Prolexbase: une ontologie pour le traitement multilingue des noms propres

Thierry Grass, Denis Maurel & Mickaël Tran
Université de Tours

Proper names often constitute a problem in translation. This contribution deals with an ontology which represents the basis for a multilingual database of proper names, Prolexbase. It is being set up for treatment of proper names in the framework of the Prolex project, a research programme supported by the French Ministry of Industry in collaboration with two firms working on the market of language technologies: Systran and Exalead. The aim of this collaboration is to create a multilingual database of proper names containing information applicable to machine translation, computer aided translation, data research as well as spelling dictionaries. These particular aims guided the creation of the ontology whose description will follow. Beside a set of language-dependent and language-independent relations associated with a logical model, the database is founded on a four level ontology: the level of instances (the proper names such as they appear in a written text in a specific language), the linguistic level (the level of so called "prolexemes"), the conceptual level (the numerical pivots) as well as the metaconceptual level (types and supertypes). We will describe here the different levels of the ontology and their implementation in the database using French and German examples.

1. Introduction

Le traitement automatique des langues (désormais TAL) est fortement dépendant des ressources linguistiques adaptées aux applications envisagées et aux méthodes utilisées. Certains composants logiciels n'utilisent que des corpus d'apprentissage pour permettre ensuite la mise en œuvre de traitements statistiques. Mais un grand nombre d'applications possèdent une composante linguistique qui nécessite une liste de mots avec des codes morphologiques, syntaxiques ou autres, autrement dit un dictionnaire électronique. La taille de ces dictionnaires électroniques est très variable suivant que le système utilise un modèle de langue hybride, c'est-à-dire composé d'une partie linguistique et d'une partie stochastique, ou uniquement linguistique.

L'effort de la communauté scientifique TAL a porté jusqu'ici principalement sur des ressources dictionnairiques de noms communs comme le système DELA de dictionnaires électroniques (Courtois & Silberztein, 1990) et sur des ressources terminologiques spécialisées (Sager 1990). Or, une catégorie spécifique de mots, les *noms propres*, est souvent

absente de ces dictionnaires, alors qu'ils constituent une partie importante des textes. En matière de traduction, il semble aussi prévaloir une idée fausse, celle que les noms propres ne se traduisent pas (Grass 2000).

Les avis sur la constitution de ressources spécifiques aux noms propres sont encore plus partagés que pour le vocabulaire commun ou technique. Il est certes possible en français de déclarer *nom propre a priori* tout mot inconnu capitalisé, c'est-à-dire commençant par une majuscule à l'instar de Ren et Perrault (1992), mais la recherche et l'extraction d'informations ou l'aide à la traduction nécessitent de délimiter précisément les noms propres, de les catégoriser et même, parfois, de les relier entre eux.

Dans un premier temps, nous essaierons de définir brièvement ce que nous entendons par nom propre, puis nous évoquerons en détail l'ontologie proprement dite, pour parler enfin des relations entre les noms propres, le tout aboutissant au modèle conceptuel et des données. Nous montrerons ainsi qu'un dictionnaire électronique de noms propres ne peut se contenter d'être une simple liste, mais doit permettre de mettre en relation les noms qu'il contient. Ce dictionnaire peut à la fois être une aide à l'utilisateur (aide à la traduction, correction d'orthographe…) ou un outil de traitement automatique des langues (étiquetage, traitement des coréférences, recherche d'information…).

2. Qu'est-ce qu'un nom propre?

Il n'est pas facile de définir précisément le nom propre et les linguistes ne sont pas unanimes. La définition classique retenue en France est celle du *Bon Usage* (Grevisse & Goosse 1986: 751): "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". Elle est relayée par les propos de Gary-Prieur (1994: 7): "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".

Cette définition correspond bien aux noms de personnes, de pays ou de villes, qui sont souvent sous-jacents à l'idée de noms propres et que Jonasson (1994) appelle des noms propres "purs" par opposition à des noms propres "descriptifs". Ceux-ci résultent souvent de la composition d'un nom propre avec une expansion, comme *Tour Eiffel*. Certains semblent être des descriptions définies figées ou en cours de figement, comme *Jardin des Plantes* ou *Organisation Mondiale de la Santé*. La classe des noms propres purs est relativement fermée, alors que celle des noms propres descriptifs est ouverte à la création lexicale.

Nous prendrons donc comme définition de référence celle de Jonasson (1994: 21) qui déclare *nom propre*: "Toute expression associée

dans la mémoire à long terme à un particulier en vertu d'un lien dénominatif conventionnel stable". Cette définition, plus souple que la précédente, permet d'inclure les noms propres descriptifs, qu'il est convenu dans le monde du *Traitement automatique des langues*, depuis les conférences *MUC*, d'appeler entités nommées (Chinchor 1997).

Une autre caractéristique importante du nom propre est son caractère de "désignateur rigide": selon Kripke (1980), un nom propre, doit "désigner le même particulier dans tous les mondes possibles" et, selon Kleiber (1981: 316), il n'est pas lié "aux situations passagères et aux propriétés accidentelles que peut connaître un particulier". Ainsi le référent d'un nom propre peut évoluer sans que le nom propre utilisé ne soit changé, ce qui permet de distinguer le nom propre (notamment descriptif) de la description définie.

Enfin, en français, comme dans d'autres langues, mais non en allemand, le critère formel le plus remarquable du nom propre, qui le distingue du nom commun ou de la description définie, est la majuscule. Celle-ci "indique le début d'une expression linguistique qui, d'un certain point de vue, constitue un tout" (Gary-Prieur 1991: 22). Mais cette majuscule n'apparaît pas sur tous les termes d'une entité nommée, ce qui pose le problème de sa limite droite: donnons pour exemples *Banque centrale européenne* ou *Banco Real* (*Le Monde* du 15 janvier 1999), dans ce dernier exemple, *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 composé *Banco Real* qu'il faut catégoriser.

3. L'ontologie: une hiérarchie à quatre niveaux

La gestion cohérente d'une base de données passe par la définition d'un modèle d'organisation des noms propres, autrement dit d'une ontologie. Il faut étudier dans ce cadre les différents aspects du domaine des noms propres, identifier les concepts, les relations et les attributs sous-jacents à ce domaine.

Selon Gruber (1995), "Une ontologie est une spécification explicite et formelle d'une conceptualisation faisant l'objet d'un consensus". Ainsi, à partir d'un processus d'abstraction il devient alors possible de définir les concepts, c'est- à- dire les termes essentiels du domaine.

Une définition est proposée par Charlet, Bachimont et Troncy (2003) selon laquelle: "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 – p.ex. entités, attributs, processus -, leurs définitions et leurs interrelations."

A partir d'une étude approfondie dans le domaine des noms propres, notamment à travers les différents mécanismes morphologiques et dérivationnels qui peuvent s'appliquer aux noms propres dans différentes langues et sans oublier les relations qui lient les noms propres entre eux,

nous avons établi une ontologie hiérarchisée en quatre niveaux: le niveau des instances (les noms propres tels qu'ils apparaissent dans les textes dans une langue sélectionnée), le niveau linguistique (les formes canoniques et leurs différentes formes dérivées), le niveau conceptuel (pivot) et le niveau méta-conceptuel (types et supertypes). Nous avons aussi introduit un nom propre conceptuel (le pivot, voir section 3.3).

Chaque niveau peut lui-même être hiérarchisé en plusieurs sous niveaux, chacun disposant de ses propres relations. Les concepts de notre ontologie sont regroupés au sein d'une arborescence et reliés entre eux par des relations de type *is-a*. Il existe deux plans principaux:

Un plan supérieur commun aux langues traitées englobant le niveau conceptuel et méta-conceptuel. Ce plan regroupe les principales relations (Chef, Méronymie, Synonymie synchronique et diachronique) entre les noms propres. Tous les concepts apparaissant ici s'appliquent à toutes les langues. La *Figure 1* présente les différents concepts de notre ontologie.

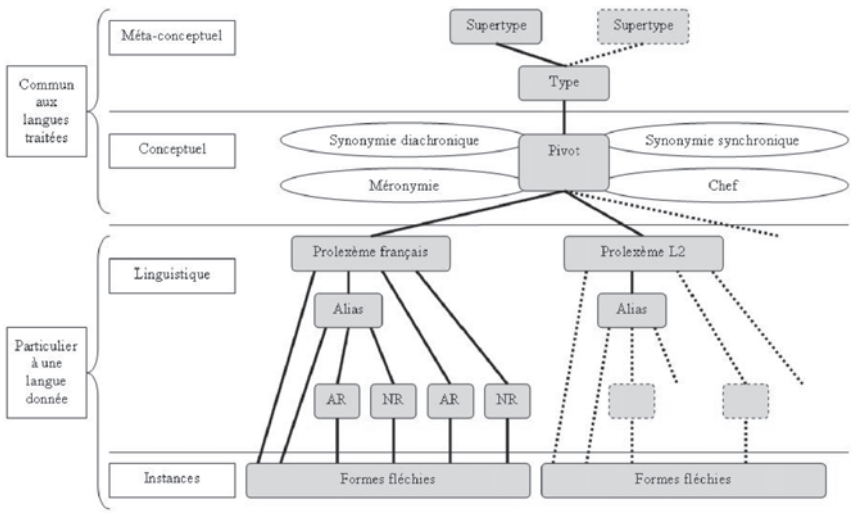

Figure 1: l'architecture en quatre niveaux de Prolexbase

3.1. Le niveau des instances

Le niveau des instances correspond au niveau inférieur et contient l'ensemble des formes fléchies d'un nom propre. En effet, les noms propres, comme les noms communs, suivent des règles morphologiques dont le degré de complexité est fonction de la langue étudiée. Un nom propre dans un texte peut apparaître sous différentes formes appelées mots-formes ou instances. C'est au niveau des instances que l'on retrouve l'ensemble des lexies liées aux noms propres, c'est-à-dire "l'ensemble des mots-formes ou construction linguistique qui ne se distingue que par la

flexion" (Polguère 2002). Ainsi, la lexie *Helvète* regroupe les mots formes *Helvète* (MS et FS) et *Helvètes* (MP et FP).

La flexion d'un nom propre joue un rôle très important dans les langues à cas comme l'allemand et surtout dans les langues slaves comme le polonais, le russe ou le serbe.

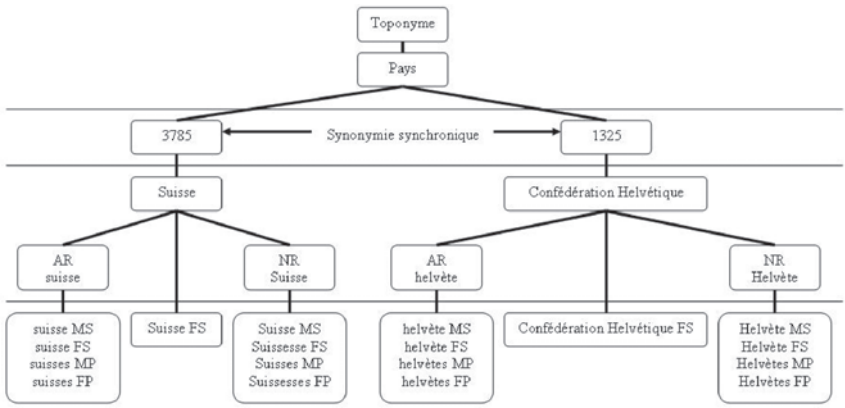

Figure 2: exemple avec le nom propre "Suisse"

3.2 Le niveau linguistique

Le niveau linguistique se décompose en trois sous-niveaux:

- Le "prolexème" correspond à la forme canonique d'un nom propre.
- Les alias représentent les différentes formes de substitution possibles du prolexème.
- Les dérivés des alias ou du prolexème: pour le français, les adjectifs relationnels et les noms relationnels.

Chaque langue disposera de son propre prolexème. Deux homographes dans différentes langues seront dupliqués. Par exemple, il y aura deux entrées *France* dans la base de données, une pour la partie française et une pour la partie anglaise.

Par **relation de polysémie** et **relation d'homonymie**, nous entendons le rattachement de plusieurs types conceptuels à une même forme écrite. Prenons l'exemple des noms propres qui s'écrivent en français et en allemand *Paris*; ils désignent (a) une ville en France (type *noms de villes*), (b) un département (type *noms de régions*), qu'on appelle aussi *la Seine*, (c) d'autres villes à travers le monde (aux Etats-Unis, au Gabon, à Haïti, au Togo, aux Philippines, à São Tomé et au Canada).

Entre les exemples (a) et (b), il s'agit du même lieu, il y a polysémie. Par contre, entre (a) et (c), il y a seulement une identité de désigna-

tion au sein d'un même type (type *noms de villes*), donc homonymie. En allemand, ce mot désigne aussi (d) un personnage mythique, *Pâris* en français. Il y a alors, pour l'allemand, entre (a) et (d), homonymie dans des types différents (type *noms de villes* et type *noms de divinités, de personnages mythiques ou fictifs*). La relation d'homonymie est propre à une langue. L'indication du type s'avère ici discriminante pour la traduction entre l'allemand et le français. La relation de polysémie (pour les noms propres) se conserve lors de la traduction. Les noms propres polysémiques dans la langue de départ le sont également dans la langue d'arrivée.

Autour du prolexème, nous avons inséré des relations comme la relation blark, détermination et expansion (voir section 4.2). Nous avons estimé que l'indice de popularité d'un nom propre varie en fonction de la langue du fait que certains noms propres peuvent être très connus dans un pays alors qu'ils le sont peu ou pas du tout dans un autre. La détermination peut aussi exister dans certaines langues et pas dans d'autres (*la France* = *Frankreich* en allemand). Les expansions d'un nom propre varient aussi énormément d'une langue à l'autre: on trouve en français les expansions *ruisseau*, *torrent*, *rivière*, *fleuve* pour les cours d'eau alors qu'en anglais il n'existe que *river*.

Les alias réfèrent à la structure interne définie par MacDonald (1996), ils peuvent être: des variantes sur les caractères comme la hauteur de casse (*Siemens* ou *SIEMENS*) ou les ligatures (*Elbœuf sur Seine* ou *Elboeuf sur Seine*); des acronymes ou abréviations prononcées de façon syllabique comme dans *ONU* (Organisation des Nations Unies) prononcé [Ony]; des sigles prononcés lettre par lettre comme *ASTM* pour *Action Solidarité Tiers Monde*; des abréviations comme *Microsoft Corp.* pour *Microsoft Corporation*; des transcriptions comme *Чехов* en russe qui peut s'écrire *Tchekhov* ou *Tchékhov* en français.

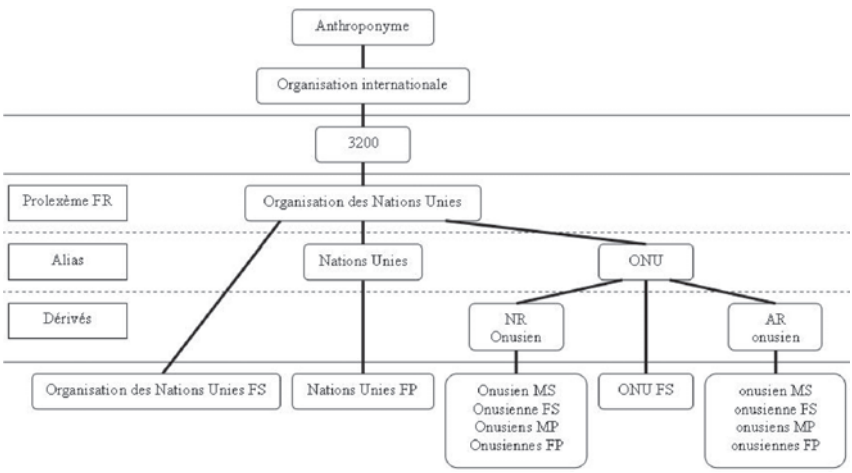

Figure 3: exemple avec le nom propre "Organisation des Nations Unies"

A partir de chaque alias, il est possible de produire pour le français deux types de formes différentes: des adjectifs relationnels et des noms relationnels. Le nombre de formes que l'on pourra générer en fonction d'un alias varie selon la langue.

3.3. Le niveau conceptuel

Le cœur de l'ontologie reste le niveau conceptuel à travers la notion de pivot. Les pivots correspondent à des numéros d'identité uniques (ID) qui sont associés à chaque prolexème de chaque langue, tels qu'ils sont définis au niveau linguistique. Le niveau conceptuel est un niveau inter-langue. Ainsi, chaque nom propre représentant le même concept possèdera le même pivot. Un nom propre n'est la traduction d'un autre nom propre dans une autre langue que si tous deux partagent le même pivot. Nous avons décidé d'intégrer dans le niveau conceptuel des relations entre les noms propres comme la relation Chef, la méronymie, la synonymie synchronique ou diachronique (cf. section 4.1.) Ces relations entre les noms propres sont communes à chaque langue. Afin d'éviter de créer autant de relations que de langues, nous avons estimé qu'il était préférable de les insérer au niveau conceptuel.

3.4. Le niveau méta-conceptuel

La désambiguïsation sémantique passe par l'utilisation de classifieurs et a priori, plus le système de ces classifieurs est élaboré et consistant, plus il est efficace. Pourtant, il se révèle à l'usage qu'un système de traduction automatique comme Reverso avec dix descripteurs sémantiques fonctionne mieux que Systran qui en utilise plus de cent. Ce qui pose problème apparemment pour Systran, c'est la non consistance du système, ce que relevait déjà Gross (1994: 22).

La typologie que nous avons créée résulte de l'absence d'une typologie adaptée au traitement linguistique. Nous avons étudié différentes classifications concernant les noms propres: Bauer (1985), MUC (Chinchor 1997), Paik *et al*. (1996), Zabeeh (1968) et Ballard (2001). Il s'agit toujours de classifications *ad hoc* dont les critères de regroupement sont créés en fonction des objectifs recherchés qui sont la plupart du temps extralinguistiques. Nous ne prétendons pas fournir ici une classification idéale, mais une classification des noms propres cohérente du point de vue du TAL. Notre typologie est à deux niveaux:

Les *supertypes*, qui reprennent en fait les traits syntactico-sémantiques classiques: les anthroponymes (trait *humain*), les toponymes (trait *locatif*), les ergonymes (trait *inanimé*) et les pragmonymes (trait *événementiel*). Les *types*, qui sont plus des *commentaires* portés sur les supertypes, commentaires destinés principalement à l'aide à la traduction ou à la recherche d'information.

Nous avons testé la pertinence de cette typologie en classant manuellement tous les noms propres contenus dans un numéro du journal *Le Monde*. Nous avons aussi recherché des noms propres à partir d'indices et de verbes supports dans le corpus allemand *COSMAS*. Nous avons aussi développé, en français, des outils de reconnaissance et de typage automatiques, principalement pour les deux premiers supertypes (Friburger & Maurel, 2001; Friburger, 2002). Ces outils sont en cours d'adaptation pour l'allemand. Les types sont reliés aux supertypes par la relation d'hyponymie, une des relations de base de *wordnet*, ce qui signifie que les supertypes sont des hyperonymes de plusieurs types et les types des hypéronymes des noms propres conceptuels (les pivots). Cette relation entre les types et les pivots est indépendante de la langue. Chaque nom propre est associé à un seul type, sinon il y a homonymie et duplication des pivots, c'est le cas pour *Washington* en tant que ville/toponyme et *Washington* en tant que célébrité/anthroponyme. Nous ne décrirons pas ici l'ensemble des types, description que l'on retrouvera de façon exhaustive dans Grass (2002), mais nous nous limiterons à les présenter dans le tableau suivant.

Types	**Exemples et traduction allemand-français**
Supertype anthroponyme (trait humain)	
Patronyme	*Dupont = Dupont*
Prénom	*Jacques = Jacques*
Célébrité	*Sokrates = Socrate*
Personnage mythique ou fictif	*Vishnu = Vichnou*
Association ou parti	*die Grünen = les Verts allemands*
Ensemble artistique, club sportif	*die Beatles = les Beatles* *die Borussia Dortmund = le Borussia de Dortmund*
Organisation internationale	*die NATO = l'OTAN*
Entreprise	*General Electric = General Electric*
Etablissement public ou privé	*Universität Freiburg = Université de Fribourg*
Pseudo-anthroponyme	
Supertype toponyme (trait locatif)	
Pays	*Belgien = la Belgique*
Région	*Ostflandern = la Flandre-Orientale*
Ville	*Antwerpen = Anvers*
Quartier, voie	*der Rote Platz = la Place rouge*
Bâtiment	*das Weisse Haus = la Maison blanche*
Groupe de pays	*die Europäische Union = l'Union européenne*
Hydronyme	*die Schelde = l'Escaut*
Géonyme	*die Ardennen = les Ardennes*
Objet céleste	*der Saturn = Saturne*
Lieu mythique ou fictif	*Atlantis = l'Atlantide*

Supertype ergonyme (trait inanimé)	
Appellation commerciale: marque & produit	*Seat Cordoba = Seat Cordoba*
Œuvre: roman, pièce de théâtre, tableau, ...	*Die Mona Lisa = La Joconde*
Objet mythique ou fictif	*der Gral = le Graal*
Vaisseau	*Ariane 5 = Ariane 5*
Supertype pragmonyme (trait événementiel)	
Evénement historique ou politique	*Waterloo = Waterloo*
Phénomène météorologique	*El Niño = El Niño*
Catastrophe	*Tschernobyl = Tchernobyl*
Manifestation artistique ou sportive	*die Champions league = la Ligue des champions*
Fête	*Ostern = Pâques*

Figure 4: liste des types et supertypes

4. Les relations entre noms propres

Ce dont on a le plus besoin en TAL, c'est de disposer de lexiques à large couverture qui expliquent suffisamment de propriétés linguistiques et de relations entre les entrées pour qu'on puisse les utiliser comme instruments de traitement non seulement de la phrase simple, mais aussi de la phrase complexe et du discours. Etant donné une instance identifiée dans un type quelconque, les informations linguistiques intéressantes découlent des deux questions suivantes: Quels sont ses synonymes ou ses mérony-mes, c'est-à-dire les autres noms propres qui peuvent se substituer à cette instance par anaphore? Quelles sont les principales expansions qui caractérisent les éléments d'un type donné et qui peuvent être repris par anaphore lexicale?

Il apparaît nécessaire en TAL de disposer à la fois d'une représentation explicite des relations sémantiques au niveau interlingual ainsi que de relations transformationnelles entre structures équivalentes sémantiquement au niveau d'une langue donnée.

4.1. Les relations indépendantes de la langue

Pour définir les relations sémantiques dans notre base de données multilingue de noms propres, nous nous sommes inspirés des fonctions lexicales de Mel'čuk (1984, 1988, 1992) et du système *Wordnet* de Miller *et al*. (1990), également implémenté sous une forme multilingue dans *EuroWordnet* (Vossen 1998). Notre modèle comprend aujourd'hui cinq

relations de type sémantique: *synonymie synchronique, synonymie diachronique*, *méronymie*, *hyponymie* et *chef.* La relation de *méronymie* a été empruntée au système *Wordnet*, la relation *chef* est une fonction lexicale. Toutes ces relations permettent de situer un nom propre dans un réseau lexical. Rappelons que les relations sémantiques définies dans *Wordnet* s'appliquent sur les *synsets* et non sur les noms eux-mêmes. Pour les noms propres, qui ont une fonction essentiellement référentielle, il existe un certain nombre de sèmes sous-jacents aux types définis ci-dessus et par rapport auxquels les relations s'organisent. C'est aussi à cause de cette fonction référentielle que d'autres relations sémantiques, comme l'antonymie ou la relation converse, telles qu'elles sont définies chez Lyons (1971: 455), ne s'appliquent pas aux noms propres.

Les relations sémantiques se situent au niveau conceptuel et concernent deux pivots. Nous allons développer ces relations. La **synonymie diachronique** est une relation qui envisage un nom propre dans son histoire. Cette relation concerne les noms propres qui ont été renommés pour des raisons historiques. C'est le cas de la ville *Karl-Marx-Stadt* de l'ancienne *RDA* qui a repris son nom d'origine *Chemnitz* après la réunification de l'Allemagne. La **synonymie synchronique** est une relation entre deux pivots. La relation de synonymie, contrairement à une idée reçue concerne de nombreux noms propres. Par exemple, presque tous les noms de pays ont en français une forme longue et une forme courte (*France* et *République française* en français *Bundesrepublik Deutschland* et *Deutschland* en allemand). Mais ils sont seulement en relation de synonymie synchronique dans un contexte politique, on ne dit pas dans un registre courant "j'ai passé mes vacances en République française". La **méronymie** est une des relations principales du système *Wordnet* développée à partir des travaux de Miller (*et al*. 1990). C'est aussi une relation fondamentale dans la structuration du dictionnaire mental. La relation de méronymie est une relation d'inclusion, une relation partie-tout (x est un méronyme de y *si* x est une partie de y). Cette relation génère une hiérarchie à plusieurs niveaux entre éléments contenants ou *holonymes* et éléments contenus ou *méronymes*. La relation de méronymie relie certains types entre eux, comme les villes, les régions et les pays: *Tours* $\subset$ *Indre-et-Loire* $\subset$ *Région Centre* $\subset$ *France.* Bien que ne jouant pas de rôle déterminant au niveau de la traduction, cette relation est précieuse pour la recherche d'informations. **Chef** est une fonction lexicale (appelée Cap) que l'on trouve dans le *Dictionnaire Explicatif et Combinatoire du français contemporain* (Mel'čuk 1984, 1988, 1992). Cette relation est tout à fait originale et désigne une entité qui se trouve à la tête de quelque chose. De par sa nature *a priori* humaine, cette relation ne devrait concerner que des humains par rapport à d'autres humains (souvent collectifs), mais elle intervient aussi en toponymie (on parle bien de *chef-lieu* en français), avec les noms d'entreprises (société mère et filiale), d'où de nombreuses implications en matière de recherche d'informations. On utilise souvent en politique internationale cette relation

entre un pays et sa capitale, cette dernière étant censée représenter le pays:

> Les Britanniques, qui ont eu des échanges avec Washington sur leur nouvelle initiative, pensent que les États-Unis pourraient revoir leur position.
> (*Le Monde,* 23 octobre 1998)

Une spécificité du français est de désigner une haute personnalité officielle par le nom de l'endroit où elle exerce ses fonctions: *Palais de l'Élysée* peut être, dans un registre politique, un synonyme de *Président de la République* et *Matignon* de *Premier ministre*. On peut concevoir des phrases avec un verbe exigeant un agent et un patient humain du type "*l'Élysée n'a pas consulté Matignon"* où des toponymes synonymes d'anthroponymes perdent leur sémantisme locatif pour acquérir le trait *humain*. Nous remarquerons au passage que dans cette phrase, on peut remplacer *l'Élysée* ou *Matignon* par un titre (*le Président* et *le Premier ministre*) ou par un nom propre actualisant l'énoncé (par exemple *Chirac* et *Raffarin*). Une telle possibilité n'existant pas en allemand, une traduction exige une transposition où la relation de synonymie qui existe entre *Matignon* et *Premier ministre* disparaît.

4.2. Les relations dépendantes de la langue

Dans notre dictionnaire, nous avons décrit trois relations dépendantes de la langue: *détermination, blark* et *expansions typiques.* Ces relations sont importantes du point de vue traductionnel dans la mesure où les données qu'elles véhiculent varient d'une langue à l'autre.

La détermination concerne un grand nombre de noms propres. En allemand comme en français et dans de nombreuses langues, le déterminant est possible avec tous les noms mais il est soumis à des restrictions pour les noms propres. Notons que le déterminant peut être absent dans une langue et présent dans une autre comme pour *Spanien = l'Espagne*. La traduction automatique devra bien entendu rétablir le déterminant là où il est nécessaire ou le supprimer lorsqu'il n'a pas lieu d'être.

Blark (Basic LAnguage Ressources Kit) (Cucchiarini et al. 2000) est un indicateur de notoriété pour un nom propre. Il existe une catégorie de noms propres dont la notoriété est liée à un patrimoine culturel national ou international (*Mozart, Bouddha, Socrate, Tokyo, Paris*…), alors que d'autres semblent avoir une notoriété liée à l'actualité. La notoriété d'un nom propre varie en fonction du lieu et du temps. Les noms propres indispensables pour l'étude d'un corpus journalistique d'une année donnée dans un pays donné peuvent se révéler inutiles quelques années plus tard ou dans un autre pays.

Certaines expansions à droite ou à gauche du nom propre sont véritablement typiques et notées dans la base. La traduction des noms propres implique souvent la prise en compte du contexte élargi ou expan-

sions et notamment des amorces (expansions à gauche). Les amorces permettent un étiquetage et la création automatique de dictionnaires.

Novell-Chef Ray Noorda = le patron de Novell Ray Noorda

5. Modèle Conceptuel de Données (MCD)

L'ontologie une fois définie, il reste à la mettre en œuvre dans un modèle conceptuel de données destiné à définir la structure de la base de données en tant que telle. Ce MCD est constitué d'une partie commune à toutes les langues et d'une partie distincte pour chacune des langues traitées. Nous donnerons ici à titre d'exemple le MCD français avec ses 11 entités et 16 associations, nous ne nous étendrons pas sur la partie commune, ce qui dépasserait le cadre qui nous est imparti ici.

Chaque entité du MCD a un identifiant abstrait indépendant de tout autre attribut. Cet identifiant commence par "NUM_" (ex: NUM_PIVOT, NUM_TYPE…) et correspond à un numéro séquentiel qui sera incrémenté au fur et à mesure des insertions.

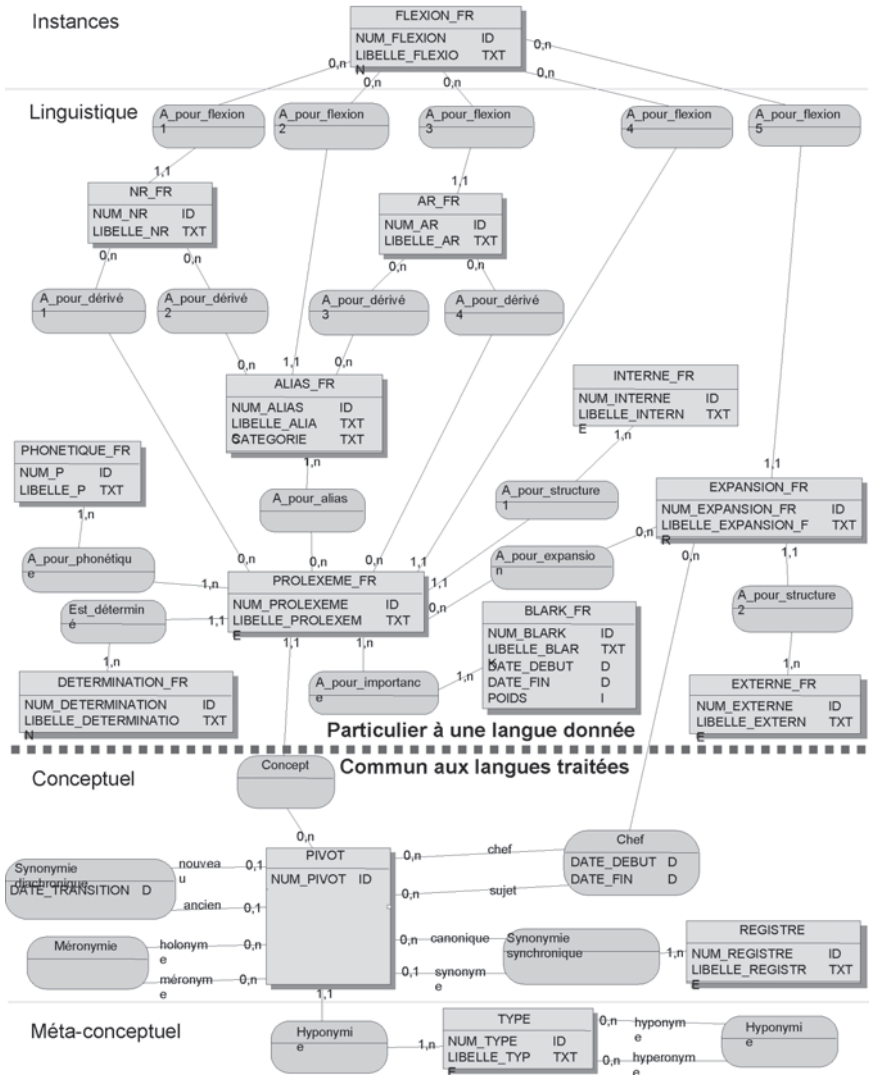

Figure 5: le MCD français.

5.1. Les entités

PROLEXEME_FR: Cette entité regroupe les formes canoniques d'un nom propre. Dans l'exemple de la figure 3, le nom propre "Organisation des Nations Unies" sera conservé dans le champ LIBELLE_NOMPROPRE.

ALIAS: C'est dans cette entité que l'on retrouvera les formes qui sont liées à la forme canonique d'un nom propre. On pourra définir des

règles de construction des alias ou bien lister l'ensemble des alias. L'attribut CATEGORIE permet de préciser la nature de l'alias: *variante*, *acronyme*, *abréviation*, *transcription*.

FLEXION: Les flexions peuvent être incluses dans la base ou considérées comme un module indépendant de la base s'il existe déjà des outils capable de générer les flexions. Pour le français, on utilisera les codes du DELAS pour les flexions des mots simples.

PHONETIQUE: L'entité PHONETIQUE est caractérisée par l'attribut LIBELLE_PHONETIQUE dans lequel on va enregistrer la phonétique d'un prolexème. Nous avons décidé de n'ajouter dans notre base de données que la phonétique des prolexèmes.

BLARK: Chaque langue définira sa propre échelle de notoriété qui sera stockée dans le LIBELLE_BLARK. Chaque nom propre disposera d'un indice de notoriété avec une date de début (DATE_DEBUT) et une date de fin (DATE_FIN) pour les noms propres ayant une notoriété liée à l'actualité. Il sera aussi possible d'attribuer un poids, en utilisant l'attribut POIDS.

DETERMINATION: L'entité détermination en français offre deux choix possibles (*oui* et *non*). Pour l'anglais, cette entité pourrait ne pas exister.

NOM_RELATIONNEL: Cette entité permet de lister les règles de formation des noms relationnels ou les noms relationnels eux-mêmes. Pour des langues où il existerait des outils capables de générer les noms relationnels à partir d'un nom propre canonique, il ne sera pas nécessaire de les stocker dans la base les règles de formation des noms relationnels ou les listes de noms relationnels.

ADJECTIF_RELATIONNEL: On pourra stocker des règles de formation des adjectifs relationnels ou stocker les adjectifs relationnels sous formes de liste. Pour des langues où il existerait des outils capables de générer les adjectifs relationnels à partir d'un nom propre canonique, il ne sera pas nécessaire de les stocker dans la base les règles de formation des adjectifs relationnels ou les listes de adjectifs relationnels.

EXPANSION: Cette entité liste les expansions des noms propres. Par exemple pour le français: *région*, *capitale*, *ville.*

EXTERNE: Cette entité renseigne sur la construction d'un nom propre avec son expansion. Les valeurs possibles pour cette table dans le cas du français sont: *CONTEXTE NP*, *CONTEXTE DE LE NP*, *CONTEXTE DE NP*.

La ville de Paris = *CONTEXTE DE NP*

INTERNE: Elle fournit des informations sur la structure interne d'un nom propre. Cette entité possède un champ NUM_INTERNE et un champ LIBELLE_INTERNE.

5.2. Les associations

Est_déterminé est une relation de type 1: N. Chaque prolexème a une seule et unique détermination. Chaque détermination est en relation avec un ou plusieurs prolexèmes.

A_pour_phonétique est une relation de type N: M. Un prolexème peut avoir une ou plusieurs phonétiques. C'est notamment le cas de la ville de Metz qui possède en français deux prononciations différentes [mɛts] ou [mɛs]. Comme il est difficile de savoir si l'une des deux prononciations est plus utilisée que l'autre, nous avons décidé de lister toutes les phonétiques possibles pour un prolexème donné. Il peut arriver que deux prolexèmes, bien qu'écrits de manières différentes, se prononcent de la même façon.

A_pour_expansion est une relation de type N: M. Chaque prolexème peut avoir zéro ou plusieurs expansions. Une expansion concerne zéro ou plusieurs prolexèmes.

A_pour_alias est une relation de type N: M qui lie un alias à un prolexème. Un alias peut provenir d'un ou plusieurs prolexèmes. Cela est dû au fait que l'entité ALIAS_FR peut regrouper des règles de générations d'alias. Il peut arriver qu'une même règle de générations d'alias puisse concerner plusieurs prolexèmes. Un prolexème peut engendrer au minimum zéro et au maximum plusieurs alias.

A_pour_flexion1 est une relation de type 1: N. Une flexion peut être appliquée à zéro ou plusieurs noms relationnels. Un nom relationnel ne suit qu'une seule et unique flexion. De même pour les relations : A_pour_flexion2, A_pour_flexion3, A_pour_flexion4, A_pour_flexion5.

A_pour_structure1 est une relation de type 1: N. Un prolexème a une seule et unique structure interne. Chaque structure interne est reliée à un ou plusieurs prolexèmes.

A_pour_structure2 est une relation de type 1: N. Une expansion possède une seule et unique structure externe. Chaque structure externe est reliée à une ou plusieurs expansions.

A_pour_dérivé est une relation de type N: M. Le MCD français possède quatre relations A_pour_dérivé (A_pour_dérivé1, A_pour_dérivé2, A_pour_dérivé3 et A_pour_dérivé4) qui relient un prolexème ou un alias à un adjectif relationnel ou un nom relationnel.

A_pour_importance est une relation de type N: M. Un prolexème a un ou plusieurs blarks. Chaque blark est lié à un ou plusieurs prolexèmes.

6. Conclusion

Cet article présente les aspects principaux de notre projet. Notre modélisation repose sur une ontologie structurée en quatre niveaux. Nous avons introduit dans ce modèle la notion de pivot représentant soit un nom pro-

pre conceptuel, soit un référent, selon le point de vue. Les relations s'organisent autour de ce pivot.

Notre dictionnaire a été implémenté sous la forme d'une base de données relationnelles, ce qui permet de l'utiliser pour du traitement automatique (recherche ou extraction d'information, analyse, traduction automatique, alignement de corpus, etc.), mais aussi pour une consultation humaine ou une aide à la rédaction ou à la traduction. Il suffit pour cela de générer différents états ou de créer une interface permettant des requêtes appropriées. Actuellement, la partie française du dictionnaire comporte plus de 323.000 entrées et 55.000 relations; la partie allemande comprend environ 13.000 entrées avec un lien de traduction vers le français.

Bibliographie

Ballard, M. (2001). *Le nom propre en traduction*. Paris: Ophrys.

Bauer, G. (1985). *Namenkunde des Deutschen*. Bern: Germanistische Lehrbuchsammlung.

Charlet J., B. Bachimont & R. Troncy (2003). "Ontologie pour le Web sémantique." Rapport final de l'Action spécifique 32 CNRS/STIC.

Chinchor N. (1997). *Muc-7 Named Entity Task Definition*. On line at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendices.

Courtois B. & M. Silberztein. (1990). "Dictionnaires électroniques du français." *Langue française* 87, 11-22.

Cucchiarini C., W. Daelemans & H. Strik (2000). "Strengthening the Dutch Human Language Technology Infrastructure." On line at: http://www.elda.fr/article48.html.

Friburger, N. (2002). *Reconnaissance automatique des noms propres; application à la classification automatique de textes journalistiques*, thèse de doctorat en informatique, Université de Tours.

Friburger, N. & D. Maurel (2001). "Elaboration d'une cascade de transducteurs pour l'extraction de motifs: l'exemple des noms de personnes." *Huitième conférence annuelle sur le traitement automatique des langues naturelles (TALN 2001),* 183-192.

Gary-Prieur, M.-N. (1994). *Grammaire du nom propre.* Paris: Presse Universitaire de France.

Grass, T. (2000). "Typologie et traductibilité des noms propres de l'allemand vers le français." TAL 41(3), 643-669.

Grass, T. (2002). *Quoi ! Vous voulez traduire "Goethe"? - Essai sur la traduction des noms propres allemand – français*. Bern: Peter Lang.

Grevisse, M. & A. Goosse (1986). *Le Bon Usage*. Gembloux: Duculot.

Gross, G. (1994). "Classes d'objets et descriptions des verbes." *Langages* 115, 15-31.

Gruber T. R. (1995). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." *Int. Journal of Human Computer Studies* 43, 907-928.

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

Kleiber, G. (1981). *Problèmes de référence: descriptions définies et noms propres*. Paris: Klincksieck.

Kripke, S. (1980). *Naming and Necessity.* Cambridge: Harvard University Press.

Lyons, J. (1971). *Introduction to Theoretical Linguistics*. Cambridge University Press.

MacDonald, D. (1996). "Internal and external evidence in the identification and semantic categorisation of Proper Names." *Corpus Processing for Lexical Acquisition*, Massachussetts Institute of Technology, 21-39.

Mel'čuk, I. (1984-I, 1988-II, 1992-III). *Dictionnaire explicatif et combinatoire du français contemporain*. Presses de l'Université de Montréal.

Miller, G., R. Beckwith, C. Fellbaum., D. Gross and K. Miller (1990). "Introduction to WordNet: an on-line lexical database." *International Journal of Lexicography* 3, 235-244.

Paik W., E. D. Liddy, E. Yu, & M. McKenna (1996). "Categorizing and Standardi-zing Proper Nouns for Efficient Information Retrieval." *Corpus Processing for Lexical Acquisition*, Massachussetts Institute of Technology, 61-73.

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

Ren X. & F. Perrault (1992). "The typology of Unknown Words: An Experimental Study of Two Corpora." *COLING 92*, Nantes.

Sager J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Tran, M., T. Grass & D. Maurel (2004). "An Ontology for Multilingual Treatment of Proper Names." OntoLex 2004, Lisbonne.

Vossen, P. (1997). "EuroWordNet: a multilingual database for information retrieval." *DELOS workshop on Cross-language Information Retrieval*, Zurich. http://www.ercim.org/publication/ws-proceedings/DELOS3/

Zabeeh, F. (1968), *What's in a Name, An Inquiry into the Semantics and Pragmatics of Proper Names*. La Haye: Martinus Nijhoff.