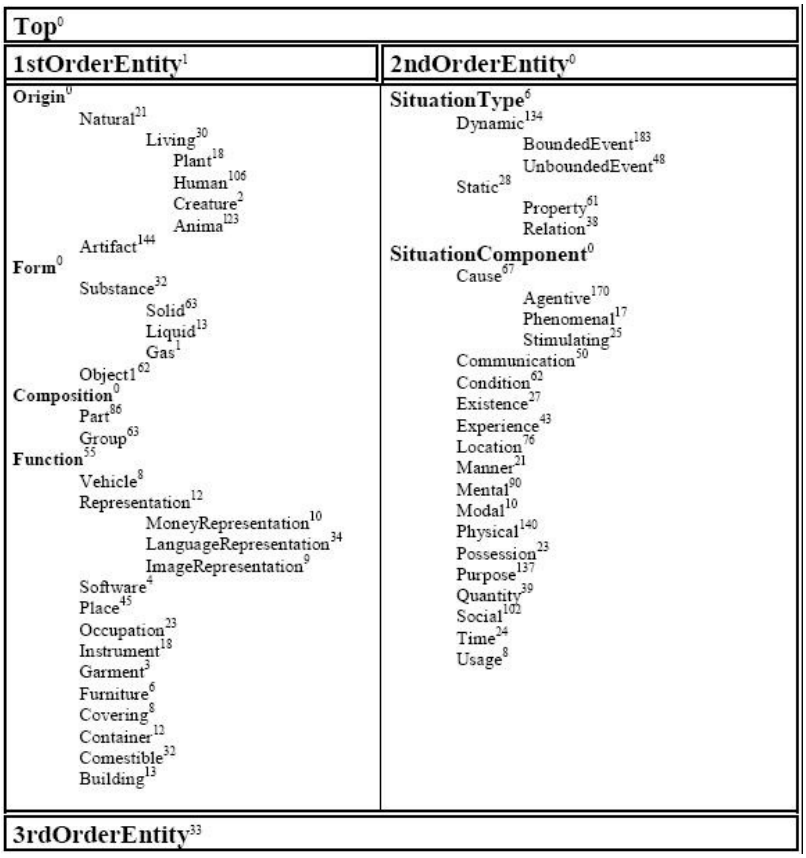

FIG. 2.9 – EuroWordNet Top-Ontology

2.2.4 Discussion

L'étude du projet WordNet et de ses extensions EuroWordNet et Balkanet nous a permis de constater l'importance de la relation de synonymie par rapport aux autres relations sémantiques. Comme dans ces projets, nous nous sommes inspiré de la relation de synonymie pour développer nos différents concepts du domaine des noms propres (voir chapitre 3 page 53). Cette étude nous a aussi permis de connaître les relations sémantiques qui peuvent exister entre différents concepts. Pour le cas des noms propres, nous avons retenu de WordNet les relations de méronymie, de synonymie et d'hyperonymie.

On retrouve dans ces projets de nombreux noms propres. Il s'agit essentiellement de noms propres les plus connus, comme *Victor Hugo*, *Paris*, *Europe*, etc.

Les projets EuroWordNet et Balkanet utilisent un niveau interlingue, basé sur la notion de pivot, qui s'avère indispensable pour créer une base de données lexicale multilingue. Les éléments qui sont reliés à un même pivot correspondent entre eux à une traduction d'une langue vers l'autre. Nous ferons de même (voir section 3.1.1 chapitre 3 page 54).

Dans ces projets, on ne trouve pas d'informations syntaxiques et flexionnelles associées à chaque entrée. De plus, ils ne précisent pas si une entrée correspond à la forme dérivée d'une autre entrée. Nous n'avons pas d'informations sur le contexte d'une relation de synonymie et pourtant ce contexte (politique, savant, familier, etc.) peut s'avérer utile dans l'aide à la traduction. Par exemple, il serait incorrect de traduire *j'ai passé mes vacances sur les plages de France* par *I spent my holidays on the beaches of the French Republic*. Il serait

SEMANTIC RELATION	EXAMPLE
near_synonym	tools <> instrument
xpos_near_synonym	movement <> move
has_hyperonym	mercedes > car
has_hyponym	car > mercedes
has_xpos_hyperonym	election > to vote
has_xpos_hyponym	to fear > paranoia
has_holonym	
has_holo_part	wheel > car
has_holo_member	player > team
has_holo_portion	liquid > drop
has_ holo_made_of	wood > stick
has_holo_location	center > city
has_meronym	
has_mero_part	car > wheel
has_mero_member	team > player
has_mero_portion	drop > liquid
has_mero_made_of	stick > wood
has_mero_location	city > center
antonym	man > woman
near_antonym	to give > to take
xpos_near_antonym	to love > hate
causes	to try > to succeed
is_caused_by	to succeed > to try
has_subevent	to sleep > to snore
is_subevent_of	to pay > to buy
role	hammer > to hammer
role_agent	dog > to bark
role_instrument	sail > to sail
role_patient	learner > to teach
role_location	school > to teach
role_direction	
role_source_direction	ship > disembark
role_target_direction	casa > rincasarse
role_result	ice > to freeze
role_manner	loudly > shout
involved	to hammer > hammer
involved_agent	to teach > teacher
involved_patient	to learn > learner
involved_instrument	to hammer > hammer
involved_location	to teach > school
involved_direction	to pass > place
involved_source_direction	to race > the start
involved_target_direction	to collapse > ground
involved_result	to crystalize > crystal
...	

FIG. 2.10 – Relations internes d'une langue entre les synsets dans EuroWordNet.

indispensable dans le cas des noms propres d'avoir une relation sémantique qui puisse relier un auteur à ses œuvres, une capitale à un pays, etc.[2]

L'utilisation des synsets pour modéliser les noms propres présente quelques inconvénients. Un synset regroupe un ensemble de mots ou de groupes de mots qui sont en relation de synonymie. On ne peut associer directement à un synset des informations spécifiques à un élément du synset (la flexion, les dérivés, la phonétique, les règles de création d'alias ou de dérivés, etc.). Voici quelques exemples de synsets :

{Paris, City of Light, French capital, capital of France} (1)
{Musset, Alfred de Musset, Louis Charles Alfred de Musset} (2)
{France, French Republic} (3)

Nous ne pouvons associer le dérivé *Parisian* au synset (1) car *Parisian* est uniquement le dérivé de *Paris* et non le dérivé de *French capital*. Il faudra préciser que cette information sera uniquement liée à l'élément *Paris* de ce synset. De plus, si l'on souhaite associer la relation d'accessibilité *Paris* est la capitale de la *France*, cela risque de poser un problème. Cette relation devra s'appliquer à tous les élements du synset. Ainsi, on aurait la relation *City of Light* est la capitale de la *France*. Cette relation est peut-être sémantiquement correcte mais elle n'apparaîtra pas dans la plupart des textes.

Peut-on dire que *Paris* et *French Capital* sont des synonymes ? Si jamais le pays change de capitale cela risque de ne plus être vrai. Cela devient compliqué pour le cas du *Jammu-et-Cachemire* qui possède deux capitales : *Srinagar* en été et *Jammu* en hiver.

A cause de ces raisons, nous avons décidé de séparer chaque élément d'un synset en plusieurs éléments différents. Parmi les relations de synonymie (voir section 3.3.1 chapitre 3 page 63), on distingue deux types de synonymie : *Musset* et *Alfred de Musset* versus *Paris* et *City of Light*.

En anglais, dans le cas de l'exemple (1), nous allons créer deux prolexèmes différents : un pour le nom propre *Paris* et un pour le nom propre *City of Light*. Cela permettra d'associer le dérivé *Parisian* au prolexème *Paris*. Celui-ci sera en relation de synonymie avec un contexte stylistique avec le prolexème *City of Light*. La relation de synonymie entre *Paris* et *capital of France* ou *French capital* sera modélisée sous la forme d'une relation d'accessibilité et d'une relation d'expansion classifiante (voir section 3.3.4 chapitre 3 page 69).

Dans le cas de l'exemple (2), les noms propres *Musset* et *Alfred de Musset* sont des variantes (voir alias section 3.2.1 page 57) de la forme vedette *Louis Charles Alfred de Musset*. Nous créerons un prolexème *Louis Charles Alfred de Musset* et nous associerons ces deux alias à ce prolexème, soit directement, soit par règles.

2.3 Le Trésor de la Langue Française informatisé

Le Trésor de la Langue Française informatisé (TLFi) est un dictionnaire monolingue français principalement destiné aux humains. Il est accessible gratuitement sur Internet à l'adresse suivante : `http://atilf.atilf.fr`.

Ce dictionnaire est une adaptation électronique des 16 volumes du Trésor de la Langue Française (TLF) [Dendien and Pierrel, 2003], qui a débuté en 1993 dans les locaux du laboratoire CNRS de l'Institut National de la Langue Française (INALF), puis s'est poursuivie dans le laboratoire d'Analyse et Traitement Informatique de la Langue Française (ATILF) de Nancy. Entièrement encodé dans un format XML (figure 2.11), le TLFi comporte environ

[2] voir relation d'accessibilité section 3.3.3 page 66.

100 000 mots vedettes, 270 000 définitions et 430 000 exemples extraits de la base textuelle Frantext[3].

<art><ved><mot>RÉMITTENT, -ENTE, </mot><cod>adj.</cod></ved>
<sync><H><paramage/><B><dom> MÉD. </dom><cro>[En parlant d'une affection, d'un trouble, d'un symptôme] </cro><def n="t"> Qui présente des poussées et des atténuations successives. </def> <exe n="e"> On a décrit un tétanos discontinu ou rémittent (<aut> CAMUS, GOURNAY </aut><tit> ds Nouv. Traité Méd. <ct> fasc. 2 1928 </ct></tit><loc> , p. 803 </loc><dum>). </dum></exe><syntita n="i"> Psychose rémittente (<so> POINSO-GORI 1972 </so><dum>). </dum></syntita></B><H>
...
<rbbg> BBG. ARVEILLER (R.). Doc. lexicogr. tirés des dict. In : [Mél. Wartburg (W. von)]. Tübingen, 1968, p. 268. QUEM. DDL t. 9.</rbbg>
</art>

FIG. 2.11 – Extrait de l'article *RÉMITTENT* du TLFi.

Discussion

Le TLFi ne comporte pas de noms propres. Nous retiendrons essentiellement l'idée d'une interface de consultation performante, adaptée à un large public (voir section 7.2 chapitre 7 page 118) et celle d'associer une transcription phonétique à chaque nom propre (voir section 5.2.5 chapitre 5 page 97). Notre but n'étant pas de créer une encyclopédie sur les noms propres mais de créer une ressource linguistique destinée à des applications du TAL, nous n'allons pas associer à chaque nom propre une définition, des commentaires ou des exemples comme le ferait le TLFi.

Nous envisageons cependant de faire apparaître des liens vers des encyclopédies, comme Wikipédia, ou d'autres ressources lexicales, comme EuroWordNet et Framenet (voir section 3.3.4 chapitre 3 page 69) en utilisant la relation d'expansion classifiante.

2.4 Dictionnaire Explicatif et Combinatoire

Le Dictionnaire Explicatif et Combinatoire (DEC) est un dictionnaire développé par [Mel'čuk, 1999] pour le russe et le français. Chaque article de ce dictionnaire est élaboré suivant les méthodes définies dans la Lexicologie Explicative et Combinatoire, issue de la théorie Sens-Texte [Mel'čuk et al., 1995].

Le terme explicatif dans le nom du dictionnaire insiste sur le fait que chaque article du dictionnaire est décomposé suivant ses différents sens et selon une méthode rigoureuse. Le DEC est un dictionnaire combinatoire car les combinatoires lexicales et syntaxiques de chaque unité lexicale sont exhaustivement détaillées. Ce dictionnaire contient environ 558 vocables répartis sur quatre volumes.

La construction d'une entrée du DEC doit obligatoirement respecter à la fois une certaine microstructure, organisant la structure des articles (définition, connotations, régimes, etc.), et une certaine macrostructure, régissant l'ensemble des articles.

La macrostructure du DEC s'articule autour de deux notions : la lexie et le vocable. Une lexie ou unité lexicale est définie soit comme un sens particulier d'un mot (lexème),

[3]Frantext est un corpus constitué de plus de 3 600 œuvres littéraires datant du XIX^{e} jusqu'au XX^{e}.

soit comme une locution (phrasème), alors qu'un vocable correspond à un regroupement de lexies. Voici l'exemple d'un vocable du DEC :

> **MÉPRIS**, nom, masc.
> I. Attitude émotionnelle défavorable...[*le mépris pour ce corrupteur*]
> II. Opinion selon laquelle quelque chose n'a pas d'importance...[*le mépris du danger des convenances*]

La première ligne indique la forme graphique du vocable suivie de sa morphologie (catégorie, genre et/ou nombre). En dessous du vocable, une liste de lexies est présentée sous forme d'arborescence.

Chaque article est structuré en trois zones : la zone sémantique, la zone combinatoire et la zone phraséologique.

La zone sémantique, dont l'objectif est de fournir une définition du contenu sémantique d'une lexie, est elle-même divisée en deux parties. La première partie donne une définition lexicographique, dont voici un exemple :

> I. *Mépris de X envers Y pour Z* = Attitude émotionnelle défavorable de X à l'égard de Y causée par le fait suivant : X croit que les actions, l'état ou les propriétés Z de Y causent que Y n'a pas de valeur morale ou sociale ; cette attitude est celle qu'on a normalement dans de pareilles situations.

Le défini, généralement en italique, est un phrasème contenant la lexie vedette et ses actants sémantiques, c'est-à-dire les arguments de prédicat sous forme de variables (X, Y, Z). A droite de l'égalité (=) se trouve le définissant, qui explicite, en utilisant les actants sémantiques, le sens de la lexie.

Après la définition lexicographique, une liste de connotations, qui parfois peut être absente, est donnée :

Connotations

> 1) Cœur I.1a est le siège des sentiments [voir CŒUR I.4a].
> 2) Cœur I.1a est le siège de l'intuition [voir CŒUR I.4b].
> 3) Cœur I.1a qui bat représente la vie [voir les phrasèmes correspondants dans le CŒUR I.1.a].

La zone combinatoire, elle-même divisée en deux parties, renseigne le lecteur sur toutes les combinaisons de syntaxe possibles d'une lexie et sur les liens qu'entretient cette lexie avec d'autres lexies.

La première partie, appelée zone de combinatoire syntaxique, se présente sous la forme d'un tableau de régime, où les colonnes représentent les actants sémantiques et les lignes listent les valeurs possibles que peuvent prendre ces actants, et une liste de contraintes, dont voici un extrait :

Régime

1=X	2=Y	3=Z
1. *de* N 2. A_{poss} 3 A	1. *de* N 2. *pour* N 3. *envers* N 4. *à l'égard de* N	1. *pour* N

$\left.\begin{array}{l} 1) C_3 \textit{ sans } C_2 \\ 2) C_{1.1} + C_{2.1} \\ 3) C_{1.2} + C_{3.1} \end{array}\right\}$: impossible

$4) C_{1.2} + C_{2.1}$: impossible si $C_{2.1}$ désigne une personne

$5) C_{1.3} + C_{2.1}$: non souhaitable

C_1 : *le mépris de Paul, son mépris, le mépris populaire*

C_2 : *le mépris de pour, envers, à l'égard de ce collègue son hypocrisie*

$C_1 + C_2 + C_3$: *le mépris de Paul, son mépris, le mépris populaire envers à l'égard de ce ministre pour son hypocrisie ses propos diffamatoires, son mépris de l'art pour son inefficacité*

La dernière partie de la zone combinatoire, appelée zone de combinatoire lexicale, liste les fonctions lexicales qui peuvent être appliquées sur cette lexie. Une fonction lexicale est indépendante des langues et se présente sous la forme suivante : f(x) = $\{y_1, y_2...y_k\}$. y_i est une valeur de f sur x. Par exemple, *Magn(fièvre)={forte, élevée ; de cheval}*. Voici un extrait d'une liste de fonctions lexicales et de leurs résultats pour la lexie *MÉPRIS* :

Syn : dédain, irrespect, condescendance, arrogance, hauteur II, morgue, **litt** mésestime, **litt** superbe
Anti : respect I
Anti : considération, égard ; différence, estime
Magn : grand, profond, absolu, souverain, sans bornes ; hautain, froid

L'article se termine par des exemples extraits de textes littéraires contenant la lexie vedette :

> *L'Anglaise reconnut sa rivale et fut glorieusement anglaise ; elle nous enveloppa d'un grand regard plein de son mépris anglais et disparut dans la bruyère avec la rapidité d'une flèche [H. de Balzac]. Je vais peut-être vous paraître vieux jeu, mais j'ai un mépris sans bornes pour ces femmes qui vont d'amant en amant, le plus souvent sans amour, pour des raisons de prestige ou de caractère [A. Maurois]. Rien ne m'a donné un absolu mépris du succès que de considérer à quel prix on l'obtient [G. Flaubert]. Le mépris âcre et froid des passants lui pénétrait dans la chair et dans l'âme comme la bise.*

lexie vocable	lexie num	lexie cgs	lexie formuleEtiquette	FL formuleFL	FL lexie	phraseme phraseme	exemple exemple
LION	I	nom, masc	animal sauvage ou espèce animale	QSyn	_Roi des animaux_	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	fauve	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	félin	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Gener	carnivore	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Femelle du L.	lionne	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.
LION	I	nom, masc	animal sauvage ou espèce animale	Petit du L.	lionceau	_avoir bouffé du lion_	Les lions, gavés d'antilopes, ne semblaient pas vouloir nous attaquer.

FIG. 2.12 – Dico

Le projet Dico (Dictionnaire combinatoire) (figure 2.12), dont l'objectif est de construire une base lexicale pour le français comprenant 3 000 vocables, est une simplification du DEC. La finalité de ce projet est de pouvoir fournir des lexiques destinés au TAL et de produire un dictionnaire accessible au grand public (`http://www.olst.umontreal.ca/`).

Discussion

Le DEC ne possède pas de noms propres parmi ses entrées. Bien que la syntaxe soit un sujet intéressant, nous n'avons pas l'ambition de faire la même chose. Nous avons retenu du DEC l'association possible d'une entrée avec un phrasème (voir la relation d'éponymie dans la section 3.3.5 chapitre 3 page 69). Nous nous sommes inspiré du DEC pour définir la relation d'accessibilité[4].

2.5 Papillon

Le projet Papillon [Mangeot-Lerebours et al., 2003] [Mangeot-Lerebours, 2001] est né en 2000 à la suite d'une collaboration entre le GETA-CLIPS et le NII (National Institute of Informatics) de Tokyo. L'objectif du projet est de créer une base lexicale multilingue à usage humain et pour des agents logiciels. Destiné au départ au français et au japonais, ce projet s'est étendu à d'autres langues, comme l'allemand, l'anglais, le malais, le laotien, le thaï, le vietnamien et le chinois.

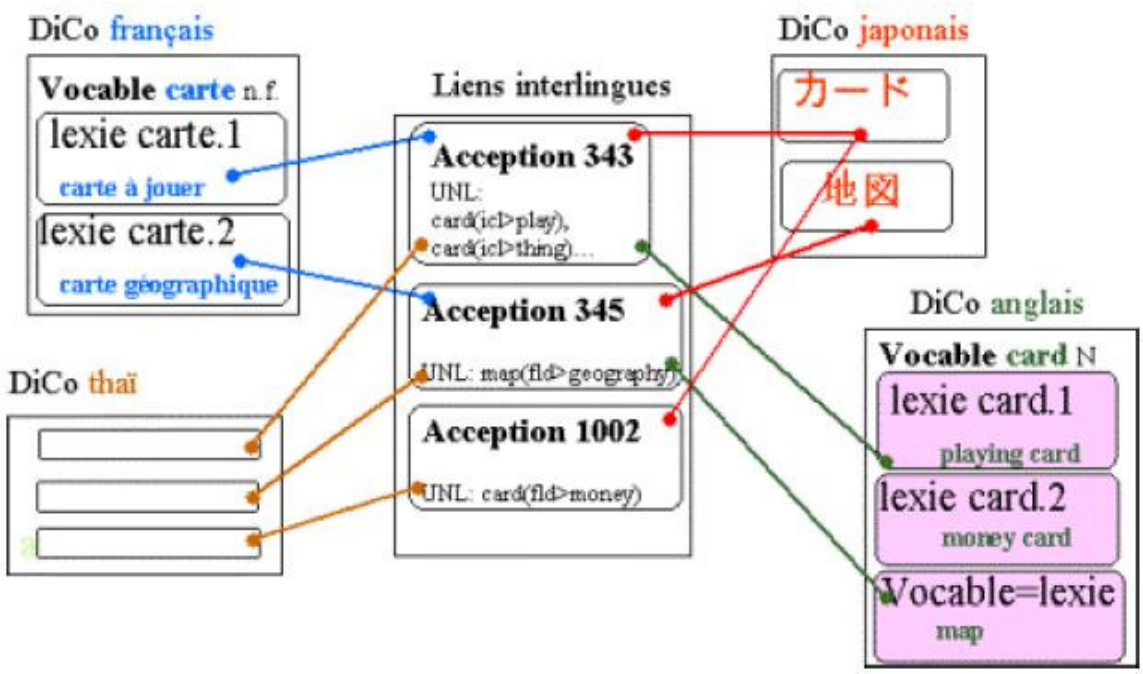

FIG. 2.13 – Macrostructure de Papillon

La macrostructure de la base Papillon (figure 2.13) repose sur la notion d'axie, c'est à dire d'acception interlingue. Dans sa thèse, [Sérasset, 1994] définit cette notion ainsi :

> *Une acception monolingue est une unité sémantique d'une langue. Elle est locale à une langue de la base. [...]*
> *Le but essentiel de la base lexicale est de fournir un lien entre les acceptions monolingues des différents dictionnaires. Pour cela, nous définissons l'ensemble des acceptions interlingues comme étant l'union des ensembles d'acceptions monolingues de différents dictionnaires de la base.*

[4]L'origine de cette dernière était la relation de chef de [Mel'čuk, 1999] ; voir section 3.3.3 du chapitre 3 page 66.

Dans la figure 2.13, le vocable anglais *card* possède trois lexies : *playing card*, *money card* et *map*. Ces trois lexies sont reliées à des acceptions différentes dans le niveau interlingue. Ces relations permettent de trouver la traduction d'une lexie dans une autre langue. Ainsi, la lexie *map* sera traduit en français par la lexie *carte géographique*.

La microstructure de la base lexicale Papillon (figure 2.14) s'est inspirée de celle utilisée dans la base DiCo.

```
<element name="lexie">
<complexType>
<sequence>
<element ref="d:headword" minOccurs="1" maxOccurs="1"/>
<element ref="d:writing" minOccurs="0" maxOccurs="1"/>
<element ref="d:reading" minOccurs="0" maxOccurs="1"/>
<element ref="d:prononiation" minOccurs="0" maxOccurs="1"/>
<element ref="d:pos" minOccurs="1" maxOccurs="1"/>
<element ref="d:langage-level" minOccurs="0" maxOccurs="1"/>
<element ref="d:semantic-formula" minOccurs="1" maxOccurs="1"/>
<element ref="d:government-pattern" minOccurs="0" maxOccurs="1"/>
<element ref="d:lexical-functions" minOccurs="0" maxOccurs="1"/>
<element ref="d:examples" minOccurs="0" maxOccurs="1"/>
<element ref="d:full-idioms" minOccurs="0" maxOccurs="1"/>
<element ref="d:more-info" minOccurs="0" maxOccurs="1"/>
</sequence>
<attribute ref="d:id" use="required"/>
</complexType>
<element>
```

FIG. 2.14 – Schéma XML des lexies.

Papillon-CMD comporte plus d'un million d'entrées dans huit langues différentes comprenant à la fois des noms communs et des noms propres. Papillon-NADIA, basé sur le format DiCo, contient toutes les fonctions lexicosémantiques de la lexicographie explicative et combinatoire entre les entrées de dictionnaires monolingues et n'est pas limité aux noms communs.

La stratégie de construction de Papillon passe par deux étapes. La première consiste en une construction automatique au cours de laquelle des dictionnaires ou ressources existantes sont récupérés et intégrés dans Papillon. Au cours de la deuxième étape, des contributeurs pourront travailler à partir des entrées obtenues lors de l'étape précédente.

Une interface de consultation est proposée aux internautes. Elle permet aussi à toute personne extérieure au projet de contribuer au développement de la base lexicale. Une fois validées par des experts, leurs contributions pourront être intégrées définitivement dans la base Papillon. Elle est accessible à l'adresse suivante : `http://www.papillon-dictionary.org/Home.po`.

Discussion

La construction d'un dictionnaire peut se faire suivant plusieurs stratégies. Il peut s'agir d'une construction manuelle (comme le DEC), automatique ou mixte (à la fois automatique et manuelle, comme dans le projet Papillon). La construction manuelle nécessite gé-

néralement un temps de construction plus long et un coût plus cher qu'une construction automatique.

Notre stratégie de peuplement de Prolexbase consiste dans un premier temps à récupérer manuellement les noms propres d'un dictionnaire papier (voir section 8.3 du chapitre 8 page 141) et des listes de toponymes des précédents travaux du projet Prolex. Il s'agit ensuite de les convertir suivant un format spécifique (voir section 7.18 du chapitre 7 page 127) et de les intégrer dans notre dictionnaire. Nous avons prévu d'utiliser le programme d'extraction automatique de noms propres de [Friburger, 2002] pour remplir automatiquement notre dictionnaire.

La construction d'un dictionnaire électronique multilingue nécessite un travail collaboratif. Pour cela, il est nécessaire de posséder une interface permettant de remplir la base de données (voir section 7.3 chapitre 7 page 119). Étant donné le manque de moyens et de personnes, nous n'avons pas d'experts pour vérifier chaque donnée rentrée dans notre base de données. De ce fait, dans notre projet, seuls des experts peuvent contribuer au développement de notre dictionnaire électronique.

Nous constatons aussi qu'un niveau interlingue est indispensable pour un dictionnaire multilingue. Nous n'avons pas retenu l'idée d'une acception par langue car notre nom propre conceptuel ne correspond pas à un sens d'un nom propre mais à un certain point de vue sur le référent de ce nom propre suivant un diasystème (voir chapitre 3 page 53).

Conclusion

L'étude des différents modèles de dictionnaires électroniques a été très enrichissante et nous a permis d'observer les différentes stratégies mises en place dans la conception de leur structure. Il existe de nombreux autres modèles de dictionnaires électroniques, mais nous n'avons présenté dans cette partie que ceux qui nous ont paru les plus intéressants pour notre projet.

La conception d'un dictionnaire électronique nécessite de spécifier au départ plusieurs paramètres. En fonction de ces paramètres, l'architecture à adopter pour construire le dictionnaire peut varier énormément.

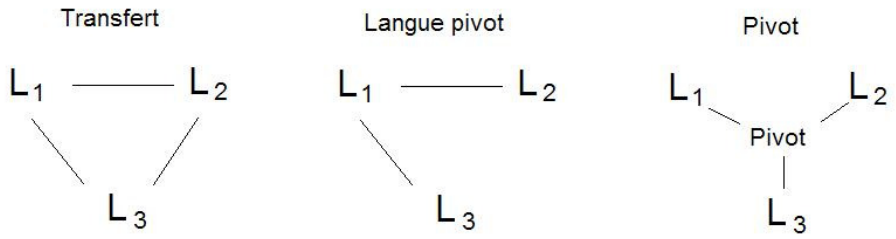

Fig. 2.15 – Différentes architectures

Le premier paramètre, sans doute le plus important, est le nombre de langues que l'on compte inclure dans le dictionnaire électronique, car, selon que le dictionnaire soit monolingue, bilingue ou multilingue, son architecture ne sera pas la même. La macrostructure d'un dictionnaire monolingue se présente sous la forme d'un unique volume, tandis que celle d'un dictionnaire bilingue nécessite au moins deux volumes, un contenant les entrées d'une langue avec les liens vers l'autre langue et vice-versa. La figure 2.15 présente les différentes architectures que l'on peut utiliser pour développer un dictionnaire contenant plus de deux

langues. L'approche multilingue par pivot a été mise en œuvre dans le projet EuroWordNet et dans le projet Papillon sous la forme d'axie. L'approche par transfert a été utilisée dans le projet Eurotra pour la traduction automatique [Danlos, 1989]. L'approche par langue pivot a été utilisé dans le projet Balkanet. Cette approche est recommandée par l'Afnor [Francopoulo, 2003] :

> *Dans un dictionnaire bilingue, nous avons besoin d'un lien pour traduire un sens en un autre et on pourrait imaginer qu'il suffit d'un simple lien entre deux sens [...]. Si cette stratégie est viable pour deux langues, elle est intenable pour un nombre de langues plus important [...]. Nous représentons les traductions via un objet intermédiaire [CN RNIL N 7 : 2003-11-25].*

Il faut aussi définir le format des entrées et les informations linguistiques que l'on souhaite intégrer dans notre dictionnaire. Selon l'utilisation envisagée, des informations syntaxiques, sémantiques ou morphologiques peuvent se révéler indispensables.

Deuxième partie

Modélisation des noms propres

Chapitre 3

Autour du nom propre conceptuel

Introduction

Cette partie est entièrement consacrée à décrire en détail tous les concepts clés de la modélisation des noms propres et les relations qu'entretiennent ces différents concepts entre eux. Ces concepts nous ont finalement permis de proposer une modélisation sous la forme d'un graphe orienté structuré sur plusieurs niveaux, pouvant parfois être très complexe en fonction de la langue que l'on souhaite intégrer à notre dictionnaire électronique.

C'est à partir des différents travaux réalisés en TAL et en linguistique sur les phénomènes dérivationnels et morphologiques qui se rapportent aux noms propres dans plusieurs langues, telles que le français, l'anglais, l'allemand et particulièrement le serbe, et aux relations que les noms propres entretiennent entre eux, que nous avons pu définir ces concepts. Le serbe constitue sans doute une des langues européennes les plus riches au niveau dérivationnel et au niveau morphologique que nous ayons pu étudier au cours de nos travaux.

Nous avons défini deux principaux concepts : le nom propre conceptuel et le prolexème. Autour de ces deux concepts, nous avons défini d'autres concepts, des relations qui ne dépendent pas des langues et des relations qui dépendent de la langue.

Nous avons ensuite regroupé nos différents concepts et relations du domaine des noms propres sous la forme d'une arborescence qui peut se décomposer en quatre niveaux distincts :

- le niveau méta-conceptuel : la typologie et l'existence, que nous présenterons dans le prochain chapitre.
- 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.

Dans les débuts de nos travaux [Tran et al., 2004], nous avions créé notre premier modèle des noms propres en nous basant uniquement sur les langues française, anglaise et allemande.

Nous nous sommes vite rendu compte de la limite du modèle que nous proposions en travaillant avec Duško Vitas et Cvetana Krstev, dans le cadre du projet Égide Pavle-Savic, sur l'intégration du serbe dans notre dictionnaire électronique de noms propres. Ce modèle, basé essentiellement sur ces trois langues, était bien trop simple et ne prenait pas en compte les mécanismes très compliqués de dérivation et de morphologie produits par la langue serbe. Un voyage à Belgrade, à la fois inoubliable et extrêmement enrichissant, nous a permis de remettre en cause notre modèle et de le perfectionner en rajoutant de nouveaux concepts

[Krstev et al., 2005].

3.1 Les deux principaux concepts

3.1.1 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 que ce phénomène se retrouve généralement d'une langue à l'autre. C'est le cas, par exemple, des noms propres *Jean-Paul II* et *Karol Józef Wojtyła* en français, des noms propres *John Paul II* et *Karol Józef Wojtyła* en anglais, etc. qui désignent tous la même personne mais à travers différentes langues.

Or, si le nom propre *Jean-Paul II* apparaît dans les lignes d'un quotidien français, un système de traduction automatique ne devra pas traduire ce nom propre en espagnol par *Karol Józef Wojtyła*, mais devra plutôt le traduire par le nom propre *Juan Pablo II*. Les noms propres *Jean-Paul II* et *Karol Józef Wojtyła* en français correspondent tous les deux à un certain point de vue sur un même et unique référent.

Nous définissons 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] et de [Blanco, 2001].

[Coseriu, 1998] propose un diasystème basé essentiellement sur quatre variétés distinctes (figure 3.1) :

> *On constate, dans chaque état de langue (c'est-à-dire, même en faisant abstraction du développement de cette langue dans le temps) trois types fondamentaux de variétés (à savoir : la variété dans l'espace, la variété relative à la stratification socio-culturelle de la communauté parlante et la variété concernant les occasions, circonstances et finalités de l'emploi de la langue dans le discours).*

DIA-S	DEFINITION
Diachronique	variété dans le temps
Diatopique	variété dans l'espace
Diastratique	variété relative à la stratification socio-culturelle
Diaphasique	variété concernant les finalités de l'emploi

FIG. 3.1 – Le diasystème de Coseriu.

Pour [Blanco, 2001], la marque diasystématique, associée à chaque unité lexicale, constitue une information indispensable qu'un dictionnaire électronique doit contenir. Dans son dictionnaire électronique de l'espagnol de langue générale, il distingue onze étiquettes différentes (figure 3.2).

La variation est un phénomène linguistique qui se retrouve dans toutes les langues et permet à une même entité d'apparaître sous des noms différents.

DIASYSTÈME	MARQUES
diastratique	soutenu, familier, vulgaire
diatopique	américanisme, dialectal
diachronique	vieilli, néologisme
diaintégratif	latinisme, argot
dianormatif	incorrect
diaconnotatif	péjoratif, enfantin
diamédiatique	oral, écrit
diaphasique	formel, informel
diatextextuel	journalistique, administratif, littéraire
diatechnique	langue spécialisée
diafréquence	rare

FIG. 3.2 – Le diasystème de Blanco.

Le nom propre conceptuel nous servira de pivot entre différentes langues. Les noms propres associés à un même nom propre conceptuel seront considérés comme des traductions possibles d'une langue à l'autre.

Le nom propre conceptuel sera représenté dans notre modèle par un numéro d'identité unique (ID), le pivot.

Nous avons décidé de ne pas prendre comme pivot le référent d'un nom propre, car cela risque de poser un certain nombre de problèmes qui ne sont pas évidents à résoudre. Prenons par exemple le nom propre *Paris* et le nom propre *Lutèce*. Supposons que ces deux noms propres correspondent à un même référent. Si nous ajoutons une relation entre le pivot correspondant au nom propre *Paris* et au nom propre *France*, indiquant que *Paris* est la capitale de la *France*. Cela impliquera que *Lutèce* qui est relié au même pivot que *Paris* soit aussi la capitale de la *France*, ce qui n'est pas vrai. Nous serions donc obligés de déplacer cette relation au niveau linguistique. Au lieu d'avoir une relation unique au niveau indépendant des langues, nous aurions autant de relations qu'il y aurait de langues. De plus, la notion de référent est une notion extra-linguistique difficile à définir. Il n'est pas toujours évident d'associer un unique nom propre à un référent[1]. Est-ce que l'on peut dire que le nom propre *France* et le nom propre *Gaule* correspondent à un même référent ? On sait que la *Gaule* ne correspondait pas surfaciquement à la *France* car elle comprenait la France actuelle, le nord de l'Italie et la Belgique. Pouvons-nous aussi dire que la Grèce antique et la Grèce d'aujourd'hui ont le même référent ?

3.1.2 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.

[1] Une solution serait peut-être de considérer que le référent est le synset de nos noms propres conceptuels. Mais contrairement à WordNet, nous ne modélisons pas cette notion de synset.

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 associés à trois types différents. Par contre, dans le cas de toponymes correspondant à la fois à un lieu et une entité administrative (comme par exemple *Paris* qui est à la fois une ville et un département), nous avons décidé de ne pas dupliquer les prolexèmes pour éviter l'abondance d'homographes [Piton and Maurel, 2004]. Cette information sera rajoutée au niveau des expansions classifiantes du prolexème (voir section 3.3.4 page 69).

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*), c'est-à-dire selon [ONU, 1977] :

> *Nom propre employé dans une certaine langue pour désigner un objet géographique situé à l'extérieur du territoire dans lequel cette langue a un statut officiel, et différent dans sa forme du nom propre utilisé dans la ou les langues officielles du territoire où l'objet géographique est situé.*

3.2 Au niveau du prolexème

Les différents alias provenant d'un même nom propre peuvent être considérés comme des synonymes. Ainsi, les phrases (1), (2) et (3) 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.

> *(1) De leur côté, en effet, les* ***États-Unis*** *considèrent que les frontières doivent être le produit d'un accord. (Libération, le mercredi 24 mai 2006)*
> *(2) De leur côté, en effet, les* ***États-Unis d'Amérique*** *considèrent que les frontières doivent être le produit d'un accord.*
> *(3) De leur côté, en effet, les* ***USA*** *considèrent que les frontières doivent être le produit d'un accord.*

Certains dérivés d'un même nom propre peuvent aussi être considérés comme des synonymes de constructions contenant celui-ci, à la condition que le nom propre puisse être reconstruit en utilisant une transformation [Harris, 1968] [Harris, 1976]. Ainsi, le remplacement du dérivé *brésilien* dans la phrase (4) par le groupe prépositionnel contenant le nom propre *Brésil* (5) ne change pas le contenu sémantique de la phrase :

> *(4) Le président* ***brésilien*** *Luiz Inacio Lula da Silva a de fortes chances d'être réélu en octobre pour un second mandat. (Liberation, le mardi 23 mai 2006)*
> *(5) Le président* ***du Brésil*** *Luiz Inacio Lula da Silva a de fortes chances d'être réélu en octobre pour un second mandat.*

Il existe bien sûr des cas ou le dérivé n'est plus le synonyme exact du nom propre dont il dérive, car il n'existe pas de transformation qui les relie :

> *(6) Petite firme familiale de distribution de lait* ***pasteurisé*** *établie dans les environs de Parme dans les années 1960... (Le Monde Diplomatique, février 2004)*[2]
> *(7) * Petite firme familiale de distribution de lait de* ***Louis Pasteur*** *établie dans les environs de Parme dans les années 1960...*
> *(8) ?* Petite firme familiale de distribution de lait* ***traité comme le préconise Louis Pasteur*** *établie dans les environs de Parme dans les années 1960...*

Dans cet exemple, la phrase (6) n'est plus sémantiquement équivalente à la phrase (7) et ni semble-t-il à la phrase (8), puisque dans la phrase (6) le dérivé lexicalisé *pasteurisé* signifie d'après le Petit Larousse :

> *Lait frais pasteurisé : lait frais ayant subi l'opération de pasteurisation par chauffage à une température de 72 à 85 °C pendant 15 à 20 secondes*

Dans ce cas là, nous n'intégrerons pas ce dérivé dans notre dictionnaire électronique, puisque celui-ci relève plutôt d'un dictionnaire de noms communs.

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*.

3.2.1 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 et 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.

Variantes de caractères

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*[3]
- 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 +*
- l'ajout, le remplacement ou la suppression d'une ou plusieurs lettres : *Jean-François Delharpe* ou *Jean-François Delaharpe*
- etc.

[2] extrait du site `http://www.monde-diplomatique.fr/2004/02/RAMONET/10686` le 24/05/06

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

Variantes orthographiques

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 trouvent sur Google.

Abréviations

Les alias peuvent être aussi des abréviations du prolexème, c'est-à-dire :

> *Désignation formée par suppression de mots ou de lettres dans une forme plus longue désignant le même concept [ISO 1087-1 :2000].*

Pour la plupart des noms de célébrités, il sera possible de modéliser la création de leurs alias en utilisant des règles basées sur leur structure interne. Par exemple, le nom propre *François Mitterrand* pourra être associé à la structure interne *Prénom Nom*, qui suit les règles d'aliasisation suivantes :

- *Prénom Nom -> Prénom_ abrégé Nom* : *F. Mitterrand* (a)
- *Prénom Nom -> Nom* : *Mitterrand* (b)

Il existe malheureusement des exceptions parmi les noms de célébrités 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 moins évident d'appliquer la règle (a) à des noms propres non contemporains, comme *Marco Polo* (*?*M. Polo*) ou *Jules César* (*?*J. César*).

Nous pouvons aussi créer des règles pour les noms d'entreprise. A partir de *Sony Corporation*, on applique la règle *Nom Raison_ sociale -> Nom* pour créer l'alias *Sony* et la règle *Nom Raison_ sociale -> Nom Raison_ abrégée* pour obtenir l'alias *Sony Corp.*

De même, pour certains noms de villes françaises, il sera possible d'utiliser les deux règles suivantes :

- *Nom Préposition Hydronyme -> Nom* (c)
- *Nom Préposition Hydronyme -> Nom / Hydronyme* (d)

Ainsi, en appliquant la règle (c) et (d) sur la ville *La Roche sur Yon*, nous obtenons les deux alias suivant : *La Roche* et *La Roche / Yon.*

[Belleil, 1997], dans sa thèse, liste les principales abréviations (figure 3.3) qui peuvent s'appliquer sur les noms de lieu.

Acronymes et sigles

Les alias peuvent être des acronymes (*Sofres* pour *Société française d'enquêtes par sondages*) :

> *Abréviation formée des premières lettres des éléments constituant la forme complète de la désignation, ou des premières syllabes de la forme complète, et prononcée de façon syllabique [ISO 1087-1:2000].*

Ils peuvent aussi être des sigles (*OCDE* pour *Organisation de coopération et de développement économiques*) :

> *Abréviation formée des premières lettres des éléments constituant la forme complète de la désignation et prononcée lettre par lettre [ISO 1087-1 :2000].*

Lemme	Abréviation	Exemple
Saint	St	St-Cloud
Sainte	Ste	Ste-Hermine
Grand	Gd	Gd-Rullecourt
Grands	Gds	Les Gds-Chézeaux
Grande	Gde	La Gde-Verrière
Pont	Pt	Pt-Noyelles
Ponts	Pts	Saint-Denis-les-Pts
Mont	Mt	Dompierre-sur-Mt
Monts	Mts	Notre-Dame-de-Mts
sous	s	Clichy-s-Bois
sur	/	Vitry / Seine
-	espace	Dompierre sur Mt

FIG. 3.3 – Les variantes graphiques.

[Nakos, 1990] définit le sigle de la façon suivante :

> *nom (unité lexicale) formé d'initiales et de syllabes provenant : a) d'un mot (lexie simple), par exemple "T" pour "Titus" à l'époque romaine, b) d'un mot composé (lexie construite à partir d'au moins deux composantes), par exemple "ECG" pour "électrocardiogramme" ou c) d'un groupe de mots (lexie complexe), par exemple "CIA" pour "Central Intelligence Agency". Comme nous le savons déjà, le sigle se prononce soit lettre à lettre, par exemple "ECG" et "CIA", soit comme un mot, par exemple "UNESCO" ou "FNAC", plus difficile à prononcer ; dans ce dernier cas, il devient acronyme ou sigle acronymique.*

La siglaison est un phénomène que l'on retrouve déjà à l'époque romaine. Par exemple, le sigle *INRI.* qui est inscrit sur certaines croix dans nos maisons et dans les églises signifie *Jesus Nazarenus Rex Iudaeorum.* C'est essentiellement à partir de la Seconde Guerre Mondiale, en raison de la création de nombreux organismes et d'entreprises multinationales, que le nombre de sigles s'est rapidement multiplié.

Certains sigles, tels que *UNESCO* (*Organisation des Nations unies pour l'éducation*) ou *OMS* (*Organisation mondiale de la santé*), sont des sigles internationaux. D'autres sont spécifiques à certaines langues, par exemple *OEA* (*Organisation des États américains*) pour le français et *OAS* (*Organization of American States*) pour l'anglais.

Certains sigles sont formés d'initiales provenant de noms composés d'une autre langue, comme par exemple en français *ESA* (*European Space Agency*) pour *Agence spatiale européenne.*

En français, les acronymes s'écrivent théoriquement toujours en majuscule ou avec une majuscule initiale (nouvelle écriture) et les sigles tout en majuscule 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. Par exemple, on trouve deux écritures possibles pour l'acronyme *Association pour l'emploi dans l'industrie et le commerce* : *ASSEDIC* ou *Assédic.* C'est le cas aussi pour l'*Organisation des nations unies* (*ONU*, *Onu* et *O.N.U.*) et l'*Institut National des Langues et Civilisations Orientales* (*INaLCO* et *INALCO*).

Transcriptions

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.

Un même nom propre russe peut posséder en français plusieurs transcriptions différentes. Les transcriptions ne sont pas identiques d'une langue à l'autre. Voici un exemple[4] avec le nom propre russe Владимир Владимирович Маяковский qui se transcrit :

- *Vladimir Vladimirovitch Mayakovski*, *Maïakovski*, ou *Mayakovsky* en français
- *Wladimir Wladimirowitsch Majakowski* en allemand
- *Vladimir Vladimirovich Mayakovsky* en anglais
- *Vladimir Vladimirovitsj Majakovski* en néerlandais
- *Vlagyimir Vlagyimirovics Majakovszkij* en hongrois

Le phénomène de transcription concerne aussi les noms propres chinois qui apparaissent souvent dans les textes journalistiques français. Actuellement, la transcription d'un nom chinois peut se faire suivant trois systèmes de transcription totalement différents : l'EFEO, le Wade-Giles et le pinyin. L'EFEO, système mis au point par l'École française d'Extrême-Orient, est basé sur l'alphabet latin et est très adapté pour les utilisateurs francophones. Le Wade-Giles est un système de transcription notamment utilisé dans la plupart des pays anglo-saxons. Enfin, le pinyin est devenu, depuis 1958, le système de transcription officiel adopté par le gouvernement chinois.

Les noms propres *Pékin*, *Tchang Kaï-chek* et *Mao Tsé-toung* transcrits avec l'EFEO sont beaucoup plus connus des Français que leur forme pinyin *Beijing*, *Jiang Jieshi* et *Mao Zedong*.

En serbe, 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.

Synonymes diastratiques

Certains alias peuvent être le résultat d'une transformation, d'un ajout ou d'une réduction : *Mère Angélique* (pour *Marie Jacqueline Angélique Arnaud*), *le Second Pitt* (pour *William Pitt*), etc.

Ces alias constituent des cas discutables. Nous avons décidé de les placer dans la partie qui dépend de la langue, car leur formation dépend souvent de la culture liée à une langue.

Synonymes diatopiques

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.

Certains pays présentent une grande diversité linguistique, notamment à travers la présence d'une ou plusieurs langues régionales. La France possède de nombreuses langues régionales telles que le catalan, le corse, le breton, le basque, etc. Par exemple, la ville de *Nantes* est appelée par les Bretons *Naoned*, qui est aussi un nom de la langue française. Il existe aussi quelques villes françaises qui possèdent un nom basque, par exemple, *Saint-Jean-de-Luz* qui donne en basque *Donibane Lohitzun*.

[4]Cet exemple à été trouvé sur le site `http://fr.wikipedia.org/wiki/Vladimir_Mayakovski` le 23/05/06

En français, *Naoned* sera considéré comme un synonyme diatopique de *Nantes*. En breton, *Naoned* serait le prolexème associé au prolexème français *Nantes*.

3.2.2 Les dérivés

Les dérivés de notre base sont considérés comme des synonymes, à une transformation près, des prolexèmes dont ils proviennent. Parmi les types de dérivés existant en français, nous avons principalement deux catégories : les noms relationnels et les adjectifs relationnels. [Daille, 1999] définit ainsi les adjectifs relationnels :

> *Les adjectifs relationnels possèdent les propriétés linguistiques suivantes : relation morphologique avec un nom, [...] possibilité d'équivalence avec un complément prépositionnel d'un nom de tête au sein d'un syntagme nominal (acidité sanguine, acidité du sang) [...] et un certain nombre d'autres propriétés comme le fait d'être inusités comme attribut, l'incompatibilité avec le degré, la postposition immédiate après le nom dans une séquence d'adjectifs postposés, etc.*

Les noms relationnels en français débutent normalement par une majuscule (par exemple : *Parisien*, *Marseillais*, etc.). Leur classement au sein de la classe des noms propres est dû à leur identité référentielle fortement connotée. Cependant, cela ne se passe pas de la même manière dans toutes les langues.

Contrairement au nom relationnel, l'adjectif relationnel ne commence pas par une majuscule (par exemple : *parisien*, *marseillais*, etc.). A ce détail près, l'adjectif relationnel reste identique au nom relationnel, sauf quelques rares exceptions. Par exemple, *suisse* est l'adjectif relationnel féminin et *Suissesse* est le nom relationnel féminin provenant du prolexème *Suisse*.

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 effective, ces règles s'appliquent à une forme supplétive. [Adouani, 1993] définit ce mécanisme ainsi :

> *Une paire de mots est ici dite supplétive si ses deux membres sont liés entre eux par une relation dérivationnelle dont la partie sémantique est régulière mais dont la partie formelle est, soit inexistante, soit profondément altérée. Cela peut être des mots complètement différents ou étymologiquement apparentés ou encore dont l'un a une forme latinisée ou réduite.*

Les *Stéphanois* sont les habitants de la ville de *Saint-Étienne*. Étienne fut surnommé le couronné (en grec *stephanos*), car il fut le premier à recevoir la couronne de martyr.

Quand un prolexème appartient à la classe des toponymes, les noms relationnels qu'il engendre sont soit des gentilés (ou noms désignant les habitants d'une ville), soit des ethniques (noms désignant les habitants d'un pays ou d'une région). Selon [Eggert et al., 1998] et [Eggert, 2002], la création des gentilés en français est un phénomène très irrégulier. La figure 3.4 présente la répartition en pourcentage des suffixes apparaissant dans la formation d'un gentilé à partir de l'étude d'un corpus comprenant environ 2 757 gentilés.

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

Suffixes	Pourcentage	Suffixes	Pourcentage
ois	36,1 %	aire	0,3 %
ais	25,1 %	iste	0,29 %
ien	18,7 %	at	0,29 %
éen	3,9 %	ar	0,22 %
in	3,3 %	asque	0,22 %
ain	3,2 %	enc	0,22 %
en	1,3 %	ol	0,22 %
on	0,9 %	ant	0,18 %
ard	0,7 %	and	0,15 %
ot	0,7 %	ate	0,11 %
an	0,6 %	ite	0,07 %
aud	0,4 %	iote	0,07 %
ier	0,4 %	autres	1,9 %
aux	0,4 %		

FIG. 3.4 – Statistique des suffixes de gentilés.

plus riche et complexe sur le plan morphologique, distingue deux catégories de noms relationnels : les noms relationnels féminins et les noms relationnels masculins. A 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.

[Aljovic, 2000], dans sa thèse, explique la création des adjectifs possessifs de la façon suivante :

> *Les possessifs sont des éléments adjectivaux au même titre que les adjectifs lexicaux [...]. Le suffixe possessif ne s'attache qu'aux noms propres, ou aux substantifs utilisés comme des noms propres (à référence unique) [...]. Le suffixe s'attache de préférence aux noms au trait [+humain] [...]. Le suffixe possessif a deux formes -ov et -in. Le premier s'attache aux noms masculins, et le deuxième aux noms féminins.*

A 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 belgradois).
  - Београђанинов (Beogradaninov) : adjectif possessif (la maison d'un Belgradois)
- Београђанка (Beogradanka) : nom relationnel féminin (Belgradoise)
  - београђански (beogradanski) : adjectif relationnel (les habitudes des Belgradoises).
  - Београђанкин (Beogradankin) : adjectif possessif (la maison d'une Belgradoise)

3.3 Les relations

Une fois que les différents concepts du domaine des noms propres ont été identifiés et définis, il s'agit maintenant de rechercher les relations qui lient les noms propres entre eux. Dans cette partie, nous présenterons donc 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. Parmi les relations syntagmatiques, on trouve par exemple la relation de 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 le prochain chapitre, 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 et l'éponymie).

3.3.1 La relation de synonymie

Le célèbre projet WordNet et ensuite le projet européen EuroWordNet ont fait de la relation de synonymie le pilier central sur lequel repose l'ensemble de leur architecture, notamment en la modélisant à travers le concept de "synset". Pour [Miller et al., 1990], la synonymie se définit de la façon suivante :

> *According to one definition (usually attributed to Leibniz) two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made. By that definition, true synonyms are rare, if they exist at all. A weakened version of this definition would make synonymy relative to a context : two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value.*

Pour définir notre relation de synonymie, nous nous sommes inspirés de la définition que propose [Polguère, 2003], présentant la synonymie de la façon suivante :

> *La synonymie, c'est-à-dire l'identité de sens, est la relation lexicale sémantique par excellence :*
>
> *Deux lexies L_1 et L_2 appartenant à la même partie du discours sont des synonymes exacts (ou synonymes absolus) si $(L_1) \equiv (L_2)$.*
> *Ce sont des synonymes approximatifs si $(L_1) \cong (L_2)$. Dans ce dernier cas, il y a soit intersection soit inclusion de sens telle que L_1 et L_2 peuvent être considérées comme ayant une valeur sémantique suffisamment proche pour que l'une puisse être utilisée à la place de l'autre pour exprimer sensiblement la même chose.*
> *Il est essentiel de noter que la synonymie lexicale exacte est rarissime.*

Deux noms propres NP_1 et NP_2 appartenant à une même langue L peuvent être considérés comme étant en relation de synonymie si les trois conditions suivantes sont remplies :

- NP_1 et NP_2 correspondent tous les deux à un point de vue différent sur un même et unique référent, c'est-à-dire qu'il est possible de remplacer NP_1 par NP_2 (ou NP_2 par NP_1) pour exprimer sensiblement la même chose dans un contexte particulier.
- NP_1 ne peut se déduire de NP_2, ni NP_2 de NP_1, en appliquant des règles de morphologie dérivationnelle ou des règles d'aliasisation (création d'alias), c'est-à-dire que NP_1 n'est pas un alias de NP_2 et NP_1 n'est pas un dérivé de NP_2, et vice-versa.

Par exemple en français, le nom propre *Algérie* est en relation de synonymie avec le nom propre *République algérienne démocratique et populaire*. Ces deux noms propres se traduisent respectivement en roumain par *Algeria* et *Republica Democrată Algeria*.

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 3.2.1 page 57). Il nous reste donc à présenter les trois variations restantes : diachronique, diastratique et diaphasique.

Diachronique

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. Voici un exemple[5] de synonymie diachronique dans différentes langues :

- *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
- *Zair* et *Demokratyczna Republika Konga* en polonais
- *Zaire* et *República Democrática del Congo* en espagnol
- *Zaire* et *Demokratische Republik Kongo* en allemand

Certaines transformations permettent aussi un passage entre ces synonymes, par exemple[6] :

> *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)*

Diastratique

La seconde variation est diastratique, c'est-à-dire liée à la classe socio-culturelle. 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 (*Paris* et *Paname*) et les variations savantes, dont voici un exemple dans plusieurs langues :

[5] La plupart de nos exemples ont été extraits du site `http://fr.wikipedia.org/`

[6] `http://www.monde-diplomatique.fr/1997/12/GARRETON/9657` consulté le 8/06/06

– *Jean-Paul II* et *Karol Józef Wojtyła* en français
– *John Paul II* et *Karol Józef Wojtyła* en anglais
– *Johannes Paul II.* et *Karol Józef Wojtyła* en allemand
– *Juan Pablo II* et *Karol Józef Wojtyła* en espagnol
– *Gioan Phaolô II* et *Karol Józef Wojtyła* en vietnamien

De même, 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.

Diaphasique

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

– *Paris* et *Ville lumière* en français
– *Paris* et *City of Light* en anglais
– *Paris* et *Stadt des Lichtes* en allemand
– *Parigi* et *Città delle luci* en italien
– *Paris* et *Cidade das Luzes* en portugais

Dans le cadre d'un article traitant de politique internationale, on considérera comme équivalent :

– *Allemagne* et *République fédérale d'Allemagne* en français
– *Alemania* et *República Federal de Alemania* en espagnol
– *Deutschland* et *Bundesrepublik Deutschland* en allemand
– *Germany* et *Federal Republic of Germany* en anglais
– *Niemcy* et *Republika Federalna Niemiec* en polonais

3.3.2 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 (figure 3.5).

Le projet européen EuroWordNet utilise ces six catégories de méronymie, tandis que WordNet se base uniquement sur trois catégories de méronymie :

– *part of* qui correspond à la catégorie *component/integral object.*
– *substance of* qui correspond à la catégorie *stuff/object.*
– *member of* qui correspond à la catégorie *member/collection.*

Dans le cadre de nos travaux sur les noms propres, seulement deux catégories de méronymes parmi les six catégories proposées par [Winston et al., 1987] s'appliquent :

– lieu/zone
– membre/collection

Type de méronymie	Exemple
component/integral object	handle/cup
member/collection	tree/forest
portion/mass	grain/salt
stuff/object	steel/bike
feature/activity	dating/adolescence
place/area	oasis/desert

FIG. 3.5 – Taxonomie des relations de méronymie.

Il faudrait aussi ajouter la relation de méronymie temporelle [Van Campenhoudt, 1996].

La figure 3.6 donne des exemples de relations de méronymie entre les différentes classes de noms propres.

3.3.3 La relation d'accessibilité

Les premiers travaux du projet Prolex, avant le début de cette thèse, portaient uniquement sur des 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. La relation Chef permettait de préciser si une entité est à la tête d'un groupe d'entité. 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, 1994].

Si nous cherchons, par exemple, le nom commun *fourchette* dans un dictionnaire de langue française, on obtient le résultat suivant :

> ***fourchette***
> *nom féminin*
> *1. Ustensile de table à dents pointues, dont on se sert pour piquer les aliments.*
> *(Le Petit Larousse)*

En cherchant le nom propre *Tours* dans le même dictionnaire, nous pouvons lire :

> ***Tours*** *[tur]*
> *chef-lieu du département d'Indre-et-Loire, sur la Loire, à 225 km au S.-O. de Paris*
> *(Le Petit Larousse)*

En lisant la définition du nom commun *fourchette*, tout lecteur humain sera capable d'imaginer à quoi ressemble une fourchette et de la reconnaître dès qu'il la verra.

Dans le cas du nom propre *Tours*, il est difficile à partir de la définition que nous donne le dictionnaire de nous représenter la ville de Tours. Par contre, nous pouvons situer l'emplacement de cette ville sur une carte géographique. Dans le cas de la définition d'un personnage célèbre, nous aurions des informations sur le métier qu'il exerçait ou exerce, sur certains événements de sa vie, sa date de naissance, ses œuvres s'il s'agit d'un artiste, etc. Nous aurions dans ce cas tout l'historique de ce personnage célèbre, tandis que dans le cas d'un nom commun nous n'aurons pas d'information sur sa date d'apparition, son histoire, etc. Il s'agit d'une description encyclopédique de l'entité plutôt que d'une définition

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 / Supranational	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
...	...

FIG. 3.6 – Relation de méronymie entre les noms propres.

classique. 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 renvoi n'est pas systématique pour tous les noms propres du dictionnaire. Prenons par exemple le cas du nom propre *Aaron* dans le dictionnaire :

> ***Aaron***
> *XIIIe s. av. J.-C.*
> *Frère aîné de Moïse et premier grand prêtre des Hébreux.*
> *(Le Petit Larousse)*

En lisant cet article, nous constatons une référence vers le nom propre *Moïse*. En recherchant l'article sur le nom propre *Moïse*, nous obtenons :

> ***Moïse**, en hébreu **Moshé***
> *XIIIe s. av. J.-C.*
> *Libérateur et législateur d'Israël. La Bible le présente comme le chef charismatique qui a donné aux Hébreux leur patrie, leur religion et leur loi. Né en Égypte, il fut l'âme de la résistance à l'oppression que subissaient les Hébreux : il les fit sortir d'Égypte (l'Exode, vers 1250 av. J.-C.) et unit leurs divers groupes en un même peuple autour du culte de Yahvé. Il posa les éléments de base de la Loi (Torah).*
> *(Le Petit Larousse)*

Nous remarquons en lisant entièrement l'article que le nom propre *Aaron* n'y figure à aucun moment. 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 bijective. Cela montre bien que l'accès au nom propre *Aaron* se fait par l'intermédiaire du nom propre *Moïse* et non l'inverse.

A 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. Nous associons à chaque relation d'accessibilité un repérage (terme emprunté à [Jonasson, 1994]). Celui-ci précise la relation. Nous avons listé les différents 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 de d'*Henri IV*, etc.
- créateur : les auteurs et les œuvres. *Richard Wagner* est le compositeur de *l'Anneau du Nibelung*, *Victor Hugo* est l'auteur de *Ruy Blas*, etc.
- capitale : les toponymes et leurs capitales. *La Rochelle* est le chef-lieu de la *Charente-Maritime*, *Bangkok* est la capitale de la *Thaïlande*, etc.
- dirigeant politique : les hommes politiques et les pays. *Jacques Chirac* est le président de la *République française*, etc.
- 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, etc. *J. Escrivá de Balaguer* est le fondateur de l'*Opus Dei*, *Richelieu* est le fondateur de l'*Académie française*, etc.
- élève : les disciples et leurs maîtres. *Aristote* est le disciple de *Platon*, etc.
- siège : les entreprises, associations ou organisations et le toponyme correspondant au siège social. *Peugeot* est une firme *sochalienne*, etc.

– locataire : les bâtiments officiels et les dirigeants. *Dominique de Villepin* est le locataire de *Matignon*, etc.
– etc.

3.3.4 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*.

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 and Mathieu-Colas, 1998], EuroWordNet, Framenet [Fillmore et al., 2003]).

Prenons l'exemple de l'expansion *écrivain*, on pourra lui associer :

1. la Frame *Text_ creation* (figure 3.7)
2. la grammaire locale de la figure 3.8
3. le concept *writer* dont le numéro ILI est *06438760n* (figure 3.9)
4. la classe d'objet correspondante.

En français, quelques noms propres, tels que les toponymes, se construisent dans une phrase avec des prépositions locatives [Constant, 2003] : *à*, *dans*, *en*, *sur*, etc. Nous avons aussi envisagé d'intégrer ces informations sous forme de grammaires locales.

3.3.5 L'éponymie

La relation d'éponymie se compose de trois relations : l'antonomase, le figement et la terminologie. Contrairement aux autres relations, l'objectif de la relation d'éponymie est d'empêcher une reconnaissance abusive des noms propres dans des textes. Par exemple, la *loi Pasqua* devra être reconnue comme terme mais pas *Pasqua* tout seul.

L'antonomase

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 en tant qu'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 :

une mégère = une femme violente
un bic = un stylo-bille
un kleenex = un mouchoir en papier

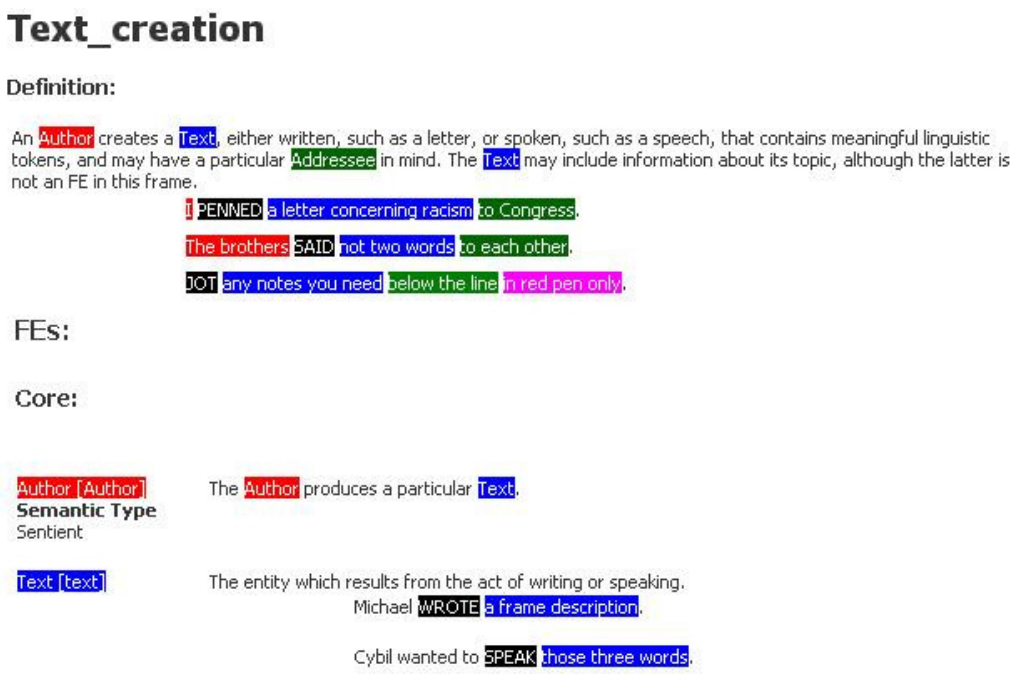

FIG. 3.7 – Extrait de la Frame *Text_ creation.*

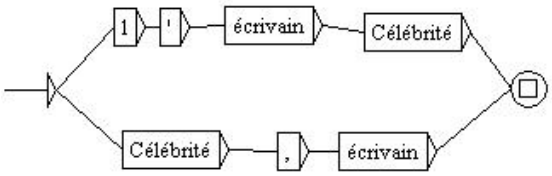

FIG. 3.8 – Exemple d'une grammaire locale.

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.

Le figement

Pour le dictionnaire anglais *Cobuild Dictionary of Idioms*, les idiomes sont définis de la façon suivante :

> *An idiom is a special kind of phrase. It is a group of words which have a different meaning when used together from the one it would have if the meaning of each word were taken individually. [...] Idioms are typically metaphorical : they are effectively metaphors which have become 'fixed' or 'fossilized'.*

entity
object, physical object
living thing, animate thing
organism, being
person, individual, someone
communicator
writer

FIG. 3.9 – Hiérarchie de *writer* dans EuroWordNet.

Nous définissons le figement comme des tournures idiomatiques construites à partir 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. 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'un figement peut varier d'une langue vers une autre :

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

La terminologie

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*. Quelques fois, on sera amené à traduire un nom propre par son dérivé :

le théorème de Pythagore = der pythagoreische Lehrsatz
la maladie de Parkinson = die parkinsonsche Krankheit

D'une langue à l'autre, les noms propres utilisés dans une terminologie peuvent être sujets à des variations :

Comme nous l'avons montré, les noms propres utilisés dans les termes médicaux peuvent être composés, notamment reliés par un trait d'union, et l'ordre de composition des noms propres pour un même terme peut varier d'une langue à l'autre

("maladie de Legg-Perthes-Calvé" ; "Legg-Perthes-Calvé disease" ; "Perthes-Legg-Calvé-Krankheit").
[Bodenreider and Zweigenbaum, 2000b]

Finally, compound proper names found in different translations of ICD-10 sometimes show variation in the order or even in the number of the names (e.g. an alternate term for "relapsing panniculitis" is "Weber-Christian disease" in English, "maladie de Weber-Christian" in French, but "Pfeifer-Weber-Christian-Krankheit" in German).
[Bodenreider and Zweigenbaum, 2000a]

3.4 Représentation sous forme d'un schéma

Nous représentons les différents concepts du domaine des noms propres sous la forme d'une arborescence (figure 3.10) qui peut se décomposer en deux niveaux : un niveau indépendant des langues et un niveau dépendant de la langue.

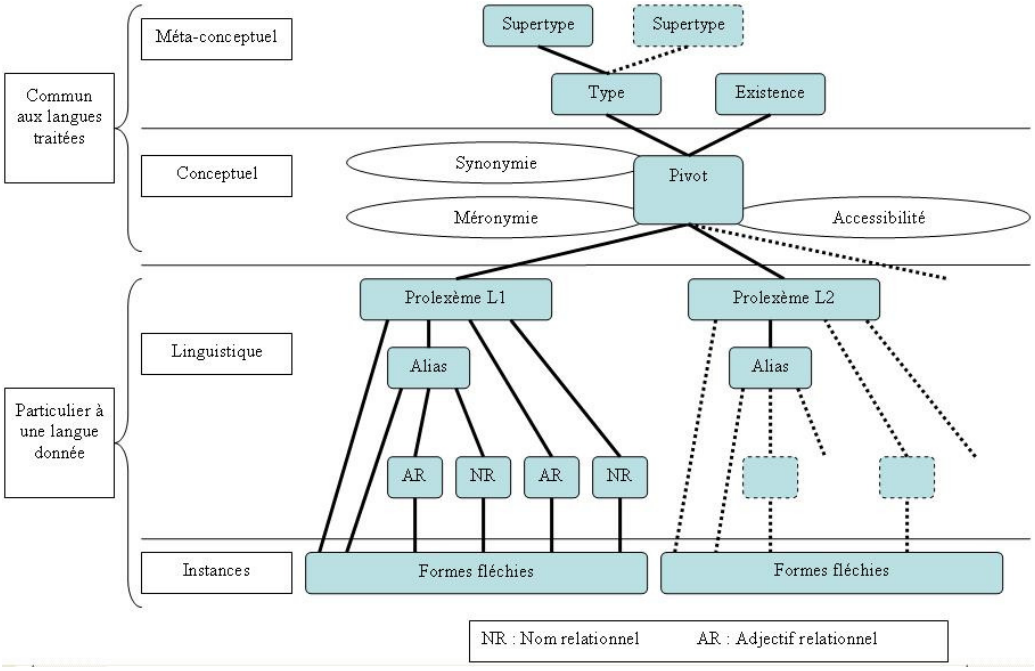

FIG. 3.10 – Les quatre niveaux.

Le niveau indépendant 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 (voir chapitre suivant). 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 dépendant 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 indépendant de la langue à travers un ensemble de noms propres conceptuels. En raison des grandes divergences et de la complexité des mécanismes s'appliquant sur les noms propres,

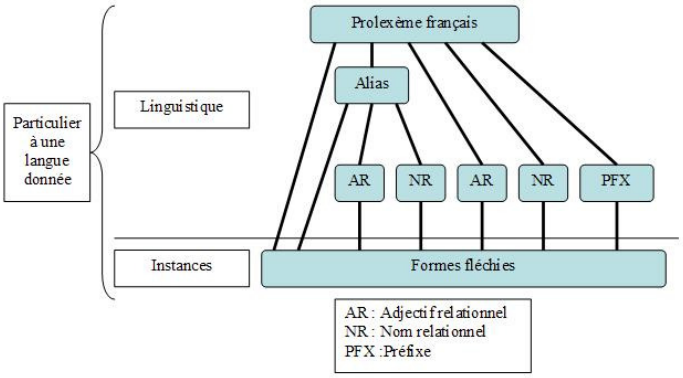

FIG. 3.11 – Le prolexème français.

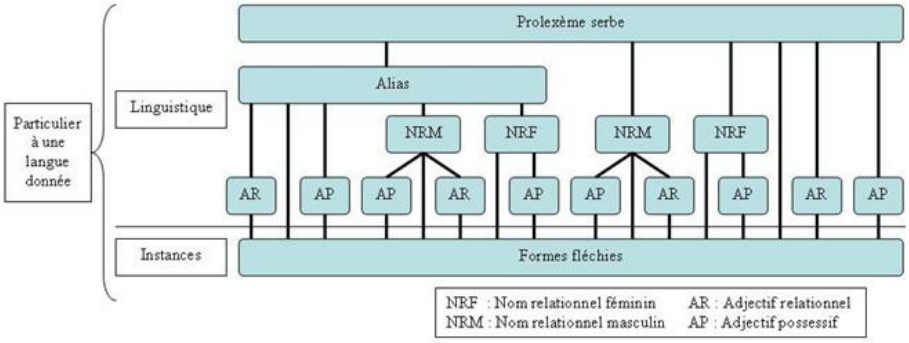

FIG. 3.12 – Le prolexème serbe.

nous ne pouvons définir une arborescence générale qui pourrait s'appliquer pour toutes les langues. Selon la langue, cette arborescence pourra être plus ou moins complexe. La figure 3.11 présente l'arborescence du prolexème français et la figure 3.12 l'arborescence plus complexe du prolexème serbe.

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. L'ensemble de ces formes fléchies, qui correspondent aux mots que l'on retrouve dans un texte, forme ce que [Polguère, 2003] appelle les lexies d'un nom propre :

> *Une lexie, aussi appelée unité lexicale, est un regroupement 1) de mots-formes ou 2) de constructions linguistiques qui ne se distinguent que par la flexion.*

La figure 3.13 montre l'exemple détaillé du nom propre *Belgrade* en cyrilique serbe. La figure 3.14 détaille les prolexèmes français *Suisse* et *Confédération helvétique*.

Conclusion

Dans ce chapitre, nous avons défini nos deux principaux concepts : le nom propre conceptuel et le prolexème. Autour de ces deux concepts, nous avons ajouté d'autres concepts et défini des relations que nous avons représentés sous la forme d'un graphe divisé en deux niveaux (un niveau indépendant de la langue et un niveau dépendant de la langue).

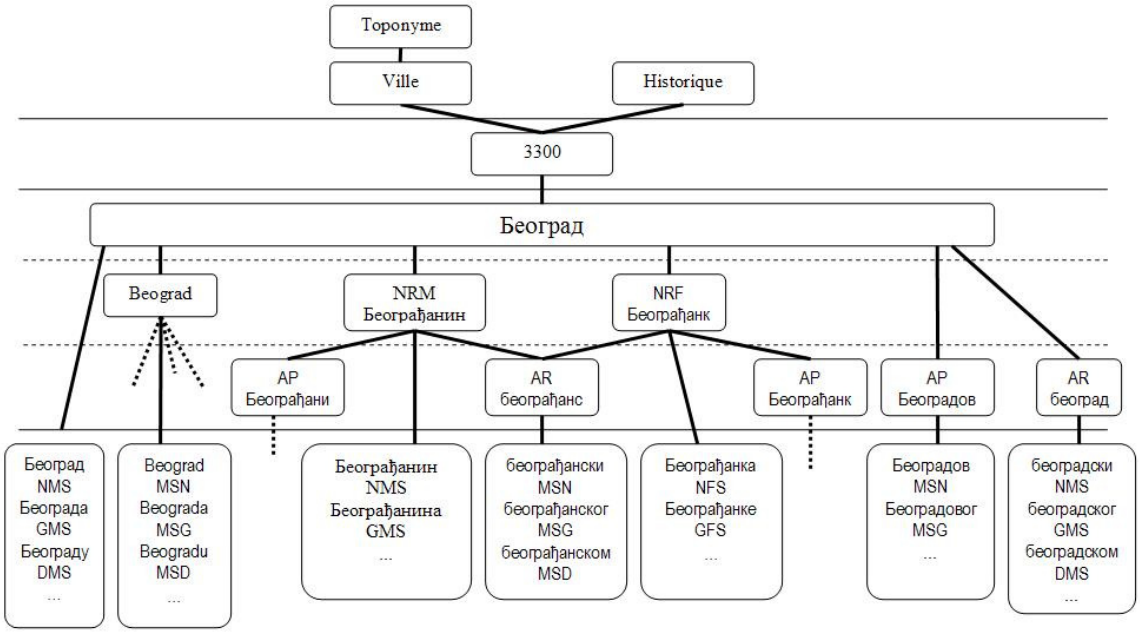

FIG. 3.13 – La partie cyrillique du prolexème serbe *Belgrade*.

Le modèle que nous présentons dans cette partie est donc le résultat de plusieurs modèles successifs. Nous espérons que la comparaison avec d'autres langues permettra de le valider définitivement.

Un essai avec le coréen a été concluant et nous a permis d'intégrer à Prolexbase une première langue non indo-européenne.

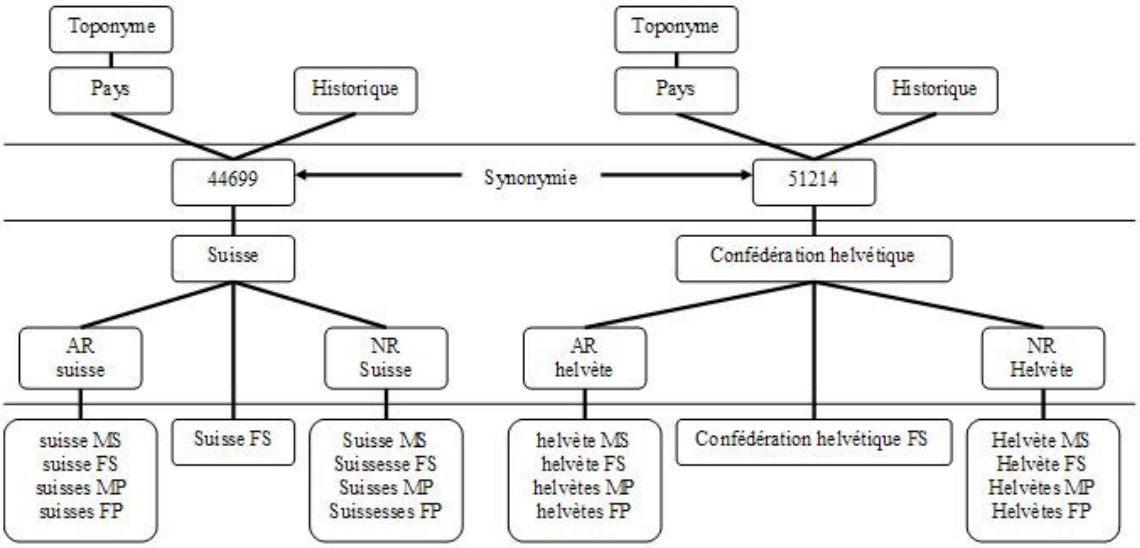

FIG. 3.14 – Les prolexèmes français *Suisse* et *Confédération helvétique*.

Chapitre 4

Ontologie des noms propres

Introduction

Ce chapitre a pour objet de présenter dans un premier temps la notion d'ontologie, notamment à travers les nombreuses définitions trouvées dans la littérature scientifique. Il s'agit ensuite de détailler les caractéristiques d'une ontologie et l'intérêt d'utiliser une approche ontologique pour modéliser l'ensemble des connaissances relatives à un domaine particulier, comme le nôtre. Nous verrons aussi une méthodologie pour construire une ontologie, qui se base sur sept étapes.

Enfin, nous présenterons notre ontologie des noms propres.

4.1 Ontologie

La notion d'ontologie est apparue la première fois il y a environ 2 300 ans sous la Grèce antique en philosophie avec Aristote et même avec Platon. Les ontologies sont, depuis 1990, au cœur de nombreux travaux dans le domaine de l'organisation des connaissances. En Intelligence Artificielle, en Ingénierie des Connaissances, dans le Web Sémantique, dans le Traitement Automatique des Langues, etc., les approches ontologiques connaissent beaucoup de succès et apportent des solutions novatrices. Cela s'explique par le besoin et la recherche d'une modélisation du monde et du sens des mots qui soit accessible aussi bien par des humains que par des agents logiciels.

4.1.1 Définition d'une ontologie

Il n'est pas évident de définir précisément ce qu'est une ontologie. Il existe bien sûr de nombreuses définitions de la notion d'ontologie, mais nous allons présenter seulement quelques définitions que nous avons trouvées dans le domaine de la recherche en informatique et qui nous ont paru intéressantes. L'une d'entre elles, dans le domaine de l'intelligence artificielle, citée fréquemment, revient à [Gruber, 1993] :

Définition 1 *An ontology is a formal, explicit, specification of a shared conceptualization. (Une ontologie est une spécification formelle explicite d'une conceptualisation.)*

Construire une ontologie consiste dans un premier temps à mener un travail de conceptualisation, qui nécessite d'identifier les concepts du domaine à modéliser en se basant sur l'étude de corpus relatif à ce domaine. De nombreux autres travaux se sont basés sur cette définition. [Charlet et al., 2003] donnent la définition suivante :

Définition 2 *Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts —e.g. entités, attributs, processus —, leurs définitions et leurs interrelations. On appelle cela une conceptualisation.*
[...]
Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification.
[...]
Une ontologie est une spécification rendant partiellement compte d'une conceptualisation.

Pour [Roche, 2005] une ontologie possède les caractéristiques suivantes :

Définition 3 *Définie pour un objectif donné et un domaine particulier, une ontologie est pour l'ingénierie des connaissances une représentation d'une modélisation d'un domaine partagée par une communauté d'acteurs. Objet informatique défini à l'aide d'un formalisme de représentation, elle se compose principalement d'un ensemble de concepts définis en compréhension, de relations et de propriétés logiques.*

Selon ces différentes définitions, toute ontologie doit au moins posséder les caractéristiques suivantes :

- des concepts : un concept peut être un objet concret ou abstrait, qui apparaît dans le domaine à modéliser.
- des propriétés : il s'agit de caractéristiques qui permettent de décrire plus précisément les concepts.
- des relations : les relations permettent relier les différents concepts de l'ontologie entre eux. Il existe de nombreuses relations : la relation de méronymie, la relation de synonymie, la relation de subsomption (*is-a*), etc.

Ces différentes définitions nous renseignent sur la notion d'ontologie dans un contexte informatique, mais elles ne nous donnent aucune méthodologie pour construire une ontologie relative à un domaine spécifique.

4.1.2 Méthodologie de construction d'ontologie

Il existe évidemment de nombreuses méthodologies qui permettent de développer des ontologies, mais aucune d'entre elles n'est admise ou reconnue par l'ensemble de la communauté scientifique.

Certaines méthodes relèvent parfois plus de l'intuition que de la rigueur scientifique. La plupart admettent qu'il est nécessaire d'identifier dans un premier temps les concepts et les relations. Mais on constate que, selon la méthodologie utilisée pour modéliser un même domaine, le résultat obtenu ne sera pas forcément le même [Mizoguchi, 2005], en raison des nombreux choix et critères que chaque ontologiste est amené à prendre au cours de cette première phase. La plupart des méthodes ne décrivent pas de manière précise les décisions à prendre ou les règles qu'il faut appliquer durant le processus de conceptualisation.

Nous allons présenter une méthodologie qui nous a semblé intéressante, celle proposée par [Noy and McGuinness, 2003].

Méthodologie de Noy et McGuinness

Selon [Noy and McGuinness, 2003], il est nécessaire au cours de la conception de l'ontologie de toujours se rappeler, particulièrement lorsque l'on est confronté à un problème, les trois règles de base suivantes :

1. Il n'y a pas qu'une seule façon correcte pour modéliser un domaine - il y a toujours des alternatives viables. La meilleure solution dépend presque toujours de l'application que vous voulez mettre en place et des évolutions que vous anticipez.
2. Le développement d'une ontologie est nécessairement un processus itératif.
3. Les concepts dans une ontologie doivent être très proches des objets (physiques ou logiques) et des relations dans votre domaine d'intérêt. Fort probablement, ce sont des noms (objets) ou verbes (relations) dans des phrases qui décrivent votre domaine.

Leur méthodologie de construction d'une ontologie relative à un domaine particulier repose sur une série de sept étapes.

Dans la première étape, il faut commencer par faire une description précise et détaillée du domaine sur lequel on va travailler afin de mieux percevoir ses limites, c'est-à-dire où il commence et où il s'arrête. Il faut aussi déterminer les applications que l'on souhaite faire de cette ontologie.

La deuxième étape consiste à rechercher dans des bibliothèques d'ontologies mises à disposition, par exemple sur Internet, s'il n'existe pas déjà une ontologie qui correspondrait à ses besoins. Si l'on n'a pas eu la chance de trouver son bonheur dans les travaux existants, il va falloir passer à l'étape suivante.

Dans la troisième étape, il faut lister les différents mots importants du domaine. Il ne faut surtout pas s'inquiéter si cette liste est extrêmement longue.

Dans la quatrième étape, on définit les différentes classes et on établit une hiérarchisation entre elles. On peut soit commencer par définir le concept le plus général pour finir par les concepts les plus spécialisés (méthode descendante, en anglais *top down*), soit appliquer la méthode inverse ascendante (*bottom up*), soit choisir une méthode mixte qui combine les deux précédentes.

L'étape cinq permet de décrire les classes plus précisément, en cherchant pour chacune ses propriétés ou attributs.

L'étape six consiste à définir la cardinalité et le type (chaîne, bouléen, etc.) associés à chaque attribut.

La dernière étape correspond au moment ou l'on pourra créer des instances (ou individus) de l'ontologie.

4.2 Typologie des noms propres

Dans cette partie, nous allons nous intéresser au domaine de la typologie des noms propres. Il s'agit maintenant de définir les différents concepts de notre typologie et les relations entre ces concepts sous la forme d'une ontologie.

Nous allons appliquer les quatre premières étapes[1] de la méthodologie de Noy et McGuinness. Pour décrire notre domaine, nous nous sommes basés sur les différentes typologies, utilisées dans le domaine de la linguistique et celles qui ont conduit à des systèmes de reconnaissance de noms propres, que nous avons décrits en détail au cours du premier chapitre. A partir de ces différents travaux, nous avons ensuite établi une liste de types de noms propres. Nous avons appliqué la méthode descendante pour définir et hiérarchiser nos différents concepts, que nous allons présenter dans cette partie. Ces différents concepts entretiennent entre eux une relation d'hyperonymie.

[1]Les étapes cinq et six seront présentées au chapitre 5 sur l'implémentation de notre modèle.

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

4.2.1 Les quatre premiers supertypes

Situés juste en dessous du concept de nom propre, les quatre premiers supertypes classent les noms propres suivant des traits syntaxo-sémantiques assez généraux. Ces traits peuvent facilement 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.

Dans notre ontologie, nous avons distingué :

– les anthroponymes : trait humain
– les ergonymes : trait inanimé
– les pragmonymes : trait événement
– les toponymes : trait locatif

La figure 4.1 montre la représentation des différents concepts de supertype à l'aide du logiciel *Protégé 3.1*[2], permettant de créer des ontologies.

FIG. 4.1 – Les supertypes.

Les anthroponymes

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 [Gross, 1995] : les anthroponymes individuels (*Lassie*, *George Orwell*, etc.) et les anthroponymes collectifs (*Mérovingiens*, *Organisation mondiale de la santé*, etc.). [Dubois, 1973], dans le *Dictionnaire de linguistique*, distingue les noms animés non humains, c'est-à-dire les animaux, et les noms animés, sous-catégorie dans laquelle il classe le trait humain. Cette distinction se fera au niveau des types célébrité et pseudo-anthroponyme (voir section 4.2.2).

[2] `http://protege.stanford.edu/`

Les toponymes

[Lepesant, 2000] définit les toponymes ainsi :

> *Les noms locatifs constituent une catégorie de noms d'objets dimensionnels, tels que leurs méronymes d'espace ont pour hyperonyme le mot lieu.*

Nous avons rassemblé sous le concept de toponyme tous les noms de lieu au sens général. Les toponymes regroupent diverses entités qui possèdent chacune une taille 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 environ quelques millions de galaxies. Il est possible de diviser les toponymes en deux classes différentes : les toponymes naturels et les toponymes bâtis par les hommes.

Les systèmes de reconnaissance automatique de noms propres arrivent à extraire les toponymes dans un texte journalistique [Friburger, 2002], car la plupart du temps, ils apparaissent dans ces textes, accompagnés de preuve externe (*la* ***ville de*** *Tours*) ou de preuve interne (*le* ***Mont*** *Blanc*) [MacDonald, 1996].

Les ergonymes

Ergonyme (du grec *ergon* : travail, force) est un mot emprunté à [Bauer, 1985] :

> *Noms des installations créées par l'homme servant à la production, [...] noms de produits créés par et pour l'homme.*

Sous le type ergonyme, on peut retrouver des noms propres qui se rattachent soit au trait sémantique inanimé concret (*Coca-Cola*), soit au trait inanimé abstrait (*Alice au pays des merveilles*). Nous distinguons dans cette catégorie les ergonymes à caractère économique de ceux à caractère artistique.

Les pragmonymes

Les pragmonymes peuvent être définis comme des noms d'événements (comme *le 14 juillet*) ou de catastrophes naturelles (comme par exemple *Katrina*) ou non (comme par exemple *Tchernobyl*).

4.2.2 Type

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 29 types que nous allons présenter dans cette partie. La figure 4.2 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 que nous appellerons Groupement et qui rassemble les anthroponymes collectifs correspondant à une association ou à une institution (politique, religieuse, culturelle, nationale, internationale, etc.). Ce supertype contient les types association, ensemble, entreprise, institution et organisation.
- un supertype que nous appellerons Territoire car il n'est parfois pas évident de faire une distinction entre les pays (au sens états indépendants) et les régions incluses ou non dans les pays. Ce supertype contient les types pays, région et supranational.

En cas de polysémie (voir section 3.1.2 page 55), comme par exemple pour le nom propre *Michelin* qui correspond à la fois à une célébrité et à une entreprise et pour le nom propre *Tempelhof* qui correspond à la fois à un faubourg de Berlin et à un de ses aéroports, nous avons décidé de créer deux noms propres conceptuels différents. Nous associerons à chacun de ces noms propres un unique type.

Rappelons aussi que les homonymes correspondent de même à des noms propres conceptuels différents même s'ils ont le même type. Par exemple, le nom propre *Vienne*, capitale de l'Autriche, sera lié au type ville, ses homonymes correspondant à une ville d'Isère et de Poitou-Charentes le type ville.

La figure 4.3 présente la hiérarchie des types correspondant à la relation d'hyperonymie primaire (voir section 4.3.2)

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 IIIe Reich*, *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.
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.

FIG. 4.2 – Les types

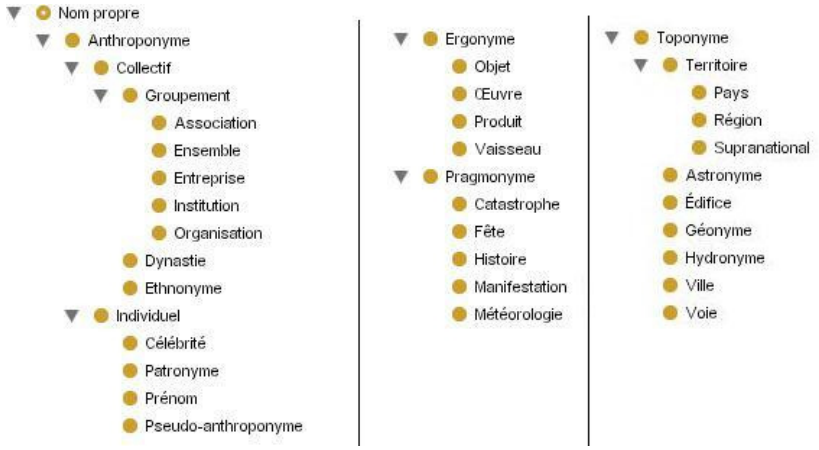

FIG. 4.3 – La hiérarchie des types.

Association

Le terme association désigne un groupe comprenant plusieurs personnes visant un but commun. Dans le type association nous avons regroupé les associations nationales, à caractère social et à but non lucratif, les partis politiques et les syndicats nationaux.

Astronyme

Les astronymes regroupent uniquement les noms propres relevant du domaine de l'astronomie. Il s'agit des noms que l'on attribue aux objets célestes, c'est-à-dire les planètes, les galaxies, les étoiles, les comètes, les constellations, etc.

Catastrophe

Les catastrophes rassemblent les noms de désastres ou tragédies entraînant le plus souvent la mort ou la destruction qui sont soit d'origine naturelle (les catastrophes climatiques comme les cyclones, ouragans, tempêtes, etc., les catastrophes sismiques, les éruptions volcaniques, etc.), soit d'origine humaine (les catastrophes industrielles, etc.).

Célébrité

Ce type regroupe les humains célèbres. Les célébrités constituent sans aucun doute une classe très vaste par rapport aux autres classes de notre typologie des noms propres et en perpétuelle expansion. Nous avons aussi regroupé dans cette classe les pseudonymes ou noms d'emprunt généralement utilisés par des artistes.

Les noms de célébrités apparaissent dans les textes sous des formes très diverses : un nom, un prénom, un prénom et un nom, etc.

Dynastie

La classe des dynasties correspond aux humains collectifs. La majorité des éléments de cette classe est constituée de noms de familles royales de divers pays ou empires, regroupant une succession de monarques qui ont marqué l'histoire. Nous avons aussi étendu cette classe aux noms de familles ayant un lien avec le pouvoir politique ou économique.

Édifice

Le type édifice correspond à des constructions humaines de toute sorte, telles que les bâtiments historiques ou officiels, les monuments, les châteaux, les ponts, les parcs, les musées, les bibliothèques, les théâtres, les bâtiments religieux (églises, basiliques, mosquées, temples, etc.), les aéroports, les prisons, les stades, les hôpitaux, les murs, etc.

Ensemble

Les ensembles sont essentiellement formés de noms de groupe de personnes relevant soit du domaine artistique, soit du domaine sportif.

Entreprise

Le type entreprise rassemble les sociétés industrielles, financières ou commerciales, qui peuvent être nationales ou multinationales.

Ethnonyme

Les ethnonymes sont des noms de peuples. Comme pour les gentilés, leur statut de nom propre est parfois contesté même s'ils prennent une majuscule initiale.

Fête

Le type fête comprend les noms des événements festifs et cycliques, qui ont pour but de rappeler certaines traditions ou faits historiques marquants.

Géonyme

La classe des géonymes est formée de noms donnés aux espaces géographiques naturels, tels que les déserts, les montagnes, les massifs montagneux, les glaciers, les plaines, les gouffres, les plateaux, les vallées, les volcans, les canyons, etc.

Histoire

La classe histoire comprend les événements qui ont marqué la mémoire des hommes. Il s'agit de traités ou accords signés entre différents pays, de batailles, de périodes historiques, de manifestations sociales, de révoltes, de crises, les ères, etc.

Hydronyme

Les hydronymes renvoient à des noms d'étendue d'eau. Il peut s'agir par exemple de rivières, de fleuves, d'étangs, de marais, de lacs, de mers, d'océans, de courants marins, de canaux, de sources, etc.

Institution

Le type institution regroupe :
- les instituts d'enseignement et de recherche
- les instituts religieux
- les fondations
- les instituts politiques

– les juridictions
– les instituts militaires
– les administrations
– etc.

Les noms d'institution sont généralement traduits d'une langue à l'autre.

Manifestation

Les manifestations regroupent toutes sortes d'activités ou d'événements sportifs ou culturels.

Météorologie

Il s'agit d'évènements météorologiques naturels et récurrents, tels que les vents, les phénomènes climatiques, etc.

Objet

Cette classe regroupe uniquement les noms d'objet qui sortent souvent de légendes, de la littérature ou qui relèvent du domaine religieux.

Œuvre

Nous avons regroupé dans cette classe toutes les formes d'œuvres artistiques. Il peut s'agir de sculptures, de livres, de tableaux de grands maîtres, de films, de pièces de théâtre ou d'opéra, de partitions de musique, etc.

Organisation

Le type organisation comprend seulement les organismes à caractère international et qui ne se rattachent pas à un gouvernement en particulier. Si l'organisation se rattache à un gouvernement, nous la classerons parmi les institutions.

Patronyme

Il s'agit de noms de famille. Les patronymes se traduisent rarement d'une langue à l'autre. Lorsqu'un nom de famille appartient à un système d'écriture différent, il est sujet à des translittérations ou à des transcriptions.

Pays

Nous avons rassemblé dans le type pays les noms d'états indépendants, des royaumes ou empires qui sont apparus au fil de l'histoire. Il peut aussi s'agir de noms de pays fictifs.

Prénom

Il s'agit de prénoms. Un même prénom peut se retrouver d'une langue à l'autre sous des formes différentes. C'est le cas du prénom français *Marie* qui donne :

– *Mary* en anglais
– *Maria* en espagnol
– etc.

Les prénoms évoluent d'une époque à l'autre (*Johan*, *Jehan*, *Jean*, etc.).

Produit

Il s'agit de noms de produits ou de marques qui ont été développés uniquement dans un but commercial. On peut trouver des noms de voitures, d'avions, de produits de consommation, de vêtements, d'outils, etc.

Pseudo-anthroponyme

Les pseudo-anthroponymes regroupent tous les noms propres qui peuvent être classés parmi les êtres vivants ou considérés comme tels et qui ne font pas partie de la catégorie des êtres humains. Le profil des entités que l'on retrouve dans cette classe est extrêmement varié. Il peut s'agir de noms donnés aux animaux (zoonymes), aux robots, aux machines, aux êtres venus d'une autre planète ou d'une autre dimension, etc.

Région

Les régions correspondent à un découpage, parfois administratif, d'un pays en plusieurs espaces géographiques de taille variable. Il peut s'agir par exemple de comté ou Land pour l'Allemagne, d'état pour les États-Unis, de canton pour la Suisse, de province pour le Canada, la Belgique et la Chine, etc.

Supranational

Le concept intitulé supranational est défini comme un regroupement de différents pays ou contenant des parties de différents pays.

Vaisseau

Cette catégorie regroupe les véhicules pouvant circuler sur l'eau (paquebots, navires de guerre, etc.) ou dans l'espace (fusées, stations spatiales, etc.). D'autres noms de véhicules peuvent rentrer dans cette catégorie. C'est le cas par exemple du nom propre *Batmobile*.

Ville

Nous avons regroupé sous le type ville les noms de villes et les noms de quartiers. Les quartiers correspondent à un découpage d'une ville, ayant parfois une densité de population bien supérieure à certaines villes. De plus, certains quartiers dans les grandes métropoles étaient autrefois des villes.

Voie

Le concept de voie est principalement formé de noms de rues, de places, de routes, d'autoroutes, etc.

4.3 Ontologie des noms propres

Les relations de notre ontologie des noms propres comprennent des relations qui ne dépendent pas de la langue vues au chapitre 3 et de la relation d'hyperonymie (voir section 4.3.2). Les concepts de notre ontologie comprennent les types, les supertypes, l'existence (voir section 4.3.1) et le nom propre conceptuel.

4.3.1 Existence

Le concept d'existence permet de préciser le domaine d'appartenance d'un nom propre. La majorité des noms propres appartiennent au domaine historique (*Mozart*, *le Danube*, *Paris*, etc.), qui se définit dans le Larousse 2004 comme :

> ***Historique** : [...] dont l'existence est considérée comme objectivement établie.*

D'autres relèvent plutôt du domaine de la croyance (*Zeus*, *Adam*, etc.) ou du domaine de la fiction (*Tintin*, *Utopie*, *Atlantis*, etc.).

La distinction entre des noms propres historiques et les autres s'avère utile pour la traduction, car ces derniers se traduisent d'une langue à l'autre. C'est le cas, par exemple, de Blanche-Neige qui devient :

- *Sneeuwwitje* en néerlandais
- *Biancaneve* en italien
- *Schneewitchen* en allemand
- *Snövit* en suèdois
- etc.

4.3.2 La relation d'hyperonymie

[Polguère, 2003] définit la relation d'hyperonymie de la façon suivante :

> *La lexie L_{hyper} est un hyperonyme de la lexie L_{hypo} si*
> *– le sens (L_{hyper}) est inclus dans le sens (L_{hypo})*
> *– et si (L_{hypo}) peut être considéré comme un cas particulier de (L_{hyper}).*
> *La lexie L_{hypo}, quant à elle, est appelée hyponyme de L_{hyper}.*

Hyperonyme secondaire	**Type**
Anthroponyme	Pays Région Supranational
Anthroponyme Ergonyme	Ville
Ergonyme	Édifice Voie
	Fête Histoire Manifestation
Ergonyme Toponyme	Association Ensemble Entreprise Institution Organisation
Toponyme	Vaisseau

FIG. 4.4 – Hyperonymie secondaire.

Les types et les supertypes de notre ontologie sont reliés par une relation d'hyperonymie (figure 4.3), que nous appellerons relation d'hyperonymie primaire.

Certains types peuvent être en relation d'hyperonymie secondaire avec d'autres supertypes (figure 4.4). Par exemple, les types association et organisation sont à la fois en relation d'hyperonymie primaire avec le supertype anthroponyme collectif et en relation d'hyperonymie secondaire avec les supertypes ergonyme et toponyme.

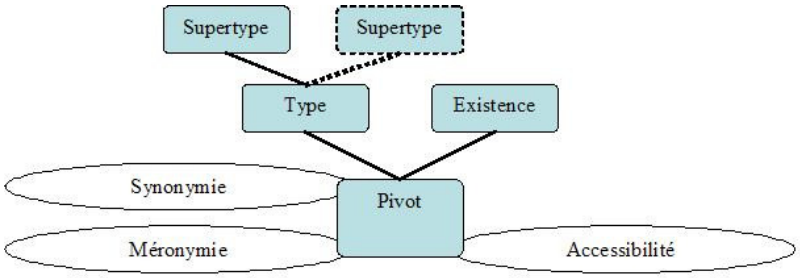

FIG. 4.5 – Ontologie des noms propres.

Chaque nom propre conceptuel (ou pivot) sera en relation d'hyperonymie avec un type et une existence (figure 4.5).

Conclusion

Nous avons présenté dans ce chapitre la modélisation de notre typologie et de notre partie qui ne dépend pas de la langue sous la forme d'une ontologie, composée de plusieurs concepts et relations d'hyperonymie permettant de hiérarchiser ces différents concepts.

Pour créer cette typologie des noms propres, nous nous sommes beaucoup inspiré des travaux de Thierry Grass, avec qui nous avons eu de nombreuses occasions de travailler durant ces trois années.

Les chapitres 3 et 4 nous permettent donc d'obtenir la définition complète de notre ontologie des noms propres (voir figure 4.5). Celle-ci comprend quarante concepts (le nom propre conceptuel, sept supertypes, vingt-neuf types et trois existences) et quatre relations (hyperonymie, synonymie, méronymie et accessibilité).

Maintenant que l'architecture de notre modèle a été définie, il s'agit d'implémenter ce modèle sur machine afin de pouvoir entrer des données. Les parties suivantes seront donc applicatives.

Troisième partie

Implémentation de Prolexbase

Chapitre 5

La base de données

Introduction

Ce chapitre est destiné à présenter l'implémentation des différents concepts et relations du domaine des noms propres sous la forme d'une base de données relationnelle, que nous avons appelée Prolexbase.

Pour créer notre base de données des noms propres, nous avons utilisé la méthode Merise. Dans la première partie, nous allons présenter quelques notions de base sur cette méthode. Ensuite, nous décrirons le modèle conceptuel de données et le modèle logique de données que nous avons mis en place pour les noms propres.

5.1 La méthode Merise

Développée en France en 1978, la méthode Merise (Méthode d'Étude et de Réalisation Informatique pour les Systèmes d'Entreprise) [Matheron, 1998] propose une démarche pour analyser et concevoir un système d'information. Dans cette partie, nous allons présenter brièvement deux étapes de cette méthode : le modèle conceptuel de données et le modèle logique de données.

5.1.1 Le modèle conceptuel de données

Le modèle conceptuel de données (MCD) permet de décrire les objets de la réalité et les dépendances ou associations entre ces objets. Un MCD aboutit à la création d'un schéma d'entité/association (E/A).

Les entités

Une entité est définie comme un objet concret ou abstrait du monde réel. Dans le modèle E/A, on représente une entité sous la forme d'un rectangle (figure 5.1) dans lequel on inscrit le nom de l'entité.

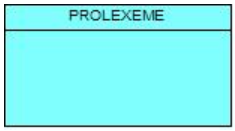

Fig. 5.1 – Représentation d'une entité.

Chaque entité peut posséder un ou plusieurs attributs, dont on devra préciser le type (Date, Entier, Booléen, Texte, etc.). Pour pouvoir identifier chaque occurrence d'une entité de manière unique, il faudra obligatoirement désigner parmi ses différents attributs un attribut ou un ensemble d'attributs qui jouera le rôle d'identifiant ou de clé primaire. Il arrive souvent que l'on rajoute un attribut fictif (un numéro) qui servira de clé primaire. Nous avons associé à chaque identifiant le type ID dans nos schémas E/A (figure 5.2).

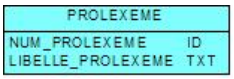

FIG. 5.2 – Représentation des attributs.

Les associations

Les associations sont des liens qui unissent les entités du modèle. Elles apparaissent dans un schéma E/A sous la forme d'un ovale (figure 5.3). On associe à chaque entité d'une association une cardinalité qui précise si une entité peut participer dans l'association zéro, une ou plusieurs fois.

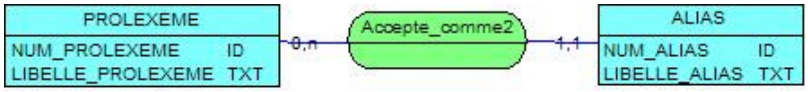

FIG. 5.3 – Représentation d'une association.

Dans la figure 5.3, les entités *PROLEXEME* et *ALIAS* sont reliées par l'association *Accepte_ comme2*. La cardinalité (0,n) indique qu'un prolexème accepte au minimum zéro alias et au maximum plusieurs alias. La cardinalité (1,1) précise qu'un alias correspond à un seul et unique prolexème.

5.1.2 Le modèle relationnel de données

Le modèle relationnel de données (MLD) correspond à une traduction des entités et des associations du MCD sous la forme de relations. Les principales règles de passage d'un MCD vers MLD sont les suivantes :

- Règle 1 : une entité du MCD se transforme en relation. Ses propriétés deviennent des attributs. La clé primaire de la relation sera représentée par l'identifiant.
- Règle 2 : soit R une association de type un-à-plusieurs reliant deux entités E1 et E2 (une occurrence de E1 peut être en relation avec au maximum une occurrence de E2 et une occurrence de E2 peut être en relation avec plusieurs occurrences de E1). R ne devient pas une relation. L'identifiant de E2 et les éventuelles propriétés de R sont rajoutés dans la relation E1.
- Règle 3 : soit R une association de type plusieurs-à-plusieurs reliant deux entités E1 et E2 (plusieurs occurrences de E1 peuvent être en relation avec plusieurs occurrences de E2 et plusieurs occurrences de E2 peuvent être en relation avec plusieurs occurrences de E1). R devient une relation et ses éventuelles propriétés seront des attributs. Les identifiants de E1 et E2 deviennent les clés primaires de R.

En appliquant ces règles sur la figure 5.3, nous obtenons le modèle relationnel suivant :

PROLEXEME (NUM_PROLEXEME, LIBELLE_PROLEXEME)
ALIAS (NUM_ALIAS, LIBELLE_ALIAS, NUM_PROLEXEME)

qu'il est possible de représenter sous la forme d'un schéma relationnel (figure 5.4).

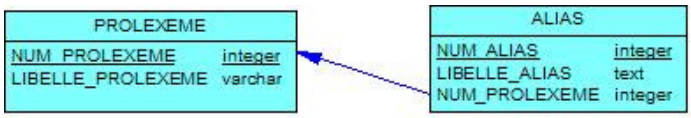

FIG. 5.4 – Schéma relationnel.

5.2 Modèle conceptuel de données

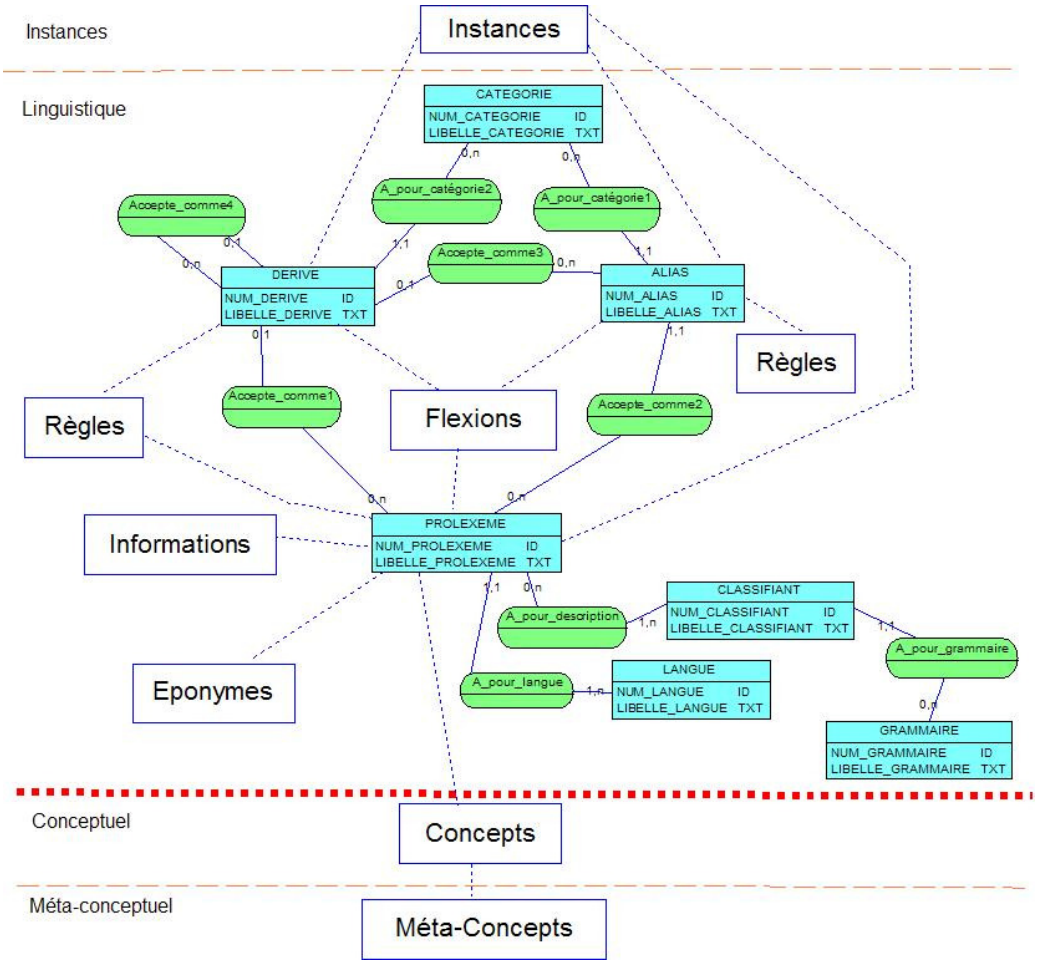

FIG. 5.5 – Le modèle conceptuel de données.

Nous avons établi notre MCD des noms propres à partir des différents concepts et relations définis dans le chapitre 2 et le chapitre 3. La figure 5.5 présente une version simplifiée de notre MCD (voir figure A.3 de l'annexe A page 151 pour un MCD complet).

Notre MCD peut être regroupé en quatre niveaux (méta-conceptuel, conceptuel, linguistique et instances) et comprend au total 28 entités et 41 associations.

Nous avons créé une entité pour chaque concept du domaine des noms propres (prolexème, alias, dérivé, etc.). L'entité *DERIVE* permet de stocker les dérivés de prolexème, d'alias ou d'autres dérivés (dans le cas du serbe). Nous avons associé à chaque alias, à travers la relation *A_pour_catégorie1*, une catégorie qui précise s'il s'agit d'une variante de caractères, d'une abréviation, d'acronymes ou sigles, d'une transcription, d'un synonyme diastratique ou d'un synonyme diatopique. Nous avons aussi associé à chaque dérivé, à travers la relation *A_pour_catégorie2*, une catégorie qui indique s'il s'agit d'un nom relationnel, d'un préfixe, d'un adjectif relationnel ou possessif. Les expansions classifiantes sont stockées dans l'entité *CLASSIFIANT* et chaque classifiant sera en relation avec une description (entité *GRAMMAIRE*) sous forme de grammaire locale, lien vers EuroWordNet ou Framenet (voir section 3.3.4). La relation *A_pour_langue* permet d'associer à chaque prolexème une langue.

5.2.1 Le niveau conceptuel

La partie conceptuelle (figure 5.6) est formée de quatre entités et cinq relations. L'entité *PIVOT* permet de stocker les noms propres conceptuels. La relation *Concept* associe à chaque nom propre conceptuel un ou plusieurs prolexèmes. On retrouve dans ce niveau la relation de méronymie, de synonymie et d'accessibilité.

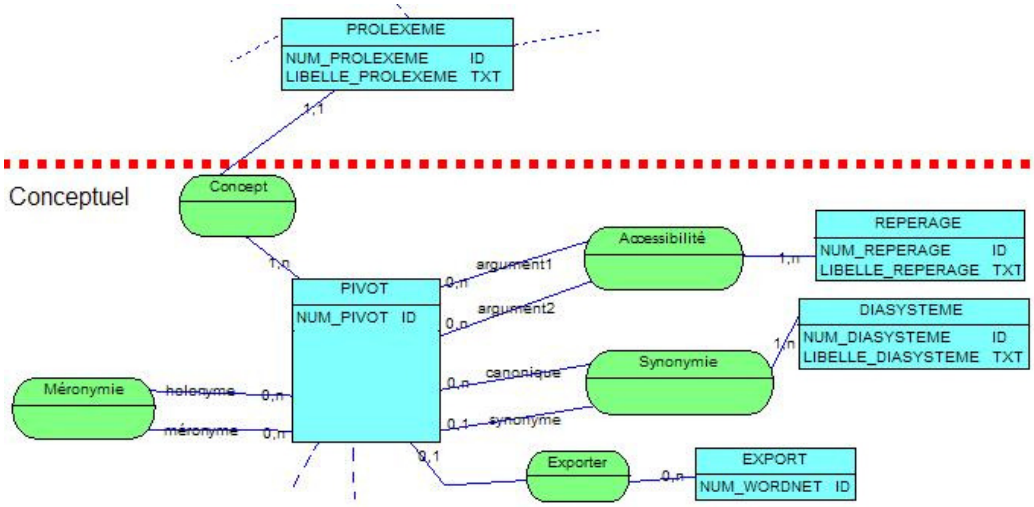

FIG. 5.6 – Le niveau conceptuel.

Un nom propre conceptuel sera en relation de synonymie avec un autre nom propre conceptuel suivant un diasystème (entité *DIASYSTEME*), qui peut être diachronique, diastratique ou diaphasique. La figure 5.7 présente la liste de repérages pour la relation d'accessibilité. Pour une relation de synonymie, nous avons imposé que chaque pivot peut être la forme canonique de plusieurs autres pivots et que chaque pivot peut être le synonyme d'une seule forme canonique.

L'entité *EXPORT* sert à relier les noms propres conceptuels de notre dictionnaire vers d'autres bases de données lexicales ou vers des encyclopédies. Des liens vers l'encyclopédie Wikipédia[1] et vers EuroWordNet ont été envisagés.

[1]Wikipédia est une encyclopédie gratuite accessible à l'adresse suivante : `http://www.wikipedia.org/`.

Repérage	**Exemple**
Capitale	Paris est la capitale de la France
Créateur	Auguste Rodin est le sculpteur du Penseur
Dirigeant non politique	Ray Norda est le patron de Novell
Dirigeant politique	Jacques Chirac est le président de la République française
Élève	Platon est l'élève de Socrate
Fondateur	Dardanos est le fondateur mythique de Troie
Héritier	Charles, prince de Galles, héritier du Royaume-Uni
Locataire	Jacques Chirac est le locataire de l'Elysée
Parent	Aaron est le frère de Moïse
Siège	Le Bureau Veritas a son siège à Paris
...	

FIG. 5.7 – Les repérages.

Le lien vers l'encyclopédie Wikipédia n'est pas conservé dans Prolexbase. Ce lien est généré dynamiquement sur le site de consultation en concaténant le code iso de la langue de consultation (fr, en, etc.), une url (wikipedia.org/wiki/Special:Search/) et le prolexème sélectionné par le visiteur. Pour le nom propre *France*, on produit ainsi le lien suivant :

http://fr.wikipedia.org/wiki/Special:Search/France

La génération automatique des liens vers l'encyclopédie Wikipédia présente un inconvénient. Tous les liens générés automatiquement n'ont pas été testés, l'interface de consultation peut par conséquent produire des liens qui n'existent pas, car certains articles ne sont pas présent dans cette encyclopédie, ou des liens vers un mauvais article. Pour éviter les liens incorrects, il faudrait vérifier manuellement chaque lien et les conserver dans la base de données. Il s'agit d'une tâche extrêmement longue. Par manque de temps, nous avons décidé de générer automatiquement les liens vers l'encyclopédie Wikipédia. Cette encyclopédie est en cours de développement : un lien incorrect aujourd'hui pourrait devenir correct le jour suivant.

Le lien vers la base lexicale EuroWordNet est conservé dans notre base de données grâce à l'entité *EXPORT*. Si le nom propre conceptuel existe dans EuroWordNet, son numéro ILI (Inter-Lingual-Index) apparaîtra dans l'entité *EXPORT*. Par exemple, on associera au nom propre *Paris* le numéro d'ILI *0558236n* (figure 5.8).

entity
location
region
area, country
center, middle, heart
seat
capital
national capital
Paris, City of Light, French capital, capital of France

FIG. 5.8 – Le nom propre *Paris* dans EuroWordNet.

5.2.2 Le niveau méta-conceptuel

La partie méta-conceptuelle (figure 5.9) comprend deux entités et quatre associations. L'entité *EXISTENCE* contient trois occurrences : historique, fictif et religieux. Nous avons regroupé les types et les supertypes dans une seule entité (*TYPE*), afin de pouvoir associer à un nom propre conceptuel un supertype, si l'on n'a pas d'information sur son type. Cela nous permet d'insérer dans notre dictionnaire des noms propres qui ont été trouvés par des systèmes de reconnaissance automatique de noms propres.

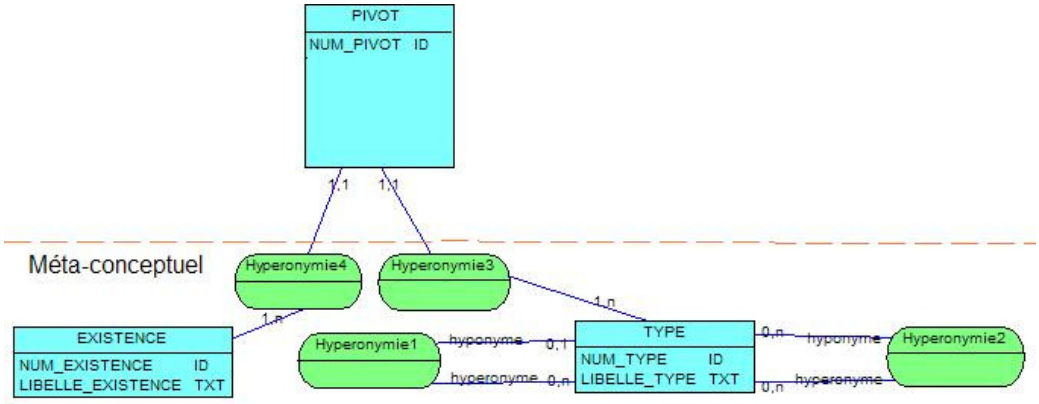

FIG. 5.9 – Le niveau méta-conceptuel.

5.2.3 L'éponymie

L'éponymie (figure 5.10) regroupe les entités *IDIOME*, *TERMINOLOGIE* et *ANTONOMASE*.

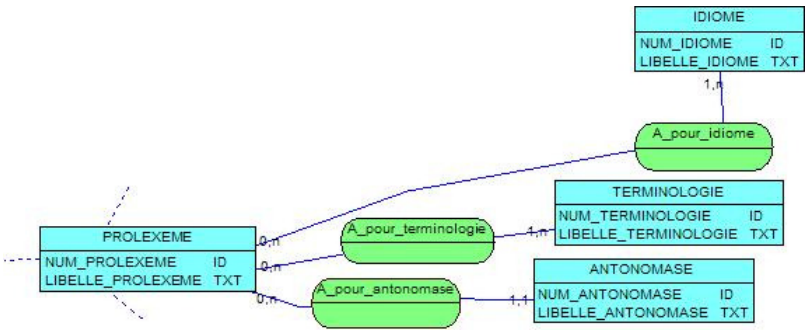

FIG. 5.10 – L'éponymie.

5.2.4 Les règles

L'entité *ALIASISATION* (figure 5.11) permet de stocker les règles de création d'alias à partir d'un prolexème. L'entité *DERIVATION* permet de stocker les règles de création de dérivés à partir d'un prolexème, d'un alias ou d'un dérivé.

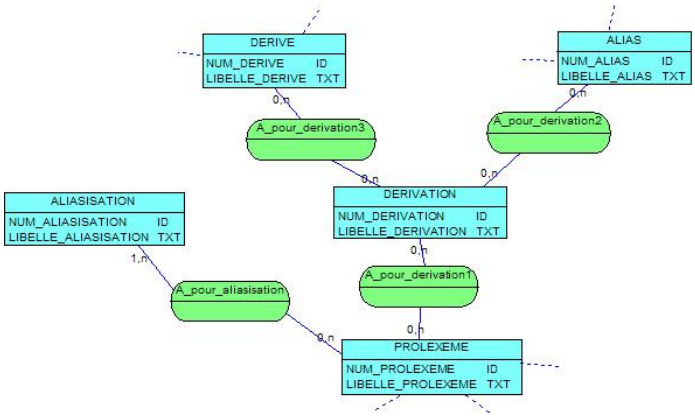

FIG. 5.11 – Les règles.

5.2.5 Les autres informations

Les informations supplémentaires (figure 5.12) sont formées de cinq entités et de cinq associations.

L'association *A_pour_statistique* permet d'associer à chaque prolexème des informations relatives à ses fréquences d'apparition (attribut *POIDS*) au sein d'un corpus donné (attribut *LIBELLE_STATISTIQUE*). Il peut s'agir, par exemple, d'étudier les fréquences d'apparition de noms propres sur quelques années d'un corpus journalistique. Certains noms propres apparaissant durant une année donnée pourront ne plus réapparaître quelques années plus tard. Cette étude statistique peut prendre en compte les différentes formes d'un même prolexème (ses alias, ses dérivés et leurs formes fléchies).

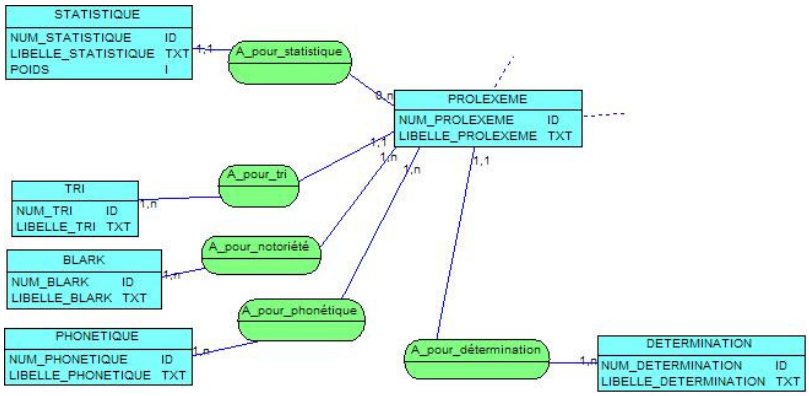

FIG. 5.12 – Les informations.

S'il est normal pour le simple mortel que nous sommes de ne pas posséder d'entrée dans les dictionnaires de noms propres, on peut parfois se demander pourquoi certains chanteurs ou chanteuses de variété française (*Johnny Hallyday*, etc.), actrices chinoises (*Michelle Yeoh*, etc.) ou autres célébrités ne figurent pas dans ces dictionnaires. On pourra aussi s'étonner que des villes telles que, par exemple, *Sainte-Enimie*, qui est le chef-lieu de canton de la *Lozère* comprenant à peine moins de 600 habitants et dont la majorité des Français ignore même l'existence, puisse apparaître dans le dictionnaire, alors que des villes de Russie, de Chine, et d'autres pays ayant une population nettement supérieure n'y figurent pas. Comme

le fait remarquer [Leroy, 1994], le choix d'inclure ou non un nom propre dans un dictionnaire repose essentiellement sur un critère de notoriété :

> *[Les dictionnaires] recueillent en effet uniquement les noms propres notoires : loin de recenser tous les noms propres, comme les dictionnaires de langue visent à recenser l'ensemble du lexique général, ils ne contiennent que les noms propres dont le référent a une certaine notoriété.*

Cette notoriété dépend beaucoup de différents facteurs extralinguistiques, tels que la culture, la région, la période considérée, etc. La figure 5.13 présente les indicateurs que nous avons utilisés. Ces indicateurs pourront servir à définir des dictionnaires de base dans une langue donnée, selon l'idée de [Cucchiarini et al., 2000] (BLARK pour *Basic LAnguage Ressources Kit*). Notons qu'un même nom propre pourra recevoir plusieurs indicateurs Blark. Par exemple, le prolexème français *Paris* correspond aux lignes 1, 2, 4, 5, 7 et 8 de la figure 5.13.

Sources des données françaises		Indicateur BLARK
INSEE (consultation Internet de juin 2005)	Les cinquante-sept villes françaises comportant plus de dix mille habitants	NATIONAL
Base de données Géopolis (consultation Internet de juin 2005)	Les quarante villes de l'Union Européenne comportant plus d'un million d'habitants	EUROPEEN
	Les soixante-quatorze principales villes du monde comportant plus de trois millions neuf cent mille habitants	INTERNATIONAL
Prolex (travaux antérieurs)	Toutes les villes françaises	DETAIL
	Les départements et les régions françaises	NATIONAL
	Hydronymes mondiaux	DETAIL
	Les pays de l'ONU, leurs capitales	INTERNATIONAL
	Tous les pays, régions et capitales mondiales	DETAIL
Dictionnaire Larousse du Collège, édition 2004	Extraction des noms propres en entrée d'un article	NATIONAL

FIG. 5.13 – Les indicateurs actuellement utilisés pour le BLARK.

La relation *A_pour_détermination* permet de spécifier si le prolexème comporte ou non un déterminant. On trouve la détermination dans de nombreuses langues, comme le français, l'anglais, l'allemand, etc. On constate que dans certaines langues, comme le serbe, ce phénomène n'existe pas. Un nom propre se construisant dans une langue avec une détermination peut apparaître dans une autre langue sans détermination. C'est le cas du nom propre *Spanien* en allemand qui devra être traduit en français par *l'Espagne*. Ce phénomène devra être pris en compte dans le cadre de la traduction automatique.

La phonétique permet de proposer une transcription d'un nom propre lorsque celui-ci ne possède pas de traduction et qu'il appartient à un autre système d'écriture. C'est le cas pour le prénom Paul qui se transcrira par :

- *Pol* en serbe alphabet latin
- *Пол* en serbe alphabet cyrillique
- *Поль* en russe

Il existe des prolexèmes ayant plusieurs prononciations différentes. Par exemple, les parisiens prononcent [*mɛts*] pour la ville de Metz alors que les Lorrains prononcent [*mɛs*].

La relation *A_pour_tri* donne des informations sur la façon de trier les noms propres polylexicaux. Nous avons attribué à chaque prolexème de notre dictionnaire un numéro qui correspond au début du cycle de tri de celui-ci. Par exemple, on associera au nom propre polylexical *mer des Philippines*, classé dans les dictionnaires sous la lettre P, le numéro de tri 3. Il devra être trié comme le mot polylexical *Philippines mer des*. C'est une simplification du modèle que nous avons présenté dans [Tran et al., 2005].

5.2.6 Les flexions

Un code flexionnel est attribué à chaque prolexème, alias et dérivé (figure 5.14). Pour la flexion des noms monolexicaux français, nous avons décidé d'utiliser les codes flexionnels du DELA [Courtois, 1992] [Paumier, 2006]. Une liste des codes flexionnels des noms monolexicaux (entité *FLEXION*), utilisée dans Prolexbase, est donnée en annexe B.

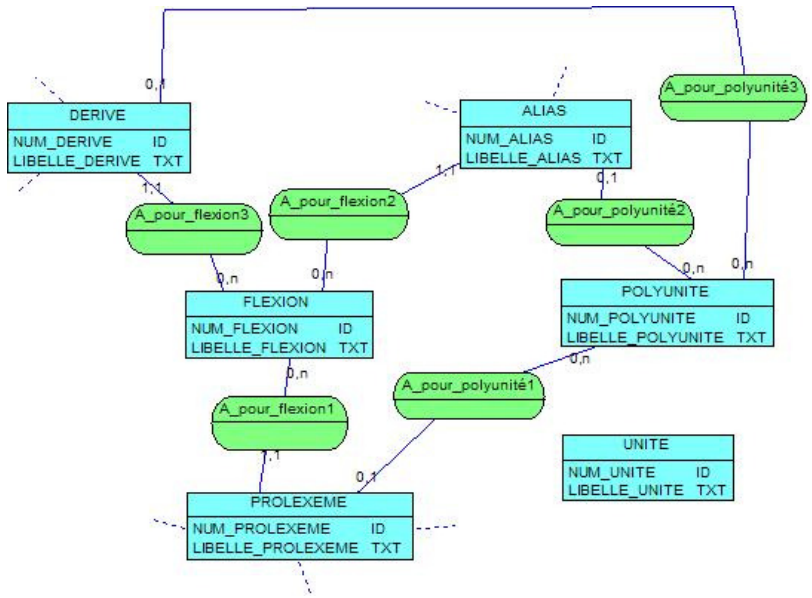

FIG. 5.14 – La flexion.

Nous avons prévu d'utiliser les codes flexionnels inspirés de [Savary, 2006] pour les noms propres polylexicaux. Chaque nom propre polylexical sera en relation avec une polyunité (entité POLYUNITE) qui correspond à une concaténation d'unités *UNITE*. Par exemple, nous associerons au nom relationnel *Antiguais-et-Barbudien*, dont le prolexème est *Antigua-et-Barbuda* :

– un code flexionnel pour les deux unités : $Antiguais.N61 : ms$, $Barbudien.N41 : ms$.
– un graphe de flexion (figure 5.15).

5.2.7 Les instances

L'entité *INSTANCE* (figure 5.16) regroupe l'ensemble des formes fléchies des prolexèmes, des alias et des dérivés. Selon la langue, à travers l'association *A_pour_morphologie*, nous indiquons pour chaque instance des informations morphologiques :

– CLASSE : nom, adjectif, etc.

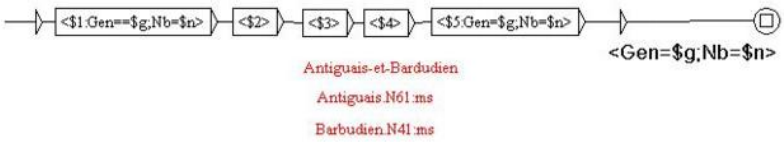

FIG. 5.15 – Le graphe de flexion d'*Antillais-et-Barbudien*.

– GENRE : masculin, féminin, etc.
– CAS : nominatif, accusatif, etc.
– NOMBRE : singulier, pluriel, etc.

Cette entité est utilisée pour les recherches de noms propres à partir de leurs formes fléchies à travers une interface web de consultation (voir section 7.2). Le visiteur rentre un nom propre fléchi et l'interface lui affiche le prolexème correspondant avec ses informations.

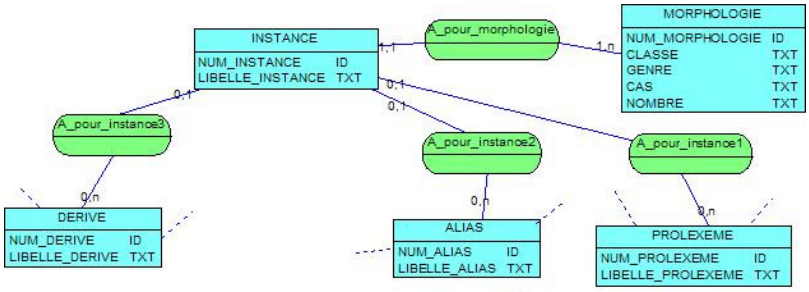

FIG. 5.16 – Les instances.

Les instances pour le français sont générées en utilisant le module *Inflect* d'Unitex [Paumier, 2003]. Le programme accepte en entrée une liste de mots au format DELA (voir chapitre 2). Voici un exemple d'un fichier que nous fournissons au module *Inflect* :

Onusien,N41+48226+48226+20394+22

Sur cette ligne, *Onusien* est le lemme et *N41* le code flexionnel ; les quatre traits qui suivent sont les numéros respectifs du pivot et du prolexème, puis éventuellement de l'alias et du dérivé.

Le programme fournit pour cette ligne le résultat suivant :

Onusiennes,Onusien.N+48226+48226+20394+22:fp
Onusiens,Onusien.N+48226+48226+20394+22:mp
Onusienne,Onusien.N+48226+48226+20394+22:fs
Onusien,Onusien.N+48226+48226+20394+22:ms

En récupérant ces informations et d'autres informations contenues dans la base, nous pouvons créer automatiquement les occurrences de l'entité *INSTANCE*.

Pour certaines langues, cette génération des instances risque de coûter cher en espace mémoire en raison d'un nombre important de cas, de genres et de nombres. Nous avons décidé que chaque langue développera sa propre stratégie pour générer ses instances. Chaque langue pourra soit stocker les formes fléchies dans cette entité, soit utiliser des systèmes morphologiques externes.

5.3 Modèle logique de données

Pour pouvoir créer les tables de notre base de données, nous avons traduit notre MCD des noms propres en MLD en appliquant les trois règles présentées dans la première partie. En traduisant directement notre MCD, nous aurions un modèle physique de données qui pourrait présenter les inconvénients suivants :

- La plupart des SGBD (MySQL, Access, etc.) possèdent une limitation sur la taille des tables. Si le nombre de langues et de données que l'on souhaite intégrer à notre base de données devenait assez grand, la taille des tables *PROLEXEME*, *ALIAS*, *DERIVE* et *INSTANCES* risquerait de dépasser cette taille limite.
- Certaines requêtes SQL, comme la recherche de données, la mise à jour, etc., risqueraient d'être longues.
- Selon les langues, certaines entités, associations ou propriétés ne seront jamais utilisées. En français, les associations *Accepte_comme4* et *A_pour_derivation3* ne seront pas utilisées, car les prolexèmes ne possèdent pas de dérivés de dérivés[2]. La propriété *CAS* de l'entité *MORPHOLOGIE* n'est pas utile pour le français. L'entité *DETERMINATION* et l'association *A_pour_détermination* ne seront jamais utilisées pour le serbe, car dans cette langue les groupes nominaux ne possèdent pas de déterminant.

A cause de ces diverses raisons, nous avons décidé de séparer chaque langue de notre dictionnaire afin de mettre en place pour chacune une structure plus adaptée. Notre modèle logique de données final comprendra deux parties : une partie commune aux langues traitées (figure 5.17) et une partie spécifique à chaque langue (figure 5.18). Le niveau méta-conceptuel et le niveau conceptuel appartiennent à la partie commune aux langues et seront reliés à la partie spécifique d'une langue donnée par la relation *Concept*.

Un schéma relationnel de données pour le français et un pour le serbe sont donnés en annexe A (voir page 151).

Conclusion

Nous avons utilisé la méthode Merise pour définir un MCD qui s'applique pour toutes les langues traitées. La traduction de ce MCD en MLD soulève un certain nombre de problèmes : limitation sur la taille des tables, rapidité des requêtes SQL, absence ou présence de tables spécifiques à certaines langues. A cause de ces différentes raisons, nous avons décidé de transformer ce MCD en un MLD comprenant deux parties : une partie commune aux langues et une partie particulière à chaque langue.

Le principal objectif de nos travaux est de développer des ressources linguistiques multilingues sur les noms propres qui puissent être mises à disposition de la communauté des chercheurs en TAL. Il nous paraît donc indispensable de développer un format d'exportation de notre base de données. Ce format devra être indépendant des systèmes d'exploitation et compatible avec la plupart des outils existants. Nous présenterons dans le chapitre suivant le modèle XML d'exportation de Prolexbase.

[2]En effet, comme nous l'avons expliqué au chapitre 3, les dérivés d'un prolexème sont des synonymes de celui-ci à une transformation près. Le dérivé *pasteurisation* est un dérivé du dérivé *pasteurisé*, mais ceux-ci ne figurent pas dans Prolexbase.

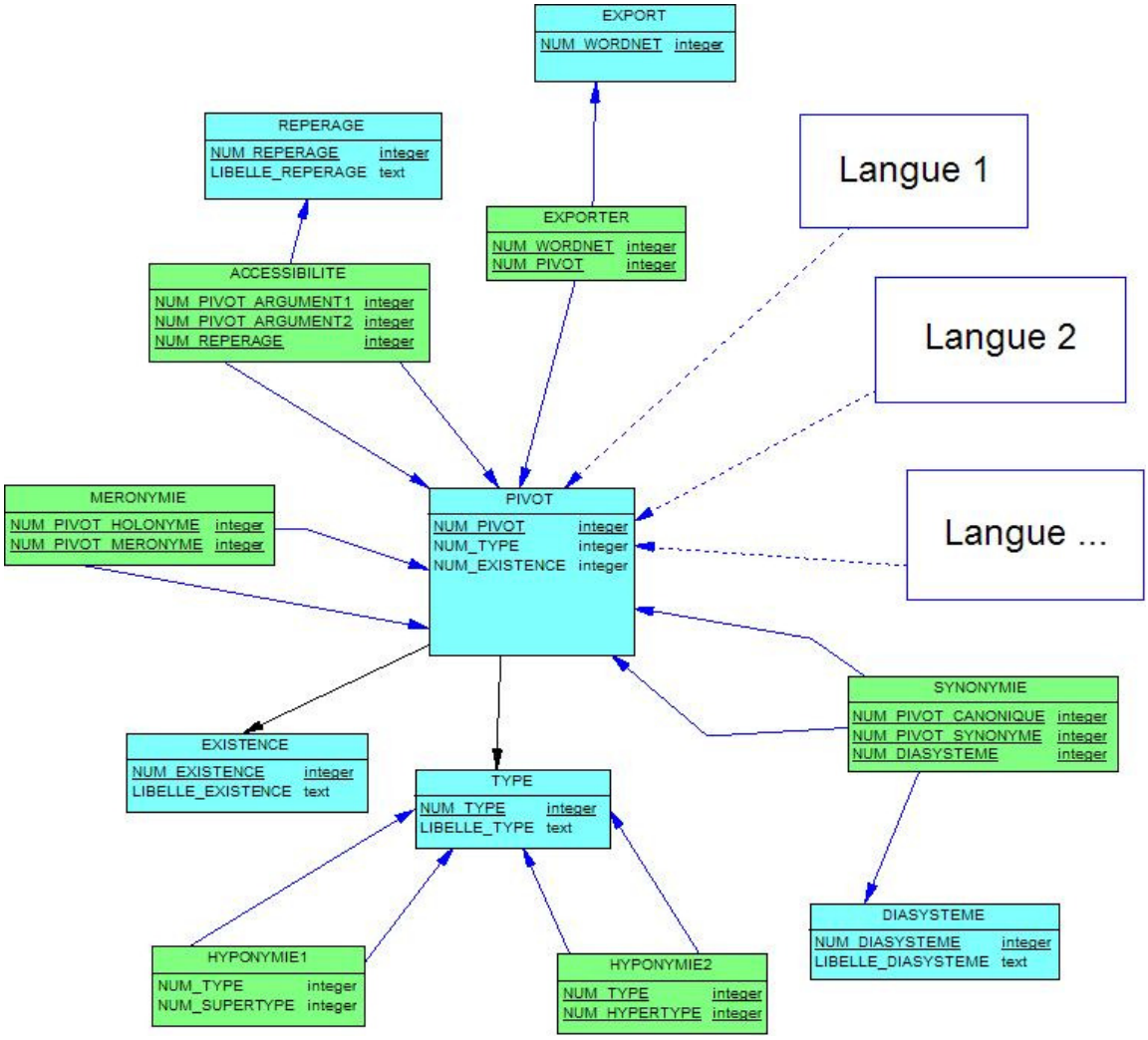

FIG. 5.17 – Le modèle relationnel de données : partie commune.

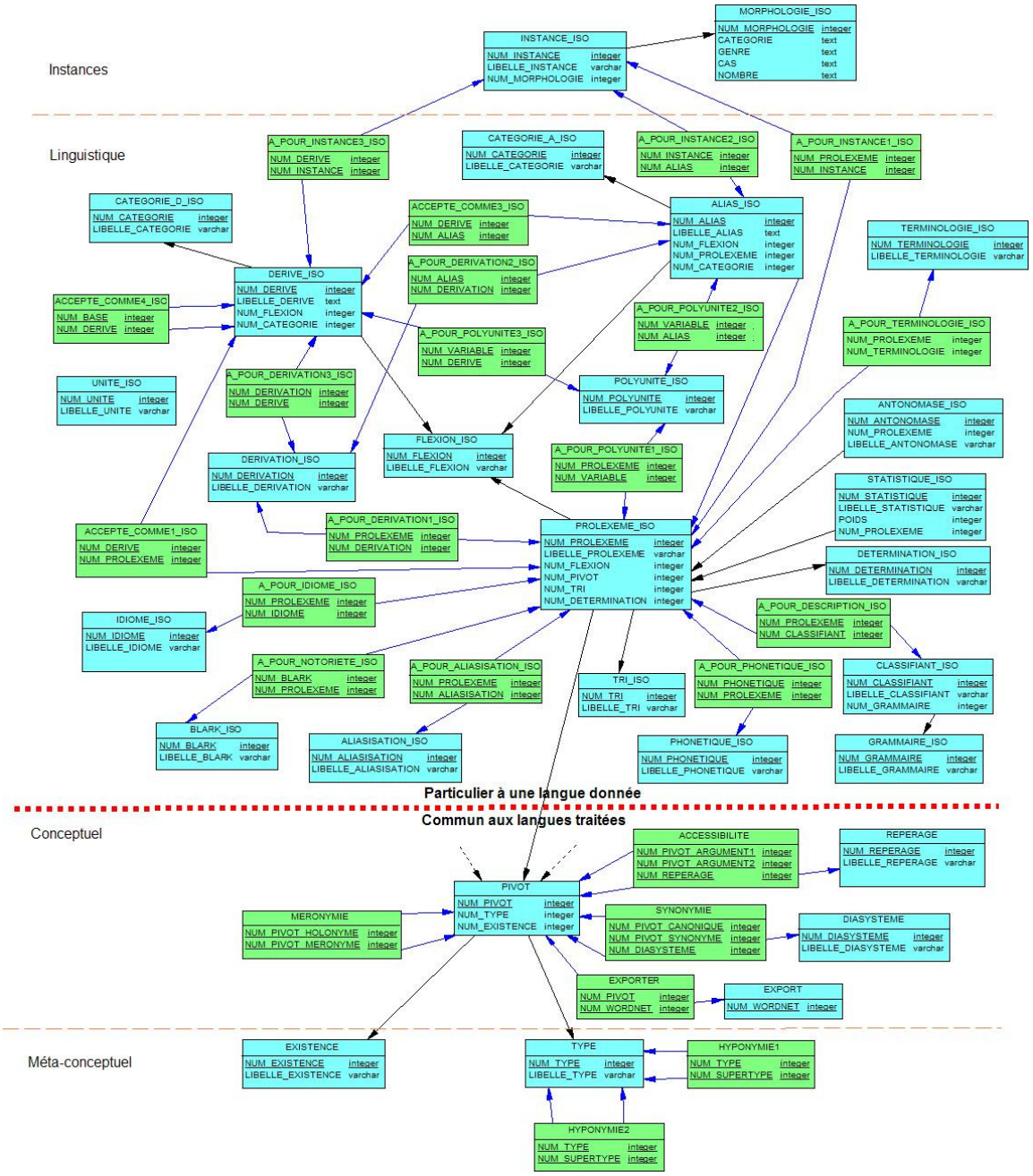

FIG. 5.18 – Le modèle relationnel de données : partie spécifique.

Chapitre 6

Exportation des données

Introduction

Dans l'article [Bouchou et al., 2005], nous avons proposé une version XML de notre base de données. Ce chapitre est consacrée à présenter une version respectant la norme TMF. Nous allons dans un premier temps présenter les caractéristiques de deux standards pour le partage et l'échange de données sur la Toile : la TEI et la TMF. Ensuite, nous présenterons le modèle d'échange et d'exportation de données pour Prolexbase que nous avons défini en nous inspirant de la TMF. Enfin, nous décrirons notre contribution au projet Outilex.

6.1 État de l'art

6.1.1 TEI

Apparue officiellement en 1988, la TEI[1] (Text Encoding Initiative) est née du besoin de normalisation du balisage de textes électroniques. L'objectif du projet est de développer un format indépendant des systèmes ou des logiciels informatiques, simple et facile pour les utilisateurs, permettant le partage et l'échange de données textuelles. Ce format doit être assez riche pour permettre à des chercheurs, provenant de différents domaines et de divers pays, de baliser leurs textes électroniques. Le projet est soutenu par de nombreuses institutions et associations, telles que l'Association for Computational Linguistics, l'Association for Literary and Linguistic Computing, le Social Sciences and Humanities Research Council du Canada, la Commission européenne, etc.

La DTD[2] de la TEI repose sur une architecture formée d'un ensemble de trois modules :

- *core tag sets* : il s'agit d'un ensemble de balises obligatoires pour toute DTD TEI (en-tête, paragraphe, divisions, etc.).
- *base tag sets* : ce module contient un ensemble de balises de base spécifique pour six catégories de textes : prose, poésie en vers, œuvre théâtrale, transcription du discours, dictionnaire et base terminologique. La figure 6.1 présente un exemple des balises TEI utilisées pour coder l'article *abandon* d'un dictionnaire.
- *additional tag sets* : il fournit un ensemble de balises additionnelles qui peut être utilisé pour n'importe quel type de textes. On trouve, par exemple, des balises pour repérer les noms de personnes (<*persName*>, etc.), de lieux (<*placeName*>, etc.) et d'organisations (<*orgName*>, etc.). La figure 6.2 donne un exemple avec la balise

[1] est accessible sur le site www.tei-c.org

[2] Le schéma est également défini en XML-Schema ou RELAX-NG

<*persName*> ; l'attribut *key* permet d'identifier le nom propre de façon unique dans un texte.

a.ban.don 1 /@"b&nd@n/ v [T1] 1 to leave completely and for ever; desert: The sailors abandoned the burning ship. 2 ... **abandon** 2 n [U] the state when one's feelings and actions are uncontrolled; freedom from control: The people were so excited that they jumped and shouted with abandon / in gay abandon. [LDOCE]

```
<superEntry>
  <form>
    <orth>abandon</orth>
    <hyph>a|ban|don</hyph>
    <pron>@"b&amp;nd@n</pron>
  </form>
  <entry n="1">
    <gramGrp>
      <pos>v</pos>
      <subc>T1</subc>
    </gramGrp>
    <sense n="1">
      <def>to leave completely and for ever ... </def>
      <!-- ... -->
    </sense>
    <sense n="2"> <!-- ... --> </sense>
  </entry>
  <entry n="2">
    <gramGrp>
      <pos>n</pos>
      <subc>U</subc>
    </gramGrp>
    <def>the state when one's feelings and actions are
      uncontrolled; freedom from control</def>
    <!-- ... -->
  </entry>
</superEntry>
```

FIG. 6.1 – Exemple de balise pour les dictionnaires.

```
<persName key="FDR1">
  <foreName>Franklin</foreName>
  <foreName>Delano</foreName>
  <surname>Roosevelt</surname>
</persName>
```

FIG. 6.2 – Exemple avec la balise <*persName*>.

La DTD proposée par la TEI pour les dictionnaires est destinée à la création de dictionnaires pour les humains. Les concepteurs de la TEI ont rencontré de nombreux problèmes lors de l'élaboration de cette DTD [Ide and Véronis, 1996]. Il était difficile de proposer une DTD qui soit à la fois assez générale pour décrire tous les dictionnaires et assez précise pour faire ressortir la spécificité de chaque dictionnaire.

Le chapitre 13[3] sur les bases terminologiques est devenu obsolète en raison de la sortie de la norme ISO 16642 ou TMF (Terminological Markup Framework) :

> *Since its first publication, this chapter has been rendered obsolete in several respects, chiefly as a result of the publication of ISO 12200, and a variant of it*

[3]sur le site www.tei-c.org consulté le 16/06/2006

(TBX) which has been recently adopted by LISA, the Localisation Industry Standard Association. Work is currently ongoing in the ISO community to define a generic platform for terminological markup (ISO CD 16642, TMF : Terminological Markup Framework), in the light of which it is anticipated that the recommendations of the present chapter will be substantially revised.

6.1.2 TMF

Le Terminological Markup Framework (TMF) ou la norme ISO 16642 [Romary, 2002] [Romary and Van Campenhoudt, 2001] propose un standard de représentation pour des données terminologiques multilingues en XML. Il définit un ensemble de contraintes que chaque langage de description de données terminologiques (TML) doit suivre. Chaque TML se caractérise par :

- un méta-modèle : il s'agit d'un squelette (figure 6.3) décrivant la structure de tout TML. Chaque entrée (TE) correspond à un ou plusieurs termes (TS) dans plusieurs langues (LS).
- un lien vers des catégories de données de la norme ISO 12620. Par exemple, la propriété *Content* précise le type du contenu, *DCName* un nom, etc.

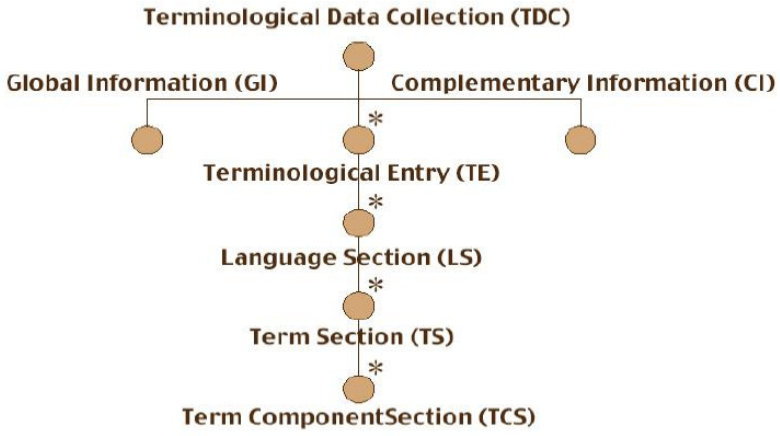

FIG. 6.3 – Le méta-modèle du TMF.

GMT (Generic Mapping Tool) est un outil permettant de représenter la structure du TMF sous le format XML. La figure 6.4 donne un exemple de représentation GMT. La représentation GMT repose sur deux balises :

- <struct> : permettant de définir chaque niveau du squelette structurel.
- <feat> : permettant de préciser le trait d'un nœud.

6.1.3 Discussion

La TEI propose un format de balisage figé qui n'est pas adapté pour modéliser notre base de données avec ses différents niveaux (niveau indépendant de la langue et niveau dépendant de la langue). Le balisage pour décrire les dictionnaires de la TEI peut être intéressant pour créer des dictionnaires monolingues pour les humains. Ce balisage ne permet pas de modéliser le niveau interlangue et nos relations sémantiques. Pour les bases terminologiques, la TEI recommande l'utilisation de la TMF.

```
<struct type="TE">
   <feat type="id">ID67</feat>
   <feat type="subjectField">manufacturing</feat>
   <feat type="definition">A value between 0 and 1 used in ...</feat>
   <struct type="LS">
      <feat type="lang">en</feat>
      <struct type="TS">
         <feat type="term">alpha smoothing factor</feat>
         <feat type="termType">fullForm</feat>
      </struct>
   </struct>
   </struct>
</struct>
```

FIG. 6.4 – Exemple de représentation avec le Generic Mapping Tool.

La TMF propose une architecture en deux niveaux : niveau dépendant de la langue et niveau indépendant de la langue, plus proche de notre modèle. Cependant, il manque une profondeur dans le niveau linguistique (indispensable pour la traduction des noms propres) et une modélisation pour des relations qui ne dépendent pas de la langue (synonymie, méronymie et accessibilité).

Comme la GMT propose un format de balisage assez flexible permettant de modéliser facilement n'importe quelle structure de données grâce à deux balises (*<struct>* et *<feat>*), nous pourrons l'adapter à notre architecture.

6.2 Le modèle XML de Prolexbase

Dans cette partie, nous allons présenter le format des fichiers de requête XML et le format des fichiers d'exportation de la base.

6.2.1 Fichier de requête XML

Toute application souhaitant extraire des données de Prolexbase doit envoyer une requête au format XML. Voici les balises qui doivent apparaître :

- *Request* : contient la structure de la requête.
- *Libelle* : nom propre sur lequel on souhaite faire une recherche.
- *RequestLanguage* : langue et le système d'écriture dans laquelle est écrit le nom propre que l'on recherche.
- *ProperName* : regroupe les informations que l'on souhaite avoir sur le nom propre :
  1. *Prolexeme* : spécifie si l'on souhaite récupérer le prolexème du nom propre que l'on recherche.
  2. *Type* : spécifie si l'on souhaite avoir le type du nom propre.
  3. *Existence* : spécifie si l'on souhaite avoir l'existence du nom propre.
  4. *Alias* : si l'on souhaite avoir tous les alias du prolexème.
  5. *Derivative* : si l'on souhaite avoir tous les dérivés du prolexème.
- *Lemmas* : regroupe les informations sur la catégorie et la classe du prolexème, de chaque alias et chaque dérivé.

```
<?xml version='1.0' encoding='ISO-8859-1' standalone='no' ?>
<!DOCTYPE Request SYSTEM "../requetes/Requete_DTD.dtd">
<Request>
      <Libelle >Organisation des nations unies</Libelle>
      <RequestLanguage>fr</RequestLanguage>
      <ProperName>
            <Prolexeme  status='ON'/>
            <Type status='ON'/>
            <Existence status='ON'/>
            <Alias status='ON' />
            <Derivative status='OFF' />
      </ProperName>
      <Lemmas>
            <Lemma status='ON'/>
            <Pos status='ON' />
            <Category status='ON' />
      </Lemmas>
      <Inflexions>
            <Form status='ON'/>
            <Gender status='ON' />
            <Number status='ON' />
      </Inflexions>
      <AnswerLanguage>
            <Language>fr</Language>
      </AnswerLanguage>
</Request>
```

FIG. 6.5 – Requête XML.

1. *Lemma* : libellé du prolexème, de l'alias ou du dérivé.
2. *Pos* : classe du prolexème, de l'alias ou du dérivé.
3. *Category* : catégorie du prolexème, de l'alias ou du dérivé.

– *Inflexions* : précise si l'on souhaite avoir les formes fléchies.

1. *Form* : libellé de la forme fléchie.
2. *Gender* : genre de la forme fléchie.
3. *Number* : nombre de la forme fléchie.

– *AnswerLanguage* : regroupe les langues dans lesquelles on souhaite faire une recherche.

1. *Language* : langue dans laquelle s'effectue la recherche.

La figure 6.5 présente un exemple de requête avec le nom propre *Organisation des nations unies*. La valeur *ON* de l'attribut *statut* de la balise *Alias* précise que le résultat doit faire apparaître les alias. La valeur *OFF* de l'attribut *Derivative* indique que les dérivés ne doivent pas apparaître dans le fichier résultat.

6.2.2 Fichier d'exportation XML

Pour créer le modèle XML d'exportation de notre base de données, nous nous sommes inspirés de la représentation GMT de la TMF, car l'utilisation d'une norme garantit la portabilité et la compatibilité de notre format avec la plupart des systèmes existants. De plus, cette norme permet de modéliser facilement n'importe quelle structure de données en se basant sur deux balises. La figure 6.6 présente le modèle sous la forme d'une arborescence.

Voici la liste des balises pouvant apparaître dans ce fichier :

– *Prolex* : contient les entrées.

– *Pivot* : Il s'agit d'une entrée qui ne dépend pas de la langue et qui correspond à la TE du méta-modèle de la TMF. Cette balise comprend les attributs suivants :

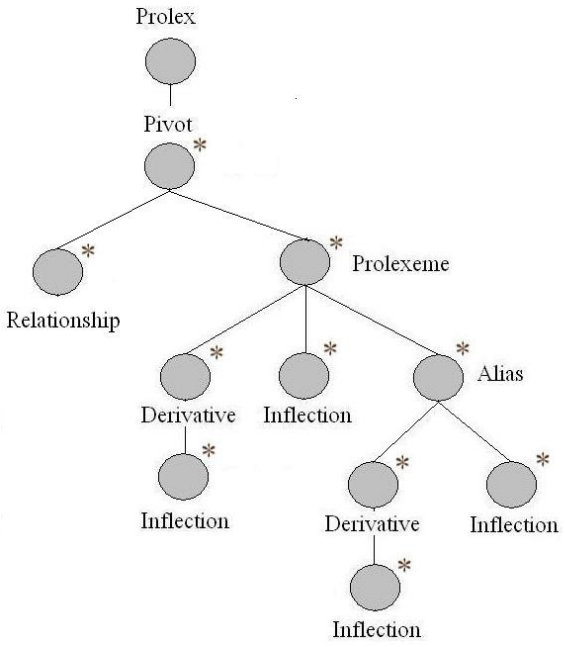

FIG. 6.6 – Le modèle de Prolexbase.

1. *type* : indique type du pivot.
2. *existence* : indique existence du pivot.
3. *identifier* : précise le numéro du pivot.

– *Relationship* : précise les relations du pivot avec d'autres pivots. On y trouve les attributs :

1. *relation* : nom de la relation (méronymie, synonymie ou accessibilité).
2. *identifier* : indique le numéro du pivot en relation.
3. *argument* : précise quel est l'argument 1 et l'argument 2 de la relation. Soit deux pivots P_1 et P_2 et R une relation entre deux arguments telle que R(P_1, P_2). Dans le cadre d'une relation de synonymie, P_1 est appelé le synonyme et P_2 est appelé le canonique. Pour une relation de méronymie, P_1 est appelé le méronyme et P_2 est appelé l'holonyme. Dans une relation d'accessibilité, P_1 est appelé l'argument 1 et P_2 l'argument 2.
4. *context* : précise le diasystème pour une relation de synonymie, le repérage pour une relation d'accessibilité.

– *Prolexeme* : Cette balise correspond à une entrée dans une langue donnée (équivalente à la LS du méta-modèle de la TMF). Elle accepte les attributs suivants :

1. *language* : indique la langue à laquelle appartient le prolexème.
2. *lemma* : libellé du prolexème.
3. *pos* : donne la catégorie grammaticale du prolexème.
4. *category* : précise qu'il s'agit d'un nom propre.

– *Alias* : regroupe les alias du prolexème. On trouve les attributs suivants :

1. *lemma* : lemme de l'alias.

2. *pos* : donne la catégorie grammaticale de l'alias.
3. *category* : précise la catégorie de l'alias (acronyme ou sigle, abréviation, etc.).

– *Derivative* : regroupe les dérivés du prolexème ou des alias. On trouve les attributs suivants :
1. *lemma* : lemme du dérivé.
2. *pos* : donne la catégorie grammaticale du dérivé.
3. *category* : précise la catégorie du dérivé (nom relationnel, adjectif relationnel, préfixe, etc.).

– *Inflection* : contient une forme fléchie pouvant provenir soit d'un dérivé, soit du prolexème, soit d'un alias. On trouve les attributs suivants :
1. *form* : libellé de la forme fléchie.
2. *gender* : donne le genre de la forme fléchie.
3. *number* : contient le nombre (singulier, pluriel, etc.).
4. *case* : pour les langues casuelles.

La figure 6.7 donne le résultat de la requête XML de la figure 6.5. Un exemple complet avec le prolexème *États-Unis d'Amérique* est donné en annexe C. La figure 6.8 donne un exemple de relation de méronymie (la *France* est un méronyme de l'*Europe*) et de relation d'accessibilité (*Paris* est la capitale de la *France*).

6.3 Implémentation

Nous avons encadré un stage d'une étudiante de Master qui a développé une interface en PHP permettant à chaque visiteur ou à chaque application de soumettre des requêtes à notre base de données.

La figure 6.9 présente l'architecture globale permettant d'exporter nos données. Le visiteur précise dans une interface web les données qu'il souhaite extraire de Prolexbase. L'interface génère à l'aide d'un module de construction de requête un fichier de requête en format XML (voir section 6.2.1) qui est envoyé à un module de traitement de requête. Ce module interroge la base de données et renvoie le résultat à l'interface sous la forme d'un fichier XML (voir section 6.2.2). Toute application peut directement soumettre un fichier de requête XML au module de traitement de requêtes et obtenir un fichier XML résultat.

6.4 Une contribution effective

Avant la conception de ce format d'exportation, nous avions contribué au Projet Outilex par la création des dictionnaires Prolex-Toponymes et Prolex-PaysCapitales. Ces dictionnaires ont été réalisés en extrayant des toponymes français de notre base de données sous le format DELAF (voir la section 2.1 page 32).

Le dictionnaire Prolex-Toponymes comprend 9 225 entrées, dont 2 110 toponymes, 3 415 gentilés, 3 407 adjectifs toponymiques, 12 préfixes toponymiques et 281 hydronymes. Voici un extrait de ce dictionnaire :

Lyon,.N+PR+DetZ+Toponyme+Ville:ms:fs
lyonnais,lyonnais.A+Toponyme+Ville:ms:mp
Lyonnais,Lyonnais.N+PR+Hum+Toponyme+Ville:ms:mp
Mékong,.N+PR+Hydronyme:ms
Vallée des Rois,.N+PR+Toponyme+Region:fs

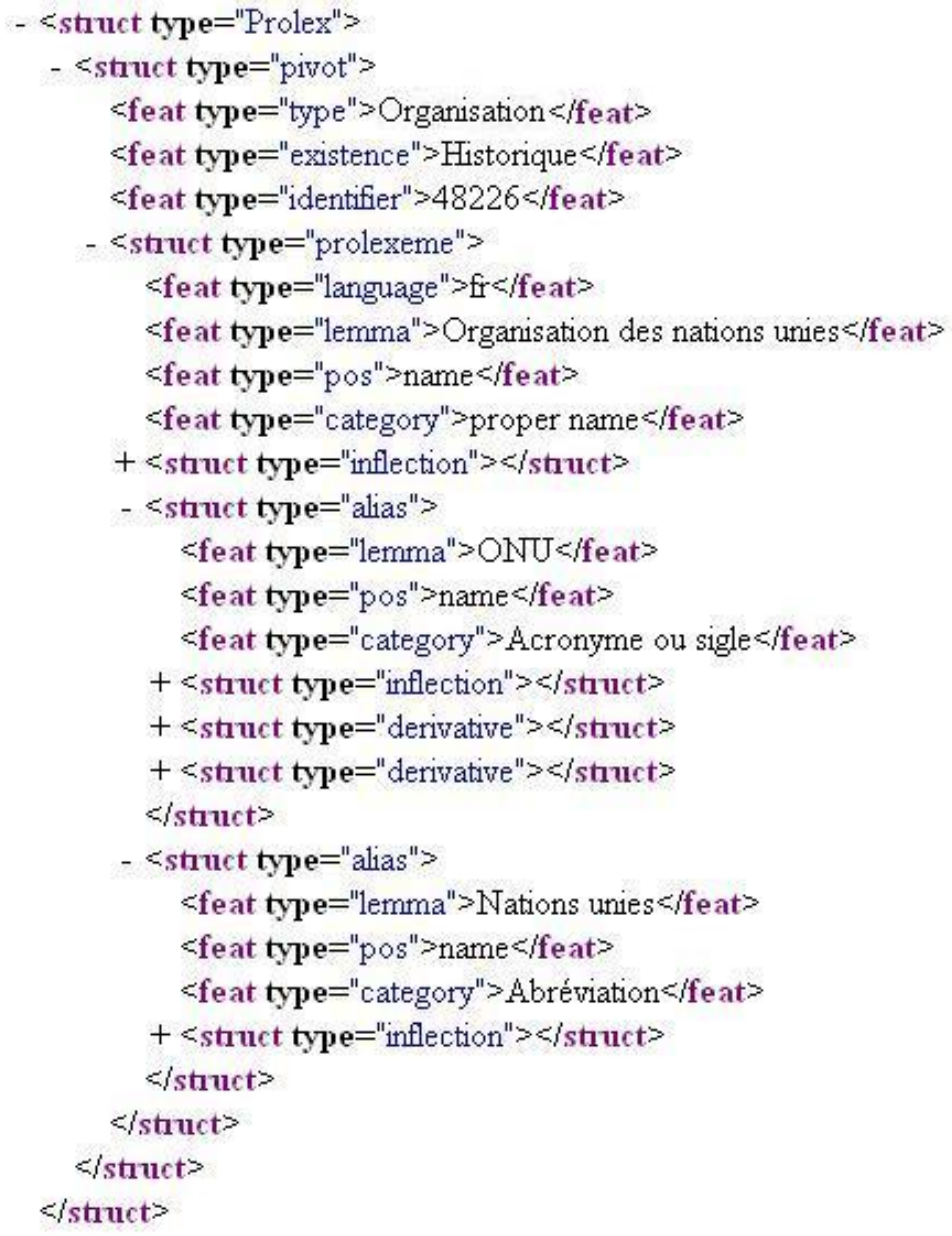

FIG. 6.7 – Résultat d'une requête XML.

```
<struct type="Prolex">
    <struct type="pivot">
        <feat type="type">Country</feat>
        <feat type="existence">Historical</feat>
        <feat type="identifier">27</feat>
        <struct type="relationship">
            <feat type="relation">meronymy</feat>
            <feat type="identifier">47947</feat>
            <feat type="argument">arg1</feat>
        </struct>
        <struct type="relationship">
            <feat type="relation">accessibility</feat>
            <feat type="identifier">38558</feat>
            <feat type="argument">arg2</feat>
            <feat type="context">capital</feat>
        </struct>
        <struct type="prolexeme">
            <feat type="language">fr</feat>
            <feat type="lemma">France</feat>
            ...
        </struct>
    </struct>
    <struct type="pivot">
        <feat type="type">Supranational</feat>
        <feat type="existence">Historical</feat>
        <feat type="identifier">47947</feat>
        <struct type="prolexeme">
            <feat type="language">fr</feat>
            <feat type="lemma">Europe</feat>
            ...
        </struct>
    </struct>
    <struct type="pivot">
        <feat type="type">City</feat>
        <feat type="existence">Historique</feat>
        <feat type="identifier">38558</feat>
        <struct type="prolexeme">
            <feat type="language">fr</feat>
            <feat type="lemma">Paris</feat>
            ...
        </struct>
    </struct>
</struct>
```

FIG. 6.8 – Exemple de relations.

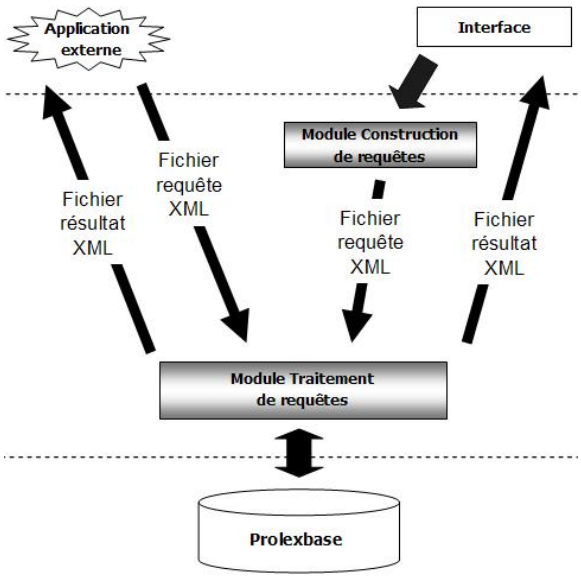

FIG. 6.9 – Architecture d'exportation de Prolexbase.

La figure 6.10 présente la liste des traits utilisés dans le dictionnaire Prolex-Toponymes.

+PR	Les noms propres
+Hum	Les humains (les gentilés)
+Toponyme	Les toponymes et les gentilés
+Ville	Les villes et les gentilés
+Region	Les régions et les gentilés
+Pays	Les pays (indépendants) et les gentilés
+Hydronyme	Les hydronymes
+DetZ	Les noms sans déterminant

FIG. 6.10 – Traits du dictionnaire Prolex-Toponymes.

Le dictionnaire Prolex-PaysCapitales regroupe les 191 pays indépendants avec leur capitale, les gentilés et les adjectifs toponymiques. Nous avons aussi ajouté dans ce dictionnaire les régions qui sont assimilées à des pays, comme par exemple le Royaume-Uni, les émirats, etc. Il comporte 3 092 entrées, dont 592 toponymes, 1 250 gentilés, 1 240 adjectifs toponymiques et 10 préfixes toponymiques. Voici quelques entrées de ce dictionnaire :

Athènes,.N+PR+DetZ+Toponyme+Ville+IsoGR:ms:fs
athénien,athénien.A+Toponyme+Ville+IsoGR:ms
Roumains,Roumain.N+PR+Hum+Toponyme+Pays+IsoRO:mp
Roumanie,.N+PR+Toponyme+Pays+IsoRO:fs
Suède,.N+PR+Toponyme+Pays+IsoSE:fs
suédois,suédois.A+Toponyme+Pays+IsoSE:ms:mp

Nous avons précisé pour chaque entrée le code ISO de son pays. Par exemple, *IsoRO* pour la *Roumanie*, *IsoSE* pour la *Suède*, etc.

Ces deux dictionnaires sont gratuitement mis à disposition des utilisateurs d'Unitex et téléchargeables à l'adresse suivante : `http://tln.li.univ-tours.fr/Tln_Unitex.html`.

Ils sont stockés dans des fichiers enregistrés au format Unicode Little-Endian ou UTF-16, qui est le format des fichiers utilisés par Unitex [Paumier, 2006].

Conclusion

Il est possible de formuler une requête en format XML pour rechercher des noms propres dans Prolexbase. Le résultat des requêtes est un fichier XML dans le format GMT de la TMF.

Chapitre 7

Interface web

Introduction

Dans la première partie de ce chapitre, nous allons établir le cahier des charges dans lequel nous présenterons les besoins et les solutions que nous avons adoptés. Les deux autres parties seront consacrées à décrire en détail les menus du site web de consultation et de l'interface de travail collaboratif de Prolexbase. Nous présenterons des calculs de complexité sur quelques menus de l'interface de travail collaboratif.

7.1 Cahier des charges

Spécification des besoins

L'aspect coopératif du projet nécessite le développement d'une interface qui soit accessible sur Internet à tous les participants du projet et à toute personne extérieure au projet. L'interface web que nous devons développer doit respecter ces différentes contraintes :

- fonctionner sur tout système d'exploitation et tout navigateur Internet.
- présenter une interface conviviale qui permet aux utilisateurs de travailler efficacement à la fois sur des données simples et sur des listes de données. Pour cela, il faudra proposer un format simple de mise en page des fichiers de données qui permettra à tout utilisateur de travailler chez lui sans accès à l'interface.
- permettre aux utilisateurs de travailler dans des langues différentes.
- mettre en place un serveur de base de données accessible sur Internet.
- permettre la traduction des menus et des messages de l'interface de travail dans plusieurs langues.

Implémentation

Nous avons décidé de développer deux interfaces web. La premiere, beaucoup plus simple (voir section 7.2), permet aux visiteurs de consulter les données. La deuxième permettra aux participants du projet de travailler sur les données (voir section 7.3).

Il existe de nombreux systèmes de gestion de base de données. Parmi ceux-ci, nous avons décidé de créer notre base de données sous MySQL 4.1, car il s'agit d'un outil gratuit, très répandu et disponible au sein de notre laboratoire. MySQL est un système de gestion de bases de données libre possédant une architecture multiutilisateur et multitraitement. De nombreux langages de programmations (C, C++, Java, Perl, PHP, etc.) peuvent être utilisés pour créer des interfaces avec une base de données MySQL.

Travailler dans un environnement multilingue nécessite l'utilisation d'un format d'encodage universel : Unicode. Tous les mots stockés dans les tables, comme la table *PROLEXEME*, la table *ALIAS*, etc., de notre base de données sont codés avec la norme UTF-16. Nous avons choisi d'adopter UTF-16, car il s'agit de la norme utilisée par le logiciel UNITEX développé par le laboratoire de Marne-la-Vallée avec lequel nous avons plusieurs projets en commun. Nous espérions qu'une version de MySQL intégrant l'UTF-16 serait disponible avant la fin de la thèse. Comme la version 4.1 de MySQL n'est pas compatible avec UTF-16, nous avons été obligé de stocker provisoirement les noms propres sous format binaire (Blob).

L'interface de travail sur Prolexbase a été développée en Java sous forme d'applet, car le langage Java intègre la norme Unicode et une applet peut s'exécuter sur tout navigateur Internet et tout système d'exploitation possédant une machine virtuelle java.

Le site de consultation a été entièrement développé en PHP car celui-ci présente deux avantages : le script PHP est exécuté uniquement par le serveur et cette exécution ne réclame aucune installation sur l'ordinateur du visiteur.

7.2 Site de consultation de Prolexbase

FIG. 7.1 – Page d'accueil de l'interface de consultation.

Le site de consultation de Prolexbase est accessible au public à partir de l'adresse suivante : `http://tln.li.univ-tours.fr/tln_prolex/prolex.php`.

Sur la page d'accueil (figure 7.1), il est possible de sélectionner la langue de l'interface : français ou anglais.

7.2.1 Le menu *Recherche*

Après avoir choisi la langue, l'utilisateur arrive sur la page de recherche des noms propres (figure 7.2). Il suffit de taper une instance d'un prolexème, d'un alias ou d'un dérivé et de cliquer sur le bouton *Rechercher* pour que le résultat de la recherche soit affiché. La figure 7.2 montre l'exemple du résultat obtenu en recherchant le nom propre *Paris*. Le programme a trouvé dans la base de données un seul nom propre correspondant à *Paris* et affiche aussi les informations sur celui-ci. En dessous une liste de noms propres contenant la séquence

Paris est affichée. Il est possible d'avoir des informations sur ceux-ci en cliquant sur leur libellé.

FIG. 7.2 – Recherche de noms propres.

Le visiteur peut aussi effectuer une recherche avancée (figure 7.3) dans la base de données en cliquant sur le lien *Recherche avancée*. Il est possible de faire une recherche sur des noms propres commençant, contenant et/ou finissant par une certaine séquence.

7.2.2 Le menu *Texte*

Ce menu (figure 7.4) permet au visiteur d'écrire, copier ou sélectionner depuis son ordinateur un texte puis de faire une recherche des noms propres dans ce texte en double-cliquant sur ceux-ci.

Par exemple, par un double-clic sur le nom propre *Paris* dans le texte, on obtient une fenêtre identique à celle de la figure 7.2 et en double-cliquant sur *luxembourgeois* la fenêtre de la figure 7.5 apparaît.

7.2.3 Autres menus

Le menu *Ontologie* fournit le schéma de l'ontologie de Prolexbase et le schéma d'un exemple. Le menu *Glossaire* donne les définitions des différents concepts utilisés pour décrire Prolexbase. Une description du projet est donnée dans le menu *Projet*. Les coordonnées du responsable du projet sont disponibles dans le menu *Contact*. Le menu *Commentaire* permet aux visiteurs de faire des remarques ou suggestions sur le site. Enfin, le menu *Remerciement* liste toutes les personnes ayant participé au projet.

7.3 Interface de travail

Elle est accessible à partir de l'adresse suivante : `http://tln.li.univ-tours.fr/tln_prolexbase/`.

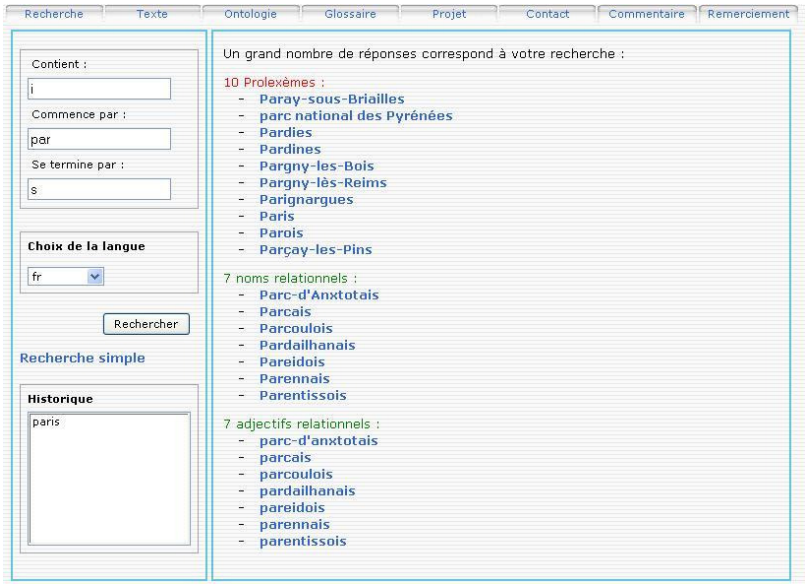

FIG. 7.3 – Recherche avancée.

FIG. 7.4 – Menu *Texte*.

FIG. 7.5 – Informations sur le mot *luxembourgeois*.

7.3.1 Les menus simples

Page d'accueil

Sur la page d'accueil (figure 7.6) l'utilisateur doit d'abord sélectionner la langue de l'interface. Trois langues sont disponibles pour l'interface : l'anglais, le français et le serbe.

Après avoir cliqué sur le bouton *OK*, les personnes ayant un compte doivent rentrer leur identifiant et leur mot de passe (figure 7.7), sinon ils peuvent cliquer sur le bouton *Visiter*. Seules les personnes ayant un compte peuvent travailler sur la base de données.

L'utilisateur doit choisir sa langue de travail, puis cliquer sur le bouton *Valider* (figure 7.8). Il peut à tout moment changer de langue de travail ou de consultation en revenant sur

FIG. 7.6 – Choix de la langue de l'interface.

FIG. 7.7 – Login et mot de passe.

FIG. 7.8 – Choix de la langue de travail.

l'onglet *Choix de la langue*.

Consultation

L'onglet *Consultation* (figure 7.9) comporte trois zones. Dans la zone gauche de l'onglet, l'utilisateur peut rechercher des prolexèmes de la base de données suivant différents critères. Les trois listes déroulantes permettent de faire une recherche de prolexèmes à partir de leur détermination, de leur flexion et de leur type. Il y a aussi la possibilité de faire une recherche de prolexèmes commençant par une certaine séquence, contenant une certaine séquence et/ou se terminant avec une certaine séquence. On peut interroger la base de données sur les derniers prolexèmes ajoutés en précisant un nombre dans le champs *derniers ajouts*. Enfin, une recherche suivant le numéro de pivot d'un prolexème est aussi possible.

Après avoir précisé les critères de recherche, l'utilisateur doit cliquer sur le bouton *Lancer la consultation* pour démarrer la recherche. Le programme affiche les résultats de la recherche dans la liste déroulante située en haut et au milieu de l'onglet. Par exemple, pour une recherche d'un prolexème commençant par *Franc*, le programme donnera les résultats suivants : *France*, *Franche-Comté*, *Francis Ford Coppola*, etc. En précisant que le type du prolexème est pays, on obtiendra seulement le prolexème *France*.

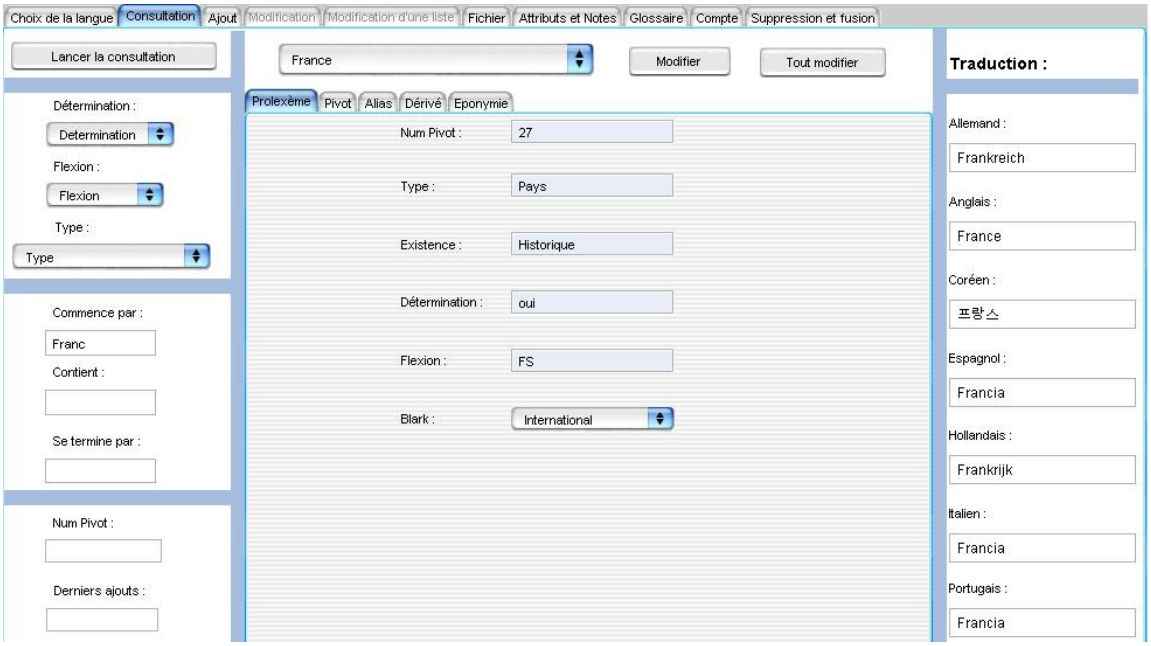

FIG. 7.9 – Onglet *Consultation*.

En sélectionnant un prolexème dans la liste déroulante en haut de la zone au milieu de la fenêtre, le programme affiche les informations sur celui-ci. Nous pouvons lire sur la figure 7.9 que le prolexème France a le numéro *27* comme pivot, son type est *pays*, son existence est *historique*, etc. L'onglet *Pivot* de la zone au milieu de la fenêtre renseigne sur les relations (méronymie, synonymie et accessibilité) que ce prolexème entretient avec d'autres prolexèmes. Les alias et les dérivés sont affiché dans l'onglet *Alias* et *Dérivés*. L'onglet *Eponymie* renseigne sur les expressions idiomatiques, les antonomases et la terminologie du prolexème.

Des traductions du prolexème sélectionné sont affichées dans la zone droite de l'onglet.

Dans cet exemple, la traduction du prolexème *France* en espagnol est *Francia*.

En cliquant sur le bouton *Modifier*, l'utilisateur pourra modifier des informations sur le prolexème sélectionné. Le bouton *Tout modifier* permet de changer les informations des prolexèmes contenus dans la liste déroulante de la zone du milieu.

Si le prolexème recherché n'existe pas dans la base de données, l'utilisateur pourra l'ajouter en utilisant l'onglet *Ajout* de l'applet.

Ajout

Ce menu (figure 7.10) permet d'ajouter un nom propre dans la base, ainsi que ses relations avec d'autres noms propres (figure 7.11), ses alias (figure 7.12), ses dérivés (figure 7.13), ses éponymes (figure 7.14), et la source d'où a été extrait celui-ci (figure 7.15).

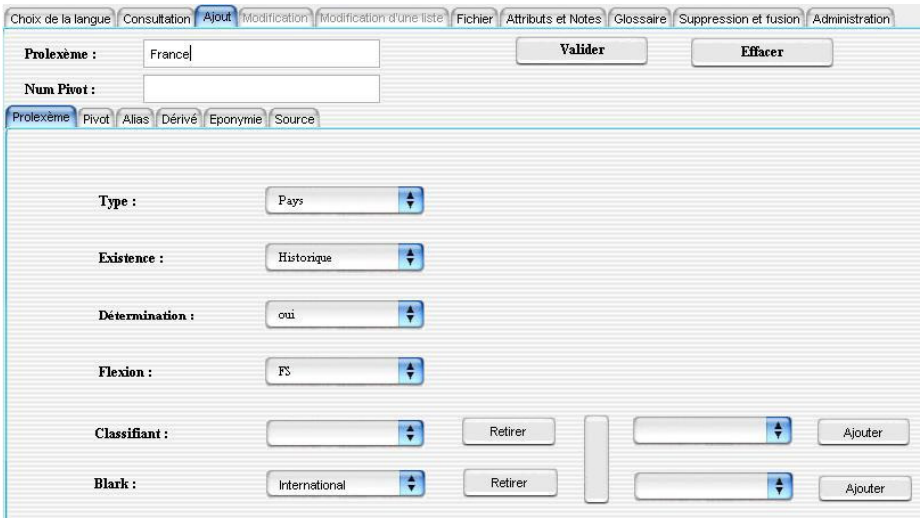

FIG. 7.10 – Onglet *Ajout*.

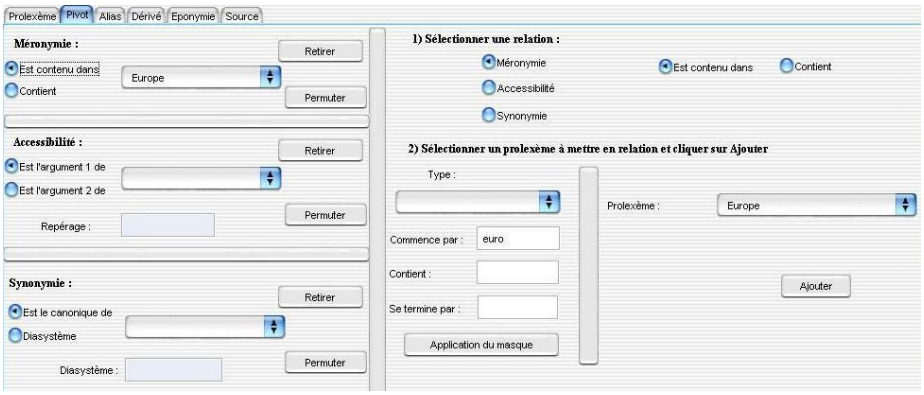

FIG. 7.11 – Onglet *Pivot*.

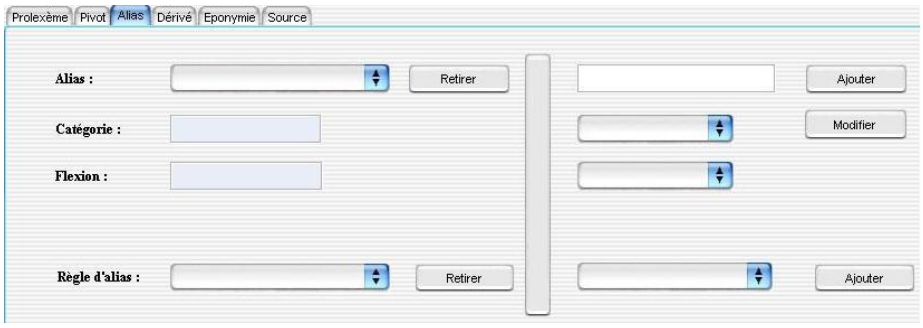

FIG. 7.12 – Onglet *Alias*.

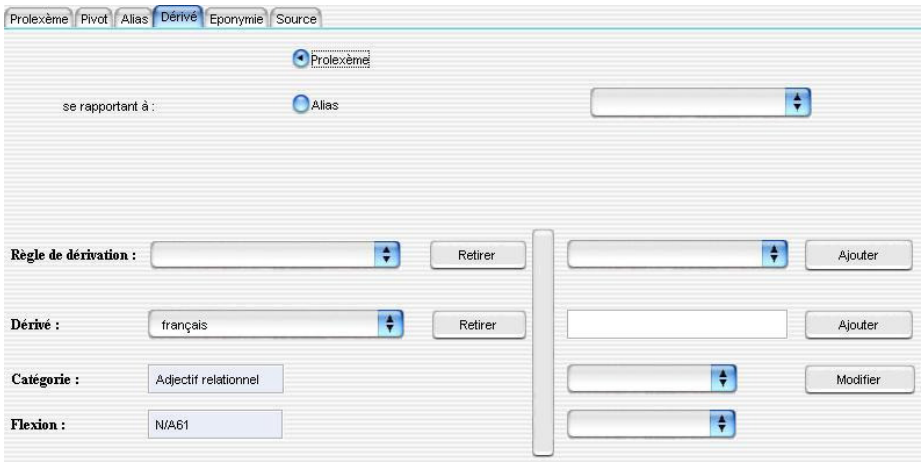

FIG. 7.13 – Onglet *Dérivé*.

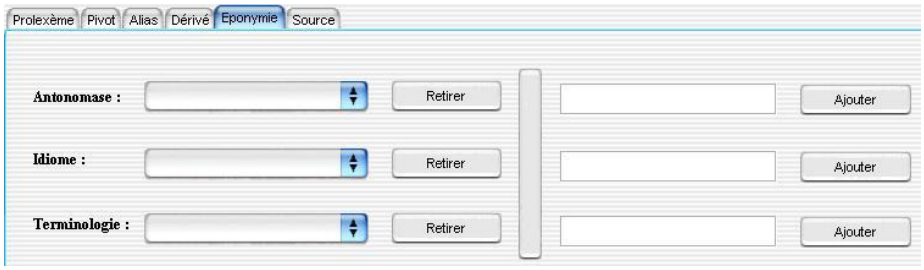

FIG. 7.14 – Onglet *Eponymie*.

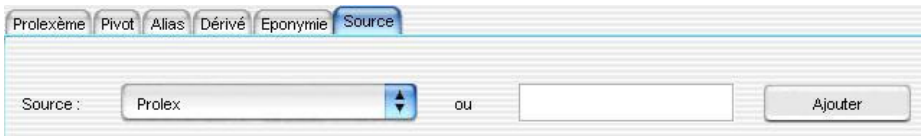

FIG. 7.15 – Onglet *Source*.

Modification

Ce menu (figure 7.16) permet de modifier le libellé d'un prolexème et d'éditer ses informations et ses relations.

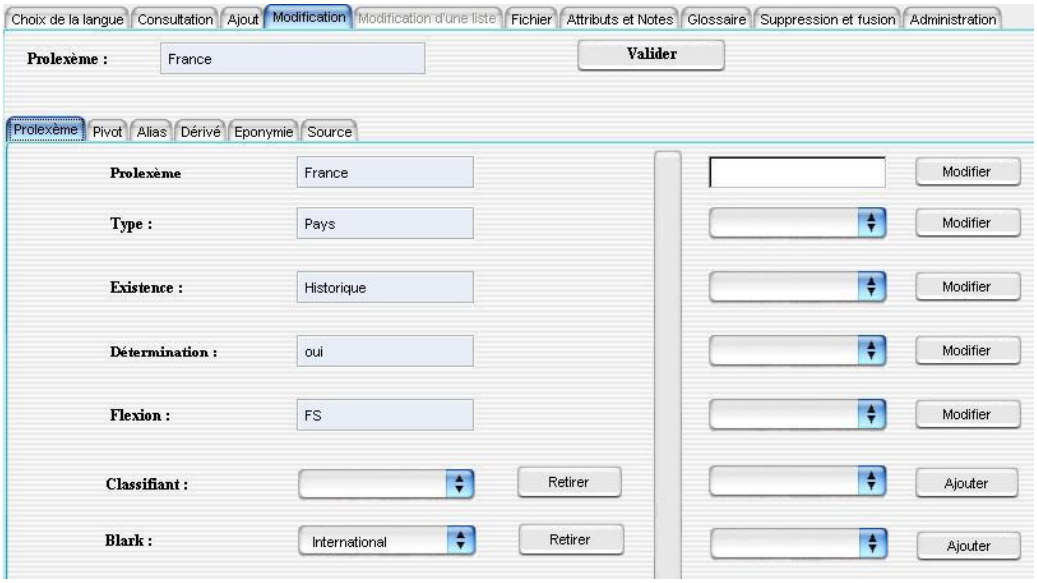

FIG. 7.16 – Onglet *Modification*.

Modification d'une liste

Ce menu (figure 7.17) permet d'éditer des informations et relations d'une liste de prolexèmes. Par exemple, si on sélectionne le classifiant *département* et le code flexionnel *FP*, ils seront associés à chaque prolexème de la liste.

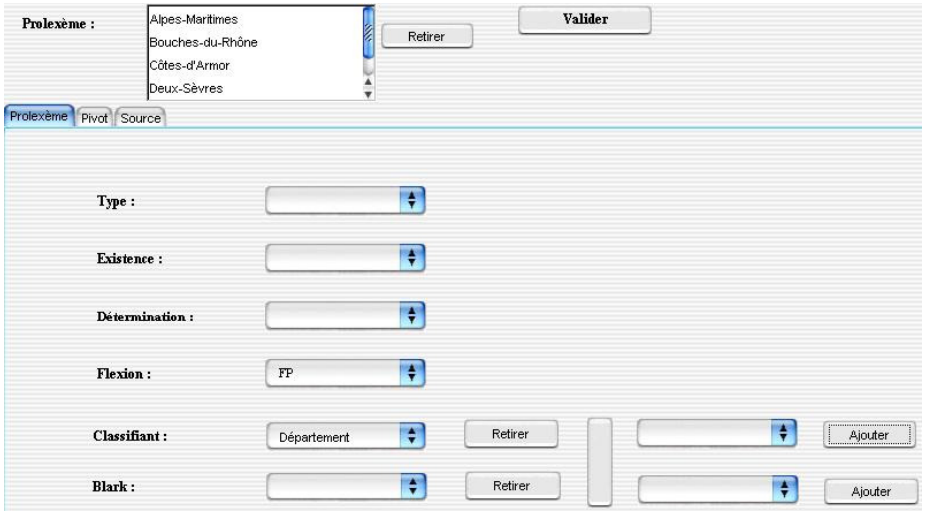

FIG. 7.17 – Onglet *Modification d'une liste*.

7.3.2 Le menu fichier

La plupart de nos collaborateurs possèdent des listes de noms propres sous forme de fichiers. Rentrer tous ces noms propres un par un en utilisant l'onglet *Ajout* risque de prendre beaucoup de temps. Pour éviter cela, nous avons développé dans l'onglet *Fichier* des outils pour permettre de travailler facilement et efficacement à partir de fichiers.

La première fonction proposée par cet onglet permet aux utilisateurs d'insérer un fichier contenant une liste de noms propres avec leurs informations. Ce fichier doit être enregistré au format Unicode UTF-16. La figure 7.18 donne un exemple de format de fichier. Ce fichier est organisé en colonnes. Une tabulation permet de séparer les données de chaque colonne. Dans cet exemple, la première colonne correspond aux prolexèmes, la seconde à la détermination, la troisième au type, la quatrième à la flexion, la cinquième aux dérivés et à la dernière la flexion des dérivés. La première ligne du fichier indique que le nom propre *France* se construit avec un article, son type est *Pays*, son code de flexion est *FS* et il a pour dérivé *Français*, dont le code de flexion est *N/A61*.

France	Oui	Pays	FS	Français	N/A61
Paris	Non	Ville	MFS	Parisien	N/A41
Belgique	Oui	Pays	FS	Belge	N/A31
Bruxelles	Non	Ville	MFS	Bruxellois	N/A61

FIG. 7.18 – Exemple de fichier.

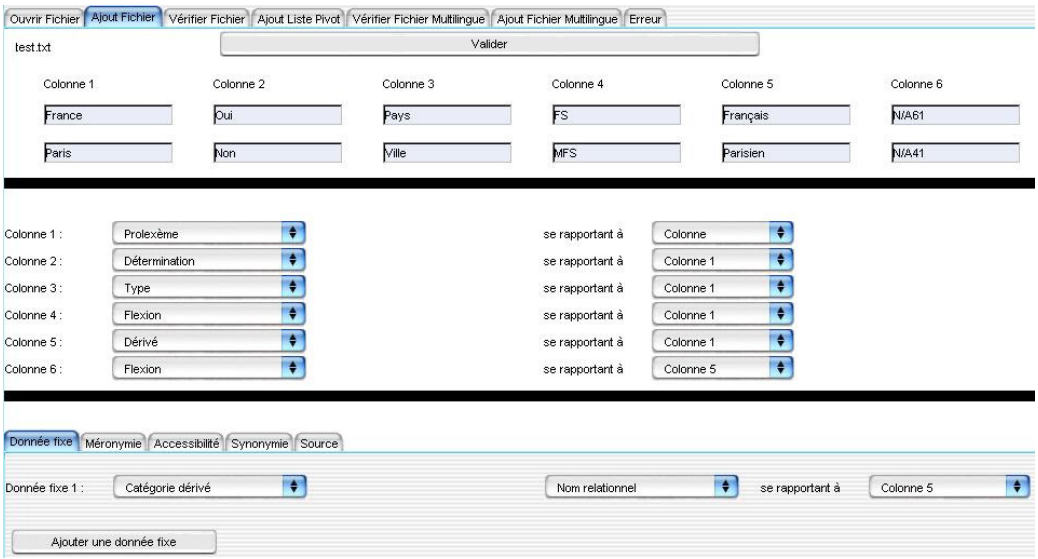

FIG. 7.19 – Ajout d'un fichier.

La figure 7.19 donne un aperçu de l'onglet *Ajout Fichier*. L'onglet comporte trois parties. La première partie, située en haut de l'onglet, affiche les deux premières lignes du fichier sur lequel on travaille. Dans la partie du milieu, l'utilisateur va pouvoir préciser la signification de chaque colonne. Si nous reprenons notre exemple, la colonne 1 correspond au prolexème, la colonne 2 à la détermination du prolexème de la colonne 1, la colonne 3 à son type, la colonne 4 à sa flexion, la colonne 5 au dérivé du prolexème et la dernière colonne à la flexion

du dérivé de la colonne 5. Pour éviter d'avoir une colonne avec une donnée unique, nous avons rajouté dans la dernière partie de cet onglet la possibilité d'affecter une valeur unique à une colonne. Par exemple, nous savons que tous les dérivés de la colonne 5 sont des noms relationnels. Nous allons déclarer qu'une donnée fixe correspondra à la catégorie des dérivés. Elle aura comme valeur *Nom relationnel* et se rapportera à la colonne 5. Il suffit, enfin, de cliquer sur le bouton *Valider* pour insérer les données dans Prolexbase.

L'onglet *Ajout Fichier* permet aussi d'ajouter des relations de méronymie, de synonymie ou d'accessibilité entre les noms propres d'un même fichier. La figure 7.20 donne un exemple d'ajout d'une relation d'accessibilité entre les prolexèmes de la colonne 2 avec ceux de la colonne 1 (*Berlin* est la capitale de l'*Allemagne*, *Lisbonne* est la capitale du *Portugal*, etc.).

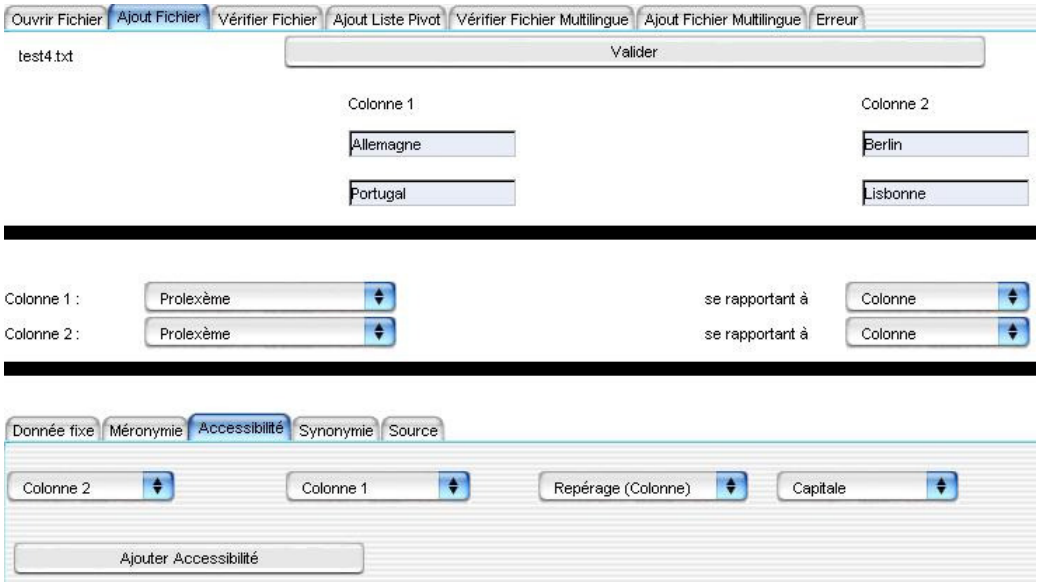

FIG. 7.20 – Relation de méronymie.

27	France	France	Frankreich	Francia	Francia	França	Frankrijk
38558	Paris	Paris	Paris	Parigi	París	Paris	Parijs
45883	Belgique	Belgium	Belgien	Belgio	Bélgica	Bélgica	België
45884	Bruxelles	Brussel	Brüssel	Brussel	Bruselas	Bruxelas	Brussel

FIG. 7.21 – Exemple de fichier multilingue.

Il arrive souvent que plusieurs personnes travaillent dans une même langue. Pour éviter qu'une personne ne rentre des noms propres qui existent déjà dans la base de données, nous avons créé l'onglet *Vérifier fichier* pour contrôler si une liste de prolexèmes existe déjà dans la base de données puis ajouter uniquement ceux qui ne sont pas présents. Cet onglet est pratiquement identique à l'onglet *Ajout Fichier*.

Une autre fonction du menu fichier permet de récupérer les numéros de pivot d'une liste de prolexèmes que l'on a déjà rentrée dans la base de données. Le programme effectue une recherche sur les prolexèmes de la base et renvoie le numéro de pivot si le prolexème existe. Pour les prolexèmes s'écrivant de la même manière, le programme renverra plusieurs numéros de pivot. L'utilisateur récupère les résultats du programme dans un fichier.

Les utilisateurs peuvent modifier les informations relatives aux prolexèmes de la base

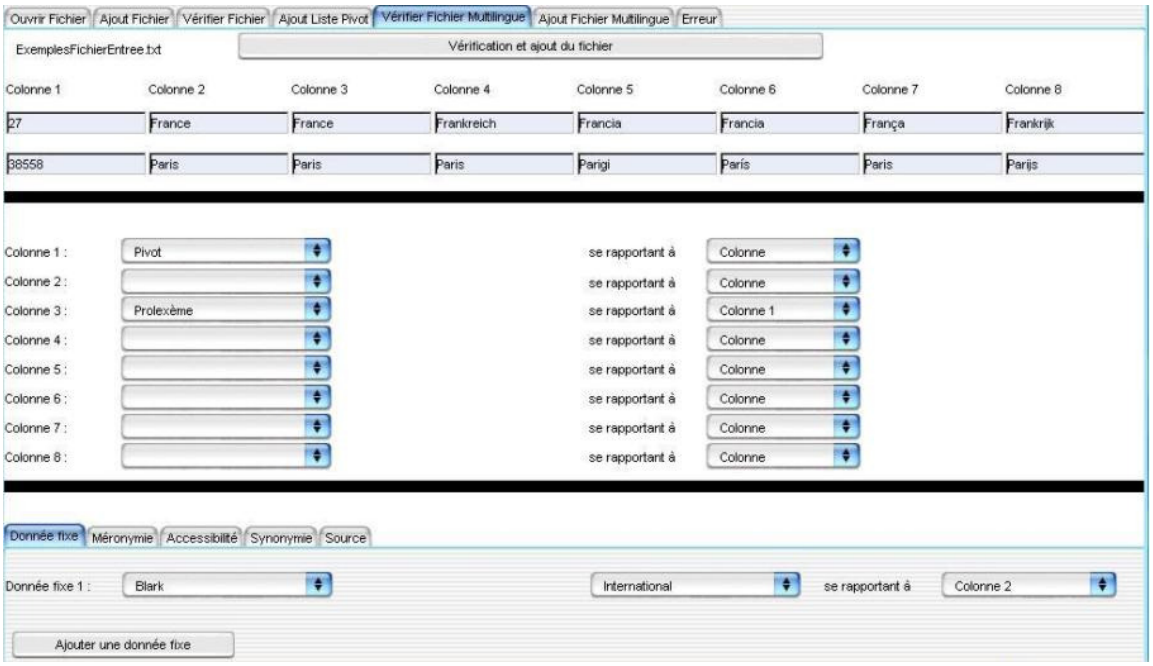

FIG. 7.22 – Traduction de prolexèmes dans une autre langue.

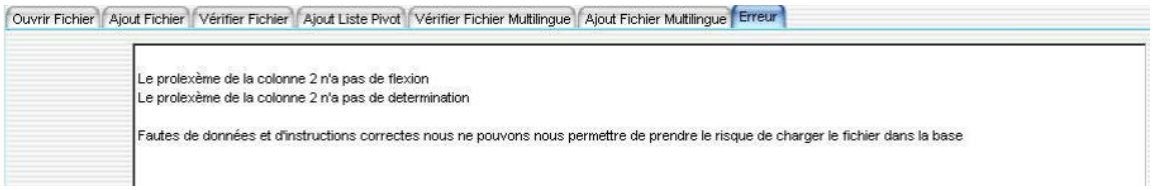

FIG. 7.23 – Les erreurs.

à partir d'une liste de numéros de pivot. Pour cela, ils doivent indiquer au programme la colonne du fichier correspondant aux numéros de pivot et sélectionner les informations qu'ils souhaitent modifier.

Les utilisateurs possédant un fichier de prolexèmes avec leur numéro de pivot peuvent ajouter dans la base la traduction de ceux-ci vers une autre langue. La figure 7.21 donne un exemple de prolexèmes français (colonne 2) avec leur numéro de pivot (colonne 1) traduit en anglais (colonne 3), en allemand (colonne 4), etc. Ils doivent d'abord sélectionner dans le menu fichier la langue vers laquelle ils souhaitent traduire et ensuite indiquer au programme le fichier de travail. Dans la figure 7.22, l'utilisateur doit préciser que la colonne 1 correspond aux numéros de pivot et la colonne 3 est la traduction des prolexèmes français en anglais. Le programme vérifie s'il n'existe pas de prolexème anglais en lien avec ces numéros de pivot et crée dans ce cas-là les prolexèmes anglais correspondants. Pour ajouter la traduction vers l'allemand, l'utilisateur doit revenir au menu fichier et changer la langue de traduction.

Les messages d'erreurs sont affichés dans l'onglet Erreur (figure 7.23).

7.3.3 Les menus d'administration

Attributs et notes

Ce menu (figure 7.24) comporte trois onglets. L'onglet *Variable* permet aux utilisateurs de définir les libellés des données qui sont spécifiques à chaque langue (comme les règles de flexions, les libellés de Blark, les catégories d'alias...). L'onglet *Fixe* permet de définir les données qui sont communes pour toutes les langues (comme les types, l'existence...). L'onglet *Interface* permet à l'utilisateur de traduire la langue de l'interface, c'est-à-dire les messages et textes affichés dans l'applet, vers une autre langue.

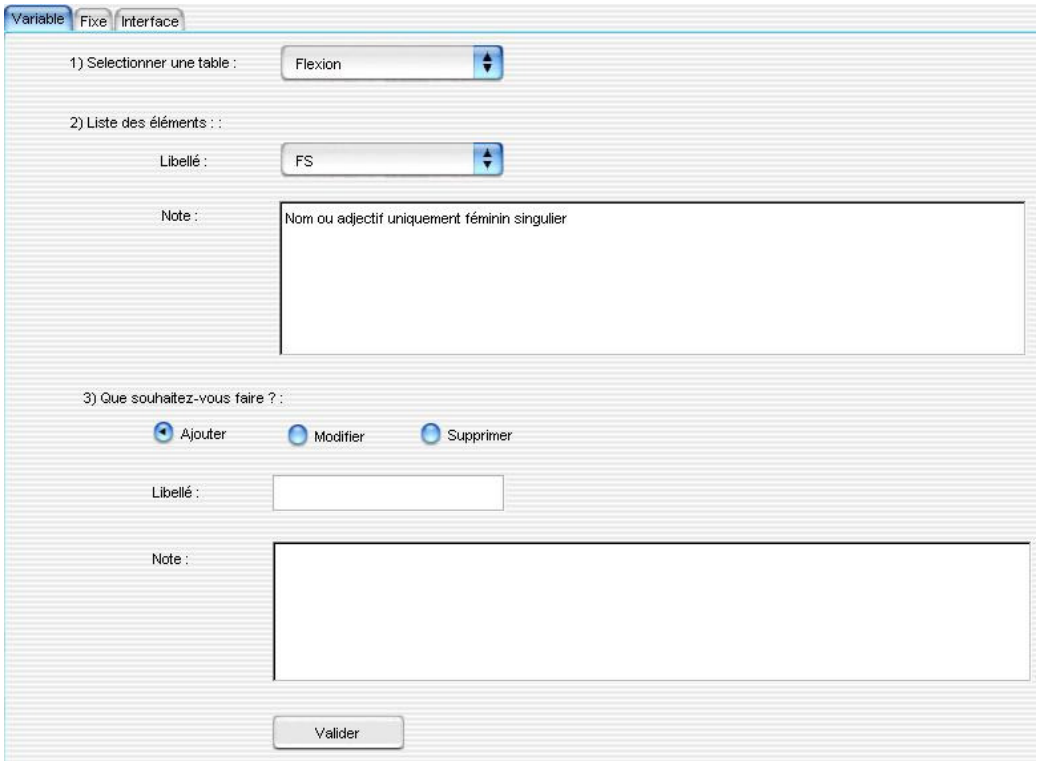

FIG. 7.24 – Onglet *Attributs et Notes*.

Compte

L'onglet *Compte* (figure 7.25) permet de créer des comptes pour les personnes qui travaillent sur Prolexbase. Il faut préciser pour chaque compte le nom, le prénom, l'identifiant, le mot de passe, les dates de validité du compte, les droits du compte et la langue dans laquelle le compte peut travailler. Nous avons défini trois types de droits :

- *A* : il s'agit d'un compte administrateur qui possède tous les droits sur Prolexbase. L'administrateur peut travailler dans toutes les langues de la base de données. Il peut créer tout type de compte.
- *C* : ce compte est destiné au chef de projet. Un chef de projet peut travailler dans une seule langue, dans laquelle il possède tous les droits y compris celui de créer des comptes utilisateurs sans pouvoir.
- *U* : ce compte est destiné aux utilisateurs sans droits, qui ne peuvent pas créer de compte.

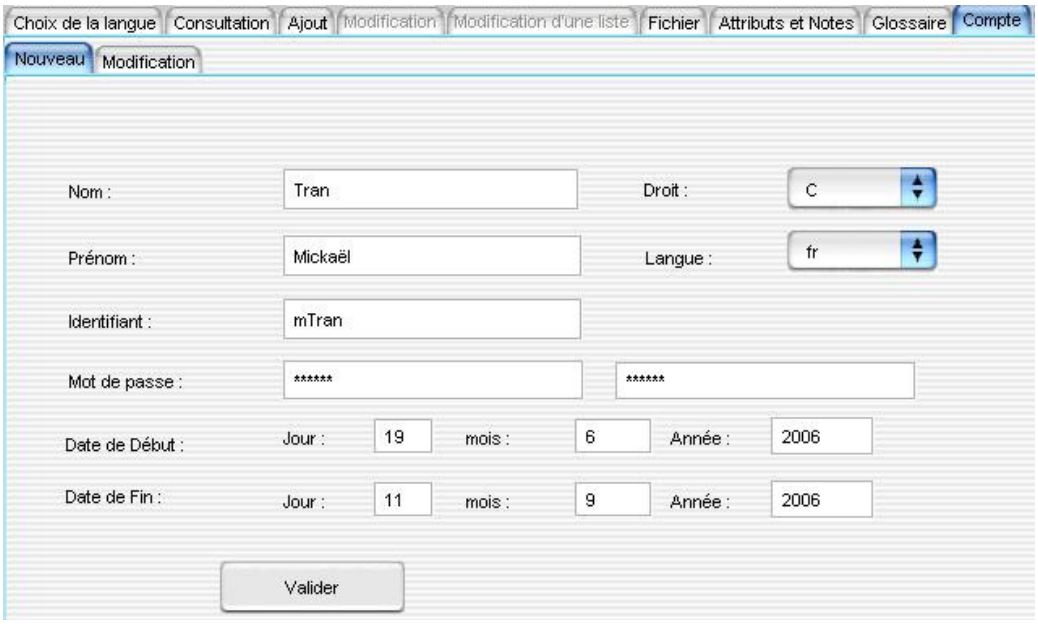

FIG. 7.25 – Onglet *Compte*.

7.3.4 Suppression et fusion

L'onglet *Suppression et fusion* permet de supprimer ou de fusionner des prolexèmes de la base de données. La suppression est possible :

- dans toutes les langues pour les administrateurs.
- dans leur langue pour les chefs de projet.
- dans leur langue de travail pour les utilisateurs. Tant qu'un autre utilisateur ne s'est pas connecté dans sa langue de travail, l'utilisateur pourra supprimer les données qu'il a créées. Cela permet d'éviter qu'il ne supprime des données qui ont été utilisées par d'autres.

Lors de la fusion de deux prolexèmes P_1 et P_2 d'une langue L, deux cas peuvent survenir :

- Il n'existe pas de langue L' dans laquelle P_1 et P_2 se traduisent par P'_1 et P'_2. Dans ce cas, la fusion se fait automatiquement.
- Les deux prolexèmes P_1 et P_2 possèdent tous les deux une traduction dans une autre langue L' (P'_1 et P'_2). La fusion nécessite l'accord du chef de projet de la langue L', sauf si les deux prolexèmes P'_1 et P'_2 et si toutes les informations associées sont identiques.

7.4 Calcul de complexité

Nous présentons dans cette partie quelques calculs de complexité sur le menu *Consultation* et le menu *Ajout fichier* de l'interface de travail collaboratif.

Menu Consultation

Soit $\alpha(T)$ le coût d'une requête SQL sur la table T. La complexité d'une requête dans le menu *Consultation* se calcule de la façon suivante :

$$O(\alpha(P) + a\ \alpha(A) + d\ \alpha(D) + m\ \alpha(M) + s\ \alpha(S) + c\ \alpha(C))$$

Avec :
- $\alpha(P)$: coût d'une requête SQL de type *SELECT* de recherche d'un prolexème suivant différents critères (flexion, commence par, type, etc.).
- a : le nombre d'alias que possède ce prolexème.
- $\alpha(A)$: coût d'une requête SQL de type *SELECT* pour récupérer des informations sur les alias.
- d : le nombre de dérivés.
- $\alpha(D)$: coût d'une requête SQL de type *SELECT* pour récupérer des informations sur les dérivés.
- m : le nombre de pivots en relation de méronymie avec le prolexème de la recherche.
- $\alpha(M)$: coût d'une requête SQL de type *SELECT* pour récupérer des informations sur les relations de méronymie.
- s : le nombre de pivots en relation de synonymie avec le prolexème de la recherche.
- $\alpha(S)$: coût d'une requête SQL de type *SELECT* pour récupérer des informations sur les relations de synonymie.
- c : le nombre de pivots en relation d'accessibilité avec le prolexème de la recherche.
- $\alpha(C)$: coût d'une requête SQL de type *SELECT* pour récupérer des informations sur les relations d'accessibilité.

Actuellement le nombre d'éléments des tables alias, synonymie et accessibilité est négligeable par rapport au nombre d'éléments des autres tables (voir section 8.4 page 144). La complexité devient alors :

$$O(\alpha(P) + d\ \alpha(D) + m\ \alpha(M))$$

Supposons que $\alpha = \alpha(P) \approx \alpha(D) \approx \alpha(M)$; comme d est négligeable par rapport à m, nous obtenons alors :

$$O(m\ \alpha)$$

Menu Ajout fichier

Nous allons calculer la complexité de ce menu sur un fichier de n prolexèmes (sans alias, ni dérivés). Ce fichier comporte donc n lignes et cinq colonnes (prolexème, type, existence, flexion et détermination).

Lors de l'ajout d'un fichier, l'utilisateur doit obligatoirement préciser au programme le contenu de chaque colonne. Par exemple, la colonne numéro un correspond à un prolexème, la colonne numéro deux, à son type, etc. Le programme commence par parcourir chaque colonne pour récupérer sa signification. La complexité de cette boucle est $O(1)$.

Il vérifie si les données du fichier (type, existence, flexion, etc.) correspondent bien à celles de la base de données. Par exemple, si l'utilisateur précise que la colonne numéro deux correspond à des types, le programme vérifie si la valeur de la colonne existe déjà dans la base de données. En cas de valeur non valide, c'est-à-dire non présente dans la base de données, le programme avertit l'utilisateur de l'erreur. Étant donné la taille des tables (30 types, 3 existences, 38 flexions et 2 déterminations), nous supposons que le coût de ces vérifications dans la base de données est quasiment identique pour chaque table et égal à β. La complexité de cette vérification est donc $O(n\ \beta)$.

Une fois que la vérification a été faite, le programme ajoute ligne par ligne les données du fichier dans la base de données. Le programme lance une requête SQL qui ajoute le prolexème et ses informations. Soit γ le coût de cette requête. La complexité de cette partie est $O(n\ \gamma)$.

La complexité totale du menu ajout fichier est donc de :

$O(n\ \beta + n\ \gamma)$

Si nous supposons que $\beta \approx \gamma$, la complexité totale du menu ajout fichier devient alors $O(n\ \beta)$.

Quatrième partie

Synthèse

Chapitre 8

Évaluation

8.1 Le modèle

Dans cette partie, nous présentons une évaluation de la modélisation des noms propres que nous avons proposée et nous en discutons quelques limites.

8.1.1 Prolexème et forme vedette

Théoriquement, nous aurions dû considérer le prolexème comme un identificateur : le couple pivot-langue. La table des alias aurait alors regroupé toutes les formes possibles du nom propre. Dans la pratique nous avons préféré définir le prolexème par une forme vedette, ce qui facilite la manipulation des données par des linguistes, en simplifiant l'accès au dictionnaire.

Il n'est cependant pas évident de la choisir. Parmi les noms propres *Organisation des Nations Unies*, *Nations Unies* et *ONU*, lequel devons-nous prendre comme forme vedette ? Et sur quel critère devons-nous le choisir ?

L'utilisation future de règles d'aliasisation et de dérivation nous a conduit à prendre la forme la plus longue comme prolexème. Nous pensons qu'il sera plus facile ainsi d'établir les règles de formation d'alias et de dérivés. Par exemple, il est plus évident de concevoir des règles pour former les noms propres *Nations Unies* et *ONU* à partir du nom propre *Organisation des Nations Unies* que d'utiliser les deux autres formes pour retrouver la première. Nous pourrions créer une règle qui consiste à effacer la partie générique *Organisation* pour obtenir le nom propre *Nations Unies* et une autre règle qui consiste à prendre les premières lettres de chaque mot plein pour générer le nom propre *ONU*. A la fin de cette thèse, ce travail sur la création de règles d'alias reste un projet.

Suivant les applications, les stratégies pourront être différentes. Dans la traduction de l'anglais vers le français, il faudrait probablement proposer *ONU* pour *UNO* alors que dans la recherche d'information le prolexème et tous ses alias seront également intéressants.

8.1.2 Date

Nous nous sommes posé la question de savoir si nous devions ajouter une date pour certaines relations, comme la synonymie et la méronymie.

Faut-il préciser la date dans une relation de synonymie diachronique ? Nous avons longtemps hésité sur cette question.

La ville de *Saint-Pétersbourg* a été fondée vers 1703. Elle a pris le nom de *Petrograd* de 1914 à 1924, puis de 1924 à 1991 elle a porté le nom de *Leningrad*. Elle a repris le nom

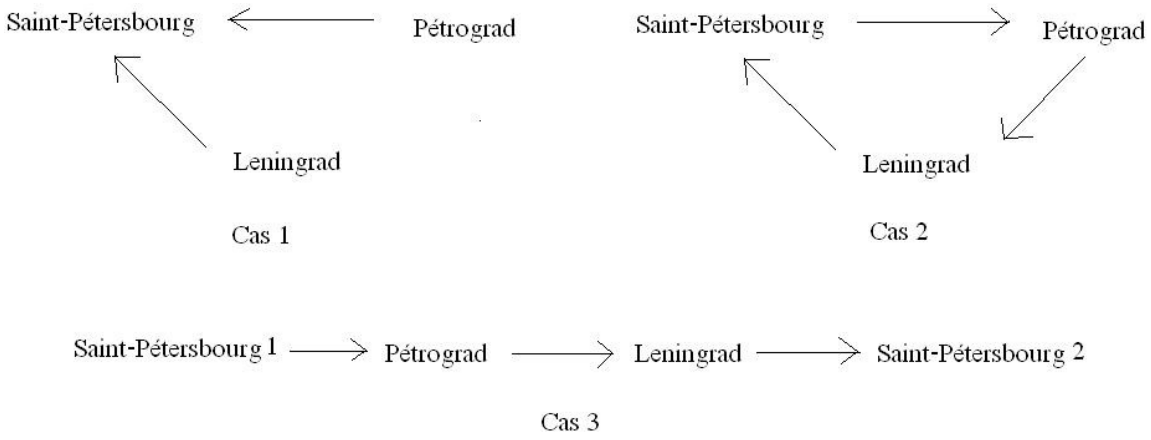

FIG. 8.1 – Exemple de cycle pour la relation de synonymie diachronique.

de *Saint-Pétersbourg* à partir de 1991. Ces relations peuvent se modéliser de trois façons différentes (voir figure 8.1). Le cas 2 représente une modélisation de ces trois relations de synonymie avec la présence d'une date. Nous avons un cycle entre ces noms propres. Il faudra peut-être préciser une date de fin et une date de début à laquelle un nom propre a été renommé. Comme notre but n'est pas de créer une encyclopédie, nous n'avons pas précisé de date pour la synonymie diachronique. De plus, imposer une date (ou deux dates ?) dans cette relation peut compliquer le travail de remplissage de la base, car l'utilisateur devra faire une recherche sur la date pour laquelle un nom propre a été renommé. Il s'agit plutôt d'un travail d'historien.

La présence de la date n'est pas utile pour toutes les applications. Pour des applications de traduction de textes datant de 1918, il faudrait plutôt proposer *Petrograd* comme traduction. Par contre, pour des applications de recherche d'information, la présence de la date n'a aucun intérêt. Si un utilisateur fait des recherches sur le nom propre *Saint-Pétersbourg*, l'application devrait non seulement lui fournir des textes ou apparaît ce nom propre mais aussi des textes ou apparaissent aussi les noms propres *Petrograd* et *Leningrad*. Si la présence de la date est nécessaire à une application donnée, nous pourrons toujours ajouter un attribut date dans la relation de synonymie diachronique dans le modèle conceptuel de données.

Le cas 3 représente une modélisation où l'on distingue la ville de *Saint-Pétersbourg* à sa création (*Saint-Pétersbourg1*) et celle d'aujourd'hui (*Saint-Pétersbourg2*), sans préciser de date. Nous n'avons pas retenu ce cas, car il suppose une duplication inutile dans notre base de données. Dans notre modélisation, nous ne considérons qu'une ville de *Saint-Pétersbourg*.

Le cas 1 correspond à notre modélisation de ces relations où nous n'avons gardé qu'une forme canonique, *Saint-Pétersbourg*. Pour l'étude de textes de l'après-guerre, il serait possible de modifier le canonique qui deviendrait *Leningrad*. Pour les textes actuels, *Petrograd* renvoie directement à *Saint-Pétersbourg*.

Le problème de la date se pose aussi dans le cadre d'une relation de méronymie.

La Bretagne est-elle est en France ? L'Alsace et la Lorraine sont-elles en France ? Selon la date, les réponses à ces questions peuvent varier. L'importance numérique de la méronymie rend cette information impossible.

Avant le référendum du 5 juin 2006, nous avions dans notre base de données les noms propres : *Serbie et Monténégro* (type Pays), *Serbie* (type Région) et *Monténégro* (type

Région). Suite au référendum, nous avons modifié les entrées *Serbie* et *Monténégro* en changeant leur type Région en Pays. Cet exemple justifie la création de notre supertype Territoire. Si nous avions associé directement aux noms propres *Serbie* et *Monténégro* ce supertype, nous n'aurions pas eu besoin de faire de modification.

8.1.3 Synonymie et forme canonique

Dans le modèle conceptuel de données, nous avons précisé que la relation de synonymie relie une forme canonique (forme ayant la plus grande notoriété) et une forme synonyme (forme moins connue). Un nom propre peut être la forme canonique de plusieurs autres noms propres et chaque nom propre peut avoir au plus une seule forme canonique. Le modèle conceptuel de données n'interdit pas les cycles dans une telle relation. Par exemple, si un nom propre P_1 est le canonique d'un nom propre P_2 et si P_2 est le canonique d'un nom propre P_3, il est possible que P_3 soit le canonique de P_1 (voir figure 8.2). Ce cas peut poser des problèmes lors de la recherche d'une forme canonique, car en partant du nom propre P_1 nous risquons de retomber sur celui-ci. Pour interdire les cycles, il faut créer une fonction de vérification des cycles dans l'interface de travail lors de la saisie des relations de synonymie. Actuellement, nous n'avons pas rencontré de noms propres vérifiant ce cas dans la base de données.

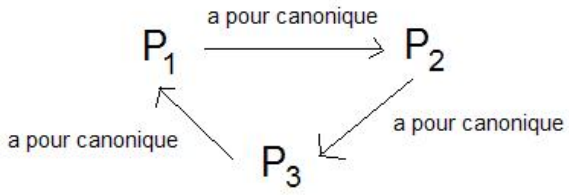

FIG. 8.2 – Exemple de cycle pour la relation de synonymie.

Soit le nom propre P_1 qui est le canonique de deux autres noms propres P_2 et P_3. Si les noms propres P_2 et P_3 sont en relation de synonymie avec le nom propre P_1 suivant un même diasystème, notre modèle ne propose alors aucun ordre entre P_2 et P_3. Pour différencier entre P_2 et P_3, il faudrait peut-être associer à chacun de ces noms propres des informations statistiques (voir table *STATISTIQUE* dans la section 5.2.5 du chapitre 5) pour indiquer leur notoriété.

8.1.4 Les relations d'hyperonymie et de méronymie

La méronymie, la synonymie et la relation d'accessibilité sont les seules relations qui relient des noms propres entre eux. Nous nous sommes posé la question de savoir si nous pouvions trouver une relation d'hyperonymie entre des noms propres.

Dans le modèle, nous avons des relations d'hyperonymie entre des noms propres et des types. Par exemple, le *Jardin des Plantes* est en relation d'hyperonymie avec le type Édifice.

Nous avons essayé de chercher des exemples d'hyperonymie entre des noms propres, mais cela semble être un phénonème rare. Voici les exemples que nous avons trouvés :

- Pouvons-nous dire qu'une *Mégane II Estate* est une *Mégane* ?
- Pouvons-nous dire qu'une *Mégane* est une *Renault* ?
- Pouvons-nous dire que *Charlemagne* est un *Carolingien* ?
- Pouvons-nous dire que *Georges Bush* est un *Bush* ?

Dans les deux premières questions, *Mégane II Estate* est un nom de produit, *Mégane* est un nom de gamme et *Renault* est un nom de marque. Nous ne faisons pas cette distinction dans notre typologie, car nous souhaitions avoir une typologie réduite. Ces trois noms sont classés dans le type Produit. Peut-on dire que les noms de produits, de gammes et de marques sont des noms propres ? Il n'est pas évident de répondre à cette question. Peut-on alors dire qu'il s'agit de noms communs ? Là aussi, la question reste ouverte. Il s'agit de noms se trouvant à la frontière entre les noms propres et les noms communs. Nous les avons considérés dans notre modèle comme des noms propres.

Dans les deux dernières questions, nous avons une relation entre des célébrités et leur dynastie. Il s'agit de cas limites et discutables. Pour le cas des noms de dynastie, cette relation d'hyperonymie s'applique bien. Mais est-ce que les noms de dynastie sont des noms propres ?

Nous pouvons aussi dire que le nom propre *Mégane* est en relation de méronymie avec le nom propre *Renault*. De même pour les noms propres *Charlemagne* et *Carolingien*, où nous avons une relation de méronymie entre un membre et une collection. Étant donné la rareté de la relation et son statut parfois discutable, nous avons décidé de considérer ces cas comme des relations de méronymie.

8.1.5 La relation d'accessiblité

Les relations classiques, telles que la méronymie et la synonymie, ne sont pas les seules relations qui relient les noms propres. Il existe d'autres relations entre les noms propres :

- *Abel* est un fils d'*Adam*
- *Wilbur Wright* est un frère de *Orville Wright*
- *Alcibiade* est un élève de *Socrate*
- *Pierre* est un apôtre de *Jésus-Christ*
- *Platon* est le fondateur de l'*Académie*
- *Pierre de Coubertin* est l'inventeur des *Jeux Olympiques*
- *Bram Stoker* est l'auteur de *Dracula*
- *Abd Allah II* est un roi de la *Jordanie*
- *Seti Ier* est un pharaon de l'*Égypte antique*
- *Mehmed II* est un sultan de *Turquie*
- *Anne Stuart* est une reine d'*Angleterre*
- *Copenhague* est la capitale du *Danemark*
- *Pyongyang* est la capitale nationale de la *Corée du Nord*
- *Lyon* est le siège d'*Interpol*
- ...

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.
...	

FIG. 8.3 – Repérages et expansions.

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 décidé de les regrouper dans une seule et unique relation que nous avons appelée relation d'accessibilité, à laquelle nous avons ajouté des repérages généraux (voir figure 8.3). 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.

8.2 L'interface de travail de Prolexbase

L'interface de travail sous la forme d'une applet pose un certain nombre de problèmes :
- il faut installer la machine virtuelle java pour pouvoir exécuter une applet.
- une applet est exécutée du côté du client. Parfois, des règles au niveau du pare-feu chez le client peuvent empêcher l'applet de se connecter à la base de données se trouvant dans le laboratoire LI de l'université de Tours.

Pour résoudre ces problèmes, il faudrait développer une nouvelle interface en utilisant JSP (Java Server Pages). JSP nécessite l'utilisation d'un serveur Tomcat, dont nous ne disposons pas dans notre laboratoire.

Le travail de remplissage de Prolexbase ne peut se faire que par des spécialistes car l'utilisation de l'interface de travail collaboratif peut être compliquée pour les utilisateurs ne connaissant pas notre projet. Pour pouvoir l'utiliser correctement, chaque utilisateur doit obligatoirement connaître la structure de notre modèle et les termes employés. Il faut envisager de mettre en place une formation détaillée sur les fonctions de l'interface de travail pour leur permettre d'utiliser efficacement l'interface.

Les points forts de l'interface sont sans doute les fonctions de travail sur les fichiers. Nos collaborateurs possèdent des données sous forme de fichiers. Ces fonctions leur permettent de faciliter le travail de remplissage de la base de données. De plus, le menu fichier accélère considérablement la phase de remplissage, car il est plus rapide et économique d'ajouter en une seule fois un fichier comprenant une centaine de noms propres avec leurs informations en utilisant le menu fichier que d'ajouter un par un chaque nom propre avec ses informations en utilisant le menu ajouter.

Ce travail de remplissage se fait en deux étapes. La première étape consiste à travailler sur un fichier sous format Unicode en utilisant un tableur (Excel, Calc, etc.). Durant cette première phase, l'utilisateur n'a pas besoin de se connecter à l'interface de travail. Il peut travailler chez lui et sous n'importe quel système d'exploitation. La deuxième phase consiste à se connecter à l'interface de travail et à utiliser le menu fichier pour préciser les données qu'il souhaite intégrer à Prolexbase.

8.3 Analyse quantitative

Au début de la phase de remplissage, nous nous sommes posé la question de savoir quels noms propres nous devions rentrer dans notre base de données et comment les classer suivant un critère de notoriété. Sur quels critères de notoriété devons-nous sélectionner ou non un nom propre ? Nous avons présenté à une étudiante de Licence de Lettres une

liste de noms de personnages célèbres. Nous lui avons demandé de sélectionner ceux qu'elle connaissait. En lisant sa liste, nous nous sommes rendu compte qu'elle connaissait des noms de peintres, d'auteurs, de philosophes, etc. que nous ne connaissions pas et que les noms de scientifiques dans sa liste étaient peu nombreux. Selon la culture et le parcours de la personne, nous obtenons des listes très différentes. Or ce travail a déjà été fait par les éditeurs de dictionnaires.

Nous avons décidé de prendre tous les noms propres du dictionnaire Larousse Collège et de les considérer comme les noms propres que tout français est censé connaître. Ce dictionnaire précise sur sa couverture qu'il comporte environ 6 000 noms propres. Ce travail nous a permis de vérifier et de tester la pertinence de notre modèle en traitant manuellement les 6 000 noms propres contenus dans le dictionnaire *Larousse Collège*.

La première partie du travail consistait à parcourir le dictionnaire pour sélectionner tous les noms propres. Les noms propres sont recopiés dans un fichier Excel. A chaque nom propre, nous associons un type, une règle de flexion, les relations qu'il entretient avec d'autres noms propres, ses alias, les expansions, la détermination, etc. La figure 8.4 présente une partie des données de ce fichier extrait du dictionnaire Larousse Collège.

Dét	Flexion	Nom	Type	Expansion	Alias	Catégorie alias	Essence
NON	MS	Alvar Aalto	Célébrité	architecte	!	!	historique
|	|	|	|	designer	!	!	historique
OUI	FS	Aar	Hydronyme	rivière	Aare	Variante	historique
NON	MS	Aaron	Célébrité	grand-prêtre	!	!	religieux
NON	MFS	Abadan	Ville	port	!	!	historique
NON	MS	Abbas Ier le Grand	Célébrité	!	!	!	historique
NON	MS	Abu al-Abbas Abd Allah	Célébrité	calife	!	!	historique
NON	MFS	Abbeville	Ville	ville	!	!	historique
NON	MS	Abd al-Aziz III Ibn Saud	Célébrité	!	Ibn Séoud	Transcription	historique
NON	MS	Abd Allah II	Célébrité	!	Abdallah II	Variante	historique

FIG. 8.4 – Extrait d'une partie du fichier de travail sur le Larousse Collège.

Nous avons rencontré quelques problèmes pour classer certains noms propres suivant un type. Notre typologie résulte des travaux du projet Prolex avec Thierry Grass. Nous avons hésité à garder les types Dynastie et Ethnonyme, car ces types ne nous semblaient peu pertinents. Nous avons considéré au début que les noms de Dynastie et d'Ethnonyme sont des dérivés, donc de les intégrer à des prolexèmes. Nous avons donc décidé de les supprimer de notre typologie. En travaillant sur le dictionnaire Larousse Collège, nous avons constaté que toutes les dynasties ne sont pas des dérivés d'un nom propre. Nous pouvons dire que Carolingien est un dérivé de *Charlemagne*, mais comment faire le lien entre *Louis II Le Bègue* et *Charlemagne*. L'accessibilité avec un repérage *Descendant* ne semblait guère convenir. De même pour les ethnonymes : Turc est un nom de nationalité, inclus dans le prolexème Turquie, mais il y a aussi des Turcs qui n'ont pas la nationalité turque. Donc Turc sera aussi une entrée du dictionnaire de type Ethnonyme. De plus, des ethnonymes comme *Sioux*, *Incas*, *Celtes* ne sont pas des dérivés. Nous avons donc conservé les types Dynastie et Ethnonyme à notre typologie.

Nous nous sommes posé la question de savoir si nous pouvions classer les noms suivants dans notre dictionnaire : *christianisme*, *hindouisme*, *islam*, *judaïsme*, *shintoïsme*, *New Age*, etc. S'agit-il de noms propres ? La question ne se pose pas pour le nom *New Age*, car il est considéré par le Larousse 2005 comme un nom propre. Il n'est pas évident de répondre à cette question pour les autres noms. Ces noms renvoient à un référent unique. Certains

Anthroponyme	4 043
Association	31
Célébrité	3 734
Dynastie	50
Ensemble	14
Entreprise	3
Ethnonyme	133
Institution	55
Organisation	20
Patronyme	0
Prénom	0
Pseudo Anthroponyme	3
Toponyme	2 755
Astronyme	24
Edifice	93
Géonyme	202
Hydronyme	295
Pays	230
Région	744
Supranational	47
Ville	1 106
Voie	14
Ergonyme	166
Objet	0
Œuvre	76
Produit	86
Vaisseau	4
Pragmonyme	216
Catastrophe	1
Fête	11
Histoire	200
Manifestation	3
Météorologie	1
Total	7 180

FIG. 8.5 – Nombre de prolexèmes extraits du Larousse Collège.

peuvent s'écrire avec ou sans majuscule. Ils sont en général considérés par le dictionnaire comme des noms communs. Comme ils respectent la définition de Jonasson, nous avons décidé de classer ces noms comme des noms propres et de créer un nouveau type absent de la typologie initiale que nous appellerons *Pensée* (voir figure 8.6). Nous n'avons pas défini de véritables critères de création de types. Nous souhaitons au contraire les limiter afin qu'ils soient le plus général possible. Mais de nouvelles créations sont possibles.

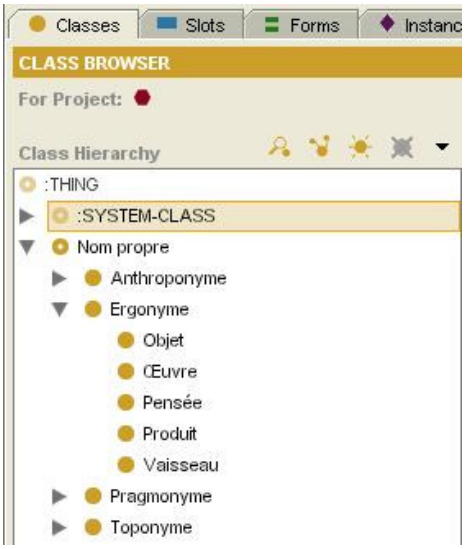

FIG. 8.6 – Le type Pensée.

Nous avons aussi réussi à classer tous les noms propres que nous avons trouvés dans le Larousse Collège suivant notre typologie. Ce qui représente une bonne validation de notre modèle.

Ce travail nous a permis de récupérer une liste de 7 180 noms propres. Le figure 8.5 présente leur répartition suivant notre typologie. Nous avons donc trouvé 1 180 noms propres de plus par rapport au nombre indiqué par le dictionnaire. Cette différence s'explique par plusieurs raisons :

- certaines célébrités possèdent deux noms. Par exemple *Molière* et *Jean-Baptiste Poquelin*. Le dictionnaire considère qu'il s'agit d'une unique entrée.
- les noms de personnes issues d'une même famille sont regroupés sous un même article.
- certains noms propres apparaissant dans les définitions d'un nom propre ne possèdent pas d'entrée dans le dictionnaire.

La deuxième partie du travail consistait à rentrer le fichier dans la base de données. Nous avons associé à ces noms propres le libellé "national" comme indicateur BLARK. Cette partie nous a permis de tester les fonctions du menu fichier et de faire leur mise au point.

8.4 Le contenu de Prolexbase

Les données de Prolexbase proviennent principalement de deux sources différentes : le dictionnaire Larousse Collège et les toponymes du projet Prolex[1]. Nous avons d'abord inséré

[1]Une bibliographie complète du projet Prolex se trouve sur le site suivant : `http://tln.li.univ-tours.fr/Tln_Bibliographie.html`.

dans la base de données les 49 732 prolexèmes du projet Prolex. Sur les 7 180 prolexèmes du dictionnaire Larousse Collège, 4 432 prolexèmes ont été ajoutés à la base de données et les 2 748 prolexèmes restants sont des toponymes qui sont déjà présents dans la base de données. En septembre 2006, la base de données contient 54 164 prolexèmes français. La figure 8.7 présente le nombre de prolexèmes pour la partie française en fonction de leur type.

Voici le nombre de relations (qui ne dépendent pas de la langue) que nous possédons dans Prolexbase :

– 641 liens de synonymie, dont 223 proviennent du dictionnaire Larousse Collège.
– 44 260 liens de méronymie, dont 6 235 proviennent du dictionnaire Larousse Collège.
– 2 244 liens d'accessibilité, 393 proviennent du dictionnaire Larousse Collège.

La base de données contient pour le français 493 alias. Voici la répartition de ces alias en fonction de leur catégorie :

– 143 abréviations, dont 123 proviennent du dictionnaire Larousse Collège.
– 31 acronymes ou sigles provenant du dictionnaire Larousse Collège.
– 6 acronymes ou sigles étrangers, dont 5 proviennent du dictionnaire Larousse Collège.
– 41 diastratiques, dont 2 proviennent du dictionnaire Larousse Collège.
– 3 diatopiques, dont 2 proviennent du dictionnaire Larousse Collège.
– 239 transcriptions provenant du dictionnaire Larousse Collège.
– 30 variantes, dont 15 proviennent du dictionnaire Larousse Collège.

Nous avons 20 609 dérivés, dont 30 proviennent du dictionnaire Larousse Collège.

En septembre 2006, nos collègues serbes ont inséré 606 prolexèmes.

Anthroponyme	4048
Association	32
Célébrité	3735
Dynastie	50
Ensemble	14
Entreprise	3
Ethnonyme	134
Institution	57
Organisation	20
Patronyme	0
Prénom	0
Pseudo Anthroponyme	3
Toponyme	49566
Astronyme	24
Edifice	93
Géonyme	205
Hydronyme	4348
Pays	398
Région	2627
Supranational	53
Ville	41804
Voie	14
Ergonyme	166
Objet	0
Œuvre	76
Produit	86
Vaisseau	4
Pragmonyme	216
Catastrophe	1
Fête	11
Histoire	200
Manifestation	3
Météorologie	1

FIG. 8.7 – Nombre de prolexèmes français de Prolexbase.

Conclusion

Bilan

Cette thèse est destinée à présenter les différentes étapes de la création d'un dictionnaire relationnel multilingue de noms propres. Elle est organisée en trois parties.

La première partie de nos travaux permet de répondre aux questions suivantes :

– qu'est-ce qu'un nom propre ?
– quelles informations devons-nous inclure dans notre dictionnaire de noms propres ?
– comment construire un dictionnaire multilingue ?

Une définition des noms propres nous semblait être indispensable pour commencer nos travaux et savoir quels noms nous devions ou non rentrer dans notre dictionnaire. Parmi les différentes définitions des linguistes et des dictionnaires, nous avons décidé d'adopter la définition de [Jonasson, 1994], car elle possède une couverture beaucoup plus large que les autres définitions.

L'étude détaillée des caractéristiques et des propriétés des noms propres nous a permis de réfléchir sur les informations que nous devions inclure dans notre dictionnaire. Il nous parait indispensable d'avoir des informations graphiques (précisant l'écriture des noms propres), syntaxiques (la détermination et ses constructions au sein d'une phrase) et flexionnelles. De plus, ce dictionnaire doit obligatoirement inclure des informations sémantiques et pragmatiques, c'est-à-dire des relations entre les noms propres.

Pour répondre à la dernière question, nous avons étudié les différents modèles de bases lexicales multilingues comme EuroWordNet, Balkanet et le projet Papillon. Nous avons constaté que tous ces projets multilingues utilisent une approche par pivot : *ILI* (EuroWordNet et Balkanet) et axie (projet Papillon). Chaque langue du dictionnaire est en relation avec ce pivot.

La deuxième partie de nos travaux consiste à modéliser le domaine des noms propres. Pour cela, nous devions d'abord essayer de répondre à la question : comment définir notre pivot ? Pour définir notre concept de pivot, nous avons étudié en détail la relation de synonymie, car cette relation est le pilier central du projet WordNet. Cette étude nous a permis de définir les deux concepts centraux de notre modèle : le nom propre conceptuel et le prolexème. Nous n'avons pas défini le nom propre conceptuel comme le référent, mais comme un point de vue sur ce référent.

Un nom propre conceptuel correspond dans chaque langue à un unique prolexème. Autour de ces deux concepts, nous avons défini d'autres concepts (alias, dérivé, etc.) et relations (méronymie, éponymie, etc.). Nous avons relié les prolexèmes et les alias par une relation de synonymie qui dépend de la langue. Une typologie des noms propres sous la forme d'une ontologie a été créée.

Notre modèle des noms propres peut se représenter sous la forme d'un graphe. Ce graphe qui hiérarchise nos différents concepts, se décompose à quatre niveaux différents. Les deux premiers niveaux, le niveau conceptuel (le nom propre conceptuel et ses relations séman-

tiques) et le niveau méta-conceptuel (typologie et existence), forment la partie qui ne dépend pas des langues. La partie qui dépend des langues est constituée du niveau linguistique (le prolexème, les alias, les dérivés et leurs relations spécifiques) et du niveau des instances (morphologie flexionnelle).

La dernière partie de nos travaux est consacrée à la description de l'implémentation de notre modèle. Nous avons appliqué la méthode Merise sur notre modélisation des noms propres pour définir un modèle conceptuel de données s'appliquant à toutes les langues de notre dictionnaire. En raison de certains problèmes (limitation des tables, rapidité des requêtes, etc.), nous avons décidé de ne pas appliquer les règles classiques de passage du MCD vers le MLD. Nous avons privilégié un MLD construit sur deux parties : une partie qui ne dépend pas des langues et une partie spécifique à chaque langue.

Le but de nos travaux étant de développer des ressources linguistiques pour la communauté des chercheurs du TAL, nous avons eu besoin de créer un format d'échange de nos données que nous avons conçu après l'étude de la TEI et de la TMF. La TEI est un format figé qui n'est pas adapté à la structure de nos données, alors que la TMF possède une structure plus souple. Cependant, comme la TMF ne permettait pas de modéliser nos relations qui ne dépendent pas des langues, nous avons adapté la TMF pour prendre en compte la structure de nos données et de nos relations.

Nous avons développé une interface de travail collaboratif pour permettre à nos partenaires de travailler sur notre dictionnaire. La plupart de nos collaborateurs possèdent déjà des listes de noms propres sous forme de fichiers. Nous avons ajouté à notre interface un menu qui leur permet de manipuler efficacement et facilement leurs fichiers.

Nous avons enfin testé et validé la pertinence de notre modèle en travaillant sur les noms propres du dictionnaire Larousse Collège, qui ont tous trouvé leur place dans notre modélisation. Ce travail nous a permis de répondre à certaines hésitations (garder ou non tel ou tel type, créer un nouveau type...).

Perspectives

Actuellement, nous travaillons sur la création d'une interface permettant l'exportation de Prolexbase sous le format défini au chapitre 6. Le site destiné aux requêtes XML et d'exportation XML est en cours de développement.

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

Dans le cadre du projet Prolex, cette thèse a été précédée de deux autres thèses :

- *Reconnaissance automatique des noms propres* ; application à la classification automatique de textes journalistiques [Friburger, 2002].
- *La dérivation toponymes-gentilés en français* : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique [Eggert, 2002].

Nous comptons utiliser leurs résultats respectifs pour ajouter de nouvelles entrées à Prolexbase et pour créer des règles de dérivation et d'aliasisation.

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...).

Liste de publications

Publications internationales avec comité de lecture

Grass T., Maurel D., **Tran M.** (2004), Une ontologie pour le traitement multilingue des noms propres in : *Linguistica Antverpiensia NS 3-2004 : "The translation of domain specific languages and multilingual terminology management"*, p. 293-309.

Tran M., Maurel M., Savary A. (2005), Implantation d'un tri lexical respectant la particularité des noms propres, *Lingvisticae Investigationes*, XXVIII-2.

Publications nationales avec comité de lecture

Maurel D., **Tran M.** (2005), Une ontologie multilingue des noms propres, Revue CORELA -Cognition, Représentation, Langage-, publication électronique.

Communications internationales avec comité de lecture

Grass T., Maurel D., **Tran M.** (2004), Un dictionnaire électronique multilingue de noms propres pour la traduction, *Third International Conference on International Translation*, Barcelone, Espagne, 4-6 mars, p. 165-174.

Krstev S., Vitas D., Maurel D., **Tran M.** (2005), Multilingual Ontology of Proper Names, *Second Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, 21-23 avril, p. 116-119.

Bouchou B., **Tran M.**, Maurel D. (2005), Towards an XML Representation of Proper Names and Their Relationships, *Tenth International Conference on Applications of Natural Language to Information Systems (NLDB'2005)*, Alicante, Spain, 15-17 juin, in Lecture Notes in Computer Science, 3513, p. 44-55.

Communications dans un atelier d'une conférence internationale avec comité de lecture

Tran M., Grass T., Maurel D. (2004), An ontology for multilingual treatment of proper names, *Ontologies and Lexical Resources in Distributed Environments (OntoLex 2004)*, in Association with LREC2004 (Actes p.75-78), Lisbonne, Portugal, 29 mai.

Tran M., Maurel D., Vitas D., Krstev S. (2005), A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names, *Papillon 2005 Workshop on Multilingual Lexical Databases, in association with the Sixth Symposium on Natural Language Processing (SNLP 2005)*, Chiang Rai, Thaïlande, 12-14 décembre, 2 :67-71.

Maurel D., **Tran M.**, Friburger N. (2006), 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, 13 avril.

Autres communications

Maurel D., **Tran M.**, Grass T., Vitas D. (2005), Prolexbase : un dictionnaire relationnel multilingue de noms propres, Colloque Traitement lexicographique des noms propres, Tours, 24 mars.

Rapports techniques

Maurel D., **Tran M.**, Vitas D., Grass T. et Savary A. (2004), *Prolexbase : Une ontologie multilingue des noms propres*. Rapport Interne du Laboratoire d'Informatique de l'Université de Tours (EA 2101). Rapport 279, 34 p.

Tran M., Maurel D. (2004), *Prolexbase : Le modèle conceptuel de données*. Rapport Interne du Laboratoire d'Informatique de l'Université de Tours (EA 2101). Rapport 275, 20 p.

Maurel D., **Tran M.**, Vitas D., Grass T., Savary A. (2004), P*rolexbase : Proposition d'une ontologie multilingue des noms propres*, Rapport interne du Laboratoire d'Informatique de l'Université de Tours, n°274, 32 p.

Maurel D., **Tran M.**, Vitas D., Grass T., Savary A. (2006), *Prolex : Implantation d'une ontologie multilingue des noms propres*, Rapport interne du Laboratoire d'Informatique de l'Université de Tours, n°286, 47 p.

Tran M., Maurel D. (2006), *Prolexbase : le modèle conceptuel de données et son implantation*, Rapport interne du Laboratoire d'Informatique de l'Université de Tours, n°287, 21 p.

Tran M., Maurel D. (2006), *Prolexbase : les interfaces de consultation et de travai*l, Rapport interne du Laboratoire d'Informatique de l'Université de Tours, n°289, 24 p.

Posters et démonstrations

Maurel D., **Tran M.** (2005), Prolexbase : Un lexique syntaxique et sémantique de noms propres, affichage à la Journée d'étude de l'ATALA : Interface lexique-grammaire et lexiques syntaxiques et sémantiques, Paris, 12 mars.

Tran M., Maurel D., Vitas D., Krstev S. (2005), Prolex : a demo, Papillon 2005 Workshop on Multilingual Lexical Databases, in association with the Sixth Symposium on Natural Language Processing (SNLP 2005), Chiang Rai, Thaïlande, 12-14 décembre.

Séminaires

Tran M. (2004) An ontology for multilingual treatment of proper names. *Forum de l'Ecole Doctorale Santé Sciences et Technologies*, Tours, France.

Annexe A

MCD et MLD

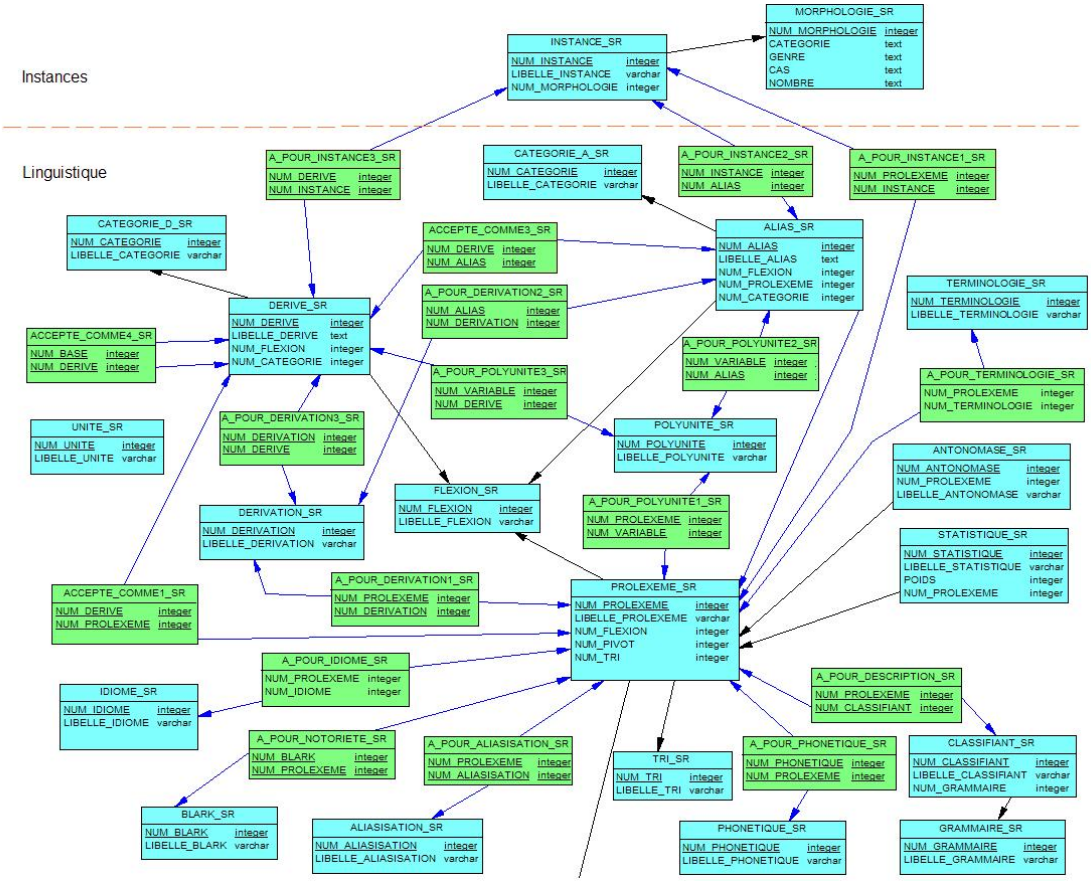

FIG. A.1 – Le MLD serbe.

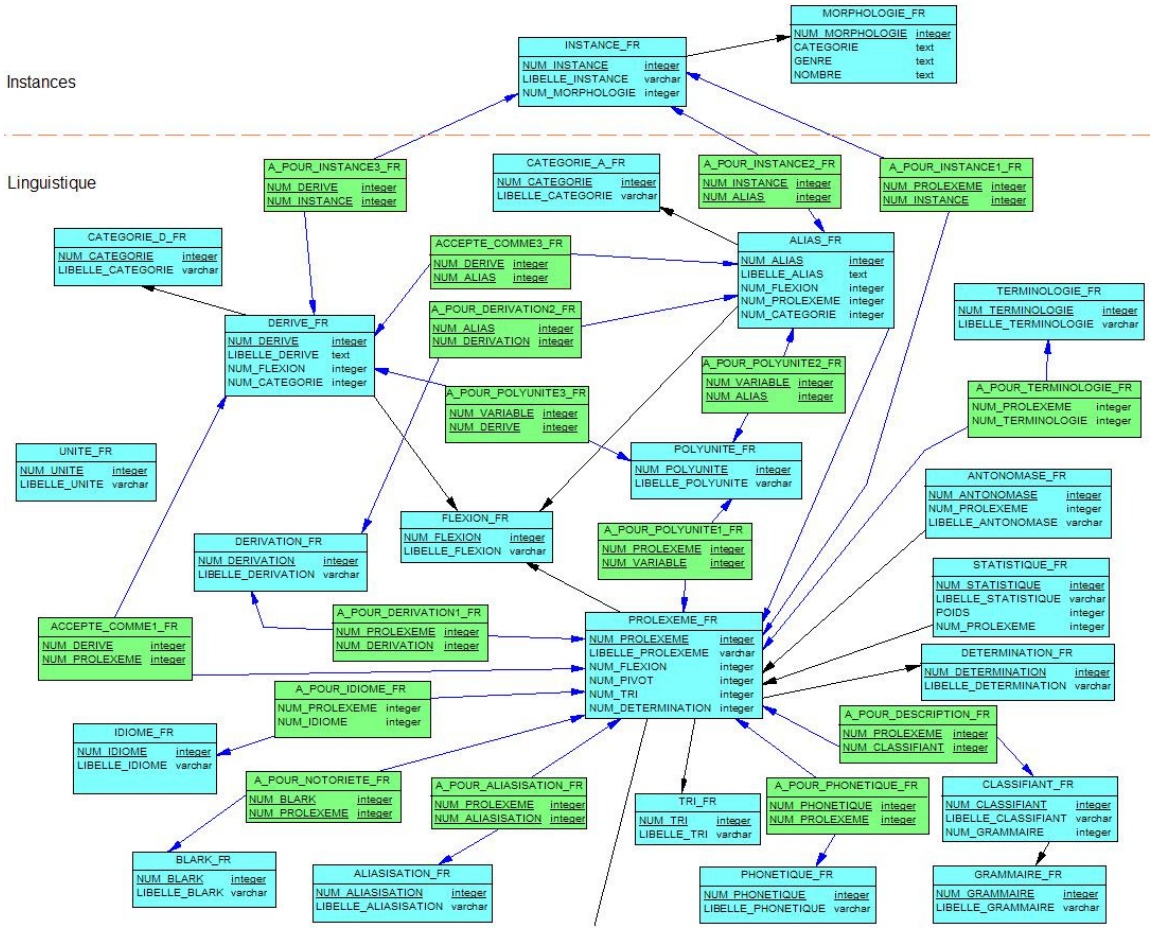

FIG. A.2 – Le MLD français.

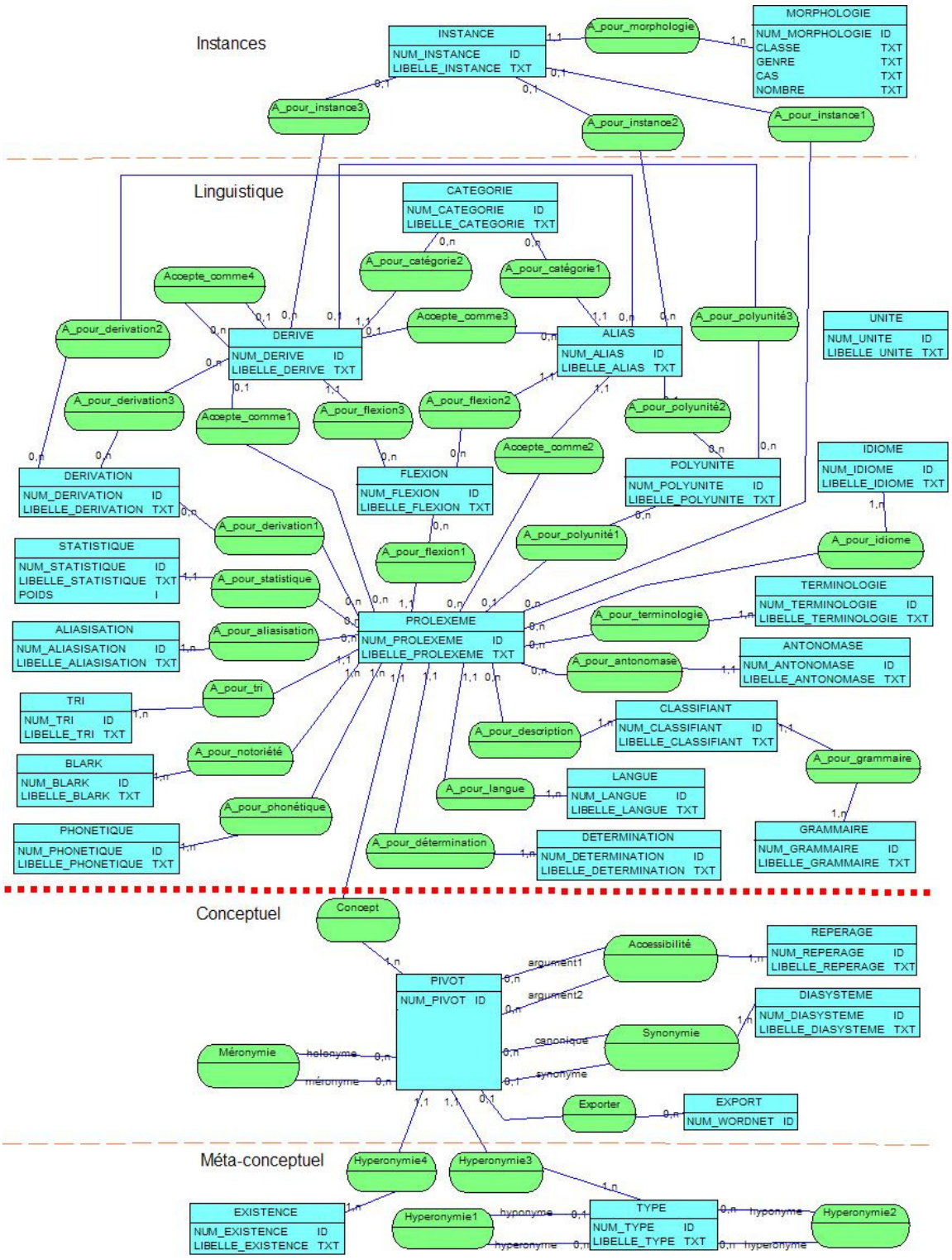

FIG. A.3 – Le MCD des noms propres.

Annexe B

Codes flexionnels du DELA

Modèle de flexion

N/A00	=	:ms, :fs, :mp, :fp

Masculins sans féminin (sauf classe 6)

N/A1	=	0,-,s,-	ballon, ballons
N/A2	=	0,-,0,-	engrais, engrais
N/A3	=	0,-,x,-	bureau, bureaux
N/A4	=	l,-,ux,-	cheval, chevaux/ciel, cieux
N/A5	=	il,-,ux,-	travail, travaux
N/A6	=	0,-,s,s	amour/délice/orgue
N/A7	=	us,-,i,-	naevus, naevi
N/A8	=	um,-,a,-	quantum, quanta
N/A9	=	homme,-,shommes,-	bonhomme, bonshommes
N/A10	=	man,-,men,-	barman, barmen
N/A11	=	y,-,ies,-	lobby, lobbies
N/A12	=	0,-,es,-	coach, coaches
N/A13	=	o,-,i,-	carbonaro, carbonari
N/A14	=	0,-,im,-	goy, goyim
N/A15	=	0,-,m,-	sefardi, sefardim
N/A16	=	e,-,i,-	nuraghe,nuraghi
N/A17	=	0,-,er,-	län, läner
N/A18	=	0,-,in,-	moudjahid, moudjahidin

Féminins sans masculin

N/A21	=	-,0,-,s	balle, balles
N/A22	=	-,0,-,0	brebis, croix
N/A23	=	-,0,-,x	peau, peaux/eau, eaux
N/A24	=	-y,-,ies	lady, ladies
N/A25	=	-,man,-,men	recordwoman, recordwomen
N/A26	=	-,a,-,ae	nova, novae

Pluriels : ajout de 's'

N/A31	=	0,0,s,s	artiste, artistes
N/A32	=	0,e,s,es	ami, amie, amis, amies
N/A33	=	0,se,s,ses	andalou, andalouse
N/A34	=	0,te,s,tes	falot/favori,favorite
N/A35	=	eur,euse,eurs, euses	danseur, danseuse
N/A36	=	eur,rice,eurs,rices	acteur, actrice
N/A37	=	ur,resse,urs,resses	demandeur, eresse
N/A38	=	f,ve,fs,ves	actif,ive/veuf,veuve
N/A39	=	0,sse,s,sses	maitre,esse/bonze,bonzesse
N/A40	=	l,lle,ls,lles	colonel,elle/nul,nulle
N/A41	=	n,nne,ns,nnes	ancien, ienne/champion,onne
N/A42	=	er,ère,ers,ères	boucher, bouchère
N/A43	=	et,ète,ets,ètes	inquiet, inquiète
N/A44	=	ef, ève,efs,èves	bref, brève
N/A45	=	ec,èche,ecs,èches	sec, sèche
N/A46	=	c,que,cs,ques	caduc, caduque/turc,turque
N/A47	=	c,che,cs,ches	blanc, blanche/franc, franche
N/A48	=	c,chesse,cs,chesses	duc, duchesse
N/A49	=	g,gue,gs,gues	long, longue
N/A50	=	gu,guë,gus,guës	ambigu, ambiguë
N/A51	=	0,sque,s,sques	maure, mauresque
N/A52	=	n,gne,ns,gnes	malin, maligne
N/A53	=	u,lle,us,lles	fou,folle/mou,molle
N/A54	=	r,use,rs,uses	streaker, streakeuse
N/A55	=	0,ine,s,ines	feuillant, feuillantine
N/A56	=	0,esse,s,esses	clown, clownesse
N/A57	=	o,a,os,as	aficionado ,aficionada
N/A58	=	ète,étesse,ètes,étesses	poète, poétesse
N/A59	=	c,cque,cs,cques	grec, grecque

...

Annexe C

Exemple XML : le prolexème *États-Unis d'Amérique*

En lançant une requête de recherche sur le nom propre *USA*, nous obtenons le fichier XML ci-dessous. Nous avons deux prolexèmes : un de type hydronyme et un de type pays.

```
<struct type="Prolex">
  <struct type="pivot">
    <feat type="type">Hydronyme</feat>
    <feat type="existence">Historique</feat>
    <feat type="identifier">42622</feat>
    <struct type="prolexeme">
      <feat type="language">fr</feat>
      <feat type="lemma">Usa</feat>
      <feat type="pos">name</feat>
      <feat type="category">proper name</feat>
      <struct type="inflection">
        <feat type="form">Usa</feat>
        <feat type="gender" />
        <feat type="number" />
      </struct>
    </struct>
  </struct>
  <struct type="pivot">
    <feat type="type">Pays</feat>
    <feat type="existence">Historique</feat>
    <feat type="identifier">46929</feat>
    <struct type="prolexeme">
      <feat type="language">fr</feat>
      <feat type="lemma">États-Unis d'Amérique</feat>
      <feat type="pos">name</feat>
      <feat type="category">proper name</feat>
      <struct type="inflection">
        <feat type="form">États-Unis d'Amérique</feat>
        <feat type="gender">masculine</feat>
        <feat type="number">plural</feat>
      </struct>
```

```
<struct type="alias">
   <feat type="lemma">États-Unis</feat>
   <feat type="pos">name</feat>
   <feat type="category">Abréviation</feat>
   <struct type="inflection">
      <feat type="form">États-Unis</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
   </struct>
</struct>
<struct type="alias">
   <feat type="lemma">USA</feat>
   <feat type="pos">name</feat>
   <feat type="category">Acronyme ou sigle étranger</feat>
   <struct type="inflection">
      <feat type="form">USA</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
   </struct>
</struct>
<struct type="alias">
   <feat type="lemma">US</feat>
   <feat type="pos">name</feat>
   <feat type="category">Acronyme ou sigle étranger</feat>
   <struct type="inflection">
      <feat type="form">US</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
   </struct>
</struct>
<struct type="alias">
   <feat type="lemma">USA</feat>
   <feat type="pos">name</feat>
   <feat type="category">Acronyme ou sigle étranger</feat>
   <struct type="inflection">
      <feat type="form">USA</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
   </struct>
</struct>
<struct type="Derivative">
   <feat type="lemma">américano</feat>
   <feat type="pos" />
   <feat type="category">Préfixe</feat>
   <struct type="inflection">
      <feat type="form">américano</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
   </struct>
```

```
</struct>
<struct type="Derivative">
   <feat type="lemma">Américain</feat>
   <feat type="pos">nom</feat>
   <feat type="category">Nom relationnel</feat>
   <struct type="inflection">
      <feat type="form">Américaines</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">Américains</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">Américaine</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">singular</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">Américain</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">singular</feat>
   </struct>
</struct>
<struct type="Derivative">
   <feat type="lemma">américain</feat>
   <feat type="pos">adjectif</feat>
   <feat type="category">Adjectif relationnel</feat>
   <struct type="inflection">
      <feat type="form">américaines</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">américains</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">américaine</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">singular</feat>
   </struct>
   <struct type="inflection">
      <feat type="form">américain</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">singular</feat>
```

```
    </struct>
  </struct>
  <struct type="Derivative">
    <feat type="lemma">États-Unien</feat>
    <feat type="pos">nom</feat>
    <feat type="category">Nom relationnel</feat>
    <struct type="inflection">
      <feat type="form">États-Uniennes</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">États-Uniens</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">États-Unienne</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">singular</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">États-Unien</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">singular</feat>
    </struct>
  </struct>
  <struct type="Derivative">
    <feat type="lemma">états-unien</feat>
    <feat type="pos">adjectif</feat>
    <feat type="category">Adjectif relationnel</feat>
    <struct type="inflection">
      <feat type="form">états-uniennes</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">plural</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">états-uniens</feat>
      <feat type="gender">masculine</feat>
      <feat type="number">plural</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">états-unienne</feat>
      <feat type="gender">feminine</feat>
      <feat type="number">singular</feat>
    </struct>
    <struct type="inflection">
      <feat type="form">états-unien</feat>
      <feat type="gender">masculine</feat>
```

```
                <feat type="number">singular</feat>
            </struct>
        </struct>
        <struct type="Derivative">
            <feat type="lemma">Yankee</feat>
            <feat type="pos">nom</feat>
            <feat type="category">Nom relationnel diastratique</feat>
            <struct type="inflection">
                <feat type="form">Yankees</feat>
                <feat type="gender">masculine feminine</feat>
                <feat type="number">plural</feat>
            </struct>
            <struct type="inflection">
                <feat type="form">Yankee</feat>
                <feat type="gender">masculine feminine</feat>
                <feat type="number">singular</feat>
            </struct>
        </struct>
        <struct type="Derivative">
            <feat type="lemma">Amerloque</feat>
            <feat type="pos">nom</feat>
            <feat type="category">Nom relationnel diastratique</feat>
            <struct type="inflection">
                <feat type="form">Amerloques</feat>
                <feat type="gender">masculine feminine</feat>
                <feat type="number">plural</feat>
            </struct>
            <struct type="inflection">
                <feat type="form">Amerloque</feat>
                <feat type="gender">masculine feminine</feat>
                <feat type="number">singular</feat>
            </struct>
        </struct>
    </struct>
</struct>
```

Index

éponymie, 87

abréviation, 54
acronyme, 54
alias, 53
aliasisation, 88
anthroponymes, 72
association, 84
axie, 42

Balkanet, 37
BLARK, 89

dérivation, 88
détermination, 89
DEC, 40
DELAS, 30
diachronique, 51
diaphasique, 51
diastratique, 51
diasystème, 48, 86
Dico, 41

entité, 84
ergonymes, 73
EuroWordNet, 32, 35, 86
existence, 78, 87

GMT, 99

hyperonymie, 34

ILI, 36, 87
Inter-Lingual-Index, 36

lexie, 40, 48

méronymie, 34
MCD, 83
Merise, 83
MLD, 84
MUC, 23
MySQL, 107

nom propre conceptuel, 52
notoriété, 89

ontologie, 69

phonétique, 90
pragmonymes, 73
projet Papillon, 42
prolexème, 52

repérage, 86

sigle, 54
statistique, 88
supertype, 72
synonymie, 34, 48
synset, 34

TEI, 97
TLFi, 39
TMF, 99
toponymes, 73
transcription, 56
tri, 90
type, 73
typologie, 22, 71

Wikipédia, 86
WordNet, 32

Bibliographie

[Adouani, 1993] Adouani, A. (1993). Traitement dérivationnel des supplétismes lexicaux. In *Cahiers de lexicologie*, volume 63, pages 87–98.

[Aljovic, 2000] Aljovic, N. (2000). *Recherches sur la morpho-syntaxe du groupe nominal en serbo-croate*. Thèse de doctorat en sciences du langage, Université Paris 8.

[Assia, 2006] Assia, K. (2006). *Extraction de données Prolex au format XML*. Mémoire de Master Informatique, Université François-Rabelais de Tours.

[Bauer, 1985] Bauer, G. (1985). *Namenkunde des Deutschen*. Germanistische Lehrbuchsammlung Band 21, Berlin.

[Belleil, 1997] Belleil, C. (1997). *Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel*. Thèse de doctorat en informatique, Université de Nantes.

[Blanco, 2001] Blanco, X. (2001). Dictionnaires électroniques et traduction automatique espagnol-français. In *Langages 143 (Lexicologie contrastive espagnol-français), Larousse*, volume 143.

[Bodenreider and Zweigenbaum, 2000a] Bodenreider, O. and Zweigenbaum, P. (2000a). Identifying proper names in parallel medical terminologies. In *Medical Infobahn for Europe (MIE2000)*, pages 443–447.

[Bodenreider and Zweigenbaum, 2000b] Bodenreider, O. and Zweigenbaum, P. (2000b). Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles. In *Traitement automatique des langues*, volume 41-3, pages 727–757.

[Bouchou et al., 2005] Bouchou, B., Tran, M., and Maurel, D. (2005). Towards an XML Representation of Proper Names and Their Relationships. In *Tenth International Conference on Applications of Natural Language to Information Systems (NLDB'2005), published in LCNS 3513*, pages 44–55, Alicante, Spain.

[Charlet et al., 2003] Charlet, J., Bachimont, B., and Troncy, R. (2003). Web sémantique, rapport final de l'action spécifique 32 CNRS/STIC. Rapport technique.

[Chinchor, 1997] Chinchor, N. (1997). Overview of MUC-7/MET-2. *http ://www.itl.nist.gov/iaui/894.02/related_ projects/muc/proceedings/muc_ 7_ toc.html#appendices*.

[Coates-Stephens, 1993] Coates-Stephens, S. (1993). The Analysis and Acquisition of Proper Names for the Understanding of Free Text. In *Computers and the Humanities*, volume 26, pages 441–456, Hingham, MA.

[Constant, 2003] Constant, M. (2003). *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.

[Coseriu, 1998] Coseriu, E. (1998). Le double problème des unités dia-s. In *Les Cahiers δia. Etudes sur la diachronie et la variation linguistique*, pages 9–16.

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

[Cucchiarini et al., 2000] Cucchiarini, C., Daelemans, W., and Strik, H. (2000). Strengthening the Dutch Human Language Technology Infrastructure. Technical report, http ://www.elda.fr/article48.html.

[Daille, 1999] Daille, B. (1999). Identification des adjectifs relationnels en corpus. In *TALN 99*, pages 105–114.

[Danlos, 1989] Danlos, L. (1989). La traduction automatique. In *Annales des télécommunications*, volume 44, pages 94–100.

[Dendien and Pierrel, 2003] Dendien, B. and Pierrel, J.-M. (2003). Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de la langue de référence. In *TAL*, volume 44, pages 11–37.

[Dubois, 1973] Dubois, J. (1973). *Dictionnaire de linguistique.* Larousse, Paris.

[Eggert, 2002] Eggert, E. (2002). *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.

[Eggert et al., 1998] Eggert, E., Maurel, D., and Belleil, C. (1998). Allomorphies et supplétions dans la formation des gentilés. Application au traitement informatique. In *Cahiers de lexicologie*, volume 73, pages 167–179.

[Fillmore et al., 2003] Fillmore, C., Johnson, C., and Petruck, M. (2003). Background to Framenet. In *International Journal of Lexicography*, volume 16, pages 235–250.

[Francopoulo, 2003] Francopoulo, G. (2003). CN RNIL N 7. *AFNOR.*

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

[Garrigues, 1993] Garrigues, M. (1993). *Méthode de paramétrage des dictionnaires et grammaires électroniques : Application à des systèmes interactifs en langue naturelle.* Thèse de doctorat en Sciences du Langage, Université Paris VII.

[Gary-Prieur, 1994] Gary-Prieur, M.-N. (1994). *Grammaire du nom propre.* Presse Universitaire de France, Paris.

[Grass, 1999] Grass, T. (1999). Typologie et traductibilité des noms propres de l'allemand vers le français à partir d'un corpus journalistique. *Journée d'Etude de l'ATALA " Le traitement automatique des noms propres ", Université Paris 7, France.*

[Gravier et al., 2004] Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., Tait, K. M., and Choukri, K. (2004). ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Journées d'Etude de la Parole*, Fèz (Maroc).

[Gross, 1995] Gross, G. (1995). À propos de la notion d'humain. In *Linguisticae Investigationes Supplementa*, volume 17, pages 16–19.

[Gross, 1989] Gross, M. (1989). The Use of Finite Automata in the Lexical Representation of Natural Language. In *Electronic Dictionaries and Automata in Computational Linguistics, LNCS*, volume 377, pages 34–50.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. In *Knowledge Acquisition*, volume 5, pages 199–220.

[Harris, 1968] Harris, Z. S. (1968). *Mathematical Structures of Language*. Interscience Publishers, Dunod, Paris.

[Harris, 1976] Harris, Z. S. (1976). *Notes du cours de syntaxe*. Interscience Publishers, Le Seuil, Paris.

[Ide and Véronis, 1996] Ide, N. and Véronis, J. (1996). Codage TEI des dictionnaires électroniques. In *Cahiers Gutenberg n° 24 (spécial TEI)*, pages 170–176.

[Jansen, 2004] Jansen, P. (2004). Lexicography in an Interlingual Ontology : An Introduction to EuroWordNet. *Canadian Undergraduate Journal of Cognitive Science*.

[Jonasson, 1994] Jonasson, K. (1994). *Le nom propre. Constructions et interprétations*. Duculot, Paris.

[Kleiber, 1996] Kleiber, G. (1996). Noms propres et noms communs : un problème de dénomination. In *Meta*, volume 41-4, pages 567–589.

[Krstev et al., 2004] Krstev, C., Pavlovic-Lazetic, G., Vitas, D., and Obradovic, I. (2004). Using Textual and Lexical Resources in Developing the Serbian Wordnet. In *Romanian journal of Information science and technology*, volume 7-1-2, pages 147–161.

[Krstev et al., 2005] Krstev, S., Vitas, D., Maurel, D., and Tran, M. (2005). Multilingual Ontology of Proper Names. In *Second Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 116–119.

[Laporte, 1990] Laporte, E. (1990). Le dictionnaire phonémique DELAP. *Langues française, Larousse*.

[Le Meur et al., 2004] Le Meur, C., Galliano, S., and Geoffrois, E. (2004). Conventions d'annotations en Entités Nommées - ESTER. Rapport technique, http ://www.afcp-parole.org/ester/docs/convention_en_old.pdf.

[Le Pesant and Mathieu-Colas, 1998] Le Pesant, D. and Mathieu-Colas, M. (1998). Introduction aux classes d'objets. In *Langages*, volume 131, pages 6–33.

[Lepesant, 2000] Lepesant, D. (2000). *Six études de sémantique lexicale sur les noms communs de lieux*. Mémoire d'HDR, Université Paris 13.

[Leroy, 1994] Leroy, S. (1994). *Le nom propre en français*. Ophrys, collection l'essentiel français.

[M. Grevisse, 1986] M. Grevisse, A. G. (1986). *Le Bon Usage*. Duculot, Gembloux, Belgique.

[MacDonald, 1996] MacDonald, D. (1996). Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names. In *Corpus Processing for Lexical Acquisition*, pages 21–39, Massachussetts Institute of Technology.

[Mangeot-Lerebours, 2001] Mangeot-Lerebours, M. (2001). *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de doctorat, Université Joseph Fourier Grenoble 1, GETA-CLIPS.

[Mangeot-Lerebours et al., 2003] Mangeot-Lerebours, M., Sérasset, G., and Lafourcade, M. (2003). Construction collaborative d'une base lexicale multilingue - Le projet Papillon. In *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*, volume 44(2), pages 151–176.

[Matheron, 1998] Matheron, J. (1998). *Comprendre Merise ; outils conceptuels et organisationnels*. Éditions Eyrolles, Paris.

[Maurel, 2004] Maurel, D. (2004). Les mots inconnus sont-ils des noms propres ? *Septièmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004), Louvain-la-Neuve, Belgique, 10-12 mars*, pages 1–8.

[Maurel et al., 1996] Maurel, D., Belleil, C., Eggert, E., and Piton, O. (1996). Le projet PROLEX, séminaire Représentations et outils pour les bases lexicales. In *Morphologie Robuste de l'action Lexique du GDR-PRC CHM, Grenoble*, pages 164–175.

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

[Mel'čuk et al., 1995] Mel'čuk, I., Clas, A., and Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

[Mikheev et al., 1999] Mikheev, A., Moens, M., and Grover, C. (1999). Named Entity Recognition without Gazetteers. *EACL'99*, pages 1–8.

[Miller, 1995] Miller, G. (1995). Wordnet : A lexical database for English. In *Communication of the ACM*, volume 38(11), pages 39–41.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to WordNet : an On-line Lexical Database. In *Journal of Lexicography*, volume 3, pages 235–244.

[Mizoguchi, 2005] Mizoguchi, R. (2005). Ontological Engineering. In *SNLP 2005*, pages 13–22, Thailand, Chiang Rai.

[Molino, 1982] Molino, J. (1982). Le nom propre dans la langue. *Langage, n°66 Paris Larousse*.

[MUC-6, 1995] MUC-6 (1995). *Proceedings of the 6th Message Understanding Conference*. Columbia, USA. Morgan Kaufmann.

[Nakos, 1990] Nakos, D. (1990). Sigles et noms propres. In *Meta 35*, volume 2, pages 407–413.

[Noy and McGuinness, 2003] Noy, N. F. and McGuinness, D. L. (2003). Ontologie pour le Web sémantique. *Web sémantique, rapport final de l'Action spécifique 32 CNRS/STIC*.

[ONU, 1977] ONU (1977). Troisième conférence des Nations Unies sur la normalisation des noms géographiques. Athènes.

[Paik et al., 1996] Paik, W., Liddy, E. D., Yu, E., and McKenna, M. (1996). Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. *Corpus Processing for Lexical Acquisition*, pages 61–76.

[Paumier, 2003] Paumier, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat en informatique, Université de Marne-la-Vallée.

[Paumier, 2006] Paumier, S. (2006). UNITEX 1.2 ; Manuel d'utilisation. Rapport technique, Université de Marne-la-Vallée.

[Piton and Maurel, 2004] Piton, O. and Maurel, D. (2004). Les noms propres géographiques et le dictionnaire Prolintex. In *Cahiers de la MSH Ledoux, Série Archive, Bases, Corpus, n° 1*, pages 53–76.

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

[Ren and Perrault, 1992] Ren, X. and Perrault, F. (1992). The Typology of Unknown Words : an Experimental Study of Two Corpora. In *COLING 92, Nantes*, pages 408–414.

[Roche, 2005] Roche, C. (2005). Terminologie et ontologie. In *Larousse-Revue Languages*, volume 157, pages 48–62.

[Romary, 2002] Romary, L. (2002). De la sémantique des contenus à la sémantique des structures. In *La recherche d'information sur les réseaux, Sciences de l'information, série Études et techniques, ADBS Éditions*, pages 203–230.

[Romary et al., 2004] Romary, L., Salmon-Alt, S., and Francopoulo, G. (2004). Standards going concrete : from LMF to Morphalou. In *Workshop on Electronic Dictionaries, COLING 2004*, pages 22–28, Geneva.

[Romary and Van Campenhoudt, 2001] Romary, L. and Van Campenhoudt, M. (2001). Normalisation des échanges de données en terminologie : le cas des relations dites conceptuelles. In *Conférence TIA-2001*, pages 77–86, Nancy.

[Sang, 2002] Sang, E. F. T. K. (2002). Introduction to the CoNLL-2002 Shared Task : Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002, Taipei, Taiwan*, pages 155–158.

[Savary, 2000] Savary, A. (2000). *Recensement et description de mots composés - méthodes et applications*. Thèse de doctorat en informatique, Université de Marne-la-Vallée.

[Savary, 2005] Savary, A. (2005). Towards a Formalism for the Computational Morphology of Multi-Word Units. In *The 2nd Language & Technology Conference (LTC'05)*, pages 21–23, Poznan.

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

[Sekine et al., 2002] Sekine, S., Sudo, K., and Nobata, C. (2002). Extended Named Entity Hierarchy. In *Proceedings of the third International Conference on Language Ressources and Evaluation (LREC'2002)*, volume 5, pages 1818–1828.

[Silberztein, 1990] Silberztein, M. (1990). Dictionnaires électroniques des mots composés. In *Langues française, Larousse*, volume 87, pages 71–83.

[Sérasset, 1994] Sérasset, G. (1994). *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse de doctorat, Université Joseph Fourier Grenoble 1.

[Tran et al., 2004] Tran, M., Grass, T., and Maurel, D. (2004). An ontology for multilingual treatment of proper names. In *OntoLex 2004, in Association with LREC2004*, pages 75–78.

[Tran et al., 2005] Tran, M., Maurel, D., and Savary, A. (2005). Implantation d'un tri lexical respectant la particularité des noms propres. In *Lingvisticae Investigationes*, volume 28-2, pages 303–323.

[Tufis et al., 2004] Tufis, D., Cristea, D., and Stamou, S. (2004). BalkaNet : Aims, Methods, Results and Perspectives. A General Overview. In *Romanian journal of Information science and technology*, volume 7-1-2, pages 9–44.

[Van Campenhoudt, 1996] Van Campenhoudt, M. (1996). Recherche d'équivalences et structuration des réseaux notionnels : le cas des relations méronymiques. In *Terminology*, volume 3:2, pages 53–83.

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

[Vossen, 1999] Vossen, P. (1999). EuroWordNet General Document. Technical report, Available at `http://www.illc.uva.nl/EuroWordNet/docs.html`.

[Vossen et al., 1998] Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., and Alonge, A. (1998). The EuroWordNet Base Concepts and Top Ontology. Technical report, Available at http ://www.illc.uva.nl/EuroWordNet/docs.html.

[Vossen et al., 1997] Vossen, P., Diez-Orzas, P., and Peters, W. (1997). Multilingual Design of EuroWordNet. In *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for Natural Language Processing Applications, Madrid*, pages 1–8.

[Winston et al., 1987] Winston, M. E., Chaffin, R., and Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. In *Cognitive Science*, volume 11, pages 417–444.

[Zabeeh, 1968] Zabeeh, F. (1968). What's in a Name, An Inquiry into the Semantics and Pragmatics of Proper Names. *La Haye, Martinus Nijhoff.*

Résumé :

Cette thèse présente les différentes étapes de la conception et de l'implémentation d'un dictionnaire électronique relationnel multilingue de noms propres, destiné à des processus automatiques. Une étude de différents travaux dans le monde des dictionnaires électroniques et dans le domaine des noms propres en linguistique et en TAL nous a permis de proposer une modélisation des noms propres. Cette modélisation repose sur une architecture en deux parties : une partie qui ne dépend pas des langues et une partie qui dépend de la langue. La première partie est formée d'un niveau métaconceptuel, regroupant les concepts de type, de supertype et d'existence, et d'un niveau conceptuel, qui comprend le concept de nom propre conceptuel et quatre relations (la synonymie, l'hyperonymie, la méronymie et l'accessibilité). Le seconde partie comprend le niveau linguistique (prolexème, alias, dérivés et les relations qui dépendent de la langue) et le niveau des instances (formé de l'ensemble des formes fléchies du prolexème, des alias et des dérivés).

Nous avons implémenté notre modèle sous la forme d'une base de données relationnelle. Une interface de consultation (`http://tln.li.univ-tours.fr/tln_prolex/prolex.php`) et une interface de travail collaboratif (`http://tln.li.univ-tours.fr/tln_prolexbase/`) ont été développées. Nous avons créé deux formats XML (un format de requête et un format d'exportation) pour l'échange de nos données.

Mots-clés : noms propres, dictionnaire électronique, typologie, ontologie, synonymie, méronymie, accessibilité, XML.

Abstract :

This thesis presents the different stage of the design and the implementation of an electronic relational multilingual dictionary of proper names for automatic process. A study of different work in the world of electronic dictionaries and in the domain of proper names in linguistic and in NLP has allowed us to propose a modelling of proper names. This modelling is based on an architecture in two parts : a commun part for languages and a specific part for a given language. The first part contains a metaconceptual level, which regroups the concepts of type, supertype and existence, and the conceptual part, which regroups the conceptual proper name and four relations (the synonymy, the hyperonymy, the meronymy and the accessibility). The second part contains a linguistic level (prolexeme, aliases, derivatives and relations which depend on a language) and a instances level (with the inflected form of the prolexeme, the aliases and the derivatives).

We have implemented our model trough a relational database. An interface of consultation (`http://tln.li.univ-tours.fr/tln_prolex/prolex.php`) and a collaborative work interface (`http://tln.li.univ-tours.fr/tln_prolexbase/`) have been developed. We have created two XML format (a format of request and a format of exportation) for the exchange of our data.

Keywords : proper names, electronic dictionary, typology, ontology, synonymy, meronymy, accessibility, XML.