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

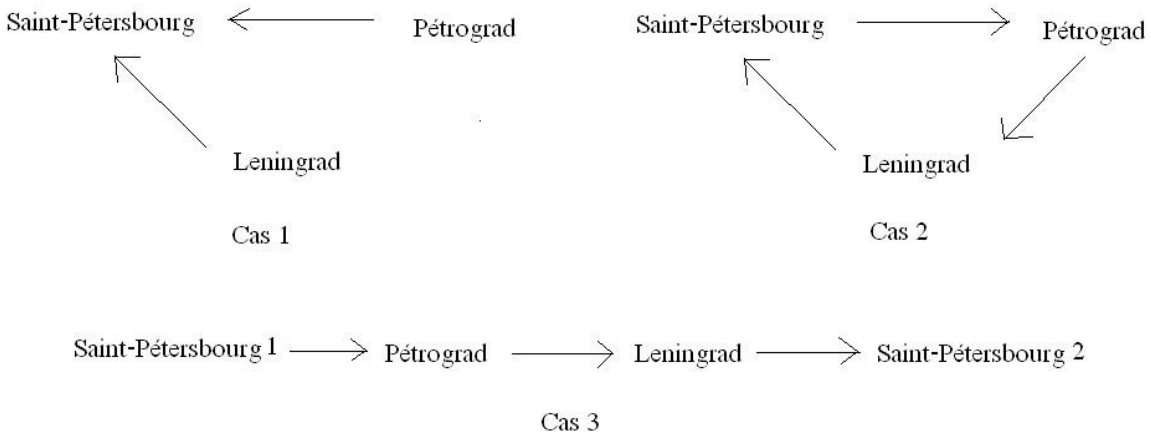

FIG. 8.1 – Exemple de cycle pour la relation de synonymie diachronique.

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.

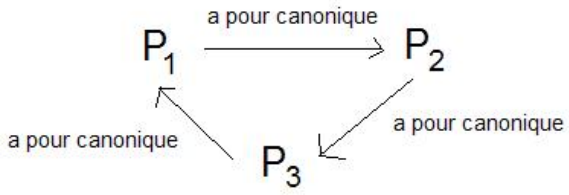

FIG. 8.2 – Exemple de cycle pour la relation de synonymie.

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.

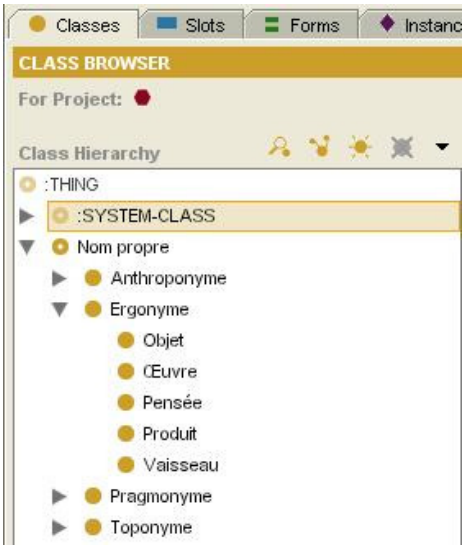

FIG. 8.6 – Le type Pensée.

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

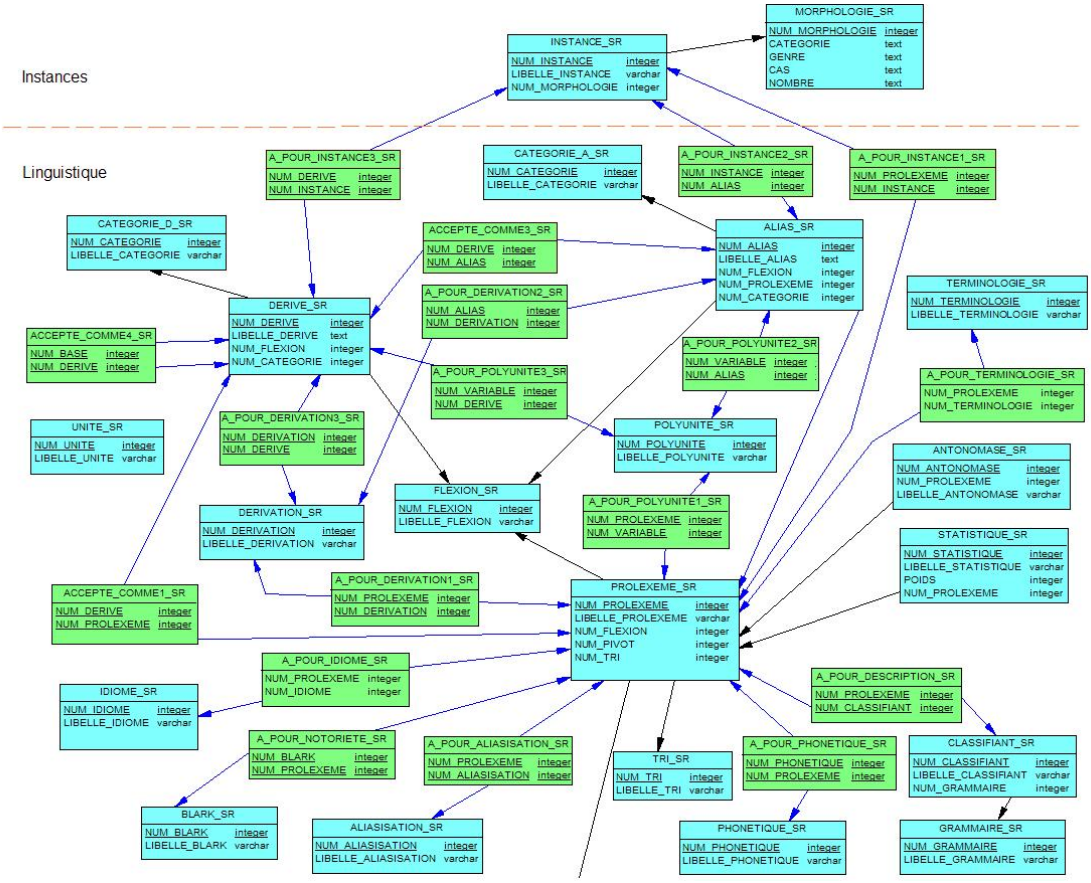

FIG. A.1 – Le MLD serbe.

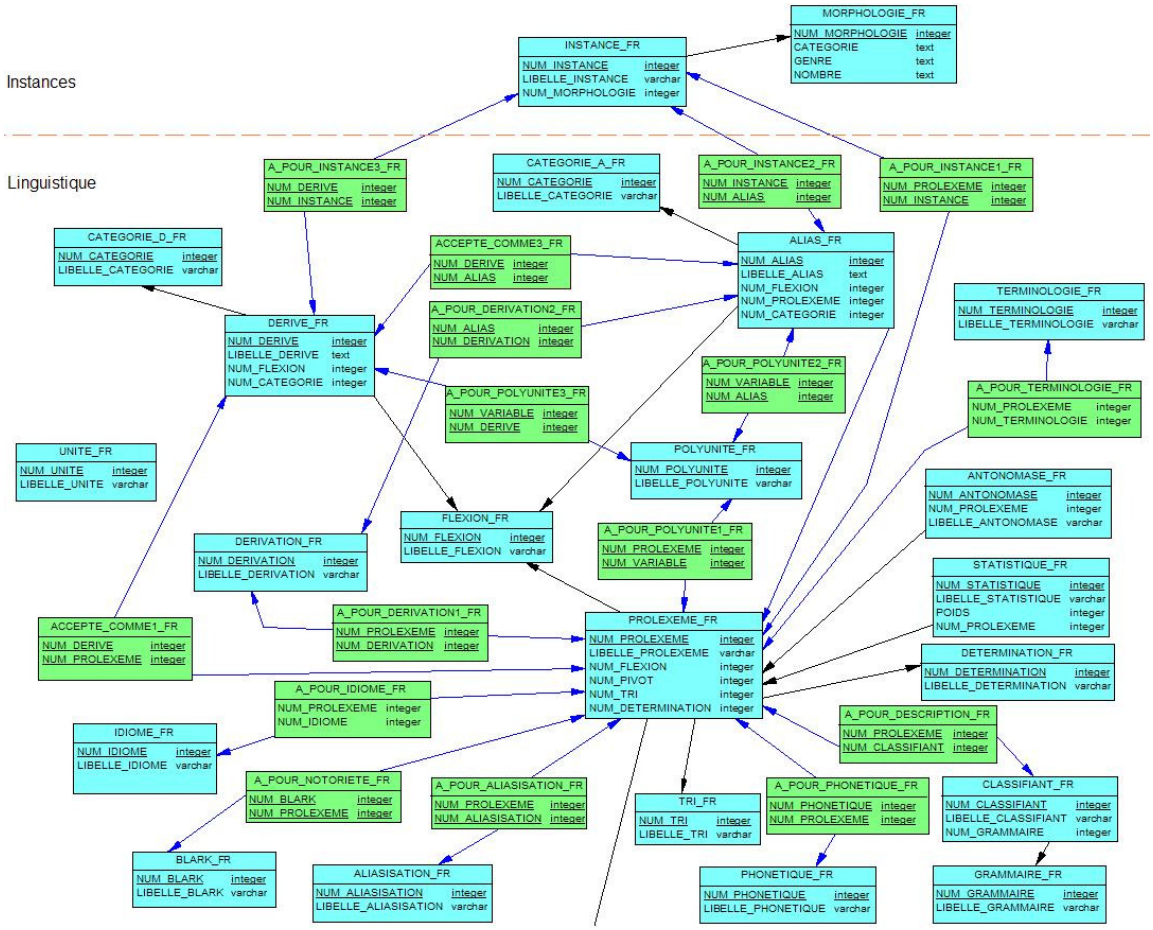

FIG. A.2 – Le MLD français.

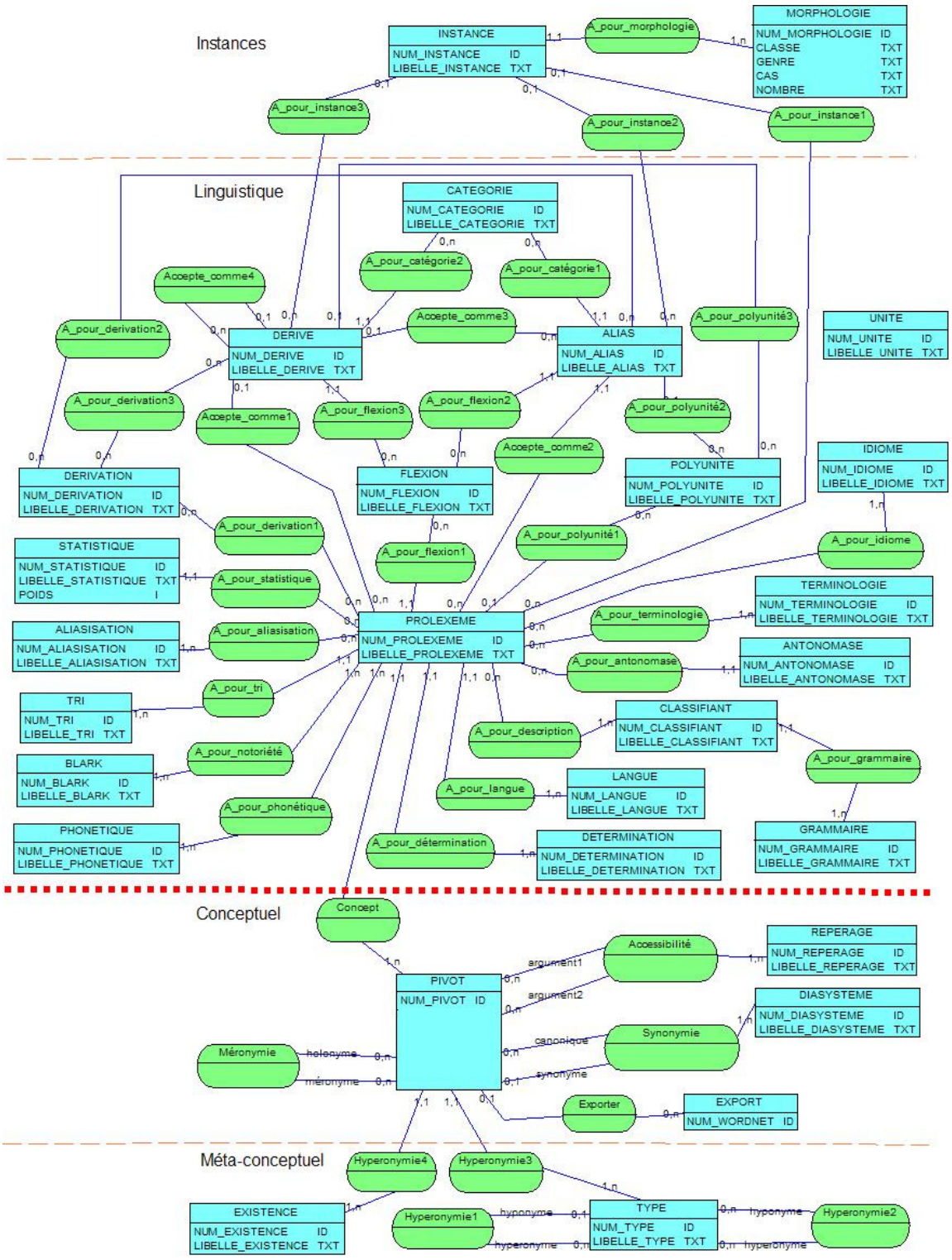

FIG. A.3 – Le MCD des noms propres.

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>
```