# Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation

Maud Ehrmann

**UNIVERSITE PARIS 7 - DENIS DIDEROT**

ÉCOLE DOCTORALE DE SCIENCES DU LANGAGE

**LaTTICe**

LAngues, Textes, Traitements Informatiques, Cognition

**DOCTORAT**

Linguistique théorique, descriptive et automatique

Maud EHRMANN

# LES ENTITES NOMMÉES, DE LA LINGUISTIQUE AU TAL : Statut théorique et méthodes de désambiguïsation.

Soutenue le 2 Juin 2008

**JURY :**

| | | |
|---|---|---|
| Mme Laurence DANLOS | Présidente | (Université Paris 7 Diderot- Lattice) |
| Mme Adeline NAZARENKO | Rapporteure | (Université Paris-Nord 13 - LIPN) |
| M. Pierre ZWEIGENBAUM | Rapporteur | (Université Paris-Sud 11 - LIMSI) |
| Mme Caroline BRUN | Examinatrice | (Xerox Research Centre Europe) |
| M. Marcel CORI | Examinateur | (Université Paris 10 Nanterre - Modyco) |
| M. Bernard VICTORRI | Directeur | (CNRS - Lattice) |

Thèse réalisée dans le cadre d'une convention CIFRE (Association Nationale de la Recherche Technique) au sein du Centre de recherche Xerox (XRCE) à Grenoble.

# Résumé

Le traitement des entités nommées fait aujourd'hui figure d'incontournable en Traitement Automatique des Langues. Apparue au milieu des années 1990 à la faveur des dernières conférences MUC (*Message Understanding Conferences*), la tâche de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. apparaît en effet comme fondamentale pour diverses applications participant de l'analyse de contenu et nombreux sont les travaux se consacrant à sa mise en œuvre, obtenant des résultats plus qu'honorables. Fort de ce succès, le traitement des entités nommées s'oriente désormais vers de nouvelles perspectives avec, entre autres, la désambiguïsation et une annotation enrichie de ces unités. Ces nouveaux défis rendent cependant d'autant plus cruciale la question du statut théorique des entités nommées, lequel n'a guère été discuté jusqu'à aujourd'hui.

Deux axes de recherche ont par conséquent été investis durant ce travail de thèse : nous avons, d'une part, tenté de proposer une définition des entités nommées et, d'autre part, expérimenté des méthodes de désambiguïsation. À la suite d'un état des lieux de la tâche de reconnaissance de ces unités et d'un exposé des difficultés pouvant se présenter à l'occasion d'une telle entreprise, il fut avant tout nécessaire d'examiner, d'un point de vue méthodologique, comment aborder la question de la définition des entités nommées. La démarche adoptée invita à se tourner du côté de la linguistique, avec les noms propres et les descriptions définies, puis du côté du traitement automatique, ce parcours visant au final à proposer une définition tenant compte tant des aspects du langage que des capacités et exigences des systèmes informatiques. La suite du mémoire rend compte d'un travail davantage expérimental, avec l'exposé d'une méthode d'annotation fine tout d'abord, de résolution de métonymie enfin. Ces travaux, combinant approche symbolique et approche distributionnelle, rendent compte de la possibilité d'une double annotation (catégories générales et catégories fines) et d'une désambiguïsation des entités nommées.

*Introduced as part of the last Message Understanding Conferences dedicated to information extraction, Named Entity extraction is a well-studied task in Natural Language Processing. The recognition and the categorization of person names, location names, organisation names, etc. is regarded as a fundamental process for a wide variety of natural language processing applications dealing with content analysis and many research works are devoted to it, achieving very good results. Following this success, named entity treatment is moving towards new research propects with, among others, disambiguation and fined-grained annotation. However, this new challenges make even more crucial the question of named entity definition, which was not much discussed until now.*

*Two main lines were explored during this PhD project : first we tried to propose a definition of named entities and then we experimented disambiguation methods. After a presentation and a state of the art of the named entity recognition task, we had to examine, from a methodological point of view, how to tackle the question of the definition of named entities. Our approach led us to study, firstly, the linguistic side, with proper names and definite descriptions and, secondly, the computing side, this development aiming at, finally, proposing a named entity definition that takes into account language aspects but also informatic systems capacities and requirements. The continuation of the dissertation is about more experimental works, with a presentation of experiments about fined-grained named entity annotation and metonymy resolution methods.*

# Remerciements

Je tiens à remercier Bernard Victorri pour avoir accepté de diriger cette thèse, pour son soutien, sa disponibilité et les nombreuses discussions qui m'ont permis d'y voir plus clair au sujet du sens, de la réference et, bien sûr, des entités nommées.

Un grand merci également à Caroline Brun qui, depuis mon premier stage à Xerox et tout au long de ces trois années de thèse, m'a accompagnée, conseillée et surtout, surtout, fait confiance.

Je souhaite remercier Laurence Danlos pour avoir accepté de faire partie du jury et de le présider. Merci également aux rapporteurs, Adeline Nazarenko et Pierre Zweigenbaum, pour leurs lectures attentives et leurs remarques constructives ; merci enfin à Marcel Cori pour avoir consacré du temps à l'examen de ce document.

Merci à Frédérique Segond pour m'avoir chaleureusement accueillie dans son équipe et donné l'occasion de réaliser cette thèse.

Merci à Bernard Combettes qui, il y a bien longtemps, m'a donné mon premier cours de linguistique et donné l'envie de continuer dans cette voie.

Merci à tous les membres de *ParSem*, pour leur soutien et leur bonne humeur. Guillaume Jacquet, pour son aide, ses conseils et les nombreuses discussions qui n'ont cessé de ponctuer notre collaboration ; « Monsieur Xip » (Claude Roux) pour ses coups de pouce sur des programmes en situation critique ; Caroline Hagège pour les nombreux cafés matinaux, pleins de gaîté et d'encouragements.

Merci à Hervé Déjean et Xavier Tannier, pour avoir eu la patience d'écouter mes hésitations, de relire mon travail et de me conseiller.

Enfin, merci à tous ceux qui, de près ou de loin, m'ont encouragée et soutenue tout au long de ce travail. Une pensée pour tous, et un mot pour : mes parents, toujours rassurants ; Ysée, amie et remonteuse de moral exceptionnelle ; Romain, qui, tout simplement, et c'est déjà beaucoup, était là.

# Table des matières

# Introduction générale

## Enjeux

Au lendemain d'une révolution technologique majeure née de la conjonction de la numérisation de l'information et de sa mise en réseau à un niveau planétaire, nombre des activités humaines se trouvent modifiées, au sein d'une société désormais dite de l'information ou de la connaissance. Qu'il s'agisse de travailler, d'apprendre, d'être citoyen ou encore de se divertir, tout un chacun est en effet amené à créer, échanger et rechercher des informations au moyen de documents numériques. Par le biais de cet espace privilégié de communication qu'est le réseau Internet, est ainsi produite et « consommée » une quantité sans cesse croissante de documents. Au regard de la diversité et de l'étendue de ce matériau qui l'occupe, cette activité soutenue de production-consommation d'information exige la mise au point de systèmes automatisés, à l'élaboration desquels participent plusieurs disciplines, dont le Traitement Automatique du Langage Naturel (TALN ou TAL) pour les documents comportant des données linguistiques.

Cet essor de la documentation électronique, s'il n'est bien sûr pas à l'origine de la recherche en Traitement Automatique des Langues (amorcée dès les années cinquante, notamment aux États-Unis), a pour le moins fortement dynamisé la recherche dans ce domaine, notamment au regard des travaux s'intéressant aux données de nature textuelle. C'est en effet à la faveur d'une demande accrue d'outils de traitement de l'information que de nouvelles méthodes ont été réfléchies, de nouvelles ressources constituées et de nouvelles applications mises au point. Au sein de cette dynamique, l'enjeu majeur est, à l'heure actuelle, de capter l'information portée par les textes et d'accéder à leur sens. C'est ainsi que, la composante syntaxique ayant été, dans une certaine mesure, « traitée » et menée à un niveau de maturité durant cette dernière décennie, la composante sémantique est aujourd'hui dominante dans les travaux de TAL.

Que recouvre cette composante sémantique ? Parler du sens dans le traitement automatique des langues n'est pas chose facile, à l'instar de parler du sens tout court. En linguistique, la notion de sens renvoie de manière générale au phénomène

de compréhension d'une unité linguistique, mot, énoncé ou texte. Ce processus de compréhension est complexe (jouant à tous les niveaux : lexical, syntaxique, sémantique et pragmatique) et fait intervenir diverses dimensions (linguistique, cognitive, psychologique, culturelle, etc.). Ces phénomènes impliqués dans le processus de compréhension du sens d'un texte, si nombreux et complexes qu'ils soient, peuvent néanmoins être objectivés et faire l'objet d'une étude scientifique. Pour ce qui est du traitement automatique des langues, il s'agit en quelque sorte de reproduire ce processus de compréhension et de représenter, dans un format informatique exploitable, le sens d'une unité linguistique, ou une des composantes de ce sens. La composante sémantique en TAL correspond donc, de manière générale, à la mise en œuvre de traitements permettant de représenter formellement le sens véhiculé par un texte, ces traitements s'appliquant de manière complémentaire sur des unités de niveaux différents.

Le cours de l'histoire, ou plutôt de la recherche, a voulu que l'on désigne un certain nombre de ces unités sous le nom d' « entités nommées » . Ces dernières correspondent traditionnellement à l'ensemble des noms propres présents dans un texte, qu'il s'agisse de noms de personnes, de lieux ou d'organisation, ensemble auquel sont souvent ajoutées d'autres expressions comme les dates, les unités monétaires, les pourcentages et autres. Contemporain des travaux en Extraction d'Information initiés au début des années 1990, le traitement des entités nommées s'articule en deux processus : *identification* ou reconnaissance de ces unités dans les textes tout d'abord, *catégorisation* ou typage selon des catégories sémantiques larges prédéfinies ensuite. C'est sur ce type d'unités et leur traitement que ce mémoire se propose de se concentrer.

## Contexte

La tâche de reconnaissance des entités nommées a fait cette dernière décennie l'objet d'une attention plus soutenue et suscite aujourd'hui un intérêt certain ; elle apparaît en effet comme fondamentale pour diverses applications de TAL participant de l'analyse de contenu, à l'instar de la recherche et l'extraction d'information, la tâche de question-réponse, le résumé automatique ou encore le fonctionnement des moteurs de recherche, et nombreux sont les travaux se consacrant à cette tâche, obtenant des résultats plus que probants, et ce pour diverses langues. Aussi, il est désormais possible d'affirmer qu'il s'agit d'un des incontournables du traitement automatique des textes.

Forte de son succès, cette tâche s'oriente dorénavant vers de nouvelles perspectives de recherche, selon plusieurs orientations. La plus évidente peut se ré-

sumer par la formule suivante : annoter plus et plus précisément. Orientation « pratique » s'il en est, cette dernière touche aux objectifs et méthodes de traitement automatique des entités nommées. S'agissant des objectifs, il existe en effet une volonté d'améliorer et d'enrichir l'annotation des entités nommées avec, d'une part, l'annotation de types de plus en plus diversifiés d'entités et, d'autre part, une annotation plus fine allant au-delà des catégories générales désormais aisément reconnues, c'est-à-dire une annotation indiquant précisément, du point de vue de la référence, les caractéristiques distinctives de l'entité considérée et permettant (si besoin) sa désambiguïsation. Pour ce qui est des méthodes, les recherches actuelles bénéficient de la disponibilité de corpus de très grande taille, ces derniers pouvant servir à la mise au point de systèmes d'annotation ou à la constitution de ressources.

Une autre orientation possible nous semble se situer davantage du côté théorique. Notre tout premier questionnement, né à l'occasion d'un stage sur l'évaluation d'un module de reconnaissance d'entités nommées (qu'est-ce donc que ces unités ?) fut par la suite conforté par quelques lectures laissant filtrer la même impression tout d'abord, l'absence de réponse ensuite, et un appel de la revue *Linguisticae Investigationes*[1] enfin, tout ceci témoignant de la nécessité d'une réflexion autour de la notion d'entité nommée, avec la question de leur définition. En effet, répondant expressément à des besoins de TAL, force est de constater que ces unités linguistiques ne bénéficient, à l'heure actuelle, d'aucune véritable assise théorique dans la littérature. Il semblerait qu'il s'agisse d'un des premiers « retours à l'envoyeur » du TAL vis-à-vis de la théorie linguistique, amenée aujourd'hui à considérer un objet qu'elle n'avait nullement défini auparavant. D'évidence moins manifeste, cette orientation n'en est pas moins décisive quant à la bonne réalisation de la première.

## Contribution

Notre travail s'est prioritairement consacré à clarifier la notion d'entité nommée, avec pour objectif, sinon de résoudre tous les problèmes se posant à l'encontre de l'appréhension de ces unités, au moins d'apporter quelques éléments de résolution. La discipline qu'est le TAL ayant ceci de particulier d'être à la frontière de plusieurs domaines scientifiques, nous avons, pour ce faire, choisi d'explorer deux dimensions impliquant les entités nommées : la linguistique d'une part et le TAL

[1] Le dernier numéro spécial de la revue (2007) avait en effet pour thème « Named Entities : Recognition, Classification and Use » ; l'appel à communication affirmait, entre autres, que « *The definition of what is a Named Entity (NE), however, still remains an overt question* » (voir à l'adresse suivante : http://www.benjamins.com/cgi-bin/welcome.cgi).

(qui comporte une part d'informatique) d'autre part. Ce faisant, nous avons tenté de recueillir divers éléments caractéristiques de ces unités, lesquels nous ont en définitive permis de formuler une proposition de définition des entités nommées. Ce travail « théorico-exploratoire » au regard des entités nommées ne se veut en aucun cas une résolution pleine et entière de la question du statut théorique des entités nommées mais, bien plus, un point de départ à une réflexion plus poussée sur un « objet TAL » certes devenu incontournable mais dont on ne connait au final pas grand chose.

Une autre voie de recherche consista à explorer les moyens d'obtenir une annotation plus poussée des entités nommées. Nous avons choisi de nous concentrer sur deux phénomènes particuliers, avec une méthode d'annotation fine d'une part et une méthode de résolution de métonymie d'autre part. Contribution plus « pratique » que la précédente, les travaux réalisés ne tentent pas moins de faire écho à la définition proposée auparavant.

## Organisation du mémoire

Cette étude comporte trois parties. La première offre une vue d'ensemble de la tâche de reconnaissance des entités nommées et les deux suivantes rendent compte de nos travaux.

### La problématique des Entités Nommées

Cette première partie est consacrée à un tour d'horizon de la problématique des entités nommées. Le chapitre 1 s'applique à présenter un « état des lieux » de la tâche de reconnaissance de ces unités, retraçant son historique (section 1.2.3), s'intéressant à ses applications (section 1.3) et examinant les systèmes existants ainsi que les performances obtenues (section 1.4). Le chapitre 2 propose quant à lui de rendre compte des difficultés pouvant advenir quant à l'annotation et l'appréhension des entités nommées, révélant de la sorte ce que l'on pourrait appeler l' « envers du décor » de la tâche de reconnaissance des entités nommées. Sont ainsi inventoriées des difficultés de catégorisation (section 2.1), des difficultés d'annotation (section 2.2) et des difficultés autour de phénomènes de polysémie (section 2.3). Au terme de cet exposé, nous dressons un bilan de la tâche (section 2.4), pointant plus particulièrement deux aspects nous paraisssant mériter attention : la question de la définition des entités nommées et celle des nouveaux traitements pouvant leur être appliqués.

## Les Entités Nommées : une création TAL

Cette deuxième partie s'intéresse à l'objet « entité nommée » d'un point de vue théorique, avec pour objectif de proposer une définition de ces unités linguistiques. Elle est divisée en trois chapitres.

Se posant la question de « *comment définir les entités nommées ?* » , le chapitre 3 est en fait un préambule méthodologique visant à déterminer la démarche à adopter pour la définition d'un objet TAL. Prenant acte des caractéristiques constitutives du Traitement Automatique des Langues, la méthode choisie propose tout d'abord un détour du côté de la linguistique avant de considérer les impératifs des traitements informatisés. Ces deux volets constituent naturellement les propos des chapitres suivants.

Le chapitre 4 explore donc quelques pistes linguistiques pouvant aider à une meilleure appréhension des entités nommées. Sont tout d'abord précisées, telles un cadre théorique, les notions de sens et de référence (section 4.1). Ensuite, conformément à ce qui a été dégagé dans le préambule méthodologique, la catégorie linguistique du nom propre est analysée (section 4.2), tout comme celle des descriptions définies (section 4.3). À ce stade de l'analyse, il est alors possible de dégager des caractéristiques linguistiques des entités nommées.

Ceci n'est cependant pas suffisant et il importe de replacer ces unités au sein du cadre dont elles sont issues. Le chapitre 5 examine ainsi ce que deviennent le sens et la référence en TAL (section 5.1), pour ensuite proposer une définition des entités nommées comportant dès lors une base linguistique et une dimension plus proprement TAL (section 5.2). Sont ensuites analysées les conséquences de cette définition (section 5.3), avec des points de méthodes et la question des polysémies.

## Les Entités Nommées : nouveaux traitements

Cette dernière partie présente deux expérimentations menées en vue d'un traitement plus fin des entités nommées. Elle ne constitue pas une parfaite illustration du statut théorique des entités nommées discuté lors de la partie précédente mais se veut plutôt une exploration de pistes « pratiques » interagissant avec un cadre théorique. Le chapitre 6 rend compte d'un travail mené avec Guillaume Jacquet autour de ce que nous avons appelé une « double annotation » des entités nommées, associant des étiquettes sémantiques fines aux catégories conceptuelles classiques. Enfin, le chapitre 7 donne à voir une méthode de résolution de métonymie des entités nommées réalisée dans le cadre de la campagne d'évaluation SemEval, en collaboration avec Guillaume Jacquet et Caroline Brun.

Au terme de ces investigations, nous revenons pour finir sur le chemin parcouru

et évoquons diverses améliorations et perspectives au regard de notre travail et du champ de recherche des entités nommées dans son ensemble.

# Première partie

# La problématique des Entités Nommées

# Introduction

L'analyse automatique de textes telle que la met en œuvre le TAL aujourd'hui vise à rendre compte de leur sens. Cet objectif n'est bien sûr pas nouveau, il a motivé de nombreux travaux pionniers du domaine, mais il revêt à l'heure actuelle une dimension plus tangible compte-tenu des nombreux acquis de la discipline. Ces dernières années ont en effet vu s'affirmer diverses techniques (modélisation symbolique et apprentissage statistique) au sein de nouvelles méthodes (linguistique de corpus) pour la résolution de tâches désormais bien définies et établies (analyse syntaxique, désambiguïsation lexicale, traitement de l'anaphore, analyse du discours, etc.). Le paysage de la recherche se précise peu à peu, les performances s'améliorent et si tout n'est pas résolu, loin s'en faut, il devient ainsi possible d'envisager des traitements relevant d'une dimension pleinement sémantique.

C'est dans cette perspective que s'inscrit la problématique des entités nommées. Accéder au sens d'un texte et en donner une représentation implique plusieurs opérations à différents niveaux. Quoi qu'il en soit, il importe avant tout de traiter dans le texte les éléments qui « portent son sens » et ces derniers correspondent en premier lieu à l'ensemble des unités lexicales qui le composent. Ces dernières ne sont cependant pas toutes égales eu égard à leur portée sémantique et certaines comportent une valeur informative plus importante que d'autres. Quelques-unes de ces unités ont été regroupées en un ensemble baptisé « entités nommées » et font l'objet de travaux, nombreux depuis la dernière décennie, cherchant à les reconnaître ainsi qu'à les catégoriser. La tâche de reconnaissance des entités nommées est aujourd'hui bien connue et les performances affichées par les systèmes de traitement de ces entités plus qu'honorables. Cette tâche, cependant, si satisfaisante et prometteuse qu'elle soit au regard de son exécution actuelle et de ses diverses applications, n'en comporte pas moins certaines difficultés et zones d'ombre : est-il si facile que cela d'annoter ce type d'unités ? Que recouvre exactement cette notion d'*entité nommée* ? Ces questions paraissent d'autant plus cruciales que les entités nommées font désormais l'objet d'un traitement plus poussé.

Cette première partie introductive sur la problématique des entités nommées comporte deux chapitres. Le premier s'attache à définir et à décrire la tâche de reconnaissance des entités nommées en un état des lieux se voulant le plus général possible. Après une rapide définition de la tâche, il sera ainsi question de son historique, de ses applications et des systèmes mis en œuvre pour sa réalisation. Par jeu de contraste, le second chapitre examine quant à lui les difficultés sous-jacentes de cette tâche, explorant des problèmes d'annotation et de polysémie.

# Chapitre 1

# La tâche de reconnaissance des entités nommées : état des lieux

## 1.1 Aperçu général

La tâche de reconnaissance d'entités nommées s'intéresse à un certain nombre d'unités lexicales particulières, que sont les noms de personnes, les noms d'organisation et les noms de lieux, ensemble auquel sont souvent ajoutés d'autres syntagmes comme les dates, les unités monétaires et les pourcentages[1]. Son objectif est double : il s'agit, d'une part, d'*identifier* ces unités dans un texte, et, d'autre part, de les *catégoriser* en fonction de types sémantiques prédéfinis. Le résultat de ces processus correspond à l'*annotation* des entités, laquelle se matérialise le plus souvent *via* des balises encadrant l'entité. C'est ainsi que, pour la phrase suivante,

> L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

il importe de reconnaître les entités *Lionel Jospin*, *RTL*, etc., puis de leur attribuer un type sémantique correspondant. L'identification et la catégorisation des entités L'annotation peut alors se présenter comme suit :

> L'ancien premier ministre socialiste <PERS>Lionel Jospin</PERS> a confirmé, <DATE>jeudi 28 septembre</DATE>, sur <ORG>RTL</ORG>, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de <DATE> 2007</DATE>.

[1]Cette énumération correspond peu ou prou à la « définition » traditionnelle des entités nommées ; cette acception sera de mise dans ces premiers paragraphes, avant une discussion plus approfondie de ce point.

La reconnaissance des entités nommées peut être mise en œuvre pour tout types de textes. Si historiquement ce processus a été appliqué sur des corpus journalistiques traitant de sujets géopolitiques (cf. section 1.2), il est aujourd'hui également appliqué sur d'autres types de corpus portant sur des domaines plus spécifiques. Ceux de la biologie et de la médecine sont par exemple fort demandeurs de ce genre d'analyse, la reconnaissance des noms de gènes, de protéines ou de maladies aidant au traitement de l'importante quantité d'information produite par ces communautés. Dans le domaine médical, il s'agit de reconnaître des expressions telles que *maladie de Parkinson* ou *fièvre de Lassa* [Bodenreider et Zweigenbaum, 2000] et, en biologie, des noms tels que *p53* ou *interleukin 1 (IL-1)-responsive kinase* [Fukuda *et al.*, 1998]. Si diverses expressions peuvent ainsi être annotées dans une tâche de traitement des entités nommées, les entités relevées le plus couramment appartiennent aux types sémantiques généraux de PERSONNE, ORGANISATION et LIEU.

La tâche de reconnaissance des entités nommées s'attache donc à extraire et à typer certains éléments informationnels d'un texte, en parfaite dépositaire de l'extraction d'information dont elle est issue. Il importe de retracer cette filiation et, pour ce faire, de considérer la longue tradition de campagnes d'évaluation, américaines et centrées sur ce domaine de l'extraction d'information aux origines, puis s'internationalisant et se diversifiant par la suite.

## 1.2 Historique

### 1.2.1 L'extraction d'information

C'est en effet à la faveur du développement de la tâche d'extraction d'information que la tâche de reconnaissance des entités nommées est apparue. La recherche pour la conception de systèmes d'analyse de textes a, depuis les débuts du TAL, exploré diverses voies. C'est dans ce cadre que l'extraction d'information a succédé aux systèmes génériques de compréhension de textes, aux visées sensiblement trop ambitieuses, comme le souligne T. Poibeau et A. Nazarenko [Poibeau et Nazarenko, 1999]. L'extraction d'information, ne cherchant plus à comprendre l'ensemble du texte, vise à extraire d'un texte donné des éléments pertinents d'information, dont la nature a été spécifiée préalablement. Il s'agit ainsi d'identifier des occurrences d'événements particuliers, d'en extraire les arguments impliqués pour ensuite en donner une représentation structurée. L'analyse s'effectue au niveau local et seule une partie du texte est considérée. Cette tâche peut alors se définir, selon la formule de T. Poibeau, comme « l'activité qui consiste à remplir automatiquement une banque de données à partir de

textes écrits en langue naturelle » [Poibeau, 2003, p.13].

Si le principe sous-jacent de l'extraction d'information n'était pas nouveau [Grishman, 1997], cette tâche a gagné en maturité et s'est singulièrement précisée grâce à la série des conférences MUC (*Message Understanding Conferences*[1]). Ce cycle de conférences, organisé par diverses institutions américaines et financé par la DARPA (*Defense Advanced Research Projects Agency*), s'est déroulé de 1987 à 1998, motivant de la sorte de nombreuses équipes de recherche pendant plus d'une décennie. Comme leur nom l'indique, l'objectif de ces conférences était à l'origine d'encourager la recherche autour de la compréhension automatique de messages militaires. Baptisées « conférences », ces dernières sont en réalité des campagnes d'évaluation, au cours desquelles un certain nombre de participants se voient remettre, dans un premier temps, un corpus d'entraînement et des instructions précises sur les informations à en extraire automatiquement, puis, dans un second temps, un corpus de test sur lequel ils doivent appliquer leurs systèmes. Les résultats sont ensuite évalués et présentés lors de la conférence finale, à laquelle seuls les participants à l'évaluation ont le droit d'assister. L'histoire de ces conférences est désormais bien connue ; [Grishman, 1997, Hirschman, 1998, Poibeau, 2003] permettent d'en apprécier l'évolution de façon détaillée. Nous en retraçons ici les grandes lignes afin de mieux situer l'apparition de la tâche qui nous occupe, la reconnaissance des entités nommées, et avant d'examiner d'autres événements ayant eux aussi contribué à l'émergence de cette dernière.

### 1.2.2 Les conférences MUC

#### 1.2.2.1 Les trois « cycles » de conférences

Il est possible de distinguer trois « cycles » au sein des 7 conférences qui se sont succédées, en fonction de la définition et de la difficulté de la tâche d'extraction à mettre en œuvre tout d'abord, de la taille et de la nature des corpus à analyser ensuite, et du degré d'aboutissement du processus d'évaluation enfin. Les deux premières conférences (1987 et 1989) forment un cycle liminaire que l'on peut qualifier d'« exploratoire » . Les corpus sont des messages de la *Navy* de style télégraphique et, après l'absence de toute instruction précise quant aux données à en extraire lors de la conférence de 1987, une premier formulaire simple de structuration de données (en anglais *template*) fait son apparition lors de la suivante en 1989. Sont également adoptées les premières mesures d'évaluation, précision et rappel, issues de la recherche d'information. Ces deux sessions pionnières, si elles n'ont révélé aucune méthode ou système particulier, ont le mérite

[1] http://www-nlpir.nist.gov/related_projects/muc/.

d'avoir rassemblé autour d'une même tâche plusieurs participants, ainsi amenés à discuter de leur travail et des moyens de l'évaluer.

Les conférences MUC-3, MUC-4 et MUC-5 constituent le second cycle, au cours duquel la tâche d'extraction d'information, telle que présentée ci-dessus et initiée par les précédentes conférences, s'est progressivement définie, gagnant en précision mais également en complexité. MUC-3 (1991) et 4 (1992) ont travaillé sur des corpus de nature journalistique, traitant d'événements ou d'actes terroristes en Amérique Centrale et du Sud. Les *templates* comportent alors de plus en plus de champs à remplir, ces derniers pouvant atteindre le nombre de 24. La figure 1.1[1] montre un exemple de formulaire à remplir pour MUC-3 : à partir d'une dépêche sur un acte terroriste, il importait d'en extraire le type d'incident, le lieu, la date, les exécutants, la cible ainsi que les effets sur cette dernière. Si les textes sont mieux écrits (moins de problèmes de casse, rédaction plus soignée et plus homogène), ils sont en revanche plus difficiles à analyser (plus longs, l'information à en extraire est plus difficile à identifier). Les collections de textes d'apprentissage sont distribuées en grand nombre et les premiers systèmes à base d'automates [Appelt *et al.*, 1993] ainsi que d'autres basés sur des méthodes statistiques font leur apparition. MUC-4 introduit également une nouvelle mesure d'évaluation, la F-mesure, qui combine les taux de précision et de rappel et rend ainsi plus facile les comparaisons entre systèmes. MUC-5 suit de près (un an) ces deux conférences et gagne encore en complexité : deux domaines sont proposés (technologique avec la microélectronique et commercial avec la vente d'entreprises) pour deux langues, anglais et japonais. Cette diversification correspond à une volonté d'améliorer la portabilité[2] des systèmes ; néanmoins, les temps de développement de ces derniers sont extrêmement longs (6 mois) et les niveaux de performance ne dépassent pas les précédents. Vue par certains comme un échec, cette dernière conférence de 1993 infléchit néanmoins de manière significative la vision de la tâche d'extraction d'information : devant traiter plusieurs domaines en plusieurs langues, les participants sont amenés à rendre plus génériques leurs architectures et certains modules d'analyse apparaissent comme nettement indépendants. MUC-5 marque ainsi un point d'aboutissement de ce deuxième cycle de conférence, révélant la nécessité de fragmenter en fonctionalités indépendantes une tâche d'extraction d'information devenue trop complexe.

Le dernier cycle est composé des conférences MUC-6 et MUC-7. La première, qui a lieu en 1995, marque un profond tournant pour ces campagnes d'évaluation : trois nouvelles tâches sont introduites et la tâche « traditionnelle » est simplifiée,

[1] Cette figure est issue de [Grishman, 1997].

[2] La portabilité désigne la capacité d'un système à être utilisé sans modification importante pour une autre application que celle pour laquelle il a été conçu.

19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unoffial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

| | |
|---|---|
| INCIDENT TYPE | bombing |
| DATE | March 19 |
| LOCATION | El Salvador : San Salvador (city) |
| PERPETRATOR | urban guerrilla commandos |
| PHYSICAL TARGET | power tower |
| HUMAN TARGET | - |
| EFFECT ON PHYSICAL TARGET | destroyed |
| EFFECT ON HUMAN TARGET | no injury or death |
| INSTRUMENT | bomb |

FIG. 1.1 – Exemple de formulaire d'extraction d'information pour des actes terroristes (MUC-3).

afin de répondre aux nouveaux objectifs fixés par le comité scientifique. Retenant les leçons du « cycle » précédent, MUC-6 se donne en effet pour programme de démontrer la possibilité de concevoir des systèmes indépendants pouvant facilement être utilisables, d'encourager à la réalisation de systèmes plus portables, et de favoriser des travaux pouvant contribuer à des modules de « compréhension profonde » [Grishman et Sundheim, 1995, Grishman et Sundheim, 1996]. Concernant ce dernier objectif, trois tâches sont envisagées : résolution de coréférence[1], désambiguïsation lexicale et détection de structure prédicat-argument. Baptisé « SemEval », ce pôle regroupant plusieurs tâches ne voit cependant se réaliser que celle de coréférence lors de la conférence. Au regard de la portabilité, un « mini-MUC » est adopté, avec la simplification du formulaire traditionnel d'extraction d'information reflétant les relations entre entités (*Scenario Template*), et un formulaire d'entité est proposé (*Template Element*), comportant 6 champs. Enfin, et surtout, la volonté de transformer certains modules impliqués dans le processus d'extraction d'information en véritables fonctionnalités indépendantes d'analyse de texte se traduit par la création de la tâche de reconnaissance des entités nommées (*Named Entities*). MUC-6 comporte donc au final les tâches suivantes :

- Entités Nommées
- Coréférence
- Formulaire des entités (*Template Element*)
- Formulaire des scénarios (*Scenario Template*, correspondant au » template« tra-

[1]Voir définition en 1.3.1.

ditionnel des MUCs)

Ce programme, certes ambitieux, reste cependant réalisable et correspond bien à l'esprit de la conférence de MUC-6 : en définissant des tâches séparées, le comité d'organisation cherche à encourager des participations pour telle ou telle tâche seulement, chacune d'elles se focalisant sur l'extraction d'un type d'information. Cette organisation de la tâche d'extraction d'information sous la forme de modules constitue la principale innovation de cette sixième conférence.

Le déroulement lui-même de cette conférence est en rupture avec les pratiques précédentes : les corpus, portant sur les changements de positions dans les entreprises, sont de taille inférieure et distribués pour une durée moindre[1] (temps d'apprentissage et temps de test), incitant ainsi les participants à concevoir des systèmes indépendants et nécessairement portables. Partant, cette conférence voit se multiplier les méthodes probabilistes et les systèmes à base d'apprentissage. Les niveaux de performances, pour chacune des tâches définies, sont encourageants, voire très encourageants : pour la tâche sur les entités nommées, des F-mesure supérieures à 0,9 sont atteintes par plusieurs systèmes (nous revenons plus précisément sur ces résultats un peu plus loin) ; le formulaire des entités (*Template Element*) est quant à lui rempli selon des taux de 75 % pour le rappel et de 86 % pour la précision et, enfin, le formulaire de scénario (ou mini-MUC) voit se réaliser des performances de l'ordre de 50 % pour le rappel et de 70 % pour la précision. Ces résultats prometteurs, au sortir de MUC-6, attestent donc de la nécessité (ou pour le moins de l'utilité) de décomposer la tâche d'extraction d'information pour, d'une part, obtenir des résultats probants et, d'autre part, faciliter la conception de systèmes génériques indépendants. Si d'importants progrès sont encore à réaliser pour permettre la portabilité des systèmes, le bilan de MUC-6 est en grande partie conforme aux objectifs fixés préalablement par le comité.

Organisée trois ans plus tard, MUC-7 ne poursuit pourtant pas le renouveau de la tâche d'extraction d'information avec autant de brio et d'enthousiasme que la précédente. Mise à part la généralisation des techniques d'apprentissage, peu de choses nouvelles sont à noter pour cette conférence de 1998, qui marque en réalité la fin des conférences MUC. Parallèlement à MUC-6 et MUC-7, une campagne d'évaluation multilingue en extraction d'information a été organisée : MET ou *Multilingual Entity Task*. Cette conférence a permis, entre autres, le développement de systèmes de reconnaissance d'entités nommées pour l'espagnol, le japonais et le chinois, expatriant avec succès cette tâche du monde anglophone vers d'autres langues.

[1] Se reporter aux figures 11 et 12 de [Hirschman, 1998].

Ce dernier cycle de la série des *Message Understanding*, s'il s'achève brutalement, n'en reste pas moins fondamental au regard de l'évolution de la tâche d'extraction d'information en général, et de la tâche de reconnaissance des entités nommées en particulier : celle-ci fait une apparition remarquée, et celle-là gagne en maturité. Après ce rapide survol des conférences MUC, considérons à présent la tâche sur les entités nommées d'un peu plus près.

#### 1.2.2.2 MUC et la reconnaissance d'entités nommées

C'est ainsi à l'occasion d'une refonte de la tâche d'extraction d'information (MUC-6) qu'apparaît la tâche de reconnaissance des entités nommées. Pourquoi les entités nommées ? L'intérêt pour ce type de noms s'explique par le fait qu'ils sont présents dans tous types de textes, quel que soit le domaine ; ils constituent ainsi un point de passage obligé pour tout système cherchant à rendre compte de l'information contenue dans un texte. En effet, qu'il s'agisse de messages militaires ou de dépêches journalistiques portant sur des actes terroristes, sur des fusions d'entreprises ou encore sur de la microélectronique, l'essentiel est de repérer les actants ainsi que les coordonnées des événements relatés. La tâche d'extraction et de reconnaissance d'entités nommées lors de MUC-6 s'est ainsi focalisée sur les trois types d'entités suivants :

**ENAMEX :** pour les noms d'entités correspondant à des noms de personnes, d'organisations et de lieux. Les sous-types sont : PERSON, ORGANISATION et LOCATION.

**TIMEX :** pour les expressions temporelles. Les sous-types sont : DATE et TIME.

**NUMEX :** pour les expressions numériques, de monnaie et de pourcentage. Les sous-types sont : MONEY et PERCENT.

Relatant la préparation de la sixième conférence, les auteurs de [Grishman et Sundheim, 1995] illustrent la tâche d'extraction d'entités nommées à l'aide du texte de la figure 1.2, où sont annotées des entités de type ENAMEX (personne et organisation) et NUMEX.

Instaurée et définie de la sorte lors de la conférence de 1995, cette tâche est également présente lors de MUC-7 et de MET. Pour chaque campagne d'évaluation, les performances atteintes sont remarquables. Quinze participants (pour 20 systèmes) s'attellent à reconnaître ces entités dans des dépêches portant sur des changements de position dans les entreprises lors de MUC-6. Sur la totalité des systèmes, plus de la moitié affichent des taux combinés de rappel et de précision supérieurs à 0,9, et le premier d'entre eux réalise une F-mesure de 0,96. Ces résultats inattendus sont d'autant plus impressionnants qu'ils ap-

Mr. <ENAMEX TYPE=« PERSON » > Dooner </ENAMEX> met with <ENAMEX TYPE=« PERSON » > Martin Puris </ENAMEX>, president and chief executive officer of <ENAMEX TYPE=« ORGANIZATION » > Ammirati & Puris </ENAMEX>, about <ENAMEX TYPE=« ORGANIZATION » > McCann </ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE=« MONEY » > $400 million </NUMEX>, but nothing has materialized.

FIG. 1.2 – Exemple d'annotation d'entités nommées (MUC-6).

prochent des performances humaines. Néanmoins, comme le font remarquer B. Sundheim [Sundheim, 1995] et R. Grishman [Grishman et Sundheim, 1996], ces performances sont à apprécier en tenant compte du fait que les corpus d'évaluation sont homogènes, de très bonne qualité rédactionnelle, et en nombre relativement restreint (30). La conférence MUC-7 voit se réaliser des performances du même ordre, quoique moins impressionnantes : le meilleur système n'obtient « que » 92 % pour le rappel et 95 % pour la précision. Ceci s'explique par l'ajout des expressions temporelles relatives dans la catégorie TIMEX, et par le fait que le corpus d'entraînement soit d'un autre domaine (accidents d'avions) que celui de test (lancements de satellites).

Force est donc de constater[1] que la tâche de reconnaissance des entités nommées fut un réel succès pour les conférences MUC. Menée à bien avec d'excellents résultats dès son lancement, cette tâche suscite dès lors un engouement certain, de la part et des participants, et des utilisateurs potentiels. En effet, à l'issue de MUC-6, deux systèmes sont commercialisés et d'autres intégrés dans des systèmes gouvernementaux d'analyse de textes, démontrant ainsi la possibilité de concevoir des systèmes génériques rapidement utilisables. En définitive, cette tâche répond positivement aux ambitions des dernières MUC, soucieuses, à juste titre, de définir des fonctionnalités indépendantes pour la tâche d'extraction d'information.

Au final, avec cette série de conférences américaines *Message Understanding*, ce sont dix années d'évolution de la tâche d'extraction d'information qui peuvent ainsi être appréciées. Coupant court aux systèmes de compréhension de textes, l'extraction d'information prend à la fin des années 1980 un parti plus modéré et ne s'intéresse, dans un texte, qu'à certains éléments précis et préalablement définis. Ainsi posée, cette tâche est-elle plus simple ? Plus facilement concevable certainement, mais plus simple, rien n'est moins sûr. La concentration de l'effort sur un type précis d'information n'empêche pas ce dernier d'être complexe et difficile à extraire. Preuve en est, si cette tâche gagne en définition au fur et à me-

[1] Ce constat est à la fois un jugement personnel et la reprise d'un jugement des organisateurs de la tâche (cf. R. Grishman).

sure des conférences, elle gagne également en complexité, les éléments à extraire étant de plus en plus nombreux. C'est ainsi que, compte tenu de cette évolution et de la nécessité de prendre en compte la réalité des applications, cette tâche a progressivement pris un caractère modulaire, se décomposant en plusieurs fonctionnalités autonomes. La tâche de reconnaissance des entités nommées apparaît alors, connaissant un succès immédiat, tant par ses résultats que par l'enthousiasme suscité auprès des participants nombreux. Aussi, avec la caractérisation progressive de la tâche, la réalisation de systèmes relativement performants et la promotion d'un véritable processus d'évaluation, ce cycle global de conférences constitue à coup sûr l'occasion d'un important progrès pour la tâche d'extraction d'information[1] et, au-delà, pour le traitement automatique du langage en général.

Au regard de la tâche de reconnaissance des entités nommées, cette évolution est d'autant plus manifeste que de nombreuses autres campagnes d'évaluation ont succédé à cette première série de conférences, consolidant et généralisant cette tâche nouvellement apparue.

### 1.2.3 Les autres conférences

Dans la droite ligne des conférences MUC, de nombreuses autres campagnes d'évaluation ont en effet poursuivi l'organisation de compétitions autour des entités nommées. Nous avons déjà évoqué les deux campagnes MET[2], organisées parallèlement à MUC-6 puis MUC-7, et dédiées à la reconnaissance des entités nommées dans d'autres langues que l'anglais [Merchant *et al.*, 1996]. Informelles et anonymes, ces deux campagnes ont motivé diverses expérimentations et favorisé la conception de systèmes indépendants de la langue, pouvant reconnaître de manière satisfaisante des entités nommées dans des corpus de nature journalistique en espagnol, chinois et japonais. Immédiatement après cette première généralisation de la tâche à d'autres langues, a été organisé au Japon le projet d'évaluation IREX[3]. L'objectif des organisateurs [Sekine et Isahara, 1999] était de réaliser quelque chose de similaire à MUC, en évaluant des systèmes d'extraction d'entités nommées sur des bases communes et en partageant données et expériences. Une quinzaine de systèmes ont participé avec des textes extraits de quotidiens nippons, obtenant pour certains des scores honorables avec des F-mesures supérieures à 0,8 [Sekine et Eriguchi, 2000]. En 2002 et 2003, CoNLL[4] a proposé une tâche de reconnaissance d'entités nommées, pour l'espagnol et le hollandais tout

[1] Même si les taux de précision et de rappel n'augmentent pas franchement entre MUC-3 et MUC-6, paradoxe relevé et expliqué par [Hirschman, 1998].

[2] Multilingual Entity Task.

[3] Information Retrieval and Extraction Exercise.

[4] Conference on Natural Language Learning.

d'abord, l'anglais et l'allemand ensuite, réunissant plus d'une dizaine de participants à chaque fois ([TjongKimSang, 2002] et [TjongKimSang et Meulder, 2003]). En France, la campagne ESTER[1], intégrée au projet EVALDA et organisée de 2002 à 2006, a proposé l'évaluation de systèmes de transcription d'émissions radiophoniques en langue française, ces transcriptions devant être enrichies par un ensemble d'informations annexes dont le marquage des entités nommées. Poursuivant encore la diffusion de cette tâche à d'autres langues, la campagne HAREM[2] proposa quant à elle une évaluation pour le portugais (portugais du Portugal, du Brésil, d'Afrique et d'Asie), réunissant près de 10 participants autour de corpus de natures diverses, issus de journaux, de courriers électroniques, de fictions, de rapports techniques ou encore de pages web [Santos *et al.*, 2006]. Enfin, il importe d'évoquer le programme ACE[3], mis en œuvre de 2000 à 2004 : prenant le relais des campagnes MUC pour l'anglais, ce programme de recherche entend poursuivre des travaux dans la même direction, à savoir détection des entités, des événements et des relations entre ces éléments, avec cependant un esprit différent, plus centré sur la mise au point de technologies que sur le traitement d'applications spécifiques[4]. Au regard de l'analyse de texte elle-même, ce programme envisage les choses différemment de MUC, faisant le choix d'une perspective plus « sémantique » que « linguistique » [Maynard *et al.*, 2005]. L'objectif dans ACE n'est effet plus d'extraire des entités nommées mais des *entités* tout court, nommées ou non : l'intérêt n'est plus pour la chaîne de caractères mais pour le *concept* de l'entité elle-même, détectable au travers de ses diverses *mentions*, ces dernières pouvant être des noms propres, des expressions nominales ou encore des pronoms. La chaîne référentielle d'une entité est donc explorée et annotée dans son intégralité ; néanmoins, il n'est pas question de reconnaître tous les types d'entités dans les textes et ces derniers sont limités aux désormais classiques PERSONNE, ORGANISATION et LIEU, légèrement remaniés toutefois, et auxquels sont ajoutés d'autres types. Présentant ainsi des modifications quant à la vision de la tâche avec une dimension plus sémantique — voire ontologique — mise en avant, le programme ACE met en œuvre un travail de recherche somme toute traditionnel autour des entités nommées, s'intéressant à l'extraction et au typage de certains types d'entités dans les textes, à l'instar des premières MUC et des autres conférences évoquées ci-avant.

L'évolution finale des conférences MUC, rappelons-le, fut celle de la valori-

[1] Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophonique.

[2] Avaliação de sistemas de Reconhecimento de Entidades Mencionadas

[3] Automatic Content Extraction.

[4] « The ACE program is a « technocentric » research effort, meaning that the emphasis is on developing core enabling technologies rather than solving the application needs that motivate the research. », [Doddington *et al.*, 2004].

sation des tâches indépendantes et des systèmes génériques. Atteignant au-delà de toute espérance des niveaux d'autonomie et de performance satisfaisants, la reconnaissance des entités nommées constitua la meilleure illustration de ce mouvement, d'ailleurs confirmé et poursuivi au travers d'autres conférences ou campagnes d'évaluation (un tableau récapitulatif de ces dernières figure en annexe A). Celles-ci se sont en effet succédées régulièrement de l'arrêt de MUC jusqu'à aujourd'hui, « démocratisant » et généralisant cette tâche à d'autres langues, à d'autres domaines et à des corpus de natures différentes. Si la tâche d'extraction d'entités nommées conserve son principe de base (repérage et typage d'éléments de types prédéfinis dans le texte), elle varie cependant quelque peu d'une conférence à l'autre, qu'il s'agisse des types d'entités à reconnaître, des occurrences à annoter ou des métriques d'évaluation utilisées. Ces particularités rendent difficile la comparaison des performances d'une campagne l'autre ; il n'empêche, cette tâche s'est de toute évidence affirmée durant cette dernière décennie, son utilisation dans de nombreuses et diverses applications marquant d'autant plus son succès.

## 1.3 Applications

Si la reconnaissance des entités nommées a connu un tel succès, offrant aux compétitions MUCs une indirecte postérité et donnant lieu à de nombreuses autres évaluations, c'est non seulement en raison de leur apparente facilité de traitement mais aussi et surtout en raison de l'important profit que peuvent en tirer de nombreuses applications. Il est possible de distinguer deux types d'applications, de natures différentes : la reconnaissance des entités nommées peut, d'une part, faire partie d'un composant TAL qui bénéficie alors de cette information (application que l'on pourrait qualifier d' « indirecte ») et, d'autre part, faire partie d'une chaîne de traitement avec une application directe particulière. Il importe de considérer successivement ces deux types d'applications.

### 1.3.1 La reconnaissance des entités nommées comme composant interne au TAL

#### 1.3.1.1 L'analyse syntaxique

La reconnaissance des entités nommées peut constituer un module fort bénéfique pour la réalisation d'étapes intermédiaires dans une chaîne de traitement TAL. L'analyse syntaxique peut ainsi bénéficier de ce type de module, comme le montrent [Brun et Hagège, 2004] et [Osenova et Kolkovska, 2002]. C. Brun et C.

Hagège [Brun et Hagège, 2004], examinant les tenants et aboutissants de l'intégration d'un module de reconnaissance d'entités nommées au sein d'un analyseur syntaxique robuste, identifient plusieurs niveaux du processus d'analyse pouvant mettre à profit ce module. En amont tout d'abord, les étapes de prétraitement que sont la segmentation et l'étiquetage morpho-syntaxique peuvent gagner en précision et en rapidité : il peut en effet être utile de savoir que la virgule et le point dans *HyOx, Inc.* ne constituent pas des séparateurs à proprement parler puisqu'ils sont parties intégrantes d'une entité de type organisation, et que *Seat* dans *Seat and Porsche had fewer registration in July 1996* est une organisation et ne peut par conséquent pas être un verbe. Ensuite, l'analyse syntaxique proprement dite peut également éviter des erreurs, notamment pour ce qui est du traitement de la coordination ; sur la base de types similaires, les entités *Egypt* et *Jordan* peuvent être coordonnées dans :

> He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to <LOC> Egypt</LOC> and <LOC>Jordan/LOC>.

En revanche, dans la phrase :

> He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to <LOC> Egypt</LOC> and <ORG>Likud party</ORG> politician.

il importe de coordonner *politician* avec *envoy*, et ce grâce à l'information apportée par les entités nommées, de deux types différents (LOC et ORG) cette fois-ci. Enfin, la construction de dépendances syntaxiques (ou relations grammaticales), peut « gagner en sémantique » grâce aux entités nommées, avec par exemple la construction d'une dépendance LOCALISATION entre *met* et *Baghdad* dans la phrase *They met in Baghdad*, et ce grâce à l'information géographique connue concernant l'entité *Baghdad*. Ainsi, c'est presque à toutes les étapes de traitement d'un analyseur syntaxique qu'un module de reconnaissance d'entités nommées peut être utile, venant soutenir des processus de base de l'analyse ou permettant d'enrichir cette dernière.

#### 1.3.1.2 La coréférence

Une autre application « interne » du TAL bénéficiant de la reconnaissance des entités nommées est la résolution de coréférence, ou le traitement des chaînes anaphoriques. Dans un texte, ou dans tout autre format de communication, les référents ou objets du monde dont il est question sont évoqués le plus souvent *via* diverses expressions référentielles. Ces dernières peuvent être des noms propres (*Lionel Jospin*), des descriptions définies complètes ou incomplètes (*l'ancien premier ministre socialiste*), ou encore de simples pronoms (*il*). Le fait de repérer

et de grouper des expressions référant à une même entité correspond à la résolution de coréférence et les systèmes automatiques réalisant une telle opération comportent très souvent un module de reconnaissance des entités nommées. En effet, il s'avère que la plupart des antécédents des pronoms personnels (80-85 % selon [Bontcheva *et al.*, 2004]) sont des noms de personnes, exprimés peu avant dans le texte. Par exemple, pour résoudre la coréférence du pronom neutre *it* en anglais, il peut être utile de discriminer parmi ses antécédents possibles ceux référant à une personne ou non, afin d'empêcher les rattachements malheureux. Considérons l'exemple suivant :

John bought a new computer. It is able to process XML.

Ici, le fait de savoir que *John* est une entité de type PERSONNE interdit le rattachement du pronom *it* à celle-ci, même si sa position sujet lui attribue l'importance d'un bon candidat. Ayant participé à la tâche de détection de relations entre entités définie par ACE, [GuoDong *et al.*, 2005] rendent compte de l'apport de diverses sources d'information pour cette tâche et montrent que celle du type d'entité nommée permet d'augmenter la F-mesure de leur système de 8.1 points. Ainsi, les entités nommées semblent constituer une source d'information non négligeable pour un système s'attachant à calculer des liens de coréférence.

#### 1.3.1.3 La désambiguïsation lexicale

Il convient par ailleurs d'examiner un autre processus TAL pour lequel les entités nommées sont d'une participation fructueuse : la désambiguïsation lexicale. Cette tâche, connue en anglais sous le nom de *Word Sense Disambiguation*, est primordiale pour tout système de traitement automatique des langues. Par désambiguïsation lexicale on désigne l'opération consistant à déterminer le sens d'un mot en contexte. En effet, un nombre important des mots de toute langue naturelle est susceptible de recevoir plusieurs interprétations ; ce phénomène, loin d'être un handicap pour les locuteurs d'une langue, qui savent bien au contraire se l'approprier en en explorant toutes les possibilités, constitue à l'égard des langues un gage de productivité et d'expressivité. La chose est cependant nettement plus compliquée pour un système automatique : comment un ordinateur peut-il sélectionner le sens à attribuer à un mot en contexte ? De nombreux travaux portent sur ce domaine et, si la tâche est loin d'être résolue, les pistes explorées sont multiples. C'est en tant qu'information sémantique que les entités nommées peuvent prendre part à un processus de désambiguïsation lexicale. Plus exactement, elles peuvent intervenir comme « filtre » au niveau des restrictions de sélection (ou sous-catégorisation sémantique) des sens d'une unité lexicale. En écho à la sous-catégorisation syntaxique, on appelle restriction de sélection le conditionnement

sémantique qu'une unité lexicale peut imposer à ses compléments, ou arguments pour les verbes, pour un sens donné. Pour donner, si l'on peut dire, du sens à tout cela, prenons les exemples suivants :

> Il est difficile de quitter Paris le vendredi soir.
> Certains se posent la question de quitter le Parti Socialiste.

Dans les phrases ci-dessus, le verbe *quitter* apparaît avec deux sens différents : dans la première il s'agit du sens de s'éloigner d'un lieu et dans la seconde de celui de se désengager. Pour les différencier, il peut être utile de prendre en compte le type sémantique des arguments convoqués par ce verbe, d'où l'intérêt de la reconnaissance des entités nommées : la présence d'une entité de type « lieu » en objet du verbe de la première phrase et d'une entité de type « organisation » en objet du verbe de la seconde peut en effet permettre de choisir le bon sens parmi les sens possibles pour le verbe *quitter*. Cette méthode d'exploitation des entités nommées pour discriminer le sens d'une unité lexicale relativement à ses restrictions de sélection fait d'ailleurs partie du cadre méthodologique de désambiguïsation lexicale déterminé par Mark Stevenson, lequel repose sur le principe de la combinaison des sources d'information, qu'elles soient d'ordre morphologique, syntaxique, sémantique, contextuel, etc. [Stevenson, 2003, Ehrmann, 2005].

#### 1.3.1.4 La traduction automatique

Enfin, il importe de dire un mot de la traduction automatique, tâche pour laquelle la reconnaissance des entités nommées constitue également une amorce importante. Un système de traduction automatique permet de traduire un document original en langue source en un document final en langue cible. Pour ce faire, diverses méthodes peuvent être utilisées, dont la plupart se concentrent à un moment ou à un autre sur les entités nommées. Ces dernières font l'objet de deux processus au cours d'une traduction : la translittération et la traduction. La translittération correspond à la mise en correspondance de signes d'entités nommées, sorte de traduction autonymique, entre *Londres* et *London* par exemple. La traduction des entités nommées correspond quand à elle à la mise en correspondance des signifiés, ou plutôt des référents des entités nommées. Si la translittération des entités nommées est un processus relativement facile et sans incidence sur la traduction globale du texte considéré, leur traduction peut en revanche conditionner la justesse du résultat final à l'échelle du texte. À titre d'illustration, considérons la traduction suivante, réalisée par le logiciel Systran[1] :

[1] http ://trans.voila.fr/voila

TEXTE SOURCE : Jack London was an american writer.
TEXTE CIBLE : Jack Londres était un auteur américain.

Ici, le patronyme *London* a été traduit avec le nom de la capitale *Londres*, ce qui n'est pas très adéquat. Cette mauvaise traduction aurait pu être évitée si la séquence *Jack London* avait été reconnue comme une entité de type personne. Un module de reconnaissance des entités nommées peut donc soutenir des systèmes de traduction automatique.

Analyse syntaxique, coréférence, désambiguïsation lexicale, traduction automatique, les entités nommées peuvent de fait venir soutenir divers composants internes au TAL. Leur utilité ne s'arrête cependant pas à cet aspect que nous avons qualifié d'« indirect », mais s'étend bien sûr au delà, avec de nombreuses applications directes.

### 1.3.2 Les applications directes

#### 1.3.2.1 L'extraction d'information et la veille

L'extraction d'information, application historique s'il en est, a déjà été décrite auparavant (cf. section 1.2.1), point n'est besoin donc de revenir longuement sur la question. Rappelons simplement que, pour remplir des bases de données avec une collection de formulaires comportant des informations ciblées relativement à un objet informationnel prédéfini, la reconnaissance des entités nommées est indispensable, ces dernières permettant de déterminer les actants de l'information à extraire. Une autre application, proche de l'extraction d'information en cela qu'elle cherche à donner une vision rapide et synthétique d'une collection de documents, est la veille documentaire. Selon l'Atalapédie[1], la veille correspond à l'activité qui vise à « surveiller, actualiser et prédire l'évolution des connaissances sur un domaine donné » . Pour ce faire, des analystes ont besoin de prendre connaissance de documents en grande quantité et en un temps relativement limité, ces documents portant le plus souvent sur des sujets économiques ou technologiques. Cette prise de connaissance demande bien sûr à être épaulée par des outils de fouille de texte ou d'extraction d'information, et c'est ici qu'interviennent de nouveau les entités nommées. Comme le développe T. Poibeau [Poibeau, 1999], celles-ci peuvent en effet constituer un véritable « enjeu pour les systèmes de veille », permettant de repérer rapidement les personnes ou entreprises dont il est question dans un do-

[1]Basée sur le principe du « wiki », l'Atalapédie est une encyclopédie en ligne offrant des informations sur le TAL, lancée par l'Association pour le Traitement Automatique des Langues. http ://www.atala.org/AtalaPedie/index.php ?title=Accueil.

cument, et donc de déterminer la pertinence de ce dernier au regard du type de veille mis en œuvre.

#### 1.3.2.2 La tâche de question-réponse

Au-delà de cette recherche documentaire basée pour l'essentiel sur des techniques d'indexation, d'autres modes d'accès à l'information sont apparus (dits de « troisième génération », cf. [Enjalbert et Bilhaut, 2005]), parmi lesquels les sytèmes de question-réponse. Les systèmes de question-réponse (tâche connue sous le nom de *Question-Answering* en anglais) s'inscrivent ainsi dans la continuité des systèmes de recherche d'information, à la différence que ceux-ci renvoient un ensemble de documents en réponse à une requête formulée à l'aide de mots-clés tandis que ceux-là renvoient une réponse précise ou un court extrait de document en réponse à une requête formulée à l'aide d'une question en langue naturelle. Cette tâche, mise en avant lors de la huitième édition de TREC[1] (*Text REtrieval Conference*) en 1999, s'est progressivement complexifiée et étendue à d'autres langues que l'anglais, avec notamment les campagnes NTCIR (*Evaluation of Information Access Technologies*) pour les langues asiatiques et CLEF (*Cross Language Evaluation Forum*) pour les langues européennes. À l'heure actuelle, les systèmes de question-réponse doivent pouvoir répondre à trois types de question[2] :

- des questions factuelles, telles que :
  When did Hawaii become a state ?
- des questions définitionnelles :
  What is Francis Scott Key famous for ?
- et des questions portant sur des listes :
  List musical compositions by Aaron Copland.

Sans entrer plus avant dans les détails de cette tâche, il convient de dire un mot de l'architecture classique des systèmes de question-réponse, afin de mieux évaluer le rôle des entités nommées. De tels systèmes procèdent traditionnellement en trois étapes[3] [Ayari, 2007] : analyse de la question, traitement des documents, puis extraction de la réponse. La première a pour mission d'identifier le *focus* de la question (ou l'élément important de la question) et de typer la réponse attendue (une personne, un lieu, etc.). Pour le premier exemple donné ci-dessus, le *focus* correspond à *Hawaï* et le type de la réponse attendue à une *date*. La deuxième étape a pour objet d'identifier, au sein d'une collection plus ou moins importante, des documents pertinents par rapport à la question puis, à l'intérieur

[1] Pour plus de détails sur ces conférences : http ://trec.nist.gov/

[2] Ces exemples sont repris de G. Marton, [Marton, 2003].

[3] Ou quatre, cela dépend de la complexité des traitements ; moteur de recherche et traitement des documents peuvent constituer deux modules séparés.

même de ces documents des passages comportant des éléments de réponse ; cette étape est le plus souvent réalisée à l'aide d'un moteur de recherche classique, auquel sont ajoutés des modules d'analyse sémantique. Enfin, la dernière étape est consacrée à l'analyse des extraits sélectionnés dans les textes afin d'en dégager la réponse souhaitée. Au cours de cette chaîne de traitement, les entités nommées interviennent à plusieurs reprises : leur reconnaissance est utile pour spécifier le type de la réponse attendue tout d'abord, pour repérer la réponse ensuite. [Ferret *et al.*, 2001] détaille cette utilisation à partir de l'exemple suivant :

QUESTION : How many people live in the Falklands ?
RÉPONSE : Falklands population of 2,100 is concentrated...

où il importe de caractériser la réponse attendue comme étant de type NUMBER, puis de repérer dans les textes des occurrences d'entités de ce même type. La même chose peut être réalisée pour notre premier exemple, où la réponse à chercher dans les documents est de type DATE. Bien sûr cette opération ne suffit pas à elle seule à découvrir la bonne réponse : il est également essentiel de travailler sur le focus et d'exploiter d'autres indices mais il n'empêche, la reconnaissance des entités nommées joue un rôle non négligeable au sein d'un système de question-réponse. À cet égard, l'étude du rôle et de l'importance des différents modules présents dans un système de *question-answering* par [Moldovan *et al.*, 2001], cités par [Marton, 2003], est probante, montrant qu'un système peut perdre plus des deux tiers de performance en l'absence de reconnaissance d'entités nommées (68 % plus exactement pour leur système). [Toral *et al.*, 2005] arrivent eux aussi à des conclusions similaires : faisant l'expérience de l'utilisation ou non d'un système de reconnaissance d'entités nommées en aval du moteur de recherche, ils démontrent que l'utilisation d'un tel module permet de réduire significativement le nombre de textes à considérer pour l'extraction de la réponse. Les systèmes de question-réponse constituent ainsi une des applications majeures de la reconnaissance des entités nommées.

#### 1.3.2.3 L'anonymisation

Autre application directe des entités nommées : l'anonymisation. Ce processus, défini avec précision par [Medlock, 2006], correspond à « l'identification et la neutralisation de références confidentielles dans un document ou un ensemble de documents »[1]. Cette tâche revêt toute son importance au regard de nombreuses activités où le partage de données textuelles comportant des informations confi-

[1] Traduction de [Medlock, 2006] : *« Anonymisation is the task of identifying and neutralising sensitive references within a given document or set of documents »* .

dentielles est indispensable. Au premier rang de ces dernières, figurent celles relevant des domaines juridique et médical. Le besoin d'anonymisation peut être motivé par la nécessité d'échanger ou de travailler réellement sur des données comportant des éléments confidentiels (cours de médecine pouvant s'appuyer sur des cas réels) ou par la nécessité d'appliquer des processus de TAL pour extraire des informations ou connaissances à partir d'une base textuelle (étude des décisions de justice sur une période donnée par exemple). Quoiqu'il en soit, certains éléments doivent être neutralisés et cela peut se faire soit par effacement, soit par remplacement par la catégorie à laquelle ils appartiennent, soit encore par pseudonymisation [Medlock, 2006]. Ces éléments confidentiels, on l'aura deviné, correspondent la plupart du temps à des noms de personnes, de lieux, des dates, etc., soit des entités nommées. C'est ainsi que D. Kokkinakis et A. Thurin exploitent un module de reconnaissance d'entités nommées dans leur systèmes d'anonymisation des lettres de décharge dans les hôpitaux [Kokkinakis et Thurin, 2007], tout comme L. Plamondon qui s'intéresse aux décisions de justice [Plamondon *et al.*, 2004]. Si A. Medlock souligne que la tâche d'anonymisation ne peut se permettre d'exploiter tels quels des modules de reconnaissance des entités nommées[1], ces derniers sont tout de même indispensables à un tel processus.

Enfin, pour clore ce tour (non exhaustif) des applications directes des systèmes de reconnaissance des entités nommées, il est possible d'évoquer, mais seulement d'évoquer, les moteurs de recherche cherchant à prendre davantage en compte la composante sémantique dans leur analyse de textes. Cette application est émergeante et ne correspond pas exactement à la reconnaissance des entités nommées telle qu'elle a été définie jusqu'à maintenant ; il en sera question de manière plus précise dans le reste de l'exposé (chapitre 6 notamment).

Aussi, qu'il intervienne en tant que composant interne au TAL ou bien directement, un module de reconnaissance d'entités nommées peut de toute évidence servir de nombreuses applications. De l'analyse syntaxique aux questions-réponses en passant par la traduction, l'extraction d'information ou encore l'anonymisation, le traitement des entités nommées trouve en effet sa place, plus ou moins importante, au sein des processus mis en œuvre. L'intérêt et le bénéfice de la reconnaissance d'entités nommées ayant ainsi été détaillés, il est temps d'en considérer la mise en œuvre et d'examiner les méthodes et systèmes permettant de reconnaître ces unités dans les textes.

[1]Les éléments à neutraliser peuvent ne pas être des entités nommées, cela dépend du domaine, et certains modules peuvent ne pas fonctionner correctement sur certains types de corpus comme par exemple ceux de courriers électroniques.

## 1.4 Systèmes et performances

Placées sur « le devant de la scène » TAL à l'occasion des conférences MUC, les entités nommées n'ont depuis cessé de motiver d'autres projets ou campagnes d'évaluation (cf. 1.2.3), favorisant de la sorte de nombreux travaux sur des systèmes permettant leur reconnaissance. Ces derniers ont donné lieu à de multiples publications et, un aperçu des plus importants étant effectué avec précision dans [Sekine et Eriguchi, 2000, Daille et Morin, 2000, Poibeau, 2001, Poibeau, 2003] et [Friburger, 2002], nous ne procéderons ici qu'à une description rapide de leur fonctionnement. Au delà d'un recensement, nous tenterons en effet de considérer certains principes de base de tout système de reconnaissance d'entités nommées, afin d'en permettre l'appréhension et la juste appréciation. Aussi, il sera question dans un premier temps des indices manifestant la présence d'entités nommées à l'écrit ainsi que des différentes méthodes possibles pour leur exploitation ; suivra dans un second temps un rapide descriptif de quelques systèmes et, enfin, un point sur les éléments et aspects essentiels de la reconnaissance d'entités nommées.

### 1.4.1 Reconnaissance d'entités nommées : indices et méthodes

Comment reconnaître des entités nommées dans des textes ? La question, préalable à toute conception de système de reconnaissance de ces unités, a déjà été largement explorée. Seront ici considérés les indices et moyens de reconnaissance d'entités nommées tout d'abord, leurs différentes méthodes d'exploitation ensuite.

#### 1.4.1.1 Comment identifier une entité nommée ?

D. McDonald distingue, dans un article abondamment cité depuis sa parution [McDonald, 1996], les désormais célèbres « preuve interne » (*internal evidence*) et « preuve externe » (*external evidence*). La première se rapporte à ce qui, à l'intérieur même d'une unité lexicale ou d'un syntagme, peut indiquer qu'il s'agit d'une entité nommée[1]. Ces indices, appelés « marqueurs » ou « mots déclencheurs » (*trigger words*), correspondent à des mots ou abréviations accompagnant régulièrement une entité nommée et permettant, dans la plupart des cas mais pas toujours, de la catégoriser. Le premier indice de cette sorte est bien sûr la majuscule, marque (typo)graphique que portent d'ordinaire les noms de personnes, de lieux ou d'organisation, soit la plupart des entités nommées. Cet indice est cependant à manipuler avec précaution, pour deux raisons principalement : d'une part, le premier mot d'une phrase comporte toujours une majuscule, par conséquent,

[1] « *Internal evidence is derived from within the sequence of words that comprise the name* », [McDonald, 1996].

même s'il s'agit d'un nom, ce n'est pas forcément une entité nommée et, d'autre part, cette marque de la majuscule n'est pas de règle dans toutes les langues (en allemand les noms communs prennent aussi une majuscule). La majuscule manifeste donc la présence potentielle d'une entité nommée mais ne permet pas de la catégoriser. Autres indices, cette fois-ci plus certains : les prénoms et les indicateurs générationnels pour les noms de personnes, des mots ou affixes de type classifiant pour les noms d'organisation et de lieux, ou encore des sigles ou des esperluettes, pour les noms d'organisation seulement. Ces marqueurs, tous catégorisant, sont surlignés dans les exemples suivants :

***Lionel*** *Jospin*
***L.*** *Jospin*
*Benoit* ***XIII***
*la* ***Banque*** *Populaire, Crédit Agricole* ***SA***
*Microsoft* ***Inc.***
*l'****avenue*** *des Champs Elysées*
*le* ***Mont*** *Granier*

Outre ces indices situés « à l'intérieur » des entités nommées, le contexte d'apparition de ces dernières peut également constituer un moyen de les reconnaître : il s'agit là de la preuve externe[1]. En effet, tout discours, ou autre format de communication, lorsqu'il réfère à une personne, un lieu ou une autre entité nommée le fait *via* son nom, lui adjoignant aussi le plus souvent des informations supplémentaires afin d'en indiquer les propriétés spécifiques. C'est ainsi qu'un nom de personne est souvent accompagné (surtout en première mention) d'un titre ou d'un grade, et un nom d'organisation d'un mot-clé de type classifiant :

***Monsieur*** *Jospin,* ***Mme*** *Denise*
***Général*** *Leclerc*
*l'****entraîneur*** *Aimé Jacquet*
*le* ***groupe*** *Sanofi-Aventis*
*the Coca-Cola* ***company***

Ces indices, internes et externes, constituent des éléments importants à considérer pour un système de reconnaissance d'entités nommées. Les premiers, exploités par la plupart des systèmes, peuvent toutefois entrer en conflit avec les seconds ; c'est la preuve externe qui l'emporte dans ce cas, le type d'une entité nommée étant contraint au final par son contexte d'apparition (conflit fréquent entre les types PERSONNE et ORGANISATION, les occurrences de ce dernier portant

[1] *« External evidence is the classificatory criteria provided by the context in which a name appears »*, [McDonald, 1996].

couramment soit le nom de leur fondateur, soit le nom d'une autre personne).

Preuve interne et preuve externe résument à peu près ce qui, dans la matière textuelle, peut aider un système de reconnaissance d'entités nommées. Ceci n'est cependant pas suffisant, et un autre moyen de compléter ces informations pour un système est le recours à des lexiques. Le terme de lexique en TAL renvoie à la notion (lexicographique) de recueil de mots et non à celle (linguistique) de l'ensemble des mots d'une langue. Un lexique, ou base de connaissance lexicale, a pour objet de décrire des mots dans leurs différents sens, leurs relations et leurs emplois et peut prendre différentes formes suivant l'organisation de cette description, dictionnaire, thésaurus ou terminologie (voir [Habert *et al.*, 1997] pour plus de détails). Pour ce qui est de la reconnaissance d'entités nommées, la notion de lexique renvoie à son interprétation la plus simple, à savoir une liste de mots auxquels sont associées des catégories sémantiques indiquant s'il s'agit d'une personne, d'un lieu ou autre. L'utilisation de lexiques a été initiée dès MUC-6 et, bien que controversée (cf. 1.4.3 ci-après), demeure très répandue à l'heure actuelle.

Information contextuelle et information lexicale constituent les deux sources traditionnelles de connaissances exploitées par les systèmes d'annotation d'entités nommées. La plupart des travaux se sont concentrés sur ces informations, à juste titre au demeurant, ces dernières s'avérant très fiables (cf. performances obtenues lors des campagnes d'évaluation). De nouvelles tentatives se font jour cependant pour trouver d'autres indices et il importe à cet égard de considérer les dernières expériences de H. Shinnou et S. Sekine. Cherchant à reconnaître des entités nommées rares pour lesquelles il est difficile d'obtenir les connaissances nécessaires à leur identification, [Shinnou et Sekine, 2004] se proposent d'exploiter la synchronicité d'apparition de mots dans des corpus dits « comparables ». Leur hypothèse est la suivante : étant donnés deux corpus de textes journalistiques alignés sur la base de leur date de parution (articles alignés au jour le jour donc), un mot pour lequel il est possible d'observer un « pic » de fréquence similaire dans les deux corpus a de fortes chances d'être une entité nommée. En effet, contrairement aux autres mots du lexique, une entité nommée conserve une forme identique d'une occurrence à l'autre, et donc d'un article à l'autre, ne pouvant que difficilement faire l'objet de paraphrase. Utilisant deux corpus journalistiques de 1995, *Los Angeles Times* et *Reuters*, les auteurs ont observé la distribution « temporelle » de deux mots : « killed » et « yigal », soit un verbe et le nom de l'assassin du premier ministre israëlien Yitzhak Rabin, décédé le 7 novembre 1995. Le premier apparaît de manière régulière dans de nombreux articles dans les deux corpus tout au long de l'année, tandis que le second voit sa distribution augmenter fortement au même moment, mouvement dont on peut déduire qu'il s'agit d'une entité nommée. H. Shinnou et S. Sekine ont mené à bien d'autres

expériences similaires, raffinant leurs calculs par la prise en compte de divers paramètres (variation du nombre d'articles publiés par jour, délai de publication différents pour un même événement, etc.) et validant leur hypothèse de départ. Partant, la date d'apparition d'une entité nommée peut constituer un bon indice, d'ordre temporel donc, pour son repérage. Cette démarche comporte néanmoins quelques biais : les entités nommées reconnues ne sont pas nombreuses, et surtout ne sont pas catégorisées. Les auteurs insistent à cet endroit sur la complémentarité de leur approche avec les moyens traditionnels de reconnaissance de ces unités. Ainsi, si les preuves internes et externes, tout comme les lexiques, constituent d'excellents indices et moyens de reconnaître des entités nommées dans un texte et sont abondamment utilisés, des travaux s'attèlent encore, avec succès, à la découvertes d'indices supplémentaires.

Les diverses sources de connaissances (contenues dans le texte à analyser ou ailleurs) nécessaires à prendre en compte par un système de reconnaissance d'entités nommées ayant été décrites, il convient d'examiner les différentes manières de les acquérir et de les exploiter.

#### 1.4.1.2 Comment annoter une entité nommée ?

Pour la plupart des processus de TAL, on distingue traditionnellement deux grandes approches : l'approche dite linguistique ou symbolique, et l'approche dite statistique ou à base d'apprentissage. Pour réaliser un système automatique d'analyse linguistique, que cette analyse soit d'ordre morphologique, syntaxique, sémantique, voire pragmatique, il importe de prendre en compte des informations, de les modéliser et de les manipuler *via* un formalisme adéquat quant à l'analyse escomptée. Ce qui distingue les approches citées ci-avant, ce n'est pas tant la nature des informations prises en compte que leur acquisition et leur manipulation.

La première repose sur l'intuition humaine, avec la construction manuelle des modèles d'analyse, sous la forme de règles contextuelles le plus souvent. Ces règles prennent la forme de patrons d'extraction, c'est-à-dire de descriptions d'enchaînements possibles de syntagmes nominaux ou verbaux, attendu qu'ils expriment l'information à repérer. Ces patrons exploitent généralement des informations d'ordre morpho-syntaxique, ainsi que celles contenues dans des ressources (lexiques ou dictionnaires). Pour ce qui est des entités nommées, ce type d'approche linguistique fut largement répandu, voire majoritaire durant les années 1990, au temps des premières conférences MUC. Un système de reconnaissance d'entités nommées basé sur une telle méthode comporte par exemple les règles suivantes (exprimées ici « verbalement ») : si un prénom connu (connaissance issue d'un lexique) précède un mot inconnu commençant par une majuscule, alors le syntagme peut être

étiqueté comme un nom de personne ; ou bien : si un mot inconnu est suivi du mot (ou de la forme) *Inc.*, alors il s'agit d'un nom d'organisation. Les choses ne sont bien sûr pas si simples, il faut savoir par exemple attribuer les bonnes frontières aux entités nommées, mais l'essentiel de cette approche est là.

L'autre type d'approche a pour principe de base la mise au point *automatique* de modèles d'analyse à partir de volumes importants de données. Ces méthodes sont dites statistiques ou à base d'apprentissage car elles *apprennent*, à partir de corpus annotés, des modèles d'analyse de textes, ces derniers pouvant prendre différentes formes, arbres de décision, ensembles de règles logiques, modèles probabilistes ou encore chaînes de Markov cachées. Au regard de la reconnaissance d'entités nommées, un système « observant » plusieurs fois la présence de l'abréviation *Mme* devant un mot annoté comme nom de personne dans le corpus d'apprentissage pourra facilement en déduire un modèle d'analyse. Ces systèmes à base d'apprentissage se sont considérablement multipliés ces dernières années.

Les avantages et inconvénients respectifs de ces deux types d'approches sont connus : les premiers reprochent aux seconds, entre autres, l'indispensable disponibilité de corpus annotés, et les seconds critiquent les premiers pour leur temps de développement ainsi que leur coût. Il est vrai qu'un travail de plusieurs mois d'un linguiste-informaticien est nécessaire pour l'écriture de règles, mais l'inverse est vrai également pour l'annotation de corpus qui peut être tout aussi longue, même si cela peut se faire par des gens moins experts. Hormis ces querelles de conception, l'intérêt se situe véritablement dans ce que chaque type de système est capable de faire et comment il peut fonctionner : si un concepteur de règles ne peut bien sûr pas penser à toutes les exceptions, il peut en revanche prévoir des patrons plus ou moins complexes pour le captage d'éléments difficiles, ce qu'un système probabiliste ne peut faire. La précision est ainsi d'ordinaire plus importante pour les systèmes symboliques tandis que les systèmes à base d'apprentissage présentent l'avantage d'être plus flexibles quant à leur adaptation à une tâche similaire mais portant sur un autre domaine et d'être plus robustes sur des corpus difficiles (ou bruités). Cette partition entre bienfaits et écueils de telle ou telle approche se reproduit bien sûr pour les systèmes de reconnaissance d'entités nommées. A. Borthwick , proposant une méthode de reconnaissance d'entités nommées à partir de calculs d'entropie maximale, ne manque pas de remarquer que, lors de la compétition MUC-7, le second système (IsoQuest, construit à partir de règles) est en de nombreux points similaire au dernier système symbolique (FACILE), à l'exception du temps de développement annoncé par les concepteurs respectifs [Borthwick, 1999]. L'auteur pointe là le coût de développement de tels systèmes ; il poursuit par ailleurs en soulignant leur difficile adaptation à d'autres domaines ou d'autres langues, les règles devant être, selon lui, totalement ré-

écrites. Inversement, [Poibeau, 2003] attire l'attention sur le volume de données annotées nécessaires pour entraîner un système, remarquant que celui de BBN [Miller *et al.*, 1998] exige un corpus annoté de 30 000 mots pour obtenir 0,81 de F-mesure, mais de 1,2 millions de mots pour atteindre 0,91. D'autres observations similaires l'amènent à la conclusion que le coût de l'écriture de règles n'est guère plus élevé que celui de l'annotation de corpus.

Enfin, au-delà de ces deux types d'approches et des désaccords de leurs partisans respectifs, il existe une troisième voie consistant à coupler l'approche symbolique et l'approche statistique en une approche alors qualifiée de *mixte* ou d'*hybride*. Cette dernière, rendue possible grâce à la maturité acquise par les deux autres, est sans doute la plus prometteuse.

Ayant ainsi présenté les divers types de processus utilisables pour la reconnaissance d'entités nommées[1], il est temps de les illustrer par l'analyse de quelques systèmes.

### 1.4.2 Un échantillon de systèmes

N. Friburger détaille abondamment un grand nombre de systèmes de reconnaissance d'entités nommées, qu'ils soient symboliques, à base d'apprentissage ou mixtes [Friburger, 2002]. Par conséquent, nous ne présenterons ici qu'un système par approche, renvoyant pour le reste au travail cité ci-avant [Sekine et Eriguchi, 2000, Daille et Morin, 2000].

#### 1.4.2.1 Un exemple de système symbolique : LaSIE

Développé à l'Université de Sheffield, le système LaSIE[2] (voir Gaizauskas *et al.*, [Gaizauskas *et al.*, 1995]) fut initialement conçu dans le cadre d'un projet de recherche autour de l'extraction d'information ou, plus généralement, autour de traitement du langage naturel. Ses concepteurs, participant à la compétition

[1]Cette présentation fait état d'une opposition entre des approches symboliques et des approches statistiques, opposition pouvant paraître quelque peu réductrice. En effet, si ce classement binaire permet une vision simple et rapide des types de traitements mis en œuvre en TAL, il n'est peut-être pas le plus approprié et met seulement en valeur l'utilisation ou non du quantitatif. Si l'on considère les choses d'un autre oeil, il est possible de dire que tout système de traitement automatique de données linguistiques nécessite des connaissances et qu'il doit, premièrement, les acquérir et, deuxièmement, les utiliser. Ce sont pour ces deux opérations d'acquisition et d'utilisation de connaissances que divers processus peuvent être mis en œuvre, d'orientation symbolique ou statistique. Acquérir, utiliser, symbolique, statistique, nous avons donc deux opérations et deux processus, ce qui fait au final quatre combinaisons possibles (dont seulement trois sont réellement mises en œuvre). Les différences observables entre ces différentes combinaisons (et qui permettent donc de choisir entre telle ou telle combinaison) ont trait à la granularité et l'échelle des traitements, le tout ayant des conséquences sur la fiabilité des systèmes.

[2]Large Scale Information Extraction.

MUC-6, firent le choix d'un système intégré, réunissant dans une même architecture plusieurs modules capables de répondre aux différentes tâches proposées à l'évaluation. C'est ainsi que, déclinées en plusieurs sous-traitements, trois étapes se succèdent : analyse lexicale, analyse puis interprétation sémantique et enfin interprétation du discours. L'entrée de ce processus est un texte (dont les paragraphes sont marqués en SGML[1]), la sortie une représentation sémantique de ce dernier, permettant d'offrir des réponses à chacune des tâches MUC (entités nommées, coréférence, « template element » et « scenario template »).

La phase de traitement lexical comprend les opérations suivantes : segmentation en mots (transformation du flux de caractères en suite d'unités ou *tokenization*), étiquetage des parties du discours (à l'aide de la méthode d'E. Brill [Brill, 1995]), analyse morphologique (ici lemmatisation) et étiquetage d'entités nommées sur la base de lexiques de noms propres et de lexiques spécialisés comprenant des mots déclencheurs. Durant cette phase, correspondant en quelque sorte à l'acquisition des connaissances nécessaires concernant les éléments de base d'un texte, le seul processus faisant intervenir un apprentissage automatique est l'étiquetage morpho-syntaxique (l'étiqueteur d'E. Brill est une méthode à base de règles avec apprentissage, voir [Paroubek et Rajman, 2000] pour plus de détails) ; l'étiquetage d'entités nommées est réalisé *via* une simple interrogation de lexique (ou *look-up*).

La deuxième étape s'attache pour sa part à la mise en œuvre de ces connaissances, opérant une analyse sémantique par le biais de l'utilisation de grammaires hors contexte : grammaire pour les entités nommées tout d'abord, pour l'analyse de phrases ensuite. Ces dernières sont écrites manuellement en Prolog et exploitent les informations recueillies lors de la phase initiale. La grammaire dédiée aux entités nommées comporte un peu plus de 200 règles au total, dont presque la moitié pour les noms d'organisations. Ci-après un exemple de règle de reconnaissance d'entité nommée utilisée dans LaSIE :

```
ORGAN_NP -> NAMES_NP '&' NAMES_NP
```

Cette règle (cf. [Gaizauskas *et al.*, 1995]) indique qu'un nom non encore classifié ou un nom propre ambigu (`NAMES_NP`, indication recueillie et attachée au token lors de la phase de traitement lexical), suivi d'une esperluette, puis de nouveau suivi d'un `NAMES_NP` est un nom d'organisation (`ORGAN_NP`). Des règles de facture similaire existent également pour l'analyse des phrases, à partir de laquelle est ensuite produite une représentation sémantique.

Enfin, la troisième phase de traitement se livre à une analyse du discours,

[1] *Standard Generalized Markup Language*. SGML est un langage générique de balisage des informations.

transformant la représentation sémantique générée par la phase d'analyse en une représentation d'instances, auxquelles sont associées des classes sémantiques (les auteurs parlent de « *ontological classes* ») ainsi que des propriétés. C'est à partir de cette représentation que la résolution de coréférence est réalisée. Interviennent également à cet endroit diverses heuristiques permettant de compléter l'annotation des entités nommées : si, par exemple, un nom d'organisation est reconnu en tant que tel puis, dans la suite du texte, est repérée une forme abrégée de ce même nom ou une variante, alors cette dernière se voit attribuer le même type que sa forme parente ou coréférente (propagation). Les connaissances utilisées lors de cette étape sont construites manuellement (objets et attributs de l'ontologie peu nombreux), et les mécanismes d'exploitation de ces dernières ne font nullement intervenir de données quantitatives (utilisation de mécanismes d'héritage). On le voit, cette dernière passe du traitement permet de « revenir en arrière » et d'enrichir des analyses précédentes. Cette possibilité correspond à une réelle visée des concepteurs du système, cherchant à exploiter des informations linguistiques de tous niveaux pour mener à bien leurs analyses.

Au final, LaSIE est un système intégré d'analyse de textes entièrement fondé sur une approche linguistique ; les résultats obtenus lors de MUC-6 pour la tâche de reconnaissance des entités nommées furent de 0,91 de F-mesure, et pour MUC-7 de 0,85 (performance réalisée par LaSIE-II, pour plus de précision, se reporter à [Humphreys *et al.*, 1998]). À la suite de cette description d'un système de reconnaissance d'entités nommées de type « symbolique », considérons à présent un exemple de système à base d'apprentissage.

#### 1.4.2.2 Un exemple de système à base d'apprentissage : MENE

Le système MENE[1] a été présenté pour la première fois à MUC-7 (1998) par l'Université de New-York. Conçu par [Borthwick *et al.*, 1998], MENE opère la reconnaissance d'entités nommées dans les textes en exploitant le principe de l'entropie maximale. À l'origine grandeur thermodynamique mesurant le degré de désordre de la matière, le principe de l'entropie a été adaptée à la théorie de l'information par C. Shannon. Ce dernier correspond alors à la mesure de la quantité d'incertitude liée à la distribution d'un événement aléatoire. Au regard du problème de la reconnaissance d'entités nommées, l'entropie maximale peut être vue comme le degré de certitude de l'information dérivable d'un corpus d'apprentissage relativement à une unité lexicale donnée, unité étant reconnue comme entité nommée et dont on cherche à établir le type. La mise en œuvre du calcul d'entropie maximale nécessite, comme tout système probabiliste, la détermina-

[1] Maximum Entropy Named Entity.

tion d'un certain nombre de traits (ou éléments d'information attachés à l'objet étudié) sur lesquels travailler. A. Borthwick *et al.*, décrivant le fonctionnement de leur système, font état de l'utilisation des traits suivants :

- traits binaires : traits dont la valeur est soit positive soit négative, ils correspondent grossièrement à ceux utilisés par le système Nimble de BBN [Bikel *et al.*, 1997] (capitalisation, présence de caractères numériques, etc.).
- traits lexicaux : traits issus du contexte lexical de l'unité considérée. Soit $U_n$ l'unité à catégoriser, on peut alors lui associer le trait $U_{n-1}$ correspondant par exemple à *Mrs.* ou le trait $U_{n+1}$ correspondant par exemple à *to* ; il est ensuite possible d'observer des régularités pour ces traits (le premier indique plutôt que l'unité est de type « personne », et le second qu'elle est de type « lieu »).
- traits « textuels » : traits véhiculant l'information de la structure du texte, *i.e.* localisation de l'unité dans le titre, le résumé, le corps du document, etc.
- traits issus de dictionnaires : traits signalant l'appartenance d'une unité à l'un de ces cinq états possibles, « start, continue, end, unique, other », sur la base d'une récupération des informations contenues dans des dictionnaires de noms propres. Ainsi, à partir de la connaissance de l'unité « British Airways » contenue dans le dictionnaire, si l'expression « on British Airways Flight 962 » est rencontrée, alors elle se verra attribuer les traits « other, start, end, other, other ». MENE exploite de la sorte des dictionnaires de noms de personnes, d'organisations, d'universités, de régions, ainsi que d'affixes de noms d'entreprises. Les auteurs insistent sur l'importance de ce type de traits pour leur système.
- traits externes : traits issus d'autres systèmes de reconnaissance d'entités nommées, soit l'indication de types sur les unités.

Une fois collecté l'ensemble de ces traits, MENE procède au calcul d'entropie maximale pour l'annotation d'entités dans de nouveaux textes[1]. Lors de la compétition MUC-7, ce système afficha 0,92 de F-mesure sur le corpus d'entraînement et 0,84 sur le corpus de test. Rappelons, à l'instar des concepteurs de ce système, que le changement de domaine entre le corpus d'entraînement (sur les accidents d'avion) et le corpus de test (sur les lancements de satellites), ne fut pas annoncé par les organisateurs de la compétition et prit par surprise la quasi totalité des systèmes à base d'apprentissage participant. Nous avons ici l'illustration de l'un des points faibles de ce type de systèmes, nécessitant de nouveaux corpus d'apprentissage pour pouvoir fonctionner sur de nouveaux domaines, *a contrario* des systèmes symboliques qui, si leur règles sont bien structurées et s'ils possèdent

[1]Sur le concept et le calcul d'entropie maximale, voir l'introduction de [Ratnaparkhi, 1997].

une bonne ergonomie (cf. [Poibeau, 2003]), sont facilement adaptables. Quoi qu'il en soit, les résultats obtenus par MENE sont fort bons. Leur niveau varie bien sûr en fonction de la quantité de données disponibles pour l'apprentissage : avec 20 documents, MENE propose 0,8 de F-mesure, mais cette dernière monte à 0,92 avec 425 documents. Toutefois, la meilleure amélioration du système provient de sa combinaison avec d'autres systèmes, de type symbolique, tels que IsoQuest [Krupka et Hausman, 1998], Manitoba [Lin, 1998] ou Proteus [Grishman, 1995] : utilisant les sorties fournies par ces trois systèmes, MENE affiche alors, sur le corpus d'entraînement de MUC-7, 0,97 de F-mesure. Ainsi, A. Borthwick *et al.* plaident à la fin de leur article pour l'utilisation combinée des deux approches, symbolique et à base d'apprentissage, l'une pouvant pallier les défauts ou insuffisances de l'autre. C'est ce postulat que D. Lin s'applique à mettre en œuvre.

#### 1.4.2.3 Un exemple de système mixte : travail de D. Lin.

Pour la compétition MUC-7, D. Lin ([Lin, 1998]) a présenté une « extension » du système symbolique réalisé par l'Université de Manitoba lors de MUC-6, extension consistant en l'addition au système initial d'un mécanisme d'enrichissement et de génération automatique de règles de reconnaissance d'entités nommées sur la base de données quantitatives. En un mot, le cœur du travail repose sur la construction d'une base de données de collocations. Le processus d'annotation des entités nommées opère en deux passes : la première construit, à partir d'un corpus annoté par le système initial et analysé syntaxiquement, une base de données de collocations de mots à partir de laquelle sont ensuite générées de nouvelles règles, et la seconde exploite le fruit de cette acquisition pour annoter, cette fois-ci définitivement, des entités nommées dans un texte.

La construction de la base de collocations consiste plus concrètement en la collection, dans un corpus donné, de l'ensemble des mots reliés deux à deux sur la base de leur relation grammaticale. Ces relations sont identifiés grâce à un parseur (D. Lin utilise MINIPAR), et sont stockées avec leur fréquences d'apparition. Une collocation (baptisée « dependency triple » par D. Lin) prend alors la forme suivante : (`word, relation, relative`), où le champ `word` correspond à un mot du corpus, celui de `relative` à un modifieur de ce mot, et celui de `relation` au type de la relation grammaticale reliant les instances des deux champs précédents, avec l'indication de leurs parties du discours. Chaque mot du corpus peut ainsi être stocké dans la base, en fonction de sa partie du discours, du mot auquel il est relié, et du type de relation opérant ce lien. Afin de se faire une idée plus précise du contenu de cette base, examinons quelques informations relatives à l'entrée `review` présentée par D. Lin :

`[review, V:comp1:N, acquisition = 3]` signifie que le verbe « to review » a eu trois fois le nom « acquisition » en relation objet dans le corpus analysé.

`[review, N:nn:N, admission = 2]` signifie que le nom « review » a eu deux fois le nom « admission » en relation modifieur de nom dans le corpus analysé.

Ce mot apparaît également dans d'autres collocations, au sein d'autres relations grammaticales ; la base de données contient au total 22 millions de mots. Au final, même si D. Lin ne fait pas référence à Z. Harris, il semble bien s'agir ici d'analyse distributionnelle.

Une fois construite, cette base de données de collocations est exploitée selon deux mécanismes. Le premier permet de générer automatiquement des règles d'annotation d'entités nommées et le second d'annoter des entités inconnues. Rappelons avant tout que le système initial utilisé est purement linguistique : il exploite des patrons à base d'automates à états finis, tirant profit de ressources lexicales tout comme des formes de surface du texte à traiter. D. Lin part du principe que le contexte d'apparition d'une entité nommée constitue un bon indice pour sa catégorisation (c'est la preuve externe de MacDonald). Ces contextes, accompagnés de leurs fréquences, sont contenus dans la base de données ; il suffit dès lors d'observer des régularités pour en déduire des schémas d'annotation. L'auteur explique par exemple que, sur 33 occurrences de noms propres apparaissant en tant que pré-modifieurs du syntagme « managing director », 26 sont classées comme relevant de la catégorie organisation. La déduction est relativement simple, et la règle permettant d'annoter les 7 entités restantes peut être produite. Ce mécanisme peut être mis en place pour 3600 contextes environ, pour lesquels la fréquence d'apparition d'un nom propre permet de déduire une règle. Cependant, il est des contextes où la déduction n'est pas si évidente et pour lesquels aucune classification sûre ne peut être proposée. Intervient alors le deuxième mécanisme, cherchant donc pour sa part à classifier des entités inconnues pour lesquelles le contexte d'apparition n'est pas suffisamment discriminant. Le système utilise à cet endroit un classificateur bayésien naïf : cet outil est basé, comme son nom l'indique, sur le théorème de Bayes et permet de calculer des probabilités conditionnelles. L'objectif est donc de prédire la catégorie d'une entité nommée à partir de ses caractéristiques. Ces dernières correspondent ici aux contextes d'apparition d'une entité donnée. Prenons l'exemple suivant (toujours issu de [Lin, 1998]) : étant donné le mot « Xichang », apparaissant à de nombreuses reprises dans le corpus mais inconnu du lexique, il suffit de relever tous ses contextes d'apparition ainsi que leurs fréquences, en autant de traits à partir desquels le classificateur bayésien peut prédire une catégorie. Par ce mécanisme, de nombreux mots inconnus peuvent être ajoutés au lexique. L'inconvénient de ce mécanisme est l'absence de contrôle : un mot peut être mal catégorisé et ajouté

au lexique, provoquant par la suite de nombreuses erreurs. Malgré cela, le système ainsi mis en œuvre par D. Lin reconnaît les entités nommées de MUC-7 avec un taux de F-mesure de 0,86, performance saluée lors de la conférence.

Nous venons de décrire trois systèmes, chacun d'eux participant d'une des approches possibles pour la reconnaissance d'entités nommées. Se reposant sur les mêmes types d'indices mais les exploitant de manières différentes, ces systèmes parviennent tous à passer les 0,85 de F-mesure lors des compétitions MUC. Au-delà du constat d'excellentes performances, est-ce à dire que toutes les approches se valent ? Que retenir de ces différentes méthodes de reconnaissance d'entités nommées ? Le paragraphe suivant tentera de donner des éléments de réponse à ces questions.

### 1.4.3 Points de controverse et lignes de force

Cela a déjà été dit plus haut et nous venons de l'illustrer en partie : la reconnaissance des entités nommées est une tâche qui, telle que définie et réalisée lors des conférences MUC-6 et MUC-7, recueille de très bonnes performances. En effet, les trois systèmes décrits ci-dessus franchissent tous les 0,85 de F-mesure, comme la plupart d'ailleurs des autres systèmes présentés à la compétition lors de MUC-6 et MUC-7 (cf. [Sundheim, 1995]). À mieux observer ces résultats cependant, une chose peut paraître déconcertante, à savoir le fait qu'aucune approche ou système ne se démarque résolument des autres. Bien plus, systèmes linguistiques, à base d'apprentissage ou bien encore mixtes se partagent régulièrement le podium, interdisant tout arbitrage absolu en faveur de telle ou telle méthode. Les organisateurs de la campagne IREX, homologue japonaise de MUC, S. Sekine et Y. Eriguchi parviennent aux mêmes conclusions : « *Il est intéressant de noter que les trois systèmes les plus performants appartiennent chacun à une catégorie différente. (...) Par conséquent il n'est pas possible de déterminer quel type de système est le meilleur.* »[1] [Sekine et Eriguchi, 2000]. Si l'ensemble des systèmes font usage des mêmes types d'indices et exploitent les mêmes sources de connaissance (preuve interne, preuve externe et lexique), les manières de les exploiter diffèrent cependant fortement. Au-delà de ces performances similaires, n'est-il pas possible d'examiner plus en détails l'apport de tel ou tel mécanisme et de dégager quelques « lignes de force » de la reconnaissance d'entités nommées ? Deux travaux analysant le fonctionnement de systèmes de reconnaissance de ces unités nous semblent

[1] Dans [Sekine et Eriguchi, 2000] : *« It is interesting to see that the top three systems came from each category ; the best system was a hand created pattern-based system, the second was an automatically created pattern-based system and the third system was a fully automatic system. So we believe we can not conclude which type is superior to the others ».* Cité également par T. Poibeau ([Poibeau, 2003]).

instructifs à cet égard. Le premier a suscité ce que nous appelons, peut-être le terme est-il trop fort, la « controverse des lexiques » : il s'agit de l'article « Named Entity Recognition without gazetteers » de A. Mikheev, M. Moens et C. Grover [Mikheev *et al.*, 1999] ; le second est l'« évaluation transparente » d'un système de reconnaissance d'entités nommées réalisée par T. Poibeau.

#### 1.4.3.1 La controverse des lexiques

L'utilisation de lexiques pour la reconnaissance d'entités nommées est fortement répandue, la plupart des systèmes ayant recours à un moment ou à un autre de leur processus à cette source de connaissance. Il s'agit en effet d'un mécanisme sûr, efficace et rapide, fonctionnant par une simple confrontation de la chaîne de surface à analyser avec une liste d'unités lexicales préalablement catégorisées. Si l'interrogation d'un lexique se fait le plus souvent conjointement à d'autres mécanismes, ce processus est de toute évidence facile à mettre en œuvre, séduisant en ce point les concepteurs d'applications commerciales (remarque de [Friburger, 2002]). Leur rôle ainsi que leur utilisation peuvent néanmoins être discutés, et ce sur plusieurs points. Leurs premiers détracteurs sont les partisans des méthodes à base d'apprentissage, incriminant leur coût de construction et leur opposant la possibilité d'acquérir automatiquement des listes de noms sur corpus annoté. Cette acquisition a cependant elle aussi un coût, celui de l'annotation de corpus. D'autres encore mettent en garde sur ce qu'implique, informatiquement parlant, l'exploitation d'une ressource de cette nature ; N. Wacholder *et al.* [Wacholder *et al.*, 1997] refusent ainsi de recourir à des lexiques si leur « chargement » est trop coûteux et qu'aucune méthode rapide d'interrogation n'est implémentable[1].

Pour pallier le premier écueil dénoncé, le coût de construction manuelle de lexique et le coût d'annotation de corpus, une solution peut être d'acquérir automatiquement de telles listes sans corpus annoté. Comment ? La méthode est décrite par M. Collins et Y. Singer dans [Collins et Singer, 1999]. Ces derniers proposent de se débarrasser du « fardeau » de l'annotation préalable en utilisant quelques « points d'amorce » ou *seed rules* à partir desquels le système peut ensuite acquérir d'autres amorces, et ainsi apprendre récursivement des listes de noms. M. Collins et Y. Singer, à partir de 7 amorces[2] et d'un corpus d'appren-

[1] Dans [Wacholder *et al.*, 1997] : « *A reliable database provides both accuracy and efficiency, if fast look-up methods are incorporated.* »

[2] Les *seed rules* du système de [Collins et Singer, 1999] sont les suivantes :
- *New York, California* et *U.S.* sont des noms de lieux.
- un nom contenant *Mr.* est un nom de personne.
- *I.B.M.* et *Microsoft* sont des noms d'organisations.

tissage de 90 000 exemples non annotés, parviennent ainsi à annoter des entités nommées avec 91 % de précision. Le facteur principal de réussite de cette méthode est la régularité et la redondance du corpus d'apprentissage. À l'instar de T. Poibeau, saluant ce travail, on peut néanmoins se demander quelles seraient les performances avec un corpus d'apprentissage moins homogène et surtout s'interroger sur les possibilités de correction des connaissances acquises (aspect interactif). Corpus annoté ou corpus homogène, la contrainte semble ainsi être du même ordre et, si la solution proposée par M. Collins et Y. Singer n'est bien sûr pas à négliger, il importe d'en prévoir l'utilisation en combinaison avec des ressources minimales. Au final, parmi ces différentes possibilités d'acquisition de lexiques, constitués à la main, acquis sur corpus annotés ou non annotés, l'essentiel étant de privilégier une construction peu coûteuse mais apportant une bonne précision ainsi qu'une couverture suffisante, la meilleure solution paraît être d'utiliser au départ une liste « manuelle » des noms les plus courants (de nombreuses listes sont disponibles sur Internet), puis de la compléter par un processus d'apprentissage sur corpus non annoté (en prenant soin de choisir ce dernier en fonction de l'application visée).

Examinons maintenant la deuxième critique énoncée ci-avant, à savoir le coût d'utilisation de tels lexiques. Implicitement, il est question ici de l'utilité de ces listes (mise en balance avec le temps de traitement) et de leur taille. L'article de A. Mikheev *et al.*, au titre somme toute percutant, propose une recherche en ce sens. Leur propos vise en effet à répondre à trois questions initiales : Au sein d'un processus de reconnaissance d'entités nommées, quelle est l'importance des lexiques ? Quelle doit-être leur taille ? et Quels sont les critères de construction de lexiques ? Afin de répondre à ces questions, les auteurs ont conduit diverses expériences, dont le principe consiste à faire varier la taille et la complétude des lexiques utilisés par un système d'annotation d'entités nommées. Quatre cas de figures ont été testés. Le premier fait usage des lexiques complets utilisés par les auteurs lors de MUC-7, soit près de 5 000 noms de pays, 30 000 d'organisations et 10 000 de personnes, lexiques constitués pour une partie automatiquement à partir des corpus d'entraînement, et pour l'autre à partir de listes disponibles sur internet. Ce mode d'utilisation de leur système permet une bonne annotation des entités nommées, chaque type considéré (ici personne, lieu et organisation) ayant une précision et un rappel supérieurs à 90 % (cf. résultats dans le tableau 1.1). La deuxième expérience fut de faire tourner le système sans lexique aucun : pour les noms de personnes, précision et rappel sont encore supérieurs à 90 %, aux alentours de 85 % pour les noms d'organisation et totalement catastrophiques pour les noms de lieux. Ces résultats indiquent que les lexiques ne sont pas d'une grande utilité pour la reconnaissance des noms de personne et d'organisation, ces

| | Full gazetteer | | Ltd gazetteer | | Some locations | | No gazetteers | |
|---|---|---|---|---|---|---|---|---|
| | recall | prec | recall | prec | recall | prec | recall | prec |
| organisation | 90 | 93 | 87 | 90 | 87 | 89 | 86 | 85 |
| person | 96 | 98 | 92 | 97 | 90 | 97 | 90 | 95 |
| location | 95 | 94 | 91 | 92 | 85 | 90 | 46 | 59 |

TAB. 1.1 – Résultats obtenus par A. Mikheev *et al.* pour leur système testé sur les corpus de MUC-7, avec des lexiques de tailles et de contenus différents.

derniers ayant par ailleurs de bonnes preuves internes et externes, mais qu'ils paraissent en revanche primordiaux pour les noms de lieux. Le troisième essai consista en l'utilisation d'un lexique de taille réduite (200 noms) et limité aux lieux : l'annotation des entités de ce type retrouve alors de bonnes performances. Enfin, la dernière expérience fut l'utilisation de lexiques « limités » (cf. colonne *Ltd* dans le tableau 1.1), construits à partir d'un tiers du corpus d'entraînement de MUC-7. Acquis à peu de frais, ces derniers permettent au système de retrouver des performances fort honorables pour chacun des types annotés.

Ces quatre expériences d'annotation d'entités nommées avec des lexiques plus ou moins gros et plus ou moins complets apportent quelques éléments de réponses aux questions liminaires d'A. Mikheev. À la question de l'importance des lexiques lors d'un processus de reconnaissance d'entités nommées, il convient de répondre que, si ces derniers ne sont pas primordiaux, leur rôle est tout de même non négligeable : l'absence totale de lexiques fait baisser de manière significative les performances. Une fois démontrée et admise la nécessité de lexiques, la deuxième interrogation porte à considérer la taille de ces derniers. Les expériences décrites ci-avant, faisant intervenir des lexiques limités (généraux ou de lieux seulement), plaident en la faveur de lexiques de taille minimale, ne comportant que les noms les plus courants. A. Mikheev *et al.* remarquent en effet que ces derniers, étant les plus connus des locuteurs, sont introduits « sans ménagement » dans le discours, tandis que les noms les moins connus ont de plus grandes chances d'être accompagnés de preuves internes et/ou externes[1]. T. Poibeau , menant à bien une expérience similaire de variation de taille de lexiques pour un système travaillant sur les corpus MUC-6 [Poibeau, 2003], voit décroître les résultats à mesure de la réduction de la taille des listes, mais constate également que la courbe des performances atteint un certain plafond autour d'un seuil d'environ 25 000 mots par lexique. Ceci confirme le fait que, s'il n'existe pas bien sûr de taille « idéale », pour les lexiques en reconnaissance d'entité nommées, rien ne sert d'utiliser 200 000 ou 100 mots, le raisonnable semblant se situer entre 15 000 et 30 000 mots. Enfin,

[1]Dans [Mikheev *et al.*, 1999] : « *When collecting these gazetteers, one can concentrate on the obvious examples of location and organisation, since these are exactly the ones that will be introduced in text without much helpful context* » (nous soulignons).

sur quels critères se baser pour constituer des lexiques ? Le propos précédent y a déjà répondu en partie : il semble judicieux de s'intéresser aux mots les plus communs car les moins spécifiés en discours. Par ailleurs, le type d'entité nommée paraît être digne de considération, les noms de lieux étant de toute évidence plus sensibles à la présence ou non de lexiques. En dernier lieu, le domaine d'application constitue sans conteste un critère de constitution de lexiques, un corpus de nature journalistique ne mobilisant pas les mêmes unités lexicales qu'un corpus scientifique.

Qu'il soit question de leur acquisition ou bien de leur exploitation, les lexiques n'ont de toute évidence pas manqué de susciter de nombreux débats. Il s'agit d'une composante essentielle de tout système de reconnaissance d'entités nommées et cela est donc tout naturel. De tout cela il se dégage que, manifestement indispensables, ces lexiques sont tout de même à constituer avec précaution, afin que leur utilisation soit judicieuse et pleinement complémentaire des autres composants ou mécanismes impliqués dans la reconnaissance des entités nommées. Ces derniers méritent également attention, T. Poibeau en a proposé une étude qu'il convient dès lors d'examiner.

#### 1.4.3.2 L'évaluation transparente de T. Poibeau

« Deconstructing Harry » [Poibeau, 2001], l'article, entreprend « *d'évaluer sur une base objective les différentes composantes d'un système de reconnaissance d'entités nommées* ». Menant à bien une sorte d'évaluation transparente (évaluation de type *glass box* et non plus de type *black box*), T. Poibeau propose ainsi un protocole de déconstruction d'un système générique de reconnaissance d'entités nommées afin d'en évaluer les différentes composantes ; il entreprend, au cours de ce protocole, d'examiner quatre points ou éléments principaux que sont : les lexiques, les grammaires, les techniques d'apprentissage et les mécanismes de révision. Les lexiques ayant été abondamment discutés, nous ne considérerons ici que les trois derniers éléments. Il convient avant tout de signaler que les expériences réalisées dans [Poibeau, 2001] l'ont été sur trois types de corpus : le corpus MUC-6 (dépêches du *Wall Street Journal*), un corpus Reuters (dépêches financières) et un corpus de courriers électroniques. Cet assortiment de corpus de natures différentes vise à nuancer l'appréciation des composantes évaluées en fonction de la nature des textes traités.

Une grammaire comporte des règles qui, sur la base de connaissances lexicales issues d'une analyse préalable ou de lexiques, permettent de prédire le type d'une entité nommée à partir de l'enchaînement des éléments lexicaux qui la composent et/ou l'encadrent. Sans revenir sur le fonctionnement précis de cette composante,

il importe de s'interroger sur son rôle et son importance au sein d'un système. T. Poibeau est parvenu à décomposer une grammaire et à chiffrer le taux d'utilisation des diverses règles. Il ressort de cette expérience que l'activation des règles suit la loi de Pareto, dite des 80/20 : un petit ensemble de règles permet d'annoter une grande partie des entités, tandis que de nombreuses règles supplémentaires sont nécessaires pour couvrir les entités restantes. Si l'on observe ce qu'il se passe d'un type de corpus à un autre, il apparaît que moins le corpus est régulier[1], plus l'activation des différentes règles est d'un ordre comparable.

Autre point évalué par T. Poibeau : les mécanismes d'inférence, de généralisation et de révision. Ces mécanismes, implantés dans la plupart des systèmes, offrent la possibilité d'annoter des occurrences d'entités inconnues, ou pour lesquelles le contexte n'est pas suffisamment discriminant, en se servant de leur éventuelle annotation déjà effectuée précédemment dans le texte. Un système peut par exemple rencontrer l'entité suivante, *Kosciusko-Morizet*, et ne pas pouvoir l'annoter faute d'indices et d'information suffisants ; un mécanisme d'induction et de généralisation peut toutefois l'aider à le faire, en lui rappelant son annotation précédente de *Mme Kosciusko-Morizet* en tant que personne. Cette annotation dynamique peut encore se transformer en véritable connaissance, avec le stockage de cette entité dans un lexique intermédiaire puis, après validation humaine, dans un lexique proprement dit. Cette induction-généralisation ne présente, d'après les résultats des expériences, que peu d'intérêt pour les corpus réguliers et homogènes. Les performances d'annotation sur le corpus MUC-6 ne varient en effet qu'imperceptiblement avec ou sans ce mécanisme. Celui-ci est en revanche beaucoup plus productif pour les deux autres corpus : les performances augmentent presque de 20 % pour l'annotation des courriers électroniques, montrant ainsi que plus la qualité rédactionnelle diminue, plus il importe d'identifier des mots inconnus sur la base d'annotations précédentes. Le mécanisme de révision consiste quant à lui à « corriger » une annotation en fonction d'indices contraires. L'exemple le plus original en est bien sûr l'entité *Washington* qui, annotée le plus souvent par défaut comme un nom de lieu, peut se voir attribuer le type personne si rencontrée dans le contexte *Mrs. Washington*. Il est alors possible de réviser en conséquence les autres occurrences de cette entité dans le reste du document. Relativement aux performances, ce mécanisme ne permet qu'un gain minime.

Enfin, le dernier élément proposé à l'évaluation correspond aux techniques d'apprentissage. En effet, pour se dégager quelque peu des discussions apparem-

[1]Benoît Habert définit la notion de corpus de la manière suivante : « *un corpus est une sélection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminées d'une langue* » [Habert, 2000]. La régularité à laquelle nous faisons référence ici concerne plus particulièrement l'aspect linguistique des corpus, c'est-à-dire le niveau de langue et la manière dont sont introduites en discours les entités nommées.

ment sans fin autour des avantages et inconvénients des méthodes symboliques *vs.* les méthodes à base d'apprentissage, le plus profitable est de déterminer l'importance de chacune des approches en situation réelle. Pour ce faire, T. Poibeau a imaginé une mesure de prédiction des gains possibles dûs aux mécanismes d'apprentissage. La mesure prend en compte les éléments suivants :

1. Proportion d'entités nommées déjà répertoriée dans le dictionnaire relativement à la taille du corpus.
2. Proportion de séquences constituées d'une amorce suivie d'un mot inconnu relativement à la taille du corpus.
3. Proportion de mots inconnus commençant par une majuscule relativement à la taille du corpus.

Nous ne reproduisons pas ici la formule de calcul utilisée par l'auteur. Cette dernière permet néanmoins de chiffrer l'apport possible des techniques d'apprentissage pour les trois corpus, ce dernier étant bien sûr le plus élévé pour les courriers électroniques. En effet, moins un corpus est régulier et homogène, moins son traitement est assuré par les grammaires et les lexiques.

De ce panel d'évaluations, il ressort qu'aucune méthode et qu'aucun mécanisme n'est prépondérant ou préférable aux autres dans un système de reconnaissance d'entités nommées. Lexiques, grammaires, inférences, apprentissage, tout est efficace et nécessaire, en un panachage prenant en compte les spécificités du corpus avant tout.

Ainsi, si la reconnaissance d'entités nommées peut s'appuyer sur de nombreux indices, elle peut également être mise en œuvre selon différentes méthodes. Les multiples combinaisons possibles de divers composants et mécanismes ont conduit à la réalisation de nombreux systèmes et alimenté moult discussions. Au-delà de ces (presque) polémiques, le plus probant s'avère de combiner les différentes approches, en tenant compte des particularités du contexte de réalisation du système. Arrivés au terme de la description des systèmes de reconnaissances d'entités nommées, dernier point de ce chapitre, il est temps de rassembler l'ensemble du propos de celui-ci en un bilan récapitulatif.

## 1.5 Bilan

De sa définition à la description des différents types de systèmes s'attelant à sa réalisation en passant par ses applications, ce chapitre a ainsi présenté la tâche de reconnaissance d'entités nommées, en un état des lieux se voulant le plus complet possible. Le propos, retraçant les évolutions de cette tâche depuis son apparition jusqu'à ses performances actuelles, a ainsi permis de dégager les points suivants.

La tâche de reconnaissance des entités nommées, initialement définie pour améliorer la modularité et la portabilité des systèmes d'extraction d'information, a fait une apparition remarquée, voire inespérée, au milieu des années 1990 dans la communauté TAL. Remarquée, cela est certain : en un espace de temps assez bref (une décennie à peine) à l'échelle de l'histoire de la discipline, les entités nommées ont su occuper le devant de la scène des compétitions MUC et consœurs, suscitant un intérêt toujours plus soutenu. Inespérée, il faut en convenir également : qu'il s'agisse du renoncement momentané à la compréhension globale de textes ou des difficultés des premières campagnes d'évaluation autour de l'extraction d'information, la tâche de reconnaissance des entités nommées a su, par ses performances pour le moins remarquables, éclairer quelque peu l'horizon des recherches en TAL et redonner confiance et enthousiasme, si tant est que ces derniers eussent alors fléchi, à de nombreuses équipes de recherche. C'est d'ailleurs ce qu'affirme en substance B. Sundheim lorsqu'elle conclut son rapport sur MUC-6 : « *The introduction of two new tasks into the MUC evaluations and the restructuring of information extraction into two separate tasks have infused new life into the evaluations.* ». Résultats prometteurs et engouement furent ainsi les ingrédients d'un réel succès pour cette tâche, et ce dès son introduction.

Ce succès fut d'autant plus avéré qu'après s'être ainsi progressivement affirmée, la tâche de reconnaissance des entités nommées se révéla être de toutes les applications : en tant que module à l'intérieur d'un système d'analyse, offrant son concours tant pour l'analyse syntaxique que pour la résolution de coréférence ou encore la désambiguïsation lexicale, ou tout aussi bien en tant qu'application visant une tâche donnée, à l'exemple de l'extraction d'information naturellement, de la tâche de question-réponse ou de l'anonymisation. Ce faisant, de nombreux systèmes furent réfléchis pour sa mise en œuvre, conduisant à l'apparition de modules variés, purement linguistiques, à base d'apprentissage ou mixtes. Cette recherche ne fut pas, nous l'avons vu, sans de nombreux débats autour de l'importance de telle ou telle composante. Profitant de cet irremplaçable moteur de recherche (au sens propre) que sont les campagnes d'évaluation faisant se réunir autour d'une même tâche et pour un temps limité différentes équipes alors assignées à un but précis, les entités nommées ont ainsi vu se diversifier leurs moyens de reconnaissance, ceux-ci gagnant en maturité au fur et à mesure des années et des controverses, s'avérant au final d'autant plus efficaces que mêlant plusieurs approches.

Performances remarquables, applications nombreuses, exploration de diverses méthodes, force est de constater que, depuis le vide, ou l'absence de recherche, dénoncé en 1992 par S. Coates-Stephens [Coates-Stephens, 1992] à l'endroit de ce type d'unité, un véritable acquis s'est constitué, faisant désormais apparaître

cette tâche de reconnaissance des entités nommées comme un incontournable du TAL.

Ainsi, tâche « performante » et utile s'il en est, la reconnaissance d'entités nommées, telle que définie par les conférences MUC, ne semble accuser aucun bémol. Cependant, que sait-on au juste de cet ensemble d'unités ? Sait-on les définir ? Est-il si « facile » de savoir ce qu'il y a à reconnaître ? À y regarder de plus près, et c'est l'objet du chapitre suivant, nous verrons que la situation n'est pas aussi limpide que ce qui est, à l'instar de l'aperçu présenté ci-avant, communément admis et qu'il convient de considérer d'un œil attentif ce que recouvre précisément cette tâche avant d'en imaginer de nouvelles perspectives de recherches.

# Chapitre 2

# Difficultés

Au terme du chapitre précédent, il se dégage tout naturellement un sentiment, sinon d'achèvement, au moins de maturité, au regard de la tâche de reconnaissance des entités nommées. Il est en effet possible de considérer que cette dernière, telle que définie originairement par les conférences MUC, a été menée à bien, offrant satisfaction à nombre d'acteurs du domaine, qu'il s'agisse de comités d'organisation de conférences (donc d'instances gouvernementales), d'équipes de recherche ou encore d'industriels. Ainsi établi, ce succès, si justifié soit-il, ne doit toutefois pas masquer l'existence d'un certain nombre de difficultés auxquelles tout un chacun est confronté lors de la conception et réalisation d'un projet de reconnaissance d'entités nommées. En d'autres termes, il importe de ne pas méconnaître quelques-unes des pierres d'achoppement de cette tâche, et ceci constitue l'objet du présent chapitre.

Au-delà de sa mise en œuvre *via* des approches et systèmes différents (mais complémentaires), la tâche de reconnaissance des entités nommées soulève de prime abord deux questions essentielles, à savoir quelles entités annoter et comment. Si l'on cherche à comprendre de quelle manière cette tâche a été conçue et définie lors des conférences durant la décennie de recherche exposée ci-avant, on rencontre un ensemble de *guidelines* ou directives d'annotation, apportant précisément des éléments de réponses à nos questions. C'est donc tout naturellement que nous nous appuierons sur ces directives pour tenter d'éclaircir le problème de la détermination des catégories à prendre en compte d'une part, et le problème de l'annotation des entités nommées telles qu'apparaissant dans le corps du texte d'autre part. À la considération de ces deux difficultés, il importera d'ajouter les éventuelles complications dues à des phénomènes de sens ; en effet, la langue réserve d'ordinaire bien des surprises et il n'est pas étonnant que celles-ci soient également de mise à l'encontre des unités faisant l'objet de cette étude.

À bien y regarder, les difficultés que nous nous proposons d'examiner ici par-

ticipent de trois niveaux différents que sont le monde, le texte et la langue. La détermination des catégories d'entités nommées à prendre en compte relève d'une considération de la réalité du monde ; les résolutions quant à la manière d'annoter les entités nommées relèvent d'une considération de la matérialité du texte ; enfin, les problèmes de « polysémies » autour de ces unités relèvent des potentialités de la langue. Ces trois points ainsi énoncés correspondent au déroulement de ce chapitre, qui se terminera par la mise en relation de ces types de difficultés via un questionnement sur la définition, au sens propre du terme, des entités nommées.

## 2.1 Les entités nommées dans le monde : le problème de la détermination des catégories

Quiconque se met dans la position de réaliser un système de reconnaissance d'entités nommées dans les textes, ou même d'organiser une évaluation autour de cette tâche, se voit dans l'obligation première de s'interroger sur ce qu'il cherche à reconnaître précisément, et donc de déterminer un ensemble de catégories. Est-on intéressé par les noms de personnes, les noms de bateaux, de continents, d'acteurs ? Beaucoup d'éléments peuvent être pris en compte, et cette brève énumération laisse déjà deviner l'étendue des possibles quant à la détermination de catégories.

### 2.1.1 De quoi est-il question ?

Précisons tout d'abord ce que l'on entend ici par *catégorie* et selon quel niveau ce terme est utilisé. En effet, différents types de catégorisation peuvent avoir lieu à l'endroit des unités composant l'ensemble de nos entités nommées : pour ne citer que ces deux travaux, B. Daille [Daille et Morin, 2000] parle de « catégorisation linguistique » et de « catégorisation pragmatique » , tandis que N. Friburger [Friburger, 2002] évoque pour sa part des « typologies morpho-syntaxiques » et des « typologies sémantiques ». Selon sa définition la plus simple, le terme de *catégorie* renvoie à « *une classe dans laquelle on range des objets de même nature* »[1], soit « *un ensemble de choses qui ont un certain nombre de caractères en commun* »[2]. Les caractères en commun des objets ici étudiés sont de nature sémantique, il n'est pas question de particularités morpho-syntaxiques ou autre. Aussi, la détermination de catégories quant aux entités nommées correspond, dans la présente section, à la spécification de différentes classes de nature sémantique.

[1]Définition du Petit Robert.
[2]Définition du TLFi (http://atilf.atilf.fr/tlf.htm).

Du point de vue du traitement effectif du texte, ce sont ces catégories qui servent à annoter les entités nommées, tenant ainsi lieu d'étiquettes.

La détermination de catégories sémantiques pour l'appréhension des objets du monde n'est bien sûr pas nouvelle. Concernant les unités qui nous intéressent, de nombreuses classifications ont déjà été proposées, et ce bien avant les conférences MUC. Les travaux en onomastique ont en effet conduit à la spécification de typologies de noms propres, dont la plus connue semble être celle du linguiste germanophone G. Bauer ([Bauer, 1985], cité par [Daille et Morin, 2000] et également présentée par [Tran, 2006, p.22-23]). Cette typologie est constituée de six classes, énumérées ci-après, chacune comportant des sous-catégories :

1. Anthroponymes : noms de personnes individuelles ou de groupes.
2. Toponymes : noms de lieux.
3. Ergonymes : noms d'objets et de produits manufacturés.
4. Praxonymes : noms de faits historiques, de maladies, d'événements culturels.
5. Phénonymes : noms de phénomènes météorologiques.
6. Zoonymes : noms d'animaux familiers.

Cette typologie fait écho, à peu de choses près, à celle élaborée par W. Paik [Paik *et al.*, 1996], qui comporte 30 catégories réparties en 9 classes : Géographique, Affiliation, Organisation, Humain, Document, Equipement, Scientifique, Temporelle et Divers. Nous évoquons ces catégorisations dans la mesure où elles devancent de quelques années l'éclosion de la tâche de reconnaissance des entités nommées et parce qu'elles contiennent déjà les principaux éléments considérés par la suite en extraction d'information. Cependant, il sera question ici essentiellement des catégorisations déterminées à partir des conférences MUC, catégorisations imaginées dans le cadre précis du TAL et non dans celui de l'onomastique, s'intéressant à ce qui fut alors baptisé « entités nommées » et non aux (seuls) noms propres. Cette partition, ainsi motivée, ne doit toutefois pas être exclusive, étant entendu que les taxonomies d'hier ont marqué celles d'aujourd'hui. Examinons à présent les diverses catégorisations proposées lors des campagnes d'évaluation autour des entités nommées ou définies lors de projets particuliers soucieux d'établir des directives. Plus précisément, il sera question des directives d'annotation formulées lors des conférences MUC[1], CoNLL[2], IREX[3], ACE[4], HAREM[5], ESTER[6], ainsi que des décisions adoptées lors d'un projet autour du suédois,

[1] [MUC-6, 1995] et [Chinchor, 1997]
[2] [TjongKimSang, 2002] et [TjongKimSang et Meulder, 2003]
[3] [Sekine et Isahara, 1999]
[4] [ACE, 2000] et [ACE, 2005]
[5] [Santos *et al.*, 2006]
[6] [Meur *et al.*, 2004]

*Nomen Nescio Project*. Il est à noter que ces évaluations et projets affichent tous une volonté de traiter des textes relevant d'un domaine général et qu'ils ne visent aucune application particulière.

### 2.1.2 Catégorisations : de ressemblances en dissemblances

#### 2.1.2.1 Du choix des catégories...

Quelle catégorisation adopter pour la reconnaissance d'entités nommées ? Si la réponse peut de prime abord paraître facile à donner, on remarque bien vite qu'au-delà de quelques catégories principales, il est malaisé d'atteindre un consensus. À cet égard, la première conférence MUC ayant défini une tâche sur les entités nommées (organisée, rappelons-nous, en 1995) exigeait la reconnaissance des entités relevant des catégories suivantes, elles-mêmes organisées en trois classes :

| | |
|---|---|
| ENAMEX | organisation<br>personne<br>lieu |
| TIMEX | date<br>expressions de temps |
| NUMEX | valeur monétaire<br>pourcentage |

La classe ENAMEX est définie comme comprenant des noms d'entités, la classe TIMEX des expressions temporelles et la classe NUMEX des expressions numériques. Relativement à ces dénominations, il est intéressant de noter un certain flottement autour de l'appellation « entité nommée » : dans ses *guidelines*, tout comme dans l'ensemble des articles publiés dans ses actes, MUC parle de *Named Entity task* de manière générale, soit « Tâche de reconnaissance d'entités nommées ». Elle prend soin en revanche, lors de la description en détail de cette dernière, de distinguer les *named entities* (de la classe ENAMEX) des *temporal expressions* et *number expressions*. Les entités nommées ne correspondraient-elles qu'aux noms de personnes, lieux et organisations comme le laisse entendre cette dernière remarque, ou au contraire comprendraient-elles également les expressions numériques et celles de temps ? L'usage de la dénomination « entités nommées » dans les nombreux articles publiés par la suite semble donner raison à l'interprétation « globale ». Amalgame inopportun ou affaire de nuances en fonction de statistiques d'occurrences, l'usage a tranché dans le flou originel de la terminologie MUC, nous reviendrons par la suite plus en détails sur cette question (cf. chapitre 5 avec une proposition de définition des entités nommées).

Cette catégorisation, reprise telle quelle lors de MUC-7 en 1998, semble avoir pour partie servi de base à d'autres catégorisations. IREX, organisée quelques années plus tard au Japon et revendiquant l'héritage des conférences MUC, reprend effectivement les 7 catégories énumérées ci-dessus, avec cependant une catégorie supplémentaire, ARTIFACT, pour annoter, entre autres, les noms de produits (*Pentium Processor*) ou de prix (*Nobel Prize*). Il est également prévu, lors de cette évaluation, une étiquette OPTIONAL afin d'annoter une entité pour laquelle il serait difficile de déterminer la catégorie. Ainsi, même s'inscrivant explicitement dans le sillage de la définition d'une tâche, il semble difficile de se calquer sur une catégorisation, l'observation d'entités dans les textes conduisant naturellement à l'évocation, et donc à la détermination, de nouvelles catégories. Les conférences CoNLL 2002 et 2003 procèdent elles aussi à quelques changements par rapport à la catégorisation de MUC : reprenant les catégories PERSONNE, LIEU et ORGANISATION, elles oublient les expressions numériques et temporelles et ajoutent la catégorie MISCELLANEOUS pour annoter toutes les entités n'appartenant pas aux trois autres catégories. Cette tâche comprend donc un nombre limité de catégories, soit quatre étiquettes possibles assignables aux entités identifiées dans le texte.

La première modification importante de la catégorisation de MUC fut apportée durant la campagne ACE. Nous avons déjà évoqué la différence de perspective adoptée à l'égard des unités dans ce projet : il ne s'agit pas de reconnaître des entités nommées mais des entités, c'est-à-dire l'ensemble des réalisations lexicales des entités nommées, soit des noms propres, des noms communs et des pronoms. En dépit de ce large spectre d'annotation, il ne convient pas de tout annoter et des catégories sont donc déterminées afin de limiter les objets à prendre en compte. La tâche pilote de ACE prévoyait les catégories suivantes : PERSONNE, ORGANISATION, BÂTIMENTS (*facility*), GPE et LIEUX. GPE signifie *Geo-Political Entities*, et concerne les entités géographiques également définies par des aspects politiques et sociaux. Cette catégorie, non prévue au départ, est née de l'expérience d'annotation des organisateurs de la tâche, se trouvant démunis face à certaines occurrences telles que « *Miami imposed a curfew*[1] » , où une entité réfère simultanément à plusieurs types (dans notre exemple : lieu géographique et organisation municipale). Si, du dire des auteurs[2], cette catégorie composite peut paraître un peu « fourre-tout » , elle permet néanmoins de rendre compte de la sémantique de certaines entités et d'éviter bien des hésitations lors de l'annotation. Par contraste, la catégorie LIEU ne désigne plus que les entités strictement géographiques. En définitive, lors de la définition de la tâche réelle, deux catégories

[1] Miami a imposé un couvre-feu.

[2] *This type can be viewed as somewhat synthetic and ad hoc, but there is also support for its conceptual reality (...).*

supplémentaires furent encore ajoutées : VÉHICULE et ARME.

Poursuivons l'exploration des différentes catégorisations existantes au-delà des campagnes les plus célèbres. Peu de temps après ACE, est organisé conjointement avec plusieurs universités scandinaves le projet *Nomen Nescio*[1]. L'objectif est de reconnaître, classifier et annoter des noms (*names*) dans des textes courants. Par conséquent, des directives d'annotation d'entités nommées sont éditées pour clarifier la tâche [Kokkinakis et Thurin, 2007], dans lesquelles D. Kokkinakis précise en préambule : « *Le contenu de ce document reprend des idées formulées lors des conférences MUC, qui identifient et utilisent sept types de noms, lors de ACE, qui utilise cinq types de noms, et lors de* IREX*, qui utilise huit types de noms ; il s'inspire également de discussions initiées lors de réunions (...), ainsi que de l'expérience de l'auteur (...).*[2]. Malgré cette ascendance clairement affirmée, seules les catégories « universelles » sont présentes dans ce projet, et trois autres sont ajoutées : EVENT, OBJECT et WORK&ART. La diversification s'accentue. Regardons encore du côté des campagnes ESTER et HAREM : la campagne française ESTER impose la catégorisation des entités nommées en 8 catégories, que sont : PERSONNE, ORGANISATION, GROUPE GEO-SOCIO-POLITIQUE (GSP), LIEU, BÂTIMENT, PRODUIT, TEMPS et QUANTITE, auxquelles est ajoutée, en cas d'incertitude, la catégorie INCONNU ; l'héritage de ACE se fait ici sentir au travers des catégories GSP (proche de GPE) et BÂTIMENT, celui de IREX au travers de la catégorie PRODUIT. La campagne portugaise HAREM propose elle aussi un nombre relativement conséquent de catégories : aux trois fondamentales sont ajoutées les catégories ÉVÉNEMENT, CHOSE, ŒUVRE, ABSTRACTION, TEMPS, VALEUR, ainsi que le désormais habituel DIVERS.

Ainsi, à mesure que la tâche de reconnaissance des entités nommées se développe et suscite diverses campagnes, les catégories selon lesquelles ces unités sont classées et annotées deviennent de plus en plus diverses. Hormis la triade universelle personne-organisation-lieu issue de la catégorisation initiale de MUC, les variantes sont légions, allant des armes aux abstractions en passant par les produits. Cette multiplication des catégories est due avant tout à la confrontation des annotateurs avec des textes et traduit une volonté de généraliser la tâche. Commentant l'ajout de la catégorie artefact dans IREX, S. Sekine [Sekine *et al.*, 2002] explique en effet que « *this was done because, in the project, no application was presupposed and NE extraction was the final target. So, the generalization was taken into account* ». Il renchérit plus loin, à propos des deux nouveaux types de

[1] http ://g3.spraakdata.gu.se/nn/

[2] Dans [Kokkinakis, ] : « *The content of this document incorporates ideas from the MUC, which recognized and used seven name types,* ACE*, which used five name types,* IREX*, which used eight name types, also the discussions initiated in the two* Nomen Nescio-Fefor *meetings (...), and the author's own experience (...)* ».

ACE : « *(...) two new entities are added to pursue the generalization of the technology.* ». Au final, ce mouvement fait de changements et d'extensions témoigne de la complexité du travail de détermination des catégories à reconnaître.

#### 2.1.2.2 ... à la spécification de ce qu'elles recouvrent.

Cette difficulté n'est pourtant pas la seule. En effet, une fois déterminées les catégories à prendre en compte, encore faut-il savoir ce qu'elles recouvrent précisément. A titre d'illustration, intéressons-nous simplement à la catégorie PERSONNE et considérons les expressions suivantes :

| | |
|---|---|
| *Lionel Jospin* | *les Peuls* |
| *les Kennedys* | *Bison futé* |
| *la famille Kennedy* | *Mickey* |
| *l'epouse Chirac* | *Zorro* |
| *les Windsor* | *Hercule* |
| *les freres Coen* | *le Prince Charmant* |
| *Zizou* | *l'ours Franska* |
| *les Democrates* | *Milou* |
| *les italiens* | *St. Nicolas* |
| *les Talibans* | *Vichnu* |

Nous avons là des noms de personnes, des surnoms, des expressions familiales, des nationalités, des groupes, des personnages de fictions, des animaux, des personnages religieux, des divinités... Ces entités font-elles toutes partie de la catégorie PERSONNE ? Comment en décider ? Les divers *guidelines* ont à cet égard produit de nombreuses indications, sous la forme le plus généralement de sous-types. Ainsi, si MUC ne précise aucune sous-catégorie pour les noms de personne et indique seulement qu'il convient d'annoter les personnes et les familles, les autres campagnes fixent par la suite des sous-types de plus en plus nombreux, afin d'expliciter quelque peu les choses. ACE divise sa catégorie PERSONNE en *individus*, *groupes* et *indéfinis*, cette dernière division comprenant les saints et figures religieuses, les personnages de fiction ainsi que les noms d'animaux. Les organisateurs de la campagne ESTER choisissent une sous-classification en *humains*, *animaux* et *imaginaires*. Ne sont donc pas pris en compte ici les *groupes*, ces derniers se retrouvant néanmoins dans les sous-types définis dans le projet *Nomen Nescio*, sous la forme de *collectif*, aux côtés des *humains*, *mythiques* et autres *animaux*. Cette diversité, illustrée ici à l'aide des noms de personne, se retrouve bien sûr avec les autres types, pour lesquels il est tout aussi difficile de décider ce qu'ils recouvrent vraiment.

On le voit, déterminer les catégories à l'aide desquelles classer et annoter les entités nommées, ainsi que décider de ce qu'elles contiennent précisément n'est pas chose facile. Avoir une idée de ce que l'on veut chercher dans un texte, cela est relativement aisé de prime abord mais, confrontée aux textes, l'efficacité des catégorisations "basiques" est vite remise en cause, conduisant alors à la multiplication des catégories. L'examen de quelques directives d'annotations élaborées pour des projets sur différentes langues a permis de mettre en évidence ce phénomène, illustrant la difficulté d'atteindre, sinon un consensus[1], du moins une classification stable et suffisamment couvrante pour le domaine général le plus souvent ciblé par les tâches de reconnaissance des entités nommées. Une fois pointées ces difficultés, il peut être intéressant de considérer quelques démarches s'étant penchées en particulier sur ce problème.

### 2.1.3 Comment penser la catégorisation des entités ?

Deux choses font difficulté au regard des catégories d'entités nommées : leur choix et leur niveau de granularité. Existe-il un moyen d'amoindrir ces difficultés, autrement dit existe-t-il une méthode de détermination des catégories ? Sur ce point, deux travaux méritent attention : l'étude en corpus pour une catégorisation des noms propres[2] de B. Daille *et al.* [Daille *et al.*, 2000] ainsi que l'ensemble des travaux et publications de S. Sekine autour d'une « catégorisation étendue » des entités nommées [Sekine *et al.*, 2002, Sekine et Nobata, 2004, Sekine, 2004]. Ces études se situent toujours dans le cadre d'applications générales, les catégorisations et systèmes étant voués à traiter des textes de type journalistique.

En une étude préalable à la conception d'un système de reconnaissance d'entités nommées, B. Daille *et al.* choisissent d'explorer du corpus afin de déterminer la catégorisation la plus adéquate. Cette démarche semble retenir les leçons des hésitations rencontrées lors des conférences précédentes et se démarque donc quelque peu de ces dernières, si tant est qu'elles n'ont, à notre connaissance, rendu compte d'aucune indication méthodologique quant à la catégorisation. Le processus est le suivant : partant de la catégorisation de G. Bauer (anthroponymes, toponymes, ergonymes, etc. cf. section 2.1.1), l'objectif est de confronter les catégories prédéterminées à leur réalisation, effective ou non, sous la forme d'entités nommées dans des textes. Les auteurs ont constitué un corpus composé de deux périodiques français (la revue *La Recherche* et le journal *Le Monde* de l'année 1987) et ont ensuite tenté de placer les entités rencontrées dans les classes et catégories

[1] Si tant est que cela soit un objectif ; ce point sera de nouveau discuté par la suite, dans la partie s'intéressant au statut théorique des entités nommées, cf. partie II.

[2] Il s'agit bien de *noms propres* dans le titre de [Daille *et al.*, 2000] mais le terme d'*entités nommées* prend bien vite le dessus dans le reste de l'article.

de Bauer. Leur conclusion est que, si une grande partie des entités rencontrées trouvent place dans des catégories, il est « *nécessaire d'étendre certaines catégories et d'en créer de nouvelles* ». L'étude statistique des occurrences des entités nommées pour chaque catégorie (au sein de la hiérarchie de catégories élaborée par Bauer et étendue par les auteurs) montre cependant que la catégorisation retenue est suffisante pour assurer une bonne couverture des entités dans ce type de corpus, et avec un degré de détail acceptable. Ayant ainsi choisi, travaillé puis validé leur catégorisation, les auteurs ont pu passer par la suite à la réalisation d'un système [Fourour, 2002]. Ce travail n'apporte pas de réelle solution aux difficultés exposées ci-dessus, il atteste toutefois d'un souci méthodologique à l'encontre de la détermination des catégories, préférant l'ajustement d'une catégorisation en fonction d'une étude de corpus à l'adoption « aveugle » ou au panachage de catégorisations existantes.

Les travaux de S. Sekine prennent à bras le corps cette question des catégories, avec la proposition d'une hiérarchie de 150 catégories [Sekine *et al.*, 2002], puis de 200 catégories [Sekine et Nobata, 2004, Sekine, 2004]. Passant en revue les décisions de MUC, IREX puis ACE, Sekine [Sekine *et al.*, 2002] observe qu' *« À la suite de ces expériences, nous avons pu constater que 7 ou 8 catégories ne sont pas suffisantes pour couvrir les principaux problèmes. Beaucoup de types de choses ont des noms propres ou des classes propres d'expressions, et des distinctions plus fines sont également nécessaires pour certaines applications »* [1]. S'il prend soin de reconnaître que différents types d'entités sont nécessaires en fonction du domaine d'application, l'auteur insiste cependant sur le fait que ses travaux sont pour des applications générales (*« nous ne visons pas la couverture de domaines spécifiques »*[2]). Partant, il tente d'élaborer une hiérarchie de catégories capable de couvrir le plus d'entités nommées possibles. La construction de la classification comprend trois étapes : tout d'abord, trois hiérarchies sont construites, selon trois méthodes différentes ; une fusion de ces hiérarchies est ensuite réalisée et, enfin, il est procédé à un raffinement de la hiérarchie ainsi obtenue grâce à sa confrontation avec des corpus. S. Sekine adopte donc une démarche similaire à celle de B. Daille *et al.*, à savoir le choix d'une catégorisation puis sa confrontation avec la réalité, à la différence qu'ici la catégorisation n'existe pas préalablement à la tâche de reconnaissance des entités nommées envisagée (les travaux de G. Bauer participaient de l'onomastique), mais est constituée de toutes pièces pour ce but précis, et ce à partir de trois sources. Celles-ci correspondent à :

– du texte : quelques centaines d'entités furent extraites de journaux et caté-

[1] *'From those experiences, we recognized that 7 or 8 kinds are not broad enought to cover general issues. Many kinds of things have proper names or proper classes of expressions, and also finer disctinctions are needed for some applications. »*

[2] *« we are not aiming to cover special domains. »*

gorisées librement, en fonction de l'intuition des annotateurs. Les catégories déterminées furent ensuite organisées en une hiérarchie.

- des catégorisations utilisées dans d'autres systèmes : des tâches d'extraction d'information et de *question(s)-réponse(s)* furent analysées.
- des ressources lexicales : Wordnet et le *Roget's Thesaurus* furent explorés.

Ces trois sources correspondent à des types de connaissances différents : connaissance du monde avec les textes, connaissance des besoins du TAL avec les différentes tâches étudiées, et connaissance « normalisée » avec les ressources lexicales[1]. L'utilisation conjointe de ces dernières peut paraître un gage d'objectivité dans la détermination de la catégorisation recherchée, celle-ci ne pouvant être que la plus complète possible. Par ailleurs, lors de l'élaboration de cette catégorisation, S. Sekine énonce quelques principes permettant de cadrer le processus ; parmi ces derniers, nous mentionnons les suivants :

- pour classer une entité dans une catégorie, le critère essentiel est sa forme de surface et non son sens (*In order to minimize the ambiguity, we try to take the surface form as the primary clue as much as possible.*). Certains types introduits dans la hiérarchie (comme GEO-POLITICAL ENTITY et GROUPE SOCIO-POLITIQUE) permettent de classer les entités participant de plusieurs types simultanément.
- le degré de granularité d'une catégorie est reconnu comme pouvant être arbitraire. Le principe de détermination de sous-catégorie doit obéir au principe suivant : une division est relative à l'entité elle-même (*i.e.* à une de ses propiétés intrinsèques, le genre par exemple) et non relative à son contexte (profession d'une personne par exemple).

Au terme de ce travail, S. Sekine *et al.* sont amenés à définir 150 puis 200 catégories. On peut bien sûr discuter ce nombre, mais l'intérêt n'est pas là ; il réside bien au contraire dans la méthode employée, innovante et « complète » , contrastant avec les premières décisions de MUC.

Au final, 8 ou 200 catégories, cela ne semble pas être le plus important. L'essentiel, pour penser une catégorisation utilisable pour la reconnaissance d'entités nommées, est de prendre en compte le domaine applicatif, bien que celui-ci soit presque exclusivement général pour les entités nommées, de considérer la catégorisation comme un véritable enjeu et d'adopter une démarche méthodologique pour sa réalisation. Il n'existe bien sûr aucune catégorisation idéale ni de solution pour y parvenir ; le mieux semble être de suivre la proposition de S. Sekine, *« We believe that there is no ultimate solution, so we seek rather empirical solution »*, et de multiplier les sources d'inspiration.

[1]La construction de cette catégorisation d'entités nommées est également présentée par [Tran, 2006, p.25]

Opérer un choix dans les objets (au sens large) du monde à prendre en compte pour la reconnaissance des entités nommées constitue donc une étape difficile. L'étendue des possibles est vaste, il n'existe aucune catégorisation « idéale » , ni même de manière bien établie de concevoir les choses et de conduire le travail. Ces difficultés de catégorisation des entités nommées ne sont cependant pas les seules : de semblables complications apparaissent également à un autre niveau, celui de l'annotation de ces unités.

## 2.2 Les entités nommées dans le texte : le problème de l'annotation

Une fois déterminé ce que l'on cherche à annoter, encore faut-il savoir comment le faire exactement. En effet, si les objets à prendre en compte sont bien circonscrits relativement à leurs caractéristiques référentielles, il faut désormais s'intéresser à leur réalisation dans le texte, à leurs différentes mentions. De nouveau, les directives d'annotation publiées lors des conférences décrivent abondamment leurs prescriptions en la matière et, de nouveau, les avis divergent. Il peut être intéressant, pour la description des difficultés d'annotation que nous cherchons ici à mettre en avant, de procéder par étapes et de se rapprocher progressivement du cœur du problème. Ainsi, il sera question du comportement de l'entité nommée vis-à-vis des autres syntagmes, puis de son annotation en tant que syntagme seul et, enfin, de ses différentes mentions possibles. Pour chacun de ces niveaux, apparaîtront des difficultés ainsi que des questionnements quant à l'annotation à mettre en œuvre. Ce tour d'horizon des difficultés d'appréhension des entités nommées dans le texte s'achèvera par l'évocation d'une question complémentaire, celle de la normalisation.

### 2.2.1 Entité nommée et combinaisons de syntagmes : une ou plusieurs entités ?

Parcourant un texte, il est rapidement possible de se rendre compte de l'existence de différentes constructions dans lesquelles les entités nommées peuvent apparaître. Combinées les unes avec les autres, sous forme de coordination ou d'imbrication, les entités nommées peuvent poser difficulté au regard de leur annotation effective : quelle entité doit réellement recevoir une étiquette est une des hésitations les plus courantes.

Pour ce qui est de la coordination, examinons les exemples suivants, glanés dans divers journaux, le type de corpus par excellence pour les entités nommées :

*Bill and Hillary Clinton flew to Chicago together last month (...).* (New York Times).

*M. et Mme. Chirac en thalasso à Biarritz.* (Site de France 3)

*Les Banques centrales européenne et américaine ont été amenées à intervenir en urgence hier.* (Le Nouvel Observateur)

Pour chacune de ces expressions on observe l'effacement d'un des constituants communs des syntagmes coordonnés. L'élément restant peut être plus ou moins fort sémantiquement, on parle alors d'ellipse partielle (pour *Bill and Hillary Clinton*) ou d'ellipse totale (pour *M. et Mme Chirac*). Concernant l'annotation de ces expressions, doit-on considérer qu'il s'agit d'une ou de plusieurs entités ? Leur traitement doit-il être uniforme ? Peu de guides d'annotation se prononcent à cet endroit mais il est malgré tout possible d'observer une certaine hétérogénéité dans les décisions prises. Si MUC-6 préconise l'annotation séparée de deux entités coordonnées dont l'une est partiellement ellipsée (*North and South America* donne lieu à l'annotation de deux entités), les choses changent lors de MUC-7, avec la consigne d'une annotation conjointe (« North and South America *should be marked up as a single expression* »). La modification est identique à l'encontre des expressions numériques avec, pour l'expression *10 and 20 dollar bills*, une annotation de *10* d'une part et de *20 dollar* d'autre part en 1995, puis la réunion des deux annotations en 1998. La campagne ESTER préconise, à l'instar de la première conférence américaine, l'annotation séparée des entités coordonnées et ellipsées, avec cependant la restitution du constituant manquant. On a alors pour Bill et Hillary Clinton l'étiquette <PERSONNE> pour *Bill Clinton* d'une part et pour *Hillary Clinton* d'autre part. Pour les ex-époux présidentiels français, le patronyme ellipsé est restitué, l'annotation étant alors : *M.* <PERSONNE>*Chirac*<PERSONNE> *et Mme* <PERSONNE>*Chirac*<PERSONNE>, sans prise en compte des titres. Enfin, il est possible de citer les directives rédigées pour un système d'annotation des entités nommées pour le français au sein de XRCE[1] (A. Rebotier, [Rebotier, 2006]). Encore une fois, les choses sont différentes, A. Rebotier préconisant l'annotation de deux entités pour *Bill and Hillaty Clinton*, mais d'une seule pour *M. et Mme. Chirac*, arguant du fait que *Monsieur* ne constitue pas une entité et qu'il peut être difficile de mettre en œuvre la restitution du constituant en commun.

Une autre hésitation peut également apparaître quant à l'annotation des entités imbriquées, construction que l'on peut illustrer à l'aide des exemples suivants :

*l'Université de Corte*[2]

[1] *Xerox Research Centre Europe* à Grenoble, France.
[2] Exemple emprunté à [Meur *et al.*, 2004].

*le Comité exécutif de l'Union des associations européennes de football*[1]
*Boston Chicken Corp.*[2]

Faut-il annoter les entités contenues dans d'autres ou bien préférer une annotation globale ? MUC recommande de ne pas décomposer les entités et donc de ne pas annoter les « sous-entités », que les types de chacune d'entre elles soient similaires ou non. Suivant cette indication, *Boston Chicken Corp.* recevra l'étiquette <ORGANISATION>, mais *Boston* ne sera pas reconnue comme <LOCATION>. Les directives d'ESTER opèrent quant à elles une distinction en fonction de l'équivalence ou non des types de chacune des entités imbriquées. Dans cette optique, *l'Université de Corte* reçoit deux annotations, l'une pour *Corte* (lieu), et l'autre pour *Université de Corte* (organisation) et le *Comité exécutif de l'Union des associations européennes de football* n'en reçoit qu'une seule. Si elle s'en tient à la décision de MUC, A. Rebotier note cependant que ce mode d'annotation constitue une perte d'information.

Coordination ou imbrication, il n'est donc pas toujours simple de décider comment annoter des entités nommées en présence de telles constructions. La question qui se pose semble être celle de la distinction entre des entités évoquées et des entités sur lesquelles opère une prédication : si l'expression *M. et Mme. Chirac* ne fait réellement état que de *Mme. Chirac*, la prédication porte sur les deux personnes et si l'expression *Université de Corte* évoque la ville de *Corte*, la prédication ne porte réellement que sur l'*Université*. Le discours évoque ainsi plus ou moins d'entités, lesquelles ne sont pas toutes nécessairement le support d'une prédication. À cet endroit, les décisions des uns et des autres diffèrent, de nouveau l'hétérogénéité est de mise et bien peu d'éléments permettent de trancher en faveur de tel ou tel choix. Si la combinaison de syntagmes d'entités nommées pose difficulté, d'autres questions affleurent également à l'examen des syntagmes seuls.

### 2.2.2 Entité nommée et syntagme : quelles frontières ?

Les incertitudes portent ici sur les " atours " des entités nommées, à savoir leurs divers modifieurs. La question est celle de la prise en compte ou non des éléments entrant dans la constitution d'un syntagme dont la tête est une entité nommée. En d'autres termes, il s'agit de l'étendue de l'entité nommée et du problème de l'annotation des entités telles que :

[1] Exemple emprunté à [Rebotier, 2006].
[2] Exemple emprunté à MUC.

| | |
|---|---|
| *Le Palais Bourbon* | *la candidate Ségolène Royal* |
| *Les Rolling Stones* | *le Président Nicolas Sarkozy* |
| *Le téléphone sonne*[1] | *l'Abbé Pierre* |
| *La Mecque* | *Benoît XVI* |
| *le cardiologue Dupont* | *George W. Bush Jr.* |
| *Monsieur Fillon* | *Lord Liverpool* |
| *Professeur Paolucci* | *Secretary of State Colin Powell* |

Sont présents ici pêle-mêle des déterminants, des titres, des professions, des qualifications ou encore des désignateurs générationnels. Faut-il les prendre en compte ou non dans l'annotation des entités nommées ? Les directives de MUC sont constantes sur ce point, ne changeant pas d'une conférence à l'autre : les titres et les noms de rôle ne doivent pas être annotés avec l'entité nommée (*Monsieur*, *Président*), mais les désignateurs générationnels doivent l'être (*Jr.*, *XVI*) ainsi que les suffixes propres aux organisations (*Inc.*, *Corp.*, *SA*). ESTER précise pour sa part que "*seule l'entité doit être prise en compte dans l'étiquette*" ; déterminants (*Le Palais Bourbon*), titres (*Monsieur Fillon*) et autres fonctions (*le Président Nicolas Sarkozy*, *le cardiologue Dupont*) sont ainsi exclus de l'annotation, sauf bien sûr s'ils font partie intégrante de l'entité (*Le téléphone sonne*, *Les Rolling Stones*). A. Rebotier opère quant à elle une distinction nette entre les titres et les qualifications, considérant celles-ci comme externes au nom, mais ceux-là comme en faisant partie intégrante. Elle annote ainsi *Monsieur Fillon* dans son entier, ne conservant que le nom propre dans *le cardiologue Dupont*. Une fois de plus, les instructions divergent et, malgré quelques justifications, aucun principe fort de distinction entre ce qui fait partie d'une entité nommée et ce qui en est exclu n'apparaît. Ces quelques considérations conduisent tout droit à la question de ce qui doit être compris comme une entité nommée.

### 2.2.3 Entité nommée et... entité nommée

La considération de tous ces problèmes d'annotation amène au final à la question des différents types de mentions pouvant être considérés comme une entité nommée, soit à la question de l'entité nommée elle-même. Une fois encore, la meilleure façon de poser et de saisir la difficulté passe par la considération de quelques exemples :

*Jacques Chirac*
*le Président Jacques Chirac*
*Chichi*

[2]Il s'agit d'une émission radiophonique.

*le président français*
*le président de la République Française*

*l'Association Sportive de Saint Etienne*
*l'ASSE*
*le club forézien*

*Elisabeth II*
*la reine d'Angleterre*

Parmi ces expressions, que doit-on considérer comme une entité nommée ? La remarque qui bien sûr vient immédiatement à l'esprit porte sur la prise en compte des seuls noms propres. Est-ce à dire que les autres types d'expressions ne sont pas à même de jouer un rôle similaire ? Que faire des acronymes, des descriptions définies ou encore des noms dont il est difficile de dire s'ils sont propres ou non ? Comme de coutume, examinons ce que disent les *guidelines*. MUC précise, au regard des ENAMEX, que la tâche est limitée « *aux noms propres, aux acronymes, et peut-être à divers autres identifiants uniques* »[1]. Cette formule n'est pas d'une grande aide quant au problème d'annotation posé ci-dessus. Les directives d'ESTER, après avoir listé les différentes mentions d'entités nommées possibles, du nom propre au pronom en passant par les descriptions définies, indiquent que seuls les noms propres doivent recevoir une étiquette. Il est également mentionné un principe dit « de catalogue » permettant de prendre une décision en cas d'hésitation : « *si on peut aisément imaginer l'entité nommée comme étant une entrée d'un catalogue, annuaire, dictionnaire ou index alors celle-ci sera bien une entité nommée* ». Et, pour finir, A. Rebotier précise au début de son document que « *les entités nommées qui ne sont pas des dates doivent être des noms propres, ce qui exclut les expressions anaphoriques et les paraphrases* », avant d'indiquer plus loin que « *lorsqu'il était difficile de décider s'il s'agissait ou non d'un nom propre, [elle] a fait des choix au cas par cas, [se] demandant s'il était potentiellement utile de reconnaître l'expression* ». « Divers identifiants uniques » pour MUC, « principe du catalogue » pour ESTER, « décisions au cas par cas » pour d'autres, les directives d'annotation n'apportent de toute évidence que peu d'informations, ne déterminant que très peu de critères d'identification d'une entité nommée. Loin de se vouloir polémique, l'intention de ces observations n'est pas tant de dénoncer certains manquements des guides d'annotation que de mettre l'accent sur un point difficile, celui de la caractérisation de l'objet à prendre en

[1] *« This subtask is limited to proper names, acronyms, and perhaps miscellaneous other unique identifiers ».*

compte. Plus que l'énonciation de principes réfléchis, c'est une sorte de confiance à la compréhension intuitive de ce qu'est une entité nommée qui semble dominer. Or les problèmes d'annotation sont bien là ; aux côtés d'*Elisabeth*, de *Jacques* et de l'*ASSE*, doit-on oublier *la reine d'Angleterre*, *le président français* et *le club forezien* ? Que faire du *Tour de france* ou encore de l'*équipe Festina* ? Ces expressions peuvent être à annoter ou non, l'intérêt serait peut-être de formaliser des critères de décision.

Cette difficulté de définir une entité nommée trouve un écho dans la question de ce qu'on appelle la « normalisation ». Ce processus, difficile à définir et circonscrire tant il est évoqué dans des situations différentes, s'applique aux entités nommées dans la mesure où il convient de représenter les différentes mentions d'entités apparaissant dans un texte.

### 2.2.4 La question de la normalisation

Qu'est-ce que la normalisation ? Appliquée aux entites nommées, la normalisation est un processus qui ne semble pas recevoir de définition unanime, à l'instar des nombreuses autres « normalisation(s) » souvent invoquées lors de divers traitements automatiques de la langue. L'idée sous-jacente présente dans tout mouvement de normalisation est naturellement celle de l'application d'une norme aplanissant ou supprimant des variations, afin de réguler la forme, le comportement ou une autre caractéristique des objets considérés. La mise en œuvre d'une normalisation implique donc de déterminer ce que l'on cherche à normaliser, et en fonction de quelle norme. Au regard des entités nommées, ce procédé peut chercher à neutraliser des variantes de différentes natures :

*Chirac*
*Jacques Chirac*
*J. Chirac*
*le Président Chirac*
*le Président de la République*

De cette liste comprenant différentes réalisations lexicales d'un nom de personne, il est possible de dégager trois types de variations, jouant à des niveaux différents. La variation la plus évidente est celle de nature graphique : le nom *Jacques Chirac* peut effectivement apparaître sous des graphies différentes, avec l'initialisation (*J. Chirac*) ou la suppression (*Chirac*) d'un ou de plusieurs composants du nom. Le second type de variation est de nature lexicale, le nom *Jacques Chirac* étant alors « décliné » sous d'autres formes comme *le Président Chirac* ou *le Président de la République*. Enfin, il convient d'évoquer une troisième variation

possible, de nature référentielle cette fois-ci, la forme *Chirac* pouvant renvoyer à l'ex-Président, à une commune française en Charente, ou encore en Lozère[1]. Variations graphiques, lexicales ou référentielles, ces dernières ne sont pas pour faciliter le traitement des entités nommées et c'est ainsi que des « normalisations » s'imposent, plus ou moins complexes.

La première normalisation, la plus « basique » , correspond à une harmonisation des formes de surfaces par le biais de l'assignation, pour un type d'entité nommée donné, d'une forme canonique dans laquelle devront être converties toutes les entités de ce type. Ainsi, pour notre exemple, les graphies *Jacques Chirac, J. Chirac*, devront être normalisées sous un format préalablement déterminé, de type `[initiale prénom-nom]` par exemple, ce format étant par ailleurs appliqué aux autres noms de personnes. S'agissant de la catégorie DATE, un processus de normalisation graphique s'attachera à transformer toutes les mentions de dates, qu'il s'agisse de dates identiques ou différentes, sous un format numérique tel que `[jour/mois/année]`. Cette standardisation graphique, correspondant à une volonté de « *s'abstraire de la forme de surface pour restituer avec régularité une même information, un même type d'information sous un même format* » [Guillemin-Lanne et Six, 2006], apparaît utile pour un remplissage efficace de bases de données ou de *templates* prédéfinis.

Ce processus revêt cependant une autre dimension si l'on s'intéresse, au-delà des chaînes de caractères, au référent dénoté par une entité : la normalisation n'intervient plus seulement au niveau d'un type donné, mais au niveau d'une entité nommée donnée. Autrement dit, dans cette perspective, la forme canonique ne l'est plus relativement à un type, mais relativement à une entité, normalisant l'ensemble de ses mentions, que celles-ci affichent des variantes graphiques ou lexicales. Ainsi, ce sont l'ensemble des mentions *Jacques Chirac, J. Chirac, le Président Chirac* qui sont rassemblées sous une même entité canonique. Le choix de cette dernière peut se faire, entre autres, en fonction de la fréquence d'apparition, en fonction de la première occurrence dans le texte, de l'occurrence la plus complète ou encore d'une entité de référence issue d'une ressource externe. Cette deuxième forme de normalisation produisant des listes d'entités de même référent, si elle est plus complète que la première, est aussi plus complexe. Elle est, de toute évidence, davantage orientée vers un traitement des chaines de coréférence entre entités nommées (traitement qui, au final, pourra ajouter *le Président de la République* à la liste évoquée ci-avant). Il peut bien sûr être envisagé de combiner ces deux types de normalisation avec, pour le choix de l'entité représentative d'un référent particulier, imposition d'une forme de surface canonique.

[1]Cet exemple n'est sûrement pas le meilleur, d'autres cas sont plus flagrants, il en sera question dans le paragraphe 2.3.

Ces deux types de normalisation s'intéressant, pour l'un, à rassembler sous une même forme canonique différentes chaînes de caractères et, pour l'autre, à rassembler sous une même entité canonique différentes mentions renvoyant à un même référent, correspondent à ce que nous souhaitons, pour notre part, appeler « normalisation » , selon une interprétation quelque peu plus restrictive que la normale. En effet, cela a été évoqué ci-dessus, une variation d'un autre type peut encore advenir à l'endroit des entités nommées et donc conduire à une autre forme de normalisation. Cette dernière, sortant du cadre « texte » dans lequel s'inscrit ce paragraphe, sera évoquée dans le paragraphe suivant.

Ainsi, s'il est donc difficile de déterminer au préalable le type des objets à reconnaître (cf. section 2.1), il est tout aussi délicat de spécifier ce qui doit réellement être annoté dans le texte. Nombreux sont les points qui posent difficulté, qu'il s'agisse de constructions particulières (mais courantes), des frontières de l'unité lexicale à prendre en compte, ou le type même de cette dernière. La tâche de reconnaissance des entités nommées doit ainsi faire face à des problèmes de catégorisation et d'annotation, auxquels viennent s'ajouter de surcroît des problèmes de « polysémies ».

## 2.3 Les entités nommées dans la langue : le problème des "polysémies"

### 2.3.1 Des unités polysémiques ?

Il est nécessaire, après le « monde » et le « texte », de considérer les entités nommées du point de vue de la langue. Il serait en effet dommageable de perdre de vue le matériau premier qui occupe le TAL, lui donnant bien du fil à retordre, à savoir le langage. Ce dernier offre à ses locuteurs d'innombrables possibilités d'expression, lesquelles sont rendues possibles grâce à la pluralité interprétative dont peuvent faire l'objet la plupart des unités lexicales d'une langue donnée. Les phénomènes d'homonymie, de métonymie, etc. sont loin d'être marginaux et occupent bien au contraire une place centrale dans les langues naturelles, constituant à leurs égards des gages de productivité et d'expressivité. Largement décrits et étudiés pour les unités lexicales classiques, ces changements, transferts ou superpositions de sens le sont en revanche très peu pour les unités de type entité nommée. Or celles-ci semblent bien être régies par les mêmes phénomènes. Le sont-elles cependant de manière identique ? Comment analyser et caractériser l'ambiguïté des entités nommées ? Considérons, à titre d'exemples rapides, les énoncés suivants :

*Orange a invité M. Dupont.*
*Leclerc a fermé ses magasins en Rhône-Alpes.*
*La France a signé le traité de Kyoto.*

Que décider quant à la catégorie de ces entités ? Est-il question de la ville d'Orange ou bien de la société de téléphonie mobile ? De la personne Michel-Edouard Leclerc ou de la chaîne de supermarchés ? Faut-il préférer une annotation de France en tant qu' « organisation » ou « gouvernement » ou en tant que « lieu » ou « pays » ? Force est de constater que les entités nommées n'échappent pas aux phénomènes d'ambiguïté et sont, à l'instar des autres unités lexicales, polyréférentielles. Parallèlement à ces exemples « linguistiques », la réalité du traitement automatique de l'information révèle peu ou prou la même chose. Une interrogation du moteur de recherche Google pour les entités *Orange* et *Leclerc* propose (sur les dix premières réponses) pour l'une, des renvois aux référents « opérateur » et « ville » et, pour l'autre, des renvois aux référents « centres d'achat », « général », « char », et « homme d'affaire ».

Cette polysémie avérée des entités nommées ne semble pas avoir fait l'objet d'importantes considérations durant les premières campagnes d'évaluation et ne figurent à cet endroit que peu d'instructions particulières dans les diverses directives d'annotation. MUC fait bien une mention à de possibles phénomènes de métonymie mais, faisant la différence entre des métonymes « propres » et des métonymes « communs », recommande une annotation ne tenant majoritairement pas compte de ces phénomènes. En effet, seuls les métonymes « propres », soit des noms qui incluent par exemple « *une référence à une organisation par le biais d'un nom de bâtiment* sur la base d'une « *association (...) suffisament régulière pour justifier sa présence dans la définition dictionnairique du terme* », doivent être annotés en tant que tels. Les métonymes dits « communs » dont le glissement de sens est occasionnel gardent pour leur part une annotation selon la catégorie initiale de l'entité nommée. La Maison Blanche est ainsi une organisation dans *The White House announced*, mais l'Allemagne reste un lieu dans *Germany invaded Poland in 1939*. Ainsi, si elles évoquent bien le phénomène, les conférences MUC ne le prennent pas totalement en compte. Le programme de ACE, nous l'avons vu, a choisi d'introduire une nouvelle catégorie « geo-political entity ». Celle-ci, bien que ne permettant pas de différencier les usages, entre par exemple un nom de pays utilisé en tant que tel ou un nom de pays utilisé pour référer à son gouvernement, permet toutefois de rendre compte du phénomène. Les directives d'ESTER adoptent une position encore différente, avec la consigne d'annoter l'entité nommée employée métonymiquement à l'aide de ses deux ca-

tégories, celle de base à laquelle appartient le mot et celle issue du contexte[1]. Au final, seuls quelques cas de métonymie sont mentionnés dans les divers *guidelines* parcourus et les autres phénomènes de sens relevés ci-avant ne semblent pas faire l'objet de considérations particulières. Ils existent quoi qu'il en soit, porteurs de diverses potentialités de sens pour les entités nommées, mais également de difficultés pour l'annotateur et/ou le concepteur de système de reconnaissance de ces unités. Comment les définir précisément et comment les prendre en compte lors d'un traitement automatique ? Avant d'interroger ces phénomènes de polysémies au regard des entités nommées, il importe au préalable de les définir de manière générale.

## 2.3.2 La polysémie lexicale

### 2.3.2.1 Éléments de définition

La polysémie participe de la structuration sémantique du lexique et se rapporte plus spécifiquement aux mots auxquels sont attachées plusieurs significations. Malgré de nombreuses difficultés de définition, il existe quelques points d'accord autour de cette notion, parmi lesquels le fait de voir dans la polysémie [Kleiber, 1999b] :

(i) une pluralité de sens liée à une seule forme

(ii) des sens qui ne paraissent pas totalement disjoints, mais se trouvent unis par tel ou tel rapport.

C'est ainsi que l'on peut dire du mot *bureau* qu'il est polysémique, ayant effectivement plusieurs sens (condition (i)), de meuble à établissement en passant par la pièce, ces derniers étant par ailleurs apparentés (condition (ii)), puisque résultats de métonymies successives (du meuble, à la pièce dans laquelle il se trouve, à l'établissement composé de ce genre de pièces, etc.). Le point essentiel pour caractériser la polysémie est donc l'existence, pour un signe linguistique unique, d'une pluralité de significations ; il s'agit d'une seule unité polysémique, « *où se trouve bel et bien mis en jeu un rapport de non-biunivocité entre forme et sens* » [Fuchs, 1996].

Cette définition permet par ailleurs de distinguer la polysémie de l'homonymie qui, pour sa part, désigne une relation entre deux signes linguistiques *accidentellement* identiques et ayant deux sens différents. On peut ainsi dire des mots *bière* au sens de boisson et *bière* au sens de cercueil qu'ils sont homonymes. Il existe bien une pluralité de sens liée à une seule forme, cette dernière relevant cependant de deux signes linguistiques distincts. Seule (i) est donc de mise pour

[1] [Meur *et al.*, 2004], p.14.

l'homonymie.

La distinction entre homonymie et polysémie se traduit, non sans difficultés, par un traitement lexicographique différencié, avec plusieurs entrées lexicales distinctes pour les cas d'homonymie et des sous-divisions à l'intérieur d'une même entrée pour les cas de polysémies. La frontière entre homonymie et polysémie est cependant bien souvent malaisée à établir, tant les critères de détermination d'un lien véritable entre deux acceptions d'un même mot sont difficiles à établir (voir [Fuchs, 1996]). Quels que soient leur nature et leur nombre[1], il semble bien que la part d'arbitraire soit toujours présente et qu'il faille en conséquence renoncer à vouloir établir des frontières pour au contraire donner la préférence à une analyse en termes de continuum [Jacquet *et al.*, 2005, Victorri et Fuchs, 1996]. Il semble en effet plus révélateur du fonctionnement de la langue de postuler qu'il existe, d'une part, des unités monosémiques et, d'autre part, des unités homonymiques, deux extrêmes entre lesquelles on trouve alors différentes formes (ou degrés) de polysémie, se rapprochant de l'une ou de l'autre en fonction de la parenté plus ou moins forte entre ses différents sens. La polysémie recouvre donc des phénomènes de pluralité sémantique divers, suivant que cette pluralité fait état d'une plus ou moins grande distance sémantique entre ses éléments.

#### 2.3.2.2 Métaphore et métonymie

On considère généralement que la métaphore et la métonymie sont les deux principaux processus à l'origine de la polysémie. La métaphore consiste à employer « *un mot qui désigne habituellement une entité ou un événement d'un certain domaine pour évoquer une entité ou événement qui joue un rôle analogue dans un autre domaine*" [Jacquet *et al.*, 2005]. On peut ainsi dire « Je suis allergique aux manières de cet homme » pour signifier qu'on ne les supporte pas, par transposition d'un terme médical désignant une réaction pathologique excessive dans le domaine des relations sociales où il est alors utilisé pour évoquer un genre similaire de réaction. La métonymie consiste quant à elle à employer un mot attaché à une certaine entité pour en désigner une autre, la seconde étant liée à la première par un rapport fonctionnel ou structurel. La phrase « Il y a trois voiles à l'horizon » évoque ainsi, par usage du nom d'une partie pour désigner le tout, la présence de trois voiliers. Métaphore et métonymie donnent lieu à la création de nombreux sens nouveaux, au départ créations éphémères étroitement dépendantes d'une situation d'énonciation dont certaines, bien souvent, deviennent d'usage courant, sans même que l'on se souvienne du glissement de

[1]Les critères souvent évoqués sont les suivants : étymologie commune, éléments de sens communs et capacité des sujets parlants à considérer un certaine parenté entre diverses significations.

sens dont elles sont issues. On parle de métaphore ou de métonymie vive pour des expressions de type : *Son bureau est un hall de gare* [Jacquet *et al.*, 2005] ou encore *L'omelette au jambon est partie sans payer* [Fauconnier, 1984]. La lexicalisation est au contraire évidente pour des expressions du type *J'ai une montagne de choses à faire*, dans laquelle le mot *montagne* se trouve avoir un sens perçu comme « normal » , recensable par les lexicographes. L'innovation sémantique occasionnée par la métaphore et la métonymie est ainsi plus ou moins « enregistrée » par la langue.

Est-il possible de prévoir ces changements de sens ? Si certaines expressions naissent d'une situation très précise (un client qui a mangé une omelette au jambon), d'autres sont toutefois le résultat d'un processus lexical productif généralisé. Un tel processus correspond au fait de pouvoir engendrer de nouveaux sens à partir de sens premiers de certains mots ou expressions ayant un substrat sémantique commun. Parmi les cadres théoriques élaborés pour analyser et expliquer ce type de phénomène, il est possible de citer, entre autres, les travaux de G. Nunberg avec une analyse en termes de changements de référent puis en termes de changement de prédicat [Nunberg, 1995] et [Kleiber, 1999b, p.128-142], ceux de J. Pustejovsky avec le *lexique génératif* [Pustejovsky, 1995, Kleiber, 1999b], ainsi que ceux de G. Kleiber avec ce qu'il appelle le principe de métonymie intégrée [Kleiber, 1995, Kleiber, 1999b]. Nous souhaitons nous attarder plus précisément sur l'analyse en termes de « facettes » de D.A. Cruse[1] [Cruse, 1996, Croft et Cruse, 2004]. La notion de facette sémantique a pour objet de rendre compte de la variation en contexte de la signification d'un mot, variation ne s'apparentant ni à de l'homonymie, ni à de la polysémie, ni encore à de la simple variation contextuelle. Les exemples suivants[2] donnent à voir une variation de ce type avec le mot *livre* :

*C'est un gros livre avec des illustrations en couleurs.*
*C'est un livre très dense, difficile à comprendre.*

Il est question, dans le premier exemple, d'un objet concret et, dans le second, d'une entité abstraite. Ces deux variations du sens de *livre* sont appelées « facettes » et peuvent recevoir les noms de TOME et TEXTE respectivement. Cruse définit les facettes comme ayant un degré d'autonomie assez élevé (le *contenu* d'un livre n'est pas sa *couverture*) tout en participant d'un concept global unitaire (*contenu* et *couverture* sont liés au concept de *livre*). Le concept de facettes permet d'expliquer comment certains lexèmes, « *tout en ayant un contenu sémantique unitaire ou global, c'est à dire tout en n'étant pas polysémiques, peuvent néanmoins présenter des composants, les* facettes*, qui sont tels qu'il peuvent ap-*

[1] Le concept de *facettes* peut être rapproché de la « structure de *qualia* » de J. Pustejovsky (cf. [Kleiber, 1999b, chap. 7].

[2] Empruntés à [Kleiber, 1999b, chap. 3, p. 87].

*paraître seuls en emploi et donc donner lieu à une variation de sens non polysémique et non simplement contextelle de l'item* » [Kleiber, 1999b]. Relativement au continuum évoqué ci-dessus (cf. section 2.3.2.1), ce type de variation défini par D. A. Cruse est ainsi plus proche de la monosémie que de l'homonymie.

Cette rapide caractérisation de la polysémie lexicale a laissé entrevoir un phénomène majeur d'évolution du lexique. Qu'en est-il de la polysémie des entités nommées ?

### 2.3.3 La polysémie des entités nommées

Deux questions peuvent se poser au regard de cette problématique : peut-on définir les phénomènes de pluralité référentielle des entités nommées en usant des mêmes termes que ceux utilisés pour caractériser la polysémie lexicale d'une part, et comment les traiter de manière automatique d'autre part ?

Pour ce qui est de la définition des phénomènes, un premier point serait de définir globalement les choses. Ainsi, si l'on peut parler d'une pluralité de sens associés à une unité linguistique et ayant un rapport entre eux pour définir la polysémie, il nous semblerait plus adapté de parler d'une pluralité de référents associés à une même entité et ayant un rapport entre eux pour ce que l'on pourrait par conséquent appeler la « polyréférentialité » des entités nommées. Ce rapport privilégié à la référence des expressions linguistiques composant l'ensemble 'entités nommées' sera developpé plus largement dans la partie II. De manière plus précise maintenant, il semble bien que la polysémie des entités nommées présente les mêmes formes de glissement de sens que celle définies pour les unités lexicales « classiques ». On peut en effet observer un continuum de changements de référents ayant pour extrémités deux référents pour deux entités d'une part (homonymie) et deux référents associés à une même entité d'autre part (polysémie) et présentant des dégrés variables de « polyréférentialité ». Pour des entités nommées comme *Vienne*, désignant une ville en France ou une ville en Autriche, ou encore comme *Orange*, renvoyant à une ville ou à une entreprise, on peut penser à de l'homonymie. De la même manière, les glissements de sens entre le nom d'une entreprise (*Renault*) et le nom de ses produits (*une Renault*) ou le nom d'un pays pour son gouvernement semblent bien relever de la métonymie. Autre phénomène, non mentionné jusqu'à présent mais qu'il serait utile de pouvoir traiter : celui de Jacques Chirac en tant que Maire de Paris ou en tant que Président de la République, que l'on pourrait dès lors analyser en termes de « facettes », selon l'analyse proposée par D.A. Cruse. Et la métaphore ? Il semble que ce mode de changement de sens soit peu présent pour les entités nommées[1]. Ainsi, les phéno-

[1]On peut penser à des expressions du type « *Les Champs Elysées de Séoul* » ; cependant, la

mènes de polysémie paraissent bien être de mise pour les entités nommées et il semblerait utile de les caractériser plus avant.

Qui dit pluralité de sens dit désambiguïsation. Au niveau lexical, le processus de désambiguïsation correspond à la sélection, pour un mot donné, d'un sens particulier en fonction de son contexte d'apparition. Cette tâche de désambiguïsation lexicale (ou *Word Sense Disambiguation*) est bien connue et largement traitée par le TAL depuis de nombreuses années. S'agissant de la désambiguïsation des entités nommées, la question ne semble pas avoir été autant abordée. Il importe désormais, au vu des phénomènes de polysémie présents pour les entités nommées, d'affiner les processus et de s'attacher à reconnaître, pour une entité donnée, le référent particulier qu'elle désigne, et ce toujours en fonction du contexte d'apparition. Désambiguïsation lexicale et désambiguïsation des entités nommées sont-elles des tâches similaires ou la seconde est-elle spécifique par rapport à la première ? Nous tenterons d'apporter des éléments de réponse à ce sujet dans le chapitre 5.

La considération des entités nommées « dans la langue » donne ainsi à voir une pluralité de phénomènes qu'il serait intéressant de caractériser plus avant et de prendre en charge lors de traitements automatiques. Au final, difficultés de catégorisation, difficultés d'annotation et difficultés de « polysémies » font en quelque sorte figure de contrepoint au succès souligné lors du chapitre 1.

## 2.4 Un héritage MUC à dépasser

Objet de nombreuses attentions depuis plus d'une décennie, les entités nommées sont aussi, manifestement, des objets difficiles à cerner et à traiter. La tâche de reconnaissance de ces unités, si encourageante soit-elle pour le TAL, doit en effet faire face à diverses difficultés, ci-dessus détaillées comme relevant de perspectives « monde », « texte » ou « langue »[1]. Du choix des catégories aux divers phénomènes de sens en passant par leurs multiples réalisations lexicales possibles, les entités nommées ne se laissent de toute évidence pas facilement annoter. Ces difficultés sont naturellement interdépendantes, les « polysémies » invitant par exemple à repenser les catégories et les diverses mentions à mieux définir ce que recouvrent ces dernières. Partant, le problème sous-jacent et peut-être primordial semble être celui de la caractérisation précise d'un objet échappant incontestablement à toute tentative de limitation à une catégorie, une mention, une inter-

métaphore ne semble pas donner lieu à des glissements de sens réguliers au regard des entités nommées.

[1]Ce dernier volet apparaît au final comme la mise en relation des deux premiers.

prétation. Comment dès lors aller au devant de ces difficultés ? Deux perspectives de recherche semblent se dégager.

### 2.4.1 Un problème de définition

Catégorisation, annotation ou interprétation, toutes ces difficultés d'appréhension des entités nommées ne relèveraient-elles pas en dernier ressort d'une difficulté de compréhension, autrement dit d'un problème de définition de ces unités ? Risquons nous tout d'abord à un retour en arrière : au terme de cette première partie se voulant un état des lieux de la tâche de reconnaissance des entités nommées, quelle définition, au sens d'un ensemble de traits permettant de circonscrire un objet, est-on en mesure de dégager à l'encontre de ces unités ? Hormis la formule liminaire utilisée en introduction de cette partie (p.11), soit « *un certain nombre d'unités lexicales particulières, que sont les noms de personnes, les noms d'organisation et les noms de lieux, ensemble auquel sont souvent ajoutés d'autres syntagmes comme les dates, les unités monétaires et les pourcentages* », il n'est guère possible d'en dire plus. On connaît leur origine, on sait les reconnaître, mais en proposer une définition semble plus difficile. Cette « auto-citation », bien loin d'être présomptueuse, se veut au contraire l'illustration d'un phénomène somme toute récurrent dans nombre de travaux sur les entités nommées, à savoir la définition de l'objet étudié sous la forme d'une énumération des éléments pouvant être considérés comme tels. Sans s'attarder pour le moment à inventorier les différentes formules définitoires augurant tout article rendant compte de travaux sur ces unités (cf. chapitre 3 à venir), il peut être intéressant de s'interroger sur leur caractère « énumératoire ». Issue de l'extraction d'information, la tâche de reconnaissance des entités nommées semble en effet avoir été conçue selon une démarche extrinsèque, nous pourrions presque dire « top-down », s'intéressant exclusivement à décider des choses du monde à reconnaître, sans véritablement se soucier de leur concrétisation textuelle en tant qu'objets de la langue. Ce dernier point mérite quelque nuance et explications : les diverses directives d'annotation, pour certaines fort prolixes, dont il a été abondamment question dans cette section peuvent être considérées, précisément, comme une tentative de s'intéresser aux entités nommées en tant qu'objet de la langue. Ces directives cependant ne font que lister des exemples de bonnes annotations et mentionner les erreurs à ne pas faire, sans définir à proprement parler la notion d'entité nommée. En définitive, dans ces formules définitoires énumératives il est possible de retrouver un lointain écho du *template* ou formulaire définissant des champs à remplir. Pragmatique, justifiée par les origines de la tâche, cette démarche n'est aucunement à remettre en cause. Toutefois, ce mode énumératif est-il apte à rendre pleine-

ment compte de la notion d'entité nommée ? N'est-il une certaine dissonance à faire cohabiter dans un même ensemble divers éléments s'apparentant à des noms propres pour les uns, des appellations et des dates, entre autres, pour les autres, et ce sans aucune ébauche de dénominateur commun ? D'autres semblent souligner pareillement un manque de caractérisation, reconnaissant que « *la notion d'entité nommée est un concept riche et complexe* », qu' « *il n'existe pas de définition standard* » [Meur *et al.*, 2004] ou indiquant encore que « *la définition de ce qu'est une entité nommée est ambigu* », « *n'est pas simple* » [Sekine *et al.*, 2002]. De la sorte, ce mode d'appréhension des entités nommées, à dominante pragmatique, gagnerait sans doute à être complété d'une réflexion davantage théorique, à dominante linguistique cette fois-ci, compte tenu du matériau de départ et des difficultés pointées ci-avant. Si les entités nommées, objet plus complexe qu'il n'y paraît, invitent ainsi à une considération plus avant de leurs caractéristiques propres, elles conduisent également à penser leur traitement en d'autres termes.

### 2.4.2 Vers de nouveaux traitements

Au regard du traitement des entités nommées, l'enjeu principal a été jusqu'à présent de les reconnaître et de les catégoriser selon des ensembles « conceptuels » relativement larges. Nous l'avons vu, cette catégorisation sémantique est délicate à déterminer, pouvant être d'une couverture plus ou moins large et d'un niveau de granularité plus ou moins fin. Par ailleurs, est-elle adaptée à rendre compte des phénomènes de sens décrits ci-dessus ? La polysémie constatée des entités nommées semble en effet appeler à abandonner l'approche catégorisante unilatérale pour préférer une annotation à caractère modulaire permettant une caractérisation plus fine et plus complète du référent dénoté par l'entité en contexte.

Depuis son apparition au milieu des années 1990 jusqu'à aujourd'hui, la tâche de reconnaissance des entités nommées a ainsi connu une progression étonnante, provoquant l'enthousiasme, suscitant de nombreux travaux de recherches, arrivant à être de toutes les conférences ainsi qu'à s'immiscer dans de nombreuses applications. Définie à l'origine lors des campagnes d'évaluation MUC, elle a gardé l'empreinte de ces conférences, qu'il s'agisse de sa définition, rappelant lointainement la tâche d'extraction d'information, ou de son traitement, via des catégories sémantiques larges et prédéfinies. Une observation plus attentive de cette tâche a néanmoins fait apparaître des difficultés de tous ordres, appelées à être comblées de manière complémentaire par, d'une part, un effort théorique permettant de mieux définir la notion d'entité nommée et, d'autre part, un effort « pratique » donnant la possibilité d'annoter au mieux ces unités. Bien loin de décrier l'héritage des MUC, l'enjeu est donc au contraire de le dépasser ; les parties sui-

vantes de ce mémoire tenteront chacune d'apporter une pierre à l'édifice, avec une réflexion sur la définition des entités nommées (partie II) d'une part, et la proposition de méthodes pour un traitement plus fin de ces entités (partie III) d'autre part.

# Deuxième partie

# Les Entités Nommées : une création TAL

# Introduction

La partie précédente a permis de dégager deux faits essentiels au regard de la problématique des entités nommées : si la tâche de reconnaissance de ces unités apparaît comme un incontestable succès pour le TAL ayant, durant cette dernière décennie, suscité un intérêt soutenu et acquis une maturité technologique certaine, sa mise en œuvre n'en comporte pas moins un certain nombre de difficultés, souvent ignorées, lesquelles s'avèrent d'autant plus importantes, sinon à résoudre, au moins à débrouiller, que de nouvelles perspectives de recherche se font jour. Les divers écueils soulignés à l'endroit des entités nommées, relevant de difficultés de catégorisation, d'annotation, de représentation ou encore de considération de phénomènes de sens, renvoient tous d'une manière ou d'une autre à la difficulté d'appréhender la notion d'entité nommée. C'est pourquoi, avant d'aller de l'avant et proposer de nouveaux traitements, il apparaît nécessaire de tenter un effort théorique relativement à ces unités, une meilleure appréhension de l'objet d'étude étant susceptible, croyons-nous, d'aider à son traitement.

Nouvelle venue dans le paysage TAL, aux côtés par exemple de la traduction automatique ou de l'analyse syntaxique, la tâche de reconnaissance des entités nommées semble en effet avoir pâti de son apparition rapide[1], laquelle ne lui a pas donné l'occasion de se doter d'une véritable assise théorique. S'il est relativement simple de définir de manière intuitive une entité nommée, il apparaît cependant plus ardu de le faire de manière rigoureuse et systématique, compte tenu de la diversité et de l'étendue des unités pouvant faire partie de cet ensemble. Ensemble hétérogène s'il en est, les entités nommées semblent constituer l'un des premiers « retours à l'envoyeur » du TAL vis-à-vis de la théorie linguistique, amenée aujourd'hui à considérer un objet qu'elle n'avait nullement défini auparavant. Partant, cette partie a pour objet de s'essayer à une « définition » des entités nommées ; pour ce faire, nous tenterons de dégager certains critères et mécanismes distinctifs des entités nommées et de proposer, modestement, quelques pistes pouvant servir à l'élaboration d'une définition des entités nommées.

Cette partie s'organise en trois temps. Le premier, le plus court, présente en

[1] Rapide et pragmatiquement motivée par la tâche d'extraction d'information.

guise de préambule une « proposition méthodologique » examinant les tenants et aboutissants d'une telle démarche et précisant l'objectif poursuivi. Le second explore quelques pistes linguistiques existantes, s'intéressant aux noms propres d'une part, aux descriptions définies d'autre part. Le dernier, enfin, prend acte du précédent et avance une proposition de définition des entités nommées.

# Chapitre 3

# Préambule méthodologique : Comment définir les entités nommées ?

Se soucier d'une dimension théorique à l'endroit des entités nommées pourrait ne pas apparaître comme primordial aux yeux d'une certaine conception du TAL voyant dans la raison d'être de cette discipline d'exclusifs besoins applicatifs. C'est pourquoi cette proposition liminaire vise à donner quelques précisions quant à la démarche entreprise ici de définition « théorique » des entités nommées. Comment procéder et sur quels éléments prendre appui ? Il importera avant tout d'examiner les tenants et aboutissants d'une telle démarche et de montrer en quoi elle participe d'une des tensions spécifiques, voire constitutives, de la discipline qu'est le TAL. Suivront par la suite un examen des pratiques définitoires ainsi qu'une exploration de l' « existant », dans les discours comme dans les pratiques, pouvant servir de marchepied à une définition des entités nommées. Enfin, une fois précisés ces éléments, il conviendra de faire le point sur la méthode d'investigation adoptée et l'objectif poursuivi.

## 3.1 Une démarche propre au TAL

Le TAL[1] a ceci de particulier qu'il s'agit d'un champ d'étude relativement récent (moins de cinquante ans) et qui peine, depuis ses débuts, à se définir, ou se faire reconnaître, comme une véritable discipline scientifique. Sans entrer plus avant dans ce débat, il peut être utile cependant, dans le sillage de l'étude sur la constitution du TAL par M. Cori et J. Léon [Cori et Léon, 2002], de considérer cette confusion identitaire comme le résultat de deux lignes de tensions constantes

[1]Dénomination qui a son histoire, cf. article de M. Cori et J. Léon [Cori et Léon, 2002]

entre, d'une part, « *la cohabitation paradoxale et nécessaire des recherches théoriques et des applications à visée industrielle* » et, d'autre part, « *le* TAL *et les différentes disciplines qui le constituent* ». Le problème semble donc être l'impossible unité et des objectifs d'une part, et des moyens d'autre part, le TAL hésitant entre coiffer la casquette d'ingénieur ou celle de « chercheur », se trouvant qui plus est à la croisée de diverses disciplines, au premier rang desquelles l'informatique, la linguistique, les mathématiques et l'intelligence artificielle. Champ d'investigation au statut et à la réalité sans cesse mouvants, il est un pas facile, que M. Cori et J. Léon hésitent à franchir mais mentionnent tout de même : celui de se résigner « *à l'inanité d'une impossible quête, celle de définir un champ unifié qui, tout en englobant les applications industrielles, soit scientifiquement fondé* » [Cori et Léon, 2002]. Quoi qu'il en soit, qu'on le considère de tel ou tel point de vue, le TAL a une réalité indéniable et il importe d'exploiter la diversité constitutive de ce champ d'étude pour lui donner, sinon le statut de discipline scientifique autonome, au moins une réalité réfléchie.

C'est ce à quoi s'attellent des ouvrages comme *Sémantique et TALN*, dirigé par P. Enjalbert [Enjalbert, 2005b], ou encore *Ingéniérie des Langues*, dirigé par J.-M. Pierrel [Pierrel, 2000]. Ces publications ont pour objectif, au-delà de l'exposé de travaux et recherches divers, de questionner le lien entre la théorie et la pratique de différentes disciplines, ici la linguistique et l'informatique principalement, ces dernières étant amenées à collaborer en des aller-retours qu'il convient d'encourager. Le titre du premier ouvrage, coordonnant un des principaux champs d'étude de la linguistique théorique avec le TAL, est à cet égard révélateur ; P. Enjalbert en déroule par la suite le programme, expliquant que « *si* [les] *réalisations* [technologiques] *posent de réels et souvent passionnants problèmes informatiques ou d'ingéniérie de la connaissance, la dimension linguistique doit aujourd'hui être pleinement reconnue. Et si le recours à l'intuition du concepteur d'applications n'est certes pas à proscrire, il n'est pas possible de se replier sur une sorte de « sémantique naïve », partagée « naturellement » par tout locuteur : la langue est une affaire trop complexe pour cela. Ici comme ailleurs, le « détour » par la théorie est nécessaire et la sémantique linguistique est riche de modèles qui ne peuvent être ignorés* ». Adoptant pour sa part un point de vue d' « ingénieur », J.-M. Pierrel, voyant comme finalité première du TAL « *la mise en place d'applications concrètes*, ne manque cependant pas de souligner la possibilité de « *confronter la linguistique, qui pendant longtemps demeura descriptive, à des exigences d'opérationnalité, ou plus précisément ses modèles aux exigences opératoires de la modélisation informatique* ». Dans un sens, la linguistique apportant à l'informatique, ou dans l'autre, l'informatique apportant à la linguistique, cette dialectique apparaît donc comme essentielle. Elle n'est bien sûr pas nouvelle et

bien des réalisations de traitement automatique du langage ont eu à s'inspirer de théories ou modèles (on peut par exemple penser aux apports de la logique ou des mathématiques). Ainsi, caractéristique voire constitutif du TAL, ce dialogue entre dimension théorique et exigences opératoires ne semble toutefois pas avoir eu lieu s'agissant des entités nommées. S'inscrivant dans cette tradition de pluridisciplinarité, c'est ce à quoi cette partie souhaite s'atteler, se penchant sur la question de la définition des entités nommées.

## 3.2 Objectif définition : examen des pratiques définitoires

Objectif définition donc, mais quel genre de définition ? Si l'idée dominante est celle « *d'une limite ou d'un ensemble de traits qui circonscrivent un objet*[1] », en quels termes faut-il réfléchir les entités nommées ? Comment s'y prendre pour définir un objet TAL ? Les entités nommées sont de toute évidence nées des besoins du traitement automatique de la langue et gravitent dans le monde, au sens large, de l'extraction d'information. Cette filiation n'est cependant pas unique, dans la mesure où leur matérialité fait qu'elles relèvent également d'un monde linguistique. Définir les entités nommées implique par conséquent que l'on prenne en compte ces deux dimensions, traitement automatique du langage et linguistique, en une interdisciplinarité, nous l'avons vu, propre au TAL. Ceci étant établi, comment procéder ? Faut-il reprendre des démarches définitoires présentes dans chacune des deux disciplines ? Une définition purement linguistique aurait-elle une utilité pour la mise en œuvre de traitements automatiques ?

Regardons tout d'abord du côté de la linguistique : dictionnaires, grammaires et théories linguistiques ont pour pratique, majoritairement, d'énoncer des règles chaque fois que cela est possible, selon un esprit peu ou prou normatif[2]. Pour considérer un phénomène linguistique, quel qu'il soit, il existe deux démarches possibles, onomasiologique ou sémasiologique. La première part du contenu pour aller vers la forme du contenu : on s'intéresse d'abord à l'idée, au concept, puis on examine les signes linguistiques correspondant à ce découpage conceptuel. C'est ainsi que le signe *bateau* sera, selon cette démarche, tout d'abord étudié en fonction des relations qu'il entretient avec les autres éléments de la taxonomie des moyens de transports (par exemple), avant d'être considéré du point de vue de son fonctionnement linguistique, avec ses distributions (axe syntagmatique), ses

[1] *définition* dans le TLFi

[2] Cette affirmation est sans doute un peu trop « normative », il existe, aux côtés de discours presciptifs, de nombreux discours descriptifs. Elle a pour objectif cependant de tracer une ligne de démarcation par rapport aux pratiques du TAL, c'est pourquoi ce point est mis en avant.

oppositions (axe paradigmatique) et ses polysémies. Cet exemple concerne une unité lexicale, mais une démarche similaire est possible pour un phénomène ou une catégorie linguistique : définir le *passé* consisterait alors à le situer au sein de concepts et de relations décrivant la dimension temporelle, avant de s'intéresser à ses diverses réalisations sous forme de signes linguistiques (les terminaisons de la conjugaison). À l'inverse, la démarche sémasiologique correspond à une étude qui part du signe pour aller vers la détermination du concept : une unité lexicale est ainsi appréhendée au niveau de son comportement en surface avant d'être référée à un concept donné. Ainsi, deux types de questionnements peuvent aider à la détermination d'un phénomène linguistique ; c'est ce que réalise en substance M.N. Gary-Prieur lorsque, dans un article interrogeant l'existence du nom propre en tant que catégorie linguistique [Gary-Prieur, 1991], elle examine d'une part la perspective sémantique (sémasiologie) puis la perspective syntaxique (onomasiologie).

Observons à présent les pratiques définitoires du côté du TAL. En tant qu' « *application de programmes et techniques informatiques à tous les aspects du langage humain*[1] », le TAL apparaît davantage concerné par des processus que par des objets proprement dits. La liste des vocables figurant dans l'Atalapédie reflète bien cet état de chose, la plupart des termes renvoyant effectivement à des applications (*traduction automatique*, *recherche d'information*, etc.) ou à des processus (*étiquetage morphosyntaxique*, *alignement phrastique*, etc.) bien plus qu'à des objets manipulés. Ces derniers ne sont toutefois pas absents, mais il s'agit pour une grande majorité de termes « importés » d'autres disciplines, n'étant pas à proprement parler des « objets TAL ». Les notions de *corpus*, de *lemme*, d'*arbre* ou encore de *requête*, si leurs acceptions sont certes remaniées en fonction du cadre précis qu'est le TAL, sont en effet directement issues de la linguistique et de l'informatique. Il est ainsi difficile de trouver des objets propres au TAL, et donc d'examiner les façons dont ils sont définis. Cette pratique semble par ailleurs moins essentielle qu'en linguistique, l'*Atalapédie* n'ayant été lancée que depuis 2005, au sein d'une association, l'Atala, qui existe pourtant depuis 1959. Cependant, ce « désintérêt » révèle en filigrane un des principes de fonctionnement du TAL : le pragmatisme, ou *l'intuition du concepteur d'applications* évoquée plus haut par P. Enjalbert. Pas de « pratique définitoire » donc du côté du TAL, mais de la pratique tout court. Il conviendra donc, pour définir les entités nommées, de prendre en compte cette dimension.

À la suite de ces précisions sur les pratiques définitoires présentes en linguistique et en TAL, il convient d'examiner ce qu'il est possible d'utiliser comme point de départ à une définition des entités nommées, en explorant ce qui se dit, puis

[1]Entrée *Traitement automatique des langues* de l'Atalapédie.

ce qui se fait.

## 3.3 Exploration

### 3.3.1 Les formules définitoires existantes

La question de la définition des entités nommées a déjà été soulevée en 2.4, avec la mise en évidence de l'absence d'un véritable discours sur la nature de ces dernières. Si la plupart des gloses sur les entités nommées s'assimilent en effet à des formules définitoires énumératives, il existe tout de même quelques propos s'engageant plus avant dans une caractérisation de ces unités. Ce paragraphe a pour objet de les examiner ; pour ce faire, nous nous appuierons sur un inventaire (non exhaustif) des propos définitoires plus ou moins détaillés recueillis dans divers rapports de campagnes d'évaluation et de projets, ou encore dans des « encyclopédies » (cet inventaire figure au complet en annexe B). Relevant de travaux relativement différents et présentées ici hors de leurs contextes, il convient de ne pas « surinterpréter » les citations qui vont suivre, l'intention sous-jacente à leur recensement étant de permettre de se faire une idée du discours général sur les entités nommées.

Les différentes formules cherchant à rendre compte de manière plus ou moins précise de la notion d'entité nommée se répartissent en diverses « tendances », suivant le point de vue adopté. Ces dernières sont au nombre de trois, la première adoptant plutôt une position sémasiologique et les deux autres une position onomasiologique. Commençons par la première : dans la lignée de la définition de la tâche par les conférences MUC, on rencontre des propos adoptant un point de vue « extraction d'information ». Souvenons-nous tout d'abord des instructions de N. Chinchor à propos de MUC-7 [Chinchor, 1997], dans une vue d'ensemble de la campagne tout d'abord :

> « *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* ».

puis dans les directives d'annotation propres à la tâche ensuite :

> « *The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are « unique indentifiers » of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, per-*

*centages)* ».

Si le flottement de la portée de l'appellation est bien là, la caractérisation des entités nommées se fait bien ici sur un mode catégoriel. Les campagnes CoNLL reprennent le même genre de formule :

> « *Named entities are phrases that contain the names of persons, organizations and locations* » .[TjongKimSang et Meulder, 2003]

tout comme T. Poibeau dans son ouvrage sur l'extraction d'information :

> « *On appelle traditionnellement » entités nommées » (de l'anglais named entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérables par les mêmes techniques à base de grammaires locales. Seules les entités nommées au sens propre du terme seront abordées dans ce chapitre (...)* ». [Poibeau, 2003]

Il serait possible de multiplier à l'envi ce genre de formules qui caractérisent les entités nommées par les diverses catégories sémantiques auxquelles elles peuvent appartenir. Ce qui domine nettement ici est l'aspect « informationnel » de ces unités, l'intérêt que l'on peut avoir à les connaître, relativement au contexte dans lequel elles sont énoncées. C'est ce que résume au final la définition du *National Institute of Standards and Technology*[1] :

> *Named Entity : a named object of interest such as a person, organization, or location.*

description dont on retrouve l'esprit, en plus développé, chez S. Sekine :

> « *The names of particular things or classes, and numeric expressions is regarded as an important component technology for many NLP applications.(...) In this paper, the term* Named Entity *includes names (which is the narrow sense of Named Entity) and numeric expressions. The definition of this Named Entity is not simple, but, intuitively, this is a class that people are often willing to know in newspaper articles.* » [Sekine *et al.*, 2002]

La perception des entités nommées, telle qu'elle apparaît ici au travers de ces formules, semble donc privilégier leur aspect informationnel, les considérant avant tout comme des noms d'éléments ayant une quelconque importance relativement à un contexte donné, fût-il général.

[1] http ://www.itl.nist.gov/iad/894.02/related_projects/muc/info/definitions.html

Il est une autre vision des choses, issue de travaux s'intéressant exclusivement aux noms propres. C'est celle de N. Friburger :

> « *En fait il semble difficile des délimiter les noms propres des autres noms ; il y a une continuité entre l'ensemble des noms propres et l'ensemble des noms communs. Les informaticiens qui travaillent dans le domaine de l'extraction d'information, ont abordé le problème de manière pragmatique. Ils ont défini la notion d'entités nommées pour regrouper tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantités* ». [Friburger, 2002]

à laquelle fait brièvement écho M. Tran :

> « *Avec MUC-6, les noms propres, ainsi que les dates et les unités chiffrées sont regroupées sous le terme d'entités nommées* ». [Tran, 2006]

La notion d'entité nommée apparaît ici comme un regroupement pratiqué par le TAL, par opposition à la catégorie linguistique du nom propre, comme si, bizarrement, il avait fallu inventer autre chose. Les acteurs de projets de reconnaissance automatique des noms propres opèrent ainsi une partition relativement exclusive entre un objet de nature linguistique, le nom propre, et un autre appartenant au TAL, les entités nommées, faisant office de chapeau à toutes les unités lexicales relevant de près, et surtout de loin, de la catégorie des noms propres.

Enfin, une troisième « tendance », à dominante linguistique cette fois-ci, semble se dessiner. Il n'est plus question d'opposition vis-à-vis du nom propre mais de caractérisation des entités nommées à l'aide de notions linguistiques, en plus des seules catégories sémantiques fondamentales dans la vision « extraction d'information ».

Examinons les propos suivants :

> ESTER : « *Même s'il n'existe pas de définition standard, on peut dire que les EN sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)* ». [Meur *et al.*, 2004]

> P. Enjalbert : « [Le] *repérage et* [l']*étiquetage sémantique des entités nommées : il faut détecter toutes les formes linguistiques qui, à l'instar des noms propres, désignent de manière univoque une entité par leur pouvoir de sélectivité : noms de personnes, d'institutions et*

> *entreprises, de lieux, ainsi souvent que les dates, unités monétaires etc. Il faut aussi leur affecter une étiquette sémantique choisie parmi une liste prédéfinie ».* [Enjalbert, 2005a]

> VICENTE : « Entité Nommée *est la notion utilisée en TAL pour désigner les éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres (note : la notion d'entité nommée concerne les noms propres mais aussi les dates et les mesures) et qui suivent des patrons syntaxiques déterminés ».* [Vicente, 2005]

On s'éloigne progressivement des simples formules énumératives, avec la considération de l'aspect « référentiel » des entités nommées, également décrites comme des formes linguistiques au fort « pouvoir de sélectivité » ou encore comme des « éléments discursifs monoréférentiels ». Si l'aspect extraction d'information n'est pas loin (« *une entité du monde concret dans certains domaines spécifiques* » ), il n'empêche, cet ensemble n'est plus seulement décrit par le biais de seules catégories sémantiques mais se trouve également caractérisé au moyen de notions relevant davantage de la discipline linguistique. Cette appréhension plus complète de l'objet entité nommée se retrouve pareillement dans les définitions proposées par l'Atalapédie et Wikipédia :

> ATALAPÉDIE[1] :
>
> « *Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprises, etc. contenues dans un texte. On ajoute souvent à ces éléments les dates et d'autres données chiffrées. Par extension, les entités désignent parfois les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologie). (...) Ces séquences référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes* ».

> WIKIPÉDIA[2] :
>
> « *Named entity recognition (NER) (also known as entity identification (EI) and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, etc. (...) In the expression named entity, the word named restricts the task to*

[1] http ://www.atala.org/AtalaPedie/index.php ?title=Accueil
[2] http ://en.wikipedia.org/wiki/Main_Page

> *those entities for which one or many rigid designators, as defined by Kripke, stands for the referent ».*

Aux côtés d'une liste de catégories, liste ouverte s'achevant systématiquement par « etc. », figurent ici, entre autres, l'évocation de la notion de « désignateur rigide » de S. Kripke ainsi que la paraphrase « séquences référentielles », autant d'indications permettant de rendre compte d'une manière plus poussée de la notion d'entité nommée. Ces trois « tendances » ainsi présentées donnent à voir les différentes manières d'appréhender la notion d'entité nommée ; elles sont bien sûr perméables.

Que retenir de tout cela ? Tout d'abord le fait que la manière d'évoquer les entités nommées a évolué, slalomant entre différentes manières de voir les choses, selon un point de vue extraction d'information ou un point de vue davantage soucieux de la dimension linguistique, passant ainsi d'une description « de l'extérieur » à davantage de considération de propriétés propres à l'objet entité nommée. Il importe également de relever le fait que ces propos définitoires s'arment bien souvent de précautions verbales quant à leur contenu : [Meur *et al.*, 2004] insiste sur le fait qu' « *il n'existe pas de définition standard* », [Sekine *et al.*, 2002] souligne que « *the definition of this Named Entity is not simple* » et [Chinchor, 1997] use de guillemets lorsqu'elle assimile les entités nommées à des « *unique identifiers* ». Tout se passe comme si les mots n'étaient jamais les bons pour caractériser les entités nommées. Remarquons en outre que ces « énoncés définitoires », bien que se trouvant au sein d'une liste que nous considérons comme non exhaustive, sont bien peu nombreux relativement à la quantité de travaux et projets réalisés sur les entités nommées depuis l'apparition de cette tâche. Enfin, si chacune de ces formules contient des indices intéressants quant à la définition des entités nommées, il apparaît nécessaire de les organiser, les approfondir et les justifier. Ayant ainsi entendu « ce qu'il se dit », il convient d'examiner « ce qu'il se fait ».

### 3.3.2 Les réalisations

Si notre objectif est de définir les entités nommées, ou d'en proposer des éléments de définition, il est nécessaire de déterminer ce que l'on cherche à définir. Noms de personnes pour les uns, séquences référentielles pour les autres, la meilleure façon de décider de l'objet à définir est de constituer un échantillon typique en partant des entités nommées reconnues dans différents systèmes. De nombreux exemples ont déjà été donnés tout au long de la première partie afin d'illustrer la présentation de la tâche et d'exposer les difficultés. S'il n'est donc pas nécessaire de s'étendre trop longuement ici, il est néanmoins primordial de « mettre à plat » ce qui, dans les faits, est considéré comme une entité nommée.

Il est possible de commencer par retracer le flottement autour de l'appellation « entité nommée », déjà évoqué brièvement lors de l'analyse des difficultés de catégorisation en 2.1.2.1. Il existe en effet une certaine incertitude quant à la portée de cette dénomination. Désignant à l'origine les seuls noms propres de personne, de lieu et d'organisation rassemblés au sein de la catégorie ENAMEX définie par les organisateurs des conférences MUC (cf. 1.2.2), l'appellation « entité nommée » s'étend bien vite aux autres grandes classes des expressions numériques (NUMEX) et temporelles (TIMEX), cette dérivation ayant lieu au sein même des participants aux premières conférences. Par la suite, ce nom s'étend, nous l'avons vu, à d'autres catégories (noms de produits — armes, véhicules, artefact, etc. — et noms d'œuvre entre autres), lesquelles rassemblent des noms relativement éloignés de la catégorie (linguistique) stricte du nom propre (stricte au sens de nom propre prototypique). Ainsi, débordée de l'intérieur avec l'extension aux autres classes NUMEX et TIMEX, l'appellation originelle de MUC l'est également de l'extérieur avec la considération d'autres unités. Si le premier mouvement ne prête peut-être pas à conséquence et relève sans doute d'un abus de langage amalgamant les entités nommées « véritables » et les autres classes ajoutées vraisemblablement en raison de leur facilité de traitement, le second témoigne quant à lui d'un incontestable élargissement de la notion d'entité nommée.

Examinons de plus près en quoi consistent exactement ces débordements, avec les expressions linguistiques retenues comme entités nommées dans divers systèmes de reconnaissance. Il y a bien sûr les grands classiques, reconnus par presque tous les systèmes : *Lionel Jospin* (système de reconnaissance des entités nommées pour le français développé au sein de XRCE, [Rebotier, 2006]), *M. Saddam Hussein* (système TagEN développé par Jean-François Berroyer et T. Poibeau[1]) ou encore *Eugen Bleuler* (système YooName[2]). Dans la même veine traditionnelle figurent aussi la *Central Intelligence Agency* (YooName), le *Koweit* (TagEN) et encore *Paris* ([Rebotier, 2006]). Sont reconnues également des unités telles que : *le Festival du film de Berlin*, *la gare Montparnasse*, *l'orchestre philarmonique de New York*, le *Président de la République* et *la ligne Maginot* pour Nemesis [Daille *et al.*, 2000, Fourour, 2002], *lysozyme* et *immature CD34 + Thy-1+ subset* pour le système ABNER [Settles, 2005] ou encore *schizophrénie*, *Advil* et *Tour de France* pour YooName. Certains systèmes s'intéressent aussi à des entités telles que *Compaq Presario 12XL300*, *Intel Pentium III Processor*, *Lotus SmartSuite Millennium license* ou *20GB* (module de reconnaissance d'entités nommées dans le projet Crossmarc, [Karkaletsis *et al.*, 2003, D.Farmakiotou *et al.*, 2002]). De même, les systèmes réalisés durant la campagne HAREM [Santos *et al.*, 2006]

[1] http://www-lipn.univ-paris13.fr/~poibeau/tagen.html
[2] http://www.yooname.com/Example.html

reconnaisssent le *Ministère de l'Environnement*, le *Nazisme* et le *Décret de loi numéro 31/3 de 2005*. Mentionnons encore *le peloton voltigeur motocycliste (PVM)* [Daille *et al.*, 2000], la *sociologie* [Santos *et al.*, 2006], *le comité de grève de Nowa Huta* [Rebotier, 2006], le *Prix Nobel* (YooName) et bien sûr (le) *21 avril 2002* et *2007*. Cette énumération pourrait se continuer à l'infini (ou presque), point n'est donc besoin d'aller plus avant, l'essentiel est là : nous avons affaire à un ensemble disparate. Un échantillon « représentatif »[1] de ce qui est considéré comme entité nommée peut ainsi être représenté comme dans la figure 3.1.

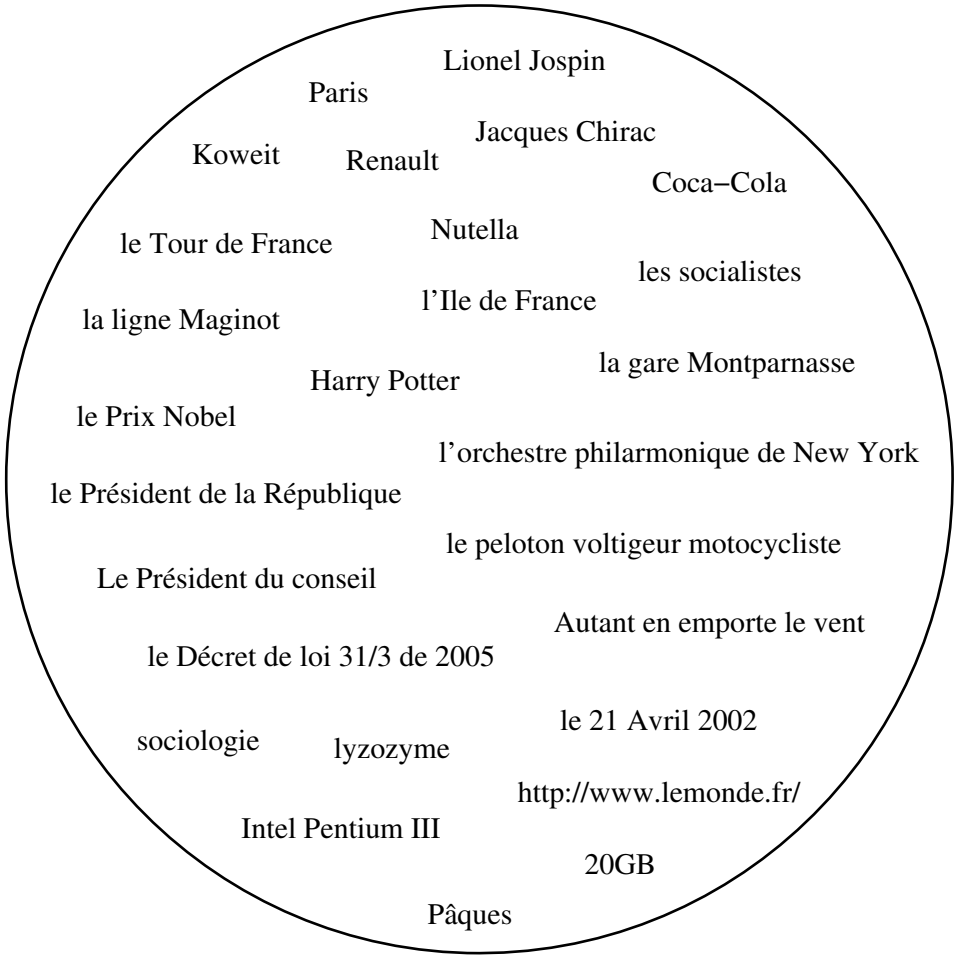



FIG. 3.1 – Echantillon d'unités lexicales considérées comme des entités nommées.

Qu'observe-t-on ? Il y a bien sûr des noms propres, mais aussi d'autres unités que l'on hésite à qualifier de noms propres, tout comme des descriptions définies. Quel peut être le dénominateur commun de ces items ? Comment comprendre et organiser cet ensemble ? C'est ce à quoi tenteront de répondre les chapitres à venir, selon la démarche suivante.

[1]Cet échantillon n'est pas *représentatif* au sens statistique du terme, c'est pourquoi le mot est ici entre guillemets. Il est par la suite qualifié de « typique ».

## 3.4 Notre démarche

Quel type de définition cherchons-nous ? Quel est l'objectif poursuivi ? Comment procéder ? Les observations précédentes ont permis de dégager quelques points significatifs de la problématique des entités nommées, sur lesquels il est possible de s'appuyer pour organiser une démarche de définition de ces unités.

L'examen des pratiques définitoires des champs disciplinaires impliqués dans la définition des entités nommées est plus riche d'avertissements que d'indications précises. En effet, si quelques démarches se dessinent du côté de la discipline historique, la dimension applicative, primordiale en TAL, invite à ne pas rester dans un bocal linguistique. C'est pourquoi il importera de proposer non pas une définition normative comme il est possible d'en voir dans les grammaires mais quelque chose de plus souple, sous la forme de critères d'identification, de "balises d'appréhension" de l'objet entité nommée. Non pas une définition d'apparat mais un cadre théorique qui prenne en compte la réalité linguistique des entités nommées tout comme les besoins réels de la discipline TAL et qui, au final, permette de limiter les hésitations ainsi que de résoudre les problèmes qui se posent lorsqu'il est question du traitement automatique des entités nommées.

L'exploration des formules définitoires existantes et des réalisations a quant à elle montré une grande hétérogénéité. Hétérogénéité des points de vue tout d'abord : le discours sur les entités nommées est en effet loin d'être consensuel et les gens semblent être gênés dans l'appréhension de ces unités, ne sachant ni par quel bout les prendre, leur forme ou leur sens, ni comment délimiter leur ensemble. Hétérogénéité dans les réalisations ensuite, confère la figure 3.1 rassemblant diverses unités considérées comme des entités nommées. Ainsi, qu'il s'agisse des « débordements » des réalisations ou des hésitations des formules définitoires, une définition des entités nommées devra si possible expliquer cette diversité et en proposer un cadre d'appréhension qui, sans nécessairement être consensuel, soit satisfaisant pour tout le monde.

Au final, en plus de dégager des propriétés conciliant aspects linguistiques et exigences du TAL, il conviendra donc de tenter d'expliquer l'hétérogénéité des entités nommées. Pour ce faire, et compte tenu de ce que l'on trouve dans la figure 3.1, il apparaît nécessaire d'aller voir tout d'abord du côté de la théorie linguistique, avec ce qui se dit du côté des noms propres, puis du côté des descriptions définies, en prenant soin de replacer ces propos dans le cadre théorique linguistique du sens et de la référence.

# Chapitre 4

# Quelques pistes linguistiques : les catégories existantes

Positivement parlant, l'ensemble baptisé « entités nommées » comporte, nous l'avons vu, des unités linguistiques fort différentes, avec des noms propres, des descriptions définies ainsi que diverses expressions numériques et temporelles. Dans un but de définition des entités nommées, il est possible de s'appuyer, dans un premier temps, sur ces unités et de tenter d'en dégager, en une démarche sémasiologique, certaines propriétés. Deux questions motivent l'étude à venir : d'une part, que peut-on déduire des unités lexicales présentes dans l'ensemble des entités nommées qui puisse servir à une meilleure appréhension de ces dernières et, d'autre part, cette perspective linguistique suffit-elle à définir les entités nommées ? C'est donc pour tenter de dégager un dénominateur commun pouvant servir de socle linguistique à une définition des entités nommées que ce chapitre reviendra successivement sur les aspects linguistiques des noms propres puis des descriptions définies. Il importera cependant de spécifier au préalable les notions appelées à être manipulées, à savoir celles de sens et de référence.

## 4.1 La linguistique et le réel : les notions de sens et de référence

La linguistique étudie scientifiquement le langage en tant que système, avec la considération de ses aspects phonologiques, lexicaux, syntaxiques, sémantiques et pragmatiques. Parmi ces différents niveaux d'investigation, la sémantique a pour objet d'étudier le langage du point de vue de la signification et donc de contribuer à démêler, dans une certaine mesure, « *la question fondamentale de la philosophie du langage* [qui est] *de comprendre comment le langage entre en relation avec le*

*réel* » (J. Searle[1], cité par M. Charolles [Charolles, 2002b]). Plus prosaïquement, il s'agit de tenter de résoudre l'énigme suivante : par quelle alchimie le langage nous permet-il de *dire* le monde et de communiquer entre nous ? Comment, en prononcant tel mot plutôt qu'un autre, est-il possible de pointer un élément du réel ? Ces questions, quelque peu maladroitement formulées ici, posent en fait celle(s), primordiale(s), du sens et de la référence.

Ces deux notions, généreusement débattues par nombre philosophes et logiciens, n'ont pas toujours été à l'honneur parmi les linguistes, la tradition saussurienne préférant en effet contourner les problèmes de référence, relevant de la parole, pour s'intéresser exclusivement aux signes, relevant du langage. Signe, sens, référent, nous avons là le célèbre triangle sémiotique reliant le signifiant (ou mot), le signifié (ou concept) et le référent (ou chose). La considération exclusive de la branche droite (signifiant-signifié), « *la branche noble, la branche vraiment linguistique* » selon G. Kleiber (ironisant), a conduit à une sémantique intralinguistique, avant qu'il ne s'avère réellement indispensable pour comprendre les mécanismes du langage de réintroduire la branche gauche, et donc la référence, dans les affaires de linguistique. G. Kleiber pose d'ailleurs, en préambule de son ouvrage intitulé *Problèmes de Sémantique* [Kleiber, 1999b], la question encore « saugrenue » de « *que faut-il faire du réel en linguistique ?* », question à laquelle il ne cesse en fait d'apporter de réponses tout au long de ses contributions, arguant que « *si l'on accepte que parler, c'est dire quelque chose, le quelque chose en question, que l'on ne peut éviter, nous pousse à répondre positivement : oui, le réel est partie prenante dans le commerce linguistique, puisque c'est sur lui que s'exerce notre dire* ». Il est ainsi nombre de travaux qui, désormais, contribuent à entériner la référence comme champ d'investigation légitime en linguistique. Cette dernière a naturellement maille à partir avec le sens : il convient ici d'éclaircir simplement ces notions, sans entrer précisément dans le détail de débats animés. Pour ce faire, nous nous appuierons principalement sur les réflexions de G. Kleiber [Kleiber, 1999a] et de M. Charolles [Charolles, 2002b].

### 4.1.1 Qu'est-ce que la référence ?

La *référence* désigne le lien qui existe entre une expression linguistique et l'élément du réel auquel elle renvoie. Cet élément du réel est un objet ou état de fait extérieur au langage et est appelé le *référent*. On peut ainsi dire d'une expression linguistique qu'elle *réfère* lorsqu'elle renvoie à une entité appartenant à un monde extérieur au langage. Le corollaire incontournable de cette idée de référence est l'idée d'existence : si je parle d'un référent, d'une entité du monde

[1] J. Searle, *L'intentionalité, essai de philosophie de l'esprit*, Paris, Minuit.

réel au moyen d'une expression linguistique, c'est bien que ce référent, cette entité, existe. M. Charolles souligne que « *l'usage naturel aussi bien que philosophique et linguistique de la notion de référence inclut nécessairement l'idée d'existence* ». Il s'agit de l'« axiome d'existence » sur lequel repose la référence : « *Tout ce à quoi on réfère doit exister* » énonce J. Searle ([Searle, 1972], cité par [Kleiber, 1999a]). Cet axiome d'existence pose la question de la conception de ce monde extérieur au langage, conception qui n'est pas sans conséquence pour la référence : si l'extra-linguistique n'existe pas, la thèse classique de la référence, telle que présentée ci-dessus, ne peut être maintenue. Autrement dit, si référer implique exister, de quelle existence s'agit-il ? G. Kleiber distingue deux paradigmes.

Une position naturelle, semble-t-il, est de croire en l'existence objective de la réalité désignée à l'aide d'expressions linguistiques. Si quelqu'un dit « *Le pain est sur la table* », cela renvoie de manière effective à deux portions de la réalité considérée, à savoir une entité *pain* posée sur une autre entité *table*. De même, pour reprendre un exemple bien connu, le nom propre *Napoléon*, dans *Napoléon est mort à Sainte Hélène*, renvoie à un individu ayant véritablement existé. Cette manière de concevoir la réalité correspond au paradigme objectiviste : les entités auxquelles réfèrent les expressions linguistiques ont une existence réelle, objective, indépendante du langage. Cet engagement ontologique en faveur de l'existence réelle de ce qui constitue le monde met cependant le dogme objectiviste dans une position difficile face aux entités fictives : si le *Père Noël*, *Peter Pan* et les *elfes* n'existent pas, comment se fait-il que l'on puisse tout de même y faire référence ? Il convient donc d'étendre la notion de référence à des « mondes possibles » et de préciser qu'il s'agit de *la propriété d'un signe linguistique lui permettant de renvoyer à un objet du monde extra-linguistique, réel ou imaginaire* » [Dubois *et al.*, 1994]. Cette conception de la référence, si elle permet de rendre compte de toutes les « références » possibles, renvoyant à des entités réelles ou élaborées par le discours, met toutefois en péril le réalisme objectiviste.

Le paradigme constructiviste, en réponse à la possibilité de renvoyer à des êtres ou choses non existants, affirme que le monde ne préexiste pas au discours et que les objets qui le composent ne possèdent pas de qualités intrinsèques. Notre vision du monde est étroitemment dépendante de notre perception, déterminée par nos capacités physiques et notre environnement culturel, et il est impossible de dire que nous avons accès au monde tel qu'il est. Afin d'illustrer cela, prenons, à la suite de Kleiber, l'exemple de l'eau : un homme, un poisson ou un ver de terre en ont des perceptions, donc des conceptions, certainement fort différentes. De même, on peut se demander avec F. Varela [Varela, 1998] « *qui voit la vraie couleur* » : les pigeons qui voient en pentachromatique, ou les abeilles qui voient dans l'ultraviolet ? Ainsi, le monde qui apparaît à l'observateur est un monde

construit, interprété ; l'individu élabore sa représentation du monde, cette dernière lui est propre et il ne peut prétendre avoir accès à la réalité *objective* des choses. Dans cette optique, les entités référentielles se voient conférer un nouveau statut : ce ne sont plus des entités réelles indépendantes du langage mais des constructions mentales, des objets de discours qui n'existent que par le biais de ce qui les énonce.

Sur le plan du langage, ces paradigmes conduisent, selon G. Kleiber, à deux impasses : d'une part l'objectivisme contraint le langage à n'être qu'une nomenclature avec d'un côté les choses du monde et de l'autre le langage qui les nomme et, d'autre part, le constructivisme nie toute référence externe à un monde qui n'est que construit pour préférer une référence totalement intralinguistique (la référence s'exerce sur des entités construites par le discours). Ces positions, trop radicales, difficilement défendables et incompatibles avec une étude linguistique de la référence en tant que pivot entre le langage et le réel conduisent G. Kleiber, et d'autres, à adopter une position intermédiaire.

Cette dernière s'articule autour de deux propositions clés : un « réalisme modulé » pour une « sémantique référentielle ». Précisons les choses. Le premier point concerne le mode d'existence des référents. Entre objectivisme et constructivisme, il est une position médiane qui consiste à accepter que « *ce que nous croyons être le monde réel n'est que le monde tel que nous le percevons ou tel que nous croyons qu'il est* ». S'il n'y a pas de réalité objective mais qu'une réalité expériencée, cela importe finalement peu, l'essentiel étant que lorsqu'on réfère à une entité par le biais d'une expression linguistique cette entité soit considérée comme existant vraiment dans un monde, réel ou seulement considéré comme tel. « *Toute réalité* [étant] *conceptualisée* », le point fondamental est qu'il y ait un certain accord sur cette réalité, soit ce que G. Kleiber appelle une « stabilité intersubjective ». Les capacités perceptives humaines (structures physiologiques et mentales) étant fortement équivalentes d'un sujet à l'autre, il en résulte un sentiment d' « objectivité » dans la conceptualisation du monde. Celui-ci, même interprété, apparaît ainsi identique d'un individu à l'autre : sa conceptualisation est partagée et ce partage donne lieu à une stabilité intersubjective qui « *contraint à considérer cette réalité conceptualisée partagée comme étant une réalité privilégiée (...) que nous croyons être la réalité* ». Une fois ce « monde réel » stabilisé, il devient possible de concevoir des mondes possibles et de faire référence à des objets non existants. C'est en effet par rapport à un réel, quel qu'il soit, que le fictif peut se positionner et être envisageable. Au sein d'une réalité intersubjectivement stable, la référence peut, au final, jouer de toutes ses cordes. Nous avons donc un « réalisme modulé » offrant la possibilité d'évoquer toutes sortes d'entités, écueil de l'objectivisme, et ce au-delà de la multitude des perceptions individuelles, enceinte phénoménologique du constructivisme.

Le second point, complémentaire du précédent, positionne les référents par rapport au langage. S'il est possible de considérer un monde, les référents de ce monde peuvent de nouveau être considérés comme extérieurs au langage. En effet, la plus malheureuse des conséquences du paradigme constructiviste est de préférer une référence interne, dans laquelle les référents sont des objets de discours, à une référence externe, dans laquelle les référents sont des entités du monde. Pour G. Kleiber, cette hypothèse « *méconnaît un point crucial, à savoir que le langage en tant que système de signes est tourné vers le dehors, vers ce qu'on appelle ou ce qu'on croit être la réalité ou encore le monde, précisément parce qu'un signe n'est signe que s'il désigne un autre que lui-même* ». Que le monde soit réel ou non, les expressions linguistiques font référence à des entités considérées comme existant en dehors du langage, ne se réduisant pas à des objets mentaux. Il s'agit là de la nécessaire prise en compte de la branche droite du triangle sémiotique, conduisant à une « sémantique référentielle ». Dans ce cadre, les expressions référentielles, « *si elles réfèrent, réfèrent à des éléments 'existants', réels ou fictif, c'est-à-dire conçus comme existant en dehors du langage : cette existence leur est garantie par cette modélisation intersubjective stable à apparence d'objectivité qui caractérise notre appréhension du monde* ». Ainsi, si notre réalité n'est que modélisée, cette modélisation n'implique pas que les entités de cette réalité ne soient que des objets mentaux. Le langage participe bien sûr de cette modélisation mais ce qu'il produit ne se réduit pas à des objets linguistiques ; il convient de ne pas perdre de vue ce pour quoi il existe et son « *rapport sémiotique fondamental avec l'extérieur* ». « *L'élément décisif dans la référence*, conclut G. Kleiber, *c'est qu'elle nous mène au dehors du langage* ». Ainsi, au terme de cette exploration, guidée par G. Kleiber, il est possible de reprendre la formule énoncée en introduction,

> La *référence* désigne le lien qui existe entre une expression linguistique et l'élément du réel (appelé *référent*) auquel elle renvoie,

et de lui adjoindre deux précisions : d'une part l'élément du réel dont il est question se trouve dans un réel conceptualisé sur la base d'une intersubjectivité stable et, d'autre part, cet élément existe en dehors du langage. Ces caractérisations portent sur la seconde partie de la formule présentée ci-dessus, il serait temps de s'intéresser à la première et de se demander par quel(s) moyen(s) une *expression linguistique* peut être mise en rapport avec un *élément du réel*. C'est en vertu, principalement, de son sens.

### 4.1.2 Qu'est-ce que le sens ?

#### 4.1.2.1 Une idée générale de la notion de sens

Pour comprendre la notion de sens, il est utile de reprendre le raisonnement mené par G. Frege [Frege, 1892] à propos de la notion d'égalité. Se demandant s'il s'agit d'une « *relation entre des objets, ou entre des noms ou signes d'objets* », G. Frege est amené à faire la différence entre la notion de *sens* (Sinn) et celle de *dénotation*[1] (Bedeuntung). Reprenons son exemple célèbre : *l'étoile du soir* et *l'étoile du matin* sont deux expressions différentes qui, cependant, désignent un même référent, à savoir la planète *Venus*. Deux formes linguistiques *dénotent*, pour reprendre la terminologie de l'auteur, la même portion de réalité ou objet du monde, mais le font différemment. Fregge explique que nous avons diverses désignations pour le même objet (« étoile du matin » et « étoile du soir » ) et que ces noms indiquent en même temps la manière dont cet objet est donné. Cette différence est particulièrement perceptible lorsque l'on pratique le test dit de substitution : si « l'étoile du matin est l'étoile du matin » est une tautologie « l'étoile du matin est l'étoile du soir » est une phrase qui apporte une « *connaissance effective* ». Ces deux expressions ont donc une *dénotation* (ou référence) identique, mais des *sens* différents. Pour Frege, le sens se rapporte au « mode de présentation » par lequel une expression livre son référent : « *Il est naturel d'associer à un signe (nom, groupe de mots, caractère), outre ce qu'il désigne et qu'on pourrait appeler sa dénotation, ce que je voudrais appeler le sens du signe, où est contenu le mode de donation de l'objet* ». C'est donc le sens des expressions qui détermine si telle ou telle chose peut par elle être désignée. Avec ce « mode de donation » nous abordons déjà l'inépuisable question de la définition du sens ; l'essentiel, avec cette démonstration de Frege, était de bien distinguer le sens (ce que véhiculent les expressions « étoile du matin » et « étoile du soir » ) de la référence (l'objet *Vénus* auquel elles renvoient). Cette présentation de la notion de sens correspond à, pourrait-on dire, la partie visible de l'iceberg ; il importe tout de même d'en évoquer la partie immergée, celle de la définition du sens.

#### 4.1.2.2 Et une vue d'ensemble des définitions du sens

Il n'existe aucun consensus sur la définition du sens, tant s'atteler à le caractériser, en s'interrogeant sur sa nature et en observant les divers rouages qu'il anime à chaque entreprise de communication, est chose difficile. M. Leiris pré-

[1] Sans qu'il soit nécessaire de revenir ici sur les imbroglios terminologiques auxquels donne lieu cette expression, il convient de préciser qu'elle correspond à ce que nous avons précédemment appelé la *référence*.

sente les choses autrement et nous dit « *sens : (sans anse pour le prendre)* » [1] ; ce dernier est d'autant plus difficile à saisir que « *tout bouge en sémantique à l'heure actuelle* » [Kleiber, 1999b]. Sans préciser tous les tenants et aboutissants des diverses théories sémantiques, nous souhaitons, brièvement et schématiquement, en préciser ici quelques éléments.

**Comment aborder la question du sens ?** Afin de faciliter l'appréhension de cet objet complexe ainsi que la compréhension des théories s'attachant à définir le sens, il convient au préalable de fixer quelques notions, oppositions ou questions clés sous-tendant la problématique du sens. Il importe tout d'abord de préciser qu'il est possible d'étudier le sens à divers niveaux : le mot, la phrase et le texte sont les trois « paliers » de la sémantique distingués par [Enjalbert et Victorri, 2005]. Ces derniers insistent bien sur le fait que cette échelle, en sémantique, présente plus des « subdivisions de méthode » qu'une réelle partition, les divers niveaux étant bien entendu perméables. Les unités que nous souhaitons étudier par la suite (nom propre et descriptions définies) relevant du premier palier, voire du deuxième (avec le syntagme), il sera question, dans cet exposé, de sémantique lexicale et non de sémantique textuelle. Peuvent être considérés à ces niveaux deux grands types d'unités, les unités lexicales d'une part, et les unités grammaticales d'autre part. Les premières correspondent à ce qu'on appelle parfois les « mots pleins », avec les noms, les verbes, les adjectifs, etc. et les secondes aux « mots outils », avec les déterminants, les conjonctions, les prépositions, etc. Les diverses unités lexicales faisant partie de l'ensemble entités nommées relèvent du lexique plein. Voici donc pour la délimitation, ou une délimitation possible, du champ d'étude du sens.

Une autre question, ou opposition, a trait à la nature du sens : celui-ci peut être perçu comme exclusivement linguistique ou bien comme faisant également intervenir un niveau cognitif. « *Les linguistes cognitifs*, nous dit D.A. Cruse [Cruse, 1996], *ont l'habitude de considérer les faits du domaine linguistique comme des reflets de faits du domaine cognitif et comme motivés par la nature, les structures et les processus de la cognition générale* ». Les approches cognitives de la linguistique considèrent que le langage, en général, et le sens, en particulier, sont étroitement liés à la façon dont nous expérimentons et conceptualisons la réalité (ou ce que nous considérons comme telle). Les moyens et les mécanismes d'acquisition des connaissances ne seraient ainsi pas étrangers à l'expression de ces dernières. Autrement dit, les représentations que chacun est à même de construire à partir du sens des mots le sont-elles grâce à un composant linguistique uniquement ou sont-elles également le fruit d'un processus cognitif ?

[1] *Langage, tangage ou Ce que les mots me disent*, Gallimard, Collection L'Imaginaire, 1995.

Une troisième question incontournable à considérer dans les définitions du sens est celle des primitives de sens. Plus que de sa nature, il s'agit là de ses « ingrédients » : le sens comporte-il un « noyau » porté par un élément linguistique ou au contraire est-il le résultat de l'interprétation d'un sujet donné dans un contexte particulier ? La prise en compte d'éléments de sens premiers n'interdit pas celle du contexte ; néanmoins, le degré d'importance accordé à l'un ou à l'autre, de manière plus ou moins exclusive, conduit à des visions du sens différentes, plus ou moins dynamiques, du sens intrinsèque au sens interprété.

Ces questions du niveau d'étude, de l'angle d'attaque et de l'existence ou non de primitives de sens permettent de donner quelques points de repères dans la problématique du sens. Quelles sont les positions adoptées par les diverses théories du sens au regard de ces critères d'appréciation ? Il est possible d'en examiner rapidement quelques unes en présentant, à la suite toujours de G.Kleiber [Kleiber, 1999b], la principale opposition entre les partisans d'un sens référentiel, et ceux d'un sens aréférentiel.

**Quelques approches du sens** L'approche référentielle propose de considérer le sens d'une expression linguistique comme un ensemble de traits objectifs déterminant les caractéristiques qu'une entité doit satisfaire pour être désignée par cette expression. Si nous reprenons les exemples donnés précédemment, une entité doit avoir, pour être désignée par le mot « table », des pieds et un plateau (grosso-modo) et il faut avoir gagné à Austerlitz pour pouvoir être appelé « le vainqueur d'Austerlitz ». Les expressions linguistiques apparaissent donc dotées d'un certain contenu sémantique précodé (des primitives de sens), par le biais duquel elles peuvent référer à des entités. Ce contenu sémantique est généralement appellé *conditions d'application* (ou encore conditions de vérité, de satisfaction) : elles sont non subjectives, analysables hors-contexte et constituent une sorte de « programme référentiel » en vertu duquel tel ou tel segment de la réalité peut être désigné. Il s'agit d'un sens *dénotatif*, stable, objectif et qui conditionne la référence, que l'on peut distinguer d'une partie moins stable de la signification, appelée sens *connotatif*, comportant des traits subjectifs et variables contextuellement. Cette distinction apparaît plus souvent aujourd'hui sous les termes de « référence virtuelle » et « référence actuelle » (J.C Milner[1], cité par [Kleiber, 1999b] et [Riegel *et al.*, 1994] : celle-ci correspond à l'actualisation, dans une situation de discours et dans un contexte donnés, des traits contenus dans la définition d'une expression linguistique sous la forme d'objets ou d'êtres particuliers désignés par cette expression, et celle-là correspond à « *l'ensemble de conditions que doit satisfaire un segment de la réalité pour pouvoir être la référence d'une séquence où*

[1]J.C. Milner, *De la syntaxe à l'interprétation*, Paris, Le Seuil, 1978

*interviendrait crucialement l'unité lexicale en question* ». Autrement dit, la référence actuelle est le segment de la réalité désigné par une expression, elle a ainsi exclusivement à voir avec le monde, la référence virtuelle est l'ensemble des conditions caractérisant une unité lexicale, elle a ainsi à voir avec le code linguistique, les deux restant bien sûr étroitement liées. Le principe du sens comme « mode de présentation » (ou donation) du référent proposé par G. Frege s'inscrit dans cette conception du sens référentiel dans la mesure où il contient un ensemble de critères permettant de déterminer la dénotation d'une expression, critères assimilables à des conditions d'application. Cette approche postule donc, au final, un sens perçu comme « *un faisceau de traits intrinsèques ou inhérents du référent, ou encore traits 'objectifs' , c'est-à-dire de traits qui sont supposés être possédés par le référent, donc des traits référentiels, en lien avec la réalité* ».

À l'opposé, se positionnant plus ou moins radicalement, se trouvent les approches que G. Kleiber a rassemblées sous l'étiquette de « *paradigme du sens aréférentiel* ». Celles-ci contestent soit le caractère stable et conventionnel du sens (pas de primitives de sens), soit son caractère référentiel (des primitives, mais non référentielles). Pour les uns, le sens est effectivement nécessairement indéterminé en raison de son caractère individuel (mon sens de « table » ne peut être le même que celui du voisin), ou il ne peut être conçu comme conventionnel, étant donné qu'il se construit à chaque utilisation (on ne peut savoir *par convention* que le mot « table » désigne une entité ayant certaines caractéristiques, ces caractéristiques sont au contraire apprises ou construites, par le biais du contexte, à chaque utilisation de ce mot). Les autres, admettant qu'il existe des éléments de sens constants et partagés par tous, remettent toutefois en cause leur caractère référentiel, et ce de diverses manières. Le sens peut tout d'abord être vu comme différentiel, c'est-à-dire comme désolidarisé de la référence, les signifiés ne se définissant plus « *positivement par leur contenu, mais négativement par leur rapports avec les autres unités du système* ». Défendue par F. Rastier, cette approche, sans être contestée outre mesure, est davantage perçue par G. Kleiber comme une théorie de l'organisation du sens (« *le principe oppositif peut certes dire quelles oppositions il y a ...* » ) que comme une théorie du sens (« *... mais ne sauraient dire en quoi elles consistent* » ). Certains [Anscombre, 1996] pensent le sens comme adescriptif : ce qu'indique le sens n'est plus un ensemble de traits descriptifs permettant d'identifier une portion de la réalité, mais un ensemble d'instructions permettant de guider vers cette entité ; il devient alors argumentatif. D'autres, sans abandonner le caractère descriptif du sens, situent les choses à un niveau supérieur avec un sens schématico-dynamique pourvoyeur de descriptions abstraites (renvoyant à une catégorie de référents) et non plus immédiates (renvoyant à un référent précis) [Franckel et Paillard, 1997] et [Kleiber, 1999b,

p.41-42]. Enfin, certains font « déborder » le sens des mots jusqu'à leur emploi en contexte : si les mots ont bien une signification qui leur est propre (appelée valeur sémantique), leur sens ne se trouve véritablement engendré que par leur interaction avec le contexte [Récanati, 1997]. Cette approche du sens correspond au contextualisme, qui peut être soit modéré, avec la conservation d'un sens propre pour les expressions linguistiques, soit radical, avec l'abandon de cette part de signification fixe. Les possibilités d'appréciation du sens, en tant que tel, sont donc nombreuses.

**Sens descriptif et sens instructionnel** L'objectif de cette rapide revue n'était pas tant de défendre une position plutôt qu'une autre mais, à la suite de G. Kleiber, de montrer qu'il est difficile de remettre intégralement en cause le caractère référentiel du sens. D'une manière ou d'une autre, le sens, qu'il soit différentiel, adescriptif, schématique ou « générationel », doit expliquer comment se fait la mise en rapport avec un référent et, pour ce faire, « *il est intuitivement préférable de concevoir le sens comme constitué de traits référentiels qui délimitent virtuellement leur référent* ». C'est pourquoi G. Kleiber propose de « rabibocher » le sens et la référence et définit, dans le cadre d'une sémantique référentielle, un modèle hétérogène offrant une double caractérisation du sens, celle-ci permettant de rendre compte de l'existence de primitives de sens tout comme de souligner l'importance du contexte.

En effet, outre l'existence d'une sorte de clivage entre partisans d'un sens référentiel et partisans d'un sens aréférentiel (partition schématique, comme le souligne G. Kleiber), il semble qu'il existe une autre ligne de force, s'articulant pour sa part autour des notions de primitive de sens et de contexte et partageant les opinions selon l'importance, voire l'exclusivité, accordée à l'une ou à l'autre dans le processus de construction du sens. Il paraît évident que le sens de certaines unités lexicales dépend pour une grande part de leur apport propre tandis que pour d'autres, le sens dépend majoritairement du contexte. Pour les premières, on parle alors d'un *sens descriptif*, car conduisant à la référence par le biais de traits descriptifs constants et objectifs, ne nécessitant qu'un apport minime, voire nul, du contexte. C'est le cas par exemple pour les mots *oursin* et *tournevis*, dont le sens ne dépend que de ce que l'on pourrait appeler une « part linguistique ». À l'inverse, pour le second type d'unité lexicale, on parle de *sens instructionnel*, celui-ci conduisant au référent par le biais d'un ensemble d'instructions indiquant les procédures à suivre pour y parvenir. Ces instructions permettent de prendre en compte une ou plusieurs des composantes de la situation d'énonciation et d'opérer ce qu'on appelle une référence indexicale. Examinons à titre d'exemple le pronom *je* : son apport propre se réduit à « celui ou celle qui parle ou qui écrit » et

il importe donc, pour comprendre à quel référent renvoie une occurence donnée du mot *je*, d'exploiter la situation d'énonciation, de voir qui est la personne qui parle effectivement. De même, la compréhension de l'adverbe *maintenant* impose la prise en compte des données temporelles de la situation d'énonciation, qui ne peuvent aucunement être données une fois pour toute dans le sens lexical de *maintenant*. L'indexicalité désigne ainsi l'incomplétude naturelle de certains mots qui ne prennent leur sens complet que dans leur contexte, que s'ils sont indexés à une situation d'échange linguistique. Cette indexation est guidée par le sens instructionnel, s'appliquant à des unités lexicales dont on peut alors dire que leur interprétation dépend d'une « part contextuelle ».

De manière générale, sens descriptif et sens instructionnel s'appliquent, pour le premier, au lexique plein et, pour le second, aux mots grammaticaux. Il ne s'agit cependant, soulignons-le, que d'une tendance très générale, dont les exemples ci-desssus, *tournevis* et *je*, représentent des cas extrêmes très prototypiques. Sans aller jusqu'à affirmer que tout sens est instructionnel, il convient de signaler que ces deux types de sens peuvent se combiner. Il existe en effet plusieurs niveaux de dépendance au contexte pour la constitution du référent : si l'indexicalité est très forte pour l'ensemble des déictiques, elle est cependant plus faible pour d'autres expressions telles que *le ministre*, laquelle donne des indications sur le référent qu'elle désigne, sans pour autant permettre son identification sans le recours au contexte. Pareillement, l'occurrence du verbe *faire* dans *Pour faire le portrait d'un oiseau* a un sens essentiellement déterminé par sa « part linguistique », le sens lexical de « fabriquer », tandis que dans *Est-ce qu'il l'a fait ?*, l'occurrence de *faire* a un sens plus anaphorique, faisant davantage jouer la « part contextuelle » du sens pour déterminer de quoi il s'agit (il n'y a plus le sens de *fabriquer*).

À la suite de G. Kleiber, il paraît ainsi pertinent de concevoir deux types de sens différents, se combinant au sein d'un modèle hétérogène du sens. Le sens instructionnel donne un moyen d'accès au référent par le biais d'un ensemble d'instructions indiquant les procédures à suivre pour y parvenir, tandis que le sens descriptif offre la même possibilité, mais par le biais d'un ensemble de traits descriptifs, ou conditions d'application. S'appliquant généralement aux mots grammaticaux pour le premier, au lexique plein pour le second, les deux types de sens peuvent néanmoins, et bien souvent, « cohabiter » pour une même unité lexicale. Ainsi, quel que soit le type d'expression linguistique, il est possible de rendre compte du fonctionnement du sens et de son articulation avec la référence. C'est dans le cadre de ce modèle hétérogène du sens, participant d'une sémantique référentielle, que nous souhaitons situer les propositions à venir concernant les entités nommées.

À la suite de ces précisions sur les notions de sens et de référence, il est temps de considérer les unités lexicales composant l'ensemble des entités nommées. Nous reproduisons pour mémoire la figure (cf. 4.1 ci-après) présentant un échantillon de ce que l'on appelle « entité nommée ».

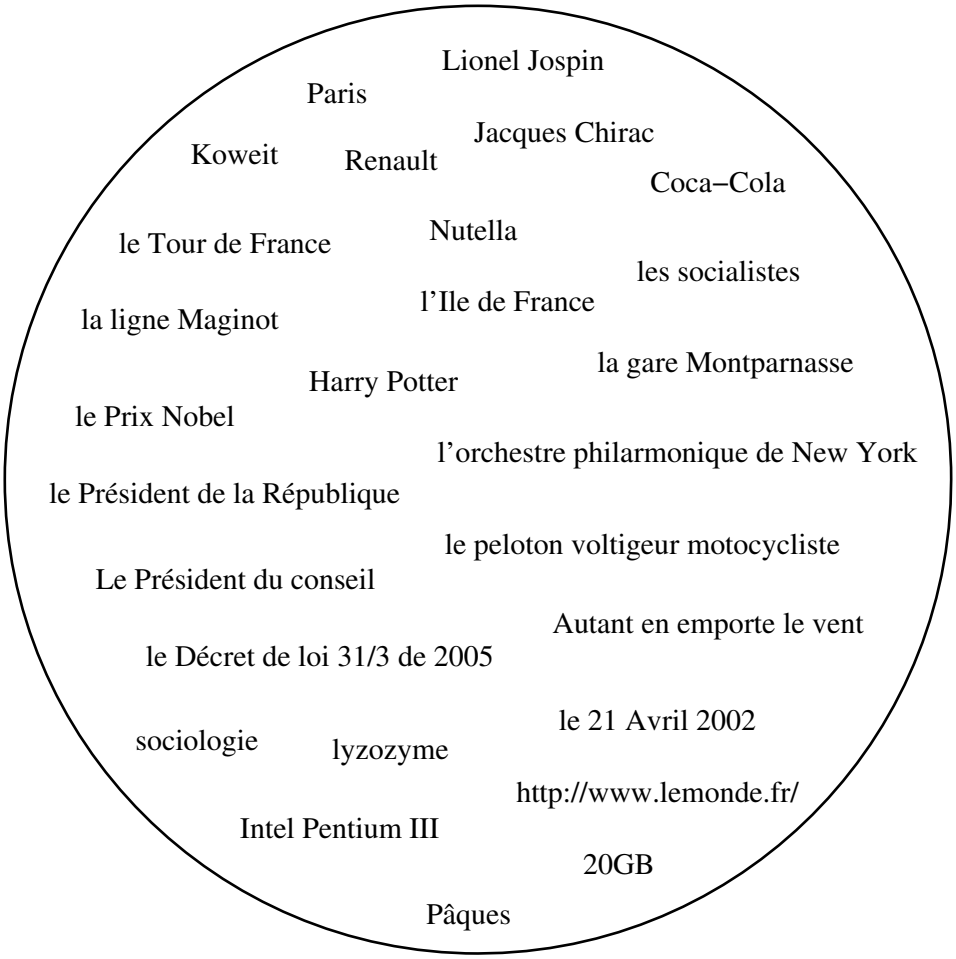



FIG. 4.1 – Echantillon d'unités lexicales considérées comme des entités nommées.

Il convient de considérer la réalité de cet ensemble et d'examiner les catégories linguistiques existantes que sont, sans surprise, les noms propres (section 4.2) et les descriptions définies (section 4.3) principalement.

## 4.2 Les noms propres

Dans la figure 4.1, les formes linguistiques telles que *Lionel Jospin, Paris* et *Renault* correspondent à ce que tout un chacun qualifie sans hésitation de nom propre. Le *Tour de France*, l'*Ile de France* ainsi que *la gare Montparnasse* peuvent de même être classées dans cette catégorie, mais cela paraît moins évident. Qu'en est-il de cette catégorie linguistique du nom propre, visiblement délicate à délimiter ?

Il n'est pas facile de présenter le nom propre, tant cet objet a posé et pose encore des questions, et ce à de nombreuses disciplines. S'il a depuis toujours sa place au sein de la catégorie nominale[1], le nom propre n'est pas, loin s'en faut, un objet d'étude de la seule discipline linguistique. Il s'agit bien au contraire d'un champ d'investigation largement ouvert, interdisciplinaire : la logique, la philosophie, l'anthropologie, la sociologie ou encore la critique littéraire composent en effet autant de disciplines que de perspectives d'analyse du nom propre. La linguistique lui a, pour sa part, longtemps tourné le dos, le considérant comme en dehors du système de la langue ([de Saussure, 1915], cité par [Leroy, 2004a], déclare que les noms propres « *ne permettent aucune analyse et par conséquent aucune interprétation de leurs éléments* »). Depuis une trentaine d'années néanmoins, cette dernière se réapproprie progressivement cet objet d'étude et, si M.N. Gary-Prieur demande encore en 1991 si « *le nom propre constitue (...) une catégorie linguistique* » [Gary-Prieur, 1991], il est possible aujourd'hui d'énumérer, reprenant le constat opéré par S. Leroy dans son ouvrage de synthèse [Leroy, 2004a], les divers travaux linguistiques sur le nom propre, allant des approches syntaxiques aux approches sémantiques, en passant par . . . la linguistique informatique. Ce renouveau des études linguistiques sur le nom propre emprunte divers chemins, dont nous allons tenter de restituer les grandes lignes, examinant les fondements et retraçant les évolutions du discours théorique linguistique sur le nom propre. Il conviendra de partir du constat classique de la difficile définition du nom propre pour ensuite examiner l'inépuisable et primordiale question du sens (et de la référence) des noms propres, et enfin dégager des points essentiels à mettre en perspective avec les entités nommées.

### 4.2.1 Nom propre : une définition difficile

Une explication possible de la désaffection de la linguistique vis-à-vis du nom propre peut être la difficile définition de cette unité lexicale par des critères d'ordre linguistique. En effet, si les grammaires consacrent quelques lignes au nom propre c'est, dans la plupart des cas, pour énumérer des critères à l'encontre desquels viennent facilement à l'esprit de nombreuses exceptions. Que l'on adopte un point de vue morphologique, syntaxique ou sémantique, le nom propre semble toujours se conformer en même temps que s'affranchir des indices évoqués. Présentons, à la suite de S. Leroy, ces divers « critères traditionnels de définition du nom propre » ; différents aspects du nom propre seront ainsi évoqués « pêle-mêle », avant d'être approfondis, pour certains, dans la section suivante.

[1] Cette insertion est remise en cause par [Flaux, 1995], qui présente l'équivalence entre nom propre et nom commun comme une illusion, préférant rapprocher celui-là du syntagme nominal.

#### 4.2.1.1 Les critères d'ordre formel ou factuel.

La majuscule, marque graphique par excellence du nom propre, apparaît comme un point de départ définitoire relativement répandu. Ce critère, pour robuste qu'il puisse paraître pour un apprentissage de la langue ou pour un système de reconnaissance automatique, n'est cependant pas principal dans la caractérisation du nom propre : il n'est pas translinguistique (usage différent d'une langue à l'autre, l'allemand par exemple met des majuscules à tous les noms, communs ou propres), n'est pas valide en diachronie (l'emploi de cette marque n'est bien établi que depuis l'apparition de l'imprimerie) et n'est pas appréciable à l'oral. Par ailleurs, il existe des noms communs ou syntagmes qui prennent la majuscule, et des noms propres qui n'en prennent pas. Pour les premiers, l'adjonction de cette marque graphique correspond à la transmission, « *pour ainsi dire par contagion, d'un ou plusieurs traits caractéristiques du nom propre* » [Jonasson, 1994]. Le nom commun doté d'une majuscule à l'initiale permet alors de désigner ou de personnifier un concept (*l'Amour, la Providence*), de présenter une réalité perçue comme particulière au sein de sa catégorie (la *Résistance* ou encore le *Débarquement* permettent de distinguer des mouvements et moments précis de l'histoire de France), de désigner une institution saillante et unique à l'échelle nationale (l'*État*, l'*Église*) ou une autre réalité conventionnellement dénommée (le *Marché Commun*, la *Loterie Nationale*), de marquer une attitude déférente vis-à-vis d'un *Professeur* ou d'une *Présidente*, ou encore d'asseoir un nom ou autre comme nom de marque (*Du Pareil au Même*, *Reflets de France*). Inversemment, on voit des noms propres abandonner leur majuscule, soit qu'ils se lexicalisent, perdant leur origine propriale par métaphorisation (*un tartuffe*) ou métonymisation (*une poubelle*), soit que l'usage soit flou, en raison principalement de la constitution du nom propre commençant par un nom commun, à l'instar des noms de rue (*rue St Honoré* ou *Rue St Honoré*) et de certains noms composés (l'*abbaye de Fontevraud* ou l'*Abbaye de Fontevraud*). Signalons encore un usage irrégulier quant aux noms de mois (il est possible d'orthographier *juillet* ou *Juillet*), hésitant à se calquer sur celui des noms de fête (*Pâques*). Ainsi, la correspondance entre nom propre et majuscule n'est pas systématique et cette dernière ne paraît pas être un critère adéquat pour cerner la catégorie des noms propres.

D'autres indices, factuels, peuvent également être invoqués pour caractériser les noms propres : la non traduction et l'absence des dictionnaires. Si l'intraduisibilité des noms propres est souvent bien réelle (*Rio de Janeiro* ne devient pas « fleuve de janvier » ni *Los Angeles* « les Anges » ), il n'en reste pas moins que certains sont traduits, pour des raisons phonétiques ou graphiques, en fonction de leur notoriété ou de leur type (traduction quasi systématique des prénoms mais

non des patronymes). Quant à l'absence des noms propres des dictionnaires, elle est pareillement à remettre en cause ; sans entrer plus avant dans le détail du traitement lexicographique des noms propres, nous nous contentons ici de souligner que ce dernier est complexe (la dichotomie dictionnaire de langue *vs.* dictionnaire encyclopédique n'est pas absolue, tous les noms propres ne figurent pas dans les ouvrages prévus à leur intention, ils apparaissent parfois dans les entrées des dictionnaires de langue, etc.) et renvoyons aux études de S. Leroy [Leroy, 2004a] et M.N Gary-Prieur [Gary-Prieur, 1991].

Majuscule, absence de traduction, absence des dictionnaires, ces critères, souvent évoqués lors de la définition du nom propre, ne permettent donc pas de délimiter ni de caractériser avec précision cette catégorie linguistique. Considérons encore d'autres indices souvent mentionnés pour une distinction du nom propre.

#### 4.2.1.2 Les critères d'ordre morpho-syntaxique

Il est question ici, pour le français uniquement, de l'absence de déterminant dans la construction du groupe nominal formé par le nom propre d'une part, et de l'absence de flexion de ce dernier d'autre part. Le nom propre s'emploie sans déterminant, toutes les grammaires vous le diront, ajoutant aussitôt après une longue série d'exceptions. La construction non déterminée est bien sûr la plus courante pour les noms propres, mais nombreux sont les usages et les emplois de cette unité avec détermination. Tout d'abord, certains noms propres « possèdent » un article défini, comme *La Rochelle* ou certains noms de personnes, et la plupart des noms de pays, régions ou fleuves apparaissent avec un défini (*le Poitou*, *la Seine*, *l'Espagne*). Par ailleurs, un usage régional ou familier peut autoriser la détermination du nom propre de personne (*la Marie*), tout comme certaines conventions pour les noms de restaurants, de bateaux ou de cantatrices (italianisme). Pour ce qui est des emplois du nom propre, certains nécessitent une détermination, que ce soit lors de l'insertion dans un syntagme complexe (*l'historien médiéviste Georges Duby*), ou lors d'un emploi figuré (*le Paris d'après-guerre*, *le Céline antisémite est un Céline souriant*[1]). La détermination du nom propre n'est ainsi pas si exceptionnelle et, si elle fait naître bien souvent une valeur sémantique différente de l'emploi non déterminé, elle ne constitue pas un critère décisif dans sa définition.

Pour sa part, l'absence de flexion fait montre d'un usage tout aussi flottant : si le nom propre ne porte d'ordinaire aucune marque de genre ou de nombre, il est possible de distinguer des formes féminines (*Yvonne*, *la Grèce*) de formes masculines (*Marc*, *le Niger*), ou d'observer des pluriels de noms propres, marquant

[1]Exemples emprunté à [Jonasson, 1994].

une pluralité des référents (*les Iles Canaries, les Martins sont venus hier, les Bourbons ont longtemps régné sur la France*, etc.), ou un emploi figuré (*J'ai acheté trois Picassos*). Il n'existe aucune régularité dans ces emplois, qu'il est dès lors difficile de normaliser ; quoiqu'il en soit, il apparaît que le critère de l'absence de flexion n'est pas plus efficace que les autres.

#### 4.2.1.3 Les critères sémantiques et pragmatiques

Le critère de la vacuité sémantique est lui aussi systématiquement présenté par les grammaires : *Le Bon Usage* [Grevisse et Goosse, 1986] avance que « *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* », et la *Grammaire Méthodique du Français* [Riegel *et al.*, 1994] indique que « *si, comme les noms communs, ils désignent des personnes, des objets, des lieux, etc., les noms propres sont pourtant dépourvus de sens lexical* ». Si ces deux grammaires pointent une même particularité du nom propre, à savoir son absence de sens, il est à remarquer cependant qu'elles ne le font pas dans les mêmes termes, la première refusant toute « sémantique » pour cette unité, la seconde étant moins catégorique, constatant l'absence d'un sens *lexical* seulement (ces nuances seront évoquées plus loin). Effectivement, les noms propres ne peuvent ni faire l'objet d'une définition lexicographique classique ni s'insérer dans les relations sémantiques structurant l'ensemble du lexique (synonymie, hyponymie, antonymie). Ces réalités doivent toutefois être nuancées, le nom propre apparaissant dans de nombreux cas comme doté, malgré tout, d'une signification. Qu'il s'agisse de repères socio-culturels (prénoms réservés plutôt à une fille, à un garçon, à un animal domestique), d'éléments descriptifs faisant partie intégrante du nom propre (l'*Arc de Triomphe*) ou encore d'un pouvoir évocateur (*Xerox* pour la xérographie, *Craquottes* pour des biscottes bien craquantes), il s'agit bien d'éléments participant de l'interprétation du nom propre. Cette interprétation ne résulte bien sûr pas du même processus de compréhension mis en œuvre par une autre unité lexicale ou par un nom commun (renvoyant pour leur part à un objet ou concept descriptible par le biais d'une définition rendant compte de son sens lexical), mais elle existe tout de même, invitant à ne pas prendre pour argent comptant et à interroger ce critère d'absence de sens.

Un autre indice, pragmatique celui-là, est l'unicité référentielle : un nom propre opère une désignation unique. Une fois encore, cette affirmation est à nuancer avec, entre autres, le fait qu'un même nom propre peut renvoyer à plusieurs individus ou réalités différents, ou qu'un nom commun puisse renvoyer à un objet ou réalité unique (le soleil). Sans évoquer plus avant ces critères pour le mo-

ment, remarquons simplement qu'ils ne permettent pas, eux non plus, de rendre pleinement compte de la catégorie du nom propre.

Ainsi, si dans les cas prototypiques le nom propre s'écrit avec une majuscule, ne se traduit pas, est absent des dictionnaires, ne se fléchit pas (ni en genre ni en nombre), n'est pas déterminé, n'a pas de sens et opère une désignation unique, force est de constater que ces critères sont loin de dessiner parfaitement la catégorie du nom propre. Leur intersection forme un ensemble bien trop étroit, et leur disjonction un ensemble bien trop large. Ces critères traditionnels de définition du nom propre, souvent évoqués par les grammaires, peinent donc à en donner une image précise et homogène. C'est en quelque sorte de ce « constat d'échec » qu'est parti le renouveau des travaux en linguistique sur le sujet, cherchant dans diverses directions à approfondir et éclaircir cette question du nom propre.

#### 4.2.1.4 Comment aborder la question du nom propre en linguistique ?

Face à un objet relativement insaisissable par des critères traditionnels mais faisant tout de même partie de la réalité langagière, la linguistique s'est depuis quelques dizaines d'années penchée sur le nom propre, ainsi qu'en attestent les nombreuses publications sur le sujet, en France comme ailleurs[1]. Au regard des problèmes de définition présentés ci-avant, il semble bien sûr difficile, voire impossible, d'établir une théorie linguistique « unifiée » du nom propre permettant de rendre compte de tous les aspects de cette unité lexicale. Il s'agit alors, semble-t-il, de décrire au plus juste cette catégorie, de cerner et d'étudier ses propriétés caractéristiques et d'analyser ses multiples emplois. Il est possible de distinguer différents niveaux et perspectives d'appréhension des noms propres, permettant de « structurer » les études sur le sujet. Ces niveaux d'études sont naturellement loin d'être indépendants.

Le nom propre peut tout d'abord être considéré comme une forme. On reconnaît alors deux grands types morphologiques ou lexicaux du nom propre, proposés à l'origine par D. J. Allerton [Allerton, 1987], puis repris par K. Jonasson :

– Les noms propres purs, correspondant à des formes nominales spécialisées dans le rôle de nom propre (c'est à dire dont aucun des composants ne pourrait être utilisé en tant que nom commun, sauf lors d'une antonomase). Il s'agit principalement de noms de personnes ou de lieux, tels que *Jacques, Paris, Iran, la Seine, le Vaucluse, l'Océanie et l'Amérique.* Ce sont ces types de noms propres que l'onomastique étudie afin de retracer leur évolution,

[1] Avec, entre autres publications : [Kleiber, 1981b, Molino, 1982, Gary-Prieur, 1991, Gary-Prieur, 1994a, Jonasson, 1994, Noailly, 1995, Kleiber, 1995, Kleiber, 1996, de Velde *et al.*, 2000, Kleiber, 2004, Leroy, 2004a, Leroy, 2005, Vaxelaire, 2007].

au travers de leur origine, leur étymologie et leur propagation au sein de la langue, tout en considérant des aspects socio-historiques.

– Les noms propres descriptifs ou mixtes, correspondant à des noms propres dont le matériau lexical est composé, tout ou partie, de noms communs. Les noms propres « à base descriptive » sont généralement constitués d'un nom commun accompagné d'un modifieur adjectival ou prépositionnel, comme le *Jardin des Plantes*, le *Commissariat à l'Énergie Atomique*, le *Massif Central* et la *Côte d'Azur*. D'autres sont composés d'un nom propre pur et d'un nom commun et sont alors appelés « mixtes » ; il s'agit du *Mont Saint Michel*, du *Collège de France*, du *Golfe du Morbihan* ou de *Guillaume le Conquérant*.

Cette typologie morphologique, pour descriptive qu'elle soit, peut néanmoins servir de base à une cartographie de la catégorie du nom propre, comportant en son cœur des noms propres prototypiques (noms propres purs), et à sa périphérie d'autres formes dont le statut proprial est plus ou moins certain (noms propres descriptifs ou mixtes).

Le nom propre peut ensuite être considéré sous un biais communicatif, par la prise en compte de sa fonction référentielle. C'est ici que se situe l'héritage du discours logique en linguistique, avec la considération exclusive des propriétés référentielles du nom propre, considéré comme un terme singulier désignant un être unique. Cette perspective logique, longtemps adoptée en linguistique, est aujourd'hui, sinon remise en cause, au moins remaniée avec des études s'intéressant d'un point de vue linguistique à la question du sens des noms propres, tout en prenant en compte la diversité de ses emplois. Nous reviendrons dans la section suivante sur cette approche sémantique du nom propre, mais il peut d'ores et déjà être utile de signaler un des apports issus de ces travaux : la distinction entre nom propre non modifié (ou nom propre standard) et nom propre modifié. Cette séparation entre différents emplois du nom propre résulte de la prise en considération par les linguistes d'énoncés authentiques, lesquels ont fait « apparaître » d'autres constructions du nom propre, jusqu'alors ignorées car absentes des exemples construits et hors contexte des logiciens. En effet, aux côtés de constructions telles que *Julien va à la piscine le mercredi*, il en existe d'autres telles que : *L'Irak est un nouveau Vietnam*. Confrontés à une réalité composite du nom propre, révélant tant des emplois prototypiques que marginaux, les linguistes se sont donc trouvés devant une « *alternative embarrassante* » [Kleiber, 1981b] : ou bien intégrer ces emplois marginaux et les considérer comme de véritables noms propres, ou bien les rejeter et les considérer comme des noms communs. Leur prise en compte a conduit à la répartition suivante du nom propre en fonction de ses emplois :

– Le nom propre non modifié ou nom propre standard. Cette appellation

renvoie à l'emploi référentiel du nom propre, ou emploi prototypique, assumant une fonction de désignation directe d'un référent unique, comme dans l'exemple suivant : *Jacques Chirac a confirmé qu'il quitterait l'Élysée.*

– Le nom propre modifié. Cette appellation renvoie aux emplois non canoniques du nom propre, alors accompagné d'un déterminant et/ou de divers modificateurs. Cette notion de modification permet de rendre compte d'emplois dénominatifs (*Il y a une Manon dans mon école*), d'emplois métaphoriques, (*l'institutrice est un vrai Napoléon*), d'emplois métonymiques (*Le Louvre a vendu deux Picassos*), de fractionnement (*le Hugo de l'exil*) ou encore exemplaires (*même un Chirac n'aurait pas tenu de tels propos*). L'analyse de la modification du nom propre peut être d'orientation syntaxique, la détermination est dans ce cas un indice fort de modification, ou d'orientation référentialiste, un nom propre peut dans ce cas être dit modifié si et seulement si, détermination ou non, il opère une désignation « oblique » de son référent, perdant sa fonction de désignation individuelle. Introduite par Kleiber puis approfondie par Jonasson, cette « modification » du nom propre est néanmoins remise en cause par S. Leroy qui s'interroge sur sa pertinence, tout en lui reconnaissant « *le mérite d'avoir été un des instruments majeurs de la réappropriation du nom propre et d'avoir contribué à la description de nombreuses constructions jusqu'alors ignorées* »[Leroy, 2004b, p.74].

Enfin, le nom propre peut être considéré comme une fonction cognitive. C'est l'approche de K. Jonasson qui affirme que « *la vraie nature du nom propre ne se laisse saisir ni au niveau du système linguistique, ni au niveau du discours, mais à un niveau plus profond, à savoir le niveau cognitif* ». La linguistique cognitive, déjà évoquée brièvement en 4.1.2.2, a bénéficié des apports de la psychologie qui a montré que le langage joue un rôle dans les capacités de conceptualisation et de catégorisation. La linguistique cognitive s'emploie dès lors à étudier les rapports entre le langage et la pensée, interrogeant le lien entre perception, expression et conceptualisation de la réalité. Ce courant de recherche a donné lieu, entre autres, à des études de sémantique lexicale en termes de prototype[1] [Kleiber, 1990] et à la théorie des espaces mentaux [Fauconnier, 1984]. K. Jonasson souligne cependant que, tout en étant indispensable à la catégorisation et à la conceptualisation, le langage ne permet pas de rendre compte intégralement de notre expérience perceptive : il est parfois difficile de décrire verbalement un endroit précis ou une personne précise et d'expliquer par un ensemble de critères déterminés la reconnaissance d'une identité perçue. C'est là qu'interviennent les noms propres,

[1] Il s'agit d'une variante du paradigme référentiel : la détermination du référent se fait sur la base d'une ressemblance avec le prototype de la catégorie.

nous permettant de désigner ces réalités particulières. En effet, contrairement aux noms communs, stockés dans notre mémoire à long terme en association avec des connaissances descriptives et nous permettant de regrouper des objets sur la base de propriétés communes, le nom propre nous « *permet d'isoler des entités uniques et spécifiques, en nommant des particuliers perçus à l'intérieur des catégories établies* ». Ainsi, « *le fondement cognitif du nom propre correspond à son association directe dans la mémoire stable à un particulier et non à un concept embrassant un nombre infini d'occurrences particulières* » [Jonasson, 1994].

Il existe donc différents angles d'attaque du nom propre, pouvant être considéré comme une forme, comme un fonction, avec la prise en compte de son rapport à la référence, ou encore sous un biais cognitif. Les travaux résultant de ces différentes approches, s'ils s'emploient à mettre en valeur tel ou tel aspect du nom propre, se concentrent tous, d'une manière ou d'une autre, sur la question de son sens et de son fonctionnement référentiel. C'est ce point nodal du nom propre qu'il importe à présent de préciser.

### 4.2.2 Sens et fonctionnement référentiel du nom propre

Si nous avons expliqué en 4.1 comment le sens conditionne la référence, le nom propre semble difficilement rentrer dans ce cadre en ce qu'il effectue bien une référence mais sans pour autant afficher explicitement de sens. S'il nous est facilement possible d'identifier le référent de *Jacques Chirac*, il nous est effectivement plus difficile de dire en quoi ce nom propre nous permet d'effectuer cette identification. D'où la question : les noms propres ont-ils un sens ? Si oui, de quelle nature est ce sens et, dans le cas contraire, comment rendre compte de son fonctionnement référentiel ? Ce problème a fait l'objet de nombreuses études et, si le débat reste encore largement ouvert aujourd'hui, il est possible d'en expliquer les fondements et d'en retracer les évolutions. Les premières propositions sur le sens des noms propres sont issues de la logique et de la philosophie. En linguistique, la théorie la plus aboutie fut celle de G. Kleiber, avant d'être partiellement remise en cause, sans pour autant être remplacée, tant il semble difficile de rendre compte intégralement, en prenant en compte tous les emplois, de la charge référentielle du nom propre. Il importera d'examiner tout d'abord les thèses logiques, puis de considérer les propositions linguistiques, pour enfin terminer avec certains points essentiels permettant de préciser, pour partie, le fonctionnement référentiel du nom propre.

#### 4.2.2.1 Les propositions logiques ou les fondements du discours théorique sur le sens des noms propres

Les travaux des logiciens sur le sens des noms propres, souvent repris en linguistique[1], ont constitué la base de nombreuses réflexions sur ce sujet. Elle se répartissent en deux positions opposées : le nom propre est vide de sens pour la première, tandis qu'il constitue, au contraire, une description de son référent pour la seconde.

**Le nom propre est vide de sens : Mill et Kripke**

La thèse du nom propre comme vide de sens a été énoncée pour la première fois par J.S. Mill , pour qui « *les seuls noms qui ne connotent rien sont les noms* propres *et ceux-ci n'ont, à strictement parler, aucune signification* » ([Mill, 1843] cité par [Leroy, 2004a]). Ainsi, selon la terminologie logicienne, les noms propres dénotent mais ne connotent rien[2]. Il est en effet impossible de leur attribuer un sens lexical sous forme de conditions d'application ou traits descriptifs spécifiant leur conditions d'emploi ou, pour reprendre notre exemple, le nom propre *Jacques Chirac* ne code aucune propriété que son porteur affiche. Le nom propre fait ainsi figure de simple « étiquette référentielle » pouvant être attribuée à tel ou tel élément du réel, sans contenir en elle-même aucune indication sur cet élément. Si les noms propres n'ont pas de sens, se pose alors le problème de l'identification du référent ou, autrement dit, si *Jacques Chirac* ne nous dit rien de cette personne, comment pouvons-nous l'identifier ? Le logicien S. Kripke [Kripke, 1982] explique alors, en complément de la thèse de Mill, que le lien qui unit un nom propre à un particulier ne repose pas sur un sens lexical mais sur une convention d'un type particulier. Tel objet porte tel nom propre en vertu d'une *chaîne causale* dont l'origine est à chercher dans une *cérémonie de baptême*, à laquelle il faut avoir été initié pour pouvoir user à bon escient le nom propre en question. Il ne s'agit pas bien sûr d'une cérémonie de baptême au sens propre mais plutôt d'une cérémonie de nomination : il suffit d'entendre une personne appeler quelqu'un ou quelque chose par un nom précis pour pouvoir ensuite réutiliser ce nom et le faire

[1] Reprise dont le « danger » a été souligné par M.N. Gary-Prieur [Gary-Prieur, 1994b] qui indique que le nom propre est, pour les logiciens, plus un « moyen » qu'un « objet d'étude » et par J. Molino [Molino, 1982] : « *Les problèmes posés sont d'autant plus complexes et les résultats d'autant plus difficiles à interpréter que, l'initiative étant venue des philosophes et des logiciens, les progrès réalisés dépendent étroitement des techniques et principes logiques qui ont permis de les obtenir ; il est très souvent dangereux de généraliser les solutions à partir du problème très précis pour lequel elles ont un sens.* »

[2] Chez Mill, pour le mot « château » par exemple, la *dénotation* correspond à la classe des châteaux, et la *connotation* à l'ensemble des propriétés nécessaires pour appartenir à cette classe.

connaître à d'autres. N'importe quel élément peut être l'objet, dans une situation ou dans une autre, d'une chaîne causale : il suffit de se mettre d'accord pour attribuer tel nom à telle personne, lieu, bâtiment ou autre pour pouvoir ensuite désigner ces éléments à l'aide de ce nom. C'est donc par le biais d'une convention et non en vertu de traits descriptifs qu'un nom propre désigne tel élément du réel. Le corollaire de cette désignation « conventionnelle » et non descriptive est que le nom propre renvoie à son porteur dans toutes les situations. S. Kripke qualifie alors les noms propres de *désignateurs rigides*, liés de manière fixe à leur référent, quelles que soient les évolutions de ce dernier. En effet, que Jacques Chirac soit ministre, maire de Paris, Président de la République, ou retraité, il porte toujours le même nom. J.S. Mill et S. Kripke ont donc posé que le nom propre est vide de sens, qu'il désigne un particulier en vertu d'une chaîne causale et non en vertu d'un sens et que cette désignation, parce que non remise en cause à chacun des changements affectant le particulier désigné, est une désignation rigide. Ces théories permettent de rendre compte efficacement de l'association d'un nom à un particulier ; on en retrouve un écho certain dans la définition du *Bon Usage* déjà citée, postulant que le nom propre « *se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière* ».

Cependant, pour séduisantes qu'elles puissent paraître, ces thèses peuvent néanmoins être critiquées sur plusieurs points. La théorie des noms propres vides de sens ne permet pas d'expliquer pourquoi des phrases équatives du genre *R. Gary est E. Ajar* ont une valeur informative ni pourquoi des expressions synonymes dans des énoncés tels que *La basilique Sainte Sophie fut élevée à Byzance / Constantinople / Istanbul* ne sont pas employées indifféremment. Elle ne permet pas non plus de rendre compte des emplois modifiés du nom propre (cf. section 4.2.1.4). Enfin, d'un point de vue sémiotique, la thèse du nom propre vide de sens pose le problème du statut de cette unité : est-ce un signe ou non ? Si oui, le nom propre est-il un signe « à une face », un signifiant dépourvu de signifié ? Vides de sens, les noms propres apparaissent tout de même porteurs d'une signification, et cela mérite quelques éclaircissements. La théorie causale de Kripke, pour sa part, ne permet pas de rendre compte de l'existence de plusieurs chaînes causales menant à un même particulier (pseudonymes) ou partant d'un même nom propre (homonymes). À l'opposé des thèses de Mill et Kripke, d'autres approches postulent l'existence d'un sens des noms propres.

**Le sens du nom propre est une description de son référent**

D'autres logiciens et philosophes ont en effet contesté les positions de Mill et Kripke pour au contraire postuler l'existence d'un sens du nom propre, équivalent

à une description de son référent. Il existe deux versions de cette approche, une version forte et une version faible. Dans la version forte, le sens du nom propre est constitué de l'ensemble des attributs du référent ou porteur du nom. Dans cette perpective, le sens du nom *Jacques Chirac* équivaut à la somme des descriptions définies exprimant les divers attributs du porteur, à savoir *l'ancien maire de Paris*, *le enième Président de la République française*, *le fondateur du RPR*, etc. La version faible pose quant à elle que le sens du nom propre correspond à certains traits descriptifs généraux du porteur du nom, comme [+/− masculin], [+/− humain], etc.

La thèse du sens du nom propre correspondant à une description de son référent, dans sa version forte comme dans sa version faible, souffre cependant elle aussi de certaines faiblesses. Surgit tout d'abord le problème du choix des descriptions définies à donner comme sens d'un nom : cherchant à définir notre cher Jacques Chirac, faut-il préférer l'expression « ex-Président de la France », « le mari de Bernadette », « l'ennemi de Nicolas », « le Ministre de l'Agriculture de 1972 », etc. ? Ce problème du choix des attributs à verser au compte du sens des noms propres se double par ailleurs du problème de leur subjectivité (voire de leur vérité, on peut imaginer des descriptions définies qui soient fausses, mais tout de même attribuées à un référent) et de leur nécessaire contingence. Autrement dit, sans avoir été Président de la République, Jacques Chirac se serait tout de même appelé *Jacques Chirac*. Il s'agit là d'une critique avancée par tous, et reprise par K. Jonasson : « *la description constituant la définition du nom propre n'est pas analytiquement vraie pour le référent* ». Une autre illustration de ce fait, proposée par G. Kleiber [Kleiber, 1996], est l'expression métalinguistique de signification : si une phrase telle que « *Une librairie est un magasin où l'on vend des livres* » a une valeur de vérité *a priori* et est paraphrasable par la description définitoire suivante : 'librairie' signifie *'magasin où l'on vend des livres'*, cela n'est plus le cas avec un nom propre tel que Jacques Chirac, dont on ne peut dire 'Jacques Chirac' signifie *'ex-Président de la France'*. Comme le souligne M. Charolles, cette thèse du sens des noms propres comme description de leur référent semble confondre valeur sémantique et connaissances des locuteurs en rapport avec un référent [Charolles, 2002b]. Dire que *untel* possède tel ou tel attribut ne constitue en rien une description du sens de son nom. Ne s'intéressant qu'à l'usage référentiel du nom propre, les auteurs de cette thèse ont, semble-t-il, « *négligé l'existence des noms propres au niveau du système linguistique, pour ne s'occuper que des noms propres dans le discours*" [Jonasson, 1994]. Pour impuissante qu'elle soit à décrire la valeur sémantique des noms propres, cette thèse permet toutefois de rendre compte de l'information véhiculée par les noms propres en discours.

Ainsi, dire que les noms propres sont vides de sens ne permet pas de rendre

compte de leurs évidentes capacités de signification, et dire qu'ils correspondent à une description de leur référent ne revient pas à analyser leur valeur sémantique mais à témoigner de connaissances de nature non linguistique. Si le nom propre n'est pas vide de sens et ne correspond pas à ses propriétes descriptives, comment donc rendre compte, d'un point de vue linguistique, du sens des noms propres ?

#### 4.2.2.2 Les propositions linguistiques

Pour un temps comme paralysée par le discours logique, la linguistique n'a su comment aborder cette épineuse question du sens du nom propre. La thèse du prédicat de dénomination de G. Kleiber [Kleiber, 1981b], élaborée dans le sillage des travaux de J. Algeo [Algeo, 1973], a sereinement éveillé le débat, suscitant critiques et commentaires, ces derniers faisant à leur tour naître de nouvelles propositions. Suivons le cours de ces évolutions.

##### La thèse du prédicat de dénomination

**Première proposition** Dans cette théorie, G. Kleiber nous dit que le sens des noms propres n'est pas à chercher du côté de propriétés descriptives associées au nom propre, mais du côté de la relation entre ce dernier et son porteur, relation dont le caractère est essentiellement *dénominatif*. Tout ce que dit le nom propre, c'est que untel s'appelle, ou est dénommé, *untel*. Le sens du nom propre correspond à l'abréviation d'un prédicat de dénomination *être appelé /N/*. Le sens de Jacques Chirac dans « *Jacques Chirac a renoncé à briguer un troisième mandat* » correspond ainsi à : *le (x) appelé /Jacques Chirac/ a renoncé à briguer un troisième mandat* ». Première précision, concernant le statut du *N* dans cette formule : celui-ci ne correspond pas lui aussi à un nom propre doté d'un sens mais seulement à la chaîne graphique et phonique représentant le nom propre en question. Sans cela, la formule est vouée à une régression infinie (*le x appelé/ le x appelé/ le x appelé...*). Autrement dit, "*le sens du nom propre (...) contient la forme du nom propre qui constitue son signifiant*" (K. Jonasson). Deuxième précision, au regard cette fois-ci du sens à attribuer au verbe *s'appeler* présent dans la formule : l'appellation ne doit pas être prise au sens métalinguistique d'une information sur la langue, comme cela pourrait alors être valable pour les unités lexicales classiques (*Il a vendu le x appelé /canapé/*), mais doit être prise au sens ordinaire d'une information sur le monde, une appellation de type « mondaine ». Il s'agit bien de deux types de dénomination différents, dont seul le second est caractéristique du nom propre. Le prédicat de dénomination, sans remettre en cause la théorie causale de Kripke, permet donc de s'arranger avec l'étrange sen-

timent que les noms propres n'ont pas de sens, postulant que ce dernier existe mais que, sans décrire l'objet dénoté, il se contente simplement de lui conférer un nom. Comme le démontre S. Leroy, il s'agit bien d'un sens, en ce que les exemples suivants :

*Comment s'appelle Kirk Douglas ?*
*Les Albert n'ont pas de nom.*

signalent bien une incompatibilité sémantique entre le verbe *s'appeler* et le nom *nom* d'un part, et la propriété dénominative inhérente aux unités *Kirk Douglas* et *Albert* d'autre part, assimilable à un sens.

La thèse du prédicat de dénomination présente les avantages suivants. Elle permet tout d'abord de rendre compte des emplois référentiels standards du nom propre, à l'instar de l'exemple donné ci-dessus contenant le nom *Jacques Chirac*, et donc d'expliquer pourquoi la phrase équative *R. Gary est E. Ajar* est informative. La personne désignée par le premier nom et celle désignée par le second sont en fait une seule et même personne ; ainsi, « *le gain de connaissance n'est pas que métalinguistique, les noms propres étant liés chacun à un référent, l'énoncé dit quelque chose du monde externe* » [Charolles, 2002b]. De même, il devient possible d'expliquer la différence sémantique entre des énoncés synonymes :

*La Basilique Sainte Sophie fut élevée à 'le x appelé /Byzance/'.*
*La Basilique Sainte Sophie fut élevée à 'le x appelé /Constantinople/'.*

Cette différence se trouve dans les deux modes de donation distincts du référent, passant par la dénomination *Byzance* pour l'un, la dénomination *Constantinople* pour l'autre, différence qui n'interdit aucunement l'identité référentielle, comme c'est le cas ici. Le prédicat de dénomination permet par ailleurs de rendre compte de certains emplois modifiés, comme l'emploi dénominatif : *Il y a une Manon dans ma classe* est paraphrasable par *Il y a un x appellé /Manon/ dans ma classe*. Enfin, il convient de le souligner, cette théorie a le mérite d'exprimer, d'un point de vue strictement linguistique, un sens des noms propres et d'en faire des signes « à deux faces ». Cela est révélateur d'une certaine conception du rapport entre sens et référence, selon laquelle une expression ne peut référer qu'en vertu de son sens ; pour G. Kleiber en effet, « *c'est le sens du nom propre qui déclenche le processus référentiel*[1] ».

**Les critiques** Cette thèse a cependant suscité des critiques dont certaines, reconnues et acceptées par l'auteur [Kleiber, 1995], ont conduit à son remaniement [Kleiber, 2004]. Ces critiques sont nombreuses et intéressent des niveaux variés, c'est pourquoi nous ne rendons compte ici que des plus importantes, laissant le

[1] cité par K. Jonasson.

soin de se reporter aux ouvrages de S. Leroy, M.N. Gary-Prieur et K. Jonasson pour plus de détails. Une première critique concerne l'explication du statut du /N/ contenu dans le prédicat de dénomination, laquelle ne paraît pas satisfaisante, allant jusqu'à être qualifiée de « contre-intuitive » par K. Jonasson : en effet, précisant qu'il ne s'agit que d'une chaîne graphique et phonique, G. Kleiber en est venu à nier également le statut de nom propre à *Paul* ou *Bernard* dans des énoncés dénominatifs tels que *Je m'appelle Paul* ou *Son nom est Bernard*. Cette remarque, acceptée par l'auteur, l'a conduit à réfléchir au statut à accorder au N de tels énoncés dénominatifs : on ne peut effectivement renoncer à y voir un nom propre, mais il n'est pas pour autant assimilable au nom propre standard. G. Kleiber ouvre deux pistes à ce sujet : la première, empruntée à J. Rey-Debove, consiste à voir cette forme comme un autonyme du nom propre, la seconde, inspirée de Van Langendonck, consiste à y voir un *lemme proprial*[1].

Une autre critique consiste à dénoncer l'inadéquation de la paraphrase dénominative pour les emplois modifiés du nom propre. En effet, bien qu'aspirant à proposer un traitement linguistique unifié du nom propre, au travers d'une « *sémantique 'naturelle' dont le but premier est d'expliquer ce que comprennent les usagers* » et donc de proposer des interprétations aux expressions « *effectivement produites en discours*, la thèse du prédicat de dénomination échoue cependant sur ce point, en ce qu'elle ne permet pas de rendre compte de tous les emplois du nom propre, voire, au final, de très peu. C'est le cas de l'emploi vocatif, pourtant de type standard (*C'est vous Charles ?*), et d'autres emplois métaphoriques, métonymiques, exemplaires ou de fractionnement. Revenant sur ce point, l'auteur explique comment même des emplois prototypiques se prêtent difficilement à la paraphrase par une description définie dénominative. Divers arguments sont avancés[2], résumables de la manière suivante : si la description définie *le x appelé /N/* constitue bien une propriété descriptive de dénomination, ce n'est pas une dénomination elle-même. Ce fait est particulièrement visible au travers des deux énoncés suivants :

*Il se trouve qu'il est à la fois (l'homme qui est) le forgeron et l'homme appelé Thomas Smith.*

**Il se trouve qu'il est à la fois (l'homme qui est) le forgeron et Thomas Smith.*

Dont seul le premier est acceptable, alors qu'ils ne se différencient seulement par les propositions *l'homme appelé Thomas Smith* et *Thomas Smith*. Si ces propositions étaient équivalentes, les deux phrases devraient avoir la même grammaticalité. L'auteur relève par la suite le problème du changement de nom qui,

[1] Il est possible de trouver des explications plus approfondies de ce point dans [Kleiber, 2004], p. 122.

[2] Pour une présentation précise, voir pp. 124-129 de [Kleiber, 2004].

entraînant un changement de prédicat dénominatif, devrait pareillement entraîner un changement de propriété du porteur de nom. Or, si *Marie s'appelait Sophie, elle ne serait plus ce qu'elle est (? à savoir 'une Marie')*, ce qui montre bien que le nom propre ne peut se réduire à un prédicat. Un autre argument, longuement développé par G. Kleiber, est issu de l'examen comparatif des deux phrases suivantes :

*L'individu appelé Charles était un grossiste en noms propres.*
*Charles était un grossiste en noms propres.*

Le fait décisif est que la paraphrase prédicative ne saisit le référent que de manière indirecte, « *par une seule de ses facettes, celle d'appellation, comme si précisément nous ne connaissions de lui que la propriété d'être appelé ainsi ou que nous voulions, pour une raison ou une autre, attirer l'attention sur le fait qu'il est nommé ainsi* ». Or cela ne correspond pas exactement à ce que véhicule le nom propre, offrant son référent « d'un seul bloc ». Cette dernière critique sur la paraphrase prédicative, approfondie par G. Kleiber lui-même, a guidé vers une reformulation de la théorie.

**Reformulation : un sens de dénomination** L'auteur, prenant en considération et reconnaissant le bien fondé de l'ensemble de ces critiques a reformulé son hypothèse du sens des noms propres, organisant sa réflexion autour des deux questions suivantes : est-il possible de maintenir un sens pour les noms propres et, si oui, peut-on parler d'un sens dénominatif tout en abandonnant la thèse inappropriée du prédicat de dénomination ? S'attachant à montrer que le nom propre a bien des conditions d'emploi (donc un sens), G. Kleiber explique alors comment ne pas « *jeter le bébé dénominatif avec l'eau du bain prédicatif* » et comment concevoir le sens des noms propres non plus comme un prédicat de dénomination mais comme un sens de dénomination. Un sens dénominatif pour les noms propres est en effet concevable, pour peu que l'on veuille bien admettre qu'un contenu sémantique n'est pas forcément descriptif ou analytique, mais qu'il peut aussi être de nature procédurale, ou instructionnelle. Dans ce cas, « *ce sens dénominatif* [des noms propres] *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* ». Le nom propre, du point de vue de son statut sémantique, se rapproche donc des symboles indexicaux, pouvant lui même être qualifié de symbole dénominatif, *symbole* dans la mesure où l'instruction est conventionnellement attachée au nom propre, *dénominatif* dans la mesure où il s'agit d'un marqueur dénominatif, insistant exclusivement sur la dénomination de telle ou telle entité. Ainsi, de même que *je* peut avoir comme sens instructionnel l'expression *l'individu qui prononce (cette occurrence) de 'je'*,

un nom propre porte comme instruction de prendre en compte dans la mémoire stable l'entité ainsi dénommée. On retrouve, pour le sens des noms propres, et selon l'expression de M.N. Gary Prieur, le « jeu de miroir » entre la forme et le sens propre aux embrayeurs dont la signification renvoie au signifiant, la seule différence étant que pour les embrayeurs, le code renvoie au message, alors que pour le nom propre, le code renvoie au code ; ce qui est au fondement de sa définition, c'est sa forme en tant que dénomination. Il est à noter que « *la paraphrase descriptive de ce sens ne peut plus être considérée comme une expression synonyme, (...) comme les symboles indexicaux, les noms propres sont irréductibles* ». Cette instruction « dénominative » constitue bien du sens dans la mesure où elle pose, comme conditions d'applications, de ne considérer comme référent que l'entité dénommée par le nom en question. De la sorte, le nom propre impose une condition sur le type de référent auquel il renvoie, et comporte bien un sens de dénomination.

Cette nouvelle proposition non prédicative du sens dénominatif des noms propres impose cependant d'aller encore plus loin dans la réflexion, attendu que les noms communs paraissent eux aussi véhiculer cette instruction. G. Kleiber, s'interrogeant alors sur la possibilité de « *sauver un sens de dénomination pour les noms propres* », répond par l'affirmative, soutenant que les noms communs n'ont pas de sens dénominatif dans la mesure où la désignation qu'ils opèrent se fait sur un mode descriptif. Il convient, pour mieux appréhender cette nuance, de préciser ce qu'on entend par dénomination et désignation. La relation de dénomination et la relation de désignation relèvent toutes deux des relations référentielles qui permettent d'établir un lien entre un signe et, pour faire simple, une chose. La dénomination se distingue toutefois de la désignation en ce qu'elle suppose un acte préalable d'attribution d'un nom à quelqu'un ou quelque chose avant de pouvoir désigner, en se servant de cette dénomination, ce quelqu'un ou ce quelque chose, c'est-à-dire l'indiquer avec précision. Noms communs et noms propres ont assurément tous deux un statut dénominatif : l'utilisation du mot 'bougie' pour un tel objet ne peut se faire que parce qu'il a été nommé ainsi, et l'utilisation du mot 'Pierre' pour telle personne ne peut se faire que parce que cette dernière a été nommée ainsi. Ces dénominations peuvent ensuite être utilisées pour désigner une bougie précise ou le Pierre en question, mais ceci en empruntant des chemins différents. En effet, si nom propre et nom commun sont des dénominations, ils diffèrent dans leur mode de désignation ou, pour le formuler autrement, dans leur manière d'atteindre leur référent : le nom commun dénomme et s'appuie sur des attributs caractéristiques du référent auquel il entend renvoyer pour le désigner tandis que le nom propre dénomme et s'appuie sur l'appellation du référent auquel il entend renvoyer pour le désigner. Plus brièvement, le nom commun dénomme et

désigne de manière descriptive, le nom propre dénomme mais désigne de manière dénominative. On a donc bien un sens de dénomination pour les noms propres, mais pas pour le nom commun, même si ce dernier est aussi une dénomination.

**Approfondissement : le nom propre réfère à un particulier** Poursuivant sa réflexion dénominative sur le sens du nom propre, G. Kleiber en vient alors à considérer un fait bien établi au regard de celui-ci, concernant le type d'entité dénommée. Tout locuteur faisant usage de noms propres peut arriver à la conclusion évidente suivante : le nom propre dénomme des entités particulières, c'est à dire considérées dans leur individualité. Effectivement, *Jacques Chirac* dénomme une personne particulière, *Versailles* un lieu particulier et *Felix* un chat particulier. Cette dénomination de particuliers est spécifique au nom propre et ce fait constitue bien souvent la base de l'opposition de ce dernier avec le nom commun qui, pour sa part, dénomme des classes d'objets ou des concepts généraux. Le mot *chat* renvoie à la classe des animaux ayant en commun certaines caractéristiques ('mammifère familier à poil doux, aux yeux oblongs et brillants, à oreilles triangulaires et griffes rétractiles, etc.'), donc à tous les chats possibles ou à la classe des chats en général. Le nom propre est lié à un individu particulier, à un être ou une chose du monde appréhendé comme possédant une unité de caractère et une existence propre tandis que le nom commun est lié à des catégories d'êtres ou de choses. Cette différence d'extension est d'ailleurs inscrite dans leur appellation même avec un nom qui est propre à une entité particulière d'une part, et un nom qui est commun à plusieurs entités d'autre part.

De nouveau, il convient ici de bien marquer la différence entre dénomination et désignation : si le nom propre est spécialisé dans la dénomination de particuliers et le nom commun dans la dénomination de concepts généraux, cela ne signifie par pour autant que leurs désignations diffèrent symétriquement. Les noms communs, tout comme les noms propres, peuvent effectivement opérer une désignation particulière : le mot *chat* dénomme le concept de 'chat' et peut ainsi désigner la classe des chats en général, mais peut aussi bien désigner un chat particulier via *le chat* (dont il est question dans le discours) ou *ce chat*. Ou, pour reprendre l'exemple précédent, la désignation d'un particulier opérée par le nom propre *Pierre* peut tout aussi bien être réalisée *via* par exemple, le syntagme *le boulanger de la Grande Rue*, (désignation d'un particulier sur un mode descriptif). Comme le précise M. Charolles, « *les syntagmes nominaux définis, démonstratifs et indéfinis permettent aux locuteurs d'attirer l'attention de leurs semblables sur les êtres singuliers et non singuliers qui les préoccupent* » ([Charolles, 2002b], p. 60). On retrouve là la question du rapport entre le nom propre et la classe nominale, celui-là trouvant son équivalent dans celle-ci auprès du syntagme nominal

(ou nom commun actualisé) plutot qu'auprès du nom commun[1]. Ainsi, le nom propre dénomme un particulier et désigne par voie dénominative un particulier (le plus souvent, nous reviendrons sur ce point plus précisément par la suite), tandis que le nom commun dénomme un concept général ou une classe d'objets et désigne par voie descriptive, soit ce concept, soit une instance particulière de ce concept.

Cette précision étant faite, revenons au sens du nom propre et à la conséquence sur ce dernier de cette dénomination exclusivement dévolue aux particuliers. Dans ses articles les plus récents sur la question ([Kleiber, 1996] et [Kleiber, 2004]), le sémanticien G. Kleiber montre comment ce statut de particulier typique des entités dénommées par des noms propres constitue en fait une restriction sémantique. C'est ainsi qu' « *au sens instructionnel de dénomination s'ajoute une partie 'descriptive' : ce n'est pas n'importe quelle entité qu'un nom propre nous demande de prendre en considération. Si le sens restreint l'extension des référents à chercher, le fait de postuler que les noms propres dénomment des particuliers équivaut, ipso facto, à postuler un sens ou du sens pour les noms propres, qu'on le veuille ou non* ». Pour les noms propres il y aurait ainsi, en plus d'un sens instructionnel ou procédural, un sens descriptif dessinant, nous pourrions dire par approximation plus que par détermination précise, le référent visé.

Cette référence à un particulier que G. Kleiber inscrit dans le sens même du nom propre, selon une approche en termes de sémantique référentielle désormais familière, fait figure d'écho sémantique (ou cela peut être l'inverse) au pivot cognitif qu'est la notion de perception d'une individualité. En effet, selon K. Jonasson, promotrice de l'approche cognitive du nom propre, le rôle du nom propre dans la structuration de notre expérience perceptive est de nous permettre « *d'isoler des entités uniques et spécifiques* » et sa fonction cognitive fondamentale est « *de nommer, d'affirmer et de maintenir une individualité* ». Ceci s'explique par le fait que nous éprouvons le besoin de parler des choses, des êtres, des événements et autres phénomènes qui nous entourent en tant que particuliers et que, s'il nous est possible d'en parler singulièrement à l'aide de syntagmes nominaux, il est certaines de ces choses, êtres, etc. avec lesquels, nous dit M. Charolles [Charolles, 2002b], « *nous entretenons un commerce suffisamment régulier et pointilleux pour que nous éprouvions le besoin de leur attacher un nom particulier capable de les désigner sous toutes leurs apparences* ». Il est possible d'aller plus loin encore, et de dire que si le nom propre réfère à un particulier, il ne s'agit pas d'un particulier en tant que particulier, mais d'un particulier en tant que saisi dans une

[1]Comme l'indique S. Leroy, « *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* » [Leroy, 2004a], p.27.

catégorie conceptuelle. K. Jonasson parle de « *particuliers établis à l'intérieur de catégories établies* », ce qui s'explique fort bien à l'aide du « commerce régulier et pointilleux » de M. Charolles qui se fait toujours avec les « mêmes catégories d'existants » qui nous importent. Elles nous permettent de structurer nos rapports aux autres, à l'espace, au temps, etc. et c'est ainsi que les noms propres sont le plus souvent dévolus aux personnes, aux lieux et aux autres éléments permettant de nous y retrouver avec la réalité. Ces catégories d'existants évoluent bien sûr selon les époques et les cultures : à l'heure d'internet, nous éprouvons moins, pour naviguer, le besoin de nommer et de connaître les étoiles que celui de nommer et de connaitre les noms de sites et autres moteurs de recherche. G. Kleiber explique les choses de manière moins pragmatique et à un autre niveau : « *nous ne concevons pas un particulier ou une entité individuelle comme étant simplement un particulier ou une entité individuelle. Les particuliers sont toujours conçus comme des particuliers d'un certain type (...), si l'on essaie de se représenter un fruit ou un animal, ce ne peut être qu'un fruit ou un animal déjà particulier* ». La conceptualisation catégorielle est indispensable à la perception de particuliers, qui ne peuvent se différencier que par rapport à d'autres. Quel que soit le point du vue adopté sur le nom propre, référentiel ou cognitif, cette notion de particulier, assimilable en quelque sorte à la nature de l'entité dénommée et désignée par un nom propre, constitue de toute évidence un point essentiel du fonctionnement référentiel du nom propre.

En définitive, les travaux de G. Kleiber présentés ci-dessus permettent de dégager un sens proprement linguistique du nom propre. Le sens d'un nom propre ne se présente pas sous la forme de traits objectifs qu'une entité devrait posséder pour pouvoir être par lui désignée mais sous la forme d'une « *instruction de chercher et de trouver dans la mémoire stable le référent qui porte le nom en question* ». Il s'agit donc d'un sens procédural dont l'instruction repose exclusivement sur une dénomination. A ce sens procédural peut être ajouté un sens descriptif comprenant comme condition d'application pour le référent le trait 'particulier'. Afin d'apprécier la portée de la théorie présentée par G. Kleiber, apportons au final une dernière indication : la thèse proprement linguistique du sens de dénomination du nom propre trouve un écho dans le discours des grammairiens avec la mention, dans la *Grammaire méthodique du français* [Riegel *et al.*, 1994], d'un « sens » (avec des guillemets cependant) pour le nom propre *Pierre*, paraphrasable par « entité qui s'appelle *Pierre* » (p. 176). Cette nuance, déjà évoquée plus haut (cf. section 4.2.1.3), est révélatrice d'un changement d'appréciation de la part de la linguistique vis-à-vis du nom propre. Il convient de se reporter à [Kleiber, 1981b, Kleiber, 1996, Kleiber, 2004] pour plus de précisions concernant les tenants et aboutissants de ce sens dénominatif suffisamment, nous l'espérons,

présenté ici.

Pour poursuivre cette revue des propositions linguistiques sur le sens des noms propres, considérons rapidement d'autres approches qui, plus récentes, tentent également de caractériser le fonctionnement sémantique du nom propre.

**Comment expliquer la 'charge référentielle' du nom propre ?**

Si la thèse du sens de dénomination permet, dans une certaine mesure, de sortir de l'impasse d'une étiquette vide de sens sans pour autant prendre la direction du « trop plein de signification » et de proposer un sens véritablement linguistique pour le nom propre, elle ne permet pourtant pas de rendre pleinement compte de sa valeur sémantique et d'expliquer pourquoi il renvoie « *à une série indéfinie d'interprétants, qui nous apparaissent comme plus riches, plus chargés d'affectivité que les interprétants évoqués par les noms communs* » [Molino, 1982]. D'autres propositions, sans pour autant rejeter la théorie formulée par G. Kleiber, s'attachent ainsi à combiner ces deux intuitions contraires (étiquette vide et étiquette pleine) et à proposer d'autres types de sens pour le nom propre afin de rendre compte de l'information sémantique non proprement linguistique attachée à ce dernier.

**Le contenu du nom propre** Selon une perspective toujours linguistique, [Gary-Prieur, 1994a] introduit la notion de « contenu » du nom propre. La profession de foi de l'auteure augurant son ouvrage fait état d'un objectif légèrement décalé par rapport aux travaux antérieurs : plus que de chercher une solution à la question du *sens* du nom propre, solution nécessairement tributaire d'une théorie sur le sens, il s'agit pour M.N. Gary-Prieur d'analyser la spécificité du *fonctionnement sémantique* du nom propre, jouant selon elle sur deux niveaux avec « *un élément régulier fixé dans la langue et des éléments variables qui se définissent en fonction de chaque énoncé* ». Reconnaissant l'efficience de la proposition de G. Kleiber pour rendre compte du fonctionnement sémantique du nom propre au niveau lexical, avec un sens dénominatif spécifiant en quelque sorte des conditions d'application et caractérisant le nom propre « *en tant qu'unité de la langue* », elle choisit de le compléter d'une dimension référentielle en définissant ce qu'elle appelle le contenu du nom propre. Ce dernier correspond à « *un ensemble de propriétés du référent initial qui interviennent dans l'interprétation de certains énoncés contenant ce nom* ». Le référent initial correspond, « *dans un énoncé,* [à] *l'individu associé par une présupposition à cette occurrence du nom propre en vertu d'un acte de batpême dont le locuteur et l'interlocuteur ont connaissance* ». Les propriétés de ce référent initial, par le biais duquel on retrouve la notion de

chaîne causale, sont des propriétés en relation avec le contexte, définies dans le cadre de la situation d'énonciation et des connaissances des interlocuteurs dans cette situation précise.

Ces propriétés sont à différencier tout d'abord des connaissances encyclopédiques qui « *servent à établir l'identité du référent initial* » : ce ne sont pas des connaissances préétablies construites en dehors du discours, elles correspondent à des connaissances discursives qui, « *certes empruntée(s) à la connaissance encyclopédique, ont leur source dans le discours lui-même* » et correspondent à l'« univers de croyance »[1]. Elles ne sont pas assimilables non plus à ce qu'on appelle des 'connotations', lesquelles véhiculent un autre type de sens, plus subjectif car lié à l'expérience personnelle des locuteurs. Si les connotations visent à rendre compte de phénomènes sémantiques qui ne relèvent pas de la dénotation, le contenu vise à « *compléter la représentation du sens des noms propres en termes de prédicat de dénomination* ».

Le contenu du nom propre, par ses propriétés, semble donc inscrire une dimension référentielle dans le fonctionnement sémantique du nom propre, complétant ainsi le sens dénominatif de nature instructionnelle inapte à expliquer, dans certains cas, l'interprétation du nom propre. Reprenons l'exemple mentionné par S. Leroy (lui même emprunté à Gary-Prieur) : si, dans la phrase *Montand était devenu Montand*, le prédicat de dénomination, ou sens de dénomination, permet bien d'expliquer le sens de la première occurrence du nom propre qui enjoint à retrouver dans la mémoire stable le référent portant ce nom, il ne permet pas de rendre compte du sens de la seconde, pour laquelle il est alors préférable de convoquer le contenu, permettant d'interpréter *Montand* comme 'le chanteur-acteur celèbre et engagé'.

**Le 'sens encyclopédique' du nom propre** M. Charolles, sans distinguer autant de nuances que M.N. Gary-Prieur, parle pour sa part de « sens encyclopédique » des noms propres. Ne cherchant pas à intégrer « coûte que coûte » l'ensemble des informations attachées à un nom propre au *sens* même du nom propre, opération qui semble difficile (témoins les guillemets employés par l'auteur), M. Charolles explique comment, dès lors qu'un nom propre peut "*servir de point de repère existentiel* », il finit par « *se charger de sens encyclopédique* ». Vide de sens descriptif, le nom propre ne donne aucune indication précise quant à son porteur mais est appelé à « *collationner l'ensemble des données se rapportant au pariculier qu'il désigne* ». Ce « sens encyclopédique » des noms propres permet ensuite d'expliquer emplois métaphoriques et inférentiels. Contenu ou sens

[1]Notion empruntée à Martin : l'univers de croyance du locuteur est l'ensemble des informations qu'il tient pour vraies au moment de l'énonciation.

encyclopédique, il est des informations indispensables à la compréhension du nom propre que l'on hésite toutefois à 'injecter' dans le sens de ce dernier.

**Le cadre classificateur** D'autres approches encore, mentionnées par K. Jonasson ([Jonasson, 1994], p.121), proposent d'expliquer l'information que le nom propre véhicule à propos de son porteur par le biais d'un « cadre classificateur ». Ce dernier exprime comment des informations de nature non lexicale, donc non inscrites dans le code linguistique, peuvent tout de même être associées de manière régulière à un nom propre. L'assignation et l'utilisation de noms propres se faisant au sein d'une société donnée, tout un système de conventions sociales et culturelles interviennent dans le fonctionnement sémantique du nom propre. Ces conventions nous permettent ainsi de prédire, à propos de noms tels que *Julien* et *François*, *Marie* et *Eléonore*, *Médor* et *Mistigri* ou encore *Super Étendard* et *Rafale*, qu'ils ont probablement pour référents, respectivement, des hommes, des femmes, des animaux domestiques et des avions.

Informations référentielles avec le contenu, sociales et culturelles avec le cadre classificateur, personnelles et subjectives avec les connotations, le fonctionnement sémantique du nom propre convoque de toute évidence des éléments divers pour permettre son interprétation. On peut néanmoins s'interroger sur le statut de ces informations : peut-on les considérer comme participant d'un sens lexical ? est-il convenable de les inscrire dans le code linguistique ? Les auteurs de ces propositions eux-même hésitent à parler de sens lorsqu'ils évoquent ces aspects entrant dans la compréhension et l'interpétation du nom propre. Gary-Prieur parle de « niveau » de sens, Charolles use de guillemets et Jonasson les introduit en tant que « types » de sens. Quoiqu'il en soit, l'ensemble de ces réflexions sur la charge référentielle du nom propre apportent un précieux supplément de précisions sur certains aspects du fonctionnement sémantique du nom propre, précisions complémentaires du sens de dénomination dégagé par G. Kleiber.

La question du sens des noms propres a donc, depuis fort longtemps, fait couler beaucoup d'encre, donnant naissance à de nombreuses propositions. L'examen de ces dernières, chacune apportant un éclairage pertinent sur la question, a fait apparaître la complexité de ce que l'on pourrait appeler une sémantique du nom propre. Ce dernier, pour sa compréhension et son interprétation, convoque de toute évidence diverses couches de signification, allant d'informations proprement lexicales, avec le sens de dénomination d'un particulier, à des informations relevant moins ou pas du tout du code linguistique, informations de nature référentielle, subjective, socio-culturelle, etc. S'il plane parfois, au dessus de la diver-

sité des propositions dont aucune ne semble exempte de critiques, un « sentiment d'insatisfaction » (le mot est de G. Kleiber) à propos de cette difficile question du sens des noms propres, il est important de souligner la réalisation de progrès notables en la matière : assimilant et dépassant le discours des logiciens, la linguistique est aujourd'hui capable, et ceci apparaît comme primordial, de proposer un sens lexical pour les noms propres, sous la forme d'un sens instructionnel de dénomination doublé d'un sens descriptif imposant la caractéristique 'particulier' à l'entité désignée, permettant de comprendre comment la route est tracée vers le référent, tout comme elle est à même de rendre compte, *via* l'élaboration de diverses notions (contenu, cadre classificateur), de la signification du nom propre. Il s'agit bien d'une « réappropriation » du nom propre en linguistique passant par l'analyse et l'approfondissement des deux intuitions premières à ce sujet avec, d'une part, la thèse du nom propre vide de sens laissant la place à un sens de dénomination et, d'autre part, celle du sens du nom propre comme une description de son référent laissant la place, de manière peut-être moins convainquante, au « contenu » ou « sens encyclopédique ». Chemin faisant, ces réflexions ont également permi d'analyser et de mettre en valeur certains aspects du fonctionnement référentiel du nom propre, pour lesquels nous souhaitons, dans un dernier temps, apporter quelques spécifications .

#### 4.2.2.3 Dénomination et référence à un particulier : quelques précisions

Nous avons vu que, contestant l'idée des noms propres comme vide de sens, G. Kleiber lui a opposé celle d'un sens de dénomination, tout en approuvant et reprenant la notion d'acte de baptême issue de la théorie causale. Dans cette perspective, les noms propres ne sont pas dotés d'un sens lexical codifié en règles de dénotation spécifiant leur conditions d'application mais d'un sens « mi-instructionnel, mi-descriptif », donnant pour instruction de chercher un référent portant le nom en question et étant un particulier. Nous nous intéressons dans un premier temps à la première « facette » de ce sens, la facette instructionnelle, avant de donner quelques précisions quant à la seconde dans un second paragraphe.

**Une dénomination plus ou moins descriptive**

Il s'agit ici d'attirer l'attention sur le rôle de la dénomination dans l'assignation du statut proprial à tel ou tel nom ainsi que son importance dans l'interprétation de l'unité (devenue) nom propre. Le sens instructionnel du nom propre nous impose de chercher un référent portant ce nom, puisque ce dernier, justement, ne nous dit rien, donc que cela. Or est-ce toujours le cas ? Dans la première section

sur la difficile définition du nom propre (cf. 4.2.1.4), nous avons évoqué la partition syntaxique entre noms propres purs et noms propres à base descriptive ou mixtes. Les premiers, prototypiques de la catégorie, regroupent des noms propres qui n'évoquent aucune caractéristique de leur porteur (*Paul, Paris*), tandis que les seconds regroupent des noms propres composés, tout (à base descriptive) ou partie (mixtes), de noms communs (*L'Assemblée Nationale, le Massif Central, le Jardin des Plantes*). C'est à l'encontre de ces derniers que se pose la question du rôle de la dénomination, que l'on serait tenté de qualifier de plus ou moins descriptive. En effet, le *Massif Central* est bien un ensemble de montagnes situé au centre de la France et le *Jardin des Plantes* est bien un jardin, ou terrain clos, où l'on trouve des plantes. Les éléments lexicaux composant ces noms ont donc de toute évidence un sens descriptif, qui véhicule des caractéristiques des référents auxquels ils s'appliquent. Toutefois, il nous semblerait tout à fait impossible d'utiliser (entre locuteurs francais) ces noms pour un autre massif montagneux au centre d'une autre aire géographique ou un autre jardin comportant lui aussi des plantes (comme tous les jardins au demeurant), sans au moins préciser qu'on parle d'un autre massif que le Massif Central en France ou d'un autre jardin des plantes que celui du cinquième arrondissement de Paris. Dans le même ordre d'idées, a-t-on jamais à l'esprit le fait que le *Pont Neuf* date de 1604 et n'est donc plus si neuf que cela ? La descriptivité des noms composant ce qu'on appelle les noms propres mixtes ne semble ainsi plus de mise pour le repérage du référent dénoté. Même composé d'éléments dotés d'un sens lexical, ce qui est essentiel pour le nom propre est la dimension dénominative, prenant le pas sur une dimension descriptive probablement efficiente à l'origine mais dont les locuteurs ont progressivement perdu conscience. K. Jonasson indique à ce propos qu' « *à la suite d'un acte de dénomination ou d'un emploi répété du nom propre comme expression référentielle, ou désignateur, associé à l'entité particulière en question, un lien plus direct a pris la relève, laissant se reculer, retirer, ou s'effacer le sens lexical descriptif devenu désormais superflu* ». Ce qui semble jouer dans l'assignation du statut proprial ou non pour un nom (et donc pour la délimitation de la catégorie du nom propre), c'est, pourrait-on dire, la force de la dénomination (appréciable *via* l'usage), laissant « reculer » ou au contraire totalement « s'effacer » le sens lexical. Les onomasticiens qualifient ce phénomène de « désémanticisation » ou « seuil du nom » ; il nous paraît cependant important de souligner que le nom ne se « vide » pas de son sens pour devenir *propre* à une entité particulière, il ne se désémanticise pas mais change simplement de sémantisme, passant progressivement de descriptif à dénominatif selon des dégrés variables. C'est ainsi que la *Côte d'Azur*, *Terre-Neuve* et la *Forêt Noire* pourront plus facilement être qualifiés de noms propres que l'*Académie Francaise* et le *Centre National de la Recherche*

*Scientifique*, unités pour lesquelles la dimension descriptive joue encore un rôle dans leur interprétation, en dépit de la stabilité de la dénomination qu'elles effectuent. La dénomination, qui établit un lien direct entre le nom et le référent et suspend peu ou prou la description, est donc capitale dans le fonctionnement référentiel du nom propre. Complétant ce sens instructionnel de dénomination, un sens descriptif pose par ailleurs comme condition d'application de l'entité dénotée par un nom propre le fait d'être un particulier, condition qui invite à quelques précisions. Après la spécificité de la dénomination du nom propre, il convient donc d'examiner celle de sa désignation.

### La notion de particulier : unité et unicité

Il importe ici d'attirer l'attention sur ce que signifient précisément la notion de particulier d'une part, et celle (liée à la première) d'unicité référentielle d'autre part.

**Le principe d'individuation** Cela a déjà été souligné, nom propre et nom commun sont tous deux des dénominations, mais se différencient en ce que la dénomination répond, pour celui-là, à un besoin d'individualisation d'une entité par rapport à d'autres et, pour celui-ci, à un besoin de classement d'une entité au sein d'autres similaires. Un nom commun sanctionne ainsi une appartenance à une catégorie tandis qu'un nom propre établit et met en avant un particulier considéré dans son individualité. C'est ce point qu'il convient de considérer : que faut-il comprendre exactement par 'particulier considéré dans son individualité' ? Selon le *Trésor de la Langue Française*, un particulier est un individu, soit un « être concret, donné dans l'expérience, possédant une unité de caractères et formant un tout reconnaissable ». Un nom propre s'intéresse donc à n'importe quel objet ou entité de la réalité appréhendable en tant qu'unité irréductible, et c'est cette dimension unitaire propre à ce qu'on a appelé 'particulier' qui, en même temps qu'elle motive la nomination par un nom propre, est renforcée par cette nomination. Soulignant la différence de conceptualisation entre nom propre et nom commun, P. Siblot résume ce mouvement d'individualisation par la formule suivante : « *Instrument et sanction d'une promotion à l'individualité, la fonction spécifique* [du nom propre] *est de réaliser une 'identification individualisante', foncièrement différente de l''identification catégorisante' du nom commun* »[Siblot, 1995]. Portant attention au même phénomène, G. Kleiber parle pour sa part de *catégorisation individuante*, mettant en avant le fait que la notion d'individu particulier « *constitue un principe organisateur, un concept rassembleur d'instances considé-*

*rées comme les instances d'un même individu*[1] ». Partant, le nom propre « *établit lui-même une 'catégorie', celle de l'individu* [Siblot, 1995].

Le fait à considérer est que cette constitution d'individu, constitution par extraction d'une entité (identification individualisante) ou par rassemblement d'occurrences d'une même entité (catégorisation individuante), est relative au niveau d'appréhension : tel homme parmi les hommes et dénommé *Jacques Chirac* est considéré comme particulier parmi les hommes, de même que tel modèle de voiture parmi les modèles de voiture et dénommé *Clio* est considéré comme particulier parmi les modèles de voitures. Ce qui est troublant dans ce dernier cas de nom de produit (duquel on peut rapprocher le nom de marque), c'est qu'en même temps qu'il y a extraction individualisante (tel modèle de voiture parmi les modèles de voitures), il y a partage de propriétés avec d'autres entités semblables (toutes les voitures *Clio* sont de petite taille, ont une ligne arrondie, une bonne maniabilité, etc.). Quel que soit le niveau considéré et la nature de l'entité, ce qui compte c'est, nous dit V. Descombres dans un article s'intéressant aux individus collectifs, « *la possibilité d'indiquer un 'principe d'individuation'* » [Descombres, 2001]. La notion d'individu particulier, relativement au nom propre, est à rapprocher de la signification initiale du terme *individu* présente chez les philosophes et les logiciens et non de celle en cours dans notre usage commun : « *La philosophie de la logique appelera* individu *tout ce qui est susceptible d'une individuation, c'est-à-dire d'une différienciation donnant lieu à un dénombrement. Par conséquent, on a des individus partout où, dans un genre de chose donné, on peut dénombrer, dire s'il y a* un *ou* plusieurs *échantillons du genre considéré* ». L'individuel ne s'oppose pas au collectif mais au général et, de la sorte, rien ne justifie de réserver la notion d'individu aux seuls êtres humains ou aux seules entités « unaires » : une personne est un individu (*Jacques Chirac* se distingue de la catégorie générale des autres hommes), mais un modèle de voiture également (*Clio*, appellation rassemblant plusieurs voitures, se distingue des autres modèles de voitures). Ainsi, en fonction de la compréhension que l'on peut avoir de la notion d'individu particulier, on comprend mieux pourquoi on hésite parfois à considérer comme noms propres certains noms désignant des entités plurielles (collectifs humains, noms de produits, de modèles, etc.), hésitation conduisant soit à leur exclusion de la catégorie propriale, soit à leur qualification d'« hybrides » [Laurent et Vicente, 2004]. Individu individuel ou individu collectif, il n'en reste pas moins que le référent dénommé et désigné par le nom propre fait l'objet d'une appréhension unitaire.

[1] Ces instances d'un même individu sont des « tranches spatio-temporelles », donnant à voir Paul faisant du vélo, Paul faisant à manger, Paul dormant, etc. Le rôle du nom propre est de faire abstraction des différences de ces occurrences et de les rassembler en un *individu*. cf. [Kleiber, 1996], p. 585-586.

**L'unicité référentielle** Si le principe d'individuation (ou la catégorisation individuante) constitue une première composante permettant de préciser la notion de particulier dans son « rapport à soi », il convient d'en examiner une seconde, l'unicité référentielle, précisant cette même notion dans son « rapport aux autres ». Si unité n'est pas unicité, ces deux composantes vont cependant de pair lorsqu'il est question du nom propre, donnant toute sa résonnance à la « référence à un particulier » opérée par ce dernier. Le nom *Jacques Chirac* permet de rassembler les occurrences spatio-temporelles d'une entité (notre ex-Président) et de la fixer en tant qu'homme-individu, tout comme il permet de la distinguer des autres entités (les autres hommes) et de la fixer en tant qu'unique. De même, mais à un autre niveau, le nom *Clio* permet de rassembler toutes les occurrences de voiture d'un certain type et de les fixer en tant que modèle-individu, tout comme il permet de le distinguer (le modèle) des autres modèles et de le fixer comme unique (la ligne arrondie de toutes les *Clio* s'oppose à celle dynamique de toutes les *Saxo*, par exemple). Un point important est à souligner : cette unicité référentielle ne vaut pas dans l'absolu mais en contexte, dans une situation de discours. Plusieurs chaînes causales peuvent en effet lier un même nom à différents référents (*Clio* modèle automobile ou déesse de l'Histoire) et un même référent peut être la cible de différentes chaînes causales (*Henry Beyle* et *Stendhal*). L'unicité référentielle dans l'absolu n'est valable que pour ce qu'on appelle les *unica*, comme le *soleil*, la *lune* ou la *terre*. On peut alors légitimement s'interroger sur l'hésitation courante qu'il y a à les considérer comme des noms propres : s'il sont uniques, pourquoi ne sont-ils pas noms propres ? C'est, nous dit Kleiber, en raison de l'absence de saisie hiérarchique, « *ils ne sont pas appréhendés comme des individus appartenant à une classe conceptuelle supérieure (...). Mars est saisi comme une planète et le nom propre marque opaquement son unicité au sein de cette catégorie, alors que la terre et le soleil apparaissent comme des entités uniques au sein des choses du monde - ce sont, en ce sens, des 'unica'* ». On retrouve donc le fait que les noms propres désignent des particuliers déjà catégorisés, c'est-à-dire pouvant se détacher et être dits uniques par rapport à d'autres. Ainsi, lorsqu'on dit qu'un nom propre réfère à un particulier, cela sous-entend une référence à une entité appréhendée de manière unitaire (de façon à en faire un *individu*) et détachée de ses semblables ou d'une classe conceptuelle (de façon à en faire un individu *unique*).

Dénomination et référence à un particulier, il est à noter que ces caractéristiques du nom propre peuvent être acceptées comme plus ou moins catégoriques, en fonction du degré de figement de l'appellation et du niveau d'appréhension de l'individualité du particulier désigné. On retrouve ici les principales oscillations du nom propre, lesquelles conduisent généralement à parler de *continuum* entre

ce dernier et le nom commun, révélant, comme le précise J.L Vaxelaire, qu' « *il n'y aurait que des degrés de 'propritude' ou de 'communitude' qui sépareraient les noms* » [Vaxelaire, 2007].

Si cette appellation a pu un temps être controversée, la *catégorie linguistique du nom propre* trouve aujourd'hui tout son sens, au terme d'une longue histoire débutée avec les logiciens, mise sous silence avec les structuralistes, puis reprise et continuée avec les sémanticiens faisant, à la suite de Frege, la part belle à la référence. Les travaux de ces derniers ont mis en lumière divers aspects du sens et du fonctionnement référentiel du nom propre avec, d'une part, un sens mi-instructionnel mi-descriptif ainsi que divers cadres d'analyse de sa charge référentielle et, d'autre part, des explications et précisions quant à la notion de particulier, tant d'un point de vue cognitif que sémantique. Dans notre exploration des catégories linguistiques existantes pouvant aider à une définition des entités nommées, il convient à présent de considérer les descriptions définies.

## 4.3 Les descriptions définies

Dans l'échantillon des unités lexicales considérées comme des entités nommées traçé en 3.3.2 (figure 4.1, p.102), on trouve des entités telles que *le Président de la République*, *le Président du Conseil*, *le Décret de loi 31/3 de 2005*, etc. assimilables à ce qu'on appelle des descriptions définies. Notre objectif dans cette section, à l'instar de la précédente sur le nom propre, est d'examiner les propriétés caractéristiques de ces dernières afin de dégager, si possible, des éléments de définition pour les entités nommées. Si l'on a pu aborder la question du nom propre en soulignant ses difficultés d'appréhension, il serait possible de faire de même à l'encontre des descriptions définies pour lesquelles il est également malaisé de trouver une caractérisation précise et consensuelle. En témoignent les propos de G. Kleiber, débutant sa partie sur les descriptions définies par des « problèmes de définition » autour du « *terme de 'description définie' (qui) se prête à des interprétations et à des usages variés parce qu'il correspond à un concept hétérogène* » [Kleiber, 1981b] ; tout comme ceux de L. Linsky au terme de son étude sur la théorie de B. Russell : « *Nous avons supposé tout au long de ce chapitre qu'il n'y pas de problème au sujet de ce que signifient les mots 'description définie'. Car en comparaison de tout ce qu'il peut avoir à nous dire sur les descriptions définies, Russell a peu de choses à dire sur ce qu'elles sont.* »[Linsky, 1974]. En effet, sur ce terrain des descriptions définies se croisent de nouveau logique et linguistique, champs disciplinaires évoquant avec plus ou moins d'insistance, au

sujet des descriptions définies, divers critères tels que la dénotation unique, la présence de l'article défini, le renvoi à un référent connu, ou encore une plus ou moins grande autonomie référentielle. La présente section sur les descriptions définies débutera donc par un questionnement sur ce que recouvre précisément le terme *description définie*, examinant notamment les travaux fondateurs du logicien philosophe B. Russell, puis se terminera par des éléments d'analyse plus proprement linguistiques avec une considération du sens et du fonctionnement référentiel de ces unités.

Afin de donner corps à cette section, nous souhaitons avant tout établir comme point de départ une liste de syntagmes et propositions nominales au sein de laquelle nous tenterons progressivement de déterminer ce qui est ou non description définie. Cette liste, reprenant des exemples connus, est la suivante :

*la baleine (est un mammifère)*
*la baleine (a heurté le bateau)*
*l'imagination*
*l'imagination de Paul*
*le fer*
*le père de Charles II*
*le Roi de France*
*l'actuel Roi de France (est chauve)*
*le cercle carré (n'existe pas)*
*le Président*
*le Président de la République*
*le Président de la République Française en 2008*
*la voiture immatriculée 478 KNB 75*

Deux remarques ou questions affleurent immédiatement au vu de ces exemples : premièrement, il est des expressions fausses (apparemment) ou, dit autrement, pour lesquelles il n'y a pas de dénotation, soit qu'elle n'existe pas, soit qu'elle soit impossible (*L'actuel Roi de France, le cercle carré*) ; deuxièmement, certaines de ces propositions paraissent renvoyer à un référent plus ou moins générique (*la baleine est un mammifère/a heurté le bateau*) et/ou de manière plus ou moins précise (*le Président de la République, le Président de la République Française en 2007*). Ces observations nous donnent déjà une idée des questionnements des uns et des autres et de la définition à venir.

### 4.3.1 Qu'est-ce qu'une *description définie* ?

Si l'on s'en tient à ce que les mots nous disent, une « description définie » correspond à un nom (N) précédé d'un article défini (*le*, *la*, *les*), donc, typiquement, à un groupe nominal en *le* + *N*[1]. Il n'est toutefois pas tout à fait juste d'assimiler complètement les descriptions définies aux expressions nominales définies dans leur ensemble et il importe, afin de caractériser au mieux le type d'unités lexicales entrant dans le champ de cette appellation, d'ajouter d'autres éléments de définition à ces critères formels. Pour ce faire, il convient tout d'abord de revenir au questionnement historique de B. Russell qui, d'un point de vue logique, a permis de poser les principaux éléments d'analyse des descriptions définies, puis de considérer les restrictions d'ordre linguistique apportées par G. Kleiber à cette définition logique.

#### 4.3.1.1 La théorie des descriptions définies de B. Russell

B. Russell, philosophe et mathématicien anglais de la première moitié du vingtième siècle, a profondément marqué la logique mathématique, la théorie de la connaissance ainsi que la philosophie du langage. Nous ne considérons ici qu'une partie de ses travaux, avec un aperçu de ce qu'on appelle la « théorie des descriptions définies ». Dans un article célèbre intitulé *On denoting*, publié en 1905 dans la revue de philosophie *Mind* [Russell, 1905], B. Russell s'intéresse aux « expressions dénotatives » (*denoting phrases*) avec pour intention de pallier les insuffisances, selon lui, de la théorie des objets d'A. Meinong (cf. ci-dessous) et de celle de la référence de G. Frege. Nous tenterons, en nous appuyant principalement sur l'article de 1905, d'exposer au mieux le raisonnement de B. Russell à propos des « expressions dénotatives », en présentant successivement : les enjeux de la théorie des expressions dénotatives, les objets auxquels elle s'applique, les énigmes référentielles qu'elle entend résoudre, la théorie elle-même et, au final, ce qu'il est intéressant, au regard des descriptions définies puis des entités nommées, de retenir.

**Les enjeux** L'approche de B. Russell s'inscrit dans le contexte de la logique moderne, dont la philosophie sous-jacente comprend une certaine conception du langage, de la signification et de la vérité [Hottois, 2002]. Ce n'est donc pas en linguiste que B. Russell s'intéresse aux descriptions définies, mais en logicien, se posant la question de la signification des symboles en termes de connaissance et

[1]Dans les langues sans article défini (comme le russe ou le chinois), ce type de dénomination peut également être exprimé, mais non sous cette forme précise de *le* + *N* à l'origine de l'appellation « description définie ».

de vérité. Son intérêt, tel qu'exposé dans son article, porte plus largement sur la *connaissance*. Relativement à la théorie de la connaissance, Russell introduit deux concepts, la connaissance directe (*knowledge by acquaintance*) et la connaissance par description. La première rend compte de notre manière de connaître directement les choses avec lesquelles nous pouvons entrer en contact et la seconde notre manière de connaître indirectement les choses pour lesquelles nous ne pouvons avoir d'expérience directe et que nous ne connaissons que par description. L'auteur prend l'exemple du centre de la masse du système solaire à un instant précis, à propos duquel nous pouvons affirmer un certain nombre de choses (*via* des descriptions), sans pour autant en avoir une connaissance directe (*via* nos sens). Son intention est alors d'expliquer la combinaison de ces deux modes de connaissance : « *All thinking has to start from acquaintance ; but it succeeds in thinking* about *many things with which we have no acquaintance* ». Par ailleurs, relativement au problème du sens et de la référence, Russell tente d'expliquer comment il est possible de dire et de comprendre ce qui n'est pas, sans pour autant abandonner un « solide sens de la réalité » (cité par [Marconi, 1997]).

**Les objets d'étude** Dans son analyse, le logicien prend pour objet d'étude ce qu'il appelle les « denoting phrases », soit des expressions qui ne sont pas des noms propres et qui peuvent être le sujet grammatical d'une phrase. Parmi ces expressions, il opère une triple distinction : il est des expressions dénotatives qui ne dénotent rien (*l'actuel Roi de France*, qui ne renvoie à rien), d'autres qui dénotent un objet défini (*l'actuel Roi d'Angleterre*, qui renvoie à certain homme), et d'autres encore qui dénotent de manière ambiguë (*un homme*, qui ne renvoie pas à une série d'homme, mais à un homme imprécis). Russell s'intéresse plus particulièrement aux expressions introduites par l'article défini, qu'il appellera par la suite « descriptions définies ».

**Les énigmes référentielles** Au point de départ de sa théorie se trouvent les énigmes ou paradoxes référentiels, faisant figure de véritables « expériences » logiques, au même titre que celles pratiquées en physique[1]. Russell se penche sur trois énigmes ; nous en détaillons une, les deux autres étant seulement évoquées :

– Le problème des énoncés d'identité vrais et informatifs[2] :
  Prenons la phrase *George IV voulait savoir si Scott était l'auteur de Waver-*

[1] « *A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since they serve much the same purpose as is served by experiments in physical science* », [Russell, 1905].

[2] On retrouve ici les mêmes interrogations déjà soulevées par Frege dans son article *Sens et Dénotation* [Frege, 1892] (à propos de l'étoile du soir et de l'étoile du matin, cf. section 4.1.2.1) ; Russell en propose cependant une autre analyse.

*ley*. Selon le principe posé par Leibniz, si *a* est identique à *b*, alors on peut substituer l'un à l'autre sans changer la valeur de vérité (*salva veritate*) de la proposition dans laquelle ils apparaissent. Or de fait, Scott *est* l'auteur de Waverley donc il serait possible d'opérer une substitution et de dire : *George IV voulait savoir si Scott était Scott*, énoncé qui n'a vraisemblablement plus la même signification que le premier, ou qui prête « au premier gentilhomme d'Angleterre » d'étonnantes interrogations.

- La loi du tiers exclu et la valeur de vérité d'une proposition telle que *L'actuel Roi de France est chauve.*
- Le problème des énoncés existenciels négatifs (*Pégase n'existe pas*) ou *« Comment une non-entité peut-elle être le sujet d'une proposition »* ?

Face à ces paradoxes, Russell défend une vision dénotative de la référence, réfutant les solutions proposées par A. Meinong et G. Frege. Le premier propose, à l'aide de sa théorie des objets, de sortir de l'impasse des entités non-dénotatives en leur conférant un référé d'un autre type que celui des entités réelles. Pour ce faire, il distingue l'*être* de l'*existence* : tous les objets ont l'être, mais tous n'ont pas l'existence ou réalité effective. Ainsi, indépendamment de savoir si l'objet d'une description est ou non une entité, il *est* en tant qu'objet donné ou ayant de l'être, tel que la description le définit. La question de l'existence d'un objet dénoté par une expression ne pose ainsi plus de problème. Russell n'est pas de cet avis (bien qu'il eut auparavant défendu une position similaire) et critique vivement cette solution. Il rejette tout autant la position frégéenne distinguant le sens de la référence, lui opposant une analyse qui se passe du concept de sens. En effet, là où Frege propose d'admettre que certaines expressions peuvent avoir un sens mais pas de dénotation, Russell pense au contraire que toute proposition ayant un sens doit avoir une dénotation[1] . Par conséquent, la proposition de Frege, si elle ne conduit pas à une véritable erreur logique, est artificielle et ne permet pas de résoudre le problème des entités qui ne dénotent rien ou des non-entités[2].

Ainsi, face à une expression dénotante comportant une non-entité, B. Russell rejette donc la proposition d'admettre qu'elles peuvent dénoter des objets qui n'existent pas (A. Meinong), tout comme celle d'admettre une dénotation purement conventionnelle (Frege). Il opte pour une autre solution, consistant à différencier la forme grammaticale de la forme logique.

[1] D. Marconi [Marconi, 1997] indique que pour Russell « *la seule propriété sémantique d'une expression linguistique qui soit de quelconque importance pour la valeur de vérité des énoncés dans lesquels elle apparaît, c'est sa dénotation* ».

[2] « *this procedure, though it may not lead to actual logical error, is plainly artificial, and does not give an exact analysis of the matter* », Russell [Russell, 1905].

**La théorie de B. Russell** Russell propose de retrouver, sous les apparences grammaticales, la structure logique sous-jacente des propositions. Dans cette perspective, les expressions dénotatives, et donc les descriptions définies, perdent leur statut de sujet (grammatical) pour devenir des variables liées au sein de fonctions propositionnelles ; elles ne peuvent dès lors avoir de dénotation par elles-mêmes, mais uniquement *contextuellement*, au sein d'une proposition. Russell expose cela synthétiquement : « *This is the principle of the theory of denoting I wish to advocate : that denoting phrases never have any meaning in themselves, but that every proposition in whose verbal expression they occur has a meaning* », et plus loin, après avoir précisé les détails de sa théorie : « *the above gives a reduction of all propositions in which denoting phrases occur to forms in which no such phrases occur* ». La thèse de Russell est donc que les propositions contenant des expressions dénotatives sont réductibles à des propositions qui n'en contiennent pas. Concrètement parlant, cela revient à faire disparaître les expressions dénotatives en les disloquant. Sans entrer dans tous les détails de la démonstration et des notations logiques, il importe d'expliquer la teneur de ces propos.

Pour Russell, le langage logique sert de norme à l'étude du langage ordinaire [Vernant, 1980]. Il présente tout d'abord son formalisme, basé pour l'essentiel sur la notion de variable et de fonction propositionnelle. Une proposition est toujours de la forme $C(x)$, où $C$ représente l'assertion, et $(x)$ le sujet à propos duquel l'assertion est faite. Prenant la proposition *J'ai rencontré un homme*, Russell propose de l'analyser en considérant, d'une part, la proposition, correspondant ici à « j'ai rencontré » et, d'autre part, l'expression dénotante « un homme ». Cette proposition a donc pour forme logique (en français) : 'Il existe (au moins) un x tel que x est un homme et j'ai rencontré x'.

Examinons à présent la proposition *Le père de Charles II fut exécuté.* Il est possible de considérer, d'une part, le prédicat « avoir été exécuté » et, d'autre part, la description définie sujet « le père de Charles II ». Le tout peut être paraphrasé logiquement par 'Il y a un $x$ qui était le père de Charles II et qui a été exécuté'. A ce point de l'analyse, la description définie 'le père de Charles II' est une variable indéterminée $(x)$ à laquelle est attribuée une propriété (être le père de Charles II) sous la forme d'une fonction prédicative, encore appelée « qualification descriptive » par D. Vernant [Vernant, 1980]. Russell considère cependant qu'il faut aller encore plus loin pour représenter logiquement et justement une description définie et que « *pour valoir comme support de la qualification descriptive* (ci-dessous $P$), *la position d'un particulier quelconque opérée par l'article 'le' et traduite par la variable d'individu* $(x)$ *doit être renforcée par la double supposition de son existence et de son unicité* » [Vernant, 1980]. La condition d'existence correspond au fait que toute description implique l'existence d'une variable d'indi-

vidu susceptible de satisfaire la propriété véhiculée par la qualifiation descriptive. Si l'on représente 'le père de Charles II' par la variable $x$ dotée d'une propriété, alors il faut qu'il existe au moins une valeur de $x$ satisfaisant la propriété en question. La description 'le père de Charles II' implique donc 'il existe au moins un $x$ tel qu'il est le père de Charles II'. A cette condition d'existence vient s'ajouter une condition d'unicité : Russell indique en effet que lorsque nous disons 'x était le père de Charles II' nous indiquons non seulement que $x$ entretenait une certaine relation (de parenté en l'occurrence) avec Charles II mais qu'en plus, en vertu de l'article défini, il était le seul à avoir cette relation. Autrement dit, cette seconde condition restreint la première, postulant que la description pose nécessairement l'existence d'*au moins* un, mais d'*au plus* un $x$ (ou un individu) satisfaisant la propriété. La paraphrase, prenant en compte cette unicité, devient alors : 'Il existe un $x$ qui était le père de Charles II et $x$ a été exécuté, et si $y$ est le père de Charles II, alors $y$ est identique à $x$' ; notée symboliquement, avec les propositions P pour 'être le père' et E 'avoir été exécuté' :

$$\exists x[Px \wedge \forall y\,(Py \rightarrow y = x)) \wedge Ex]$$

Ainsi, pour que *Le père de Charles II fut exécuté* soit vraie, il faut, premièrement, qu'il existe au moins une entité qui soit le père de Charles II (existence), deuxièmement, qu'une entité tout au plus possède cette propriété d'être le père de Charles II (unicité) et, troisièmement, que cette entité aie été exécutée, ce dernier point (contextuel) ne pouvant être possible que si les deux premiers (liés à la description définie) sont vérifiés. A l'inverse, si la proposition *Le père de Charles II fut exécuté* est vraie alors *Le père de Charles II existe* l'est également. Le point essentiel dans cette analyse est que toute expression de cette forme, soit toute description définie, implique l'existence et l'unicité du référent. Ce raisonnement par dislocation, faisant « disparaître » les descriptions définies des propositions dans lesquelles elles apparaissent, permet par ailleurs de résoudre les puzzles référentiels.

**La résolution des énigmes** Précisons tout d'abord la distinction opérée par Russell entre occurrence primaire » et « occurrence secondaire ». Dans la phrase *J'ai rencontré un homme*, l'expression dénotante *un homme* est une occurrence primaire car elle porte sur la totalité de la phrase dans laquelle elle apparaît. Elle est représentée : '$(\exists\, x)$ (j'ai rencontré $x$ et $x$ est humain)'. Dans la phrase *J'ai dit que j'ai rencontré un homme*, l'expression dénotante *un homme* est une occurrence secondaire car elle ne porte que sur la proposition subordonnée, cette dernière n'étant qu'un composant de la phrase dans laquelle elle s'insère. Elle est représentée : 'J'ai dit que $(\exists\, x)$ (j'ai rencontré $x$ et $x$ est humain)'. De fait, la dislocation d'une expression dénotante en occurrence primaire n'a pas la même

portée que celle d'une occurrence secondaire.

Au regard du premier paradoxe de l'identité référentielle, représenté par l'interrogation de George IV qui « *souhaitait savoir si Scott était l'auteur de Waverley* », la théorie russellienne propose de faire disparaître la description définie *l'auteur de Waverley*. Cette dernière devient alors : 'il existe une et une seule entité qui a écrit *Waverley*, et Scott est identique à cette entité'[1]. De la sorte, la proposition première ne contient plus véritablement de constituant auquel on peut substituer *Scott*. Russell précise que « *The phrase (the author of Waverley) 'per se' has no meaning, because in any proposition in which it occurs the proposition, fully expressed, does not contain the phrase, which has been broken up* ». L'énoncé, pleinement exprimé, devient un énoncé informatif et non un énoncé d'identité (tautologique).

Le problème de la calvitie du Roi de France tout comme celui des non-entités se trouvent également résolus.

**Bilan et critiques** Les travaux de Russell visent à montrer qu'il est possible de trouver, sous la forme grammaticale apparente des phrases des langues naturelles, la forme logique traduisant authentiquement les propositions exprimées. Au regard des descriptions définies, Russell propose de réduire une description définie aux constituants suivants :

1. une condition d'existence (liée à la description)
2. une condition d'unicité (liée à la description)
3. une qualification contextuelle

Dès lors qu'un des constituants n'est pas vérifié, la proposition ne peut être vraie. Au-delà (ou en-deçà) de la valeur des jugements comportant des descriptions définies, il semble donc important de retenir le fait qu'une description de la forme 'le tel-et-tel' indique qu'il existe une et une seule entité qui soit telle-et-telle. Ces spécificités des descriptions définies pour la première fois dégagées « logiquement » par B. Russell ont servi d'amorce à de nombreux débats et critiques, parmi lesquelles celle de P. F. Strawson notamment. Ayant choisi d'examiner la thèse de B. Russell, nous nous limiterons ici à une brève indication de cette critique, cette rapidité d'évocation ne devant en rien masquer l'importance du propos de P. Strawson.

Dans un autre article phare sur les descriptions définies, intitulé *On referring* et publié en 1950 en réponse à Russell dans la même revue philosophique *Mind*,

[1] ou, selon la formule complète, 'It is not always false of $x$ that $x$ wrote *Waverley*, that is always true of $y$ that if $y$ wrote *Waverley* $y$ is identical with $x$, and that Scott is identical with $x$' [Russell, 1905].

le philosophe du langage P. F. Strawson entend montrer que l'article du logicien comprend « quelques erreurs fondamentales » [1]. Sa critique ne porte pas sur les indications d'existence et d'unicité véhiculées par les descriptions définies, mais plus exactement sur la nature de ces indications, qu'il voit comme des *présuppositions* et non, comme le dit Russell, comme des *implications*. Partant, la valeur de vérité d'un énoncé comprenant une description définie peut être *inévaluable*, soit ni vrai ni faux.

Pour Strawson, se demander si « L'actuel Roi de France est chauve » est vrai ou faux ne rime à rien car il n'existe aucune entité ayant la propriété « actuel Roi de France ». Si l'on vous posait une telle question « *Je pense*, nous dit Strawson, « *que vous seriez enclin à répondre, après quelques hésitations, que vous n'étiez ni d'accord ni en désaccord, que la question de savoir si cet énoncé était vrai ou faux* ne se posait *tout simplement* pas *parce qu'il n'y a personne qui soit à présent le Roi de France* ». Russell, pour qui une proposition qui a du sens doit nécessairement avoir une valeur de vérité, transforme les descriptions définies en assertions d'existence, dont on peut ensuite dire si elles sont vraies ou fausses. Strawson considère qu'une description définie présuppose mais n'implique pas une condition d'existence, de fait la question de la valeur de vérité d'une entité qui n'existe pas ne se pose pas.

A l'origine de ces critiques se trouve l'observation de l'*usage ordinaire* du langage et de la distinction entre une phrase, l'usage d'une phrase et une occurrence. La phrase « L'actuel Roi de France est chauve » peut avoir un usage ordinaire, dont le plus courant se déroule entre locuteurs, s'opposant à son usage sur les planches d'un théâtre ou dans un roman par exemple, et plusieurs occurrences, selon des coordonnées spatio-temporelles différentes. Un locuteur faisant usage de cette phrase et en prononçant une occurrence en 1750 aurait certainement pu lui attribuer une valeur de vérité (la présupposition d'existence entrant en correspondance avec la réalité), tandis qu'en 2005, ce que présuppose la phrase ne correspond pas à la réalité de son occurrence, laquelle ne peut alors ni être dite vraie, ni être dite fausse, bien que la phrase ait une signification. Une description définie contient donc la présupposition qu'il existe un objet lui correspondant, mais ne l'implique pas.

La critique de Strawson à l'encontre de la théorie des descriptions définies de Russell fut d'importance, davantage pour l'étude du langage, posant qu'une phrase a une signification et que son usage a une référence, que pour celle des descriptions définies à proprement parler. Le débat Strawson-Russell (ce dernier lui fit une réponse dans un troisième article) n'est en fait pas une réelle controverse,

[1] cité par L. Linsky, [Linsky, 1974], chapitre VI : *Strawson sur la référence*, p125.

« *le propos de Russell* [étant] *d'analyser une certaine* classe de propositions *; le propos de Strawson, d'étudier* un certain usage des mots » [Linsky, 1974]. Ou comme le dit G. Kleiber : « Russell tire le langage naturel vers la logique, Strawson la logique vers le langage naturel » [Kleiber, 1981b].

Au final, implication ou présupposition, ce qu'il importe de retenir du débat philosophico-logique sur les descriptions définies[1], c'est qu'elles indiquent qu'il existe (quel que soit ce mode d'existence) un et un seul objet satisfaisant à la qualification descriptive. Au regard de la spécification de ce qu'est une description définie, s'ajoute ainsi, au critère formel de présence de l'article défini, un critère logique de dénotation unique. Si nous reprenons la liste proposée ci-dessus :

*la baleine (est un mammifère)*
*la baleine (a heurté le bateau)*
*l'imagination*
*l'imagination de Paul*
*le fer*
*le père de Charles II*
*le Roi de France*
*l'actuel Roi de France (est chauve)*
*le cercle carré (n'existe pas)*
*le Président*
*le Président de la République*
*le Président de la République Française en 2005*
*la voiture immatriculée 478 KNB 75*

on observe que tous les syntagmes en *le* (critère formel de départ) ne renvoient pas nécessairement à un référent unique (critère logique). Il faut alors supprimer les expressions qui ne nomment pas et ne renvoient pas à un et un seul objet, à savoir *la baleine* telle qu'employée génériquement et renvoyant à la classe des baleines dans sa totalité dans *la baleine est un mammifère.* Les expressions restantes semblent satisfaire le critère de renvoi à un référent unique, mais il est certaines pour lesquelles demeurent quelques hésitations, à savoir :

*l'imagination*
*l'imagination de Paul*
*le fer*

[1]Pour plus de précisions sur les travaux de Russell puis de Strawson sur les descriptions définies : [Linsky, 1974, Vernant, 1980, Vernant, 1993, Kleiber, 1981b].

G. Kleiber propose alors d'introduire deux critères suppplémentaires permettant de caractériser et d'identifier au mieux les descriptions définies.

#### 4.3.1.2 La définition des descriptions définies de G. Kleiber

Partant du critère formel et du critère logique, G. Kleiber ([Kleiber, 1981a]) propose deux restrictions supplémentaires pour une définition des descriptions définies. A l'origine de ces restrictions, se trouve l'hésitation à accepter comme descriptions définies certaines des expressions suivantes :

(1) *l'imagination*
(2) *le fer*
(3) *le petit-fils de Pierre*
(4) *le tort de fumer*

G. Kleiber fait en effet remarquer que ce sont exclusivement des exemples similaires à (3) qui servent d'illustration à la notion de description définie, soit des exemples comprenant des expressions opérant une référence particulière spécifique, c'est-à-dire des expressions renvoyant à une entité particulière d'un certain type dont l'existence est présupposée et qui est identifiable comme telle dans une situation donnée. Quelle différence existe-t-il entre les expressions mentionnées ci-dessus ? Il est possible de remarquer à première vue que *l'imagination*, *le fer* et *le tort de fumer* renvoient à des entités (abstraites ou concrètes) certes particulières, mais ayant une existence en elles-mêmes, quoiqu'il arrive pourrait-on dire, tandis que *le petit-fils de Pierre* peut renvoyer à divers petits-fils, voire à aucun. Le point essentiel, précise Kleiber, c'est que « *(1)(2) et (4) ne posent pas de problèmes d'interprétation référentielle, car l'objet de référence s'y trouve identifié par le sens (ou référence virtuelle) de l'expression elle-même.(3), au contraire, nécessite (...) des connaissances extra-linguistiques permettant de reconnaître le particulier qui satisfait à la description* ». Ces observations conduisent l'auteur à poser deux restrictions supplémentaires à la définition traditonnelle des descriptions définies :

– Restriction 1 sur la nature du substantif :
  Le substantif d'une description définie doit être soit *individuant*, c'est-à-dire doit présupposer par avance une catégorie référentielle découpée en individus (voiture, livre, ministre, etc.), soit *globalisant réifiés en discontinu*, c'est-à-dire réduisant à l'état d'objet une notion abstraite, objet supposant alors l'existence d'une classe de 'tel-et-tels'. Cette restriction élimine les exemples (1) et (2) dans lesquels figurent des substantifs globalisants. En

revanche, des expressions telles que *l'imagination de Paul est fertile* et *le fer de cette balustrade a été repeint* dans lesquelles figurent des substantifs globalisants réifiés en discontinu sont des descriptions définies. Se pose alors le problème d'expressions telles que :

(5) *le toit de la maison est généralement en tuiles*
(6) *la douceur de vivre est une vertu du soleil*

qui elles aussi comportent des substantifs globalisants réifié en discontinu, sans pour autant être des descriptions définies. Ce problème est également soulevé par L. Linsky qui, considérant l'expression *la table* dans la proposition *la table est le meuble le plus important de la salle à manger*, établit que « *ce qui ne va pas ici, c'est que l'expression 'la table' n'est pas employée en vue de renvoyer à quelque table particulière* ». Prenant ensuite l'expression « *le livre se trouve sur la table* » et proposant d' « [admettre] *que cette proposition* [est] *exprimée en un contexte dans lequel 'la table' se réfère à une table particulière* », alors « *nous serons enclins à dire qu'ici 'la table' est une description définie. Mais si nous le disons, nous abandonnons l'idée qu'une expression est descriptive en vertu de sa forme grammaticale ; car l'expression 'la table' est de même forme grammaticale dans la proposition où elle est une description définie et dans celle où elle ne l'est pas* » [Linsky, 1974, p.96]. Linsky, pour répondre à cette question, s'en remet à l'analyse logique de Russell[1] ; G. Kleiber propose quant à lui une seconde restriction, d'ordre plus linguistique.

– Restriction 2 sur les indices ou points référentiels :
Pour localiser le référent d'une description définie, il faut avoir recours à des indices ou points référentiels. Ce que G. Kleiber appelle « points ou indices référentiels » correspond « *aux facteurs qui, à partir du sens donné d'une expression, en déterminent la signification* ». La seconde contrainte est donc de nature empirique et découle naturellement de la première : si le substantif suppose une classe d'individus (restriction 1), l'article défini pointe lui sur un particulier précis et unique de cette classe, cette unicité devant être vérifiée *empiriquement*. Il y a donc contradiction entre une référence à une pluralité d'individus possibles et la nécessité d'en identifier un et un seul ; cette contradiction se trouve résolue par des indices référentiels

[1] « *Mais si la seule manière de décider si une expression donnée fonctionne comme description définie consistait à assurer que l'analyse donnée par Russell de telles propositions est (ou non) l'analyse correcte dans le cas en question, il s'ensuivrait que l'analyse de Russell ne pourrait pas être erronnée et que ceux qui, comme Strawson, ont prétendu qu'elle était erronée, ont commis une confusion au sujet d'une définition.* », [Linsky, 1974], p.97.

qui permettent de restreindre les indications apportées par le substantif individualisant. G. Kleiber présente les exemples suivants :

(5) *le toit de la maison est généralement en tuiles*
(7) *le toit de la maison a été totalement détruit*

Dans (5) l'expression nominale définie renvoie à un objet en général dont l'identification est assurée par le seul sens de l'expression, tandis que dans (7) elle renvoie à un objet particulier dont l'identification doit obligatoirement passer par des points référentiels.

Substantif individuant et indices référentiels, ces deux restrictions proposées par G. Kleiber en complément des critères logique et formel évoqués ci-dessus conduisent au final à un concept relativement restreint. L'appellation « description définie » est alors réservée aux expressions nominales de forme *le + substantif individuant (+ modificateur ou non)* qui, d'un point de vue référentiel, servent à renvoyer à un particulier dont l'identification ne peut se faire par le sens seul de l'expression mais doit obligatoirement passer par des indices ou points référentiels. On retrouve une restriction similaire dans le propos introductif de M. Charolles à son chapitre consacré aux « expressions nominales définies » [Charolles, 2002a]. Il circonscrit en effet l'étude de ces expressions à l'étude des « *emplois des SN définis dans lesquels ceux-ci réfèrent à des entités concrètes, comptables et spécifiques* », laissant de côté « *les définis désignant des entités massives (l'huile, le sable) ou abstraites (le bonheur, la paix) ou des événements (l'inflation, la pollution), ainsi que les définis renvoyant génériquement à des types (le chat est le meilleur ami de l'homme) ou à des classes (les chats craignent l'eau)* ». Il use par la suite de l'appellation « description définie » pour désigner les expressions restantes et constituant l'objet d'étude de son chapitre, effectuant ainsi la même caractérisation que G. Kleiber.

Ainsi, au terme de ce parcours débuté du côté (historique) de la logique et s'achevant du côté de la linguistique, il est possible de présenter comme descriptions définies les expressions suivantes :

*la baleine (a heurté le bateau)*
*l'imagination de Paul*
*le père de Charles II*
*le Roi de France*
*l'actuel Roi de France (est chauve)*
*le Président*
*le Président de la République*
*le Président de la République Française en 2007*

*la voiture immatriculée 478 KNB 75*

qui toutes réfèrent spécifiquement à un particulier présupposé existant et unique et dont l'identification, ne pouvant se faire par le sens seul de la description et nécessitant des connaissances empiriques, passe par le recours à des indices ou points référentiels. Ayant ainsi précisé ce que l'on entend par *descriptions définies*, il convient à présent d'examiner leur sens et leur fonctionnement référentiel, dont les principales caractéristiques découlent tout naturellement des critères ci-avant mentionnés.

## 4.3.2 Sens et fonctionnement référentiel des descriptions définies

### 4.3.2.1 Le sens des descriptions définies

Contrairement à celui des noms propres, le sens d'une description définie ne pose aucun problème. Avec ces unités lexicales, on retrouve en effet le sens '« classique » de la plupart des expressions, à savoir un sens dénotatif stable, précodé, spécifiant objectivement les conditions qu'une entité doit satisfaire pour être désignée par elle. Selon que l'expression est composée d'un ou plusieurs mots pleins, son sens est compositionnel ou non. Ainsi, la description définie '« la voiture rouge » renvoie à un véhicule de type *voiture* et de couleur rouge et *Le président de la République* renvoie à un particulier qui dirige un pays doté d'un régime politique républicain. Cette désignation est typiquement *descriptive* et transparente (d'où le nom *description* définie), à l'opposé de la désignation dénominative opérée par le nom propre, ce dernier ne donnant aucun attribut ou propriété du particulier désigné, sinon son nom (*Clio* ne nous dit rien en effet de l'objet qu'il désigne). Une des conséquences de cette référence descriptive est le fait que la production de ce type d'unité lexicale obéit aux règles lexicales et grammaticales classiques : la désignation d'un objet unique passe ici par la combinaison de mots précis et non par la libre invention ou choix d'un nom.

Ce sens descriptif des descriptions définies vient compléter les propriétés sémantico-référentielles intrinsèques que sont les présuppositions d'existence et d'unicité. De la sorte, '« la voiture rouge » renvoie à un et un seul véhicule de type *voiture* et de couleur rouge identifiable dans une situation donnée, de même que *Le président de la République* renvoie à un et un seul président d'une République donnée.

Ce sens descriptif et ces propriétés sémantico-référentielles intrinsèques ne suffisent cependant pas, la plupart du temps, à identifier correctement le référent visé ; le fonctionnement référentiel des descriptions définies appelle la prise

en compte du contexte restreint dans lequel vaut telle ou telle description. Ce contexte entre en jeu selon des degrés variables, qui permettent de faire la distinction entre descriptions définies complètes et descriptions définies incomplètes.

#### 4.3.2.2 Descriptions définies complètes et incomplètes

Reprenons la liste d'expressions nominales considérées comme des descriptions définies à l'issue de la section 4.3.1.2 et considérons les unités suivantes :

*le Président*
*le Président de la République*
*le Président de la République Française en 2005*
*la voiture rouge*
*la voiture immatriculée 478 KNB 75*

Ce qui différencie ces expressions est naturellement le fait qu'à l'énonciation de la phrase *Le Président a fait une déclaration*, le locuteur aura plus de chances d'entendre de son interlocuteur la question *Quel Président ?* qu'après l'énonciation de la phrase *Le Président de la République Française en 2005 a fait une déclaration*. La première description définie (*le Président*) a pour objectif, en tant que telle, de référer à un particulier unique ayant la propriété d'être président, mais elle ne fournit cependant pas la totalité des éléments permettant l'identification du référent désigné, tandis que la seconde (*le Président de la République Française en 2005*), à l'inverse, est riche d'informations quant à ce dernier. De la même manière, le référent de *la voiture rouge*, en l'absence d'indications ou points référentiels supplémentaires, est plus difficile à identifier que celui de *la voiture immatriculée 478 KNB 75*. La différence entre ces descriptions définies réside donc dans leur mode d'établissement de la référence, plus ou moins autonome, différence qui justifie la distinction entre descriptions définies complètes et incomplètes introduite par M. Charolles [Charolles, 2002a].

Une *description définie complète* est autonome référentiellement : sa composition restreint suffisamment sa dénotation au point qu'elle ne vaut, dans l'absolu, que pour un seul référent. En d'autres termes elle peut '« *par ses seules ressources, évoquer un référent* ». Pour ce faire, les descriptions définies complètes sont souvent composées d'un syntagme nominal défini ($le + N$) accompagné d'un complément prépositionnel (le président *de la République*) ou d'une relative déterminative (le président *qui vient d'être élu*), adjoints dans lequels figurent prototypiquement des noms propres (le président *de la France*), qualifiés d'ailleurs de '« *pivots des descriptions définies* » par G. Kleiber ([Kleiber, 1981a], p.545). Ces adjoints ont pour effet de restreindre l'extension de la description définie à

un particulier déterminé et de permettre sa représentation indépendamment du contexte.

A l'opposé, les *descriptions définies incomplètes* ne fournissent pas toutes les indications nécessaires à l'identification du référent, leur interprétation doit donc faire intervenir des éléments du contexte ou connaissances extra-linguistiques en complément d'information. L'interprétation de l'expression *le Président* requiert des connaissances extra-linguistiques et situationnelles : dans le contexte de l'assemblée générale d'une association, l'expression pourra renvoyer à Pierre Dupont par exemple, tandis que dans le contexte d'un journal télévisé français, l'expression pourra renvoyer à l'actuel Président de la République française. L'accès au référent d'une description définie incomplète est ainsi indirect, il passe par une évaluation du contexte et de l'univers de discours dans lequel le syntagme défini sélectionne une entité unique de son espèce.

Le premier type d'expressions, les descriptions définies complètes, qui ne nécessitent pas de restriction de leur dénotation pour évoquer de manière univoque un particulier dans toutes les situations, conduit à penser à une certaine contradiction avec la restriction numéro 2 apportée par G. Kleiber à la caractérisation des descriptions définies. Cette contradiction n'est qu'apparente et il importe de bien insister sur ce qu'entend G. Kleiber par indices ou points référentiels. Son objectif (croyons-nous) par l'indication de ce critère, est de rendre compte de la spécificité de la référence à un particulier par rapport à la référence générique. De fait, '« *toute phrase référant à un particulier est obligatoirement une phrase synthétique*[1] *(ou une phrase empirique non a priori), c'est-à-dire qu'elle ne peut jamais être vraie ou fausse ou ni vraie ni fausse que par son seul sens. Pourquoi cela ? Renvoyant à un individu précis, elle exige une vérification empirique* ». Ainsi, complète ou incomplète, la référence d'une description définie revêt un caractère *accidentel* qui explique que '« *si les expressions qui réfèrent à des particuliers figurent, pour certaines d'entre elles, dans des dictionnaires de langue, il est évidemment exclu que les individus particuliers trouvent place dans leur définition* ». Il s'en suit qu'une description complète dite autonome référentiellement comme *le Président de la République Française en 2007*, si elle indique bien par elle seule le chemin à prendre pour aller vers le référent (autonomie référentielle vis-à-vis des connaissances linguistiques), ne permet pas totalement l'identification de ce dernier qui passe alors par le truchement de points référentiels que sont ici, la connaissance du référent de la *République Française*.

Au niveau de la compétence des sujets parlants, les descriptions définies fonc-

[1]En philosophie, jugement synthétique : qui introduit dans un prédicat une notion qui n'est pas comprise dans le sujet et qui ne peut se vérifier que par les faits (*Trésor de la langue française informatisé*).

tionnent donc différemment desf noms propres : il n'est en effet pas nécessaire de connaître une convention de nomination pour comprendre et user d'une description définie, ce qui importe vraiment pour son interprétation, ce sont des connaissances extra-linguistiques, indiquant que tel ou tel objet ou entité possède les propriétés dénotées par le sens des composants de la description.

La distinction entre descriptions définies complètes et incomplètes nous permet ainsi de rendre compte de la différence entre les expressions suivantes :

*le Président*
*le Président de la République*
*le Président de la République Française en 2005*
*la voiture rouge*
*la voiture immatriculée 478 KNB 75*

qui exigent plus (*le Président*, *la voiture rouge*) ou moins (*le Président de la République*, *la voiture immatriculée 478 KNB 75*) de connaissances pour l'identification de leur référent, selon un continuum dont le curseur peut être défini comme l'autonomie référentielle ; à l'extrêmité de ce dernier, on retrouve des expressions dont le fonctionnement référentiel se rapproche de celui du nom propre (*le Président de la République Française en 2007*).

Les descriptions définies offrent donc la possibilité de référer à une entité particulière, quelle qu'elle soit. Elles ne sont pas cantonnées à quelques catégories d'existants et viennent ainsi compléter astucieusement les noms propres dans l'opération langagière de référence définie unique : ceux-ci sont ‘« cognitivement économiques » dans la mesure où ils permettent de référer directement (par une dénomination uniquement) et efficacement (dénomination non contingente) à des entités avec lesquelles nous entretenons un ‘« commerce régulier », et celles-là offrent une alternative à l'impossible apprentissage d'un nom propre pour chaque objet ou entité particuliers du monde.

A cette étape de l'analyse des catégories linguistiques existantes composant l'ensemble 'entités nommées', reprenons le point de départ de cette exploration avec la figure 4.2 (ci-après). Il est désormais possible de préciser que les descriptions définies *le Président de la République*, *le Président du Conseil*, *le Décret de loi 31/3 de 2005*, renvoient à un unique Président de la République, à un unique Président du Conseil ainsi qu'à un unique décret, de façon plus ou mois autonome : le nom *la République* restreint plus la dénotation de *le Président* que le nom *Conseil* (il y a en effet plus de conseils que de Républiques dans l'univers de discours par défaut d'un locuteur français), la première description définie est donc plus autonome référentiellement que la deuxième. Ce phénomène est la rai-

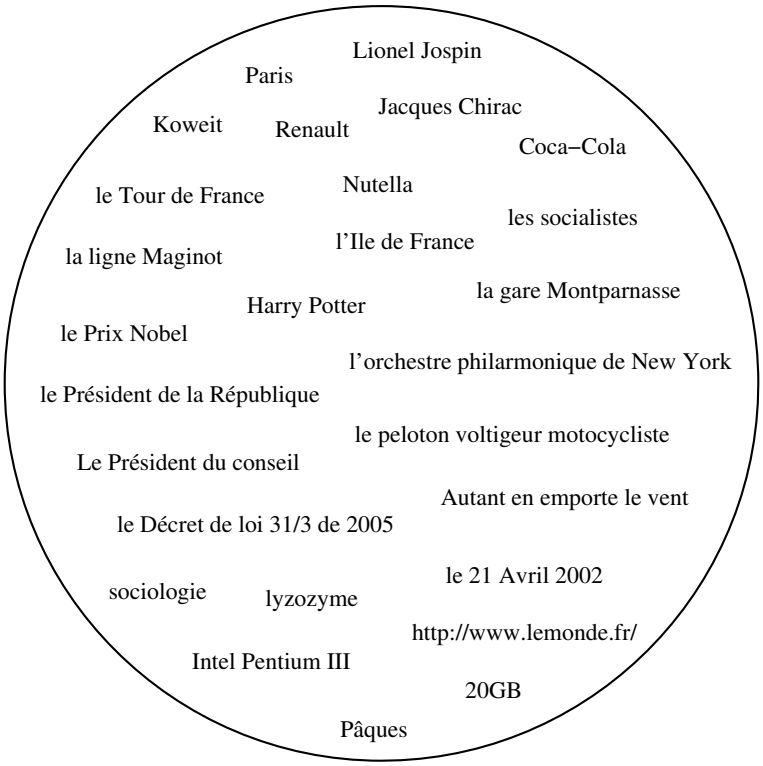



FIG. 4.2 – Echantillon d'unités lexicales considérées comme des entités nommées.

son pour laquelle B. Daille *et al.*, dans leur étude de corpus pour la catégorisation de noms propres '« *ne* [considèrent] *pas 'le Conseiller régional', à l'inverse de 'le Président de la République', comme une entité nommée car il est difficile d'identifier quelle est la personne à laquelle il est fait référence par ce titre* ». Au sein des entités nommées, on trouve donc des descriptions définies, sinon totalement autonomes, au moins fortement identifiantes, ne nécessitant que peu d'informations supplémentaires pour l'identification de leur référent. Le contexte entrant en jeu correspond alors le plus souvent à l'univers de discours habituel des locuteurs ou univers "par défaut". Les descriptions définies réfèrent le plus souvent, dans ce cadre des entités nommées, à des entités particulières relativement stables qui, soit qu'elles ne font pas l'objet d'une prédication régulière (ou que leur référent change), soit que l'usage n'impose pas une nomination, n'ont pas de nom propre, ne sont pas nommées directement mais seulement décrites.

D'autres unités faisant partie de l'échantillon typique des entités nommées sont les expressions numériques et temporelles, représentées dans la figure par les unités *20GB* et *le 21 Avril 2002*. Il convient, au final, de dire un mot sur ces expressions.

#### 4.3.2.3 Les expressions numériques

L'idée tient effectivement en quelques mots : les expressions numériques et temporelles fonctionnent de la même manière que les descriptions définies. Il s'agit, au delà des *20GB* représentés dans le cercle, des diverses expressions temporelles (absolues comme *le 21 Avril 2002* ou *1989*, ou relatives comme *l'année dernière*), des expressions monétaires telles que *20 euros* et de diverses expressions quantitatives telles que *20%*, *trois mille huit cent sept mètres* et *2kg 950*.

Les expressions temporelles renvoient à un moment particulier d'un calendrier[1], et ce de manière plus ou moins autonome. La description définie *le 21 Avril 2002* renvoie sans aucune ambiguïté à un jour précis et unique et peut donc être qualifiée de complète, tandis que la description définie *l'année dernière* renvoie à un intervalle de temps lui aussi précis et unique, mais dont l'identification et l'interprétation nécessitent des connaissances supplémentaires, à savoir l'année du moment de l'énonciation (ou de production du document) ; elle peut alors être qualifiée d'incomplète. La plus ou moins grande précision d'une description définie temporelle absolue ne joue pas, contrairement aux autres descriptions définies, au niveau de sa capacité identificatoire. Si la précision des expressions *le 21 Avril 2002*, *Avril 2002* et *2002* s'amoindrit comme pour les expressions *le Président de la République Française en 2007*, *le Président de la République* et *le Président*, leur capacité identificatoire fonctionne tout autant, avec un référent plus facile à déterminer pour *2002* (on ne se demandra pas *quelle 2002 ?* ou *2002 de quoi ?*) que pour *le Président* (pour lequel beaucoup de questions peuvent se poser). Les entités auxquelles on se réfère *via* les points d'un calendrier sont statiques et universellement connues, contrairement aux autres types d'entités. Au final, la partition expression temporelle absolue *vs.* expression temporelle relative recouvre naturellement celle de description définie complète (ou autonome) *vs.* description définie incomplète.

Les expressions composées d'adjectifs numéraux cardinaux combinés à des unités monétaires ou de mesure dénombrent des quantités précises. Il est à noter que ces quantités, si elles sont précises dans leur grandeur, restent la plupart du temps indéfinies quant à la dénotation de ce qui est quantifié (les cardinaux sont combinés à des unités de mesure et non à des noms) : si l'expression *20 euros* renvoie de manière autonome à une quantité précise d'argent, on est en droit, au regard de l'expression *2kg 950*, de demander 2kg 950 de quoi ? A l'instar des expressions temporelles, et des descriptions définies dans leur ensemble, les expressions numériques sont plus ou moins autonomes, selon qu'est pris en compte

[1]Par défaut le calendrier grégorien, utilisé dans une grande partie du monde et constituant la référence pour l'ensemble des exemples à venir dans ce paragraphe.

ou non le nom dont la quantité est dénotée.

Pouvant être considérées comme un artefact de la reconnaissance d'entités nommées[1], les expressions temporelles et numériques n'en fonctionnent pas moins comme des decriptions définies, pointant avec plus ou moins de précision sur des points temporels et des quantités précis.

Il reste malgré tout, dans la figure 4.2, des unités lexicales quelque peu ‘« marginales » par rapport aux autres et dont on hésite à dire qu'elles sont noms propres ou descriptions définies. Que penser en effet de *sociologie*, des *socialistes*, du *peloton voltigeur motocycliste* et de l'adresse web ? Cette dernière pourrait être considérée comme une description définie complète, indiquant précisément et donnant les moyens d'identifier ce à quoi elle se rapporte (*via* un chemin précis conforme aux standards de l'internet). L'analyse des autres expressions est en revanche plus délicate, nous reviendrons sur certaines d'entre elles dans le chapitre 5.

En définitive, ce qu'il importe de retenir au regard des descriptions définies, c'est qu'elle permettent de faire référence à une entité unique. Cette monoréférentialité est la conséquence, avant tout, des propriétés sémantico-référentielles intrinsèques de ce type d'unités lexicales qui présupposent l'existence et l'unicité de leur référent. Au delà de ces critères logiques, avancés par B. Russell puis discutés par P.F. Strawson, il est possible de souligner le « fonctionnement » de cette monoréférentialité d'un point de vue linguistique : cette dernière n'est réellement atteinte que contextuellement, faisant intervenir des connaissances du monde dans le cas des descriptions définies complètes (*la championne du monde de natation sur 400m nage libre en 2007*) ou celles du contexte dans le cas des descriptions définies incomplètes (*la championne*).

## 4.4 Conclusion

Au terme de cet examen des catégories linguistiques existantes, quels sont les éléments qu'il importe de retenir s'agissant d'une définition des entités nommées ? Après avoir posé un cadre d'analyse en termes de sens et de référence, le premier spécifiant linguistiquement les moyens d'accéder à la réalité du second, nous avons étudié précisément les propriétés sémantiques et référentielles du nom propre d'une part et des descriptions définies d'autre part.

L'étude du nom propre a permis de clarifier la difficile question de son sens : loin d'en être totalement dénué comme l'avait posé le logicien Mill, le nom propre

[1]Ces unités lexicales sont dites « faciles » à reconnaître par un système automatique, du fait de leur composition à base de chiffres et d'un nombre fini d'unités de mesure ou temporelles facilement isolable et schématisable.

est doté d'un sens instructionnel dénominatif, indiquant que le référent à identifier est celui qui porte le nom en question, doublé d'un sens descriptif, certes ténu, indiquant pour sa part que le référent est un particulier. L'interprétation et l'usage d'un nom propre sont le fait de la connaissance d'une convention, celle instaurée lors du « baptême » d'une entité par un nom donné. La référence du nom propre s'opère ainsi majoritairement sur un mode dénominatif et ce dernier ne véhicule aucun trait ou attribut spécifique du porteur du nom. C'est pour cela que les noms propres sont analysés comme 'désignateurs rigides' (S. Kripke), renvoyant directement au même particulier quelle que soit la situation et les changements pouvant l'affecter.

L'étude des descriptions définies a quant à elle permis de dégager deux faits essentiels : ces unités lexicales renvoient à une entité unique, ceci correspondant à une propriété intrinsèque des expressions de type 'le tel-et-tel', et ce renvoi se fait sur un mode descriptif, le sens des mots composant les descriptions définies *décrivant* précisément l'entité à identifier. De cette façon, l'interprétation et l'usage d'une description définie sont le fait de connaissances extra-linguistiques (connaissance que telle entité possède la ou les propriétés évoquées par la description). Les indications fournies par une description définie peuvent être plus ou moins nombreuses ; dans le cas où peu d'informations permettent de restreindre la dénotation de l'expression à une entité unique, le recours au contexte est obligatoire. Ce recours détermine la plus ou moins grande autonomie référentielle de l'expression, qui peut être dite description définie complète ou description définie incomplète.

Ce tour du côté de la linguistique est donc doublement instructif et nous permet de répondre aux questions posées en introduction de ce chapitre. L'étude des catégories linguistiques composant l'ensemble des entités nommées a tout d'abord permis de dégager des points caractéristiques motivant cet ensemble et permettant d'y voir plus clair à propos de la question du sens et de la référence des unités le composant. En effet, qu'il s'agisse de noms propres dotés d'une référence dénominative s'appuyant sur la connaissance d'une convention, ou bien de descriptions définies dotées d'une référence descriptive s'appuyant sur des connaissances extralinguistiques, ces unités lexicales ont toutes deux la capacité d'opérer une référence définie à un particulier unique[1]. L'analyse de ces catégories a donc permis de dégager ce qui intéresse et justifie linguistiquement l'ensemble 'entités nommées' : la référence à une entité unique et l'autonomie référentielle. C'est par des mécanismes différents que noms propres et descrip-

[1]Le fait que les logiciens se soient intéressés à ces deux types d'unités lexicales n'est d'ailleurs pas étranger à cette caractéristique partagée, ces derniers étant exclusivement soucieux des problèmes de référence.

tions définies permettent d'y parvenir ; celui mis en œuvre par les noms propres étant plus direct, on comprend pourquoi les entités nommées ont plus d'affinités avec ceux-ci. Ces dernières ne sont toutefois réductibles ni à l'une ni à l'autre de ces catégories, étant en quelque sorte "plus que les noms propres, mais moins que les descriptions définies". En effet, cette étude a d'autre part mis en valeur le fait que, contrairement aux autres unités lexicales manipulées en traitement automatique du langage, les entités nommées ne se voient renvoyer aucune image précise de la part de la théorie linguistique. Au final, si des critères linguistiques de définition peuvent être dégagés au regard des unités lexicales faisant partie des entités nommées, cet ensemble n'est pas né de la linguistique et ne peut être uniquement défini par ce biais ; il importe donc, et c'est le propos du chapitre suivant, d'identifier d'autres éléments permettant de pleinement le définir.

# Chapitre 5

# Proposition de définition des entités nommées

L'objectif établi dans le préambule méthodologique de cette partie (cf. chapitre 3) est la spécification d'un cadre théorique pour les entités nommés, cadre devant prendre en compte la réalité linguistique de ces unités d'une part, tout comme les besoins réels de la discipline TAL de laquelle elles sont issues d'autre part. Le chapitre précédent s'est intéressé au premier volet de ce programme, l'objet de celui-ci est donc de le compléter par le second.

Il a été établi que, linguistiquement parlant, les entités nommées débordent les catégories classiques et qu'il s'agit en réalité d'un objet pour lequel on ne peut spécifier qu'un type de comportement référentiel (monoréférentialité et autonomie). Du point de vue du TAL, il importe donc de prendre en compte cet aspect référentiel, et c'est ce qui constituera le fil conducteur de ce chapitre.

Ainsi, de même que nous avions défini le sens et la référence en linguistique, il conviendra tout d'abord de spécifier la réalité de leurs équivalents en TAL. Une fois posé ce cadre, non plus d'analyse mais plutôt d'action, il sera alors possible de faire une proposition de définition des entités nommées, tenant compte, et de la réalité linguistique, et des besoins et moyens d'action du TAL. Les conséquences de cette définition sur la tâche de reconnaissance des entités nommées pourront alors être évoquées, tout comme, pour finir, le problème des polysémies.

## 5.1 Sens et Référence en TAL : la notion de modèle

L'étude du langage par les logiciens, les philosophes du langage et les linguistes a déterminé un cadre d'analyse en termes de sens et de référence. La discipline du TAL, ayant pour objet de *traiter automatiquement le langage* (plus précisément les langues naturelles), doit par conséquent tenir compte de ce cadre d'analyse, notamment pour ce qui est de traitements sémantiques. Il convient dès lors de reprendre ces deux notions exposées ci-dessus (cf. section 4.1) selon une perspective linguistique et de les considérer dans le cadre précis, nous pouvons déjà dire restreint, du TAL.

### 5.1.1 La référence en TAL

La discussion sur la référence en linguistique (section 4.1.1) s'est articulée autour de deux points principalement : la question de l'existence du monde ou de la nature du réel auquel il est fait référence par le biais du langage d'une part, et la question (dont la réponse dépend directement de celle apportée à la première interrogation) de l'indépendance de ce réel par rapport au langage d'autre part. Nous avons vu qu'entre une conception objectiviste, selon laquelle la réalité évoquée a une existence objective et extérieure au langage qui sert alors à la nommer, et celle, opposée, du constructivisme, selon laquelle le réel ne peut être que subjectif et construit par le langage lui même, il est une position médiane, proposée par G. Kleiber [Kleiber, 1999b], lequel présente ce qu'il appelle un « réalisme modulé ». Cette troisième voie pose le monde comme extérieur au langage, monde dont on ne peut savoir s'il est réel ou non mais dont une conception partagée sur la base d'une intersubjectivité stable permet de le poser comme privilégié. Cette discussion, d'ordre plus philosophique et ontologique que véritablement linguistique, permet de poser comme base de la référence en linguistique un monde extérieur au langage et peuplé d'entités, réelles ou fictives. Que faire de ce monde lorsque le moyen d'y renvoyer, à savoir le langage, est non plus interprété par un humain mais par un système informatique ? Plus exactement, quelle est ou que peut être la référence en TAL ?

Il importe, pour convenablement caractériser la référence en TAL, de bien comprendre la position de ce dernier vis-à-vis de la linguistique, tant d'une manière générale, avec un rappel de ses objectifs, que particulière, avec une observation de la manière dont le premier exploite ou peut exploiter les connaissances issues de la seconde. Si la linguistique étudie le langage en tant que système, prenant en compte ses aspects phonologiques, lexicaux, syntaxiques, sémantiques et pragmatiques, le Traitement Automatique des Langues a pour objectif la mise au point

de techniques permettant d'automatiser, par le biais de systèmes informatiques, tout ou partie de ce système. Ce programme d'automatisation des divers processus engagés lors d'une activité langagière fait ainsi intervenir, pour l'essentiel, linguistique et informatique conjointement.

Prenons le domaine du traitement automatique de la parole : la reconnaissance vocale, la transcription oral-écrit ou encore la synthèse vocale sont des technologies dont la mise en œuvre s'appuie sur, d'une part, des éléments dégagés par l'analyse linguistique des sons d'une langue (phonologie) et l'analyse anatomique de l'appareil phonatoire et, d'autre part, sur des méthodes de traitement du signal. L'analyse automatique de la langue orale bénéficie donc de connaissances linguistiques de départ relativement « stables », ce qui n'enlève rien à la complexité de sa mise en œuvre.

Considérons ensuite le domaine de l'analyse syntaxique automatique : loin d'être simple elle aussi, cette dernière s'appuie pour sa part sur des éléments dégagés par l'analyse linguistique de la morphologie et de la syntaxe d'une langue, soit sur ce que l'on appelle une grammaire (au sens restreint), ainsi que sur tel ou tel formalisme de représentation de ces données. Ces connaissances de départ paraissent donc, elles aussi, relativement stables : s'il existe un nombre infini de phrases possibles, le nombre de catégories du discours est néanmoins limité, tout comme les règles de combinaison de syntagmes[1].

Enfin, observons ce qu'il se passe lors d'une analyse sémantique : l'objectif de l'automatisation n'est plus ici le traitement de sons ou d'une structure, mais le traitement ce qui est évoqué par ces sons (ou leurs correspondants graphiques) et cette structure. Sur quels types d'éléments une analyse automatique de ce genre peut-elle prendre appui ? Quelles sont les connaissances linguistiques de départ ? Ce qui est alors d'importance ici sont précisément les éléments linguistiques de sens et de référence. Ce cadre d'analyse, s'il permet de rendre compte efficacement d'un point de vue théorique du pouvoir d'évocation et de communication du langage, est cependant difficile à manipuler au sein d'un processus automatique. Nous retombons ainsi sur notre question de départ, que faut-il faire de la référence en TAL, avec cependant des données supplémentaires pour y répondre : si le traitement automatique du langage peut prendre en compte « tels quels » les éléments d'analyse fournis par la linguistique pour le traitement de la parole ou l'analyse syntaxique, il ne peut faire de même pour ce qui est de la référence. Cette dernière correspond au monde dans son entier, avec toute sorte d'entités possibles et ce monde, sans aller jusqu'à considérer ses problèmes de définition, est bien trop « grand » pour un système automatique. Il est impossible d'avoir une vision

[1]Ces observations ne veulent en aucun cas minimiser la difficulté de ces tâches.

universelle en TAL, on ne peut travailler avec le monde tel quel mais seulement avec une vision réduite de ce dernier. Autrement dit, on ne peut considérer l'intégralité de la référence, il faut la *restreindre* et n'en manipuler qu'un fragment. Cette restriction constitue un premier élément de caractérisation de la référence en TAL. Il convient d'y ajouter le fait que cette vision n'est pas directe comme elle peut l'être chez un locuteur, mais qu'elle est le fait d'une *représentation* des éléments ou traits fondamentaux du fragment du monde alors pris en compte. En définitive, il est donc possible de dire que la référence en TAL ne peut être la référence au monde mais qu'elle correspond nécessairement à une *représentation partielle* du monde. Face au « réalisme » modulé et modéré de G. Kleiber posant que les expressions linguistiques « *réfèrent à des éléments 'existants', réels ou fictifs, c'est-à-dire conçus comme existant en dehors du langage* », fait ainsi écho en TAL un réalisme tout autant modéré mais restreint, réduit à une représentation partielle du monde dans lequel ce ne sont pas toutes les entités, réelles ou fictives, qui sont prises en compte mais seulement quelques unes.

Représentation partielle du monde donc, mais comment y parvenir ? Examinons tout d'abord le problème de la restriction de la référence en TAL. Comment s'opère-t-elle ? Il est à noter que cette dernière se pratique immanquablement à chaque réalisation de traitement automatique du langage, de manière naturelle et presque implicite, avec la prise en considération de l'application. Le TAL se détermine en effet essentiellement par des besoins applicatifs correspondant au traitement de tel ou tel processus dans tel ou tel domaine, ce dernier pouvant relever d'activités extrêmement diverses. La considération de l'application est donc une réponse à la question de la restriction de la référence, c'est elle qui détermine ce à quoi on s'intéresse dans le monde, ce que l'on cherche à repérer. Pour une application concernant le domaine de la construction aéronautique, la référence est réduite aux métiers de la conception, de la fabrication et de la commercialisation des avions, tandis que pour une gestion des risques liés à l'informatique, la référence est réduite au domaine des réseaux informatiques, des logiciels, des accès internet, etc.

Se pose alors la question de comment *représenter* le résultat de cette restriction ? Il faut passer par un *modèle* de la portion du monde considérée. Du latin *modus* signifiant « mesure », un modèle renvoie à l'origine à « la représentation en petit de ce qui sera reproduit en plus grand ». Son acception scientifique, celle qui nous intéresse, apparaît dans les années 1950 et connaît depuis un grand succès dans le monde informatique : il s'agit d'un système représentant les structures essentielles d'une réalité et capable à son niveau d'en expliquer ou d'en reproduire dynamiquement le fonctionnement. S'il n'est donc plus question de taille et d'ordres de grandeur dans cette acception, la notion d'analogie ou d'équivalence

entre l'objet considéré et son modèle demeure primordiale, doublée du fait plus pragmatique que le modèle doit pouvoir répondre de manière satisfaisante aux questions ou problèmes qui se posent au regard de l'objet modélisé. Cette étape de modélisation est donc importante, permettant en quelque sorte de déterminer l' « espace de travail » de telle ou telle application TAL, tout en conditionnant déjà les moyens de sa réalisation. Pour représenter une portion de la référence, il faut la modéliser, en choisissant les entités à prendre en compte et en en déterminant les interactions. L'idée importante ici est bien celle de modélisation, il n'est pas question du formalisme permettant de l'exprimer. A titre d'illustration, prenons de nouveau l'application de la construction aéronautique : le modèle du monde comprendra des entités de type 'constructeur aéronautique', 'avion' ou 'équipement' par exemple, tandis que pour le risque informatique, le modèle du monde comprendra des entités de type 'logiciel', 'virus' ou 'site internet'.

Un autre parallèle avec la référence en linguistique est alors possible : si le renvoi, par le biais du langage, à ce que l'on considère comme le monde est garanti, nous dit Kleiber, par une modélisation intersubjective stable (à apparence d'objectivité), le renvoi à une partie du monde en TAL est au contraire soumis à variation, le monde en question étant alors modélisé différemment (subjectivement) par chaque application. Le partage de la référence en TAL ne peut se faire, en l'état actuel des choses, entre différentes applications, ce partage ne s'opère qu'entre la technologie TAL et ses concepteurs et utilisateurs[1].

Pour résumer, la référence en TAL ne peut être la référence au monde, le renvoi opéré par les expressions liguistiques ne peut se faire qu'à une portion de ce monde seulement, celui-ci étant alors représenté au moyen d'un modèle. Ainsi, si la référence en linguistique peut être définie par la formule suivante (il s'agit de celle présentée en 4.1.1) :

> La *référence* (en linguistique) désigne le lien qui existe entre une expression linguistique et l'élément du réel (appelé *référent*) auquel elle renvoie,

il est possible de proposer, pour la référence en TAL cette fois-ci :

> La *référence* (en traitement automatique du langage) désigne le lien qui existe entre une expression linguistique et l'élément du *modèle* (appelé *référent*) auquel elle renvoie,

où le *modèle* correspond à une représentation partielle du monde, ou du réel.

À la suite de cette tentative de caractérisation de la référence en TAL, il convient de s'intéresser à son indispensable complément : la notion de sens en

[1]Ce constat peut être nuancé quelque peu : lorsque l'inventaire des individus est partagé entre deux applications, alors on peut parler de partage de la référence.

TAL.

### 5.1.2 Le sens en TAL

Dans un article intitulé *Sens et traitement automatique des langues* [Sabah, 2000], G. Sabah donne une vue d'ensemble de cette problématique, de laquelle il ressort deux choses principalement : premièrement, le sens occupe une place centrale en TAL (« *le sens est (ou tout du moins devrait être) partout dans le traitement automatique des langues* »), et ce depuis les débuts de l'intelligence artificielle jusqu'à aujourd'hui, passant d'une « *vision simpliste* » à une « *prise en considération totale, dépassant la linguistique même* », mais, et c'est le second point, il reste difficile à traiter de manière automatique, en témoignent la variété des formalismes proposés pour représenter les connaissances qu'il met en jeu (des différentes logiques aux réseaux sémantiques), voire les interrogations portant sur la possibilité même du traitement du sens en machine (il n'est pas prouvé que l'homme traite le langage de façon formelle). A l'instar du sens en linguistique, le sens en traitement automatique des langues apparaît ainsi comme une donnée complexe. Nous tenterons ici de donner une caractérisation générale du sens en TAL, partant de ce qui a précédemment été évoqué dans la section 4.1.2 pour aboutir aux moyens dont dispose actuellement le TAL pour traiter du sens ainsi qu'aux usages qu'il en fait.

#### 5.1.2.1 Rappel sur le sens en linguistique

Le sens est ce qui conduit à la référence. Analysé par Frege comme contenant le « mode de donation » du référent, le sens d'une expression linguistique détermine en effet si telle ou telle chose du réel peut par elle être désignée. Au delà de cette distinction fondamentale entre ce à quoi on réfère et ce qui permet d'y parvenir, la question du sens n'est pas facile à démêler : le sens est partout et il importe de déterminer des « niveaux d'étude », du mot au texte en passant par la phrase ; le sens éveille de nombreuses interactions et il importe d'avoir à l'esprit tant les aspects linguistiques, cognitifs que psychologiques ; enfin, le sens revêt un caractère composite, étant le résultat d'une interaction entre ce que l'on appelle des primitives de sens et le contexte. Ces questions du niveau d'étude, de la nature du sens et de ses « ingrédients » ont pour conséquence la multiplication des cadres d'analyse et des théories, suivant les choix opérés et l'importance accordée à l'un ou à l'autre des critères (sémantique lexicale, sémantique textuelle, sémantique cognitive, sens référentiel ou aréférentiel, etc.). Lors de la section 4.1.2, nous nous étions placée dans le cadre d'une sémantique lexicale et avions retenu la définition d'un sens référentiel au sein d'un modèle

hétérogène, distinguant sens descriptif et sens instructionnel. Ces deux types de sens donnent des informations permettant d'identifier le référent, mais le font de manières différentes ; le premier au moyen d'une description du référent par un ensemble de traits objectifs car intersubjectivement partagés (c'est ainsi que l'on aura tendance à qualifier le sens du mot *téléphone* de descriptif ou dénotatif car il comporte les traits 'appareil permettant de communiquer', 'fixe ou mobile', etc. qu'un référent doit satisfaire pour être désigné par lui) ; le second type de sens conduit au référent au moyen de l'indication de la procédure à suivre pour le trouver (le sens du mot 'je' est ainsi instructionnel car il indique de prendre en compte la personne qui parle ou qui écrit). Dans ce cadre, nous avons observé (cf. section 4.1.2.2) que le sens comporte une « part linguistique » avec des éléments linguistiques conventionnels et préconstruits, et une « part contextuelle » avec des éléments issus du contexte, ce dernier étant lui même composé de connaissances sur la situation d'énonciation d'une part et de connaissances sur le monde d'autre part. Ainsi, ce qui nous permet de comprendre le sens et d'être conduits vers la référence correspond à une combinaison plus ou moins équilibrée entre des connaissances lexicales, des connaissances contextuelles (contextuelle au sens de situation d'énonciation) et des connaissances encyclopédiques. Étant donné ce cadre d'analyse et de fonctionnement du sens en linguistique, quel est actuellement son correspondant en TAL ?

#### 5.1.2.2 Caractérisation du sens en TAL

Si nous avons besoin de connaissances nous permettant de construire et de comprendre le sens pour référer au monde, il en est de même pour le TAL, quand bien même ses objectifs sont réduits à une vision restreinte du monde. Aux connaissances nécessaires à la compréhension humaine d'une expression linguistique dans une situation donnée correspondent ainsi en TAL des *ressources* permettant à un système informatique de « comprendre » une expression linguistique dans une application donnée. A l'image de la référence, le sens en TAL doit néanmoins être restreint par rapport à son homologue linguistique. Cette restriction est double, portant sur les ressources d'une part, et sur les usages du sens (qu'est-ce qu'un système informatique peut « comprendre ») d'autre part.

**Les usages du sens en TAL** Examinons tout d'abord les usages (actuels) du sens en TAL. Un système informatique « lisant » ou plutôt analysant un texte ne peut aujourd'hui en afficher une compréhension similaire à celle d'un humain. L'intérêt ici n'est pas de comparer les performances de la machine face à celle de l'homme, celle-là étant somme toute, par ses capacités de traitement, complémen-

taire de celui-ci, mais de déterminer à quoi correspond l'exercice du sens en TAL. Ce dernier se présente sous la forme de deux opérations principalement : classer et reformuler[1]. La classification correspond à l'attribution d'une catégorie de nature sémantique à un mot ou syntagme, à un paragraphe, à un texte ou encore à une collection de textes. La catégorie peut appartenir à un ensemble plus ou moins structuré, allant de la simple liste « plate » aux ontologies. Cet usage du sens est à l'œuvre dans la catégorisation de documents et l'annotation d'entités nommées par exemple. Le second usage du sens en TAL, la reformulation, est une opération davantage interne à la langue : il s'agit de donner, pour des mots, des paragraphes ou des textes, des mots ou paraphrases équivalents sémantiquement. Au niveau lexical, la reformulation correspond à de la désambiguïsation et peut être mise en œuvre pour les noms d'une part, et pour les autres catégories, verbes, adjectifs, etc. d'autre part. La désambiguïsation nominale rejoint en quelque sorte l'opération de catégorisation dans la mesure où on s'intéresse ici au niveau le plus fin de la classification : pour le mot *barrage* par exemple, il s'agit d'identifier l'objet du monde dont il est question, un barrage de police ou un barrage hydraulique. Pour les autres catégories syntaxiques, le rapport à la référence est moins présent et la désambiguïsation prend une dimension plus « linguistique », le problème pour les mots exprimant des qualités et des processus étant plus de comprendre le sens du mot que d'indiquer l'objet du monde en question. Au niveau lexical, la reformulation correspond ainsi à l'indication d'un synonyme ou au « pointage » vers un sens précis ; cette opération de reformulation lexicale est d'importance, elle paraît même indispensable pour la traduction automatique, la recherche d'information et les tâches de question-réponse (entre autres). Au niveau du paragraphe ou du texte, la reformulation équivaut au résumé automatique, avec la production d'un texte plus court que l'original, ou à la traduction, avec la production d'un texte dans une autre langue que celle de l'original. Cette description de quelques applications et processus TAL comportant une dimension sémantique est loin d'être exhaustive, toutefois, classer et reformuler semblent être deux opérations de base en TAL, permettant de couvrir la plupart des besoins reliés au sens. Classer et/ou reformuler ne permettent cependant pas de couvrir l'ensemble des interactions complexes mises en jeu lors de la compréhension de données langagières, les usages du sens en TAL sont ainsi, pour le moment, plus restreints que ceux envisageables par un cerveau humain et décrits dans la théorie linguistique.

**La nécessité de ressources** Pour mener à bien ces deux grands types d'usages du sens en TAL couvrant une diversité de traitements, difficiles à réaliser pour la plupart, un système informatique requiert des ressources. Ces dernières peuvent

[1]Ce à quoi on pourrait ajouter l'opération préliminaire de *segmentation*.

être interprétées symétriquement aux connaissances sollicitées lors d'un processus sémantique, connaissances présentées ci-dessus comme étant de nature lexicale, encyclopédique ou encore contextuelle (contexte toujours au sens restreint de situation d'énonciation). Qu'est-ce que les connaissances lexicales et encyclopédiques en TAL ? De même, à quoi correspond la situation d'énonciation ?

Les connaissances lexicales nécessaires à la compréhension sont reproduites en TAL sous la forme de ressources lexicales. Le terme de lexique en TAL renvoie à la notion (lexicographique) de recueil de mots et non à celle (linguistique) de la totalité des mots d'une langue. Un lexique, ou base de connaissance lexicale, a pour objet de décrire sous format numérique, donc exploitable par un système informatique, des mots dans leurs différents sens, leurs relations et leurs emplois et peut prendre différentes formes suivant l'organisation de cette description, dictionnaire, thésaurus ou terminologie (voir [Habert *et al.*, 1997] pour plus de détails). La construction de ce type de ressource prend appui sur des données linguistiques relativement stables, l'étude du lexique (au sens linguistique), et devient effective soit par simple translation d'un dictionnaire papier à un format électronique (cette correspondance imprimé-numérisé existe aujourd'hui pour presque tous les dictionnaires disponibles dans le commerce), soit par le biais d'un véritable travail d'élaboration d'une ressource lexicale numérique en fonction d'une théorie précise, à l'instar du thésaurus Wordnet conçu par une équipe de psycholinguistes de l'Université de Princeton et en cours d'élaboration depuis le début des années 1990 [C.Fellbaum, 1998, Habert *et al.*, 1997]. Ces ressources lexicales, si elles sont disponibles en grand nombre à l'heure actuelle, ne permettent cependant pas à un système automatique de tout faire, ce dernier restant nécessairement dépendant de la couverture et de l'adéquation de la ressource avec l'application visée. En somme, si un système informatique peut disposer d'informations lexicales, ces dernières ont nécessairement un aspect contingent (qu'il s'agisse de la simple acquisition — avec tous les aspects juridiques et financiers que cela peut comporter —, du domaine couvert, de l'organisation des données, etc.) qui restreint peu ou prou le domaine d'action par rapport aux connaissances lexicales mobilisées par la compréhension humaine : un système informatique ne possède, comme information, que ce qu'on lui donne.

Le processus de compréhension du sens d'une expression linguistique mobilise également d'autres connaissances, de nature non linguistique cette fois : les connaissances générales sur le monde (encyclopédique) et les connaissances sur la situation d'énonciation (pragmatique). De même que pour les connaissances lexicales, un système informatique ne peut avoir comme connaissances encyclopédiques que ce qu'on lui donne. Ces dernières sont présentes sous deux formes différentes. Il y a d'une part les ressources encyclopédiques proprement dites, se

présentant sous la forme d'ontologies ou de bases de données encyclopédiques numérisées (de type Wikipédia). Par rapport aux ressources lexicales évoquées ci-avant, la disponibilité de ce type de ressource est cependant moindre et, étant donnée la difficulté que représente leur constitution (il est pour le moment hors de portée de construire une base de connaissance encyclopédique à caractère général), leur couverture reste limitée le plus souvent à un domaine bien précis. Un système informatique peut également disposer d'informations de nature encyclopédique par le biais du modèle spécifiant les entités du monde à prendre en compte ainsi que leurs relations. À titre d'illustration, il est possible de prendre une application médicale : un système informatique pourra disposer de connaissances encyclopédiques issues d'une base de données médicale comme Medline d'une part, tout en ayant comme informations issues de son modèle le fait qu'un *patient* a généralement un *dossier médical*. De même, si le logiciel Word « sait » que tout ce qui comporte un arobase est une adresse e-mail, c'est qu'il connaît l'existence de ce genre d'entité grâce à son modèle d'une part, et qu'il possède l'information lexicale relative au « sens » d'un arobase d'autre part. Il n'en est rien pour Wordpad qui, manifestement, ne dispose pas des mêmes connaissances. Ressource et modèle, la notion de connaissance encyclopédique pour un système informatique est de fait elle aussi limitée.

Reste la situation d'énonciation, soit l'ensemble des composantes de l'environnement dans lequel s'inscrit la production-réception d'une expression linguistique. Comment un système informatique peut-il avoir accès à ces composantes ? Au regard de tous les éléments de la situation d'énonciation qui, potentiellement, peuvent infléchir l'interprétation d'une expression linguistique, on se trouve contraint, de nouveau, de qualifier de limité l'accès aux connaissances contextuelles dont peut bénéficier un système informatique. En effet, si un geste, une vision ou encore une odeur peuvent influencer l'interprétation dans une situation d'énonciation donnée, force est de constater qu'un système informatique ne peut avoir accès qu'aux seules informations relatives au corpus. Si elles existent, ces dernières représentent un ensemble de métadonnées tout de même fort utile pour la compréhension, par le système, du contenu même du corpus. On retrouve dans ces métadonnées les informations classiques relatives à toute situation d'énonciation, soit la triade référentielle de la personne, du temps et du lieu. Elles peuvent en effet offrir des informations concernant la date de production du document ou la période temporelle dont il est question dans celui-ci, des informations relatives à son auteur et à sa source (organisme éditeur), ou encore des informations à propos du lieu de production du document, de son thème et de son intention. Prenons un corpus journalistique d'un pays donné pour une période donnée, par exemple le corpus du journal *Le Monde* pour les années 2000-2004 : à l'aide des métadonnées

disponibles pour ce corpus, il est possible de savoir que, par défaut (et uniquement dans ce cas), une expression comme *le Président* renvoie à un Président de la République Française durant les années 2000-2004. En revanche, il conviendra de l'interpréter autrement s'il s'agit d'un corpus sur les relations sociales chez EDF (cf. travaux de B. Habert et H. Folch dans le cadre du projet *Scriptorium* initié par le département Recherche et Développement d'EDF dans le but d'identifier des thématiques dans le discours de l'entreprise, [Floch et Habert, 2000]). Ainsi, certains corpus sont porteurs de beaucoup d'information par défaut, d'autres moins ; quoi qu'il en soit, ces informations doivent théoriquement être utilisées comme des connaissances contextuelles.

Au final, si le sens en linguistique correspond à une donnée composite dont la compréhension par un humain engage un processus complexe s'appuyant sur diverses connaissances, force est de constater que l'exercice du sens par un système informatique ne peut en être le strict équivalent. Deux opérations résument les usages du sens en TAL, classer et reformuler ; ces opérations sont effectuées à l'aide de connaissances diverses elles aussi, mais limitées : les informations de nature lexicale, encyclopédique ou encore contextuelle dont peut disposer un système informatique ne sont que celles qui lui sont fournies au préalable.

#### 5.1.2.3 Articulation sens-référence en tal

Où en sommes-nous ? Après avoir défini la référence en TAL, nous venons d'examiner ce à quoi correspond, dans ce même domaine, le sens ainsi que les moyens de sa mise en œuvre. Dès lors, de même que nous avions souligné que le sens est ce qui conduit à la référence en linguistique, il convient de préciser en quoi, en TAL, le sens donne également accès à la référence. Dans la vie de tous les jours, la référence se fait entre le langage et le monde, s'appuyant sur des mécanismes complexes de compréhension et d'interprétation d'expressions linguistiques dans une situation d'énonciation donnée. En TAL, la référence se fait entre le langage et un modèle représentant une partie du monde seulement, s'appuyant sur deux mécanismes principalement, la classification et la reformulation. Ainsi, de même que, dans une conversation par exemple, la compréhension d'une expression linguistique permet à l'interlocuteur d'accéder au référent désigné, les deux opérations (ci-dessus mentionnées) permettent, à l'aide de ressources diverses, de calculer la référence entre une expression d'un corpus et une entité (quelle que soit sa nature) d'un modèle donné d'une application donnée. Si un système informatique parvient par exemple à classer le mot *table* en tant que meuble plutôt qu'en tant qu'outil d'apprentissage d'une opération mathématique, alors la référence peut être faite avec la catégorie « meuble » d'un modèle destiné à traiter des informations pour

le compte de magasins d'ameublement. De la même manière, prenons par exemple un modèle général du domaine de la musique, comprenant des « musiciens », des « concerts », des « instruments », etc. : le fait de classer le mot *violon* en tant qu'instrument de musique et non en tant que musicien permet de construire une occurrence de la classe « instruments » du modèle. Ainsi, du corpus au modèle, on voit bien comment opère la référence en TAL.

Le cadre d'analyse en terme de sens et de référence proposé par la linguistique trouve donc un écho en TAL sous la forme d'un « cadre d'action » à l'intérieur duquel un système informatique peut évoluer et mener à bien divers traitements. Ce cadre d'action possède deux caractéristiques principales. Il est tout d'abord restreint par rapport au cadre théorique d'origine, cette limitation étant la conséquence du fait que la référence en TAL ne peut être la référence au monde mais seulement une référence à une représentation partielle du monde d'une part, et du fait que le sens en TAL est nécessairement dépendant des ressources fournies au système informatique et correspond à deux usages principalement, classer et reformuler, d'autre part. La seconde caractéristique de ce cadre d'action est directement liée à la première : si le sens et la référence en TAL se voient limités, cette limitation doit indispensablement être déterminée et d'avance prise en compte. Ayant ainsi caractérisé, ou plutôt circonscrit, le champ d'action potentiel d'un système informatique à l'intérieur du cadre théorique sens-référence, il est temps de considérer les entités nommées.

## 5.2 Les entités nommées

Commençons tout d'abord par rappeler les données précédemment dégagées au sujet des entités nommées et disponibles à ce stade de l'analyse. Le détour du côté de la théorie linguistique (cf. chapitre 4) nous a permis de montrer que ce qui intéresse et justifie l'ensemble 'entités nommées' n'est aucunement en rapport avec la forme mais bien plus en rapport avec le fonctionnement référentiel. Il est bien sûr possible de parler de catégories syntaxiques et donc d'affirmer que cet ensemble est majoritairement composé de noms propres, de descriptions définies et d'expressions numériques et temporelles, mais ce qui paraît le plus constant, le plus stable, sont les caractéristiques référentielles de ces unités dont la principale fonction est de pointer un élément du réel. Le rapport à la référence apparaît en effet comme fondamental pour ces unités, allant jusqu'à éclipser le sens de certaines, celui-ci se réduisant alors à une instruction de dénomination, et obli-

geant, pour l'interprétation des autres, à user d'indices ou points référentiels. Ces spécificités linguistiques des entités nommées s'inscrivent dans le cadre d'analyse sens-référence, lequel trouve un équivalent restreint en TAL sous la forme d'un cadre d'action circonscrit par un modèle et disposant de ressources. C'est en prenant en compte ces différents éléments qu'il est possible de proposer une définition des entités nommées.

### 5.2.1 Proposition de définition

Une entité nommée est une expression linguistique monoréférentielle. La monoréférentialité désigne la capacité d'une unité linguistique de renvoyer à une entité extralinguistique ou référent unique. Lors de l'étude des catégories linguistiques existantes au sein des entités nommées, nous avons vu, d'une part, que le nom propre opère une identification individualisante, renvoyant à un particulier distingué de ses semblables, et, d'autre part, que les descriptions définies ont comme propriété référentielle intrinsèque d'opérer un renvoi à une entité unique. Si cette monoréférentialité des noms propres et des descriptions définies est valable en linguistique relativement à un monde intersubjectivement stable et à une situation d'énonciation donnée, il ne peut en être de même pour les entités nommées. Ces dernières sont des unités manipulées en TAL, mode de traitement automatique du langage à l'intérieur duquel nous avons vu que la référence ne peut être la référence au monde dans son entier mais seulement une référence à une représentation partielle de ce monde, ou modèle du monde. Partant, la monoréférentialité des entités nommées ne peut avoir de sens que par rapport à un modèle du monde préalablement défini, elle ne peut être effective, se préciser et s'organiser qu'à l'intérieur de ce modèle. Ce dernier est établi tout naturellement en fonction de l'application, sollicitant la prise en compte de telles entités plutôt que telles autres. Ainsi, si les deux descriptions définies suivantes, *le Président de la République* et *le costume bleu du Président* renvoient toutes deux à une entité unique, elles ne deviennent entité nommée que si un modèle applicatif requiert leur identification et annotation. Rien n'est entité nommée par « nature », seulement des unités linguistiques monoréférentielles peuvent le devenir, et ce dans le cadre d'une modélisation applicative uniquement. Cette caractéristique de monoréférentialité ne suffit cependant pas à elle seule à spécifier ce qu'est une entité nommée, le mot *je* est en effet lui aussi monoréférentiel, sans pour autant être une entité nommée.

Une entité nommée est une expression linguistique autonome. On dit d'une expression linguistique qu'elle est autonome référentiellement quand elle peut, par ses seules ressources, évoquer un référent [Charolles, 2002b]. Les noms propres

ainsi que les descriptions définies fonctionnent tous deux sur un mode autonome, pouvant à eux seuls instancier un particulier ou une entité unique, du fait d'une convention dénominative pour les premiers, d'une description plus ou moins complète du référent pour les secondes (dont l'autonomie peut alors connaître des degrés différents). Dans les deux cas, l'autonomie n'est possible que relativement à des connaissances encyclopédiques et à des connaissances sur la situation d'énonciation, dont nous avons vu qu'elles dépendent, pour les premières, de ressources fournies au système informatique et, pour les secondes, des informations disponibles à propos du corpus.

Eu égard à ces caractéristiques, il est alors possible de spécifier ce qu'est une entité nommée et de proposer la définition suivante :

> Étant donnés un modèle applicatif et un corpus, on appelle *entité nommée* toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Précisons quelque peu cette formule, en revenant sur le *modèle applicatif*, le *corpus* et les *expressions linguistiques*. Le modèle applicatif, comme cela a été spécifié dans la section 5.1.1 sur la référence en TAL, correspond à la détermination et à la structuration d'un ensemble de données pertinentes au regard d'une application. Les modèles sont souvent de type entité-relation, spécifiant, selon un niveau conceptuel, les objets à traiter d'une part, et les relations entre ces objets d'autre part. Un modèle peut être plus ou moins développé, suivant qu'il contient un grand nombre de classes avec de nombreuses relations ou non. Imaginons par exemple que le Centre d'Etude de la Vie Politique Française ait besoin d'un système automatique de traitement des informations relatives à ce domaine. Un modèle complexe peut être déterminé, représentant de nombreuses entités telles que « personne », « parti politique », « élection », « syndicat », « mouvement de grève », ces entités étant liées par des relations des type « affiliation » et autres. Un tel modèle pourra, par sa richesse, offrir une représentation plus ou moins complète des personnes politiques, avec l'indication de leur affiliation à tel ou tel parti politique ou leur victoire à telle ou telle élection. À l'inverse, il est tout aussi possible de concevoir un modèle plus simple, n'ayant comme représentation du monde que les entités « personne », « organisation » et « événement », sans relations entre elles, modèle ne pouvant dans ce cas offrir qu'une représentation restreinte des personnes, politiques ou non. L'intention ici n'est pas de présumer de l'efficience ou de la supériorité de tel type de modèle sur tel autre mais d'expliquer que, d'un point de vue théorique, une modélisation peut se faire selon divers degrés de précision, lesquels influent nécessairement sur les résultats des traitements.

Le corpus correspond quant à lui à l'ensemble des données langagières auxquelles est appliqué le traitement. Un corpus peut être de diverses natures (journalistique, technique, universitaire, littéraire, etc.) tout comme plus ou moins spécialisé. Afin d'illustrer l'influence du corpus pour ce qui est de la spécification des entités nommées, prenons l'exemple (extrême) d'un corpus épistolaire, celui de la correspondance de G. Flaubert ; dans cet ensemble de textes, le narrateur étant toujours la même personne, le pronom personnel *je* est toujours saturé référentiellement, c'est-à-dire qu'il est monoréférentiel et surtout autonome, pouvant par conséquent et en toute logique devenir une entité nommée. Il est bien peu probable de trouver une application en TAL s'intéressant aux lettres adressées à Louise Colet ou à George Sand, il n'en demeure pas moins que la nature et le degré de spécialisation du corpus jouent un rôle non négligeable dans la spécification de ce qui est ou non autonome, et donc dans la définition des entités nommées pour une application donnée.

Enfin, les expressions linguistiques correspondent, de manière privilégiée, aux noms propres : ces derniers ont un rapport direct à la référence et désignent des particuliers de manière autonome. D'autres types d'unités linguistiques possédant peu ou prou les mêmes propriétés peuvent s'adjoindre aux noms propres, telles les descriptions définies.

Issue d'explorations complémentaires menées en linguistique et en traitement automatique des langues, cette proposition de définition tente ainsi de spécifier et de consolider le statut théorique des entités nommées. Une entité nommée n'existe pas en soi, elle ne relève d'aucune catégorie linguistique préexistante dans la littérature ; si elle est caractérisable d'un point de vue linguistique par le biais de certaines propriétés — et ces dernières sont indispensables car elles balisent le champ des possibles au niveau des expressions linguistiques—, elle n'existe que relativement à un besoin applicatif précis, général ou spécifique, et donc relativement à un modèle représentant ce besoin applicatif.

Cette formule définitoire permet par ailleurs d'apporter des éléments de réponses aux observations réalisées antérieurement (cf. chapitre 2 et chapitre 3). Se trouve tout d'abord expliquée l'*hétérogénéité* de l'ensemble 'entités nommées', lequel est composé d'une collection d'expressions linguistiques diverses, réunies sur la base de caractéristiques référentielles communes, au delà de la traditionnelle partition linguistique en catégories syntaxiques. On comprend mieux, par ailleurs, la *variabilité* de composition de cet ensemble, ou pourquoi, en fonction du modèle applicatif pris en compte, une unité linguistique monoréférentielle et autonome peut être une entité nommée dans un cas, et non dans l'autre. Il est possible de parler d'entités nommées en général, mais ces dernières composent un

ensemble hétérogène et variable pour lequel il n'est pas possible de proposer de définition dans l'absolu. L'ensemble des réalisations (cf. figure 4.2) est très hétérogène, mais chaque réalisation ne considère qu'un fragment de cette hétérogénéité et se trouve une cohérence propre, en fonction de son modèle. Seul le TAL a pu « inventer » cet ensemble des entités nommées, analysable en des termes linguistiques mais nécessairement solidaire de besoins applicatifs. Cette caractérisation de la notion d'entité nommée se doit d'être complétée par une illustration.

### 5.2.2 Illustration

Avant de rentrer dans les détails de l'illustration de cette proposition de définition, il importe de considérer rapidement ce qu'il advient du calcul des entités nommées dans ce nouveau cadre. En effet, les précisions apportées à l'égard des expressions linguistiques pouvant être des entités nommées changent la donne au niveau des objectifs d'identification et de catégorisation de ces entités. Si l'on prend par exemple une application s'intéressant aux hommes politiques, alors les expressions suivantes sont toutes intéressantes : {*Jacques Chirac*, *le Président de la République de 2002*, *l'ancien maire de Paris*, *le dernier Président du RPR*}. Si ces expressions, parce qu'elles désignent une entité unique et sont autonomes relativement à un corpus donné, peuvent chacune accéder au statut d'entité nommée dans cette application précise, il n'est cependant pas possible de toutes les identifier et de les annoter. En effet, un système informatique est dépendant des connaissances qui lui sont fournies. Pour ce qui est des entités nommées, un tel système travaille généralement avec des lexiques d'une part et des règles de reconnaissance à partir d'amorces d'autre part (cf. section 1.4.1). Les lexiques se présentent sous forme de listes de noms, lesquels se voient associer une catégorie sémantique précise. Ce type de ressource de nature encyclopédique est la plus utilisée car pour le moment la plus efficace pour le traitement des noms propres, ces derniers n'ayant qu'un sens de dénomination indiquant de chercher et de trouver dans la mémoire stable le référent qui porte le nom en question. En effet, aucun calcul compositionnel ne peut être mis en œuvre pour les noms propres et une façon de les comprendre, pour l'homme comme pour la machine, est de connaître la convention qui instaure un lien dénominatif entre telle expression linguistique et tel référent, ainsi que la catégorie d'existant à laquelle il appartient. Ces informations sont présentes dans les lexiques utilisés pour le traitement des entités nommées, c'est pourquoi, relativement à notre exemple, il est possible d'identifier l'entité nommée *Jacques Chirac*. La compréhension des autres expressions, qui sont des descriptions définies, nécessite en revanche un calcul compositionnel, indiquant par exemple que *le Président de la République de 2002* renvoie à

une entité unique qui *préside* un pays dont le régime politique est de type *République*, etc. Il est plus difficile à l'heure actuelle pour un système informatique de mener à bien ce type de calcul, c'est pourquoi les descriptions définies que l'on veut annoter en tant qu'entités nommées sont majoritairement traitées à l'aide de lexiques. En somme, il existe (pour le moment uniquement) plus d'entités nommées qu'on ne sait en traiter. Il est à présent possible de revenir au point de départ et de considérer, en s'appuyant sur les éléments définitoires dégagés dans la section précédente, quelques-unes des expressions linguistiques figurant dans l'échantillon d'unités lexicales considérées comme des entités nommées (figure 5.1 reproduite ci-après pour mémoire).

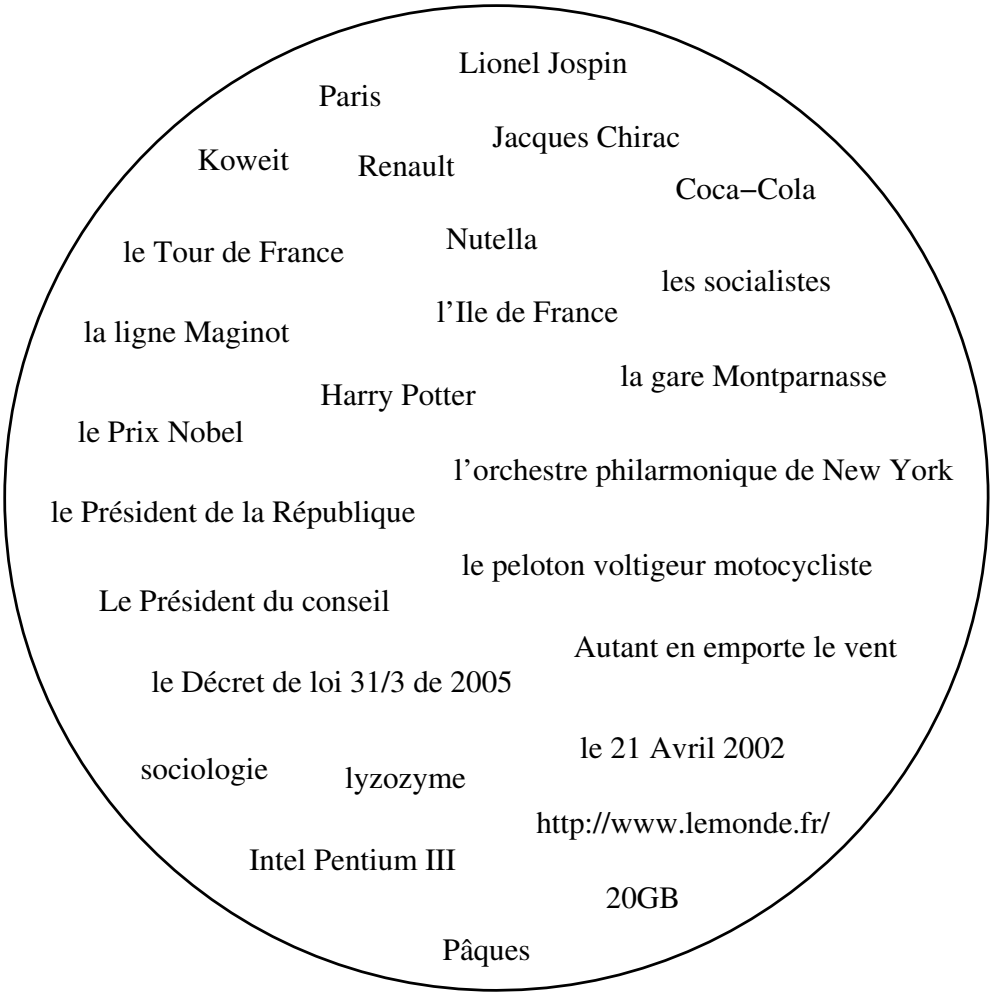



FIG. 5.1 – Echantillon d'unités lexicales considérées comme des entités nommées.

Considérons tout d'abord les unités linguistiques *Lionel Jospin* et *Paris*. Noms propres, elles réfèrent toutes deux de manière autonome à un référent unique, une personne pour la première, une ville pour la seconde (plusieurs villes différentes répondent au nom de *Paris*, nous y reviendrons plus tard dans la section consacrée à la polysémie). Par définition, ces expressions linguistiques sont ainsi d'ex-

cellentes candidates au statut d'entité nommée, à la condition néanmoins qu'un modèle applicatif s'intéresse aux noms de personnes et aux noms de lieux dans des corpus de nature journalistique (voire universitaire). Cela est très souvent le cas, la plupart des applications requièrent en effet l'identification et l'annotation de telles unités, en raison, cela vient d'être dit, de leur parfaite adéquation aux « propriétés requises » (monoréférentialité et autonomie) d'une part, et, d'autre part, parce qu'elles renvoient à des êtres ou entités singulières appartenant à des catégories d'existants « *avec lesquels nous entretenons un commerce suffisamment régulier et pointilleux* », et donc facilement modélisables. Ces expressions, ainsi que d'autres similaires (*Koweit*, *Jacques Chirac*, etc.), sont de la sorte dans de nombreuses applications, et ce quelle que soit l'organisation du modèle, distinguant seulement les personnes ou bien s'intéressant précisément aux hommes politiques, considérant globalement les lieux géographiques ou prenant en compte plus particulièrement les villes. Il s'agit en quelque sorte d'entités nommées « universelles », même si, au regard de la formule proposée ci-avant, ceci constitue un abus de langage. Malgré cette universalité, il est cependant possible d'imaginer des applications pour lesquelles les noms de personnes et de lieux, même nommés par des unités lexicales autonomes, ne sont d'aucune utilité : c'est le cas pour la reconnaissance d'entités nommées opérée dans le projet CrossMarc s'intéressant uniquement aux entités relatives aux ordinateurs portables (nom du processeur, capacité de mémoire, capacité du disque dur, etc.) sur des pages web de vente de matériel informatique [Karkaletsis *et al.*, 2003].

Examinons à présent *le Président de la République* : dans quels cas cette expression linguistique peut-elle être considérée comme une entité nommée ? Il s'agit là d'une description définie incomplète, dont l'identification du référent dépend d'un calcul compositionnel d'une part (il faut en effet comprendre qu'un Président de la République *préside* une *République*) et d'un calcul instructionnel d'autre part (les coordonnées spatio-temporelles dans lesquelles s'inscrit cette expression modifient en effet son interprétation). En théorie, si l'on parvient à interpéter correctement cette expression en calculant son côté compositionnel comme son côté instructionnel, elle possède alors les caractéristiques linguistiques nécessaires d'autonomie et de monoréférentialité pour pouvoir devenir une entité nommée. Il est possible d'imaginer deux cas de figure dans lesquels cette expression peut devenir une entité nommée, suivant les moyens dont on dispose pour mettre en œuvre un calcul compositionnel. Dans les deux cas, la prise en compte d'une telle expression en tant qu'entité nommée réclame un modèle doté d'une structuration plus détaillée que celui ne comportant par exemple que la catégorie « individu ». Examinons le premier cas de figure : *le Président de la République* peut devenir une entité nommée dans le cadre d'un modèle relativement simple, associant

un attribut de type « titres identifiants » à la catégorie « individu », attribut pour lequel on décide alors, en fonction d'une application précise s'intéressant par exemple aux faits politiques, qu'il est entité nommée. Sans plus d'information dans le modèle, le calcul proprement dit de cette entité nommée n'est pas possible, il faut par conséquent passer par l'indication contenue dans un lexique (cf. paragraphe ci-dessus sur le calcul des entités nommées). Dans ce cas, l'entité nommée n'est pas calculée mais « décidée » à l'avance en quelque sorte. Pas de calcul compositionnel donc, mais il importe tout de même de restituer l'autonomie de cette expression, c'est-à-dire d'indiquer, si possible, de quelle République et de quelle année il s'agit. Ces informations peuvent être déduites des métadonnées disponibles à propos du corpus ; si cela n'est pas possible, alors on comprend mieux pourquoi une expression telle que *le Président de la République française* est plus à même de devenir une entité nommée. Considérons maintenant le deuxième cas de figure. Cette expression peut devenir entité nommée dans le cas d'un modèle très détaillé, s'intéressant exclusivement aux institutions politiques européennes par exemple et comportant alors des entités de type « régime politique »,« institutions »,« constitution", etc. À l'aide de ressources lexicales et des informations contenues dans le modèle, il est alors possible de calculer que si quelqu'un préside une République, ce ne peut être le dirigeant du Royaume-Uni mais plutôt celui, en fonction de l'année et de la République, d'un pays comme la France ou l'Italie. Ce dernier processus d'identification et d'annotation est complexe et coûteux en terme de ressources, c'est pourquoi ce type d'expression linguistique (*le Président de la République* mais aussi *le Président du Conseil*), bien qu'intéressant du point de vue informationnel, et donc applicatif, n'est que très rarement considéré comme entité nommée. Ces unités n'ont d'utilité que dans des modèles complexes, déterminés en fonction d'applications exigeant une importante précision, et sont donc pour le moment absentes de la plupart des réalisations.

Cette discussion sur le statut d'entité nommée pour des expressions telles que *le Président de la République* conduit à réfléchir à la notion d'entité nommée relativement à ce que l'on appelle des chaînes de coréférence. Ces dernières désignent une série d'expressions linguistiques renvoyant à un même référent ou co-référent. Si un système informatique doté d'un modèle particulier annote le nom propre *Jacques Chirac* puis la description définie *le Président de la République* et relie ensuite les deux expressions, l'une étant le titre ou la fonction de l'autre, alors on est en droit de dire que ces deux expressions coréfèrent dans le modèle. Toutefois, cette proximité des entités nommées avec la notion de coréférence s'arrête là, en raison tout d'abord du fait qu'elle est dépendante d'un modèle donné, et parce que toutes les expressions d'une chaîne de coréférence ne sont pas entités nommées ensuite. Relativement à ce dernier point, on peut en effet remarquer

que, parmi les expressions suivantes, {*Jacques Chirac*, *le Président de la République*, *l'homme politique*, *il*}, seules les premières sont autonomes (au regard d'un modèle et d'un corpus toujours) et peuvent devenir entités nommées. Elles sont par ailleurs annotées indépendamment de leur appartenance à une chaîne de coréférence.

Revenons aux unités de la figure et étudions le cas du *Tour de France*. Dans quelles conditions peut-on prendre cette expression comme une entité nommée ? Il s'agit d'un nom propre, de la désignation d'une entité précise par le biais d'une description dont le sens joue un rôle mineur par rapport à la dénomination, opérant directement, du référent (cf. section 4.2.2.3). En effet, si le Tour de France parcourait certainement le territoire national au début du siècle, le circuit opéré par les cyclistes (qui passent souvent par d'autres pays) est aujourd'hui de moindre importance dans la compréhension du *Tour du France*, davantage perçu comme une événement sportif, quel que soit son parcours. Il s'agit donc d'une expression renvoyant de manière autonome à une entité unique, laquelle devient entité nommée dans le cadre d'un modèle considérant les événements. Cette catégorie rend compte d'entités référentielles facilement modélisables, les événements sportifs ou culturels à caractère périodique (*le Tour de France*, *le Tournoi des Cinq Nations*, *le Festival d'Avignon*) ou ponctuels (*le bi-centenaire de la Révolution*). De fait, l'ensemble de ces expressions linguistiques autonomes et monoréférentielles sont appellées à devenir entités nommées dans le cadre d'un modèle applicatif s'intéressant aux événements-spectacles, c'est-à-dire médiatisés et ayant un public, dans un corpus donné.

Que penser de *sociologie* ? On peut hésiter à dire de cette expression, appellée à être reconnue dans le cadre de la campagne d'évaluation HAREM [Santos *et al.*, 2006], qu'elle fonctionne de manière autonome et qu'elle renvoie à un particulier ou à une entité unique compte tenu du caractère général de la dénomination qu'elle opère. Il est possible toutefois d'imaginer un système informatique destiné à collationner et à gérer un ensemble d'informations sur les offres de formations de différentes universités, dans le modèle duquel *sociologie* pourrait être considérée comme une entité individuelle de type « discipline ». Cela demeure néanmoins une application très particulière, d'où la singularité de cette expression en tant qu'entité nommée. Il importe ici de souligner la différence entre la reconnaissance d'entités nommées et la terminologie : celle-ci s'intéresse aux termes d'une langue en tant qu'ils représentent les traces de concepts d'un domaine donné, celle-là s'intéresse à des expressions linguistiques en tant qu'elles représentent les traces de référents d'un modèle donné. S'il n'y a pas de modèle, il n'y a pas d'entités nommées. À l'instar de *sociologie*, *les socialistes* ne rendent compte d'une entité unique de manière autonome que dans un modèle très précis, d'où leur présence

par intermittence seulement dans les systèmes courants de reconnaissance d'entités nommées. Le *Parti Socialiste* est une entité possible d'un modèle applicatif, renvoyant à une entité précise de manière relativement autonome ; *les socialistes* en revanche est une expression qui exige, dans notre cadre d'analyse, que l'on considère comme une entité un ensemble d'individu. Cela est possible, selon le principe d'individuation (exposé en section 4.2.2.3), mais on se trouve alors aux limites de ce que l'on considère habituellement comme individuel et unique, d'où la rareté (mais non l'impossibilité) de la considération de ce type d'expressions comme entités nommées.

*Lyzozyme* : il s'agit d'un type d'enzyme. Cette expression est monoréférentielle et autonome, ce qui explique sa présence dans la figure. Dans quel cadre peut-elle devenir une entité nommée ? Dans les applications généralistes qui s'occupent de politique, *lyzozyme* est un mot de la langue française, simplement. En revanche, dans des modèles applicatifs médicaux ou de biologie, *lyzozyme* est une entité nommée à part entière, parmi d'autres noms de gènes, de protéines ou de maladie.

Pour ce qui est des expressions temporelles absolues telles que *le 21 Avril 2002* ou *1986*, on comprend aisément leur présence au sein des unités lexicales pouvant être considérées comme des entités nommées. Ces expressions désignent en effet de manière totalement autonome des entités uniques équivalant à des intervalles plus ou moins longs (une heure, une journée, une année, etc.). Les expressions de temps sont ainsi très proches des noms propres dans leur fonctionnement, même si elles *décrivent*, avec plus ou moins de précision, l'entité à laquelle elles se rapportent. D. Van de Velde pose d'ailleurs la question *Existe-il des noms propres de temps ?* et répond par l'affirmative, l'idée étant que « *si la base subjective inévitable de tout discours est un système ternaire* (cf. les trois repères fondamentaux de la référence du *je*, *ici* et *maintenant*), *le corrélat objectif de cette base doit être ternaire lui aussi, et par conséquent il doit y avoir des noms propres de temps* » [de Velde, 2000]. On observe effectivement qu'aux côtés des personnes et des lieux, la plupart des applications dites généralistes intègrent dans leur modèle des entités temporelles de type « date » ou « période ». Le temps est très certainement une notion complexe, mais il est pour partie, et pour partie seulement, « facilement » modélisable, d'où la présence presque assurée d'expressions temporelles absolues dans les systèmes de reconnaissance d'entités nommées. Nous avons vu que la conférence MUC-7 avait introduit l'annotation d'expressions temporelles relatives, telle que *hier* ou *l'année prochaine*. Ces expressions sont monoréférentielles mais peu autonomes : elles peuvent devenir entité nommée, mais à l'intérieur d'un modèle qui permette de calculer ce à quoi elles renvoient exactement, par rapport à la date d'écriture du document ou par rapport au temps mentionné dans la narration. Ce processus est déjà plus

complexe et nécessite une modélisation ainsi qu'un traitement précis des expressions temporelles, c'est pourquoi seulement quelques unes, parmi ces dernières, deviennent « à coup sûr » des entités nommées.

Le problème vaut également pour les expressions numériques qui, si elles ne se rapportent à rien, ne sont d'aucune utilité d'un point de vue informatif. En effet, si l'on annote dans un texte une expression telle que *20%*, sans savoir s'il s'agit d'une réduction ou d'une augmentation d'un stock de marchandises, de la superficie d'une forêt ou encore d'une cote de popularité, cela est trompeur. Les expressions numériques peuvent facilement être autonomes et renvoyer à une entité ou quantité précise mais, telles qu'elles sont annotées pour le moment, elles ne remplissent aucune de ces conditions. Pour devenir entités nommées, ce type d'expressions doit être traité dans le cadre d'un modèle très précis, spécifiant, pour chaque unité de valeur, ce à quoi elle se rapporte spécifiquement. Les expressions monétaires relèvent quant à elles d'un domaine facilement délimitable et modélisable (l'existence des différentes devises est relativement stable), il est donc possible d'en faire des entités nommées.

Au final, il apparaît bien que la notion d'entité nommée recouvre une réalité composite, caractérisable linguistiquement par son rapport à la référence, autonome et dédié à une entité considérée comme unique, et délimitable référentiellement par un modèle spécifiant la portion du monde à prendre en compte relativement à une application donnée. Si, d'un point de vue linguistique, ce sont généralement les mêmes types d'unités lexicales qui sont candidats au statut d'entité nommée, le monde des possibles est largement ouvert d'un point de vue référentiel, ce qui explique la diversité des réalisations. En effet, chacune des expressions examinées ci-avant, possédant les propriétés linguistiques d'autonomie et de monoréférentialité, peut ou non, suivant le modèle pensé, devenir une entité nommée. En définitive, les entités nommées viennent de l'applicatif et non du linguistique (ce qui ne les empêche pas d'être caractérisables linguistiquement, ce sont des unités de la langue), cette filiation expliquant par ailleurs l'existence d'entités nommées « génériques » relevant de domaines ou sous-domaines facilement modélisables et donc abondamment repris[1]. Les entités nommées sont une catégorie inventée par le TAL, une catégorie flexible pour répondre à des besoins applicatifs. À la suite de cette proposition de définition des entités nommées, il convient de considérer ce que devient la tâche de reconnaissance des entités nommées.

[1]On trouve ici une des raisons du succès rapide de cette tâche (cf. chapitre 1).

## 5.3 Conséquences de cette définition

Au regard de la tâche de reconnaissance des entités nommées, les prolongements de cette définition sont de deux ordres, avec des conséquences méthodologiques sur la mise en œuvre de la tâche, et des conséquences pratiques sur le problème des polysémies.

### 5.3.1 Méthodologie d'application

D'un point de vue méthodologique, les conséquences de la définition sont relativement claires, apportant quelques éléments de réponse aux difficultés pointées lors du chapitre 2.

Un premier point essentiel est que, pour travailler sur les entités nommées, il faut au préalable se mettre d'accord sur un modèle. Cet impératif se trouve justifié par la définition des entités nommées d'une part (ce sont des entités référentielles et il faut bien décider quelle partie de la référence on cherche à traiter), et par les difficultés qu'il permet de contourner d'autre part. Le modèle circonscrit une portion du monde et détermine les entités à prendre en compte en fonction de l'application. La notion de modèle est déjà présente dans de nombreuses réalisations, mais implicitement, et sans être le résultat d'une véritable étape de concertation sur ce qu'il importe d'annoter. Sans remettre en cause les précédents travaux, ce que nous cherchons à mettre en valeur ici, d'un point de vue méthodologique, est le fait qu'il est primordial, dans la préparation d'une tâche de reconnaissance d'entités nommées, d'*expliciter* le modèle. D'expérience, il est très difficile de trouver un consensus entre concepteurs de systèmes de traitement d'entités nommées à propos de ce qu'il importe d'annoter ; ce constat est d'ailleurs partagé par S. Sekine, qui ne manque pas de remarquer que « *Even if we had a small number of categories, there were long discussions, arguments, and a bit artificial solution to define the categories* » [Sekine, 2004]. Il semble qu'il existe une tension entre, d'une part, des gens cherchant à définir les entités nommées en général en vue de concevoir des systèmes facilement transposables d'une application à une autre et, d'autre part, des gens qui cherchent au contraire à concevoir les choses de manière très ciblée, en fonction de besoins applicatifs précis. Cette tension est normale, personne n'a raison et en même temps tout le monde a raison : il n'y a pas de bonne entité nommée en soi, il n'y a pas de bonne ou mauvaise réponse à la question « qu'est-ce qu'une entité nommée », il n'y a que des critères linguistiques et un modèle. Il est donc nécessaire, voire impératif, de préciser clairement ce modèle, qu'il soit simple ou complexe, général ou très spécialisé.

À quoi correspond cette étape d'explicitation du modèle ? Cette dernière

consiste en la détermination des catégories sémantiques des entités à reconnaître et à annoter, détermination devant fixer les choses au niveau du choix des catégories elles-mêmes, ainsi qu'au niveau de ce qu'elles recouvrent. Concernant le premier point sur le choix des catégories, la prise en compte des besoins applicatifs pose le fait qu'aucune catégorisation n'est ontologiquement supérieure aux autres et qu'il s'agit juste d'une question d'efficacité : ai-je ou non besoin de reconnaître tel ou tel type d'entité. Il est évident qu'une application financière exigera la reconnaissance d'entités différentes d'une application médicale, sans pour autant qu'une expression telle que *Livret A* soit par nature plus ou moins évidente ou intéressante que *H5N1*. Les catégories dépendent exclusivement de ce que l'on cherche à traiter, d'où l'importance d'un questionnement à ce sujet. Ce questionnement peut néanmoins être difficile à mettre en œuvre, dans la mesure où la reconnaissance d'entités nommées à l'heure actuelle se généralise fortement, cherchant à traiter tous les textes selon une perspective générale d'extraction d'information. La question se pose alors : la détermination préalable d'un modèle pour la reconnaissance d'entités nommées est-elle contradictoire avec le fait d'avoir besoin d'outils génériques ? Non, il importe seulement de savoir à quoi le système de reconnaissance d'entités nommées va être utile. Il peut s'agir d'un outil générique pour lequel on conçoit un modèle « pauvre » composé de quelques classes seulement. Le grand avantage de ce genre d'outils génériques basés sur un modèle simple est qu'ils peuvent servir pour de nombreuses applications ; il est en effet assez « systématiquement » utile de reconnaître des entités de type « personne ». Le revers est le fait que, pour pouvoir être utilisés sur des applications plus précises, cherchant à distinguer les hommes politiques des hommes d'affaires par exemple, ces systèmes génériques doivent faire l'objet de nouveaux développements et doivent être enrichis, par l'ajout de nouveaux lexiques par exemple. Une autre option peut être de chercher à concevoir des systèmes tout autant génériques mais plus riches sur certains points, faisant alors état d'une autre forme de généricité réclamant par endroits beaucoup de détails donnés de manière explicite. Un système conçu à partir d'un modèle de ce type peut alors être utilisé soit pour une application spécifique, soit pour une application générique n'exploitant pas toutes les ressources du système. Il peut alors être utile, étant donné qu'il paraît impossible de penser à tout au préalable, de disposer d'outils aidant à la découverte de nouvelles catégories, plus ou moins précises (cf. chapitre 6 à venir, présentant une méthode d'aide à la conception de catégories).

La question peut ensuite se poser de ce que recouvrent précisément les catégories. Faut-il annoter *les époux Chirac* comme *Lionel Jospin* ? La décision à prendre, une fois de plus, n'est pas donnée une fois pour toutes, il importe d'expliciter totalement le modèle en faisant des choix à propos de ce que l'on considère comme

individu ou collectif.

Un autre point méthodologique déductible de cette proposition de définition concerne l'annotation des entités nommées. Une entité nommée est une expression linguistique monoréférentielle et autonome. Ce sont ces propriétés qu'il importe de prendre en compte et de spécifier avant toute entreprise d'annotation. Au regard de la monoréférentialité, une expression linguistique peut renvoyer à une entité unique selon des échelles différentes, c'est-à-dire selon que l'on considère des entités ou des classes d'entités (comme cela vient d'être évoqué ci-dessus). Elle peut également être plus ou moins autonome, selon l'importance des connaissances à mobiliser pour son interprétation. Ainsi, s'il convient de se mettre d'accord sur la notion d'individu, il importe par ailleurs de déterminer le niveau d'autonomie souhaité pour les entités nommées. Si la plupart des applications choisissent de reconnaître des expressions linguistiques parfaitement autonomes telles que *Lionel Jospin*, il est aussi possible de faire le choix d'annoter des entités telles que *le premier ministre en 1999* dont l'autonomie ne peut être absolue que relativement à des connaissances de nature encyclopédique et contextuelle. Cette notion d'autonomie joue également au niveau, par exemple, des entités coordonnées : *M. et Mme Chirac* constitue une expression non annotable telle quelle en tant qu'entité nommée, il convient en effet de restituer l'autonomie référentielle du premier composant (*M. Chirac*). Cette restitution d'autonomie, ou complémentation dans le cas de descriptions définies, est cependant dépendante des ressources dont dispose le système informatique.

La proposition de définition constitue ainsi un cadre d'appréhension des entités nommées, spécifiant certains éléments à prendre en compte lors de la conception d'un système de reconnaissance de ces unités. D'un point de vue méthodologique, il peut en effet être utile, voire indispensable pour éviter de longues discussions et faciliter l'interaction entre différents systèmes, d'expliciter le mieux possible un modèle et de faire des choix vis-à-vis de ce que l'on considère comme monoréférentiel et autonome. Ces choix sont bien sûr dépendants des capacités de calcul des entités nommées.

### 5.3.2 Le problème des polysémies

Outre les difficultés de catégorisation et d'annotation, les entités nommées sont parfois difficiles à traiter en raison de phénomènes de pluralité référentielle. Ce point a été souligné lors de la section 2.3, au cours de laquelle nous avons précisé la notion de polysémie lexicale avant de nous interroger sur la caractérisation de la polysémie des entités nommées et son éventuel traitement automatique. Étant donné la proposition de définition des entités nommées présentée ci-avant,

l'objectif ici est, d'une part, de définir précisément les phénomènes de polysémie au regard de ces unités et, d'autre part, de montrer comment le modèle du monde nécessaire à la définition des entités à prendre en compte peut également clarifier les choses au niveau de la polysémie.

#### 5.3.2.1 Entités nommées et polysémie : éléments de définition

On appelera homonymes deux entités nommées différentes dans un modèle donné, ces deux entités étant désignées par deux expressions linguistiques accidentellement identiques. Autrement dit, il s'agit d'une même expression dans le corpus qui réfère de manière autonome et monoréférentielle parfois à une entité du modèle, parfois à une autre, sans qu'il existe de relation entre ces dernières. Pour illustrer ce type de relation entre entités il est possible de penser à une expression comme *Vienne*, pouvant renvoyer à une ville en France ou à une ville en Autriche (pour ne prendre que les plus connues et donc les plus couramment traitées en TAL), sans qu'il existe un lien entre ces deux dénominations. Il s'agit de deux entités différentes dans le monde mais de deux entités relevant de la même catégorie sémantique, manifestant alors une homonymie que l'on pourrait qualifier d' « intracatégorielle ». De la même manière, examinons l'expression *Orange*, pouvant renvoyer ou bien à une ville française (entre autres), ou bien à une société de téléphonie mobile. Ces deux entités ne participent pas du même type, on peut alors parler d'homonymie « intercatégorielle ».

La métonymie correspond en langue au fait d'employer un mot attaché à une certaine entité pour en désigner une autre, la seconde étant liée à la première par un rapport fonctionnel ou structurel (cf. section 2.3.2.2). On distingue plusieurs « types » de métonymies, suivant que leur apparition est exclusivement dépendante du contexte (métonymie contingente) ou au contraire le résultat d'un processus régulier valable pour tout un ensemble d'unités lexicales (métonymie systématique). Au regard des entités nommées, on peut parler de métonymie lorsqu'une une expression linguistique habituellement utilisée pour référer à une certaine entité du modèle est utilisée pour référer à une autre entité du modèle, la seconde étant liée à la première par un rapport fonctionnel ou structurel. Il existe de nombreux cas de métonymie pour les entités nommées, à l'encontre desquels on peut observer divers degrés de « systématicité ».

Un glissement de sens à caractère métonymique semble être à l'origine de la relation existant entre le *Général Leclerc* et le *char Leclerc*, ce dernier ayant bien été nommé en l'honneur du premier. Cette nomination honorifique ne constitue certes ni une relation structurelle ni une relation fonctionnelle mais peut s'interpréter comme une « relation de contiguïté », selon les termes employés par G. Kleiber

(à propos du principe de *métonymie intégrée*), attestant par là même d'un lien entre ces deux entités désignées par un même nom mais renvoyant à des référents différents. Cette opération de baptême entre un officier de l'armée et un engin de guerre semble cependant imprévisible ou pour le moins liée à des circonstances particulières, on se situe donc ici plus du côté de la métonymie contingente que du côté de la métonymie systématique.

Le glissement de sens entre *M. Leclerc* (l'homme d'affaire), la chaîne de magasins du même nom et le groupe financier revêt un caractère un peu plus systématique, sans que cela soit cependant le cas pour tous les fondateurs ou administrateurs d'entreprises (Ford a également donné son nom à son entreprise, mais le fondateur d'Orange n'est pas M. Orange ni celui de Carrefour M. Carrefour). Il existe toutefois bel et bien un lien entre ces entités, dont on peut par conséquent dire qu'elles sont métonymiques. Remarquons au passage que le Général Leclerc et Michel-Edouard Leclerc sont de parfaits homonymes.

Par ailleurs, certains cas de pluralité référentielle des entités nommées peuvent s'analyser en termes de métonymie systématique. Examinons les phrases suivantes :

*La France a signé le traité de Kyoto.*
*Marseille ne s'est pas qualifiée pour la demi-finale.*
*Les leçons du 21 Avril 2002 ont-elles été tirées ?*

Dans la première phrase, le nom propre *France* est utilisé pour renvoyer non pas à l'unité géographique mais au gouvernement de ce pays, à même de signer un traité. De même, la seconde phrase rend compte d'un événement sportif, par le biais du nom propre *Marseille* qui renvoie ici à une équipe sportive et non à la ville proprement dite. Enfin, la date du 21 Avril 2002 est employée comme pointeur vers un événement particulier et non vers une période temporelle du calendrier. Ces glissements métonymiques entre nom de pays et gouvernement, nom de ville ou pays et équipe sportive ou encore mention de date et événement particulier ont lieu régulièrement, c'est pour cela que l'on peut parler de métonymie systématique. D'autres schémas sont envisageables, ils seront décrits plus précisément dans le chapitre 7, l'essentiel ici étant de définir et illustrer la métonymie pour les entités nommées. Ainsi, la plupart des cas de polysémie sont des métonymies pour les entités nommées.

Au-delà de ces phénomènes de pluralité référentielle permettant de renvoyer, à partir d'une même expression linguistique, à plusieurs référents de différentes natures, il importe de mentionner une autre forme de diversité à l'endroit des entités nommées. Il s'agit ici de la conséquence d'une des principales caractéristiques du nom propre, catégorie linguistique abondamment représentée, nous l'avons vu, au

sein de l'ensemble 'entité nommée'. En effet, dans le discours théorique, le nom propre est, à la suite de S. Kripke, qualifié de « désignateur rigide » : attribué par convention à un particulier, il permet de le désigner dans toutes les situations, et ce quels que soient les changements pouvant l'affecter. De la sorte, si le particulier Jacques Chirac est Maire de Paris un jour puis Président de la République le lendemain, son nom demeure *Jacques Chirac* quoi qu'il arrive. Si cette invariabilité de la dénomination par un nom propre est justifiée d'un point de vue cognitif par un principe d'économie (cf. 4.2.2.2), elle peut néanmoins constituer un handicap pour qui cherche à capter les variations référentielles affectant un particulier. En effet, et nous arrivons là au point que nous souhaitons souligner pour les entités nommées, il peut être intéressant de vouloir leur associer une information sémantique fine, permettant de mieux circonscrire leur référent en contexte. Dans une phrase telle que :

*Arnold Schwarzenegger s'est rallié le soutien de la majorité démocrate pour s'assurer un vote favorable*,

il s'agit alors d'indiquer, au-delà de fait que A. Schwarzenegger est une « personne », qu'il est question ici du « gouverneur républicain de Californie » et non de l' « acteur » ou du « bodybuildeur ». Dans le même esprit, il importe de distinguer, dans un texte ou un ensemble de textes, les occurrences de l'entité Jacques Chirac renvoyant au « Président de la République » de celles renvoyant au « maire de Paris ». Cette spécification précise du référent auquel renvoie l'entité est à mettre en relation avec la notion de « facette sémantique » et de signification en contexte (cf. section 2.3.2.2) : de même que le sens d'une unité lexicale est activé différemment suivant son contexte d'apparition, le référent d'une entité nommée revêt ou non certaines caractéristiques.

Ainsi définis, les phénomènes de superposition de sens des entités nommées parcourent presque tout l'échantillon des possibles, à l'exception de la métaphore, s'organisant le long d'un continuum allant d'entités homonymiques à des entités monosémiques. Conformément à la définition proposée auparavant, ces phénomènes ont été caractérisés relativement à un modèle ; néanmoins, ce dernier était en quelque sorte « sous-spécifié », n'étant considéré que dans l'absolu, indépendamment de toute application. C'est pourquoi il importe à présent d'examiner comment s'organisent concrètement les cas de polysémie des entités nommées, lesquels n'existent (ou n'existent pas) que relativement à un modèle.

#### 5.3.2.2 Entités nommées et polysémie : éléments de modélisation

Comment les phénomènes ci-dessus définis se distribuent-ils suivant le modèle envisagé ?

Commençons par les cas d'homonymie (*a priori*). L'homonymie entre deux entités nommées a été définie en théorie comme le fait qu'une expression linguistique renvoie à deux entités différentes du modèle sans aucun lien entre elles. Imaginons un modèle comprenant la classe « ville » d'une part, et la classe « pays » d'autre part. Il serait possible de créer une occurrence de la classe « ville » pour une expression linguistique rencontrée, disons *Vienne1* et de la relier à l'occurrence *Autriche* de la classe « pays ». De même, on pourrait créer une seconde occurrence de la classe « ville » pour une autre expression linguistique rencontrée, disons *Vienne2*, et de la relier à l'occurrence *France* de la classe « pays ». Dans cette situation, il existe bien deux entités nommées différentes, renvoyant pour l'une à la ville de Vienne en Autriche et, pour l'autre, à la ville de Vienne en France, ces deux entités n'ayant aucune relation entre elles, sinon le partage d'une expression linguistique de même graphie. Ce type de modélisation permet donc de faire état de l'homonymie entre ces deux entités ; la contre-partie néanmoins est la suivante : le système informatique travaillant à partir d'un tel modèle doit être en mesure de faire la différence entre ces deux entités, autrement dit doit mettre en œuvre un processus de désambiguïsation.

Autre possibilité, faire un modèle généraliste avec des classes habituelles telles que « personne »,« organisation » et « lieu » par exemple. Dans ce cas, l'expression *Vienne*, quelle qu'elle soit, devient une entité dans le modèle par la création d'une instance de la classe « lieu ». Aucune différenciation n'est possible, il n'existe qu'une seule ville de Vienne dans le modèle, cette expression devenant alors monosémique. Cette configuration n'impose aucunement au système informatique de savoir désambiguïser et c'est celle qui est la plus courante malgré le handicap que cela peut constituer pour certaines applications.

Restons toujours dans les cas d'homonymie et examinons l'expression *Orange*, pouvant pour sa part renvoyer soit à une ville, soit à une entreprise. Si le modèle d'annotation prévoit une classe « ville » d'une part et une classe « organisation » (ou entreprise) d'autre part, alors deux types différents d'entités peuvent être instancié, par conséquent Orange-ville et Orange-entreprise sont homonymes. Dans le cas d'une application s'intéressant exclusivement à la géographie française, il est fort à parier qu'aucune classe « organisation » ne soit prévue. L'entité *Orange* devient en ce cas monosémique.

Avec l'étude de ces deux cas possibles d'homonymie est ainsi mis en évidence le fait que les phénomènes de sens concernant les entités nommées n'existent

que relativement à un modèle. Ce modèle sélectionne une partie du monde à traiter, ici les personnes, là les lieux, contraignant par là-même les possibilités de sens d'une entité. Ainsi, si nous savons, en tant que locuteurs français, que telle expression peut référer à tel particulier ou à tel autre, il n'en est peut-être rien pour le système informatique qui, pour sa part, ne connaît que le modèle. Si toutes les possibilités de sens ne sont pas représentées dans le modèle pour une entité nommée donnée, alors cette dernière passe d'homonymique dans la réalité à monosémique en TAL.

Poursuivons avec les cas de polysémie. Les expressions *Leclerc*, *France* et *Chirac* ont souvent servi d'illustration, nous les reprenons donc ici. Deux *Leclerc* ont été distingués jusqu'à maintenant : le Général, qui a donné son nom à un char, et l'homme d'affaire, qui a donné son nom à son entreprise, qui elle-même a donné son nom à ses établissements. Pour l'un comme pour l'autre, la pluralité de sens participe de la métonymie, la relation existant entre deux entités donnant la possibilité d'appeler l'une par le nom de l'autre. Dans un cas concret d'annotation d'entité nommée, comment annoter l'expression *Leclerc* ? Traditionnellement, il est possible de penser à deux classes, « personne » d'un côté et « organisation » de l'autre. Le Général comme l'homme d'affaire vont instancier la classe « personne » ; si aucun moyen ne permet de les différencier, l'entité *Leclerc* est dans ce cas monosémique. Si d'aventure on rajoute au modèle la classe « arme », il est alors possible de rendre compte de la relation de métonymie entre le Général et le blindé, permettant par ailleurs de différencier les deux précédents *Leclercs* alors homonymes dans ce cas. Ce cas de figure semble toutefois difficile à traiter : soit la relation de métonymie entre le Général et le char n'est pas établie, et dans ce cas il y a instanciation de la classe « arme » par une entité homonymique (le char) d'une part, et instanciation de la classe « personne » par une entité homonymique elle aussi (renvoyant indifféremment aux deux hommes), soit le char n'est même pas annoté comme tel, et dans ce cas une seule entité monosémique est présente. Le cas de figure le plus prévisible (et le plus utile) est cependant la reconnaissance d'une occurrence de *Leclerc* en tant que supermarché, ne permettant certes pas de faire la différence entre la personne et le char, mais de rendre compte de celle entre la personne et l'organisation.

Avec la *France*, on se situe du côté de la métonymie systématique. Cette expression est utilisée pour renvoyer soit à l'entité géographique, soit à l'entité politico-administrative. De nouveau, cette expression est ou non polysémique suivant le modèle envisagé. Le plus classique fait état d'une classe « lieu » ou « pays » d'une part, et d'une classe « organisation » d'autre part. Si aucune relation n'est prévue entre ces deux classes, alors l'expression *France* est soit un pays, soit un gouvernement : elle est homonymique et le système est obligé de désambiguïser. Si au

contraire il existe une relation entre ces deux classes, alors l'expression *France* a soit un sens littéral, soit un sens métonymique : cette configuration paraît être la plus proche de la réalité, elle n'est cependant pas facile à mettre en œuvre et réclame une désambiguïsation (une méthode de résolution de ce problème est proposée dans le chapitre 7). Enfin, dernière solution : avoir une sorte de classe métonymique « figée », de type GPE (classe « geo-political-entity » de la compétition ACE) ou encore « Locorg » (lieu et organisation), permettant de rendre compte des deux sens à la fois. *France* est dans ce cas monosémique, différenciée des entités géographiques proprement dites ne pouvant être des entités politiques d'une part, et des organisations proprement dites ne pouvant être des entités géographiques d'autre part. Cette solution est intermédiaire, à mi-chemin entre la non prise en compte du phénomène et son traitement complet.

Enfin, que devient *Jacques Chirac* ? Il est possible de vouloir spécifier au plus juste ce à quoi renvoie cette expression, et donc de rendre compte de ce que nous avons appellé des facettes référentielles. Jacques Chirac est en effet une personne ayant deux facettes notoires (entre autres), celle de « Maire de Paris » et celle de « Président de la République ». Avec un système et un modèle classique, l'expression *Jacques Chirac* peut instancier, sans plus d'informations, une classe « personne ». Elle est alors pleinement monosémique. Un modèle peut par ailleurs prévoir divers attributs pour une entité d'une classe donnée, dont celui de « Président de la République » par exemple. De la sorte, l'expression se trouve quelque peu précisée, et peut devenir polysémique si une autre occurrence ne se voit pas attribuer le même attribut. Il est encore possible d'imaginer un modèle consacré à l'étude de la classe politique française, faisant état d'une classe « Président de la République » d'une part et d'une classe « Maire » d'autre part. Dans ce modèle, l'expression considérée devient alors homonymique (si aucune relation n'existe entre les deux classes).

L'examen de ces quelques cas de polysémie, du plus contingent au plus systématique, montre de nouveau que les phénomènes de sens à l'endroit des entités nommées sont tributaires de la modélisation choisie pour une application donnée. En fonction du modèle, deux entités homonymiques peuvent devenir une entité monosémique, ou encore deux entités peuvent voir leur relation passer de métonymique à homonymique. La polysémie des entités nommées existe donc, comme les entités nommées elles-mêmes, relativement à un modèle. Cela confirme l'importance de l'explicitation du modèle dans toute tâche de reconnaissance d'entités nommées, lequel, de nouveau, est fonction de l'application et des possiblités de calcul de la polysémie des entités nommées.

# Bilan

À l'origine de cette deuxième partie étaient, rappelons-le, le constat de nombreuses difficultés pouvant apparaître durant le traitement des entités nommées ainsi que la perspective de nouveaux besoins s'agissant de leur annotation. Entrant en résonance avec des problèmes de définition, ces deux points ont alors révélé le besoin d'un équipage théorique un peu plus développé qu'auparavant au regard des entités nommées. Ces unités, intuitivement appréhendées de façon sémantico-pragmatique la plupart du temps, ne semblaient en effet ne bénéficier d'aucune véritable assise théorique dans la littérature, n'ayant d'autres propriétés saillantes que de constituer un ensemble hétérogène et d'être l'objet de réalisations diverses, fort différentes les unes des autres. Prenant acte de leur appartenance au TAL, cette partie a tenté de rendre compte de la double filiation des entités nommées, regardant du côté de la linguistique d'une part, du côté du traitement automatique des langues d'autre part, et ce afin de proposer une définition de ces unités.

Le « détour » du côté de la linguistique a tout d'abord permis de préciser le cadre de réflexion en définissant les notions de sens et de référence. Deux investigations ont par la suite été menées, s'intéressant à la catégorie du nom propre pour l'une, à celle des descriptions définies pour l'autre. Catégorie linguistique pour le moins singulière et difficile à définir, le nom propre resta longtemps l'apanage des logiciens, ces derniers posant les fondements théoriques du nom propre comme vide de sens d'une part, et comme description de son référent d'autre part. Cette alternative constitua le point de départ de la réflexion des linguistes qui proposèrent néanmoins de la dépasser avec une analyse du nom propre non plus dénué de sens mais doté d'un sens dénominatif de type instructionnel augmenté d'un sens descriptif de référence à un particulier (G. Kleiber) et d'une « charge référentielle », analysée en termes de contenu (M.N. Gary-Prieur) ou de sens encyclopédique (M. Charolles). Ces analyses, en même temps qu'elles contribuaient à établir pleinement le nom propre en tant que catégorie linguistique, mirent également en valeur les propriétés référentielles des noms propres, ceux-ci désignant de manière autonome un référent unique ou perçu comme tel. L'étude des descriptions définies débuta pour sa part par un exposé des travaux fondateurs du philosophe et logicien B. Russell, lesquels donnèrent lieu à une polémique célèbre avec P. F. Strawson à propos de la référence singulière. Si le premier considère en effet qu'une description définie comporte une indication d'existence et d'unicité du référent, le second préfère quant à lui parler de *présupposition* d'existence et d'unicité, selon un point de vue davantage centré sur l'usage des mots que sur les valeurs de vérité des propositions. Ces réflexions, selon un cheminement similaire

à celui constaté pour les noms propres, influencèrent les travaux des linguistes qui travaillèrent à une définition plus proprement linguistique de ce type d'unités lexicales (G. Kleiber) ainsi qu'à une détermination de leur comportement référentiel (M. Charolles). En définitive, qu'elles soient d'inspiration logique ou linguistique, ces réflexions nous ont permis de mettre en valeur le fait que les descriptions définies opèrent une référence à une entité unique, selon divers degrés d'autonomie. Ce cheminement du côté de la linguistique nous a ainsi donné l'occasion de véritablement établir les propriétés des unités composant l'ensemble 'entités nommées' : monoréférentialité et autonomie, telles un dénominateur commun pouvant servir de socle linguistique à une caractérisation de ces unités.

À la suite de cette prise en compte de la composante linguistique, s'imposait toutefois la nécessité de replacer l'étude des entités nommées dans le cadre du traitement automatique des langues. Il fut d'abord question de redéfinir le cadrage théorique initial de sens et de référence en examinant la valeur de ces notions au sein du TAL : indispensables aux traitements relevant de la sémantique, ces notions sont bien présentes en TAL mais leur définition fait intervenir la notion de modèle. C'est en considérant ce « cadre d'action » propre au TAL qu'il fut alors possible de proposer une définition des entités nommées prenant en compte leurs caractéristiques linguistiques d'autonomie et de monoréférentialité tout comme les exigences applicatives du traitement automatique du langage. Les prolongements de cette proposition de définition peuvent s'apprécier selon divers points de vue. Au regard de la notion d'entité nommée tout d'abord, cette définition propose un cadre d'appréhension de ces unités exprimé et justifié en des termes plus précis, nous l'espérons, qu'auparavant. Elle permet par ailleurs de rendre compte de l'hétérogénéité constatée tant dans les discours que dans les réalisations concernant ces unités pourvues d'une base linguistique stable mais nécessairement solidaires d'un modèle du monde soumis à variation. D'un point de vue méthodologique, cette proposition de définition met en valeur la nécessité d'une détermination préalable d'un modèle des entités à prendre en compte, cette précaution pouvant prévenir de longues discussions, simplifier la réalisation d'un système de reconnaissance ainsi que l'interaction entre diverses applications. Enfin, les problèmes de « polysémies » peuvent également être pris en compte dans le cadre d'un modèle, celui-ci prévoyant de telle ou telle manière, suivant ses objectifs et ses capacités, les possibilités de phénomènes de sens pour les entités nommées. Rassemblant diverses unités lexicales sur la base de propriétés linguistiques outrepassant les catégories traditionnelles, les entités nommées ont été « imaginées » pour répondre à des besoins applicatifs ; c'est à ce titre que l'on peut à leur égard parler de « création TAL ».

Au terme de cette deuxième partie à dominante théorique ayant pour visée

une réflexion sur la notion d'entité nommée, il est temps de passer à la deuxième perspective de recherche proposée à la fin de la partie I (cf. section 2.4) et de considérer les entités nommées sous un angle plus pratique et expérimental.

# Troisième partie

# Entités nommées : nouveaux traitements

# Introduction

La présentation de la problématique des entités nommées (cf. partie I) avait, au-delà de la narration de la bonne fortune de la tâche de reconnaissance de ces unités, pointé du doigt deux pierres d'achoppement avec, d'une part, le besoin d'une réflexion sur le statut théorique de ces unités et, d'autre part, la nécessité de mettre en œuvre de nouveaux traitements. La partie précédente (cf. partie II) ayant permis d'apporter quelques éléments de réponse concernant le premier point, l'objet de la présente est donc de s'intéresser au second touchant aux nouveaux objectifs et méthodes de traitement automatique des entités nommées et de tenter d'apporter quelques éléments de résolution.

Du point du vue des réalisations, ou plus exactement des systèmes et des performances, nous avons vu précédemment que la tâche de reconnaissance des entités nommées bénéficie à l'heure actuelle d'une certaine maturité. Les recherches menées jusqu'à ce jour ont en effet permis l'élaboration de systèmes de reconnaissance de ces unités relativement performants, permettant d'identifier dans des textes de nature journalistique des noms relevant de types généraux tels que « personne », « lieu » et « organisation », et ce avec un taux de F-mesure[1] dépassant généralement les 0.90. Dans la droite ligne de ces réalisations, de nouvelles perspectives de recherche se font jour ; il est en effet une volonté d'améliorer et d'enrichir l'annotation des entités nommées avec, entre autres, l'annotation de nouveaux types d'entités (tels les noms de produits, [Nilsson et Malmgren, 2005]), une annotation plus fine allant au-delà des catégories générales désormais aisément reconnues, et une annotation permettant de prendre en compte certains des phénomènes de polysémie observés à l'endroit des entités nommées.

Les travaux présentés ici s'intéresseront plus particulièrement à deux types de traitements pouvant améliorer l'annotation des entités nommées, avec une méthode d'annotation permettant d'effectuer une « double annotation » (annotation générale et annotation fine) d'une part, et une méthode de résolution de métonymie d'autre part. Les chapitres 6 et 7 présenteront successivement ces expériences.

[1] Moyenne harmonique de la précision et du rappel : (2× précision × rappel)/(précision + rappel).

Une chose à signaler avant tout : cette partie, si elle s'inscrit dans la continuité de la précédente, n'en constitue cependant pas une illustration ou une mise en application stricte. Elle s'attache plus modestement à rendre compte de deux expériences pouvant possiblement faciliter la détermination du modèle précédemment introduit.

# Chapitre 6

# Vers une annotation fine des entités nommées

Cherchant à aller au-delà des catégories générales d'annotation des entités nommées et à améliorer leur traitement, nous présentons une méthode ayant pour objet l'aide à la conception de catégories ainsi que la révélation de phénomènes de sens, ceci pouvant au final permettre une annotation fine d'entités nommées ainsi que la désambiguïsation de certaines de ces unités. L'approche proposée repose sur, d'une part, la construction dynamique à partir de corpus d'une ressource lexico-sémantique dédiée aux entités nommées, proposant pour ces dernières des étiquettes sémantiques fines et, d'autre part, la combinaison de cette ressource (permettant une annotation fine) avec un système standard de reconnaissance d'entité nommée (proposant des étiquettes générales classiques) afin d'obtenir une annotation enrichie, ou double annotation, des entités nommées. L'originalité de cette approche réside dans l'exploitation de relations syntaxiques profondes lors de la construction de la ressource d'entités nommées, dans l'annotation à l'aide de groupes d'étiquettes plutôt qu'à l'aide d'étiquettes et dans la mise en œuvre d'une double annotation des entités nommées offrant des informations de niveau tant général que particulier. Cette expérience a été guidée par la volonté de prendre en compte l'extrême mouvance ou instabilité référentielle des entités nommées ainsi que leur caractère polysémique. Ce travail a été réalisé avec Guillaume Jacquet et ce chapitre reprend, pour une majeure partie, un article publié avec lui [Ehrmann et Jacquet, 2006][1].

Il importera, dans la première section, de rendre compte plus précisément des motivations de cette expérience et de donner un aperçu général de l'approche. Les travaux connexes seront évoqués lors de la deuxième section, avant une description

[1]Les chapitres 6 et 7 présentent tous deux des travaux réalisés en équipe ; les contributions respectives de chacun des collaborateurs furent égales et l'ordre alphabétique fut par conséquent choisi pour chacune des publications réalisées à ces occasions.

plus précise, dans la troisième, de la méthode de construction de la ressource et de son exploitation combinée permettant une double annotation. Enfin, la dernière section présentera une évaluation de la ressource et de la méthode.

## 6.1 Des catégories fines pour une annotation enrichie

### 6.1.1 Motivation

Pourquoi des catégories fines ? Trois types d'arguments peuvent venir justifier et expliquer cet objectif, selon des perspectives relevant tantôt de la linguistique, tantôt du TAL, bien souvent des deux.

D'un point de vue linguistique tout d'abord : l'étude de la catégorie la plus fortement représentée au sein des entités nommées, celle des noms propres, a montré que, d'une part, ces derniers sont « réservés » à certaines catégories d'existants (et ce en fonction de codes socio-culturels pouvant varier selon les pays et les époques) et que, d'autre part, s'ils ne se voient attribuer qu'un sens dénominatif, ils finissent par se charger de « sens encyclopédique » au fur et à mesure de leur usage. Ces caractéristiques linguistiques permettent d'expliquer les deux types de catégorisations possibles à l'égard des entités nommées et dont il est ici question : la catégorisation « traditionnelle » au moyen de catégories sémantiques générales s'intéressant, entre autres catégories d'existants courantes, à la personne, au lieu et au temps, et la catégorisation fine, cherchant à capter la charge référentielle attachée à tel ou tel nom propre désignant rigidement telle ou telle entité. Cette charge référentielle correspond plus exactement à ce que nous avons appelé, à l'instar des facettes sémantiques de D.A. Cruse, les facettes référentielles des entités nommées. Prenons quelques exemples, pour certains déjà mentionnés auparavant : pour l'entité *Marlon Brando*, une annotation traditionnelle consisterait à utiliser la catégorie « personne », à laquelle il pourrait néanmoins être utile et informatif d'ajouter la catégorie fine « acteur ». De même, l'entité *Paris* pourrait se voir attribuer, au delà des catégories générales « lieu » ou « ville », celle de « capitale ». Pareillement, il pourrait être avantageux à l'égard de l'entité *Arnold Schwarzenegger* de préciser, au-delà de l'étiquette « personne », s'il s'agit du « gouverneur », du « bodybuildeur » ou de l' « acteur » [1]. Il serait donc possible, à l'aide de catégories fines, de proposer une annotation des entités nommées plus précise et plus proche de leur réalité linguistique et référentielle, une annotation

[1] La personne d'Arnold Schwarzenegger possède bien sûr toutes ces facettes simultanément, ces dernières correspondent au contenu du nom propre *Arnold Schwarzenegger* ; on peut néanmoins, dans un contexte particulier, ressentir le besoin de préciser s'il est question de telle ou telle facette, ce qui se rapproche d'une opération de désambiguïsation.

consistant à leur associer une information sémantique plus fine qu'auparavant et permettant de mieux circonscrire leur référent.

D'un point de vue pratique ensuite : cette annotation fine s'avère quelque peu différente de la catégorisation réalisée jusqu'à maintenant. S'il est en effet relativement facile de penser une catégorisation sémantique large préalablement à l'annotation des entités nommées, il semble plus périlleux de le faire pour des informations plus précises : les étiquettes mentionnées ci-avant, « gouverneur, bodybuildeur, acteur, etc. » n'ont de toute évidence rien de prédictible. On touche ici au problème de la conception des catégories, et donc à celui de la détermination du modèle dont nous avons expliqué la nécessité dans le chapitre 5. Pour décider des entités à annoter pour une application donnée, il faut en effet se poser préalablement la question du modèle du monde à prendre en compte. L'application, relevant d'un domaine donné, est le moyen principal de décider si telle ou telle entité fait ou non partie du modèle et quelles relations elle entretient potentiellement avec les autres entités. Suivant que l'on dispose ou non d'une théorie du domaine, il peut être intéressant, d'un point de vue méthodologique, de disposer d'un outil à même de proposer des catégories possibles mais non évidentes pour les entités nommées. Imaginons une application s'intéressant à la vie politique française : il est *relativement* facile d'anticiper le besoin de catégories sémantiques (plus ou moins fines) telles que « ministre », « secrétaire d'Etat »,« sénateur », « parlementaire » ou encore « conseiller d'Etat ». Prenons maintenant une application ayant pour objet la gestion du risque militaire : il n'est pas forcément aisé d'avoir des connaissances sur l'ensemble des éléments potentiellement impliqués dans ce domaine et donc d'être en mesure d'élaborer un modèle de reconnaissance d'entités nommées pour cette application, d'où la nécessité d'un outil d'aide à la conception de catégories. Autre exemple : la génétique. Même avec une bonne théorie de ce domaine, il peut être intéressant de se voir suggérer de nouvelles catégories, ou encore des catégories en parfaite adéquation avec l'application précise à traiter. Ainsi, si des catégories fines peuvent venir enrichir l'annotation avec des indications plus précises sur les référents des entités, elles sont également les bienvenues pour l'élaboration de modèles d'entités nommées.

Enfin, il est un troisième point pouvant justifier le recours à des catégories fines : la polysémie des entités nommées. En effet, nous avons vu qu'il est possible de multiplier à l'envi les exemples d'homonymie ou de métonymie des entités nommées, avec des entités comme *Orange* (ville ou entreprise), Vienne (ville en France ou en Autriche) ou encore Leclerc (Général, homme d'affaires, char, supermarché, groupe financier). Ce point est bien sûr très proche de la spécification du référent et de la détermination du modèle dont il a été question ci-avant et ces problématiques sont à penser comme complémentaires : pour prendre en compte

la possible polysémie d'une entité lors de l'élaboration d'un modèle tout comme pour concrètement être en mesure de désambiguïser une entité, il est indispensable de savoir qu'elle est ambiguë, et donc d'avoir des informations précises sur son référent et les diverses facettes qu'il peut revêtir. Nous posons donc le problème du traitement des entités nommées en ces termes : il importe de caractériser plus précisément le référent d'une entité nommée en contexte afin de, d'une part, apporter davantage d'information et, d'autre part, être en mesure de mettre en œuvre un processus de désambiguïsation si l'information précise mise à jour révèle une ambiguïté. Notons que certains cas d'ambiguïté peuvent s'avérer difficiles à résoudre si le contexte ne donne aucun indice, comme cela est le cas pour *Orange* dans un énoncé tel que *Orange a invité M. Dupont* (présenté en section 2.3.1) ; le modèle de traitement fin des entités nommées doit pouvoir rendre compte de ces cas d'ambiguïté. Il importe à cet endroit de pointer une spécificité de l'ambiguïté des entités nommées : cette dernière est en effet « évolutive » dans le temps dans la mesure où une entité monoréférentielle aujourd'hui peut devenir polyréférentielle demain. Pour reprendre un exemple désormais bien connu, Orange est une société de téléphonie mobile depuis 1994 seulement, et les marques de produits, sociétés et autres célébrités ne manquent pas d'apparaître, de disparaître et de changer chaque jour. Aussi, l'espace des possibles au regard des sens, ou plutôt des référents, attachés à une entité semble relativement ouvert et peu contrôlable (ensemble fort productif du point du vue référentiel, mais composé pour la plupart d'unités lexicales sans règles de production du point de vue linguistique).

Meilleure caractérisation du référent, aide à la conception de catégories, mise en valeur ou révélation de phénomène de sens, ces points sont autant de motivations à l'élaboration d'une méthode permettant la collecte de catégories fines et, à terme, une annotation enrichie des entités nommées. Un aperçu général de l'approche est présenté ci-après, davantage de précisions seront données dans les parties suivantes.

### 6.1.2 Aperçu général de l'approche

La méthode proposée repose sur la construction à partir de corpus d'une ressource lexico-sémantique dédiée aux entités nommées et, à titre d'expérimentation et de validation, sur la combinaison de cette dernière avec un système classique de reconnaissance d'entités nommées. L'objectif est, d'une part, de pouvoir suggérer de nouvelles catégories d'entités nommées lors de la détermination d'un modèle d'annotation (construction de la ressource) et, d'autre part, d'être en mesure, dans la pratique, d'associer une information sémantique fine aux entités nommées tout en conservant le fruit des travaux réalisés jusqu'à aujourd'hui (combinaison avec

un système classique) afin d'offrir une double annotation des entités nommées et de pouvoir désambiguïser certaines de ces unités.

La première étape correspond à la construction de la ressource d'entités nommées. Cette ressource permet d'associer, pour chaque entité, une ou plusieurs étiquette(s) sémantique(s) fine(s) rendant compte de certaines caractéristiques du ou des référents possibles de l'entité. En d'autres termes, elle permet, idéalement, d'associer à l'entité *Leclerc* les étiquettes (ou catégories fines) « société, homme d'affaire, char, général ». Par son mode de construction, cette ressource pallie certains écueils de la recherche actuelle sur les entités nommées tout en prenant en compte les particularités de ces unités. En effet, constituée automatiquement à partir de corpus, cette ressource peut tout d'abord être construite à moindre frais (or chacun connaît le coût de construction de lexiques spécialisés ou le coût de l'annotation manuelle d'un corpus d'apprentissage). Ensuite, son élaboration est réalisée de manière non supervisée : aucun type d'entité n'est visé particulièrement (or les lexiques concernant les entités nommées ont été jusqu'à maintenant réservés à certains types, comme les lieux) et aucune catégorie n'est définie préalablement, ce qui confère un caractère « inventif » à la ressource. Il a en effet été souligné ci-dessus que le type d'information recherché aujourd'hui correspond à des étiquettes difficilement prévisibles et qu'il peut être malaisé, sans théorie ou connaissance du domaine de l'application visée, d'expliciter un modèle d'annotation. Enfin, la ressource permet de « suivre l'actualité » des entités nommées, ces dernières pouvant être caractérisées de différentes manières en fonction de la période et du domaine du corpus. La construction de la ressource se fait pour chaque nouveau corpus, elle est ainsi exclusivement dépendante du corpus et c'est ce qui lui confère une caractéristique dynamique. Un processus incrémental de construction de la ressource serait bien sûr envisageable mais serait alors perdue l'idée d'adaptation en fonction du corpus, caractère important à nos yeux car offrant la possibilité de rendre compte avec exactitude des référents des entités telles qu'elles apparaissent dans un corpus[1].

Ainsi, cette ressource fournit une représentation fidèle des entités nommées présentes dans un corpus et a pour principal intérêt, au travers d'étiquettes sémantiques fines, de suggérer de nouvelles catégories, de mettre en valeur les facettes sémantiques des entités et de révéler leur éventuelle polysémie. Il importe ici de spécifier la nature de l'information obtenue par le biais de ces étiquettes : si nous avons distingué ci-avant deux types d'amélioration pour le traitement des

[1]Une alternative bienvenue serait d'associer aux étiquettes « récoltées » dans un corpus donné une information temporelle : l'évolutivité des référents des entités nommées aurait alors un véritable ancrage (non plus palpable par le seul biais du corpus) et une construction incrémentale de la ressource serait alors envisageable. Ceci implique cependant une maitrise fine des marqueurs temporels.

entités nommées touchant, pour l'un, à davantage de précision quant au référent et, pour l'autre, à une désambiguïsation, notre approche n'est cependant pas en mesure de dire à quel niveau jouent les étiquettes de la ressource ni de caractériser le type d'ambiguïté révélée (homonymie, facette, métonymie), s'il est question d'ambiguïté. Autrement dit, par cette ressource nous ne visons pas un phénomène de sens lié aux entités nommées en particulier : qu'il s'agisse d'homonymie, de polysémie ou de métonymie, la méthode permet de les révéler, sans nécessairement les caractériser.

La seconde étape correspond à l'exploitation de cette ressource pour l'annotation fine voire la désambiguïsation d'entités nommées. Idéalement, cet usage de la ressource se fait après qu'elle ait servi, en amont, à la détermination de la structure d'un modèle à l'aide de la suggestion de catégories (pour les objets) et de la révélation de phénomènes de sens (pour les relations). L'exploitation de la ressource menée ici correspond cependant davantage à une validation de cette dernière, avec la combinaison de l'annotation fine permise par la ressource d'une part et l'annotation plus classique d'un système à base de règles d'autre part, combinaison donnant ainsi lieu, lorsque cela est possible, à une double annotation des entités. De la sorte, la ressource est indirectement utilisée dans le cadre d'un modèle, celui du système traditionel d'annotation d'entités nommées auquel elle est couplée. Afin d'illustrer le gain de la double annotation, imaginons une application (relevant de l'extraction d'information) pour laquelle il importerait d'extraire d'un corpus toutes les personnes appartenant à l'armée. Un système classique de reconnaissance d'entités (qu'il soit symbolique ou à base d'apprentissage) serait capable de repérer dans le texte toutes les « personnes », parmi lesquelles des journalistes ou autres, non pertinentes du point de vue de l'application ; couplé à notre ressource, le système serait en mesure de pointer, par le biais de la double annotation, les « personnes » ayant aussi une étiquette « général, lieutenant, etc. », satisfaisant par là aux exigences de l'application. Cette combinaison de systèmes à grains différents a en outre permis une évaluation des résultats de la ressource.

L'objectif de cette approche est donc de fournir une double annotation des entités nommées, de « poser » ou révéler la polysémie de certaines entités, et de faire quelques pas sur le chemin de la désambiguïsation. Avant de détailler plus avant la réalisation de cette double annotation, il convient d'évoquer les travaux relatifs à ce champ de recherche.

## 6.2 Travaux connexes

Les travaux relatifs à l'identification et à l'annotation des entités nommées, pour l'essentiel fondés sur des méthodes par apprentissage, ont déjà été évoqués dans la section 1.4. L'approche ici présentée différant quelque peu dans ses objectifs des systèmes de reconnaissance d'entités nommées classiques, nous proposons de guider ce tour d'horizon des travaux connexes en fonction de deux points de vue : l'annotation fine et la désambiguïsation des entités nommées d'une part, et la construction de lexiques sémantiques d'autre part.

Bien que relativement récente, la tendance à considérer de plus près les entités nommées a déjà suscité des travaux intéressants. Il est possible de distinguer trois types de travaux : ceux dont le but est de désambiguïser les entités nommées, ceux cherchant à construire une ressource spécifique pour le traitement des entités nommées, et enfin ceux combinant les deux précédents, c'est à dire cherchant à faire de la désambiguïsation tout en exploitant une ressource spécifique. Commençons par le premier type de travaux. La plupart d'entre eux reposent sur des méthodes d'apprentissage : à partir d'un corpus dans lequel les entités nommées visées ont été annotées, l'objectif est d'« apprendre » des traits (ou caractéristiques) de nature linguistique et statistique caractérisant ces entités afin d'en déduire des modèles probabilistes de reconnaissance et de typage. Des lexiques (simple mise en correspondance, une entité = un type) peuvent aussi être utilisés par ces algorithmes. Les recherches exploitant ces méthodes se sont d'abord intéressées aux noms de lieu, avec notamment les travaux de [Fleischman, 2001], [Lee et Lee, 2004] et [Li *et al.*, 2003]. En usant de méthodes similaires [Fleischman et Hovy, 2002] ainsi que [Mann et Yarowsky, 2003] ont porté leur attention sur la sous-catégorisation de noms de personnes. Le deuxième type de travaux s'intéresse à la construction de ressource spécifique pour les entités nommées. Peu de choses (à notre connaissance) existent à ce jour : [Mann, 2002] construit automatiquement une ontologie de noms propres à l'aide de patrons de co-occurrence, avec pour objectif une intégration des données obtenues à WordNet et [Maurel et Tran, 2006] mène des travaux similaires au sein du projet Prolex, réunissant (« manuellement », à partir de ressources existantes) des informations sur des noms propres, avec une dimension multilingue. Ayant pour idée sous-jacente la notion d'ontologie, ces travaux cherchent à caractériser les noms propres de manière statique, tandis que la ressource présentée ici s'attache à refléter une information contextuelle, en prenant en compte tous ses changements. Enfin, certains tentent de traiter finement les entités nommées en s'appuyant sur des ressources spécifiques. [Bunescu et Paşca, 2007] présentent une approche fort intéressante de désambiguïsation d'entités nommées qui exploite la ressource encyclopédique Wikipédia.

Le travail présenté dans [Paşca, 2004] semble être celui qui se rapproche le plus de ce que nous présentons : il construit une ressource à partir de corpus pour annoter finement les entités nommées. Sa méthode de construction de ressource diffère cependant de celle présentée ci-après dans la mesure où elle n'utilise pas d'analyse syntaxique. Signalons pour finir [Nissim et Markert, 2003], dont les travaux sur la métonymie participent d'un domaine de recherche similaire.

Il convient de dire un mot sur l'acquisition de lexiques sémantiques, dans la mesure où nous nous inspirons de ces travaux sur les unités lexicales pour traiter les entités nommées. Cette tâche a connu un essor important depuis l'apparition de corpus de grande taille. Jusque récemment, l'acquisition, à partir de textes, de labels sémantiques associés à des unités lexicales était basée sur des patrons lexico-syntaxiques développés à l'aide d'algorithmes d'apprentissage ; seule Pasca a proposé d'appliquer ce processus aux entités nommées. L'exploitation de corpus analysés syntaxiquement (avec des relations de dépendances) est plus récente, les travaux de [Phillips et Riloff, 2002] vont dans cette direction. Enfin, il importe de replacer notre approche dans la droite ligne des nombreux travaux dit de « linguistique de corpus » [Habert *et al.*, 1997], avec ceux de Didier Bourigault[1] qui propose une méthode de rapprochement sémantique entre mots ou groupes de mots en fonction de leur distribution syntaxique dans un corpus donné ainsi que ceux de [Zweigenbaum *et al.*, 1997], ces derniers cherchant notamment à combiner connaissances du domaine et connaissances acquises sur corpus.

## 6.3 Méthode

### 6.3.1 Construction d'une ressource d'entités nommées

Ce que nous appelons ressource d'entités nommées est une liste d'entités nommées avec pour chacune d'elles une liste d'étiquettes sémantiques fines potentielles (par exemple les étiquettes « porte-avions », « maréchal », « avenue », « hôpital » pour l'entité nommée Foch) provenant d'un corpus. Le principe général de construction de cette ressource est l'identification dans le corpus de mots ou groupes de mots étant en relation avec les entités nommées et pouvant servir d'étiquettes sémantiques. Afin de repérer et d'associer pertinemment entités et étiquettes, nous proposons d'exploiter un analyseur syntaxique robuste. Le processus de construction de la ressource se déroule en trois étapes : identification des relations syntaxiques pertinentes permettant d'associer des entités avec des étiquettes, construction effective de la ressource et gestion des étiquettes par le

[1] http ://w3.univ-tlse2.fr/erss/voisinsdelemonde/

calcul de cliques. Il convient de détailler chacune de ces étapes ; nous précisons auparavant les données et les outils utilisés.

#### 6.3.1.1 Analyseur syntaxique et corpus utilisés

Pour les différentes expérimentations effectuées nous avons utilisé deux corpus en français : un corpus contenant l'ensemble des articles du journal *Le Monde* de 1992 à 1996 (2 830 180 phrases) et un corpus contenant articles et dépêches provenant de différentes sources et traitant tous de la crise en Côte d'Ivoire entre 2002 et 2003[1] (331 433 phrases). Dorénavant, nous les nommerons respectivement corpus LM92-96 et corpus CI02-03. Ces corpus ont été traités à l'aide de l'analyseur syntaxique robuste XIP (*Xerox Incremental Parser* ([Aït-Mokhtar et Chanod, 1997] et [Aït-Mokhtar *et al.*, 2002] ; une description détaillée de l'analyseur figure en annexe D.).

#### 6.3.1.2 Identification des relations syntaxiques pertinentes

Nous disposons de corpus, d'un analyseur syntaxique et avons pour objectif d'associer à des entités nommées des étiquettes sémantiques précisant leur référent. Les objets manipulés lors de la construction de la ressource sont donc des entités, des étiquettes et des relations syntaxiques : il importe dans une première étape de déterminer précisément les caractéristiques de ces objets. Pour cela, il est possible de s'appuyer sur des critères linguistiques d'une part, et de prendre en compte les conclusions d'une observation empirique des résultats de l'analyse syntaxique d'autre part. Ainsi nous cherchons tout d'abord des entités, que nous choisissons de définir de la manière suivante : est une entité nommée potentielle tout nom ou groupe nominal dont la tête commence par une majuscule. Nous cherchons ensuite des mots pouvant servir d'étiquettes à ces entités ; pour ce faire, nous nous intéressons davantage à des syntagmes modifieurs entretenant avec l'entité potentielle un rapport déterminatif et non explicatif ou descriptif. Est ainsi une étiquette potentielle tout nom ou syntagme nominal dont la tête nominale commence par une minuscule. Nous excluons donc les adjectifs (plus qualifiants que classifiants) ainsi que les expressions temporelles (nom de mois, années, etc.) et numériques. Nous observons ensuite toutes les relations syntaxiques établissant un lien entre ces deux types d'objets, étiquette potentielle et entité potentielle. Ainsi, pour chaque entité potentielle, on identifie une liste d'étiquettes potentielles rattachées à cette entité par un certain nombre de relations.

[1] Ce corpus a été constitué dans le but d'expérimentations sur les entités nommées dans le cadre du projet Infom@gic. Infom@gic est un projet national (pôle de compétitivité Cap Digital) en Ingénierie des Connaissances faisant intervenir des acteurs industriels et universitaires. XRCE est plus particulièrement impliqué dans le sous-projet *Extraction d'Information*.

Nous appellerons dorénavant chaque combinaison [étiquette potentielle-relation syntaxique] un contexte syntaxique.

Le tableau 6.1 illustre les contextes syntaxiques les plus fréquents pour trois entités nommées potentielles (Chirac, Foch, PC) dans le corpus LM92-96 à partir d'une analyse XIP :

| ENTITÉ : *Chirac* | ENTITÉ : *Foch* | ENTITÉ :*PC* |
|---|---|---|
| 1.NOUN ;président.NMOD | 1.NOUN ;porte-avions.NMOD | 1.NOUN ;PS.COORD |
| 1. NOUN ;candidat.NMOD | 1.NOUN ;avenue.NMOD | 1.noun ;secretaire general.NMOD_DE |
| 1.NOUN ;gouvernement.NMOD | 1.NOUN ;hôpital.NMOD | 1.NOUN ;congrès.NMOD_DE |
| 1.NOUN ;école.NMOD | 1.NOUN ;maréchal.NMOD | 2.NOUN ;Macintosh.COORD |
| 1.NOUN ;Balladur.COORD | 1.NOUN ;Clémenceau.COORD | 2.NOUN ;secrétaire.NMOD_DE |
| 1.NOUN ;monsieur.NMOD | 1.NOUN ;successeur.NMOD_A | 1.NOUN ;membre.NMOD_DE |
| 2.NOUN ;Jospin.COORD | 1.NOUN ;bord.NMOD_DE | 1.NOUN ;dirigeant.NMOD_DE |
| 2.NOUN ;époux.COORD | 2.NOUN ;premier.NMOD_POUR | 1.NOUN ;comité.NMOD_DE |
| 1.NOUN ;élection.COORD | 1.NOUN ;service.NMOD_DE | 1.NOUN ;ordinateur.NMOD |
| 1.NOUN ;candidat.ATTRIB | 1.NOUN ;place.ATTRIB | 1.NOUN ;conférence.NMOD_DE |

TAB. 6.1 – Contextes syntaxiques les plus fréquents pour les entités *Chirac*, *Foch* et *PC*.

Le contexte syntaxique 1.NOUN ;président.NMOD (contexte syntaxique le plus fréquent pour l'entité Chirac) se lit de la manière suivante : « NOUN ;président » décrit l'étiquette potentielle impliquée dans la relation syntaxique à l'aide de son type (NOUN) et de son lemme (président) ; « 1 » signifie que cette étiquette potentielle est le recteur de la relation syntaxique (« 2 » : étiquette en position régie) ; « NMOD » décrit le type de la relation syntaxique (NMOD : modifieur de nom ; NMOD_DE : modifieur de nom impliquant la préposition de ; COORD : relation de coordination ; etc.)

C'est à partir de ces listes de contextes que nous déterminons les relations syntaxiques pertinentes pour établir un lien entre une entité nommée et ses étiquettes potentielles. Nous avons identifié trois types de relations pertinents :

- La relation modifieur de nom sans préposition (NMOD). Dans le tableau 1, elle permet de faire le lien entre Chirac et « président, candidat, gouvernement ».
- La relation attribut (ATTRIB). Dans le tableau 1, elle permet de faire le lien entre Chirac et « candidat » et le lien entre Foch et « place ».
- La coordination (COORD) Dans le tableau 1, elle permet de faire le lien entre deux entités du même type (Chirac avec Balladur, Jospin et Juppé) .

Le choix manuel de ces relations syntaxiques est la seule étape supervisée de notre méthode. Ce choix dépend de l'analyseur syntaxique utilisé et de la langue du texte. Si ces choix sont faits empiriquement, il n'en reste pas moins que les

relations identifiées correspondent à celles habituellement reconnues au sein du groupe nominal étendu et qu'elles sont constantes au sein d'une langue.

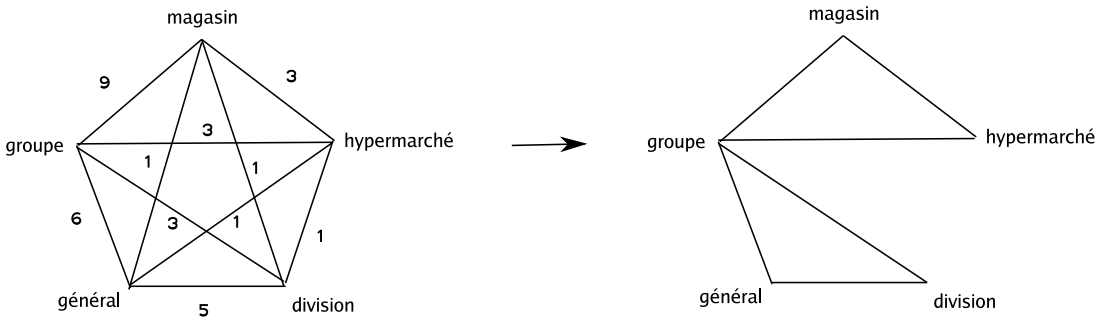
L'avantage d'une approche syntaxique par rapport à une approche de simple recherche de patrons peut être illustré par deux types d'énoncés : les énoncés du type *Le très vieux Foch* sont retenus à tort avec une approche à base de patrons puisque vieux peut être un nom mais pas avec l'approche syntaxique puisque celle-ci permet d'attribuer à vieux le type « adjectif » ce qui l'exclut des étiquettes potentielles. De même, des énoncés plus complexes peuvent être exploités. Par exemple pour l'énoncé *Bush est, pour le moment, le Président des Etats Unis*, l'analyse syntaxique profonde permet d'établir une relation « attribut » entre Bush et président, et ainsi de saisir une étiquette à distance.

#### 6.3.1.3 Construction effective de la ressource

La seconde étape correspond à la construction de la ressource à proprement parler. À partir du corpus, nous extrayons toutes les occurrences des relations syntaxiques pertinentes identifiées lors de l'étape précédente. Dorénavant, nous appellerons $R$ cette liste de relations pertinentes. Pour chacune de ces relations syntaxiques, le recteur est introduit dans la ressource en tant qu'étiquette potentielle et le régi en tant qu'entité nommée potentielle. De cette manière, nous pouvons construire une matrice $M$ (étiquettes x entités) où chaque ligne est une étiquette potentielle et chaque colonne est une entité nommée potentielle. Dans cette matrice, la valeur de (ligne=candidat ; colonne=Chirac) est égale à la somme des fréquences de chaque relation syntaxique de $R$ entre « candidat » et Chirac dans le corpus. Nous verrons dans le paragraphe suivant qu'il existe différentes manières de filtrer les informations contenues dans cette ressource. Pour nos expérimentations, nous avons choisi d'exclure simplement toutes les valeurs (ligne, colonne) égales à 1, autrement dit toutes les relations étiquette-entité présentes une seule fois dans le corpus[1].

Les étiquettes obtenues sont la plupart du temps pertinentes mais il peut y avoir un effet de surproduction de ces étiquettes. *Leclerc* peut être un char, un supermarché, un maréchal, mais les étiquettes obtenues sont les suivantes : « groupe, magasin, char, division, général, centre, programme, combat, colonne, hypermarché, maréchal, supermarché, bataille ». « Supermarché » et « magasin » semblent renvoyer à la même annotation fine alors que « groupe » reste une étiquette ambiguë (un groupe peut renvoyer à un parti politique, une entreprise, un rassemblement d'individus, etc.). Nous proposons de substituer à cette liste d'étiquettes une liste de groupes d'étiquettes, c'est-à-dire un ensemble d'éti-

[1]Ce seuil pourrait être ajusté en fonction de la taille du corpus.




Extrait du sous-graphe G(L,U) pour l'entité *Leclerc*.

Extrait du sous-graphe G'(L',U') pour l'entité *Leclerc* avec k=2.

FIG. 6.1 –

quettes pouvant renvoyer à une même annotation fine.

La méthode de construction de ces ensembles d'étiquettes est la suivante : à partir de la matrice M(étiquettes x entités) citée ci-dessus, nous construisons un graphe valué G d'étiquettes. Dans ce graphe, chaque sommet est une étiquette et le nombre d'arêtes reliant deux sommets correspond au nombre d'entités ayant ces deux étiquettes en commun. À partir de ce graphe nous effectuons des regroupements d'étiquettes par le calcul de cliques. Une clique est un sous-graphe complet maximal, soit un ensemble le plus grand possible d'étiquettes toutes reliées deux à deux. Les cliques ont déjà été exploitées pour la représentation d'espaces sémantiques d'unités lexicales à partir de dictionnaires de synonymes [Ploux et Victorri, 1998, Victorri, 2002], l'idée étant alors de considérer que chaque clique de synonymes correspond à un sens très précis de l'unité lexicale étudiée. De la même manière, dans notre cas, une clique d'étiquettes correspond à une annotation d'entité très fine. Si l'étiquette « groupe » est ambiguë, la clique « groupe, société, firme » permet de contraindre l'emploi de « groupe » en tant qu'entreprise.

Pour une entité nommée E possédant une liste d'étiquettes L nous extrayons de ce graphe G le sous-graphe G(L,U), c'est-à-dire le sous-graphe composé des sommets (étiquettes) L et des arêtes U reliant ces sommets. Nous construisons alors un graphe simple G'(L,U') dans lequel deux sommets sont reliés par une arête si dans le sous-graphe G(L,U) le nombre d'arêtes entre ces deux mêmes sommets est supérieur à k. C'est dans ce graphe simple G'(L,U') que nous calculons les cliques d'étiquettes.

Si k=0, l'entité E possédant toutes les étiquettes du sous graphe G(L,U), E permet de faire le lien entre toutes ces étiquettes et le calcul donne une et une seule clique composée des étiquettes L, ce qui n'est pas notre but ici. Nous

proposons d'illustrer les cliques obtenues avec k=1 et k=2 à l'aide de l'entité Leclerc (cf. tableau 6.2). Si l'étiquette « groupe » est ambiguë, les cliques dont elle fait partie ne le sont plus : pour k=2, les deux premières cliques « groupe ; magasin ; hypermarché » et « groupe ; division ; général » correspondent bien à deux annotations fines et non ambiguës pour l'entité Leclerc.

| Cliques pour *Leclerc* avec k=1 | Cliques pour *Leclerc* avec k=2 |
|---|---|
| groupe ; magasin ; hypermarché. | groupe ; magasin ; hypermarché. |
| groupe ; magasin ; supermarché. | groupe ; division ; général. |
| groupe ; division ; général. | groupe ; programme. |
| groupe ; division ; programme. | centre. |
| général ; maréchal. | général ; maréchal. |
| char. | char. |
| centre. | programme ; combat. |
| programme. combat | colonne. |
| colonne. | supermarché. |
| bataille. | bataille. |

TAB. 6.2 – Cliques obtenues pour l'entité *Leclerc* avec k=1 et k=2 .

Si nous insistons sur cette variable k, c'est qu'elle permet de jouer sur le nombre de cliques calculées, autrement dit sur la finesse d'annotation des entités. $k = 0$ revient à considérer que toutes les entités sont non ambiguës puisqu'elles possèdent toutes une seule clique d'étiquettes et donc une seule annotation possible. Plus on augmente k, plus on se rapproche de la liste classique d'étiquettes telle celle présentée pour l'entité Leclerc (« groupe,magasin,char », etc.). Pour les expérimentations qui suivent nous avons utilisé $k = 2$ (ce choix correspond à un choix empirique).

En utilisant cette méthode, nous avons construit une ressource à partir du corpus LM92-96. Pour cette construction, l'ensemble de relations $R$ est composé des relations « modifieur de nom sans préposition » (NMOD) et « attribut » (ATTRIB) et sont exclues toutes les relations étiquette-entité présentes une seule fois dans le corpus. La ressource obtenue contient 15 040 entités différentes et 2 547 étiquettes différentes. Le temps de construction de cette ressource correspond au temps d'analyse du parseur.

Les tableaux 6.3 et 6.4 présentent quelques entités et étiquettes extraites de cette ressource. Le tableau 6.3 présente trois entités (Chirac, Foch et PC) avec pour chacune d'elles un extrait de leurs cliques d'étiquettes. Les cliques de l'entité Chirac rendent compte de différentes facettes d'une même entité, celles de l'entité Foch décrivent différents emplois polysémiques, enfin celles de l'entité PC révèlent un cas d'homonymie (ordinateur vs. Parti communiste). Le tableau 6.4 décrit quelques étiquettes de la ressource avec les entités les plus fréquentes pour chacune

de ces étiquettes.

| ENTITÉ : **Chirac** (25 cliques d'étiquettes) |
|---|
| président,famille,ère,gouvernement<br>candidat,ère, plan,effet,vote<br>famille,époux, couple |
| ENTITÉ : **Foch** (4 cliques d'étiquettes) |
| porte-avions<br>avenue<br>hôpital<br>maréchal |
| ENTITÉ : **PC** (14 cliques d'étiquettes) |
| groupe,maire,liste,candidat,maison<br>groupe,machine, micro-ordinateur,ordinateur<br>microprocesseur,machine, puce,processeur |

TAB. 6.3 – Entités Chirac, Foch et PC avec un extrait de leurs cliques.

#### 6.3.1.4 Etape facultative

Une troisième étape facultative consistant en un filtrage de la ressource peut être effectuée. Ce filtrage peut correspondre à l'adaptation de la ressource pour une application donnée. Il est possible de supprimer certaines annotations non pertinentes comme d'en regrouper d'autres, ou encore d'ajouter certaines annotations ou entités nommées provenant d'une autre ressource. Si nous insistons sur cette étape facultative alors que nous ne l'avons pas appliquée dans nos expérimentations, c'est que même si notre ressource peut être exploitée telle quelle et déjà améliorer les résultats d'annotation d'entités nommées dans un texte, elle n'est pas une boîte noire et peut être manipulée en toute transparence en fonction d'une application donnée.

À ce stade de l'expérimentation, nous disposons donc d'une ressource associant des groupes d'étiquettes sémantiques fines à des entités nommées. Cette ressource peut, dans un projet applicatif précis, être utilisée comme aide à la conception de catégories d'un modèle. Elle peut également être utilisée telle quelle, pour ce que nous avons appelé l'annotation fine des entités nommées ; c'est cette perspective de développement qui, dans les sections suivantes, est présentée puis évaluée.

| ÉTIQUETTE : *acteur* (contient 11 entités) | ÉTIQUETTE :*assurance* (contient 19 entités) |
|---|---|
| Marina Vlady | AXA |
| Sharon Stone | GAN |
| Jean-Paul Belmondo | Allianz |
| Steve Martin | Generali |
| ÉTIQUETTE : *porte-avions* (contient 11 entités) | ÉTIQUETTE :*gouvernement* (contient 93 entités) |
| Charles de Gaulle | Balladur |
| Foch | Juppé |
| Clémenceau | Rocard |
| Indépendance | Chirac |
| ÉTIQUETTE : *ordinateur* (contient 20 entités) | ÉTIQUETTE : *avenue* (contient 50 entités) |
| Apple | Jean Jaurès |
| PC | Montaigne |
| Macintosh | Victor Hugo |
| IBM | Daumesnil |

TAB. 6.4 – Etiquettes de la ressource avec leurs entités les plus fréquentes.

### 6.3.2 Annotation fine ou première désambiguïsation

Dans cette section nous présentons l'intérêt d'une telle ressource pour l'annotation et la désambiguïsation d'entités nommées. En effet, sans faire appel à une méthode de désambiguïsation classique nécessitant de faire un choix entre plusieurs annotations possibles, il est possible, à l'aide de la ressource seule, d'annoter de manière fine et non ambiguë des entités nommées. Pour illustrer cela, nous détaillons les différents contextes d'emploi des occurrences d'entités nommées dans un corpus donné en fonction de l'information présente dans notre ressource. Nous appuyons cette étude sur le corpus CI02-03 (331 433 phrases) et toutes les fréquences d'occurrences que nous citons dans cette partie proviennent de ce corpus. La ressource a été construite à partir d'une liste de relations R, composée des relations « modifieur de nom sans préposition » (NMOD) et « attribut » (ATTRIB), et en excluant toutes les relations étiquettes-entités nommées présente une seule fois dans le corpus. La ressource construite contient ainsi 2 360 entités nommées différentes et 967 étiquettes différentes. Ces 2 360 entités représentent 284 702 occurrences d'entités, soit plus d'une entité par phrase. Nous ne connaissons pas le nombre réel d'entités nommées présentes dans ce corpus, cependant nous verrons dans le paragraphe6.3.3 que sur un extrait contenant 855 occurrences d'entités nommées, 70 % ont été annotées à l'aide de notre ressource.

Il existe différents cas de figure pour l'annotation des occurrences d'entité nom-

mée en fonction de l'information présente dans la ressource et dans le contexte :

1. L'entité nommée n'a qu'une seule clique d'étiquettes dans la ressource. Dans ce cas l'annotation fine est faite directement. C'est le cas par exemple de l'entité Daouda Konate qui ne possède dans la ressource que la clique « sergent ; sergent chef ». Dans la ressource construite à partir du corpus CI92-93, 84 % des entités de la ressource ne possèdent qu'une clique (87 % dans le corpus LM92-96). Ces entités représentent 60 195 occurrences dans le corpus, soit 21,1 % des occurrences d'entités annotées du corpus.
2. L'entité nommée a plusieurs cliques d'étiquettes dans la ressource. Trois types d'occurrences sont alors possibles :
   (a) Les occurrences ayant une relation syntaxique appartenant à R avec une étiquette appartenant à une seule clique parmi les cliques d'étiquettes possibles pour cette entité dans la ressource. Par exemple, l'énoncé « *Le président de la république, son excellence Laurent Gbagbo, déclarait que ...* » permet d'établir une relation entre l'entité Laurent Gbagbo et l'étiquette « président de la république » [1]. L'entité Laurent Gbagbo possède plusieurs cliques d'étiquettes, mais une seule (« président, candidat, chef d'état, président de la république ») contient l'étiquette « président de la république ». Dans ce cas, l'annotation fine est faite directement, ceci intervient pour 3 % des occurrences d'entités nommées annotées du corpus CI92-93.
   (b) Les occurrences ayant une relation syntaxique appartenant à R avec une étiquette appartenant à plusieurs cliques parmi les cliques d'étiquettes possibles pour cette entité dans la ressource. Par exemple, l'énoncé « *Le général Gueï prend la tête du conseil* » permet d'établir une relation entre l'entité Gueï et l'étiquette « général ». L'entité Gueï possède plusieurs cliques d'étiquettes dont deux contenant l'étiquette « général » (les cliques « président, général » et « général, division »). Dans ce cas, l'entité reste ambiguë ; cependant son ambiguïté est réduite puisque dans notre exemple, il n'y a plus que deux cliques possibles parmi les quatre que possède l'entité. Ce cas correspond à 5,6 % des occurrences d'entités nommées de CI92-93.
   (c) Les occurrences qui n'ont pas de relation syntaxique appartenant à R avec l'une des étiquettes potentielles de l'entité et dont l'entité nommée possède plusieurs cliques dans la ressource (70,3 % des occurrences d'entités nommées de CI92-93). Dans ce cas l'entité reste ambiguë :

[1] C'est parce qu'elle est présente dans la ressource que cette expression est utilisée comme étiquette.

nous gardons toutes les cliques d'étiquettes de l'entité, et une méthode de désambiguïsation pourra exploiter cette liste de cliques afin de déterminer la ou les cliques d'étiquettes adaptée(s) à l'énoncé.

Cette ressource construite à partir d'un corpus n'est donc pas seulement une ressource de référence permettant de connaître toutes les annotations fines possibles d'une entité en vue d'une désambiguïsation ultérieure, elle permet déjà à elle seule d'annoter de manière fine et non ambiguë un nombre important (24,1 %) d'entités dans le texte. Si l'on considère que l'annotation par une seule étiquette est suffisante, ce taux passe même à 29,7 % puisque l'on peut ajouter le cas de figure 2b. Ainsi, à qualité de désambiguïsation égale, l'exploitation de cliques d'étiquettes apparaît comme bénéfique (26,3 % d'entités désambiguïsées à l'aide d'étiquettes non regroupées). Cette ressource met aussi en avant le fait que 84 % des entités sont non ambiguës dans ce corpus. Avant d'évaluer la qualité de ces annotations fines, nous proposons d'introduire ce que nous appelons la double annotation.

### 6.3.3 Double annotation

L'objectif est de combiner l'information contenue dans la ressource que nous venons de présenter avec un système classique de traitement d'entités nommées afin de permettre une annotation selon deux niveaux, général et particulier, avec l'idée sous-jacente d'exploiter chaque méthode pour la tâche pour laquelle elle est le mieux adaptée. Nous avons vu qu'il existe un nombre important de systèmes capables d'annoter des entités nommées selon des catégories générales, et ce avec des taux de précision supérieurs à 90 %. Il importe cependant pour ces systèmes que les textes à annoter soient du même genre et/ou domaine que ceux ayant guidé leur conception, voire qu'ils soient issus d'une même source. Ainsi, l'utilisation de systèmes classiques pour d'autres domaines que celui pour lequel ils ont été conçu amoindrit généralement les performances, notamment en faisant chuter le rappel (ce ne sont plus les mêmes entités que l'on retrouve ni les même types d'entités : certains types d'entités comme les noms de produits peuvent être présents dans certains domaines et pas dans d'autres). De manière identique, dès que l'on cherche à affiner les annotations d'entités nommées en passant de « organisation » à « entreprise », « association » ou « parti politique », les taux de précision chutent.

Notre approche d'annotation fine à partir d'une ressource d'entités nommées dépendante d'un corpus, dorénavant approche « ressource », a les qualités inverses : étant dépendante du corpus, elle permet de proposer des annotations précises et fines pour les entités correspondant au corpus puisqu'elles en ont été

extraites. La limite de cette approche est qu'elle est difficilement rattachable à une catégorisation générale préalable de type « personne, organisation, lieu, etc. ». Elle permet de dire que Coca-Cola est une « société », mais pas que « société » est rattaché à la catégorie « organisation », du moins pas sans faire appel à une ressource de nature ontologique externe de type Wordnet.

L'idée est alors, non d'essayer de fusionner ces deux approches, mais plutôt de les rendre complémentaires. Il nous semble important de maintenir la combinaison « approche classique/approche ressource » afin d'exploiter au mieux les qualités de chacune (ceci n'excluant pas de les faire interagir) : la première permet d'annoter des entités avec des catégories connues alors que la seconde permet de proposer de nouvelles (sous-)catégories. Autrement dit, nous proposons une double annotation, l'une générale obtenue à l'aide d'une approche symbolique classique et l'autre fine obtenue à l'aide de l'approche « ressource ». Ci-dessous, nous illustrons avec l'entité Leclerc le type de résultats que nous cherchons à obtenir.

*Ce supermarché Leclerc vient d'être inauguré.*
Annotation générale = ORGANISATION
Annotation fine = supermarché

Pour cet énoncé l'approche classique est en mesure de proposer, grâce à la structure (supermarché + nom commençant par une majuscule), une annotation générale de type « organisation ». L'approche « ressource » est quant à elle en mesure de proposer une annotation plus fine : malgré l'ambiguïté de l'entité Leclerc (cf. les cliques de cette entité en section 6.3.1.3), la clique « supermarché » peut lui être attribuée directement grâce à l'identification de la relation « modifieur » avec le nom supermarché dans l'énoncé et à la présence de la clique « supermarché » dans la ressource (cas 2a dans la section 6.3.2). Ainsi, les deux méthodes se révèlent complémentaires, l'une venant enrichir l'annotation générale de l'autre par une étiquette sémantique de grain plus fin. L'objectif est, idéalement, d'avoir une double annotation pour toutes les occurrences d'entités. Cependant, certains énoncés ne permettent qu'une annotation partielle. Ainsi, si elle n'est pas possible dans tous les cas, la double annotation permet d'enrichir l'annotation de certaines entités nommées d'une part (comme pour l'énoncé ci-dessus), et d'annoter, au moins partiellement, un nombre plus important d'occurrences d'entités dans le texte d'autre part.

## 6.4 Evaluation

### 6.4.1 Evaluation de l'annotation fine

Dans cette partie nous présentons une évaluation de la qualité des annotations fines obtenues à l'aide de l'approche « ressource ». Toute la difficulté est de savoir comment juger de la pertinence ou justesse d'une annotation fine : dire que Coca-Cola est une « société » est-il plus approprié que dire que c'est une « firme » ? Est-ce que dire que Bush est un « président » permet de dire que c'est un « homme politique » ?

Nous revendiquons un système d'annotation double (classique + fin), mais afin de comparer les approches classique et « ressource » sur des annotations identiques, nous avons choisi d'établir une correspondance entre nos étiquettes fines et les catégories « personne », « lieu » et « organisation » du modèle générique adopté pour la reconnaissance d'entités nommées par XIP. Nous insistons sur le fait que cette mise en correspondance n'est faite que dans le but de l'évaluation et non dans le but de construire une ressource hiérarchique à deux niveaux interdépendants ; cette démarche confirme par ailleurs la nécessité d'un modèle pour l'appréhension des entités nommées.

Pour l'approche classique, différents systèmes d'annotation peuvent être utilisés, qu'ils soient de type symbolique ou basés sur de l'apprentissage. Nous avons choisi d'utiliser le système de reconnaissance d'entités nommées fondé sur une approche symbolique intégrée au sein de l'analyseur syntaxique XIP [**?**]. Ce choix se justifie par le fait que XIP nous donne la possibilité d'effectuer les deux types d'annotation, classique et fin, au sein d'un même processus. Outre le gain en rapidité, cela permet de maintenir une certaine cohésion entre les deux approches, les annotations étant faites sur des unités de texte provenant de la même analyse. Le système symbolique classique que nous avons utilisé a fait l'objet d'une évaluation pour le français [Rebotier, 2006]. Précisons tout d'abord que l'élaboration de ce système (constitution de lexiques et écritures de règles) a été guidée par l'étude d'un corpus composé d'articles du journal Libération (Lib1). L'évaluation a ensuite été menée à partir de deux corpus : un corpus similaire, composé d'autres articles de Libération (Lib2), et un corpus plus éloigné thématiquement afin d'évaluer la transposabilité inter-domaine du système, composé d'extraits du corpus CI92-93 (cf. section 6.3.1.1). La F-mesure est de 0,90 pour l'extrait Lib2 et de 0,80 pour l'extrait CI. Nous avons choisi l'extrait CI pour notre évaluation dans le but de montrer que la double annotation, outre sa finesse, permet de combler la perte de qualité d'un système classique lorsque l'on change de domaine. L'extrait CI contient 765 entités annotées manuellement par une catégorie

(« lieu »,« personne »,« organisation » ). Sur cet extrait, la précision obtenue par l'approche symbolique classique est de 86,58 %, et le rappel de 74,25 %. L'évaluation consiste à tester si notre approche « ressource » permet d'améliorer les résultats d'annotation de l'approche classique. C'est pourquoi nous ne nous intéressons dans cette évaluation qu'à la qualité des résultats obtenus sur les entités « oubliées » par l'approche classique. Sur les 765 entités du corpus de référence, 149 ont été « oubliées » par cette dernière. La plupart du temps, ces oublis sont dus à des contextes d'emplois difficiles à traiter pour l'approche classique (ex. Mais Hanny Tchelley ne comptait pas s'arrêter là.). Nous avons donc évalué les résultats de l'approche « ressource » sur ces 149 cas.

La correspondance entre étiquettes provenant du corpus et étiquettes classiques a été faite par 4 juges qui ont dû classer les premières en fonction des secondes, ou dans une catégorie « autre » (par exemple l'étiquette firme a été apparentée à ORGANISATION par les quatre juges, administration à PERSONNE ou ORGANISATION et voiture à AUTRE ). L'accord inter-annotateur est de 92 %. Nous avons intégré ces jugements dans chaque clique d'étiquettes (cf. colonnes 3 et 4 du tableau 6.5). Nous présentons ci-dessous un extrait des 37 entités qui ont été correctement annotées par l'approche « ressource » (cf. tableau 6.5) ainsi que la seule erreur relevée (cf. 6.6). Une entité correctement annotée correspond à une entité annotée par une seule clique (cas 1 ou 2.a. de la section 6.3.2). En terme d'extraction d'information, le tableau 6.5 illustre l'intérêt de cette annotation fine. Si l'on recherche toutes les personnes qui on eu la fonction de ministre, les entités Alassane Ouattara et Niamien Messou sont sélectionnées. Si, en revanche, on ne s'intéresse qu'aux personnes qui ont un grade militaire, seul Alassane Ouattara est sélectionné. Dans ces deux requêtes fines l'entité Hanny Tchelley n'est pas pertinente et peut facilement être exclue. L'erreur (cf. tableau 6.6) vient du fait que l'entité Lama Bamba possède dans notre ressource l'étiquette « presse ». Cette erreur provient d'énoncés tels que celui-ci : « Pour l'instant, je n'en dis pas plus, a sobrement déclaré à la presse Lama Bamba ». Dans ce cas, le lien établi entre presse et Lama Bamba est dû à une erreur de l'analyse syntaxique.

Le bilan global est le suivant : sur les 149 entités oubliées par l'approche symbolique classique, 37 ont été correctement annotées par l'approche « ressource » et une seule a été mal annotée. 111 entités n'ont pas été annotées : elles correspondent soit à des entités qui ne font pas partie de la ressource et donc ne sont pas annotées, soit à des entités qui font partie de la ressource, mais ne sont pas désambiguïsées (cas 2b ou 2c de la section 6.3.2), soit à des entités qui font partie de la ressource mais sont annotées avec l'étiquette « autre » par les quatre annotateurs. La précision de ce sous-ensemble de 149 annotations est de 97,4 % et le rappel est de 24.8 %. Rapporté à l'ensemble des entités du texte (soit 765

| Occurrences d'entités | Annotations de référence | Cliques obtenues | Correspondance |
|---|---|---|---|
| Hanny Tchelley | PERSONNE | animateur | PERSONNE(4/4) |
| M. Alassane Ouattara | PERSONNE | commandant, sergent, professeur, ministre | PERSONNE(4/4) |
| Niamen Messou | PERSONNE | ministre, délégué | PERSONNE(4/4) |
| Le patriote | ORGANISATION | quotidien | ORGANISATION(4/4) |
| Pdci | ORGANISATION | Régime | ORGANISATION(3/4) |
| Man | LIEU | capitale | LIEU(4/4) |

TAB. 6.5 – Extrait des entités correctement annotées.

| Occurrences d'entités | Annotations de référence | Cliques obtenues | Correspondance |
|---|---|---|---|
| Lama Bamba | PERSONNE | presse | ORGANISATION (4/4) |

TAB. 6.6 – Entité mal annotée.

entités), cela permet une augmentation du taux de précision, passant de 86,58 % à 87,18 %, et surtout une augmentation significative du taux de rappel, qui passe lui de 74,25 % à 79,08 % (cf. tableau 6.7).

| | Méthode classique | Méthode hybride | Gain |
|---|---|---|---|
| Précision | 86,58 % | 87,18 % | +0,6 |
| Rappel | 74,25 % | 79,08 % | +4,83 |
| F-mesure | 79,9 % | 82,9 % | +3,0 |

TAB. 6.7 – Comparaison des taux de précision et de rappel.

### 6.4.2 Remarque : étude sur les entités ajoutées

Le corpus de référence que nous avons utilisé a été annoté pour identifier des entités de type « personne », « lieu », « organisation ». Dans ces conditions, toute annotation d'entité ajoutée par l'approche classique par rapport à l'annotation de référence doit être considérée comme une erreur, ce qui a été fait. En revanche, les annotations provenant de l'approche « ressource » n'ont pas de frontière de catégorie, notre ressource peut attribuer l'étiquette « voiture » à l'entité BMW sans se soucier de savoir à quelle catégorie plus large cette étiquette appartient. Dans l'évaluation que nous avons présentée, l'étiquette « voiture » a été classée dans la catégorie « autre » par les quatre annotateurs, autrement dit, l'entité BMW a été ignorée. Or ces entités ajoutées par l'approche « ressource » sont

une très bonne illustration des annotations fines que cette approche peut proposer. Nous avons donc décidé de juger, cette fois-ci subjectivement, de la qualité de ces annotations fines. L'approche « ressource » a annoté 44 entités supplémentaires par rapport à celles qui avaient été annotées dans le corpus de référence. Sur ces 44 entités, 26 ont été jugées justes, 2 ont été jugées fausses et 16 ont été jugées neutres. Par neutre, nous entendons soit des entités ayant une étiquette ne portant aucune sémantique telle que l'étiquette « autre », soit des entités possédant plusieurs cliques d'étiquettes ne renvoyant pas à la même catégorie d'entité, autrement dit des entités encore ambiguës (voir le tableau 6.8).

| Occurrences d'entités | Annotations fines obtenues | Jugement |
|---|---|---|
| Ivoiro-Burkinabé | clan | juste |
| Krou | groupe | juste |
| Licorne | opération | juste |
| Messou | professeur | juste |
| T55 | char | juste |
| Ivoirien | ministre, gouvernement, président, télévision | neutre |
| ONG | autre | neutre |
| Républiques | deuxième | faux |
| Mano | fleuve | faux |

Tab. 6.8 – Exemples de jugements.

Ces jugements ne peuvent pas faire office d'évaluation, néanmoins ces entités « ajoutées » illustrent bien l'intérêt de l'annotation fine à partir du corpus puisqu'elle permet d'identifier des entités et des étiquettes qui n'auraient probablement pas pu être prévues a priori comme opération Licorne, clan Ivoiro-Burkinabé ou encore char T55.

## 6.5 Conclusion

Nous avons présenté une nouvelle approche pour l'annotation fine des entités nommées. La particularité de cette approche réside dans l'exploitation d'une ressource associant des cliques d'étiquettes sémantiques fines aux entités nommées : construite dynamiquement à partir de corpus, cette ressource permet de rendre compte des diverses caractéristiques (ou facettes) potentielles du référent d'une entité en contexte. A elle seule, cette ressource rend possible l'annotation fine d'une entité nommée, la mise en valeur de son éventuelle polysémie, et parfois même sa désambiguïsation (désambiguïsation pour 24.1 % des occurrences d'entités dans le corpus CI92-93). Elle présente également l'intérêt de faire émerger de nouveaux types d'entités à partir du texte. Couplée a un système standard

de reconnaissance d'entités nommées, cette ressource permet une annotation selon deux niveaux, général et fin. Nous avons parlé de l'intérêt des annotations fines obtenues pour extraire d'un texte des informations précises telles que « personnes ayant un grade militaire » ou « personnes ayant été ministre » (cf. section 6.4.1). L'intérêt de la double annotation réside dans sa double approche (classique et ressource) et dans la combinaison entre une structure imposée (personne / organisation / lieu) et des informations non-structurées provenant du texte. Malgré la difficulté de l'évaluation de cette double annotation, les résultats obtenus sont prometteurs et semblent attester l'efficience de cette approche.

La poursuite de ce travail est à envisager selon plusieurs perspectives. Il est possible tout d'abord d'affiner la construction de la ressource en considérant d'autres relations syntaxiques pour le repérage d'étiquettes sémantiques, comme par exemple les relations de coordination. Ensuite cette méthode d'annotation fine peut être exploitée pour d'autres langues. Nous envisageons également de réduire les cas d'annotation partielle (annotation seulement générale ou seulement fine) en faisant interagir les deux types d'annotations : savoir qu'une entité possède l'annotation fine « journaliste » doit pouvoir aider à proposer l'annotation générale « personne » pour cette même entité. Enfin, le traitement des entités demeurant ambiguës demeure un problème crucial. Nous présentons, dans le chapitre suivant, une méthode de résolution d'un type particulier de polysémie des entités nommées : la métonymie.

# Chapitre 7

# Résolution de métonymie

Comme annoncé au début de cette troisième partie, deux points peuvent être considérés pour enrichir l'annotation des entités nommées. Après la présentation d'une méthode pour l'annotation fine des entités nommées (cf. chapitre 6), ce chapitre aborde ainsi le problème de la résolution de métonymie pour ces unités.

La première section de ce chapitre reviendra tout d'abord sur la définition et la caractérisation du phénomène à traiter, tant du point de vue linguistique que du point de vue du TAL. La suivante s'attachera ensuite à présenter la campagne d'évaluation SemEval (*Semantic Evaluation*), dans le cadre de laquelle les travaux ici présentés ont été réalisés et évalués. Enfin, la dernière section se consacrera à la description du système mis au point.

La réalisation de ce travail ainsi que la participation à la campagne d'évaluation furent menées à bien en collaboration avec Caroline Brun et Guillaume Jacquet.

## 7.1 La métonymie des entités nommées

Il importe de caractériser le phénomène de métonymie des entités nommées avant de considérer les enjeux de son traitement en TAL.

### 7.1.1 Caractérisation linguistique

Sans revenir sur la définition de la métonymie lexicale (cf. section 2.3.2), il convient de souligner le fait que, s'il s'agit bien d'une classe ouverte permettant un nombre indéfini de glissements de sens (métonymies vives), il existe des changements de sens réguliers ou systématiques, au regard notamment des entités nommées. Examinons les exemples suivants :

*La politique américaine est plombée par l'Irak.*
*La France a gagné en demi-finale.*

Dans la première phrase, il n'est bien sûr pas question du pays proprement dit mais de l'événement qui s'y déroule, tout comme dans la seconde ou il s'agit non pas de la France en tant que telle mais d'une équipe sportive française. Cet usage des unités *Irak* et *France* en tant qu'événement et équipe sportive respectivement est possible pour d'autres noms de pays dans des situations similaires. Il existe bien d'autres exemples possibles, sur lesquels nous reviendrons par la suite, mais il est d'ores et déjà possible de postuler une certaine régularité et productivité des phénomènes de métonymie pour les entités nommées.

Outre ces caractéristiques, des études conduites par K. Markert, U. Hahn et M. Nissim [Markert et Hahn, 2002, Markert et Nissim, 2006] ont fait état de la fréquence de ce phénomène, montrant que 17 % de l'ensemble des occurrences dans un corpus de 27 magazines allemands étaient métonymiques, tout comme 20 % des occurrences des noms de pays et 30 % de celles des noms d'organisation sur des extraits significatifs du *British National Corpus*. Régulière, productive et fréquente, la métonymie des entités nommées constitue ainsi un réel intérêt pour le traitement automatique des langues.

### 7.1.2 Enjeux et moyens pour le TAL

#### 7.1.2.1 TAL et métonymie

Le traitement de la métonymie lexicale tout comme celui de la métonymie des entités nommées peut améliorer nombre de traitements dont les tâches d'extraction d'information, de question-réponse et de résolution de coréférence ; seul ce dernier cas est illustré ici. S'agissant des entités nommées, le repérage de glisse-

ments métonymiques peut en effet aider à la résolution de coréférence, comme dans l'exemple suivant (emprunté à [Markert et Nissim, 2006]) :

> **China** has agreed to let a United Nations investigator conduct an independant probe into [...] But it was unclear whether **Beijing** would meet past UN demands for unrestricted acess to [...]

où le fait de savoir que *China* et *Beijing* renvoient tous deux au gouvernement chinois permet d'établir une coréférence entre ces deux unités.

Si la résolution de métonymie constitue ainsi un réel enjeu pour le TAL, l'analyse du phénomène manque cependant d' « envergure » et les moyens disponibles pour mettre en œuvre son traitement font défaut. C'est en effet le constat opéré par [Markert et Nissim, 2006] qui, sans remettre en cause les nombreux travaux linguistiques sur le sujet, font toutefois remarquer que les exemples de métonymies sur lesquels ils se basent sont généralement imaginés pour l'occasion, avancés hors-contexte, sans laisser de place à l'hésitation quant à leur interpétation. Partant, ces exemples ne reflètent pas la réalité du phénomène de métonymie tel qu'il existe dans les textes, avec ses nombreux cas de figure et leur distribution (en terme de fréquence), réalité dont le TAL doit pourtant être informé, compte tenu des traitements sur le langage naturel bien souvent à grande échelle qu'il est amené à implémenter. Au-delà du manque d'une caractérisation objective indispensable au TAL, les auteures soulignent également l'insuffisance de ressources dédiées à ce phénomène. Noms propres et sens métonymiques n'ont en effet bien souvent que portion congrue dans la plupart des dictionnaires ou ressources lexicales telles que Wordnet, ressources à partir desquelles est par ailleurs effectuée l'annotation sémantique de corpus de référence (SenSeval-II, SenSeval-III), ces derniers étant par conséquent inexploitables pour le traitement de la métonymie. Nissim et Markert imputent, à juste titre nous semble-t-il, à ce manque de caractérisation et de ressources l'absence de véritables travaux et évaluations de traitement automatique du langage portant sur la résolution de métonymie.

#### 7.1.2.2 Les recherches de K. Markert et M. Nissim

Ce constat établi, K. Markert et M. Nissim se sont appliquées à conduire des études en corpus sur la métonymie de certaines catégories sémantiques afin de mieux les caractériser et de procéder à leur annotation dédiée en corpus. Il convient de souligner, à l'instar de [Poibeau, 2006], l'importance de ces travaux pour la compréhension et le traitement de la métonymie des entités nommées en TAL.

Nissim et Markert ont tout d'abord commencé par définir une série de prin-

cipes pour la construction de schémas d'annotation de métonymies. Parmi ces derniers, figurent des recommandations à caractère technique (encoder le texte en *XML* pour assurer l'indépendance de la plateforme) tout comme d'autres plus linguistiques, telles que, entre autres, la prise en compte de textes relevant de domaines et de genres différents (pour assurer la couverture de divers types de métonymies), l'utilisation de classes sémantiques et de patrons métonymiques[1] pour définir les catégories d'annotation ou encore la prévision d'une catégorie d'annotation pour les cas de métonymie non-conventionnelle. Se conformant à ce cadre de travail, les auteures ont ensuite déterminé un schéma d'annotation pour les métonymies en général, avant d'en élaborer deux plus précis, pour les classes sémantiques LOCATION et ORGANISATION. Le cadre général prévoit de faire la distinction entre trois types d'interprétation : interprétation littérale (*literal reading*), interprétation métonymique (*metonymic reading*), pour les cas réguliers comme pour les cas non réguliers de métonymies, et interprétation mixte (*mixed reading*) pour les cas faisant état de deux lectures métonymiques différentes ou d'une lecture métonymique et d'une lecture littérale. Les schémas plus précis d'annotation ont ensuite été réalisés à l'aide d'indications figurant dans la littérature linguistique sur ce sujet ainsi qu'à l'aide d'études en corpus [Markert et Nissim, 2005]. Nous reviendrons plus précisément sur ces schémas par la suite. Une fois établies ces instructions, les auteures ont annoté des corpus, pour la classe LOCATION [Markert et Nissim, 2002b] et la classe ORGANISATION [Markert et Nissim, 2006].

Ces travaux ont permis de dégager certains points essentiels à la compréhension et au traitement de la métonymie en TAL. L'étude en corpus a tout d'abord révélé l'existence d'autres patrons métonymiques que ceux traditionnellemnt reconnus, tel celui `org-for-index` pour les occurrences d'organisations en tant qu'indice boursier, ou encore `org-for-event`. Le travail d'annotation a ensuite permis de mieux caractériser la distribution des cas métonymiques pour les différentes classes sémantiques. Les taux satisfaisants d'accord inter-annotateurs ont également démontré la fiabilité d'annotation pour certains cas, comme la difficulté pour d'autres (interprétation mixte notamment). Enfin, deux corpus métonymiquement annotés sont désormais disponibles pour la réalisation de systèmes de résolution de métonymie pour les entités nommées. Fortes de ces travaux et recherches, Nissim et Markert ont proposé une tâche d'évaluation sur la résolution de métonymie lors de l'édition 2007 de la campagne SemEval. Avant de présenter cette dernière, nous souhaitons rendre compte rapidement des travaux existants sur la résolution de métonymie des entités nommées.

[1] Un patron métonymique permet de rendre compte du glissement de sens opéré entre le référent premier et le référent cible ; ces patrons sont de type `org-for-product` ou encore `place-for-people` et peuvent servir pour l'annotation.

### 7.1.3 Travaux existants

Comme nous venons de le voir, la tâche de résolution de métonymie des entités nommées émerge depuis peu à la faveur des recherches de M. Nissim et K. Markert. Il n'existe par conséquent que quelques travaux s'intéressant au traitement de ce phénomène, parmi lesquels, naturellement, ceux de M. Nissim et K. Markert. Ces dernières, après avoir défini un cadre de travail [**?**] ont en effet pu se pencher sur le problème de la résolution effective des glissements de sens pour les entités nommées, commençant tout d'abord par rapprocher cette tâche de celle de désambiguïsation lexicale (*word sense disambiguation*) [Markert et Nissim, 2002a]. Tout comme il importe de choisir, pour un mot donné, un sens précis parmi un ensemble de sens possibles (désambiguïsation lexicale), il importe de choisir, pour un mot appartenant à une classe sémantique, un changement de sens métonymique (ou patron métonymique) précis parmi un ensemble de partons métonymiques possibles ; l'objet de la désambiguïsation n'est plus un mot mais une classe sémantique. Ayant établi cela, M. Nissim et K. Markert ont dès lors expérimenté des méthodes de désambiguïsation lexicale pour la métonymie des entités nommées, utilisant des algorithmes d'apprentissage supervisés avec des traits dont elles avaient auparavant étudié la pertinence [Markert, 2000], pour les noms de lieux [Nissim et Markert, 2003] puis pour les noms d'entreprises [Nissim et Markert, 2005], obtenant à chaque fois des résultats prometteurs. Ces travaux ont permis de mettre en valeur l'importance du contexte avec le rôle des traits grammaticaux et de montrer leur possible généralisation au travers de contextes similaires (exploitation d'un thésaurus construit à partir de mesures de similarité entre mots). Les autres travaux existants sur le sujet sont également à base d'apprentissage. Dans la lignée des recherches de M. Nissim et K. Markert, Y. Peirsman a testé, pour les noms de lieux, des algorithmes supervisés (approche de Schütze) et non supervisés (*memory-based learning*), examinant la convenance de différents traits [Peirsman, 2006]. Enfin, T. Poibeau s'est lui aussi attelé à la résolution de métonymie des entités nommées [Poibeau, 2006], dans un cadre d'annotation différent et pour le français, s'appuyant sur des calculs de probabilité évaluant le pouvoir discriminant de tel ou tel trait pour les noms de lieux. Les méthodes présentées dans l'ensemble de ces travaux s'éloignent quelque peu de celles mises en œuvre pour la résolution de métonymie nominale ou verbale, ces dernières faisant état de glissements de sens beaucoup plus irréguliers pour lesquels il est utile de passer par une ressource lexicale de type dictionnairique ; pour plus d'informations sur ce sujet, voir l'état de l'art présenté par [Harabagiu, 1998]. À la suite de cette rapide présentation des travaux existants, il est temps de présenter la campagne SemEval.

## 7.2 La campagne SemEval

Cette section décrit l'édition 2007 de la campagne d'évaluation SemEval dans le cadre de laquelle a été élaboré le système de résolution de métonymie des entités nommées présenté ci-après. Après une présentation générale de la campagne et de la tâche, cette dernière sera explicitée plus en détail avec les catégories d'annotation.

### 7.2.1 Présentation générale

L'édition 2007 de la campagne *SemEval* s'inscrit dans une longue tradition de campagnes d'évaluation, à l'époque baptisées *SensEval*. Organisées conjointement aux conférences de l'*Association for Computational Linguistics* ou ACL depuis 1998, ces campagnes internationales avaient à l'origine pour vocation de rassembler des chercheurs autour des problèmes de polysémie et de désambiguïsation lexicale et de faire progresser les systèmes par le biais d'évaluations sur différents mots et différentes langues[1]. De campagnes en évaluations, les objectifs se sont quelque peu modifiés pour prendre une configuration plus générale, se focalisant non plus sur la seule désambiguïsation lexicale mais considérant un plus large panel de phénomènes sémantiques, d'où la nouvelle appellation *SemEval* pour *Semantic Evaluations*. SemEval 2007[2] a ainsi proposé 18 tâches d'évaluation autour de problèmes sémantiques tels que la désambiguïsation de prépositions, l'annotation d'expressions et de relations temporelles (*TempEval*), la désambiguïsation des noms de personne sur le web (*Web People Search*), ou encore la résolution de métonymie pour les entités nommées.

#### 7.2.1.1 La tâche de résolution de métonymie

La tâche proposée par K. Markert et M. Nissim est une tâche lexicale sur l'anglais, portant plus précisément sur la résolution de métonymie pour deux classes sémantiques, la classe LOCATION avec des noms de lieux et la classe ORGANISATION avec des noms d'entreprises. L'objectif pour les participants est de classer automatiquement des occurences pré-sélectionnées et en contexte de noms de lieux et d'entreprises, et ce en fonction de leur interprétation litérale ou non-litérale. Cette première alternative correspond à l'annotation "gros grain" ou *coarse-grained annotation*. Deux autres niveaux d'annotation sont possibles : un niveau "moyen" (*medium*) pour lequel il faut faire la distinction entre des in-

[1] Voir le site : http ://www.senseval.org/ pour plus de détails.

[2] *SemEval-2007, 4th International Workshop on Semantic Evaluations*. Voir le site : http://nlp.cs.swarthmore.edu/semeval/.

terprétations litérales, métonymiques et mixtes et, enfin, un niveau "fin" (*fine*), pour lequel il importe de préciser, en cas d'interprétation métonymique, le patron métonymique dont il est question. À titre d'illustration, nous pouvons reprendre les exemples donnés par les organisatrices dans le document décrivant la tâche [Markert et Nissim, 2007a] :

> At the time of **Vietnam**, increased spending led to inflation.
> **BMW** slipped 4p to 31p.
> The **BMW** slowed down.

Pour la première phrase, le nom *Vietnam* ne renvoie pas au pays mais à la guerre qui s'y est déroulée, il convient donc d'annoter cette entité comme `non-literal` (niveau 1), comme `metonymic` (niveau 2) ou comme `place-for-event` (niveau 3). De même, les occurrences de *BMW* dans les exemples suivants ne renvoient pas à l'entreprise mais à l'action de l'entreprise pour la première (`org-for-index`) et au produit de cette entreprise pour la seconde (`org-for-product`).

Les trois niveaux d'annotation sont représentés ci-après dans les tableaux 7.1 pour les noms d'entreprise et 7.2 pour les noms de pays (les catégories d'annotation seront détaillés dans la section 7.2.2). Les participants peuvent choisir d'évaluer leurs systèmes pour l'une ou l'autre des classes sémantiques ou pour chacune des deux, et en fonction d'un ou plusieurs niveaux d'annotation (6 soumissions autorisées au total). Ils peuvent également choisir de faire une évaluation partielle, se concentrant par exemple sur des noms dans telle ou telle position syntaxique.

| *coarse* | *medium* | *fine* |
|---|---|---|
| `literal` | `literal` | `literal` |
| `non-literal` | `mixed` | `mixed` |
| | `metonymic` | `othermet` |
| | | `object-for-name` |
| | | `object-for-representation` |
| | | `organisation-for-members` |
| | | `organisation-for-event` |
| | | `organisation-for-product` |
| | | `organisation-for-facility` |
| | | `organisation-for-index` |

TAB. 7.1 – Niveaux de granularité et catégories d'annotation pour la classe ORGANISATION.

| *coarse* | *medium* | *fine* |
|---|---|---|
| literal | literal | literal |
| non-literal | mixed | mixed |
| | metonymic | othermet |
| | | object-for-name |
| | | object-for-representation |
| | | place-for-people |
| | | place-for-event |
| | | place-for-product |

TAB. 7.2 – Niveaux de granularité et catégories d'annotation pour la classe LOCATION.

#### 7.2.1.2 Corpus, processus d'annotation et déroulement de la campagne

Les participants disposent d'un certain nombre de données pour participer à la tâche. Tout d'abord les corpus : un corpus d'entraînement puis un corpus de test ont été mis à disposition. Ces deux corpus, extraits tous deux du *British National Corpus*, se présentent chacun sous la forme de deux sous-corpus, un pour les noms de pays et un autre pour les noms d'entreprise. Il s'agit plus exactement d'une collection d'extraits du BNC, chaque extrait se composant d'environ quatre phrases et contenant une occurrence pré-selectionnée de nom de pays ou d'entreprise à annoter. Cette occurrence est le plus souvent présente dans la troisième phrase de l'extrait, avec deux phrases de contexte avant et une après ; d'autres configurations sont néanmoins possibles, avec cette fois-ci moins de contexte. Codés en *XML*, les extraits se présentent à la suite les uns des autres, avec comme informations un numéro identifiant (*sample id*) et le titre du fichier BNC duquel ils sont extraits. Dans le texte principal, contenu entre deux balises `<par>`, figure l'occurrence à annoter, entre les balises `<annot>`, balises contenant elles-mêmes un élément spécifiant la classe sémantique (`<org>` ou `<location>`), élément doté d'un attribut `reading` prenant comme valeur l'interprétation de l'occurrence, à savoir `literal`, `metonymic` ou `mixed`, ou encore la valeur du patron métonymique. Voici un exemple d'extrait pour mieux se représenter les choses (cf. figure 7.1).

Les corpus d'entraînement comportent l'annotation de l'interprétation de l'unité visée, les corpus de test ne comportent quant à eux que l'indication de l'unité à annoter, sans l'annotation bien sûr. Les corpus d'entraînement et de test pour les noms de pays comportent respectivement 925 et 908 extraits, ceux pour les noms d'entreprises 1090 et 842.

Outre les corpus, les participants disposent également d'informations grammaticales, avec des versions des corpus comportant l'annotation du BNC[1], et

[1]Segmentation et parties du discours.

```
 <sample id="samp8">
<bnc :title> Keesings Contemporary Archives. August 1990
</bnc :title>
<par>
Employees of the Peugeot motor company, France's largest pri-
vate company, staged a two-month strike which ended on Oct. 23,
1989, to demand higher wages and better conditions. The dead-
lock was broken when a government-appointed conciliator per-
suaded the Peugeot chairman to agree to talks as soon as a 19-
day sit-in at the Mulhouse plant was ended. <annot><org rea-
ding="metonymic" metotype="organisation-for-members" notes="OFF">
Peugeot </org></annot> said that it had lost production of 49,000
cars as a result of the strikes. Public-sector strikes
</par>
</sample>
```

FIG. 7.1 – Exemple de formulaire d'extraction d'information pour des actes terroristes (MUC-3).

d'autres comportant l'annotation des relations grammaticales impliquant les unités à annoter (annotation réalisée manuellement par les organisatrices de l'évaluation). L'utilisation de ces fichiers est optionnelle. Enfin, est fourni un script d'évaluation, permettant aux participants de tester leurs systèmes pendant la phase d'entraînement.

Pour ce qui est du déroulement de l'évaluation, les participants ont disposé d'une période d'entraînement de 2 mois, puis d'une période de deux semaines entre le téléchargement des corpus de test et l'envoi des résultats. Chacun a ensuite rédigé un article décrivant son système, pour une présentation au *Workshop* dédié à cette tâche lors de la conférence ACL à Prague en Juin 2007. À la suite de cette présentation de la tâche, il importe d'exposer les catégories d'annotation.

### 7.2.2 Les catégories d'annotation

Nous présentons ici les différentes catégories d'annotation de cas métonymiques proposées par les organisatrices et issues de leurs travaux antérieurs auxquelles devaient se conformer les participants. Il s'agit plus précisément des patrons métonymiques, c'est-à-dire des catégories intervenant au niveau le plus fin de l'annotation[1]. Nous décrirons tout d'abord les patrons métonymiques pour la classe ORGANISATION puis ceux pour la classe LOCATION. Seules les catégories

[1]Les catégories proposées ainsi que les explications les accompagnant nous ont parfois posé difficulté, ceci montrant une nouvelle fois combien il est difficile d'atteindre un consensus autour de standards pour une évalutation.

les plus fréquentes et les plus "marquantes" sont présentées, de sorte à assurer la compréhension de la tâche. Toutefois, cette présentation restant relativement succincte, nous renvoyons aux divers articles de K. Markert et M. Nissim ainsi qu'à l'annexe C pour plus de détails.

#### 7.2.2.1 Catégories d'annotation pour la classe ORGANISATION

Nous présentons successivement les annotations `literal`, `metonymic` et `mixed` avec, pour la deuxième annotation, des précisions sur quelques patrons, puis faisons état de leur statistiques.

1. Annotation `literal`

   Pour les noms d'entreprise, la catégorie d'annotation `literal` correspond aux cas où un nom d'organisation réfère à l'entreprise en tant qu'entité légale. Cette annotation couvre les cas, comme dans les exemples ci-dessous, de description de la structure d'une entreprise, d'association ou d'acquisition entre entreprises ou encore les cas de relation entre l'entreprise et les services ou produits qu'elle offre.

   - NATO countries.
   - Sun acquired that part of Eastman-Kodak Co's Unix Subsidiary.
   - Intel's Indeo video compression hardware.

2. Annotation `metonymic`

   L'annotation de type `metonymic` comprend des patrons métonymiques comme des catégories dédiées aux métonymies « créatives ».

   – `organisation-for-members`
   Cette catégorie est utilisée lorsqu'un nom d'entreprise est utilisé pour désigner ses employés. Par exemple, lorsqu'un porte-parole parle au nom de l'entreprise, ou que les employés participent à une action, comme dans les exemples suivants.

   - Last February, IBM announced [...].
   - It's customary to go to work in black and white suits. [...] Woolworths were them.

   – `organisation-for-product`
   Ce patron est à adopter lorsqu'un nom d'entreprise est utilisé pour référer aux produits qu'elle produit.

A red light was hung on the Ford's tail-gate.

– `object-for-name`
Ce patron est, comme le suivant, indépendant de la classe sémantique. Pour les entreprises, il convient de l'utiliser pour les mentions de noms d'entreprise en usage autonyme, c'est-à-dire évoquant des caractéristiques de leurs propres noms, comme dans l'exemple ci-dessous.

Chevrolet is feminine because of its sound (it's a longer word than Ford, has an open vowel at the end, connotes Frenchness).

– `orthermet`
Enfin, ce dernier patron permet de couvrir les cas pour lesquels aucune des annotations précédentes n'a pu être utilisée. Dans l'exemple suivant, *Barclays Bank* renvoie à un compte en banque.

funds [. . . ] had been paid into Barclays Bank.

3. Annotation `mixed`
L'annotation `mixed` est similaire au zeugma ; elle correspond aux cas où un nom d'entreprise convoque simultanément deux interprétations métonymiques. Dans l'exemple suivant, il est question de l'indice boursier (`organisation-for-index`) de l'entreprise et de ses employés (`organisation-for-members`).

Barclays slipped 4p to 351p after confirming 3,000 more job losses.

Les catégories d'annotation pour la classe ORGANISATION sont relativement nombreuses (prise en compte de tous les patrons, cf. annexe C). Du dire des organisatrices, l'accord inter-annotateurs est globalement bon pour l'ensemble des patrons, à l'exception de `mixed`, lequel semble difficile à distinguer à coup sûr et de manière objective. Elles reconnaissent par ailleurs la possible difficulté de distinguer l'interprétation `literal` de celle `organisation-for-members`, dans des cas notamment où le prédicat semble pouvoir s'appliquer à l'entreprise en entier ou bien à quelques uns de ses membres. La répartition de ces interprétations dans le corpus d'entraînement figure dans le tableau 7.3.

Le cas le plus fréquent, avec 63.3 %, est celui où les noms d'entreprise sont utilisés suivant leur sens premier (annotation `literal`). Cette proportion correspond à la « baseline », c'est à dire au taux de réussite minimum atteignable grâce

| Reading | N | % |
|---|---|---|
| `literal` | 690 | 63.3 |
| `organisation-for-members` | 220 | 20.2 |
| `organisation-for-event` | 2 | 0.2 |
| `organisation-for-product` | 74 | 6.8 |
| `organisation-for-facility` | 15 | 1.4 |
| `organisation-for-index` | 7 | 0.6 |
| `object-for-name` | 8 | 0.7 |
| `object-for-representation` | 1 | 0.1 |
| `othermet` | 14 | 1.3 |
| `mixed` | 59 | 5.4 |
| total | 1090 | 100.0 |

TAB. 7.3 – Distribution des cas de métonymies pour la classe ORGANISATION.

à l'assignation de l'annotation `literal` à tous les noms d'entreprise à annoter. L'autre cas le plus fréquent est `organisation-for-members` (20.2 %), et derrière lui les annotations `organisation-for-product` et `mixed`. Les autres patrons sont plus rares, mais tout de même représentés dans le corpus. Considérons à présent les catégories pour la classe LOCATION.

#### 7.2.2.2 Catégories d'annotation pour la classe LOCATION

1. Annotation `literal`

   Pour les noms de lieux, l'interprétation litérale est valable lorsqu'il s'agit de l'entité géographique ou de l'entité politique.

   - The coral coast of Papua New Guinea.
   - Britain's current account deficit.

2. Annotation `metonymic`
   - `place-for-people`

     Cette annotation est à utiliser lorsqu'un nom de pays renvoie aux personnes ou aux organisations qui lui sont associées. On peut penser aux cas où il est question du gouvernement, d'une équipe sportive ou bien de la population du pays. Ces trois cas sont illustrés par les exemples ci-dessous ; le dernier exemple illustre le cas d'une sous-spécification du référent cible, qui reste tout de même de type « people ».

     - America did once try to ban alcohol.
     - England lost in the semi-final.

> - The notion that the incarnation was to fulfil the promise to Israel and to reconcile the world with God.
> - The G-24 group expressed readiness to provide Albania with food aid.

– `place-for-event`
Annotation à utiliser lorsque le nom de pays renvoie à un événement ayant eu lieu dans ce pays (cf. exemple donné dans la section 7.2.1.1, p.219).

– `place-for-product`
Il s'agit d'une catégorie pour les cas de renvois à un produit fabriqué par le pays (ou le lieu).

> A smooth Bordeaux that was gutsy enough to cope with our food.

– `object-for-representation`
Un nom peut renvoyer à une représentation, telle une photo, une peinture, un symbole, du référent réel évoqué par ce nom. Pour les noms de pays, on peut penser à la mention pouvant exister sur une carte géographique, comme dans l'exemple suivant, dans le contexte de quelqu'un pointant du doigt un point sur une carte :

> This is Malta.

Les fréquence d'occurrences pour l'ensemble des patrons métonymiques de cette classe (cf. annexe C) sont les suivantes :

| Reading | N | % |
|---|---|---|
| `literal` | 737 | 79.7 |
| `place-for-people` | 161 | 17.4 |
| `place-for-event` | 3 | 0.3 |
| `place-for-product` | 0 | 0.0 |
| `object-for-name` | 0 | 0.0 |
| `object-for-representation` | 0 | 0.0 |
| `othermet` | 9 | 1.0 |
| `mixed` | 15 | 1.6 |
| total | 925 | 100.0 |

TAB. 7.4 – Distribution des cas de métonymies pour la classe LOCATION.

Avec seulement trois patrons spécifiques, cette classe des noms de pays semble moins diverse « métonymiquement parlant » que celle des noms d'entreprise.

On observe de nouveau une forte proportion des cas litéraux, avec une *baseline* plus élévée que pour les noms d'entreprise, de près de 80 %. Le cinquième des cas restants est représenté pour une grande partie par le patron `place-for-people`, celui-ci laissant une petite part pour `place-for-event`, `othermet` et `mixed`. Les autres cas ne sont pas représentés dans le corpus d'entraînement. Les organisatrices conviennent une fois de plus de confusions possibles entre les catégories `literal` et `place-for-people`.

En fonction de l'ensemble de ces catégories, nous avons tenté de mettre un œuvre un système automatique de résolution de métonymie.

## 7.3 Un système de résolution de métonymie pour les entités nommées

Cette section décrira tout d'abord le système élaboré puis présentera et discutera les résultats de l'évaluation.

### 7.3.1 Description du système

Notre participation, avec C. Brun et G. Jacquet, à la tâche de résolution de métonymie pour les noms de pays et d'entreprises a consisté en l'élaboration d'un système automatique hybride reposant sur la combinaison d'un composant symbolique et d'un composant distributionnel. Nous présenterons successivement ces deux composants.

#### 7.3.1.1 Composant symbolique

La méthode proposée fait tout d'abord intervenir un composant symbolique, c'est-à-dire un système d'analyse syntaxique robuste à base de règles. Nous évoquons rapidement l'analyseur syntaxique utilisé, XIP, avant de détailler son adaptation pour la tâche à accomplir.

**Analyse syntaxique robuste avec XIP** L'élément fondamental sur lequel repose notre approche est l'analyseur syntaxique robuste XIP [Aït-Mokhtar et Chanod, 1997, Aït-Mokhtar *et al.*, 2002]. *Xerox Incremental Parser* prend en entrée du texte tout venant et produit en sortie de façon robuste une analyse syntaxique profonde sous forme de relations de dépendances. Nous renvoyons à l'annexe D pour plus de précisions sur la mise en œuvre de l'ensemble de ces traitements.

**Adaptation à la tâche de résolution de métonymie** L'analyseur évoqué ci-dessus contient déjà un module de reconnaissance d'entités nommées (cf. chapitre 6), mais ce dernier ne permet pas le traitement de la métonymie. En effet, si le modèle d'annotation sur lequel il repose contient la catégorie « double » LOCORG permettant d'annoter des noms de lieux (LOC) pouvant être utilisés pour référer à des entités fonctionnant comme des organisations (ORG), telles les gouvernements ou populations, il ne prévoit pas de les différencier. C'est précisément ce qu'il importe de faire pour la tâche SemEval, avec une distinction entre les interprétations `literal` et `place-for-people`, et bien d'autres patrons encore.

**Etude de corpus** Pour mener à bien cette adaptation, nous avons commencé par effectuer une étude des corpus d'entraînement à l'aide de XIP, en tenant compte bien entendu des directives d'annotation. Pour les noms de pays comme pour les noms d'entreprises, notre attention s'est focalisée sur, d'une part, les types de relations grammaticales dans lesquelles étaient impliquées les unités à étudier et, d'autre part, sur les informations lexicales ou sémantiques attachées aux arguments de ces relations. Pour chaque patron métonymique, nous avons donc analysé l'ensemble de ses occurrences (attachées à une unité lexicale) dans le corpus d'entraînement pour dégager des configurations grammaticales et lexicales jouant le rôle d' « amorces » d'interprétations métonymiques.

Compte tenu des statistiques d'occurrences des patrons pour chacune des classes (cf. Tableaux 7.3 et 7.4), il importait de se concentrer en priorité sur les lectures `organisation-for-members` (20.2 %), `organisation-for-product` (6.8 %) et `place-for-people` (17.4 %). Bien que l'attribution par défaut de la catégorie `literal` à toutes les unités à prendre en compte permette d'annoter avec une précision parfaite les cas d'interprétations littérales, nous avons choisi d'étudier et de traiter ce patron comme les autres afin d'être en mesure de faire la différence entre des cas litéraux « par défaut » et des cas litéraux « certains ». Au fur et à mesure de la réalisation de ce composant symbolique, sont en effet apparues des configurations lexico-grammaticales concurrentes entre le patron `literal` et d'autres patrons, conflit qu'il était parfois permis de résoudre en fonction de l'analyse par défaut ou non de l'interprétation littérale. Cette précaution permettait en outre de ne pas « passer » inutilement une occurrence analysée comme littérale « certaine » à d'autres règles dédiées à d'autres patrons, s'appliquant alors aux seules occurences littérales « par défaut » (nous reviendrons par la suite sur l'organisation des traitements).

Commençons par étudier le comportement des noms d'entreprise dans le corpus d'entraînement, avec ses interprétations `literal`, `org-for-members` et `org-for-product`.

Selon l'analyse XIP, sur 690 occurrences « `literal` » de noms d'entreprise dans le corpus d'entraînement, une majorité de ces noms (28.5 %) sont des têtes de syntagmes prépositionnels (*Mr Nigel Whittaker of Kingfisher*), une grande partie (24 %) se trouvent en position de prémodifieur (*IBM engineers*), une part importante (environ 18.1 %) en position sujet (*IBM produces ...*) ou génitif (*Hungary's case*), et une petite part (5.2 %) en position objet. Il importe donc de se concentrer sur ces cas les plus importants et d'étudier leurs réalisations dans le corpus.

Pour les occurrences `literal` des noms d'entreprise en position sujet, on peut observer qu'elles sont régies par des verbes de type *launch, provide, license, pay, challenge*, cas pour lesquels on peut établir de manière certaine l'annotation `literal`. D'autres verbes sont en revanche moins clairs, tels que *let, cost, leave, bring*, pour lesquels on ne peut assurer de manière certaine qu'ils régissent un nom d'entreprise référant à l'entreprise en tant qu'entité légale. Certains verbes sont également « inutilisables » pour l'assignation certaine de la catégorie `literal` car en commun avec d'autres patrons métonymiques et faisant état d'un sémantisme pas assez marqué relativement à l'interprétation recherchée (il s'agit de verbes tels que *come, take, keep, start, make, become*).

Pour les occurrences en position de prémodifieur, compte tenu de la moindre importance de cette configuration pour les autres patrons métonymiques de cette classe, aucune information lexicale n'a été recueillie pour cette interprétation `literal` dans cette configuration (cela a été fait pour les autres patrons). Pour ce qui est des occurrences en tête de groupe prépositionnel, ce cas est normalement difficile à traiter en raison de deux types d'informations à prendre en compte, la préposition elle-même d'une part, et le terme qui précède le groupe prépositionnel d'autre part. En raison de nouveau de la plus faible proportion de cette configuration pour les autres patrons, nous avons réservé son traitement lexical pour ces autres cas.

Si l'on s'intéresse au patron `org-for-members`, une étude similaire est faisable. Ce dernier se rencontre majoritairement en position sujet (188 cas sur 220), de temps en temps comme tête de groupe prépositionnel, plus rarement en objet ou prémodifieur. En position sujet, ce patron est régi par des verbes au sémantisme marqué, tels que des verbes de communication (*declare, state, tell, respond*), des verbes d'action juridique (*subscribe, sign, ratify, consent*) et des verbes de cognition (*agree, choose, establish, understand, decide*), autant de lexique utilisable pour aiguiller une interprétation.

Pour les `org-for-product`, les indices les plus discriminants sont la détermination (*a Ford*) et la qualification (*a large BMW*). L'ensemble des déterminants a donc été étudié afin d'établir un ordre d'importance (possessif, indéfini, défini), tout

comme les adjectifs. De même que précédemment, ce patron a été étudié en position sujet (*appear, drive, come, cross*) et objet (*drive, start, switch*).

Signalons encore les informations ayant pu être recueillies pour le patron `object-for-name`, relativement peu représenté (0.7 % d'occurrences). L'acronymisation (*by selling it to British Aerospace (BAE))*, la modification sur des mots relevant d'un champ sémantique particulier (comme les noms de vêtements : *under a Coca-Cola hat)* ou encore les verbes de signification (Ford connotes « Frenchness ») nous ont permis d'extrapoler certaines configuations contextuelles propres à ce glissement métonymique.

Pour ce qui est des noms de pays, une étude analogue fut mise en œuvre, se concentrant principalement sur le patron `place-for-people`, de loin le plus fréquent. Sans entrer dans les détails de cette étude (similaire à la précédente), indiquons seulement la prise en considération de verbes tels que *import, provide, refund, repay* régissant un nom de pays en position sujet.

Ces éléments rendent compte sensiblement des informations collectées lors de l'étude de corpus, qui fut naturellement plus fructueuse pour les patrons les plus fréquents ; d'autres indices non évoqués ici furent également exploités, parmi lesquels la présence de l'unité métonymique en tête de phrase, l'indication du nombre, etc.

**Module de traitement de métonymie** Au terme de cette étude de corpus, il fut possible de prendre en compte l'ensemble des indices récoltés et d'élaborer un module de résolution de métonymie dans XIP. Intervenant à la fin de la chaîne de traitements de l'analyseur, cette adaptation consiste en l'écriture de règles traduisant les configurations discriminantes relevées pour tel ou tel patron d'une part, et à l'ajout de lexique d'autre part. À titre d'exemple, prenons l'hypothèse suivante : « *Si un nom de pays est sujet d'un verbe renvoyant à une action économique, alors le patron* `loc-for-people` *doit s'appliquer* » ; cette hypothèse se retrouve dans XIP sous la forme de la règle suivante :

```
if(LOCATION(#1) & SUBJ-N(#2[v_econ],#1))
   LOC-FOR-PEOPLE(#1)
```

Règle qui se lit de la manière suivante : si le parseur a détecté un nom de pays (#1) qui est le sujet d'un verbe (#2) portant le trait « `v_econ` », alors il créé un relation unaire LOC-FOR-PEOPLE pour le nom de pays.

En sus des indices récoltés lors de l'étude de corpus, nous avons exploité des informations lexicales déjà codées dans XIP, comme par exemple les traits attachés aux verbes de communication (*say, deny, comment)* et les catégories relevant du

cadre « experiencer » de Framenet, à savoir des verbes tels que *feel, sense, see*, les noms *despair, compassion, adoration* ou les adjectifs *sympathetic, sad,* etc. En effet, eu égard au fait que ce cadre « experiencer » renvoie à des personnes ou à des groupes de personnes, dès lors qu'un nom de pays ou d'entreprise a ce rôle, il peut être annoté en tant que `loc-for-people` ou `organisation-for-members`.

Pour chaque classe, un ordre de passage des règles fut déterminé, en fonction des fréquences de chaque patron métonymique. Faute d'indices forts de reconnaissance, les patrons `mixed` et `othermet` n'ont été traités pour aucune des classes. Environ 70 règles ont été développées pour le traitement de la métonymie des noms de pays et 100 pour les noms d'entreprises. Ci-dessous quelques exemples de sortie de l'analyseur avec le nouveau développement :

*It was the largest Fiat anyone had ever seen.*
ORG_FOR_PRODUCT(`Fiat`)
MOD_PRE(`seen,ever`)
SUBJ-N_PRE(`was,It`)
ATTRIB(`It,Fiat`)
EXPERIENCER_PRE(`seen,everyone`)
SUBJATTR(`It,Fiat`)
QUALIF(`Fiat,largest`)

Dans cet exemple, la présence de la relation QUALIF entre le nom d'entreprise *Fiat* et l'adjectif *largest*, renforcée par la présence d'un article défini, conduit à l'annotation en tant que ORG_FOR_PRODUCT.

*Malta endorsed a serie of proposals..*
LOC-FOR-PEOPLE(`Malta`)
PREP_OF(`serie,proposals`)
SUBJ-N_PRE(`endorsed,Malta`)
OBJ-N(`endorsed,serie`)
DETD(`serie,a`)
LOCORG(`Malta`)

Pour ce cas, c'est la relation sujet entre le nom de pays et le verbe *to endorse* qui oriente l'analyse vers le patron LOC-FOR-PEOPLE, l'action d'*appuyer, donner son aval à* renvoyant à des personnes associées à Malte.

Enfin, dans le but d'améliorer la couverture des lexiques, nous avons entrepris d'exploiter le corpus BNC dans son entier pour collecter davantage d'information lexicale. Ce dernier a été analysé à l'aide de XIP afin d'extraire avec leurs

fréquences les dépendances mentionnées ci-avant (sujet, objet, modifieur) impliquant un nom de pays ou un nom d'entreprise (module de reconnaissance d'entités nommées). L'analyse de ces données consista ensuite à sélectionner, au sein des relations les plus fréquentes, les contextes propres à déclencher une lecture métonymique et à les classer en conséquence dans les patrons correspondants. Par exemple, la relation PREP_OF(`invasion`,COUNTRY) apparaît 469 fois dans le BNC et la relation SUBJ(COMPANY, `decides`) 420 fois. Une fois classés, les mots apparaissant dans ces contextes furent intégrés aux lexiques.

Ce composant symbolique constitue le premier volet de notre méthode de résolution de métonymie, il est complété par un composant distributionnel, qu'il convient à présent de détailler.

#### 7.3.1.2 Composant distributionnel

Intervient en « deuxième passe » du système de résolution de métonymie un composant distributionnel. L'idée principale de cette combinaison est d'exploiter deux méthodes relevant d'approches différentes mais complémentaires, autrement dit de pallier les manquements de l'analyse symbolique, focalisée sur des données précises nécessitant une étude minutieuse, par une analyse distributionnelle, apte à récolter des informations à grande échelle sur d'importantes données textuelles. Le principe est donc, lorsqu'une unité n'a pu être traitée par le composant symbolique, d'essayer de trouver son annotation en exploitant les informations présentes à propos de cette unité ou d'une unité de même type dans un grand corpus. Il importera de présenter rapidement l'analyse distributionnelle avant de détailler sa mise en œuvre pour la résolution de métonymie.

**L'analyse distributionnelle** La notion d'analyse distributionnelle a été introduite au milieu du siècle dernier par l'américain Z. S. Harris. En linguistique de corpus, cette méthode est aujourd'hui largement exploitée, notamment dans les travaux de terminologie, de structuration de terminologie et de construction d'ontologies. L'hypothèse est la suivante : il serait possible, à partir de régularités syntaxiques observées pour un ensemble de mots, de déduire des propriétés sémantiques pour ces mots. Concrètement, il s'agit d'étudier la distribution des mots, c'est à dire les contextes lexico-syntaxiques dans lesquels ils apparaissent, pour ensuite tenter de dégager des parentés sémantiques et éventuellement construire des classes sémantiques. Prenons par exemple l'unité lexicale *pomme*. Pour savoir de quels mots il est possible de la rapprocher, on regarde les contextes syntaxiques dans lesquels elle apparaît, puis les mots occurrant avec ces mêmes contextes : le mot *pomme* se trouve dans les contextes « objet de *manger* » et « objet de

*déguster* », tout comme les mots *poire, kiwi, tarte, gâteau*. L'ensemble de ces mots n'apparaît en revanche pas du tout dans des contextes tels que « objet de *boire* » ou « objet de *poursuivre* ». Sur la base de ces comportements syntaxiques communs, il est donc possible de rapprocher les unités *pomme, poire, kiwi, tarte, gâteau*.

Durant la dernière décennie, nombreux sont les travaux en linguistique de corpus ayant utilisé cette méthode. G. Greffenstette a travaillé sur la construction automatique de classes de noms se retrouvant régulièrement arguments des mêmes verbes, en utilisant l'analyseur syntaxique Sextant. D'autres travaux ont par la suite suivi cette perspective, citons ceux de D. Faure autour du système *Asium* [Faure et Nedellec, 1999] pour l'apprentissage de cadres de sous-catégorisation de verbes, ceux de H. Assadi [Assadi, 1998], ou encore les recherches de B. Habert et A. Nazarenko (outil *Zellig*, [Bouaud *et al.*, 2000]) et, plus récemment encore, celles de D. Bourigault (outil *Upery*, [Bourigault, 2002]). Les regroupements effectués sont de plus en plus pertinents à mesure, d'une part, de l'amélioration des performances des analyseurs utilisés pour extraire les dépendances syntaxiques et d'autre part, de la plus grande disponibilité de corpus de tailles importantes. Cependant, rien n'est encore réellement possible au-delà d'une structuration du lexique utilisé dans le corpus, la mise au jour de véritables concepts ne se faisant que manuellement.

**Méthode distributionnelle pour la résolution de métonymie** Le principe est le suivant : si, pour une entité donnée à annoter le composant symbolique n'a pu établir la catégorie faute de contexte (ou configuration lexico-grammaticale) connu, alors il peut être utile de tenter de trouver des contextes sémantiquement proches pour lesquels on possède une annotation. L'objectif est donc de rapprocher des contextes et d'exploiter les résultats du composant symbolique. Cette méthode comporte deux processus. Il s'agit d'une part de construire un espace distributionnel pour être en mesure de rapprocher des contextes en fonction d'une entité donnée et, d'autre part, de capitaliser l'information du composant symbolique sous la forme d'une sorte de « base de données » de contextes avec annotation. Nous détaillons ces deux processus successivement.

**Construction de l'espace distributionnel et rapprochement des contextes** Ce premier processus comporte 5 étapes. L'objectif est d'être en mesure, pour une entité donnée apparaissant dans un contexte donné, d'établir une liste de contextes les plus proches. Le point de départ est le corpus BNC dans son entier (100 millions de mots), duquel on été extraits les corpus d'entraînement et de test de la tâche de résolution de métonymie.

La première étape correspond à l'analyse syntaxique de ce corpus (à l'aide de l'analyseur XIP) afin d'en extraire des dépendances syntaxiques. C'est à partir de ces dépendances que sont construits les contextes et les unités lexicales. Prenons un exemple, avec la proposition *provide Albania with food aid*. Les dépendances extraites par XIP sont les suivantes :

IND-OBJ-N(VERB :`provide`,NOUN :`Albania`)
PREP_WITH(VERB :`provide`,NOUN :`aid`)
PREP_WITH(VERB :`provide`,NOUN :`food aid`)

Où l'on peut voir que les arguments des dépendances sont de simples unités lexicales (*aid*) ou des syntagmes (*food aid*).

La deuxième étape de ce processus est la construction d'un espace distributionnel à partir de ces dépendances. Cet espace distributionnel est l'inverse de celui habituellement construit en analyse distributionnelle : puisque l'objectif est de rapprocher des contextes, chaque point de l'espace est un contexte syntaxique et chaque dimension est une unité lexicale. Cette inversion constitue une première différence avec l'approche de [Jacquet et Venant, 2005]. Comment est construit cet espace distributionnel ? Chaque dépendance obtenue lors de l'étape précédente permet de construire plusieurs contextes simples et/ou composés. Un contexte syntaxique comporte une relation et une unité lexicale. Pour la phrase analysée ci-dessus, les unités lexicales sont :

VERB :`provide`
NOUN :`Albania`
NOUN :`aid`
NOUN :`food aid`

Les contextes simples sont les suivants (1. pour recteur et 2. pour régi) :

1.VERB : `provide`.IND-OBJ-N
1.VERB : `provide`.PREP_WITH
2.NOUN : `Albania`.IND-OBJ-N
2.NOUN : `aid`.PREP_WITH
2.NOUN : `food aid`.PREP_WITH

La méthode prévoit également de construire des contextes composés (autre différence par rapport à [Jacquet et Venant, 2005]), soit des contextes combinant plusieurs contextes, dont le premier comporte nécessairement un syntagme verbal et le second a pour recteur ce même syntagme. Voici les contextes composés pour la phrase analysée :

1.VERB : `provide`.IND-OBJ-N + 2.NOUN : `aid`.PREP_WITH
1.VERB : `provide`.IND-OBJ-N + 2.NOUN : `food aid`.PREP_WITH
1.VERB : `provide`.PREP_WITH + 2.NOUN : `Albania`.IND-OBJ-N

Une heuristique permet de filtrer ces données en fonction de leur productivité : chaque unité lexicale doit être présente au moins 100 fois dans le corpus, tout comme les contextes (y compris les contextes composés). Avec le corpus BNC de 100 millions de mots, on obtient au final 60 849 unités lexicales et 140 634 contextes. Il est donc possible de construire un espace distributionnel comportant 140 634 points (les contextes) et 60 849 dimensions (les unités lexicales). Cet espace est le matériau de base à partir duquel les autres traitements viennent s'effectuer.

À partir de la troisième étape intervient la prise en compte d'une unité lexicale précise, pour laquelle le composant symbolique n'a pu trouver d'annotation. En fonction de cette unité et de son contexte d'apparition (soit telle qu'elle apparaît dans un extrait de SemEval), il s'agit tout d'abord de construire un sous-espace, de la manière suivante :

Pour un couple donné formé d'un contexte $i$ et d'une unité lexicale $j$,

Le sous-espace des contextes équivaut à la liste des contextes occurrant avec l'unité lexicale $j$. S'il y a plus de $k$ contextes, alors on ne garde que ces $k$ contextes les plus fréquents ;

Le sous-espace des dimensions équivaut à la liste des unités lexicales occurrant avec au moins un des contextes du sous-espace des contextes. S'il y a plus de $n$ unités, alors on ne garde que ces $n$ unités les plus fréquentes.

La quatrième étape consiste à réduire les dimensions de ce sous-espace, à l'aide d'une analyse factorielle des correspondances (AFC) [Lebart et Salem, 1994, chap. 3]. Afin d'expliquer cette étape, il importe de donner des précisions sur l'AFC.

Il existe deux types de méthodes en statistique multidimensionnelle : les méthodes factorielles, qui s'attachent à éliminer la redondance des données originales en essayant de résumer les variations à l'aide d'un nombre plus faible de variables (les facteurs) qui sont une combinaison des variables originales, et les méthodes de classification (ou clusterisation), qui visent à regrouper les points en sous ensembles. Concernant les méthodes factorielles, trois techniques fondamentales peuvent être considérées : l'Analyse en Composantes Principales (plusieurs variables quantitatives), l'Analyse des Correspondances (2 variables quantitatives dans un tableau de contingence) et l'Analyse des Correspondances Multiples (plus de deux variables quantitatives). Il s'agit de techniques de représentations de don-

nées. Pour ce qui est de l'ACP (concue par K. Pearson en 1901), cette méthode permet de synthétiser un ensemble de données en identifiant la redondance dans celles-ci. L'objectif est, étant donné l'abondance et la diversité des informations (mesures) à traiter, de condenser ces données en de nouveaux groupements synthétisant au mieux l'information. Concrètement, si on a un nuage de points représenté dans un espace à $n$ dimensions, on ne voit pas grand chose. Il importe donc de définir de nouvelles dimensions (les composantes principales) et de projeter les points initiaux dans ce sous-espace de dimension raisonnable, de sorte que cette projection retienne le plus d'information possible. L'AFC est une méthode plus récente développée par J.-P. Benzécri relevant de la même logique que l'ACP : elle permet de rechercher la meilleure représentation simultanée de deux ensembles constituant les lignes et les colonnes d'un tableau de contingence, ces deux ensembles jouant un rôle symétrique. La différence entre les deux méthodes tient au fait que l'une (ACP) utilise la distance euclidienne classique tandis que l'autre est dotée de la métrique du Chi2.

La réduction des dimensions du sous-espace obtenu à la fin de la quatrième étape fait donc intervenir la distance du Chi2. Il s'agit d'une méthode statistique de calcul de distance entre individus qui a pour principal intérêt de mettre en évidence les individus dont le comportement est le plus divergent par rapport à la moyenne. Le résultat de cette étape est un espace réduit, composé de contextes et de $n$ dimensions, dont nous avons retenu les 10 premières composantes.

Enfin, la cinquième étape consiste à rapprocher les contextes restants, c'est-à-dire ayant passé le filtre des deux étapes précédentes. Ce rapprochement est calculé à l'aide de la distance euclidienne. On obtient ainsi, au final, une liste des contextes proches de celui de l'unité considérée. Pour l'unité *Albania* dans le contexte *provide*, les contextes les plus proches sont présentés dans le tableau 7.5.

| Contexte | Distance |
|---|---|
| VERB :`provide`.IND-OBJ-N | 0.00 |
| VERB :`allow`.OBJ-N | 0.76 |
| VERB :`include`.OBJ-N | 0.96 |
| ADJ :`new`.MOD_PRE | 1.02 |
| VERB :`be`.SUBJ-N | 1.43 |
| VERB :`supply`.SUBJ-N_PRE | 1.47 |
| VERB :`become`.SUBJ-N_PRE | 1.64 |
| VERB :`come`.SUBJ-N_PRE | 1.69 |
| VERB :`support`.SUBJ-N_PRE | 1.70 |
| ... | ... |

TAB. 7.5 – Extrait des contextes les plus proches pour l'unité *Albania*.

Pour pouvoir utiliser cette liste de contextes, encore faut-il savoir s'ils ont reçu

ou non une annotation par le composant symbolique. Intervient alors le second processus.

**Capitalisation de l'information du composant symbolique** Ce second processus a pour objectif de construire une sorte de base de données de contextes avec annotations. Pour ce faire, deux étapes sont nécessaires. Il importe tout d'abord de parser le corpus BNC avec XIP, augmenté cette fois-ci du module de résolution de métonymie. Cette analyse permet d'obtenir une série de dépendances syntaxiques mettant en jeu des noms de pays et des noms d'entreprise avec leurs annotation métonymique (application du module de résolution de métonymie sur les entités nommées reconnues par l'analyseur, indépendamment des données de SemEval). Le résultat de cette première étape est donc une série de contextes impliquant des unités lexicales avec leur annotation. La deuxième étape correspond à la sélection de contextes discriminants au regard des annotations. Par exemple, si le contexte VERB :`allow`.OBJ-N se retrouve pour 10% des cas avec une unité comportant l'annotation `literal` et pour 90% des autres avec une unité comportant l'annotation `loc-for-people`, alors le contexte est considéré comme discriminant vis-à-vis de cette dernière annotation. En revanche, si un contexte se trouve dans 50% des cas avec telle annotation et 50% des cas avec telle autre, alors il n'est pas conservé. À l'issue de cette sélection de contextes discriminants, on dispose alors d'un stock de contextes avec leurs annotations, pouvant être exploité en parallèle avec le processus précédent.

**Annotation d'une unité** Le premier processus permet de déterminer une liste de contextes plus ou moins proches du contexte dans lequel apparaît une unité lexicale non annotée. Le second permet de collecter un certain nombre de contextes avec leur annotation. Ces deux types de données peuvent dès lors être croisées pour permettre l'annotation d'une unité.

Reprenons la liste de contextes proches du contexte VERB :`provide`.OBJ-N pour la proposition *provide Albania with food aid*, avec l'indication de leur distance et de leur annotation correspondante dans la base de données de contextes (cf tableau 7.6). Pour pouvoir déterminer une annotation pour le contexte `provide`.OBJ-N, il faut regarder les annotations attribuées à ses contextes les plus proches et examiner si elles sont pertinentes ou non. Pour ce faire, on peut calculer un score pour chaque annotation, valorisant sa présence dans les contextes de « tête de liste » : le score d'une annotation donnée est égal à l'inverse de la somme des distances des contextes portant cette annotation, ce qui correspond à

la formule :

$$score_{annot_i} = \sum_{C_j} \frac{1}{d(C_j)} \tag{7.1}$$

où $annot_i$ correspond à l'annotation pour laquelle on calcule le score, et $C_j$ correspond aux contextes portant cette annotation.

| Contexte | Distance | Annotation |
|---|---|---|
| VERB :provide.IND-OBJ-N | 0.00 | |
| VERB :allow.OBJ-N | 0.76 | LOC-FOR-PEOPLE |
| VERB :include.OBJ-N | 0.96 | |
| ADJ :new.MOD_PRE | 1.02 | |
| VERB :be.SUBJ-N | 1.43 | |
| VERB :supply.SUBJ-N_PRE | 1.47 | LITERAL |
| VERB :become.SUBJ-N_PRE | 1.64 | |
| VERB :come.SUBJ-N_PRE | 1.69 | |
| VERB :support.SUBJ-N_PRE | 1.70 | LOC-FOR-PEOPLE |
| ... | ... | ... |

TAB. 7.6 – Contextes proches du contexte VERB :provide.OBJ-N.

Pour notre cas, les scores sont les suivants :

| Annotation | Score |
|---|---|
| LOC-FOR-PEOPLE | 3.11 |
| LITERAL | 1.23 |
| LOC-FOR-EVENT | 0.00 |
| ... | ... |

TAB. 7.7 – Scores des annotations pour les contextes proches du contexte VERB :provide.OBJ-N.

Relativement à ces scores, il est possible d'attribuer l'annotation LOC-FOR-PEOPLE à l'unité *Albania* dans la proposition *provide Albania with food aid*.

### 7.3.2 Evaluation et analyse des résultats

Les résultats obtenus avec notre système sur les corpus de test furent assez prometteurs car au-dessus de la baseline. Au milieu de quatre autres systèmes, XRCE-M est en effet parvenu à se placer en deuxième position pour la première tâche (noms de pays) à tous les niveaux de granularité, et en troisième position pour la seconde (noms d'entreprises). Les taux de précision[1] pour l'ensemble des

[1]Il n'est pas utile de présenter le rappel pour cette évaluation concernant les différents niveaux d'annotation (*coarse*, etc.) : les occurrences à annoter étant indiquées, il est systé-

participants sont présentés ci-après, pour les noms de lieux (tableau 7.8) et pour les noms d'entreprises (tableau 7.9).

| Tâche \ Systèmes | baseline | up13 | FUH | UTD-HLT-CG | XRCE-M | GYDER |
|---|---|---|---|---|---|---|
| LOCATION-coarse | 0,794 | 0,754 | 0,778 | 0,841 | **0,851** | 0,852 |
| LOCATION-medium | 0,794 | 0,750 | 0,772 | 0,840 | **0,848** | 0,848 |
| LOCATION-fine | 0,794 | 0,741 | 0,759 | 0,822 | **0,841** | 0,844 |

TAB. 7.8 – Taux de précision pour tous les systèmes pour la classe LOCATION.

| Tâche \ Systèmes | baseline | XRCE-M | UTD-HLT-CG | GYDER |
|---|---|---|---|---|
| ORGANISATION-coarse | 0,618 | **0,732** | 0,739 | 0,767 |
| ORGANISATION-medium | 0,618 | **0,711** | 0,711 | 0,733 |
| ORGANISATION-fine | 0,618 | **0,700** | 0,711 | 0,728 |

TAB. 7.9 – Taux de précision pour tous les systèmes pour la classe ORGANISATION.

Nous revenons tout d'abord sur les résultats de notre système avant de considérer l'ensemble de ceux obtenus pour la tâche, globalement et par système.

Satisfaisants dans l'ensemble, les résultats obtenus par XRCE-M sur les corpus de test sont néanmoins inférieurs à ceux obtenus sur les corpus d'entraînement, montrant par là la possibilité d'améliorations. Comme cela était prévisible, en raison des baselines et de la diversité des patrons métonymiques à prendre en compte pour chacune des classes, les résultats sont meilleurs pour les noms de lieux (0,841 %) que pour les noms d'entreprise (0,7 %). La différence accrue de résultats entre les divers niveaux de granularité pour les noms d'entreprise (perte de 30 centièmes entre *coarse* et *fine*) par rapport aux noms de pays (perte de 10 centièmes seulement) vient par ailleurs confirmer cette caractéristique de plus grande diversité, et donc difficulté. Nous donnons dans les tableaux 7.10 et 7.11 les résultats détaillés de notre système pour chaque patron dans chacune des classes.

Pour la classe LOCATION, on peut remarquer de relativement bonnes précisions pour les patrons `place-for-people`, `place-for-event` mais également `object-for-name`, peu présent dans le corpus d'entraînement mais relativement simple à schématiser à partir de l'information et des exemples disponibles dans les directives d'annotation. Les autres patrons, non couverts par notre méthode car trop

matiquement égal à 1. Au niveau plus précis de l'évaluation de la reconnaissance de patrons métonymiques (annotation *fine*), il est en revanche intéressant de considérer le rappel et d'observer la couverture des systèmes pour chaque patron. Ces résultats sont présentés dans les tableaux 7.10 et 7.11 pour notre système ; pour obtenir plus de détails concernant ce niveau d'évaluation pour les autres participants, nous invitons à consulter les articles suivants : [Poibeau, 2007, Leveling, 2007, Farkas *et al.*, 2007, Nicolae *et al.*, 2007].

| Scores / LOCATION-fine | Nb occurrences | Précision | Rappel | F-mesure |
|---|---|---|---|---|
| `literal` | 721 | 0,867 | 0,960 | 0,911 |
| `place-for-people` | 141 | 0,651 | 0,490 | 0,559 |
| `place-for-event` | 10 | 0,5 | 0,1 | 0,166 |
| `place-for-product` | 1 | - | 0 | 0 |
| `object-for-name` | 4 | 1 | 0,5 | 0,666 |
| `object-for-representation` | 0 | - | - | - |
| `othermet` | 11 | - | 0 | 0 |
| `mixed` | 20 | - | 0 | 0 |

TAB. 7.10 – XRCE-M : résultats détaillés pour la classe LOCATION.

peu nombreux, voire absents, dans le corpus d'entraînement et surtout difficiles à configurer (`mixed` et `othermet` entre autres) ont des résultats nuls.

| Scores / ORGANISATION-fine | Nb occurrences | Précision | Rappel | F-mesure |
|---|---|---|---|---|
| `literal` | 520 | 0,730 | 0,906 | 0,808 |
| `organisation-for-members` | 161 | 0,622 | 0,522 | 0,568 |
| `organisation-for-event` | 1 | - | 0 | 0 |
| `organisation-for-product` | 67 | 0,550 | 0,418 | 0,475 |
| `organisation-for-facility` | 16 | 0,5 | 0,125 | 0,2 |
| `organisation-for-index` | 3 | - | 0 | 0 |
| `object-for-name` | 6 | 1 | 0,666 | 0,8 |
| `othermet` | 8 | - | 0 | 0 |
| `mixed` | 60 | - | 0 | 0 |

TAB. 7.11 – XRCE-M : résultats détaillés pour la classe ORGANISATION.

Les résultats pour la classe ORGANISATION présentent sensiblement la même répartition, entre des patrons bien couverts (`organisation-for-members`, `organisation-for-product` et `object-for-name`) et d'autres non traités.

Les mauvaises annotations produites sont souvent dues à des erreurs de l'analyseur syntaxique utilisé en amont de notre système. Dans certains cas en effet, la qualité de l'analyse syntaxique fait défaut, comme dans la phrase suivante : "*Many galleries in the States, England and France declined the invitation*", où l'entité *France*, en raison d'une mauvaise analyse de la coordination, est analysée comme étant le sujet du verbe *declined*, et donc comme portant l'annotation `place-for-people`.

Malgré ces imperfections, XRCE-M est relativement bien placé au sein des autres systèmes ayant participé à l'évaluation (cf tableaux 7.8 et 7.9), avec notamment une entité d'écart pour la classe LOCATION par rapport au système le

plus performant GYDER. Nous reproduisons dans le tableau 7.12 une vue d'ensemble des résultats des participants à SemEval présentée dans l'article de K. Markert et M. Nissim consacré au bilan de la tâche [Markert et Nissim, 2007b], avec les taux de précision moyen, minimum et maximum.

Cinq systèmes ont participé à la tâche sur les LOCATION et trois à celle sur les ORGANISATION, pour tous les niveaux de granularité à chaque fois. Hormis XRCE-M, les systèmes reposaient sur des méthodes statistiques par apprentissage, exploitant divers algorithmes et un échantillon de traits différents. Deux systèmes ne sont pas parvenus à atteindre la baseline pour les noms de lieux ([Poibeau, 2007] et [Leveling, 2007]), montrant par là la difficulté de la tâche. Ces derniers n'ont cependant exploité que des traits de surface sans utiliser de ressource externe, contrairement aux deux autres systèmes ([Farkas *et al.*, 2007] et [Nicolae *et al.*, 2007]) qui, pour leur part, ont utilisé des traits syntaxiques et sémantiques et fait usage de ressouces externes (Wordnet) et/ou du web, avec au final de meilleurs résultats, attestant de la pertinence de ce type de traits pour le traitement de la métonymie. Pour l'ensemble des systèmes, seulement quelques catégories ont pu être annotées avec succès, telles celles `org-for-members`, `org-for-products`, `place-for-people` et `literal`. Les patrons les plus rares n'ont pas été reconnus pour certains (cf. "undef" dans le tableau 7.12), et mal annotés pour d'autres. Le patron `mixed`, d'emblée reconnu comme délicat et difficile par les organisatrices (cf. section 7.2.2.1), ne fut traité avec succès que par le système GYDER et pour quelques cas seulement, ce fait confirmant l'absence de configuration contextuelle stable et productive pour ce type de glissement métonymique. Seul le patron `object-for-name` semble faire exception, avec un traitement relativement réussi par les systèmes UTD-HLT CG et XRCE-M.

Les perspectives de développements sont donc nombreuses, pour notre système tout comme pour la tâche de résolution de métonymie en général. À notre échelle, nous envisageons de poursuivre le travail entrepris afin d'améliorer les performances du système pour les glissements de sens déjà pris en compte et d'étendre ce traitement à d'autres types d'entités nommées. Au niveau de la tâche, les organisatrices affichent une volonté d'améliorer les données textuelles avec une meilleure prise en compte, lors de l'annotation de corpus, de phénomènes rares ainsi qu'un désir de traiter d'autres types de glissements de sens, pour les classes sémantiques présentes lors de cette campagne ainsi que d'autres, et ce sur des entités nommées comme sur des noms communs. En définitive, cette participation à la campagne SemEval fut une bonne expérience, occasionnant tout d'abord la mise au point d'un système dans de bonnes conditions (connaissance du domaine des organisatrices et mise à disposition de corpus), son évaluation dans des délais limités ensuite et, enfin, l'écriture d'un article [Brun *et al.*, 2007]

ainsi que le partage d'expériences avec les autres participants lors du Workshop à Prague (ACL 2007). Ce travail nous a également permis, aussi et surtout, de rendre compte d'un nouveau type de traitement au regard des entités nommées.

| | base | min | max | ave |
|---|---|---|---|---|
| LOCATION-coarse | | | | |
| accuracy | 0,794 | 0,754 | 0,852 | 0,815 |
| literal-f | | 0,849 | 0,912 | 0,888 |
| non-literal-f | | 0,344 | 0,576 | 0,472 |
| LOCATION-medium | | | | |
| accuracy | 0,794 | 0,750 | 0,848 | 0,812 |
| literal-f | | 0,849 | 0,912 | 0,889 |
| metonymic-f | | 0,331 | 0,580 | 0,476 |
| mixed-f | | 0,000 | 0,083 | 0,017 |
| LOCATION-fine | | | | |
| accuracy | 0,794 | 0,741 | 0,844 | 0,801 |
| literal-f | | 0,849 | 0,912 | 0,887 |
| place-for-people-f | | 0,308 | 0,589 | 0,456 |
| place-for-event-f | | 0,000 | 0,167 | 0,033 |
| place-for-product-f | | 0,000 | undef | 0,000 |
| obj-for-name-f | | 0,000 | 0,667 | 0,133 |
| obj-for-rep-f | | undef | undef | undef |
| othermet-f | | 0,000 | undef | 0,000 |
| mixed-f | | 0,000 | 0,083 | 0,017 |
| ORGANISATION-coarse | | | | |
| accuracy | 0,618 | 0,732 | 0,767 | 0,746 |
| literal-f | | 0,800 | 0,825 | 0,810 |
| non-literal-f | | 0,572 | 0,652 | 0,615 |
| ORGANISATION-medium | | | | |
| accuracy | 0,618 | 0,711 | 0,733 | 0,718 |
| literal-f | | 0,804 | 0,825 | 0,814 |
| metonymic-f | | 0,553 | 0,604 | 0,577 |
| mixed-f | | 0,000 | 0,308 | 0,163 |
| ORGANISATION-fine | | | | |
| accuracy | 0,618 | 0,700 | 0,728 | 0,713 |
| literal-f | | 0,808 | 0,826 | 0,817 |
| org-for-members-f | | 0,568 | 0,630 | 0,608 |
| org-for-event-f | | 0,000 | undef | 0,000 |
| org-for-product-f | | 0,400 | 0,500 | 0,458 |
| org-for-facility-f | | 0,000 | 0,222 | 0,141 |
| org-for-index-f | | 0,000 | undef | 0,000 |
| obj-for-name-f | | 0,250 | 0,800 | 0,592 |
| obj-for-rep-f | | undef | undef | undef |
| othermet-f | | 0,000 | undef | 0,000 |
| mixed-f | | 0,000 | 0,343 | 0,135 |

TAB. 7.12 – Vue d'ensemble des résultats des participants.

# Conclusion générale

## Synthèse

Le travail présenté dans ce mémoire avait pour double objectif de clarifier la notion d'*entité nommée* et de proposer des méthodes permettant une annotation enrichie de ces unités. À la suite d'un état des lieux de la tâche de reconnaissance des entités nommées retraçant ses succès mais pointant également ses difficultés, deux perspectives de recherche ont donc été investies, l'une à dominante théorique et l'autre à vocation pratique.

S'agissant du premier axe de recherche relatif au statut théorique des entités nommées, nous avons commencé par nous interroger sur la manière de définir un objet TAL. Eu égard aux caractéristiques constitutives de la discipline du Traitement Automatique du Langage, à la croisée de plusieurs domaines, nous avons analysé les pratiques définitoires en linguistique et en TAL avant d'explorer le discours théorique existant ainsi que les diverses réalisations. Ce point méthodologique nous a ainsi permis de préciser notre démarche (considérer tant des aspects linguistiques que des aspects d'automatique) et de déterminer l'objectif poursuivi (donner un cadre d'appréhension de ces unités et expliquer l'hétérogénité de cet ensemble).

Nous avons ainsi, dans un premier temps, examiné les entités nommées sous un angle de vue exclusivement linguistique avec, au sein du cadre théorique du sens et de la référence, l'analyse des catégories linguistiques les plus représentées dans l'ensemble 'entités nommées' : les noms propres et les descriptions définies. Deux caractéristiques linguistiques ont pu être dégagées, la monoréférentialité et l'autonomie référentielle, pouvant dès lors jouer le rôle de critères (linguistiques uniquement) d'identification des unités lexicales pouvant être considérées comme des entités nommées ou, en d'autres termes, faire office de dénominateur linguistique commun pour ces unités.

La deuxième étape du travail de définition consista ensuite à considérer les entités nommées et leurs caractéristiques linguistiques distinctives selon un angle

du vue ‘traitement automatique’. Après avoir examiné la réalité du cadre du sens et de la référence en TAL et constaté le nécessaire recours à un modèle du monde (restreint, pour le moment), nous avons pu formuler une proposition de définition des entités nommées prenant en compte le matériau de base qu'est le langage tout comme les contraintes et exigences imposées par les traitements informatiques.

Au terme de ce parcours de la linguistique au TAL et en conséquence de la proposition de définition des entités nommées, nous avons ainsi pu mettre en lumière les points suivants : d'un point de vue méthodologique tout d'abord, il est primordial, avant toute entreprise d'annotation automatique d'entités nommées, d'expliciter un modèle précisant les objets du monde à reconnaître ; d'un point de vue pratique ensuite, il peut être utile de s'appuyer sur les critères linguistiques d'autonomie et de monoréférentialité pour déterminer précisément ce qu'il convient d'annoter dans les textes et, enfin, du point de vue de l'objet entité nommée lui-même, il convient de considérer cet ensemble d'unités lexicales débordant les catégories linguistiques traditionnelles et trouvant sa raison d'être dans des besoins applicatifs comme une création de la discipline TAL.

Le second axe de recherche retenu fut celui, pratique et expérimental, de l'amélioration du traitement des entités nommées, avec une annotation rendant compte plus précisément du référent de ces unités. Deux expériences ont été menées à bien.

Il fut d'abord question d'une méthode d'annotation fine des entités nommées. Après avoir détaillé la construction d'une ressource d'étiquettes sémantiques fines pour les entités nommées, nous avons, d'une part, évoqué comment l'exploiter pour l'aide à la construction de modèle et, d'autre part, montré comment l'utiliser pour une double annotation et une première désambiguïsation des entités nommées.

La seconde expérience porta pour sa part sur la résolution de métonymie des entités nommées ; au regard des glissements métonymiques définis dans le cadre de la campagne d'évaluation *SemEval*, nous avons proposé une méthode hybride combinant approche symbolique et approche distributionnelle.

## Améliorations et perspectives

Nous espérons, avec ces travaux, avoir en partie atteint les objectifs fixés et fait quelque peu évoluer la problématique des entités nommées. Cette contribution demeure cependant perfectible et serait à poursuivre selon plusieurs directions.

Pour ce qui est de la réflexion théorique sur la notion d'entité nommée, la définition avancée précédemment n'est qu'une proposition, laquelle gagnerait as-

surément à être critiquée, questionnée, discutée. Si nous sommes convaincue du bien-fondé de l'approche adoptée, alliant exploration linguistique et analyse des capacités et besoins du TAL, il existe certainement d'autres manières d'aborder la question du statut théorique des entités nommées. Au regard des expérimentations présentées, annotation fine et résolution de métonymie, nous avons déjà évoqué des perspectives de développement pour chacune d'elle à la fin des chapitres concernés.

Une perspective possible de notre travail consisterait à poursuivre la mise en adéquation des considérations théoriques et des réalisations "pratiques". La dernière partie présente en effet des modules s'attachant à améliorer ponctuellement l'annotation des entités nommées, sans pour autant pleinement illustrer et développer le statut théorique et ses implications méthodologiques et pratiques discutés dans la partie précédente. Les éléments que nous avons tenté de poser de part et d'autre afin d'encourager un dialogue entre questions théoriques et questions pratiques pourront, nous l'espérons, soutenir le rôle que sont appelées à jouer, au sein de ce que nous avons appelé la "composante sémantique", les entités nommées.

# Annexes

# Annexe A

# Tableau récapitulatif des campagnes d'évaluation sur les entités nommées

Le tableau suivant présente les principales campagnes d'évaluation sur les entités nommées.

| Campagne | Date | Langue | Nature des corpus | Nb participants et systèmes |
|---|---|---|---|---|
| **MUC-6** | 1995 | anglais | Dépêches de presse sur les changements de positions dans les entreprises | 15 participants, 20 systèmes |
| **MET-1** | 1995 | espagnol, chinois, japonais | Dépêches de presse | 7 systèmes |
| **MUC-7** | 1998 | anglais | Training : dépêches sur les crashes d'avions, Test : lancements de satellites | 12 systèmes |
| **MET-2** | 1998 | chinois, japonais | | 2 systèmes chinois, 3 japonais |
| **IREX** | 1998-1999 | japonais | Quotidien japonais | 14 participants, 15 systèmes |
| **ACE** | 1998-1999 | anglais | Articles de presse + transcriptions | - |
| **CoNLL** | 2002 | espagnol et hollandais | Articles de presse | 12 systèmes |
| | 2003 | allemand et anglais | Articles de presse | 16 systèmes |
| **ESTER** | 2002-2006 | français | Transcriptions d'émissions radiophoniques | - |
| **HAREM** | 2005 | portugais et brésilien | Collection de textes divers (presse, fiction, technique, oral, web, etc.) | 10 participants, 15 systèmes |
| | 2008 | - | - | 16 inscrits |

# Annexe B

# Entités nommées : les formules définitoires existantes

La liste suivante regroupe différentes « définitions » ou propos descriptifs s'appliquant aux entités nommées recueillis dans divers rapports de campagnes d'évaluation, de projets, ou encore dans des « encyclopédies ». S'il convient de ne pas « surinterpréter » ces citations, relevant de travaux relativement différents et présentées ici hors de leurs contextes, leur recensement permet de se faire une idée du discours général sur les entités nommées.

## B.1 Dans les campagnes d'évaluation :

**MUC-7 :** « *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* » [Chinchor, 1998].

« *The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are « unique indentifiers » of entities (organization, persons, locations), times (dates, times), and quantities (monetary values, percentages)* » [Chinchor, 1997].

**ESTER :** « *Même s'il n'existe pas de définition standard, on peut dire que les EN sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)* » [Meur *et al.*, 2004].

**CoNNL :** « *Named entities are phrases that contain the names of persons, organizations and locations* » [TjongKimSang et Meulder, 2003].

## B.2 Dans les projets de reconnaissance d'entités nommées :

**N. Friburger :** « *En fait il semble difficile des délimiter les noms propres des autres noms ; il y a une continuité entre l'ensemble des noms propres et l'ensemble des noms communs. Les informaticiens qui travaillent dans le domaine de l'extraction d'information, ont abordé le problème de manière pragmatique. Ils ont défini la notion d'**entités nommées** pour regrouper tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantités* » [Friburger, 2002].

**M. Tran :** « *Avec MUC-6, les noms propres, ainsi que les dates et les unités chiffrées sont regroupées sous le terme d'entités nommées* ». [Tran, 2006]

**D. Weissenbacher :** *« Nous donnons une définition provisoire de ce terme, une entité nommée c'est un syntagme qui réfère à un unique objet d'une réalité supposée. Les noms propres, d'entreprises, de lieux, de dates... tombent sous cette définition. (...) Nous admettrons à l'avenir qu'une EN est un nom non ambigu pour un intervenant essentiel de l'événement à modéliser. Remarquons que l'ancienne définition est inclue dans la nouvelle. En effet, pour désigner un individu unique, il suffit de former le concept être cet individu* » [Weissenbacher, 2003].

**T. Poibeau :** *« On appelle traditionnellement 'entités nommées' (de l'anglais named entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérables par les mêmes techniques à base de grammaires locales. Seules les entités nommées au sens propre du terme seront abordées dans ce chapitre (...)* » [Poibeau, 2003].

## B.3 Dans les « ,encyclopédies » ou études sur la tâche de reconnaissance d'entités nommées :

**B. Daille et E. Morin :** « *la notion d'entité nommée représente une catégorisation bien plus large que celle des noms propres (...), puisqu'elle inclut des expressions temporelles ou numériques, des maladies ou des drogues.* » [Daille et Morin, 200

**P. Enjalbert :** « [Le] *repérage et* [l']*étiquetage sémantique des entités nommées : il faut détecter toutes les formes linguistiques qui, à l'instar des nom propres, désignent de manière univoque une entité par leur pouvoir de sélectivité : noms de personnes, d'institutions et entreprises, de lieux, ainsi souvent que les dates, unités monétaires etc. Il faut aussi leur affecter une étiquette sémantique choisie parmi une liste prédéfinie.* » [Enjalbert, 2005a].

**M.R. Vicente :** « Entité Nommée *est la notion utilisée en TAL pour désigner les éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres (note : la notion d'entité nommée concerne les noms propres mais aussi les dates et les mesures) et qui suivent des patrons syntaxique déterminés* » [Vicente, 2005].

**S. Sekine :** « *The names of particular things or classes, and numeric expressions is regarded as an importent component technology for many NLP applications.(...) In this paper, the term* Named Entity *includes names (which is the narrow sense of Named Entity) and numeric expressions. The definition of this Named Entity is not simple, but, intuitively, this is a class that people are often willing to know in newpaper articles.* » [Sekine *et al.*, 2002].

**Atalapédie :** Version du 03/12/2005 : « *Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprise, etc. contenues dans un texte. Ces entités référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes.* »

Version du 08/12/2005 : « *Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprise, etc. contenues dans un texte. On ajoute souvent à ces éléments les dates et d'autres données chiffrées. Par extension, les entités désignent parfois les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologie). (...) Ces séquences référentielles sont primordiales pour*

*beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes* ».

**Wikipédia** « *Named entity recognition (NER) (also known as entity identification (EI) and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. (…) In the expression named entity, the word named restricts the task to those entities for which one or many rigid designators, as defined by Kripke, stands for the referent.* »

# Annexe C

# Résolution de métonymie à SemEval 2007 : les catégories d'annotation

Cette annexe présente les catégories d'annotations pour la classe ORGANISATION puis pour la classe LOCATION, en considérant à chaque fois les interprétations `literal`, `metonymic` et `mixed` avec, pour la lecture `metonymic`, des précisions sur les patrons. Bien que certains patrons soient indépendants de la classe sémantique, ils sont présentés successivement pour chaque classe.

## C.1 Catégories d'annotation pour la classe ORGANISATION

1. Annotation `literal`

   Pour les noms d'entreprise, la catégorie d'annotation `literal` correspond aux cas où un nom d'organisation réfère à l'entreprise en tant qu'entité légale. Cette annotation couvre les cas, comme dans les exemples ci-dessous, de description de la structure d'une entreprise, d'association ou d'acquisition entre entreprises ou encore les cas de relation entre l'entreprise et les services ou produits qu'elle offre.

   - NATO countries.
   - Sun acquired that part of Eastman-Kodak Co's Unix Subsidiary.
   - Intel's Indeo video compression hardware.

2. Annotation `metonymic`

   L'annotation de type `metonymic` comprend des patrons métonymiques comme des catégories dédiées aux métonymies "créatives".

– `organisation-for-members`
Cette catégorie est utilisée lorsqu'un nom d'entreprise est utilisé pour désigner ses employés. Par exemple, lorsqu'un porte-parole parle au nom de l'entreprise, ou que les employés participent à une action, comme dans les exemples suivants.

> - Last February, IBM announced [...].
> - It's customary to go to work in black and white suits. [...] Woolworths were them.

– `organisation-for-event`
On use de cette annotation lorsqu'un nom d'entreprise est utilisé pour référer à un événement lié à elle (par exemple un scandale ou une banqueroute).

> A remarkable example of back-bench influence on the Prime Minister was seen in the resignation of Leon Brittan from Trade and Industry in the aftermath of Westland.

– `organisation-for-product`
Ce patron est à adopter lorsqu'un nom d'entreprise est utilisé pour référer aux produits qu'elle produit.

> A red light was hung on the Ford's tail-gate.

– `organisation-for-facility`
Un nom d'entreprise est parfois utilisé pour renvoyer aux bâtiments qui l'abritent, il faut dans ce cas annoter à l'aide de la catégorie ci-dessus.

> The opening of a McDonald's is a major event.

– `organisation-for-index`
Ce patron est pour les cas où un nom d'entreprise est utilisé pour renvoyer aux indices boursiers indiquant sa valeur.

> BMW slipped 4p to 31p.

– `object-for-name`

Ce patron est, comme les trois suivants, indépendants de la classe sémantique. Pour les entreprises, il convient de l'utiliser pour les mentions de noms d'entreprise en usage autonyme, c'est-à-dire évoquant des caractéristiques de leurs propres noms, comme dans l'exemple ci-dessous.

> Chevrolet is feminine because of its sound (it's a longer word than Ford, has an open vowel at the end, connotes Frenchness).

– `object-for-representation`
Il convient d'utiliser ce patron lorsqu'un nom d'entreprise réfère à sa propre représentation, au travers d'un logo par exemple.

> BT's pipes-of-Pan motif was, for him, somehow too British. Graphically, it lacked what King calls the world class of Shell.

– `orthermet`
Enfin, ce dernier patron permet de couvrir les cas pour lesquels aucune des annotations précédentes n'a pu être utilisée. Dans l'exemple suivant, *Barclays Bank* renvoie à un compte en banque.

> funds [. . . ] had been paid into Barclays Bank.

3. Annotation `mixed`
L'annotation `mixed` est similaire au zeugma ; elle correspond aux cas où un nom d'entreprise convoque simultanément deux interprétations métonymiques. Dans l'exemple suivant, il est question de l'indice boursier (`organisation-for-index`) de l'entreprise et de ses employés (`organisation-for-members`).

> Barclays slipped 4p to 351p after confirming 3,000 more job losses.

## C.2 Catégories d'annotation pour la classe LOCATION

1. Annotation `literal`
Pour les noms de lieux, l'interprétation litérale est valable lorsqu'il s'agit de l'entité géographique ou de l'entité politique.

> - The coral coast of Papua New Guinea.
> - Britain's current account deficit.

2. Annotation `metonymic`
   - `place-for-people`
     Cette annotation est à utiliser lorsqu'un nom de pays renvoie aux personnes ou aux organisations qui lui sont associées. On peut penser aux cas où il est question du gouvernement, d'une équipe sportive ou bien de la population du pays. Ces trois cas sont illustrés par les exemples ci-dessous ; le dernier exemple illustre le cas d'une sous-spécification du référent cible, qui reste tout de même de type "people".

     - America did once try to ban alcohol.
     - England lost in the semi-final.
     - The notion that the incarnation was to fulfil the promise to Israel and to reconcile the world with God.
     - The G-24 group expressed readiness to provide Albania with food aid.

   - `place-for-event`
     Annotation à utiliser lorsque le nom de pays renvoie à un événement ayant eu lieu dans ce pays (cf. exemple donné dans la section 7.2.1.1).

   - `place-for-product`
     Il s'agit d'une catégorie pour les cas de renvois à un produit fabriqué par le pays (ou le lieu).

     A smooth Bordeaux that was gutsy enough to cope with our food.

   - `object-for-name`
     Nous retrouvons ici les patrons valables pour toutes les classes sémantiques. Cette annotation est pour les noms de pays renvoyant à leur propre nom (mention autonymique).

     Guyana (formerly British Guiana) gained independence.

   - `object-for-representation`
     Un nom peut renvoyer à une représentation, telle une photo, une peinture, un symbole, du référent réel évoqué par ce nom. Pour les noms de pays, on peut penser à la mention pouvant exister sur une carte géographique, comme dans l'exemple suivant, dans le contexte de quelqu'un pointant du doigt un point sur une carte :

This is Malta.

– `othermet`
Comme précédemment, cette catégorie est pour les "inclassables". Dans l'exemple suivant *New Jersey* renvoie à un style particulier de musique.

The thing about the record is the influences of the music. The bottom end is very New York/New Jersey and the top is very melodic.

3. Annotation `mixed`
Ce patron équivaut toujours à la présence de deux interprétations métonymiques possibles pour un même nom de lieu.

They arrived in Nigeria, hitherto a leading critic of [...]

# Annexe D

# XIP : un analyseur syntaxique robuste

*Xerox Incremental Parser* est un outil de traitement linguistique conçu au Centre de Recherche de Xerox à Meylan permettant l'analyse syntaxique robuste de textes tout venant ([Aït-Mokhtar et Chanod, 1997], [Aït-Mokhtar *et al.*, 2002]). De manière générale, la notion de robustesse correspond en TAL à la capacité d'un système à traiter des données linguistiques réelles ; pour un analyseur syntaxique, cela correspond à la capacité d'un système de produire des analyses utiles pour des textes réels, soit des analyses au moins partiellement correctes et utilisables dans une tâche automatique (voir [Bourigault, 2007] pour un état de l'art sur l'analyse syntaxique robuste). XIP fonctionne de manière incrémentale (les traitements sont appliqués les uns après les autres) et produit en sortie une liste de dépendances et un arbre de *chunks*. Nous détaillons successivement les traitements mis en œuvre.

Le traitement liminaire consiste à découper le texte en mots : il s'agit de la segmentation ou *tokenization*, tâche beaucoup plus difficile qu'il n'en paraît, compte tenu des problèmes posés par la ponctuation. L'étape suivante est l'analyse morpho-syntaxique, soit la description de la structure interne des mots, avec l'indication, pour chaque mot analysé, du lemme, des catégories possibles, du genre, du nombre et si besoin de la personne et du temps. Ces étapes de segmentation et d'analyse morphologique sont toutes deux réalisées à l'aide de transducteurs à états finis (*finite state transducers*). À ce stade du traitement, voici la sortie de l'analyseur pour la phrase très simple " The man closed the door " :

```
The the +0+3+Det+Def+SP+DET
man man +4+7+Noun+countable+c_person+Sg+NOUN
man man +4+7+Verb+Trans+Pres+Non3sg+VERB
closed close +8+14+Verb+s_sc_pbehind+s_sc_pon+s_p_up+s_p_over+s_p_off+s_p_in+s_p_down
   +Trans+PastBoth+123SP+VPAST
closed close +8+14+Verb+s_sc_pbehind+s_sc_pon+s_p_up+s_p_over+s_p_off+s_p_in+s_p_down
```

```
   +Trans+PastBoth+123SP+VPAP
closed closed +8+14+Adj+VPap+ADJPAP
the the +15+18+Det+Def+SP+DET
door door +19+23+Noun+countable+Sg+NOUN
. . +23+24+Punct+Sent+SENT
```

Intervient ensuite l'étape de désambiguïsation syntaxique (*Part of Speech tagging*) permettant d'attribuer à chaque unité lexicale une étiquette syntaxique unique en contexte. Ce traitement est réalisé via des règles symboliques locales et des modèles de Markov cachés (*Hidden Markov Model*).

La suite de l'analyse consiste à regrouper les mots en *chunks* ou syntagmes noyaux, c'est-à-dire en structures syntaxiques minimales non récursives. Deux types de règles sont convoqués lors de ce traitement (règles successives ou avec contrainte de préséance) et permettent la création d'un arbre (chunk tree) sur lequel peuvent alors venir s'appliquer des règles de déduction, en vue d'établir des relations dépendances (n-aires). Les structures syntaxiques obtenues à la sortie de ce traitement ne sont que superficielles dans la mesure ou des ambiguïtés liées à des phénomènes syntaxiques plus complexes demeurent (verbes non finis, subordination, passivité).

Dans le prolongement de cette analyse syntaxique générale, intervient une étape de normalisation en vue d'extraire des dépendances syntaxiques profondes [Hagège et Roux, 2003]. Ce travail de normalisation s'effectue en exploitant des informations de morphologie dérivationnelle, des relations syntaxiques mises en évidence lors de l'analyse générale et des informations sémantiques lexicales limitées (classes verbales de Levin). Ce traitement permet d'extraire des liens syntaxiques non superficiels tels que les sujets et objets normalisés des propositions infinitives, des relatives et des phrases à la forme passive.

Voici un exemple de sortie, avec l'arbre des chunks et les dépendances :

> Mr Bouez said Lebanon still wanted to see the implementation of a UN resolution demanding Israel immediately take back nearly 400 Palestinians deported to southern Lebanon.

TOP{SC{NP{NOUN{Mr Bouez}} FV{said}} SC{NP{Lebanon} FV{still wanted}} IV{to see} NP{the implementation} PP{of NP{a UN resolution}} NP{AP{demanding} Israel} immediately FV{take} back nearly NP{400 Palestinians} FV{deported} PP{to NP{NOUN{southern Lebanon}}.}

MOD_PRE(wanted,still)

MOD_PRE(Israel,demanding)
MOD_PRE(resolution,UN)
MOD_POST(deported,southern Lebanon)
MOD_POST(implementation,resolution)
SUBJ-N_PRE(said,Mr Bouez)
SUBJ-N_PRE(take,Israel)
SUBJ-N_PRE(deported,Palestinians)
SUBJ-N(demanding,Israel)
PERSON(Mr Bouez)
COUNTRY(Lebanon)
COUNTRY(Israel)
COUNTRY(Lebanon)
ORGANISATION(UN)
MANNER_PRE(deported,nearly)
MANNER_PRE(take,immediately)
EXPERIENCER_PRE(wanted,Lebanon)
EXPERIENCER(see,Lebanon)
CONTENT(see,implementation)
EMBED_INFINIT(see,wanted)
OBJ-N(implement,resolution)

Au terme de ces analyses, sont disponibles des informations relatives à des relations syntaxiques superficielles et profondes, exploitables par d'autres modules faisant davantage intervenir une dimension sémantique. Ces modules complémentaires sont, grâce à une souplesse d'architecture, facilement intégrables à l'analyseur robuste, à l'instar du système de repérage d'Entités Nommées.

L'architecture globale du système est présenté dans la figure D.1.

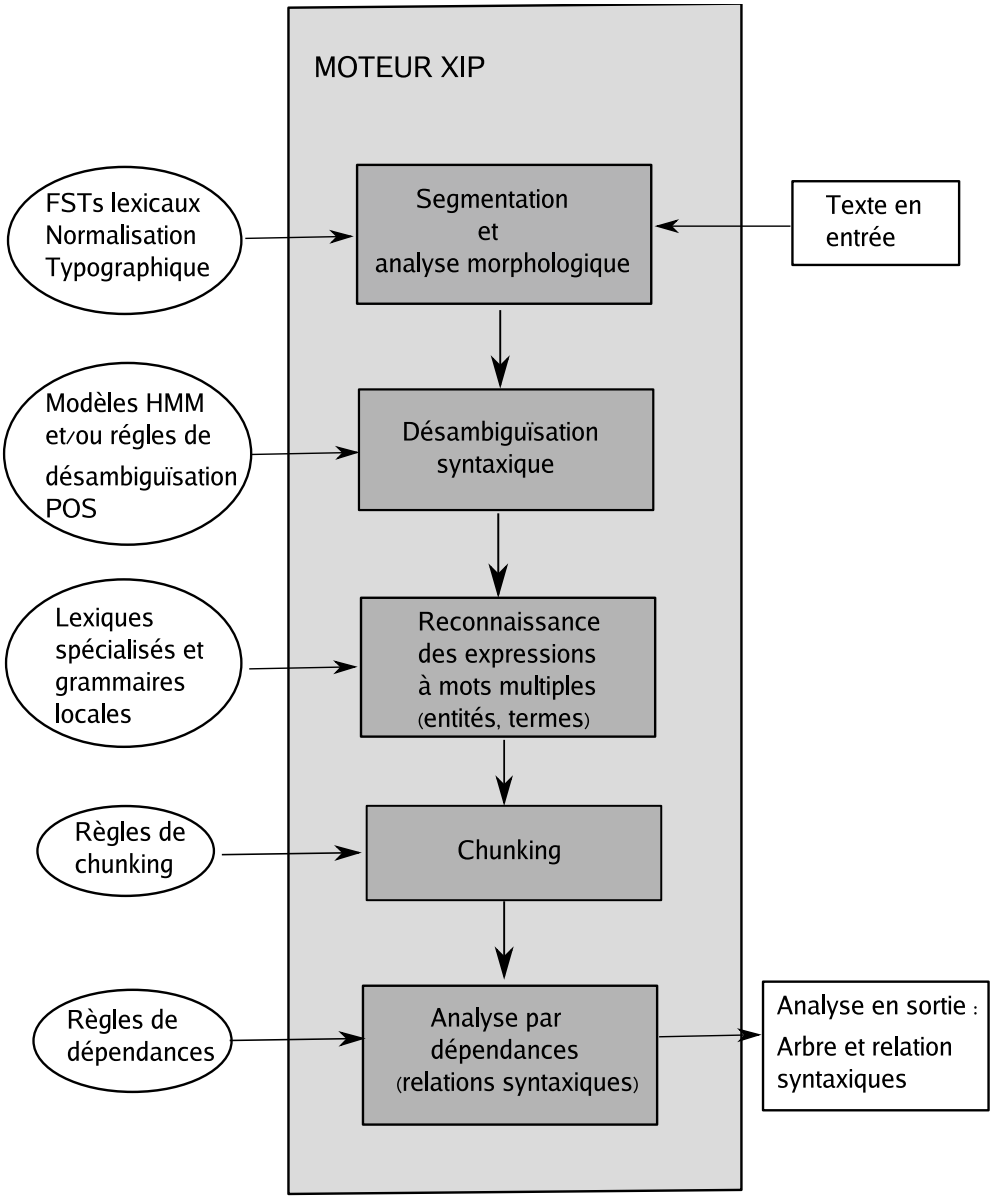



FIG. D.1 – Architecture du système

# Bibliographie


[ACE, 2000] ACE (2000). Entity Detection and Tracking - ACE Pilot Study Task Definition. Disponible à l'adresse suivante : http://www.nist.gov/speech/tests/ace/phase1/doc/index.htm.

[ACE, 2005] ACE (2005). The ACE 2005 Evaluation Plan. http://www.nist.gov/speech/tests/ace/ace05/.

[Algeo, 1973] Algeo, J. (1973). *On defining the Proper Name*. University of Florida Press, Gainesville.

[Allerton, 1987] Allerton, D. J. (1987). The linguistic and Sociolinguistic Status of Proper Names : What are they and what do they belong to ? *Journal of Pragmatics*, 11.

[Anscombre, 1996] Anscombre, J.-C. (1996). Théories et méthodes en sémantique française. Thèse d'Habilitation, Université de Paris VIII.

[Appelt *et al.*, 1993] Appelt, D., Hobbs, J., Bear, J., Israel, D. et Tyson, M. (1993). Fastus : A Finite-state Processor for Information Extraction from Real-world Text. *In IJCAI*, pages 1172–1178.

[Assadi, 1998] Assadi, H. (1998). *Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires.* Thèse de doctorat, Université Paris VI.

[Ayari, 2007] Ayari, S. E. (2007). Evaluation transparente de systèmes de question-réponses : application au focus. *In RECITAL, Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles, TALN*, Toulouse.

[Aït-Mokhtar et Chanod, 1997] Aït-Mokhtar, S. et Chanod, J. (1997). Incremental finite-state parsing. *In Proceedings of Applied Natural Language Processing*, Washington, DC.

[Aït-Mokhtar *et al.*, 2002] Aït-Mokhtar, S., Chanod, J. et Roux, C. (2002). Robustness beyond shallowness : incremental dependency parsing. *Natural Language Engineering Journal.*

[Bauer, 1985] Bauer, G. (1985). Namenkunde des deutschen. *Germanistische Lehrbuchsammlung.*

[Bikel *et al.*, 1997] Bikel, D., Miller, S., Schwartz, R. et Weischedel, R. (1997). Nymble : a high-performance learning name-finder. *In Proceedings of the fifth conference on Applied natural language processing*, pages 194–201, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Bodenreider et Zweigenbaum, 2000] Bodenreider, O. et Zweigenbaum, P. (2000). Stratégies d'identification des noms propres à partir de nomencaltures médicales parallèles. *Traitement Automatique des Langues*.

[Bontcheva *et al.*, 2004] Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V. et Cunningham, H. (2004). Shallow Methods for Named Entity Coreference Resolution. *In Actes de la Conférences Traitement Automatique du Langage Naturel (TALN)*, Nancy, France.

[Borthwick, 1999] Borthwick, A. (1999). *A Maximum Entropy Approach to Named Entity Recognition*. Thèse de doctorat, New York University.

[Borthwick *et al.*, 1998] Borthwick, A., Sterling, J., Agichtein, E. et Grishman, R. (1998). Nyu : Description of the MENE Named Entity System as used in MUC. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

[Bouaud *et al.*, 2000] Bouaud, J., Habert, B., Nazarenko, A. et Zweigenbaum, P. (2000). Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. *In* Charlet, J., Zacklad, M., Kassel, G. et Bourigault, D., éditeurs : *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, chapitre 17, pages 275–290. Eyrolles, Paris.

[Bourigault, 2002] Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN 2002*, Nancy.

[Bourigault, 2007] Bourigault, D. (2007). syntex, analyseur syntaxique opérationnel. habilitation à Diriger des Recherches.

[Brill, 1995] Brill, E. (1995). Transformation-based error-driven learning and natural language processing : a case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

[Brun *et al.*, 2007] Brun, C., Ehrmann, M. et Jacquet, G. (2007). XRCE-M : A hybrid system for named entity metonymy resolution. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Brun et Hagège, 2004] Brun, C. et Hagège, C. (2004). Intertwining Deep Syntactic Processing and Named Entity Detection. *In EsTAL*, pages 195–206.

[Bunescu et Paşca, 2007] Bunescu, R. et Paşca, M. (2007). Using encyclopedic knowledge for named entity disambiguation. *In Proceedings of the 11th Confe-*

*rence of the European chapter of the Association for Computational Linguistics*, Prague ?

[C.Fellbaum, 1998] C.Fellbaum, éditeur (1998). *WordNet, an electronic lexical database*. The MIT Press, Cambridge, Massachusetts.

[Charolles, 2002a] Charolles, M. (2002a). *Les expressions nominales définies*, chapitre IV, pages 75–104. *In* [Charolles, 2002b].

[Charolles, 2002b] Charolles, M. (2002b). *La référence et les expressions référentielles en français*. L'essentiel français. Ophrys.

[Chinchor, 1997] Chinchor, N. (1997). Named Entity Task Definition. Version 3.5 disponible à l'adresse suivante : http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

[Chinchor, 1998] Chinchor, N. (1998). Overview of MUC-7. *In Proceedings Seventh Message Understanding Conference (MUC-7)*, Fairfax, Virginia.

[Coates-Stephens, 1992] Coates-Stephens, S. (1992). *The Analysis and Acquisition of Proper Names fo Robust Text Understanding*. Thèse de doctorat, Department of Computer Science of City University, London, England.

[Collins et Singer, 1999] Collins, M. et Singer, Y. (1999). Unsupervised models for named entity classification. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[Cori et Léon, 2002] Cori, M. et Léon, J. (2002). La constitution du TAL, étude historique des dénominations et des concepts. *Revue TAL*, 43(3):21–55.

[Croft et Cruse, 2004] Croft, W. et Cruse, D. (2004). *Cognitive Linguistics*. Cambridge University Press.

[Cruse, 1996] Cruse, D. (1996). La signification des noms propres de pays en anglais. *In* Rémi-Giraudand, S. et Rétat, P., éditeurs : *Les mots de la nation*, pages 93–102. Presses Universitaires de Lyon.

[Daille *et al.*, 2000] Daille, B., Fourour, N. et Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, pages 115–129.

[Daille et Morin, 2000] Daille, B. et Morin, E. (2000). Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. *Traitement Automatique des Langues*, 41(3):601–621.

[de Saussure, 1915] de Saussure, F. (1915). *Cours de linguistique générale*. Payot, Paris, 1976 édition.

[de Velde, 2000] de Velde, D. V. (2000). Existe-il des noms propres de temps ? *Lexique*, 15:35–45. Presses Universitaires du Septentrion.

[de Velde *et al.*, 2000] de Velde, D. V., Flaux, N. et Mulder, W. D., éditeurs (2000). *Lexique. Les noms propres : nature et détermination*, volume 15, Villeneuve d'Ascq. Presses du Septentrion.

[Descombres, 2001] Descombres, V. (2001). Les individus collectifs. *Revue MAUSS*, 18(2):305–337.

[D.Farmakiotou *et al.*, 2002] D.Farmakiotou, Karkaletsis, V., Samaritakis, G., Petasis, G. et Spyropoulos, C. (2002). Named Entities Recognition in Greek Web Pages. *In 2nd Hellenic Conference on AI, SETN-2002*, pages pp. 91–102, Thessaloniki, Greece.

[Doddington *et al.*, 2004] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. et Weischedel, R. (2004). The Automatic Content Extraction (ace) program tasks, data, and evaluation. *In Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal.

[Dubois *et al.*, 1994] Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J. et Mével, J. (1994). *Dictionnaire de linguistique*. Larouse.

[Ehrmann, 2005] Ehrmann, M. (2005). Compte-rendu de lecture de M. Stevenson. *Traitement Automatique des Langues*, 46(3). (voir Stevenson, 2003).

[Ehrmann et Jacquet, 2006] Ehrmann, M. et Jacquet, G. (2006). Vers une double annotation des entités nommées. *Traitement Automatique des Langues*, 47(3). Disponible sur : http://www.atala.org.

[Enjalbert, 2005a] Enjalbert, P. (2005a). *L'extraction d'information*, chapitre 8, pages 309–334. *In* [Enjalbert, 2005b].

[Enjalbert, 2005b] Enjalbert, P., éditeur (2005b). *Sémantique et traitement automatique du langage naturel*. Hermès, Paris.

[Enjalbert et Bilhaut, 2005] Enjalbert, P. et Bilhaut, F. (2005). *L'accès assisté à l'information documentaire*, chapitre 9, pages 335–370. *In* [Enjalbert, 2005b].

[Enjalbert et Victorri, 2005] Enjalbert, P. et Victorri, B. (2005). *Les paliers de la sémantique*, chapitre 2, pages 53–96. *In* [Enjalbert, 2005b].

[Farkas *et al.*, 2007] Farkas, R., Simon, E., Szarvas, G. et Varga, D. (2007). GYDER : maxent metonymy resolution. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Fauconnier, 1984] Fauconnier, G. (1984). *Espaces mentaux*. Minuit.

[Faure et Nedellec, 1999] Faure, D. et Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system ASIUM. *In Proceedings of the 11th European Workshop,*

*Knowledge Acquisition, Modelling and Managment (EKAW'99)*, Juan-les-Pins, France.

[Ferret *et al.*, 2001] Ferret, O., B. Grau, a. M. H.-P., Illouz, G. et Jacquemin, C. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *In Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles, TALN*, Tours.

[Flaux, 1995] Flaux, N. (1995). La catégorisation du nom propre. *In* Noailly, M., éditeur : *Nom propre et Nomination*, pages 63–74, Paris. Klincksieck.

[Fleischman, 2001] Fleischman, M. (2001). Automated subcategorization of named entities. *In Proceedings of ACL Student*, Toulouse, France.

[Fleischman et Hovy, 2002] Fleischman, M. et Hovy, E. (2002). Fine grained classification of named entities. *In Proceedings of the 19th international conference on Computational linguistics (COLING)*, Taipei, Taiwan. Association for Computational Linguistics.

[Floch et Habert, 2000] Floch, H. et Habert, B. (2000). Constructing a navigable topic map by inductive semantic acquisition methods. Extreme Markup Languages. Disponible à l'adresse suivante : http://www.limsi.fr/Individu/habert/Publications/Fichiers/folch-et-habert00/folch-et-habert00.html.

[Fourour, 2002] Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *In Actes, Neuvième Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN 2002))*, volume 1, pages 265–274.

[Franckel et Paillard, 1997] Franckel, J.-J. et Paillard, D. (1997). Représentation formelle des mots du discours ; le cas de 'd'ailleurs'. *Revue de sémantique et de pragmatique*, 1:51–64.

[Frege, 1892] Frege, G. (1892). *Ecrits logiques et philosophiques*, chapitre Sens et dénotation, pages 102–126. L'Ordre philosophique. Le Seuil, Paris.

[Friburger, 2002] Friburger, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques.* Thèse de doctorat, Université François Rabelais, Tours. (sous la direction de M. Denis Maurel).

[Fuchs, 1996] Fuchs, C. (1996). *Les ambiguïtés du français.* L'essentiel français. Ophrys.

[Fukuda *et al.*, 1998] Fukuda, K., Tamura, A., Tsunoda, T. et Takagi, T. (1998). Toward Information Extraction : Identifying Protein Names from Biological Papers. *In Proceedings of the Pacific Symposium on BioComputing*, Hawaii.

[Gaizauskas *et al.*, 1995] Gaizauskas, R., Humphreys, K., Cunningham, H. et Wilks, Y. (1995). University of Sheffield : description of the LaSIE system as used for MUC-6. *In MUC6 '95 : Proceedings of the 6th conference on Message Understanding*, pages 207–220, Morristown, NJ, USA. Association for Computational Linguistics.

[Gary-Prieur, 1991] Gary-Prieur, M. (1991). Le nom propre constitue-t-il une catégorie linguistique ? *Langue Française*, 92:4–25.

[Gary-Prieur, 1994a] Gary-Prieur, M. (1994a). *Grammaire du nom propre*. Presses Universitaires de France, Paris.

[Gary-Prieur, 1994b] Gary-Prieur, M. (1994b). *L'importation des thèses logiques en linguistique*, chapitre 1, pages 14–25. *In* [Gary-Prieur, 1994a].

[Grevisse et Goosse, 1986] Grevisse, M. et Goosse, A. (1986). *Le Bon Usage*. Duculot, Paris.

[Grishman, 1995] Grishman, R. (1995). The NYU system for MUC-6 or where's the syntax ? *In MUC6 '95 : Proceedings of the 6th conference on Message understanding*, pages 167–175, Morristown, NJ, USA. Association for Computational Linguistics.

[Grishman, 1997] Grishman, R. (1997). Information Extraction : Techniques and Challenges. *In SCIE '97 : International Summer School on Information Extraction*, pages 10–27, London, UK. Springer-Verlag.

[Grishman et Sundheim, 1995] Grishman, R. et Sundheim, B. (1995). Design of the MUC-6 evaluation. *In MUC6 '95 : Proceedings of the 6th conference on Message understanding*, Morristown, NJ, USA. Association for Computational Linguistics.

[Grishman et Sundheim, 1996] Grishman, R. et Sundheim, B. (1996). Message Understanding Conference-6 : a brief history. *In Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

[Guillemin-Lanne et Six, 2006] Guillemin-Lanne, S. et Six, A. (2006). La normalisation : nouveau challenge en extraction d'information. *In Colloque Veille Stratégique Scientifique et Technologique VSST06*.

[GuoDong *et al.*, 2005] GuoDong, Z., Jian, S., Jie, Z. et Min, Z. (2005). Exploring various knowledge in relation extraction. *In ACL '05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

[Habert, 2000] Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *In* Bilger, M., éditeur : *Linguistique sur corpus. Études et*

*réflexions*, numéro 31 de Cahiers de l'université de Perpignan, pages 11–58. Presses Universitaires de Perpignan, Perpignan.

[Habert *et al.*, 1997] Habert, B., Nazarenko, A. et Salem, A. (1997). *Les Linguistiques de corpus*. coll. U. Armand Colin, Paris.

[Hagège et Roux, 2003] Hagège, C. et Roux, C. (2003). Entre syntaxe et sémantique : Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. *In TALN 2003*, Batz-sur-Mer.

[Harabagiu, 1998] Harabagiu, S. M. (1998). Deriving metonymic coercions from WordNet. *In* Harabagiu, S., éditeur : *Use of WordNet in Natural Language Processing Systems : Proceedings of the Conference*, pages 142–148. Association for Computational Linguistics, Somerset, New Jersey.

[Hirschman, 1998] Hirschman, L. (1998). The Evolution of Evaluation : Lessons from the Message Understanding Conferences. *Computer Speech and Language*.

[Hottois, 2002] Hottois, G. (2002). *Penser la logique. Une introduction technique et théorique à la philosophie de la logique et du langage*. De Boeck Université.

[Humphreys *et al.*, 1998] Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. et Wilks, Y. (1998). University of Sheffield : description of the LaSIE-II system as used for MUC-7. *In Proceedings of the 7th conference on Message Understanding*. Association for Computational Linguistics.

[Jacquet et Venant, 2005] Jacquet, G. et Venant, F. (2005). Construction automatique de classes de sélection distributionnelle. *In Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2005)*, Dourdan, France.

[Jacquet *et al.*, 2005] Jacquet, G., Venant, F. et Victorri, B. (2005). *Polysémie lexicale*, chapitre 3, pages 99–132. *In* [Enjalbert, 2005b].

[Jonasson, 1994] Jonasson, K. (1994). *Le Nom Propre. Constructions et interprétations*. Champs Linguistiques. Duculot, Louvain-la-Neuve, Belgique.

[Karkaletsis *et al.*, 2003] Karkaletsis, V., Spyropoulos, C., Souflis, D., Grover, C., Hachey, B., Pazienza, M., Vindigni, M., Cartier, E. et J.Coch (2003). Demonstration of the CROSSMARC system. *In Companion Volume of the Proceedings of HLT-NAACL 2003*.

[Kleiber, 1981a] Kleiber, G. (1981a). *Les descriptions définies*, pages 301–510. *In* [Kleiber, 1981b].

[Kleiber, 1981b] Kleiber, G. (1981b). *Problèmes de référence. Descriptions définies et noms propres*. Paris.

[Kleiber, 1990] Kleiber, G. (1990). *La sémantique du prototype*. Presses Universitaires de France.

[Kleiber, 1995] Kleiber, G. (1995). Sur la définition des noms propres, une dizaine d'années après. *In* Noailly, M., éditeur : *Nom propre et Nomination*, pages 11–36, Paris. Klincksieck.

[Kleiber, 1996] Kleiber, G. (1996). Noms propres et noms communs : un problème de dénomination. *Méta*, 41(4):567–589.

[Kleiber, 1999a] Kleiber, G. (1999a). *Du sens. En général et en particulier*, chapitre 1. *In* [Kleiber, 1999b].

[Kleiber, 1999b] Kleiber, G. (1999b). *Problèmes de sémantique. La polysémie en questions*. Presses Universitaires du Septentrion.

[Kleiber, 2004] Kleiber, G. (2004). Peut-on sauver un sens de dénomination pour les noms propres ? *Functions of language*, 11(1):115–145.

[Kokkinakis, ] Kokkinakis, D. Swedish NER in the Nomen Nescio Project. Göteborg University, Sprakdata.

[Kokkinakis et Thurin, 2007] Kokkinakis, D. et Thurin, A. (2007). Identification of Entity References in Hospital Discharge Letters. *In Proceedings of the 16th Nordic Conference of Computational Linguistics*, Tartu, Estonia.

[Kripke, 1982] Kripke, S. (1972/1982). *La Logique des noms propres (Naming and Necessity*. Minuit, Paris.

[Krupka et Hausman, 1998] Krupka, G. et Hausman, K. (1998). Isoquest : Description of the NetOwl (TM) Extractor System as Used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

[Laurent et Vicente, 2004] Laurent, B. et Vicente, M. (2004). L'anthroponyme et le nom de marque et de produit : comparaison du prototype du nom propre et d'un hybride. *In Actes des VIIèmes Rencontres des Jeunes Chercheurs de l'ED268 'Langues et Langages'*, Paris.

[Lebart et Salem, 1994] Lebart, L. et Salem, A. (1994). *Analyse statistique des données textuelles*. Dunod.

[Lee et Lee, 2004] Lee, S. et Lee, G. (2004). A bootstrapping approach for geographic named entity annotation. *In AIRS*.

[Leroy, 2004a] Leroy, S. (2004a). *Le nom propre en français*. L'essentiel Français. Ophrys.

[Leroy, 2004b] Leroy, S. (2004b). *Nom propre modifié et nom propre standard : une distinction pertinente ?*, chapitre 5, pages 67–76. *In* [Leroy, 2004a].

[Leroy, 2005] Leroy, S., éditeur (2005). *Langue Française. Noms Propres : la modification*. Larousse.

[Leveling, 2007] Leveling, J. (2007). FUH (fernuniversitĺat in hagen) : Metonymy Recognition Using Different Kinds of Context for a Memory-Based Learner. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Li *et al.*, 2003] Li, H., Srihari, R., Niu, C. et Li, W. (2003). Infoxtract location normalization : a hybrid approach to geographic references in information extraction. *In Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*. ACL.

[Lin, 1998] Lin, D. (1998). Using collocation statistics in information extraction. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

[Linsky, 1974] Linsky, L. (1974). *Le problème de la référence*. L'Ordre philosophique. Le Seuil, Paris.

[Mann, 2002] Mann, G. (2002). Fine-grained proper noun ontologies for question answering. *In Proceedings of SemaNet'02, Building and Using Semantic Networks*.

[Mann et Yarowsky, 2003] Mann, G. et Yarowsky, D. (2003). Unsupervised personal name disambiguation. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, Edmonton, Canada. Association for Computational Linguistics.

[Marconi, 1997] Marconi, D. (1997). *La philosophie du langage au XXième siècle*. L'Eclat. Disponible sur : http://www.lyber-eclat.net/lyber/marconi/langage.html.

[Markert, 2000] Markert, K. (2000). Features integration in metonymy resolution. Rapport technique, Université d'Edimburgh.

[Markert et Hahn, 2002] Markert, K. et Hahn, U. (2002). Understanding metonymies in discourse. *Artificial Intelligence*, 135(1/2):145–198.

[Markert et Nissim, 2002a] Markert, K. et Nissim, M. (2002a). Metonymy Resolution as a Classification Task. *In Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Penn.

[Markert et Nissim, 2002b] Markert, K. et Nissim, M. (2002b). Toward a corpus annotated for metonymies : the case of location names. *In Proc. of the 3rd International Conference on Language Resources and Evaluations*, Las Palmas, Iles Canaries.

[Markert et Nissim, 2005] Markert, K. et Nissim, M. (2005). Annotation scheme for metonymies AS1. Rapport technique, Université de Leeds, Université d'Edimburgh. Disponible sur : http://www.comp.leeds.ac.uk/markert/Papers/index.html.

[Markert et Nissim, 2006] Markert, K. et Nissim, M. (2006). Metonymic proper names : a corpus based account. *In* Stefanowitsch, A. et Gries, S., éditeurs : *Corpus-based approaches to Metaphor and Metonymy*, pages 152–174. Mouton de Gruyter.

[Markert et Nissim, 2007a] Markert, K. et Nissim, M. (2007a). Metonymy Resolution at SemEval I : Guidelines for Participants. Rapport technique, SemEval.

[Markert et Nissim, 2007b] Markert, K. et Nissim, M. (2007b). Semeval-2007 task 08 : Metonymy resolution at semeval-2007. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Marton, 2003] Marton, G. (2003). Sepia : Semantic Parsing for Named Entities. Rapport technique, Massuchusettes Institute of Technology.

[Maurel et Tran, 2006] Maurel, D. et Tran, M. (2006). Prolex : une ressource sémantique multilingue de noms propres. *In Journée d'Etude ATALA : Des ressources sémantiques existantes à un Framenet français ?*

[Maynard *et al.*, 2005] Maynard, D., Bontcheva, K. et Cunningham, H. (2005). Towards a Semantic Extraction of Named Entities.

[McDonald, 1996] McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In* Boguraev, B. et Pustejovsky, J., éditeurs : *Corpus processing for lexical acquisition*, pages 21–39. MIT Press, Cambridge, MA, USA.

[Medlock, 2006] Medlock, B. (2006). An introduction to NLP-based textual anonymisation. *In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, Genes, Italie.

[Merchant *et al.*, 1996] Merchant, R., Okurowski, M. et Chinchor, N. (1996). The multilingual entity task (MET) overview. *In Proceedings of a workshop on held at Vienna, Virginia*, pages 445–447, Morristown, NJ, USA. Association for Computational Linguistics.

[Meur *et al.*, 2004] Meur, C. L., Gallinao, S. et Geoffrois, E. (2004). Conventions d'annotations en Entités Nommées. http://www.afcp-parole.org/ester/docs.html.

[Mikheev *et al.*, 1999] Mikheev, A., Moens, M. et Grover, C. (1999). Named Entity recognition without gazetteers. *In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

[Mill, 1843] Mill, J. (1843). *Système de logique déductive et inductive*. Londres.

[Miller *et al.*, 1998] Miller, S., Crystal, M., Fox, H., Ramshaw, L. et Schwartz, R. (1998). Algorithms that learn to extract information-BBN :

Description of the SIFT system as used for MUC-7. *In Proceedings of the 7th Message Understanding Conference*.

[Moldovan *et al.*, 2001] Moldovan, D., Paşca, M., Harabagiu, S. et Surdeanu, M. (2001). Performance issues and error analysis in an open-domain question answering system. *In ACL '02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

[Molino, 1982] Molino, J., éditeur (1982). *Langages. Le nom propre dans la langue*. Larousse.

[MUC-6, 1995] MUC-6 (1995). Named Entity Task Definition. Version 2.0 disponible à l'adresse suivante : http://cs.nyu.edu/faculty/grishman/muc6.html.

[Nicolae *et al.*, 2007] Nicolae, C., Nicolae, G. et Harabagiu, S. (2007). UTD-HLT-CG : Semantic Architecture for Metonymy Resolution and Classification of Nominal Relations. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Nilsson et Malmgren, 2005] Nilsson, K. et Malmgren, A. (2005). Towards automatic recognition of product names : an exploratory study of brand names in economic texts. *In Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA2005)*, Finland.

[Nissim et Markert, 2003] Nissim, M. et Markert, K. (2003). Syntactic features and word similarity for supervised metonymy resolution. *In Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03)*, Sapporo, Japon.

[Nissim et Markert, 2005] Nissim, M. et Markert, K. (2005). Learning to buy a Renault and to talk to a BMW : A supervised approache to conventional metonymy. *In Proceedings of the 6th International Workshop on Computational Semantics*, Tilburg.

[Noailly, 1995] Noailly, M., éditeur (1995). *Nom Propre et Nomination*. Klincksieck.

[Nunberg, 1995] Nunberg, G. (1995). Transfers of meanings. *Journal of Semantics*, 12(2):109–132. Disponible sur : http://jos.oxfordjournals.org/cgi/content/abstract/12/2/109.

[Osenova et Kolkovska, 2002] Osenova, P. et Kolkovska, S. (2002). Combining the named-entity recognition task and NP chunking strategy for robust pre-processing. *In Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, Sozopol, Bulgaria.

[Paşca, 2004] Paşca, M. (2004). Acquisition of categorized named entities for web search. *In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04)*. ACM Press.

[Paik *et al.*, 1996] Paik, W., E.Liddy, Yu, E. et McKenna, M. (1996). Categorizing and standardizing proper nouns for efficient information retrieval. pages 61–73.

[Paroubek et Rajman, 2000] Paroubek, P. et Rajman, M. (2000). *Etiquetage morpho-syntaxique*, chapitre 5, pages 131–150. *In* [Pierrel, 2000].

[Peirsman, 2006] Peirsman, Y. (2006). What's in a name ? Computational approaches to metonymical location names. *In Proceedings of the Workshop on Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together*, Trento, Italie.

[Phillips et Riloff, 2002] Phillips, W. et Riloff, E. (2002). Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP '02)*. Association for Computational Linguistics.

[Pierrel, 2000] Pierrel, J. M., éditeur (2000). *Ingénierie des langues*. Hermès, Paris.

[Plamondon *et al.*, 2004] Plamondon, L., Lapalme, G. et Pelletier, F. (2004). Anonymisation de décisions de justice. *In Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles, TALN*.

[Ploux et Victorri, 1998] Ploux, S. et Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *TAL*, 39(1).

[Poibeau, 1999] Poibeau, T. (1999). Le repérage des entités nommées, un enjeu pour les systèmes de veille. *In actes du colloque Terminologie et Intelligence Artificielle (TIA'99)*, Nantes.

[Poibeau, 2001] Poibeau, T. (2001). "Deconstructing Harry" : une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*, 5:25–33.

[Poibeau, 2003] Poibeau, T. (2003). *Extraction Automatique d'Information. Du texte brut au web sémantique*. Hermès.

[Poibeau, 2006] Poibeau, T. (2006). Dealing with metonymic readings of named entities. *In Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Canada.

[Poibeau, 2007] Poibeau, T. (2007). up13 : Knowledge-poor Methods (Sometimes) Perform Poorly. *In 4th International Workshop on Semantic Evaluations*, Prague. ACL-SemEval 2007.

[Poibeau et Nazarenko, 1999] Poibeau, T. et Nazarenko, A. (1999). L'extraction d'information, une nouvelle conception de la compréhension de texte ? *Traitement Automatique des Langues*, 40(2):87–115.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The generative lexicon*. The MIT Press, Cambridge.

[Ratnaparkhi, 1997] Ratnaparkhi, A. (1997). A simple introduction to maximum entropy models for natural language processing. Rapport technique, Institute for Research in Cognitive Science, University of Pennsylvania.

[Rebotier, 2006] Rebotier, A. (2006). Développement d'un module d'extraction d'entités nommées pour le français. Rapport technique, Centre de Recherche Xerox, Grenoble.

[Riegel *et al.*, 1994] Riegel, M., Pellat, J. et Rioul, R. (1994). *Grammaire méthodique du français*. Linguistique nouvelle. Presses Universitaires de France, Paris.

[Russell, 1905] Russell, B. (1905). On denoting. *Mind*. Disponible sur : http://www.cscs.umich.edu/(tilde)crshalizi/Russell/denoting/.

[Récanati, 1997] Récanati, F. (1997). La polysémie contre le fixisme. *Langue Française*, (113):107–123.

[Sabah, 2000] Sabah, G. (2000). *Sens et traitement automatique des langues*, chapitre 3. *In* [Pierrel, 2000].

[Santos *et al.*, 2006] Santos, D., Seco, N., Cardoso, N. et Vilela, R. (2006). HAREM : An Advanced NER Evaluation Contest for Portuguese. *In* Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J. et Tapias, D., éditeurs : *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 1986–1991, Genève, Italie. ELRA.

[Searle, 1972] Searle, J. (1972). *Les actes de langage*. Herrmann, Paris.

[Sekine, 2004] Sekine, S. (2004). Named Entity : History and Future. http://nlp.cs.nyu.edu/sekine/Main/publications.html.

[Sekine et Eriguchi, 2000] Sekine, S. et Eriguchi, Y. (2000). Japanese named entity extraction evaluation : analysis of results. *In Proceedings of the 18th conference on Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

[Sekine et Isahara, 1999] Sekine, S. et Isahara, H. (1999). IREX project overview. *In Proceedings of the Information Retrieval and Extraction Exercise*, Japon.

[Sekine et Nobata, 2004] Sekine, S. et Nobata, C. (2004). Definition, Dictionary and Tagger for Extended Named Entities. *In Forth International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, Portugal.

[Sekine *et al.*, 2002] Sekine, S., Sudo, K. et Nobata, C. (2002). Extended named entity hierarchy. *In The Third International Conference on Language Resources and Evaluation (LREC)*, Iles Canaries , Espagne.

[Settles, 2005] Settles, B. (2005). ABNER : an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

[Shinnou et Sekine, 2004] Shinnou, H. et Sekine, S. (2004). Named Entity Discovery Using Comparable News Articles. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING04)*, Genève, Suisse.

[Siblot, 1995] Siblot, P. (1995). Nom et images de marque : de la construction de sens dans les noms propres. *In Nom propre et Nomination*, pages 147–160. Klincksieck.

[Stevenson, 2003] Stevenson, M. (2003). *Word Sense Disambiguation. The case for combinations of knowledge sources.* CSLI Publications.

[Sundheim, 1995] Sundheim, B. (1995). Overview of results of the muc-6 evaluation. *In MUC6 '95 : Proceedings of the 6th conference on Message understanding*, Morristown, NJ, USA. Association for Computational Linguistics.

[TjongKimSang, 2002] TjongKimSang, E. (2002). Introduction to the CoNLL-2002 shared task : language-independent named entity recognition. *In COLING-02 : proceeding of the 6th conference on Natural language learning*, Morristown, NJ, USA. Association for Computational Linguistics.

[TjongKimSang et Meulder, 2003] TjongKimSang, E. et Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task : language-independent named entity recognition. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, Morristown, NJ, USA. Association for Computational Linguistics.

[Toral *et al.*, 2005] Toral, A., Noguera, E., Llopis, F. et Muñoz, R. (2005). Improving Question Answering Using Named Entity Recognition. *In Proceedings of the 10th NLDB congress*, Lecture notes in Computer Science, pages 181–191, Alicante, Spain. Springer-Verlag.

[Tran, 2006] Tran, M. (2006). *Prolexbase, un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne.* Thèse de doctorat, Université François Rabelais, Tours. (sous la direction de M. Denis Maurel).

[Varela, 1998] Varela, F. (1998). Le cerveau n'est pas un ordinateur. *La Recherche*, (308). disponible sur http://www.overdream.com/html/varela.htm.

[Vaxelaire, 2007] Vaxelaire, J. (2007). Ontologie et dé-ontologie en linguistique : le cas des noms propres. *Texto ! Textes et cultures (revue électronique)*, 12(2). http://www.revue-texto.net/.

[Vernant, 1980] Vernant, D. (1980). La théorie des descritpions définies de russell ou le problème de la référence. *Revue de métaphysique et de morale*, 85(5):489–502.

[Vernant, 1993] Vernant, D. (1993). *La philosophie mathématique de Bertrand Russell*. Vrin.

[Vicente, 2005] Vicente, M. (2005). La glose comme outil de désambiguïsation référentielle des noms propres ours. *Corela*, Le traitement lexicographique des noms propres.

[Victorri, 2002] Victorri, B. (2002). Espaces sémantiques et représentation du sens. *éc/artS*, 3.

[Victorri et Fuchs, 1996] Victorri, B. et Fuchs, C. (1996). *La polysémie. Construction dynamique du sens*. Hermes, Paris.

[Wacholder *et al.*, 1997] Wacholder, N., Ravin, Y. et Choi, M. (1997). Disambiguation of proper names in text. *In Proceedings of the fifth conference on Applied natural language processing*, pages 202–208, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Weissenbacher, 2003] Weissenbacher, D. (2003). Etude et reconnaissance automatique des relations de synonymie et de renommage dans les textes de génomique. Mémoire de D.E.A., Laboratoire d'informatique de Paris Nord. (sous la direction de Mme. Adeline Nazarenko).

[Zweigenbaum *et al.*, 1997] Zweigenbaum, P., Bouaud, J., Nazarenko, A. et Habert, B. (1997). Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies. *In Actes des 1ères Journées scientifiques et techniques FRANCIL 97*, Avignon.