FACULTÉ DES SCIENCES

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES
DE L'INFORMATION ET DES MATÉRIAUX

Année 2004

N° attribué par la bibliothèque

Identification et catégorisation automatiques des entités nommées dans les textes français

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE NANTES

Discipline : INFORMATIQUE

présentée et soutenue publiquement par

Nordine FOUROUR

le 29 juin 2004

à l'UFR Sciences et Techniques, Université de Nantes

devant le jury ci-dessous

Président	:		
Rapporteurs	:	Pascale SÉBILLOT, Maître de conférence HDR	IRISA (UMR 6074), Rennes
		Benoît HABERT, Professeur des Universités	LIMSI – CNRS, Université Paris X
Examinateurs	:	Laurence DANLOS, Professeur des Universités	Lattice-Talana, Université Paris 7
		Béatrice DAILLE, Professeur des Universités	LINA, Université de Nantes
		Emmanuel MORIN, Maître de conférence	LINA, Université de Nantes

Directeur de thèse : Professeur Béatrice DAILLE
Co-encadrant : Maître de Conférence Emmanuel MORIN
Laboratoire : Laboratoire d'Informatique de Nantes Atlantique

N° ED 0366-XXX

IDENTIFICATION ET CATÉGORISATION AUTOMATIQUES DES ENTITÉS NOMMÉES DANS LES TEXTES FRANÇAIS

Automatic Recognition and Categorisation of Named Entities in French Texts

Nordine Fourour

⋈

favet neptunus eunti

Université de Nantes

Nordine FOUROUR

Identification et catégorisation automatiques des entités nommées dans les textes français

xviii+150 p.

Ce document a été préparé avec LaTeX2$_\varepsilon$ et la classe `these-LINA` version 0.92 de l'association de jeunes chercheurs en informatique LOGIN, Université de Nantes. La classe `these-LINA` est disponible à l'adresse :

`http://www.sciences.univ-nantes.fr/info/Login/`

Impression : ma-these.tex – 11/5/2004 – 22:25

Révision pour la classe : $Id: these-LINA.cls,v 1.3 2000/11/19 18:30:42 fred Exp

Ceux qui comprendront ce que je dis
sauront de quoi je parle...

— Prince Norwin, Ambre.

Résumé

La reconnaissance des entités nommées (EN) reste un problème pour de nombreuses applications de Traitement Automatique des Langues Naturelles (TALN). Si cette reconnaissance est réalisée de façon satisfaisante en extraction d'information, pour des textes journalistiques anglais, elle reste en revanche insuffisante, dès lors qu'elle porte sur des textes français, en particulier lorsque l'on souhaite obtenir une catégorisation fine. Conséquemment à une étude linguistique permettant l'émergence de paramètres définitoires opérationnels liés au concept d' « entité nommée », un état de l'art des différents travaux réalisés dans ce domaine et une étude en corpus portant sur la distribution des EN en fonction de leurs caractéristiques graphiques et référentielles, nous présentons **Nemesis**, un système d'identification et de catégorisation des EN pour le français. Ce système s'appuie sur une typologie des EN évolutive, la plus exhaustive et la plus fine possible. Son architecture logicielle se compose principalement de quatre modules (prétraitements, première reconnaissance des EN, apprentissage, seconde reconnaissance) qui effectuent un traitement séquentiel immédiat des données à partir de textes bruts. La reconnaissance des EN est réalisée en analysant leur structure interne et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles de réécriture. Dans cette version minimale, **Nemesis** atteint environ 90 % en précision et 80 % en rappel. Un gain en rappel ne pouvant faire l'économie de la mise en place d'autres techniques, nous proposons donc différents modules optionnels pour faire face à l'incomplétude des lexiques et au passage à de nouveaux corpus : examen d'un contexte encore plus large et utilisation du *Web* comme source de nouveaux contextes. Nous proposons également une étude des conflits engendrés par les règles de réécritures en vu de l'établissement d'un module de désambiguïsation et d'apprentissage de règles.

Mots-clés : noms propres, entités nommées, corpus, identification, catégorisation, mots déclencheurs, règles de réécriture, évidence interne, évidence externe, apprentissage automatique, révision, surcomposition référentielle

Remerciements

Sommaire

Introduction et cadre de la thèse

Motivations

Avec l'augmentation constante de la quantité de documents électroniques disponibles, notamment due à l'avènement du *Web*, scientifiques et industriels s'intéressent de plus en plus aux outils permettant l'extraction de connaissances à partir de corpus de textes.

Le TALN (traitement automatique des langues naturelles) apparaît comme une composante essentielle dans plusieurs grands domaines d'application :

- aide à la rédaction (correction, traduction, résumé et génération automatiques de textes) ;
- filtrage, classification d'information ;
- recherche documentaire ;
- traitement de la parole ;
- dialogue homme-machine.

La reconnaissance des entités nommées (EN, de l'anglais *Named Entity*) est un problème récurrent dans différents domaines du TALN : veille technologique, indexation de documents, modélisation de connaissances, recherche d'information, traduction automatique, Question/Réponse, ou encore résumé de textes automatique [Daille et Morin, 2000]. Cette reconnaissance a été réalisée de façon satisfaisante en extraction d'information (EI, de l'anglais *Information Extraction*), pour des textes journalistiques anglais : la majorité des systèmes de reconnaissance des entités nommées en compétition aux dernières conférences MUC (*Message Understanding Conference*) [MUC-7] ont une précision et un rappel supérieurs à 90 %.

Ces systèmes distinguent deux étapes dans la reconnaissance des entités nommées : l'identification et la délimitation à droite et à gauche de l'entité nommée, et sa catégorisation. Cette reconnaissance constitue, au sein des documents écrits, une tâche considérable pour le traitement automatique des langues naturelles. En effet, les entités nommées sont très présentes dans les textes et lorsqu'elles n'apparaissent pas dans une base lexicale, elles sont considérées par les programmes d'étiquetage morpho-syntaxique ou les analyseurs comme des mots inconnus. Spriet et coll. [1996] montrent qu'un programme stochastique d'étiquetage syntaxique produit un taux moyen de 6 % d'erreurs et que ces erreurs sont dues dans 58 % des cas à la non reconnaissance des entités nommées. De même, une étude Béchet et Yvon [2000] portant sur un corpus de textes issus du journal *Le Monde Diplomatique* a montrée que 72 % des formes inconnues d'un dictionnaire de 256 000 mots sont potentiellement des entités nommées. De plus, Béchet et Yvon [2000] souligne leur contenu informationnel très riche qui est d'une grande utilité pour toutes les applications d'extraction d'information. Enfin, elles sont à la base de la plupart des erreurs de traduction automatique.

Métriques d'évaluation

Les métriques d'évaluation classiquement utilisées en extraction d'information sont les taux de *rappel* et de *précision*. Le rappel mesure la quantité de réponses pertinentes relevées par rapport à la quantité totale de réponses pertinentes, tandis que la précision mesure la proportion de réponses pertinentes parmi les réponses fournies :

$$P = \frac{\text{Réponses correctes}}{\text{Réponses apportées}} \tag{1}$$

$$R = \frac{\text{Réponses correctes}}{\text{Réponses attendues}} \tag{2}$$

Ces métriques correspondent à deux réalités. En effet, comme le rappelle Popescu-Belis [2000] : « De façon générale, lorsqu'un système doit, à partir d'un texte ou d'un ensemble de textes, produire un ensemble de résultats, deux types d'erreurs peuvent se produire : le système peut omettre dans sa réponse des éléments qui auraient dû y figurer ou bien ajouter des éléments qui ne devraient pas y figurer. Les premières erreurs sont dites de *rappel* et les secondes de *précision* ». On peut ainsi évaluer la couverture comme le bruit d'un système.

Pour obtenir un seul score synthétisant ces deux types d'erreurs, la métrique communément adoptée est la *F-Mesure*, introduite par van Rijsbergen [1979], qui représente la moyenne harmonique du rappel et de la précision :

$$F\text{-}Mesure(R,P) = \frac{(\beta+1)\times P\times R}{\beta\times P+R} \text{ ou } 0 \text{ si } P=R=0 \tag{3}$$

L'avantage de cette métrique, par rapport à une moyenne arithmétique, tient à ce qu'elle est plus proche de la valeur la plus faible, et ce d'autant que celle-ci est proche de zéro. Elle pénalise donc les trop grandes inégalités entre les taux de précision et de rappel, et les valeurs proches de zéro.

Le coefficient β permet de paramétrer l'influence du rappel ou de la précision et ainsi avantager les systèmes qui privilégient l'un ou l'autre. En règle générale, ce coefficient est fixé à un, de sorte que la précision et le rappel aient le même poids. Dans ce cas, on obtient la formule appelée P&R :

$$P\&R = \frac{2\times P\times R}{P+R} \text{ ou } 0 \text{ si } P=R=0 \tag{4}$$

Dans la suite de notre document, lorsque nous parlons de F-Mesure sans préciser la valeur du coefficient β, il s'agit en réalité de P&R.

Cadre méthodologique

En extraction d'information, les entités nommées (de l'anglais *named entity*) regroupent traditionnellement les cinq classes de la classification MUC : les noms de personnes, les noms de lieux et les noms d'organisations, auxquels il convient d'ajouter un certain nombre d'expressions numériques (pourcentages, unités monétaires) et temporelles (dates). Dans cette thèse, nous avons décidé de ne pas prendre en compte les entités nommées composant les catégories TIMEX (expressions temporelles) et NUMEX (expressions numériques) définies dans le cadre des conférences MUC, mais de nous consacrer uniquement aux entités nommées, comme on les entend communément (noms de personnes, de lieux, d'organisations, etc.) regroupées sous l'étiquette

ENAMEX pour les conférences MUC. En effet, pour ces deux premières catégories, il existe déjà des outils, qui répondent à la reconnaissance des ces entités avec de très bons taux de rappel et de précision (95 % pour la F-Mesure) ; il n'est donc pas nécessaire d'approfondir cette problématique. De plus, compte tenu des difficultés que soulève la reconnaissance des entités nommées qui ne sont composées d'aucune majuscule, nous avons résolu de ne traiter qu'un nombre très réduit de celles-là. En revanche, toutes les autres, en particulier les mixtes, seront prises en compte.

Dans MUC et MET, une réponse est correcte lorsque la catégorie et les deux bornes (gauche et droite) sont correctement définies. Une réponse est considérée comme à moitié correcte si la catégorie est correcte et si la chaîne de caractères reconnue et la chaîne attendue se chevauchent. Elle peut également l'être si la classe (ENAMEX, TIMEX, NUMEX) et les deux bornes sont correctement définies. Pour nous, une réponse est correcte lorsque l'entité nommée est parfaitement identifiée et catégorisée. Dans tous les autres cas, elle est incorrecte.

Comme nous pourrons le voir, la reconnaissance des entités nommées de la classe ENAMEX est une problématique qui n'est pas encore parfaitement résolue, surtout pour les textes français et en particulier lorsque l'on souhaite obtenir une catégorisation fine. C'est donc sur ces deux points que nous avons choisi de nous positionner. L'objectif de notre travail est donc de créer un système de reconnaissance automatique des entités nommées pour le français, s'appuyant sur une catégorisation fine. En revanche, nous nous limitons aux corpus écrits comportant des informations de casse et de ponctuations (souvent absentes des corpus oraux).

Plan du document

Dans un premier temps, nous présentons dans ce manuscrit une étude des concepts de « nom propre » et d'« entité nommée », visant à faire émerger des paramètres définitoires opérationnels de ce dernier et à définir un principe de sélection des entités nommées en corpus (cf. chapitre 1).

Nous proposons ensuite un état de l'art des différents travaux réalisés en reconnaissance des entités nommées, afin d'en dégager les méthodes à mettre en place pour notre système (cf. chapitre 2).

À la suite de cet état de l'art, nous présentons une double étude en corpus, graphique et référentielle. L'étude référentielle conduit à l'élaboration de notre typologie des entités nommées du français, tandis que l'étude graphique donne à voir les problèmes posés par l'identification et la catégorisation des entités nommées selon leur graphie (cf. chapitre 3).

Conséquemment à ces trois études, nous détaillons l'architecture logicielle de **Nemesis**, notre système d'identification et de catégorisation automatiques des entités nommées dans les textes français. Cette architecture logicielle se compose principalement de quatre modules (prétraitement lexical, première reconnaissance des entités nommées, apprentissage automatique, seconde reconnaissance) qui effectuent un traitement séquentiel immédiat des données à partir de textes bruts (cf. chapitre 4).

Nous évaluons ensuite l'apport de chacun de ces modules, puis les performances de **Nemesis** sur toutes les catégories d'entités nommées. À partir de cette dernière évaluation, nous montrons les limites de **Nemesis** et proposons différents modules permettant d'améliorer le rappel (cf. chapitre 5).

Enfin, nous dressons le bilan du travail réalisé et dégageons les perspectives qu'offre cette thèse (cf. chapitre 6).

Chapitre 1

Noms propres et entités nommées

Le nom propre est omniprésent dans la langue, comme en témoignent sa présence importante dans nos corpus (cf. section 3.3) ou la richesse d'un dictionnaire de noms propres (environ 40 000 entrées pour *Petit Robert des Noms Propres*), presque aussi grande que celle d'un dictionnaire de noms communs (environ 60 000 entrées pour *Petit Robert des Noms Communs*) alors qu'ils sont loin d'être exhaustifs, notamment en ce qui concerne les prénoms et les patronymes (rien qu'un million de prénoms pour les États-Unis). Le nom propre tient depuis longtemps une grande place dans la vie courante, ainsi que dans de nombreux domaines des sciences humaines (la littérature, l'anthropologie, la philosophie, la logique, la sémiotique, la psychanalyse, etc.) et son statut intéresse fortement la linguistique française moderne depuis une trentaine d'années [Le Bihan, 1974 ; Kleiber, 1981 ; Molino, 1982a ; Gary-Prieur, 1991b, 1994 ; Noailly, 1994].

Depuis une dizaine d'année et les travaux de McDonald [1996], la communauté du TALN s'est également beaucoup intéressée aux noms propres et à leur traitement, au point de voir émerger le terme *named entity* – traduit par entité nommée en français – pour désigner des entités issues des noms propres, mais dans une acception plus large.

L'objectif de ce chapitre est de faire émerger des paramètres définitoires opérationnels liés au concept d'« entité nommée » afin d'obtenir une démarche de sélection des candidats au statut d'entité nommée. Pour ce faire, nous partons d'une étude lexicographique de « nom propre », puis nous étudions les fondements typologiques et linguistiques de l'intuition de « nom propre », afin d'en restituer la nature et les origines. Ensuite, à partir d'une étude des différentes typologies des entités nommées, nous situons ces dernières dans le continuum noms propres/noms communs. Enfin, nous érigeons les paramètres linguistiques caractérisant les entités nommées, en sélectionnant les plus saillants parmi ceux constitutifs du nom propre, et définissons un principe de sélection des entités nommées en corpus.

1.1 Le nom propre : concept général

Il est d'usage quand on cherche la définition d'un terme, de se référer prioritairement au discours lexicographique. L'objectif de cette démarche n'est pas de lister exhaustivement les différentes significations de « nom propre », mais plutôt de préciser au fur et à mesure du chapitre les paramètres définitoires retenus pour cette étude. Dans cette optique, nous recourrons aux grands dictionnaires de la langue française[1] (*Grand Larousse Universel*, *Grand Robert de la Langue Française*, *Trésor de la Langue Française*) destinés à un public cultivé, qui maîtrise très bien la variante la plus valorisée (la norme) de sa langue maternelle, et aux dictionnaires d'ampleur moyenne, destinés à un large public (*Petit Robert*, *Petit Larousse*).

[1] Pour une taxinomie des usages du dictionnaire nous nous référons ici à Girardin [1979].

Grand Larousse Universel (GLU)

Gramm. *Nom propre*, sous-catégorie du nom, désignant un être ou un objet considérés comme uniques, par opposition au nom commun (par ex., *Jacques*, *Bonaparte*, *Paris*).

Grand Robert de la Langue Française (GRLF)

– NOM* PROPRE (opposé à *nom commun*, ainsi qu'aux autres mots de la langue). *Jean*, *Paris*, *les Français sont des noms propres. Les noms propres prennent une majuscule en français. Pluriel** (infra cit. 3) *des noms propres* (aussi Exploitation, cit. 8 ; extrêmement, cit. 5 ; personne, cit. 37). *Noms communs issus de noms propres. Description d'un nom propre. Dictionnaire de noms propres.*

Trésor de la Langue Française informatisé (TLFi)

LINGUISTIQUE

– *GRAMM. Nom propre* (p. oppos. à *nom commun*). V. *nom* III. *Syntaxiquement, les noms propres présentent des propriétés particulières ; ils sont autodéterminés, ce qui entraîne souvent l'absence d'article défini dans l'emploi courant* (Jean, Dupont, Paris) *ou bien la présence obligatoire du seul article défini* (le Brésil, la France) (*Ling.* 1972).

Petit Robert (PR)

NOM PROPRE (opposé à *nom commun*, ainsi qu'aux autres mots de la langue) : nom qui s'applique à un individu, un objet unique, une réalité individuelle qu'il désigne (alors que le nom commun correspond à une classe, une idée générale, un sens). *Jean, Napoléon, Paris, O.N.U, France, Louvre sont des noms propres. Les noms propres prennent une majuscule. Dictionnaires de noms propres.*

Petit Larousse (PL)

Nom propre, nom qui ne peut s'appliquer qu'à un seul être ou objet ou à une seule catégorie (par oppos. à *nom commun*).

1.1.1 Définition du nom propre

Les traits communs aux définitions présentées ci-dessus – qui opposent toutes le nom propre au nom commun – s'articulent autour de trois composantes : grammaticale, référentielle et graphique.

L'aspect grammatical est présent dans les cinq définitions sous la forme « par opposition au nom commun ». Le nom propre devient une catégorie grammaticale au même titre que les noms communs, les articles ou les verbes.

Certaines définitions accentuent cet aspect définitoire et donc cet effet de catégorisation grammatical : GLU « sous-catégorie du nom », GRLF/PR « [opposé à NC] ainsi qu'aux autres mots de la langue », TLFi « ils sont autodéterminés, ce qui entraîne souvent l'absence d'article défini dans l'emploi courant ou bien la présence obligatoire du seul article défini ».

L'aspect référentiel se retrouve dans trois des cinq définitions présentées sous les formes : GLU « désignant un être ou un objet considérés comme uniques », PR « nom qui s'applique à un individu, un objet unique, une réalité individuelle qu'il désigne », PL « nom qui ne peut s'appliquer qu'à un seul être ou objet ou à une seule catégorie ».

Il est à noter que pour l'aspect référentiel, les informations recueillies présentent un caractère duel :

- quant à la nature des référents possibles : individu, être, objet, catégorie, réalité ;

- quant au principe qui génère les noms propres : unicité du référent ;

l'aspect graphique demeure dans deux des cinq définitions présentées sous un même aspect : GRLF/PR « Les noms propres prennent une majuscule (en français) ».

Nous constatons, pour cette étude lexicographique, qu'il n'y a aucune opposition ou contradiction entre les différentes définitions en présence, sinon par omission – par la présence de tel élément dans un article, non mentionné dans un autre.

De plus, parmi ces définitions, seul l'aspect grammatical – selon lequel le nom propre s'oppose au nom commun – est commun aux différentes définitions. Par conséquent, nous pouvons penser qu'il existe une sorte de définition indirecte du nom propre, par ce qu'il n'est pas, c'est-à-dire une forme de négation du nom commun. Nous nous proposons donc d'effectuer la même analyse lexicographique pour le nom commun.

1.1.2 Définition du nom commun

Grand Larousse Universel (GLU)

Gramm. *Nom commun*, substantif qui désigne un être ou une chose considérés comme appartenant à une catégorie générale, par opposition au nom propre (par ex., *enfant*, *chien*, *maison*, *courage*, etc.).

Grand Robert de la Langue Française (GRLF)

Gramm. Nom commun, qui appartient à tous les individus de la même espèce (opposé au nom propre) → *Appellatif. Nom commun masculin, féminin, épicène.*

Trésor de la Langue Française informatisé (TLFi)

– *GRAMM. Nom commun* (p. oppos. à *nom propre*). Nom qui convient à plusieurs êtres ou choses formant un genre, une espèce.

Petit Robert (PR)

LING.

NOM COMMUN : nom de tous les individus de la même espèce (opposé à *propre*). *Nom commun masculin, féminin. « Chat », « table » sont des noms communs.*

Petit Larousse (PL)

GRAMM. *Nom commun*, qui s'applique à un être, une chose considérés comme appartenant à une catégorie générale, par opposition à *nom propre.*

Les traits communs aux définitions présentées ci-dessus s'articulent autour de deux composantes seulement (grammaticale et référentielle) :

l'aspect grammatical est présenté dans les cinq définitions sous deux formes. Premièrement, l'entrée se fait par « GRAMM. » (ou « LING. ») pour chaque article, ce qui confère explicitement à « nom commun » le statut de catégorie grammaticale (sous-catégorie du nom). Deuxièmement, en terme de contenu, toutes les définitions mentionnent « [par opposition/opposé] [à/au] nom propre » ;

l'aspect référentiel se retrouve également dans toutes les définitions sous la forme : « plusieurs [êtres/choses/individus] [formant/appartenant à/de] [une catégorie générale/la même espèce/un même genre] ». Là encore, les informations recueillies présentent un caractère duel :

- quant à la nature des référents possibles : être, chose, individu ;

- quant au principe qui génère les noms communs : formant/appartenant à un(e) genre/catégorie générale/espèce.

Nous constatons, à l'issue de cette analyse lexicographique, une grande cohérence des définitions entre elles, eu égard à la redondance des informations recueillies.

Toutefois, nous pouvons déjà regretter le caractère infructueux de notre démarche. En effet, nous l'aurons remarqué, il y a, du point de vu grammatical, une circularité totale des définitions en présence (nom propre = par opposition à nom commun ; nom commun = par opposition à nom propre). Cela nous empêche d'accéder ne serait-ce qu'à des éléments définitoires de « nom propre ».

Au terme de cette analyse et des sources exploitées, nous concluons donc à une impossible définition de « nom propre », tant autonome que par opposition à « nom commun ». Pourtant, il existe une intuition manifeste de la notion de nom propre, ce que nous proposons de montrer ci-dessous.

1.2 L'intuition de « nom propre »

Aussi peu scientifique que cela puisse paraître, il existe indéniablement une intuition du nom propre, c'est à dire une possibilité de percevoir, de comprendre, de connaître, de convoquer clairement ce concept, sans avoir parcouru préalablement les étapes de la réflexion, du raisonnement et de l'analyse propres à le faire émerger.

Sans chercher à remettre en cause l'existence de cette intuition, il semble légitime d'en rechercher la nature et les fondements, dans l'espoir d'isoler ainsi des paramètres définitoires constitutifs du concept de « nom propre ».

Cette intuition de « nom propre » revêt un caractère double :

1. il s'agit d'abord d'une intuition de nature typologique, qui découle directement de l'impossibilité de définir le concept. Molino [1982b] considère que « Plutôt que d'en donner une définition *a priori*, mieux vaut offrir une géographie du nom propre, c'est-à-dire présenter et classer tous les candidats au rang de N.P. » ;
2. il s'agit ensuite d'une intuition linguistique fondée simultanément sur la notion de « nom propre prototypique » et sur l'opposition entre nom propre et nom commun. Gary-Prieur [1991a] soutient cette idée par ces mots : « pour le français en tout cas, tout locuteur adulte a une intuition claire de la différence entre nom propre et nom commun ».

1.2.1 Typologie du nom propre

Nous avons relevé deux typologies différentes pour les noms propres : la première, lexicographique, est réalisée par Alain Rey et se trouve dans la préface du *Petit Robert des Noms Propres*, tandis que la deuxième, linguistique, est donnée par Molino [1982b].

Pour Rey [2003] : « les *noms propres* [...] désignent des individus ou des réalités individuelles. Celles-ci ne sauraient être définies ; on peut seulement les décrire ». Il ajoute que : « L'opposition entre la description des mots d'une langue, celle des notions et celle des choses, réunies par les mots en classes, est loin d'être franche ». C'est pour ces raisons qu'il présente une nomenclature du *Petit Robert des Noms Propres*, qui comporte six classes :

les noms de personnes rassemblent les noms de famille et les prénoms ;

les noms de lieux admettent, d'une manière générale, toute unité géographique notable, chaîne montagneuse ou sommet, fleuve, baie ou mer, presqu'île, détroit, archipel ;

les œuvres regroupent des entités culturelles, des « œuvres » :

- œuvres littéraires ;
- textes non littéraires (lois, codes, textes religieux, ouvrages scientifiques et techniques) ;
- œuvres musicales et artistiques ;
- thèmes iconographiques qui servent de titres à des œuvres importantes (*annonciation*, *Crucifixion*) ;
- œuvres de cinéma ;
- titres de la presse mondiale (*The Times*, *Libération*).

Rey [2003] précise que « le lecteur trouvera des entrées (imprimées en minuscules, et non en capitales, à la différence des noms de personnes et de lieux) comportant plusieurs « mots » graphiques, des expressions formant des titres, et intégrant ou non des noms propres (*À la recherche du temps perdu* ; *Monsieur Teste*) » ;

les évènements, les périodes de l'histoire sont le plus souvent désignés par des noms communs et renvoient à des évènements ou une série d'évènements bien précis (p. ex. *Révolution française*, *Guerre mondiale*, *Collier (affaire du)*, *Dreyfus (affaire)*, *mai 1968*, *Thermidor*, *Renaissance*, *Paléolithique*) ;

les collectivités, groupes, institutions concernent des « noms propres par destination » et peuvent correspondre à des expressions complexes (p. ex. *groupe des Cinq*, *Cavalier bleu*, *Organisation des nations unies*, *Compagnie de Jésus*, *Bibliothèque nationale de France*) ;

les articles encadrés sont des noms communs qui assurent une homogénéité de traitement à des faits historiques comme la *résistance* ou la *collaboration*), à des mouvements artistiques ou littéraires comme le *baroque* ou le *naturalisme*), à des doctrines philosophiques, religieuses, politiques comme l'*anarchisme* ou le *bouddhisme*, par rapport à ceux qui sont désignés comme des noms propres.

Rey [2003] note que l'on trouve souvent les éléments de la classe « articles encadrés » avec une majuscule à l'initiale, ce qui les apparente totalement à ceux de la classe « évènements, périodes de l'histoire ».

La typologie linguistique proposée par Molino [1982b] a ceci de particulier qu'elle est donnée comme exhaustive. Elle permet de « classer tous les candidats au rang de N.P., soit tous les termes qui ont pu être un jour considérés comme appartenant à la catégorie ; il s'agit donc d'une liste maximale ». Il se base sur les travaux de Zabeeh [1968] et Le Bihan [1974] pour distinguer neuf catégories :

les noms de personnes ou anthroponymes : *Jean*, *Homère*, *Reagan*, etc. ;

les noms d'animaux : *Médor*, ou tout autre nom qu'une personne aurait donné à un animal ;

les appellatifs et titres : *Papa*, *Maman*, etc. ;

les noms de lieux : *Paris*, *Aix-en-Provence*, *la France*, *La Normandie*, etc. ;

les noms de temps : *midi*, *lundi*, *septembre*, *Pâques*, *la Renaissance*, etc.

les noms d'institutions : *Renault*, *la C.G.T.*, etc. ;

les noms de produits de l'activité humaine : *la 5^e Symphonie*, *Madame Bovary*, *Concorde*, etc. ;

les noms de symboles mathématiques et scientifiques : *pi*, etc. ;

les autres noms propres : pour Molino [1982b] : « tout peut, dans certaines circonstances et pour un public donné, recevoir un nom propre ».

Nous nous intéressons maintenant au deuxième pan de notre analyse : l'intuition linguistique de « nom propre ».

1.2.2 Le statut du nom propre en linguistique

Les entités pour lesquelles l'intuition de « nom propre » est la plus évidente sont les noms de personnes et les prénoms (*Jean*, *Dupond*, etc.). En effet, il est dit de ces deux classes qu'elles sont les prototypes du nom propre.

La notion de prototype, apparue dans les année 1970, a été théorisée par G. Kleiber dans le cadre de la sémantique du prototype [Kleiber, 1990]. La définition proposée par Kleiber [1990] établit que le prototype est « le meilleur exemplaire ou encore la meilleure instance, le meilleur représentant ou l'instance centrale d'une catégorie ». Il s'agit donc du meilleur candidat « *communément* associé à une catégorie », pour reprendre les propos de l'auteur. Le terme « *communément* » revêt ici une grande importance : il reflète l'aspect intuitif du concept dans la mesure où l'instance est considérée comme prototypique à l'unique condition qu'il y ait consensus, parmi les sujets (locuteurs) interrogés, pour la juger comme la plus représentative de la catégorie. Pour la catégorie *fruit*, par exemple, les sujets interrogés par Rosch [1973] ont donné la *pomme* comme prototype et l'*olive* comme le moins bon exemplaire ; pour la catégorie *oiseau*, le prototype est le *moineau*, tandis que le *pingouin* en est un moins bon représentant.

Ces exemples de « moins bon exemplaire » permettent d'inclure l'idée selon laquelle l'appartenance à une catégorie s'effectue sur la base du degré de similarité avec le prototype de cette catégorie. Cette notion de degré de similarité avec le prototype suggère une continuité, une gradualité et une hiérarchie entre les différents candidats d'une catégorie.

Traditionnellement, la sémantique du prototype permet de catégoriser des noms communs selon des paramètres sémantiques. Or, dans le cadre du nom propre, la notion de meilleur candidat « communément associé à une catégorie » subsiste, mais la hiérarchisation des entités ne s'applique plus en terme de catégories sémantiques, mais en terme de catégories grammaticales.

Concernant les noms propres Gary-Prieur [1991a] affirme que « [les] objets typiques de cette catégorie sont, de l'avis général, les noms de personnes, sur lesquels l'intuition d'une opposition entre nom propre et nom commun est très claire (*Pierre*/*pierre*, *le Havre*/*un havre*) ».

La nature prototypique de l'intuition du nom propre conduit donc à juger de l'appartenance d'un nom à la catégorie des noms propres – par opposition au nom commun – en fonction de son degré de similarité avec les noms de personnes, communément admis comme prototypiques du nom propre. Molino [1982b] écrit : « plus un nom aura un comportement qui se rapproche du prototype qu'est le prénom ou le nom de famille, plus il sera ressenti comme nom propre. Mais l'analyse linguistique ne doit pas en rester à cette échelle, et elle doit décrire avec la plus grande précision la totalité des traits qui caractérisent l'ensemble flou du nom propre ». Conformément aux propos de Molino [1982b], et dans la perspective de restituer les fondements du nom propre et d'en isoler les caractéristiques, nous allons maintenant décrire, d'un point de vu linguistique, la totalité des traits qui caractérisent cet « ensemble flou ». Or, en linguistique

comme en lexicographie, le statut du nom propre est fondé sur son opposition au nom commun. En effet, dès les origines de la grammaire occidentale, on a opposé ces deux catégories, comme Donat dans sa grammaire latine – *Donatus, de partibus orationis ars minor* (IV^{e} siècle après Jésus-Christ) : « En quoi consiste la qualité du nom ? Elle est double : ou il est le nom d'un seul et est appelé nom propre, ou il est le nom de plusieurs et il est appelé commun » [2].

Nous présentons maintenant les critères linguistiques qui régissent cette opposition, en testant la validité des caractéristiques repérées par la recherche systématique de contre-exemples.

1.2.2.1 Critères graphiques

Les noms propres possèdent des caractéristiques graphiques particulières, en terme d'orthographe, mais surtout en terme de casse.

Bien qu'une orthographe communément admise existe pour les noms propres, des variations orthographiques sont néanmoins acceptées pour certains d'entre eux :

Christelle, *Kristelle*, *Kristell*, *Christèle*.

New York, *New-York*.

Kassel, *Cassel*.

Ce critère demeure néanmoins très peu discriminant, car il ne touche qu'une très faible proportion des noms propres.

Le critère graphique principal concerne la casse de l'initiale des noms propres. En effet, ils commencent en règle générale par une majuscule. Cependant, ce critère, bien que souvent donné comme définitoire (pour Grevisse [1993] : « Les noms propres prennent toujours la majuscule. »), n'est pas un critère absolu de distinction entre nom propre et nom commun. Si les noms propres prototypiques (les prénoms et les patronymes) requièrent toujours une majuscule à l'initiale (*Jean*, **jean*, *Dupond*, **dupond*), il apparaît que tous les noms requérant une majuscule à l'initiale ne sont pas considérés comme des noms propres, car se comportant linguistiquement comme des noms communs :

les Italiens, **les italiens*.

D'autre part, certains noms communs peuvent admettre une majuscule à l'initiale – sans qu'elle soit obligatoire – car il s'agit de noms abstraits ayant tendance à fonctionner comme les noms propres prototypiques :

le surréalisme, *le Surréalisme*.

la révolution française, *la Révolution française*.

Cela explique – comme le souligne Gary-Prieur [1991a] – que « l'attribution d'une majuscule à un nom commun le fait facilement interpréter comme un nom propre », comme en témoignent ces exemples extraits de *L'échiquier du mal* (Dan Simmons) :

Et elle avait Festoyé plus souvent (p. 21).

un Noir gigantesque (p. 26).

Willi l'appelait la Chasse (p. 28).

[2] Traduction du latin « *Qualitas nominum in quo est ? Bipertita est : aut enim unius nomen est et proprium dictiur, aut multorum et appelativum* » réalisée par Molino [1982b].

Malgré tout, ce critère constitue une des caractéristiques principales de l'intuition de l'opposition entre nom propre et nom commun. D'ailleurs, n'apprend-on pas aux enfants, dès le CE1, que « Le nom propre s'écrit avec une majuscule » [3].

Ce dernier critère graphique demeure un critère d'identification possible, puisque la plupart des noms propres prennent une majuscule à l'initiale. Cependant, il ne saurait être un critère absolument discriminant, puisque certains noms communs – ou noms se comportant comme tels – prennent également une majuscule. Cette affirmation est confirmée par les éléments appartenant à la nomenclature du *Petit Robert des Noms Propres* (cf. section 1.2.1), notamment ceux des quatre dernières classes (les œuvres ; les évènements, les périodes de l'histoire ; les collectivités, groupes, institutions ; les articles encadrés).

1.2.2.2 Critères morphologiques

Selon Molino [1982b] : « Dans une langue comme le français ou l'anglais, il n'y a pas de caractéristique morphologique valable pour l'ensemble des noms propres ». En effet, si la plupart des noms propres suivent certains critères morphologiques, ces derniers ne s'appliquent qu'à des sous-ensembles de noms propres. Pour Gary-Prieur [1991a] : « la plupart des noms propres n'ont pas de genre déterminé, ne prennent pas la marque du pluriel, et s'emploient sans déterminant ». Nous avons relevé quatre caractéristiques morphologiques de natures différentes[4].

La première concerne le mode de formation particulier des diminutifs chargés affectivement (formes hypocoristiques). Contrairement aux diminutifs des noms communs (p. ex. *télévision*/*télé*), les formes hypocoristiques ne sont pas nécessairement tronquées et ne conservent pas toujours la même structure :

Anne, *Nanou*. *Philippe*, *Philou*.

La seconde caractéristique touche les phénomènes de dérivation, qui génèrent essentiellement des adjectifs relationnels issus des anthroponymes et des toponymes, ainsi que des noms d'habitants issus des toponymes, également appelés gentilés :

balzacien, *homérique*, *gaulliste*, *italien* → adjectifs relationnels.

Français, *Francomtois*, *Européen* → adjectifs relationnels.

Cex : *pasteuriser*, *italianisant* → verbe du 1er groupe, adjectif.

Les deux dernières caractéristiques portent sur les flexions en genre et en nombre :

- Comme les noms communs, certains noms propres, notamment les prénoms, subissent la flexion en genre :

  Germain, *Germaine*.

- Comme les noms communs, certains noms propres subissent la flexion en nombre lorsqu'ils prennent un déterminant[5] :

  les Dupond(s), *les Bourbon(s)-Orléans*.

Sur le plan morphologique, il semble donc très délicat de dégager une caractéristique linguistique stable participant de l'intuition de la distinction entre nom propre et nom commun, tant

[3] Extrait de la définition du nom propre dans le manuel de CE1 *La balle aux mots* aux éditions *Nathan*.

[4] Les noms propres partageant les deux premières caractéristiques s'éloignent du fonctionnement des noms communs ; au contraire, ceux qui partagent les deux dernières s'en approchent.

[5] À noter, que dans ce cas, on ne sait pas si l'on doit mettre un « s » ou pas ?

les faits présentés sont ambivalents. En revanche, Eggert et coll. [1998] ont étudié la dérivation morphologique d'un toponyme en un nom d'habitants et en ont déduit que pour la plupart des toponymes, les gentilés dérivés sont réguliers et certains possèdent un suffixe particulier aux dérivations de toponymes (p. ex. *-ois*, *-ais*). Ce critère n'est pas suffisant car il se limite à une partie des toponymes et ne caractérise pas les noms propres, mais il est intéressant dans le cadre de la mise en place de règles morphologiques permettant la reconnaissance automatique des noms d'habitants (cf. section 4.2.2.2).

1.2.2.3 Critères syntaxiques

L'élément fondamental ici concerne la présence ou l'absence de déterminant, et le caractère contraint ou non de ce dernier. Dans ce cadre, nous distinguons les noms propres modifiés et les autres, puisque leurs comportements diffèrent. En effet, en prenant un déterminant et/ou des modificateurs (adjectifs, complément du nom, etc.), les noms propres sont dits « modifiés » et acquièrent un aspect essentiel du nom commun, puisqu'ils fonctionnent comme des termes généraux qui présupposent l'existence de classes référentielles comportant plus d'un membre.

Nous avons donc relevé les constructions syntaxiques caractéristiques des noms propres.

Les noms propres non modifiés

Les noms propres prototypiques ont la caractéristique, contrairement au nom commun, d'être dépourvus de déterminant et de modificateur en position référentielle[6] :

Alain est mon frère, **le Paul est arrivé*, **je connais bien le Paris.*

Le Bihan [1974] révèle une autre construction spécifique, très productive pour certains noms propres et ne tolérant aucun nom commun[7] :

Alger la blanche, *Alexandre le grand*, *Pépin le bref*, **voiture la blanche.*

Gary-Prieur [1991a] souligne l'idée que « seule est reconnue comme typique du nom propre sa construction sans déterminant », en ajoutant en note de bas de page : « Sauf bien sûr pour les noms propres qui ont ce qu'on pourrait appeler un « déterminant lexical » ». Certains noms propres sont donc accompagnés d'un article défini (le, la, les) à l'exclusion de tout autre déterminant :

la Seine, **une Seine.*

Cex : *une Apocalypse* (extrait des *Neufs Princes d'Ambre* de Roger Zelazny).

Ces noms propres sont désignés par Gary-Prieur [1994] en termes de « noms à articles définis lexical ». Il s'agit :

- des noms propres toponymiques autres que ceux des villes (pays, régions, fleuves, etc.) :

  le Pérou, *la Seine*, *l'Asie*, *les Vosges.*

  Cex : *Cuba*, *Israël.*

- des noms de restaurants, de bateaux :

  le Montesquieu, *le Queen Mary II.*

[6] En position non référentielle, les noms communs peuvent également s'employer sans déterminants (p. ex. *je serai* [*champion du Monde*]).

[7] Il est intéressant de noter que, dans une telle construction, la casse de l'initiale de l'adjectif n'est pas universelle (*Alger la blanche* ou *Alger la Blanche*).

- des noms de cantatrices, d'actrices :

  la Callas.

  Cex : **la Crespin* (pour *Régine Crespin*).

- de certains patronymes pour lesquels on parle d'articles intégrés :

  Legrand, *Letourneur*.

- des noms propres dont l'article appartient à la morphologie du nom propre et ne dispose à ce titre d'aucune autonomie. Il est introduit par une majuscule, comme le serait un nom propre composé :

  Le Corbusier, *Le Mans*, *La Fontaine*.

  Seuls certains toponymes précédés d'une préposition autorisent la contraction avec l'article défini :

  des nouvelles du Caire, *les vingt-quatre heures du Mans*.

  Cex : *une architecture de Le Corbusier*.

Nous voyons donc que, parmi les noms propres non modifiés, il existe de nombreuses constructions spécifiques, ainsi qu'une très grande hétérogénéité des comportements syntaxiques.

Les noms propres modifiés

Bien que prenant un déterminant et d'éventuels modificateurs, ils ne perdent jamais leur statut de nom propre. Nous distinguons quatre types d'emplois pour les noms propres modifiés :

- les emplois dénominatifs, dans lesquels le nom propre renvoie à la classe de ceux qui le portent :

  Il n'y a pas de Huguette au numéro que vous demandez.

- les emplois exemplaires, dans lesquels le nom propre désigne le porteur comme représentant typique d'une classe d'individus analogues partageant une caractéristique commune :

  Reste qu'un José Bové serait plus crédible que Philippe de Villiers (Libération).

- les emplois « fractionnés », dans lesquels le modificateur opère une scission ou une division de l'individu porteur du nom propre ou d'une période de son existence :

  Le Hugo de 1825 ne vaut pas celui de la vieillesse.

- les emplois « dérivés », dans lesquels le référent n'est ni le porteur du nom, ni une « portion » du porteur du nom, mais une entité unie à ce porteur. Ces emplois dérivés peuvent prendre trois formes différentes :
  - l'ellipse :

    Il a acheté une (voiture) *Renault*. **Il a acheté un(e) Renault* (l'entreprise industrielle).

    Dans ce cas, tous les déterminants sont possibles et l'adjonction de modificateurs fréquente :

    la/les/une/des Renault(s) rouges.

  - la métonymie :

    Il écoute du (de la musique de) *Mozart*.

  - la métaphore :

un machiavel (un homme d'État sans scrupule).

Les noms propres modifiés peuvent donc apparaître dans des constructions syntaxiques identiques à celles des noms communs.

Si plusieurs comportements syntaxiques spécifiques caractérisent la catégorie des noms propres, chacun de ces comportements n'est pas applicable à l'ensemble de la catégorie de nom propre. Pour Kleiber [1981], « il n'est guère possible de définir de façon satisfaisant la catégorie syntaxique du nom propre uniquement d'après leur comportement avec les déterminants ». Gary-Prieur [1991a] confirme : « Leur syntaxe ne se limite pas à l'absence de déterminant ».

Gary-Prieur [1991a] ajoute enfin : « L'absence de statut syntaxique bien défini pour l'opposition entre N_P et N_C explique sans doute que beaucoup de linguistes ont tendance à parler de « communisation » du nom propre chaque fois que sa construction le rapproche du nom commun ».

Comme pour l'aspect morphologique, il est difficile, sur le plan syntaxique, de dégager ne serait-ce qu'une caractéristique stable de l'opposition entre nom propre et nom commun.

1.2.2.4 Critères sémantiques

Nous nous intéressons, dans cette section, au problème de la signification[8] des noms propres. Or, s'intéresser à la signification des noms propres nécessite prioritairement d'interroger le concept de référence, ce que constate Molino [1982b] : « Le point de départ de l'analyse est généralement fourni par la constatation suivante : les noms propres sont des expressions qui réfèrent uniquement, c'est à dire qui renvoient à une entité particulière, considérée comme un « individu » singulier ». Cette caractéristique du nom propre est rarement discutée. Nous avons décidé de nous en tenir à l'aspect uniquement sémantique de la référence et n'aborderons pas sa perspective logique, bien que celle-ci en découle naturellement. En effet, discuter de la dimension logique de la référence amène à des considérations propres à la nature de la référence, ce qui nous éloignerait de notre propos tant les théories en présence sont nombreuses et contradictoires (cf. S. Kripke, B. Russel, G. Frege ou encore J. S. Mill). En revanche, la problématique sémantique se justifie car elle nous apporte, par le concept de référence, une vision duelle de la nature de la signification des noms propres pour laquelle deux thèses s'opposent :

1. **les noms propres n'ont pas de signification**. En effet, il semble impossible de leur donner une définition autre que métalinguistique ; il est nécessaire de convoquer une entité intermédiaire permettant l'accès à une définition plus générale, caractérisant la nature du nom propre :

   Jean est un prénom, *Paris est une ville.*

   De même que pour l'association entre le signifiant et le signifié du nom commun, la relation qui unit le référent et le nom propre peut être qualifiée d'arbitraire, puisque l'implication réciproque entre ces deux éléments n'est pas fondée sur une correspondance naturelle entre la forme du référent et les traits définitoires du nom propre. On dit donc que le référent et le nom propre sont conventionnellement associés dans la langue.

   Selon Grevisse [1993], « 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 ».

[8] Nous ne faisons pas de différence entre signification et sens, bien qu'ayant connaissance de l'importance théorique que celle-ci revêt : nous emploierons donc indifféremment ces termes, dans leur acception courante et non linguistique.

Enfin, bien que, à l'instar des noms communs, ils désignent des personnes, des objets, des lieux, les noms propres n'entretiennent pas de relations sémantiques (synonymie, hyponymie, antonymie). Cependant, il arrive que certains noms propres soient lexicalisés :

un frigidaire, *un bic*, *un kleenex*.

Dans ces cas, ces noms sont utilisés comme synonymes de réfrigérateur, stylo à bille et mouchoir. Toutefois, ce qui s'est produit pour la marque *Frigidaire* ne s'est pas produit pour la marque *Renault* : une Renault n'est pas synonyme d'une voiture quelconque [Rey-Debove, 1994] ;

2. **les noms propres sont les entités les plus signifiantes**, parce que les plus individuelles, les plus personnelles. Il s'agit d'entités singulières : dire *Jacques est venu*, c'est viser un individu unique, même s'il existe des centaines de personnes qui portent ce prénom.

   De plus, les noms propres ont davantage de propriétés sémantiques que les noms communs. Non seulement *Jacques* prend un sens différent chaque fois qu'il est employé (seul le contexte permet de le déterminer), mais, de surcroît, ce sens est plus spécialisé dans chacun de ces emplois. Un nom propre convoque donc un plus grand nombre de caractéristiques particulières que celles mobilisées par un nom commun.

   Enfin, les noms propres sont les termes les plus chargés affectivement (p. ex. *Kiki*, *Loulou*, *Philou*). Molino [1982b] explique : « Il [le nom propre] induit une série indéfinie d'interprétants plus riches et plus chargés d'affectivité que ne le sont les interprétants des noms communs, comme l'indique la fonction littéraire et poétique des noms propres ».

Sur l'aspect sémantique du nom propre, la distinction avec le nom commun est manifeste et semble même fondée. Le critère sémantique du référent unique est largement partagé par les noms propres, indépendamment de la théorie sémantique considérée. En effet, il ne fait aucun doute que les noms propres prototypiques possèdent un référent unique. En revanche, pour des sujets comme *Renault 5* ou *Parisiens*, cela devient discutable. En fait, ces entités n'ont pas de référent unique en terme d'individu, mais en terme de classe. Par conséquent, elles ont un statut controversé, tantôt nom propre pour certains, tantôt nom commun pour d'autres.

Concernant les critères retenus par chacune des deux théories présentées, ils paraissent plus fiables, mais le problème se déplace cette fois sur l'intuition elle-même. En effet, ces deux théories sur la signification des noms propres participent d'une intuition qui revêt deux facettes contradictoires et difficilement réconciliables.

L'intuition de nom propre est donc fortement liée au critère du référent unique. Cependant, ce critère reste flou et ne saurait être suffisant pour distinguer tous les noms propres des noms communs.

1.2.2.5 Critères pragmatiques

Pour beaucoup de noms propres l'attache à un référent unique n'est assurée que dans la situation d'énonciation où ils sont employés. Comme le souligne Molino [1982b] : « la propriété sémantique de référence unique est comme soulignée et doublée par les propriétés pragmatiques ».

En situation de communication, il y a une mise en correspondance d'un nom propre avec un individu, un lieu, etc. :

Je te présente Jacques, *Voilà Paris*, *Jacques est venu*.

Dans ce même type de situation, les noms propres prototypiques peuvent servir à interpeller ; dans ce cas, ils figurent souvent en apostrophe :

Toi, tais-toi ! Jacques, tu sors ! Fourour, un mojito !

Dans cet emploi « interpellatif », les autres noms propres sont construits rhétoriquement. Les noms géographiques, par exemple, peuvent intervenir dans une synecdoque[9] :

Ici Paris, à vous Boulogne !

Parce qu'ils sont circonscrits à une situation de communication, ces critères ne sauraient être suffisants pour distinguer un nom propre en toute circonstance. Cependant, tout comme en sémantique, ce critère du référent unique en pragmatique gouverne fortement l'intuition de nom propre.

1.2.3 Synthèse

Au vu des nomenclatures des noms propres de Molino [1982b] et Rey [2003], il apparaît clairement :

- qu'il est tout à fait possible de classer les noms propres, d'en proposer une typologie exhaustive[10] ;
- que cette typologie du nom propre nous donne une idée des candidats pouvant être considérés comme tels ;
- que l'émergence des catégories se fait sur la base d'une classification des différents types de référents ;
- que le concept de « nom propre » réunit en son sein de nombreux éléments de natures différentes, ce qui explique que les analyses linguistiques parviennent difficilement à couvrir le sujet ;
- que le fait de classer les noms propres ne les rend pas plus faciles à définir, comme en témoigne la neuvième et dernière catégorie de la nomenclature de Molino [1982b]. ».

Ne pouvant définir le nom propre, nous avons alors cherché la nature et les fondements de l'intuition linguistique à l'origine du nom propre, notamment par un examen systématique de ses caractéristiques linguistiques. Nous en concluons :

- que chaque critère évoqué est un paramètre constitutif du concept « nom propre ». Toutefois, s'il est nécessaire à un sujet de partager un minimum de ces caractéristiques pour être considéré comme étant un nom propre, il n'y en a pas une qui soit suffisante et valable pour l'ensemble des noms propres. De fait, il n'y a aucun critère qui permette de distinguer sans ambiguïté le nom propre du nom commun ;
- que ces critères ne sont pas nécessairement convergents et donc cumulables, car pas de même nature. En effet, ne serait-ce que sur l'aspect sémantique, un sujet ne pourrait pas réunir à la fois les caractéristiques des deux théories : beaucoup de signification et aucune signification ;
- que plus un sujet partagera de ces paramètres constitutifs du nom propre, plus il s'approchera du nom propre prototypique : un sujet qui prendrait une majuscule à l'initiale, qui n'admettrait aucun déterminant, aucun modificateur, qui n'aurait pas de genre déterminé, qui ne prendrait pas de marque du pluriel, et qui référerait une entité unique, un tel sujet serait certainement un nom propre prototypique.

[9] La synecdoque est un cas particulier de métonymie, dans lequel on prend le tout pour la partie ou la partie pour le tout.

[10] On notera tout de même le prix à payer pour cette exhaustivité : la dernière catégorie n'a d'autre caractéristique que d'être composée des autres noms propres qui peuvent inclurent n'importe quoi.

1.3 Un continuum noms propres/noms communs

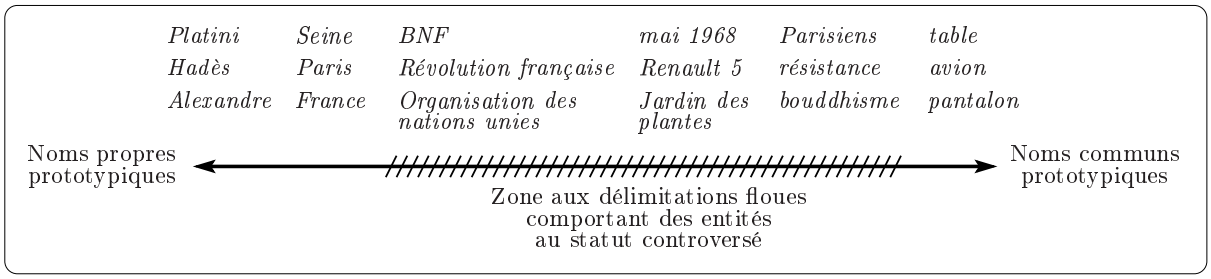


Figure 1.1 – Continuum noms propres/noms communs

Conformément à nos attentes, nous avons pu restituer la nature et les fondements de l'intuition linguistique du nom propre. Se faisant, nous avons abouti à un faisceau de paramètres constitutifs de ce concept. Cependant, il n'existe pas de critère absolu permettant une distinction claire entre nom propre et non commun.

La variabilité des possibles combinaisons de ces paramètres amène naturellement, selon les personnes, les temps ou les lieux, à accorder le statut de nom propre à des candidats différents. Il apparaît alors un continuum noms propres/noms communs, avec à chaque extrémité, les prototypes de chaque espèce ; plus on se rapproche du milieu, plus les sujets ont un statut ambigu (cf. figure 1.1). Ce concept de continuum est à rapprocher de la notion de degré de similarité avec le prototype, qui suggérait déjà une continuité et une hiérarchie entre les différents candidats d'une catégorie (cf. section 1.2.2).

Ce flou entraîne automatiquement une ambiguïté, dès lors que l'on emploie le terme de nom propre. Face à cette ambiguïté, la communauté du TALN a introduit le terme « entité nommée » à l'occasion de la conférence MUC-6.

Il est à déplorer que de nombreux travaux emploient *noms propres* pour *entités nommées* ou indifféremment les deux termes, comme l'avoue [Friburger, 2002, chap. 2] : « Dans la suite, nous utiliserons indifféremment les termes de **noms propres** ou **entités nommées** pour désigner les noms propres au sens large du terme ». Dans ce cas, on est en droit de s'interroger sur la nécessité de l'introduction d'un nouveau terme et de son acceptation par la communauté.

1.4 Du nom propre à l'entité nommée

Depuis la « définition » donnée par MUC-6 (cf. section 1.4.1.1), de nombreux termes ont été assimilés aux entités nommées par les différents systèmes qui opèrent leur reconnaissance automatique : les ethnonymes, les artefacts (noms de marques et de produits), les noms de maladies, etc. (cf. section 2).

Pour Daille et Morin [2000] : « La notion d' « entité nommée » représente une catégorisation bien plus large que celle du nom propre [tel qu'il est abordé en linguistique] ».

Nous n'avons pas trouvé de réelle définition de ce qu'est une entité nommée, mais plutôt des catégorisations qui participent, tout comme pour le nom propre (cf. section 1.2.1), à une intuition de ce qu'est une entité nommée.

1.4.1 Catégorisations des entités nommées en TALN

La très large majorité des systèmes de reconnaissance des entités nommées se fondent sur la classification MUC. Malgré tout, quelques travaux font état d'autres typologies pour les entités nommées.

1.4.1.1 Classification MUC

Sous l'impulsion des conférences MUC dans les années 1980, les systèmes d'extraction d'information ont développé pour l'anglais des techniques robustes pour identifier et catégoriser les entités nommées à partir de corpus de textes.

Dans le cadre des conférences MUC, plusieurs systèmes d'extraction d'information ont été développés pour des domaines comme le terrorisme en Amérique Latine [MUC-3 ; MUC-4], la fusion d'entreprises internationales et la fabrication de circuits électroniques [MUC-5], ou les changements de dirigeants des entreprises [MUC-6]. Lors d'une conférence MUC, les protagonistes doivent développer un système qui extrait le plus d'informations possibles sur des entités bien déterminées, puis les résultats sont évalués suivant une procédure identique pour tous. Ainsi pour les conférences MUC-4 et MUC-5 qui portaient sur le terrorisme en Amérique Latine, l'objectif était d'extraire des dépêches d'agences de presse le maximum d'informations sur des actes de terrorisme tels que le nom du groupe terroriste, le nom de la victime, le type d'agression, etc. Dans cette optique, la reconnaissance des entités nommées est un enjeu majeur.

Avant la conférence MUC-6, les systèmes d'extraction d'information reconnaissaient déjà les noms de personnes, de lieux et d'organisations, mais ce n'est que lors de celle-ci que la « tâche de reconnaissance des entités nommées » (en anglais *Named Entity Task*) a été définie. Selon MUC-6, les entités nommées sont les noms propres, les acronymes et éventuellement divers autres identificateurs uniques, qui sont catégorisés par l'un des types suivants :

- la catégorie `ORGANISATION` regroupe les entreprises, les institutions gouvernementales et les autres organisations ;
- sous la catégorie `PERSON` se trouvent les noms de personnes ou de familles ;
- la catégorie `LOCATION` réunit les noms de lieux politiquement ou géographiquement définis (villes, provinces, pays, régions internationales, entités d'eau, montagnes, etc.).

1.4.1.2 Autres typologies

En fait, les systèmes de reconnaissance des entités nommées développés dans le cadre des conférences MUC sont loin de considérer toute la palette des entités existantes.

Coates-Stephens [1993] propose une typologie constituée des trois classes MUC enrichies, de quatre nouvelles classes :

- les noms de personnes ;
- les noms de lieux ;
- les noms d'organisations ;
- les noms d'origines (noms d'habitants de pays, de villes, de régions, etc.) ;
- les noms de législations (p. ex. *loi Evin*, *taxe Tobin*), d'indices boursiers (p. ex. *Nikkei*, *CAC 40*, *Dow Jones*), etc. ;
- les noms de sources d'informations (toutes les formes de médias) ;
- les noms d'évènements (guerres, révolutions, catastrophes, salons, JO, etc.) ;
- les noms d'objets.

Paik et coll. [1996] présentent une classification des entités nommées qui est beaucoup plus complète que les simples catégories MUC. Cette classification a été réalisée à partir d'une étude du *Wall Street Journal* qui comporte trente catégories divisées en neuf classes :

géographique : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, région, fleuves, autres noms géographiques ;

affiliation : religions, nationalités ;

organisation : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations ;

humain : personnes, fonctions ;

document : documents ;

équipement : logiciels, matériels, machines ;

scientifique : maladies, drogues, médicaments ;

temporelle : dates et heures ;

divers : autres noms d'entités nommées.

Les huit premières classes couvrent 89 % des entités nommées présentes dans le corpus d'étude de Paik et coll. [1996].

Sekine et coll. [2002] présentent une hiérarchie des entités nommées, composée de 150 types différents. Ils indiquent bien que la définition de ce qu'est une entité nommée est ambiguë et qu'elle doit inclure certains éléments, car leur contenu informationnel est intéressant à extraire pour des applications en TALN. Prenant l'exemple des modèles de voitures (p. ex. *Integra*, *Odyssey*), Sekine et coll. [2002] considèrent qu'il s'agit de noms de classes et non de noms d'individus spécifiques. Pour réaliser leur hiérarchie, ils décident que les noms de classes (p. ex. couleur, type d'avion, nom de matériau, nom d'animal) sont inclus, car l'apport que peut amener leur reconnaissance peut être utile. En revanche, les mots ordinaires comme « *cup* », « *feelings* », « *tear* », ne sont pas inclus dans les entités nommées. Sekine et coll. [2002] avouent enfin que la limite est fatalement ambiguë et que des décisions arbitraires sont nécessaires.

Cette hiérarchie a été construite sur la base de trois méthodes :

à base de corpus journalistiques Après avoir extrait 3 500 expressions candidates au statut d'entité nommée, à partir de différents textes journalistiques anglais, une catégorie leur a été assignée sur la base de l'intuition de l'examinateur. D'autre part, les catégories d'entités nommées ont été fusionnées ou divisées en fonction du nombre d'éléments qu'elles regroupaient ;

à base des systèmes et des tâches préexistants Le système d'extraction d'information **REES** [Aone et Ramos-Santacruz, 1998] suggère quelques nouvelles catégories qui sont intégrées par Sekine et coll. [2002]. De plus, l'analyse de la tâche TREC-QA (*Text REtrieval Conferences – Question Answering* [Voorhees, 2003]) a permis d'induire certaines autres catégories d'entités nommées ;

à base de thésaurus Les thésaurus consultés sont *WordNet* et *Roget Thesaurus*. Leur analyse s'effectue à la fois de façon ascendante et de façon descendante. Par l'analyse descendante, Sekine et coll. [2002] cherchent à augmenter la couverture d'entités nommées, afin de ne pas manquer de catégorie importante. L'analyse ascendante sert quant à elle à trouver les types d'entités nommées déjà connus.

À l'intérieur de cette hiérarchie, se trouvent des expressions qui sont souvent associées aux entités nommées : les expressions temporelles et les expressions numériques. Il ne s'agit pas à proprement parler d'entités nommées, mais elles sont reconnues par la « tâche de reconnaissance des entités nommées » lors des conférences MUC, ce qui explique cette fréquente confusion.

À l'intérieur de cette hiérarchie, les classes de plus haut niveau – de la classe `NAME` qui contient les entités nommées à proprement parler – sont :

- `PERSON` : les noms et les prénoms ;
- `ORGANIZATION` : les sociétés, les organisations politiques, militaires, sportives, etc., les groupes ethniques et les noms de nationalités ;
- `LOCATION` : tous les lieux géographiques et les corps astraux ;
- `FACILITY` : toutes les constructions de bâtiments et travaux publics (voies de circulation, monuments, parques, aéroports, ports, gares, etc.) ;
- `PRODUCT` : les productions industrielles, les monnaies, les récompenses, les productions de la pensée humaine (théories, lois, projets, etc.), les personnages fictifs, les matières universitaires (sociologie, philosophie, économie, etc.), les catégories (poids, taille, etc.), les sports, les œuvres d'art, les œuvres littéraires, les journaux et les magazines ;
- `DISEASE` : les noms de maladies ;
- `EVENT` : les évènements, les guerres, les phénomènes naturels et les crimes ;
- `TITLE` : les titres de civilité et de situation (p. ex. *président*, *roi*, *docteur*) ;
- `LANGUAGE` : les noms de langues ;
- `RELIGION` : les religions;

Une classification, pragmatique, des noms propres a été réalisée par le linguiste germanophone Bauer [1985]. Sa typologie est constituée de cinq catégories principales, avec pour chacune, différents types. Cette typologie a été reprise et présentée par Grass [1999] lors de la journée de l'*ATALA* sur *Le traitement automatique des noms propres* et possède cinq grandes classes d'entités nommées :

les anthroponymes caractérisent les personnes individuelles ou les groupes (patronymes, prénoms, pseudonymes, gentilés, hypocoristes, ethnonymes, groupes musicaux modernes, ensembles artistiques et orchestre classique, partis et organisations, noms données aux animaux familiers) ;

les toponymes désignent les noms de lieux (pays, villes, microtoponymes, hydronymes, oronymes, installations militaires) ;

les ergonymes comprennent les objets et les produits manufacturés (marques, entreprises, établissements d'enseignement et de recherche, titres de livres, de films, de publications, d'œuvres d'art) ;

les praxonymes désignent des faits historiques, des maladies, des évènements culturels ;

les phénonymes comprennent les ouragans, les zones de haute et de basse pressions, les astres et les comètes.

1.4.1.3 Synthèse

La notion d'entité nommée peut donc être comparée à celle de prototype. En effet, les prototypes du nom propre (les noms de personnes) représentent le nom propre au sens le plus strict du terme ; leur nature n'est en aucun cas ambiguë. En revanche, les entités nommées, si elles incluent les noms propres prototypiques, n'en ont pas moins, pour certaines, un statut très controversé.

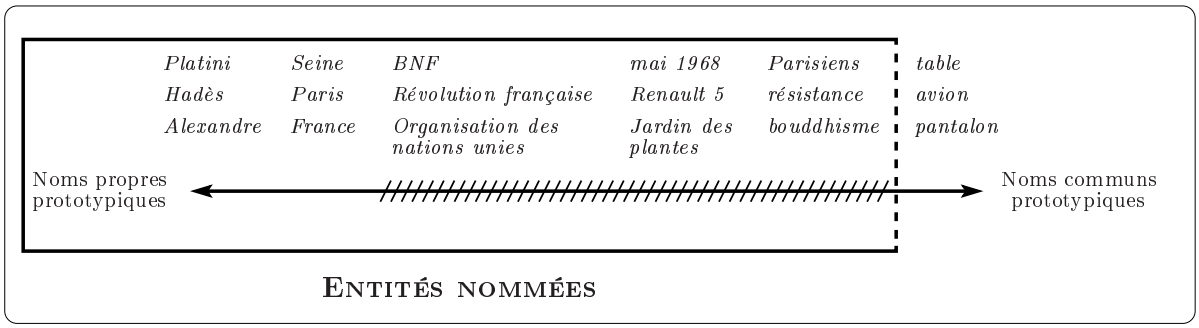


Figure 1.2 – L'entité nommée comme nom propre au sens le plus large

Au vu de ces typologies des entités nommées, nous pouvons donc définir l'entité nommée comme étant « l'acception la plus large que l'on peut faire du nom propre » (cf. figure 1.2). Les entités nommées sont, pour reprendre les termes de Molino [1982b] au sujet de sa géographie du nom propre (cf. section 1.2.1), tous les candidats au rang de nom propre, soit tous les termes qui ont pu être un jour considérés comme appartenant à la catégorie ; il s'agit donc d'une liste maximale.

1.4.2 Identification des paramètres linguistiques caractérisant les entités nommées

Si l'aspect typologique de l'intuition de l'entité nommée nous donne une bonne vision des sujets qui composent cette classe, il ne nous permet pas d'en dégager les paramètres linguistiques caractéristiques. En effet, au vu des conclusions auxquelles nous avons abouti concernant le nom propre, nous ne définissons pas de catégorie linguistique pour l'entité nommée. En revanche, nous avons décidé de sélectionner, parmi les différents paramètres établis pour restituer l'intuition de nom propre (cf. section 1.2.2), ceux qui apparaissent comme les plus saillants. En effet, la notion d'entité nommée étant l'acception la plus large que l'on puisse faire du nom propre, les éléments qu'elle regroupe ne partageront que les paramètres définitoires les plus caractéristiques des noms propres. Le but d'une telle démarche est de faire émerger des critères qui vont nous guider, lors de notre étude en corpus (cf. chapitre 3), afin de juger si un élément rencontré est une entité nommée ou non.

Parmi le faisceau de paramètres constitutifs du concept de nom propre, nous n'avons retenu que le critère graphique de la majuscule à l'initiale et le critère sémantico-pragmatique du référent unique, parce qu'il s'agit des deux critères partagés par le plus grand nombre de noms propres. De plus, dans l'optique de notre étude en corpus, certains critères doivent être immédiatement écartés, car ils nécessitent de projeter le sujet dans un autre contexte (pour les critères syntaxiques), ou d'en étudier les constructions et les dérivations (pour les critères morphologiques).

1.4.2.1 Critère graphique de la majuscule à l'initiale

Ce critère est sans doutes le plus répandu parmi les locuteurs du français et reste souvent donné comme définitoire du nom propre. De plus, la présence d'une majuscule à l'initiale d'un mot est très facile à identifier lors de la lecture d'un texte, mais également lors d'un traite-

ment automatique. Si l'application d'un tel critère présente le risque d'englober certains noms communs, il présente le gros avantage de couvrir la quasi totalité des entités nommées.

Il convient malgré tout de complexifier ce critère, en raison de la présence de nombreux éléments composés parmi les entités nommées (cf. section 3.2). En effet, les noms propres prototypiques étant très majoritairement composés d'une seule forme, il n'était pas fait cas, précédemment, des structures composées. Il ne nous est plus possible de valuer ce paramètre de façon binaire (avec ou sans majuscule à l'initiale) ; il s'agit maintenant d'examiner la proportion des formes du candidat qui prennent une majuscule à l'initiale : plus cette proportion sera grande, plus forte sera la probabilité qu'un candidat soit une entité nommée.

1.4.2.2 Critère sémantico-pragmatique du référent unique

Le critère graphique de la majuscule à l'initiale ne saurait être suffisant pour caractériser les entités nommées, car il englobe un certain nombre de noms communs. Nous avons donc choisi de l'accompagner d'un deuxième critère saillant : le critère sémantico-pragmatique du référent unique. Il s'agit de l'idée selon laquelle une entité nommée possède un référent unique.

Pour les noms propres, ce critère était discutable pour un ensemble de sujets dont l'unicité du référent se faisait en terme de classe et non d'individu (p. ex. *une Renault 5*, *les Parisiens*). Pour les entités nommées, nous avons décidé d'introduire une gradualité dans cette unicité : nous aurons donc des sujets qui possèdent un référent plus ou moins unique (cf. figure 1.3).

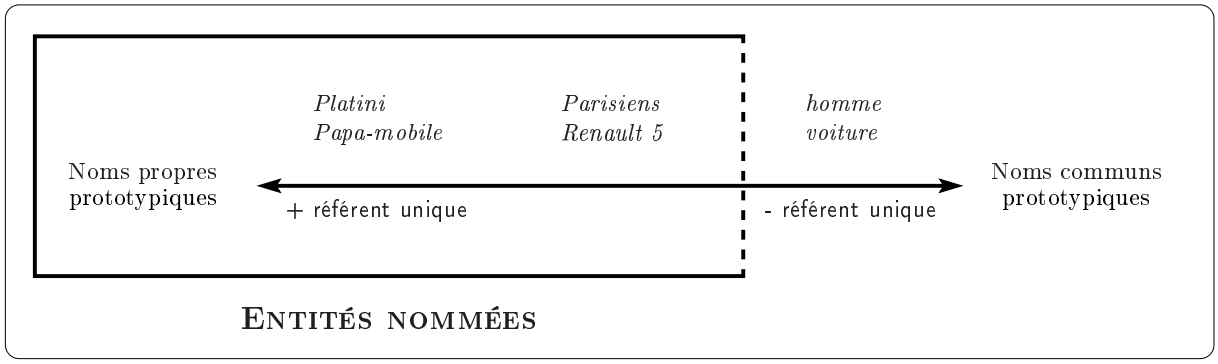


Figure 1.3 – Gradualité dans l'unicité du référent d'un sujet

1.4.2.3 Principe de sélection des entités nommées en corpus

Les deux paramètres linguistiques que nous avons choisi pour caractériser les entités nommées vont nous permettre de définir un principe de sélection des entités nommées pour notre étude en corpus (cf. section 3.2). Cette sélection consiste à accorder le statut d'entité nommée à un candidat en fonction de l'unicité de son référent et de sa graphie. Dans un premier temps, nous regardons parmi les formes qui composent le candidat au statut d'entité nommée, si au moins une d'entre elles possède une majuscule à l'initiale. Si ce n'est pas le cas, il n'est pas retenu ; sinon, il est considéré comme entité nommée en fonction de la proportion des formes la composant qui prennent une majuscule à l'initiale et du caractère unique de son référent.

Outre ces critères linguistiques, il ne faut pas oublier le paramètre typologique qui veut qu'un sujet est considéré comme entité nommée s'il s'intègre à la typologie des entités nommées. Ce

critère nécessite donc une typologie exhaustive pour être utilisé. Or, notre but est justement l'élaboration de cette typologie. Nous ne pouvons donc pas nous baser sur cet indice pour sélectionner les entités nommées en corpus, mais une fois notre typologie des entités nommées établie (cf. tableau 3.5), elle devra le permettre.

1.5 Conclusion

De l'étude lexicographique menée au début de ce chapitre, nous avons conclu à une impossible définition de « nom propre », ce qui nous a amené à considérer la notion d'intuition liée à ce concept.

L'analyse de la nature double de cette intuition – typologique et linguistique – a abouti, par la restitution de ses fondements, à un faisceau de paramètres linguistiques constitutifs du concept de nom propre.

La variabilité des possibles combinaisons de ces paramètres nous a amené à introduire l'idée d'un continuum entre les noms propres et les noms communs, et non deux ensembles distincts.

Ensuite, nous avons défini la notion d'entité nommée par rapport à celle de nom propre, en situant les entités nommées dans le continuum noms propres/noms communs, grâce à une étude des différentes catégorisations d'entités nommées en TALN.

Enfin, nous avons établi les paramètres linguistiques caractérisant les entités nommées et défini un principe de sélection des possibles entités nommées en corpus, afin d'établir notre typologie.

Chapitre 2

Les systèmes de reconnaissance des entités nommées

La reconnaissance des entités nommées en corpus est donc devenue incontournable dans la plupart des tâches du traitement automatique des langues. Dans ce chapitre, nous passons en revue les récents systèmes développés pour la reconnaissance des entités nommées appliquée à différentes langues écrites : leurs objectifs, leurs méthodes et leurs résultats. Nous montrons que dans cette problématique, un acquis s'est constitué, notamment sous l'impulsion des projets MUC (*Message Understanding Conference*), MET (*Multilingual Entity Task Conference*), ou encore IREX (*Information Retrieval and Extraction Exercise* [Sekine et Isahara, 1999]). Les méthodes élaborées ont permis la construction de systèmes aux performances intéressantes. Cependant, ces systèmes ne constituent qu'une base, ils ne sont en aucun cas une fin. En effet, comme nous le verrons, il reste dans cette problématique des lacunes à combler ainsi que des progrès à réaliser.

2.1 Introduction

2.1.1 Les différents types de systèmes

Poibeau [2002] propose une classification des systèmes de reconnaissance des entités nommées, inspirée de celle de Sekine et Eriguchi [2000]. Il distingue trois types de systèmes :

Les systèmes fondés sur des règles écrites « à la main » dont le concepteur doit élaborer manuellement un ensemble de patrons permettant la reconnaissance des entités nommées. Historiquement, cette technique fut la première à voir le jour et demeura la base de la quasi totalité des systèmes de reconnaissance des entités nommées, avant l'apparition de l'apprentissage dans ce domaine [MUC-6]. Aujourd'hui encore, la majorité des systèmes se fondent sur cette technique.

Les systèmes à base d'apprentissage automatique qui se fondent sur des techniques d'apprentissage pour créer un modèle d'étiquetage à partir d'un corpus d'entraînement. Cette approche s'est largement inspirée des travaux réalisés sur la langue orale et les techniques d'apprentissages utilisées sont variées (p. ex. modèle de Markov cachés, maximisation de l'entropie, arbres de décision).

Les systèmes mixtes qui utilisent le plus souvent un ensemble de lexiques initiaux. Parmi ces systèmes, Poibeau [2002] distingue deux approches. La première consiste à apprendre automatiquement des règles, puis à utiliser un expert pour les réviser. Dans la seconde, un ensemble de règles de bases est constitué par le concepteur, puis étendu (semi-)automatiquement par inférence, afin d'obtenir une meilleure couverture.

Le tableau 2.1 recense les différents systèmes de reconnaissance des entités nommées que nous avons particulièrement étudiés, regroupés selon les différents types de cette classification.

Sekine et Eriguchi [2000] constatent que les trois systèmes qui arrivent en tête des évaluations du projet IREX viennent des trois catégories et concluent ne pas pouvoir en déduire la supériorité d'une des méthodes sur les autres.

2.1.2 Problèmes linguistiques

Quelle que soit l'approche utilisée, différents types de problèmes sont à résoudre au cours des phases de la reconnaissance des entités nommées en corpus.

Pour exprimer ces problèmes, nous reprenons la terminologie introduite par Daille et Morin [2000] concernant les entités nommées, fondée sur des critères graphiques. Nous appellerons :

Entités nommées pures les entités nommées constituées d'une seule forme commençant par une majuscule comme *France*, *FFF*, *États-Unis*.

Entités nommées composées les entités nommées constituées de plusieurs formes pleines comportant toutes une majuscule (qu'elles soient pures ou descriptives) comme *Quai Branly*, *Grand Palais*.

Entités nommées mixtes les entités nommées constituées de plusieurs formes qui peuvent ou non commencer par une majuscule, mais dont au moins une est entièrement en minuscules comme *Parti socialiste*, *palais de Chaillot*.

Cette distinction est importante, car en anglais on est uniquement confrontée à la difficulté d'identifier les entités nommées composées, alors que pour les textes français se posent des problèmes liés à la fois aux entités nommées composées et aux entités nommées mixtes.

La première phase de la reconnaissance des entités nommées réside dans la reconnaissance d'entités nommées connues. Cette étape consiste à rechercher les entités nommées qui se trouvent présentes dans les dictionnaires utilisés par le système et les catégoriser en résolvant les problèmes liés à la polysémie.

2.1.2.1 Reconnaissance d'entités nommées connues

Cette reconnaissance se fait à l'aide de bases lexicales qui sont projetées sur un texte. Dans cette phase, les programmes sont confrontés à l'ambiguïté sémantique des entités nommées pures : par exemple, *Paris* est enregistré comme un nom de ville, mais il peut aussi apparaître au sein d'entités nommées composées ou mixtes. Il désigne alors une personne *Comte de Paris*, un nom de rue *rue de Paris*, un parfum *Paris d'Yves Saint-Laurent*, une société *Paris International*, etc. Une catégorisation correcte, même pour les entités nommées connues, nécessite l'examen du contexte puisque des éléments linguistiques situés à l'intérieur d'une même phrase influencent significativement son interprétation. Une reconnaissance correcte dépendra évidemment de l'exhaustivité des dictionnaires recensant les entités nommées composées ou mixtes, mais surtout, d'une analyse linguistique permettant de calculer le changement de catégorisation en fonction du contexte. Dans la plupart des cas, une analyse locale suffit comme pour *Comte de Paris* où l'apparition de *Comte* induit une catégorisation humaine, mais d'autres cas demandent une analyse de la phrase comme pour l'exemple de Kleiber [1981, page 320] : *Georges Sand est sur l'étagère de gauche*.

À ce problème d'ambiguïté s'ajoutent les variations que peuvent subir les entités nommées :

Tableau 2.1 – Les différents systèmes de reconnaissance des entités nommées

Systèmes à base de règles		
Nom	**Auteur(s)**	**Année**
LaSIE	Gaizauskas, Wakao, Humphreys, Cunningham et Wilks	1995
Proteus	Grishman	1995
Exoseme	Wolinski, Vichot et Dillet	1995
PNF	McDonald	1996
DR-LINK	Paik, Liddy, Yu et McKenna	1996
	Trouilleux	1997
Nominator	Wacholder, Ravin et Choi	1997
LaSIE-II	Humphreys, Gaizauskas, Azzam, Huyck, Mitchell, Cunningham et Wilks	1998
Sprach-R	Renals, Gotoh, Gaizauskas et Stevenson	1999
NERC	Demiros, Boutsis, Giouli, Liakata, Papageorgiou et Piperidis	2000
	Nenadić et Spacić	2000
ExtracNP	Friburger	2002
Systèmes à base d'apprentissage automatique		
Nom	**Auteur(s)**	**Année**
Nymble	Bikel, Miller, Schwartz et Weischedel	1997
	Palmer et Day	1997
	Sekine, Grishman et Shinnou	1998
MENE	Borthwick	1999
	Collins et Singer	1999
Sprach-S	Renals, Gotoh, Gaizauskas et Stevenson	1999
	Niu, Li, Ding et Srihari	2003
Systèmes mixtes		
Nom	**Auteur(s)**	**Année**
	Gallippi	1996
	Cucchiarelli, Luzi et Velardi	1999
	Mikheev, Moens et Grover	1999
	Béchet, Nasr et Genet	2000
SemTex	Poibeau	2001
	Quasthoff, Biemann et Wolff	2002

- les variations graphiques : *Parti Socialiste*, *Parti socialiste*, *parti socialiste* ; *École Normale Supérieure*, *Ecole normale supérieure* ;
- les sigles, acronymes ou abréviations : *École Normale Supérieure* → *ENS* ;
- les variations morphosyntaxiques : *les habitants de Nantes* → *les nantais*, *la politique de Jospin* → *la politique jospinienne* ;
- les coordinations : *le couple Montand-Signoret*, *le Grand et le Petit Palais* ;
- les ellipses : *École Normale Supérieure* → *Normale*, *Normale sup* ;
- les métaphores : *l'Everest* → *le toit du monde*.

Hormis la reconnaissance des entités nommées connues, l'identification et la catégorisation des entités nommées consiste dans la découverte de nouvelles entités nommées (c.-à-d. absentes des dictionnaires).

2.1.2.2 Découverte de nouvelles entités nommées

La plupart des entités nommées se distinguent des autres catégories linguistiques par l'emploi systématique de la majuscule. Cependant, cette caractéristique graphique ne s'applique pas aux noms communs et aux adjectifs de type relationnel reflétant une origine ethnique associée à une langue (*lévites*) ou encore à une école de pensée, de religion (*gaullistes*, *scientologues*). Ces noms et ces adjectifs sont très productifs et font référence à un humain ou à un collectif d'humains et sont assimilés à des entités nommées de la classe des anthroponymes par Bauer [1985].

Pour les entités nommées mixtes, elles sont, par définition, constituées d'une ou plusieurs formes comportant une majuscule et de noms communs. La difficulté réside dans la délimitation des bornes de l'entité nommée, à gauche et à droite de la ou des formes identifiées à l'aide de leur majuscule.

La délimitation à gauche porte sur le problème d'inclure ou non le nom commun précédant la forme comportant une majuscule au sein de l'entité nommée mixte. Ces noms communs sont de type classifiant et précisent un rôle social *le président Chirac*, *le capitaine Troy*, des statuts particuliers *l'exilé Musil*, une catégorisation géographique *la région Aquitaine*, *la rue de Paris*, etc. Pour certains, ces noms communs font partie intégrante de l'entité nommée comme *rue de Paris*, pour d'autres, ils permettent de constituer une entité nommée « complète » et non-ambiguë [Jonasson, 1994], mais qui ne constitue pas une entité nommée mixte comme *le président Chirac*. Ces entités nommées complètes sont soumises à des variations, ainsi *le président Chirac* peut être cité dans les textes sous les variantes suivantes : *Jacques Chirac – le Président de la République française, Jacques Chirac – Jacques Chirac, le président de la France* – etc.

La délimitation à droite se heurte aux problèmes de la modification, de la résolution de l'attachement prépositionnel et de la portée de la coordination [Wacholder et coll., 1997]. L'adjectif modifiant une entité nommée comme *fédéral* dans *Allemagne fédérale* constitue une entité nommée mixte, ce qui n'est pas vrai avec *lointain* dans *Chine lointaine*. De même, le groupe prépositionnel *des nations* dans *Le championnat d'Europe des nations a eu lieu...* fait partie intégrante de l'entité nommée, ce qui n'est pas le cas avec *Les Premiers Ministres des nations se sont rencontrés...* Enfin, *le Mouvement contre le racisme et pour l'amitié entre les peuples* est une entité nommée mixte à l'inverse de *Mouvements populaires et État furent longtemps attelés ensemble....*

En plus de cette ambiguïté sur la limite à droite des entités nommées, se pose le problème de la sur-composition : une entité nommée mixte peut contenir une entité nommée d'une autre catégorie (p. ex. *Guerre d'Algérie*, *Université de Nantes*).

Pour faire face à ces différents problèmes, les systèmes de reconnaissance des entités nommées font appel à des solutions tantôt linguistiques, tantôt à base d'apprentissage (issues de la théorie de l'information), voire une conjonction des deux.

2.2 Méthodes linguistiques

Les méthodes linguistiques pour la reconnaissance des entités nommées permettent non seulement de lever les ambiguïtés portant sur les entités nommées connues, mais également d'identifier et de catégoriser de nouvelles entités nommées. Ces méthodes reposent essentiellement sur :

1. L'utilisation de dictionnaires.
2. L'analyse de la structure interne de l'entité nommée.
3. L'analyse du contexte dans lequel l'entité nommée apparaît.
4. La résolution de certaines coréférences.

Les analyses de la structure interne et du contexte s'effectuent soit à l'aide d'heuristiques utilisant des mots-clés [Wacholder et coll., 1997], soit grâce à des informations morpho-syntaxiques fournies par un programme d'étiquetage [Trouilleux, 1997 ; Grishman, 1995 ; Friburger, 2002], soit en mettant en œuvre des méthodes plus élaborées comme une analyse syntaxique [McDonald, 1996] ou un modèle discursif [Wakao et coll., 1996].

2.2.1 Lexiques spécialisés

Il existe trois grandes familles de lexiques spécialisés intervenant dans la reconnaissance des entités nommées :

- les dictionnaires électroniques d'entités nommées ;
- les lexiques de mots déclencheurs (*trigger words*) ;
- les dictionnaires de synonymes.

2.2.1.1 Les dictionnaires d'entités nommées

Les dictionnaires d'entités nommées permettent, à l'aide de transducteurs par exemple, de repérer facilement et rapidement en corpus les entités nommées connues. Actuellement, de telles dictionnaire sont indispensables à tous les systèmes de reconnaissance d'entités nommées, même si la façon de les obtenir peut varier (utilisation de corpus d'apprentissage ou de listes préétablies).

Un dictionnaire électronique recensant les gentilés de la région Ouest a été réalisé par Belleil [1997] et permet donc l'identification des noms communs reflétant une origine ethnique associé à un lieu géographique ne commençant pas par une majuscule.

Senellart [1998] semi-automatise la construction de transducteurs à automates finis pour la reconnaissance des entités nommées, de leurs variations et de leurs contextes d'apparition en corpus. Cet apprentissage interactif s'effectue à l'aide d'un algorithme d'analyse fondé sur une indexation du corpus. À partir d'une concordance obtenue sur un mot simple souvent ambigu comme *Officer*, l'utilisateur identifie le contexte droit ou gauche permettant de lever cette ambiguïté comme *Army officer*, *intelligence officer*, *Marine officer*. Ce contexte est alors inséré dans l'automate de manière à raffiner la reconnaissance. Les états de l'automate sont associés soit à une

ou plusieurs formes lexicales, soit à une liste de mots partageant une même fonction sémantique comme par exemple les adjectifs de nationalité : *french*, *soviet*, *american*, etc.

Stevenson et Gaizauskas [2000] montrent que les lexiques de noms de personnes souffrent fortement d'un changement entre leur période de création et celle dont datent les corpus sur lesquels ils sont utilisés, alors que les noms d'organisations et de lieux sont plus stables dans le temps.

Mikheev et coll. [1999] montrent que, pour des textes journalistiques, seuls de petits dictionnaires d'entités nommées sont suffisants à la reconnaissance de celles-ci. Il rappelle également que cette tâche ne peut être réalisée uniquement en utilisant des listes de noms de personnes, lieux et organisations, et ce pour plusieurs raisons :

- il n'est pas possible de lister toutes les entités nommées, en témoigne le nombre de prénoms recensés pour les seuls États Unis (1,5 million) ;
- quand bien même, certaines de ces listes seraient immédiatement surannées (p. ex. des organisations se créent en permanence) ;
- de plus, toutes les variations d'une même entité nommée devraient être listées (p. ex. les noms d'organisations peuvent apparaître sous différentes formes : *The Royal Bank of Scotland plc*, *The Royal Bank of Scotland*, *The Royal* ou *The Royal plc*) ;
- il y a toujours des problèmes de conflits entre les listes dus à la polysémie de certaines entités nommées (p. ex. *Washington* ou *Emerson* peuvent désigner le lieu ou la personne) ;
- certaines entités nommées peuvent être composées (de plusieurs formes), particulièrement lorsque composées de conjonctions ;

2.2.1.2 Les lexiques de mots déclencheurs

Associés à des heuristiques, les lexiques de mots déclencheurs vont contribuer à la découverte de nouvelles entités nommées et à la désambiguïsation de la catégorie assignée à l'entité nommée lorsque cela s'avère nécessaire. La très grande majorité des systèmes de reconnaissance des entités nommées requièrent l'utilisation de tels lexiques. Ces listes de mots peuvent intervenir à la fois dans l'analyse de la structure interne de l'entité nommée, ainsi que dans celle de son contexte. La présence de tels mots dans un corpus fournit non seulement une forte présomption sur la présence d'une entité nommée dans les mots qui les entourent (cette entité pouvant contenir ou non le mot déclencheur), mais permet également de lui assigner le plus souvent une catégorie non ambiguë.

Dans le cadre des conférences MUC et HUB4 [Renals et coll., 1999], Stevenson et Gaizauskas [2000] présentent un système de reconnaissance des entités nommées, fondé sur celui de Wakao et coll. [1996], qui n'utilise qu'un module de recherche d'éléments appartenant aux lexiques. Ce module marque tous les mots contenus dans chaque liste d'entités nommées comme telle. Toute forme appartenant à plusieurs listes se voit assigner la catégorie de la première liste dans laquelle elle est retrouvée.

Ces listes sont composées d'éléments extraits des lexiques fournis par MUC-7 et enrichies manuellement. Elles incluent des noms d'entreprises, d'organisations, de pays, de villes, de continents, de régions, ainsi que des prénoms, des titres de personnes et des mots-clés d'entreprises. Pour comparer cette approche avec un système qui collecte simplement toutes les entités nommées d'un corpus d'entraînement, Stevenson et Gaizauskas [2000] ont implémenté un programme

qui analyse le texte annoté par les balise *SGML* de MUC et créent une liste pour chaque type d'entité nommée trouvée.

Une première série d'expériences montre que :

- des résultats parfaits ne peuvent être obtenus par cette seule méthode, même en utilisant les listes sur le corpus dont elles sont issues ;
- sur un corpus de test, les résultats sont meilleurs avec les lexiques de base ;
- la combinaison des deux types de lexiques donne des résultats intéressants, mais peut s'avérer négative si les lexiques dérivés de corpus sont obtenus à partir d'un corpus trop large.

Stevenson et Gaizauskas [2000] filtrent ensuite les lexiques dérivés de corpus de différentes manières :

1. À l'aide de dictionnaire : retirer les éléments qui apparaissent dans un dictionnaire.
2. De façon statistique : retirer les éléments qui apparaissent plus souvent comme noms communs dans le corpus.
3. En combinant les deux premières approches : *combinaison ou* et *combinaison et*.

L'utilisation des lexiques dérivés de corpus, filtrés par la combinaison *et*, donne les meilleurs résultats : 87 % de F-mesure, soit 4 % de mieux qu'avec les lexiques de base.

En conclusion, un filtrage adéquat des lexiques dérivés de corpus peut fournir des lexiques plus efficaces que ceux créés manuellement. De plus, bien que 5 % en dessous des meilleurs systèmes, les performances de Stevenson et Gaizauskas [2000] restent intéressantes au regard des techniques utilisées.

Dans le cadre des conférences MUC, Mikheev et coll. [1999] proposent une étude sur l'impact des lexiques dans la reconnaissance des entités nommées. En utilisant le système que nous décrivons à la section 2.4.2 sans avoir recours au moindre lexique, Mikheev et coll. [1999] parviennent à des résultats corrects pour les noms de personnes et d'organisations (plus de 85 % de rappel et de précision), mais mauvais pour les noms de lieux (environ 46 % de rappel et 59 % de précision).

Ils testent ensuite un système minimal, semblable à celui de Palmer et Day [1997], qui n'a recours qu'à l'utilisation de lexiques (entités nommées et mots déclencheurs). Les performances atteintes par ce système sont de 90 % pour la précision et entre 26 et 76 % pour le rappel, selon la catégorie (26 % pour les noms de personnes, 76 % pour les noms de lieux, 49 % pour les noms d'organisations). Malgré sa simplicité, un tel système présuppose l'existence de corpus d'entraînement[1] pour créer ces lexiques. En remplaçant les lexiques ainsi obtenus par des listes d'entités nommées les plus courantes, les performances sont comparables pour les noms de lieux. En revanche, elles sont bien moindres pour les noms de personnes et mauvaises pour les noms d'organisations. En combinant ces deux types de lexiques, Mikheev et coll. [1999] obtiennent des performances raisonnables pour les noms de lieux (90-94 % de précision et 75-85 % de rappel), mais insuffisantes pour les noms de personnes et surtout les noms d'organisations (75-85 % de précision et moins de 50 % de rappel).

Cette étude suggère donc que la constitution des lexiques n'est pas un goulet d'étranglement pour la reconnaissance des entités nommées : des lexiques limités – dont la création demeure un problème surmontable –, conjugués à une utilisation judicieuse des évidences interne et externe sont suffisants pour obtenir des bons taux de rappel et de précision.

[1] Ici, ce corpus est composé de 30 articles du même domaine.

2.2.1.3 Les dictionnaires de synonymes

Les dictionnaires de synonymes permettent de découvrir de nouvelles entités nommées en substituant aux constituants des entités nommées déjà connues leurs formes synonymiques [Cucchiarelli et coll., 1999] ou en standardisant les différentes formes d'une même entité nommée [Paik et coll., 1996 ; Wolinski et coll., 1995].

Cucchiarelli et coll. [1999] utilisent un dictionnaire des synonymes pour engendrer des contextes similaires aux contextes d'entités nommées déjà reconnues. Par exemple, si le nom de lieu *Kilimanjaro* est trouvé dans le contexte *le mont Kilimanjaro*, le dictionnaire de synonymes permet de produire les contextes similaires à celui-là, avec les synonymes de *mont* (*montagne*, *hauteur*, *massif*, etc.) et le même le lien syntaxique élémentaire (cf. section 2.4.1.1).

Paik et coll. [1996], lors de la phase de classification des entités nommées, regardent, pour chacune d'entre elles qui n'a pu l'être préalablement, si cette entité nommée aurait, dans une base de donnée d'alias, une forme alternative qui permette de la classifier. Ensuite, les différentes formes sont standardisées.

Wolinski et coll. [1995] utilisent des listes contenant les différentes formes que peuvent prendre les entités nommées, afin d'autoriser certaines fautes d'orthographes considérées comme des variations (p. ex. *Boris Eltsine*, *Boris Elstine*, *Boris Etlsine*, *Boris Yeltsine*). Toutes ces variations sont regroupées autour d'un « équivalent », à la manière de Hayes [1994]. Les « mots équivalents » sont exprimés dans la base de connaissance à travers une relation sémantique (cf. figure 2.1). Il ne s'agit pas à proprement parler d'un dictionnaire de synonymes, mais cela s'en approche beaucoup, surtout si l'on considère qu'il n'y a pas de fautes d'orthographe sur les nom propres (cf. section 1.2.2.1).

De la même façon, Wolinski et coll. [1995] traitent les divers synonymes d'une même entité nommée (p. ex. *Hexagone/France* ou *Rue d'Antin/Paribas*) à l'aide d'un dictionnaire. Tous les synonymes sont regroupés autour d'une seule référence (cf. figure 2.2). Ce regroupement permet la représentation des différentes formes et la factorisation de leurs attributs.

Malgré leur utilité, les dictionnaires de synonymes sont très peu utilisés par les systèmes de reconnaissance des entités nommées. De plus, ils le sont presque systématiquement pour rechercher les formes synonymiques des entités nommées, mais pas des éléments de leur contexte.

2.2.2 Évidence interne

L'évidence interne (*internal evidence*), comme la définit McDonald [1996], se situe au niveau de la séquence de mots comprenant l'entité nommée. Elle peut consister dans des critères immuables, comme la présence de termes clefs qui indiquent des entreprises (p. ex. *Ltd.* ou *Inc.*), ou dans des critères heuristiques, comme des abréviations ou des prénoms connus qui indiquent généralement des personnes. La grande majorité des systèmes de reconnaissance des entités nommées ont recours à l'étude de cette évidence interne. Elle est communément réalisée grâce à l'utilisation de volumineux dictionnaires d'entités nommées et de listes de mots déclencheurs.

Il peut être parfois délicat de savoir si un terme appartient ou non à l'entité nommée (cf. section 4.1.1) et donc s'il fait partie de l'évidence interne ou externe (p. ex. *les montagnes Rocheuses*, *les Montagnes Rocheuses*, *les Rocheuses*). Par conséquent, nous avons décidé, dans ce chapitre, de considérer les noms communs de type classifiant (mots déclencheurs) comme faisant partie de l'évidence interne, qu'ils précèdent directement l'entité nommée ou qu'ils en fassent intégralement partie.

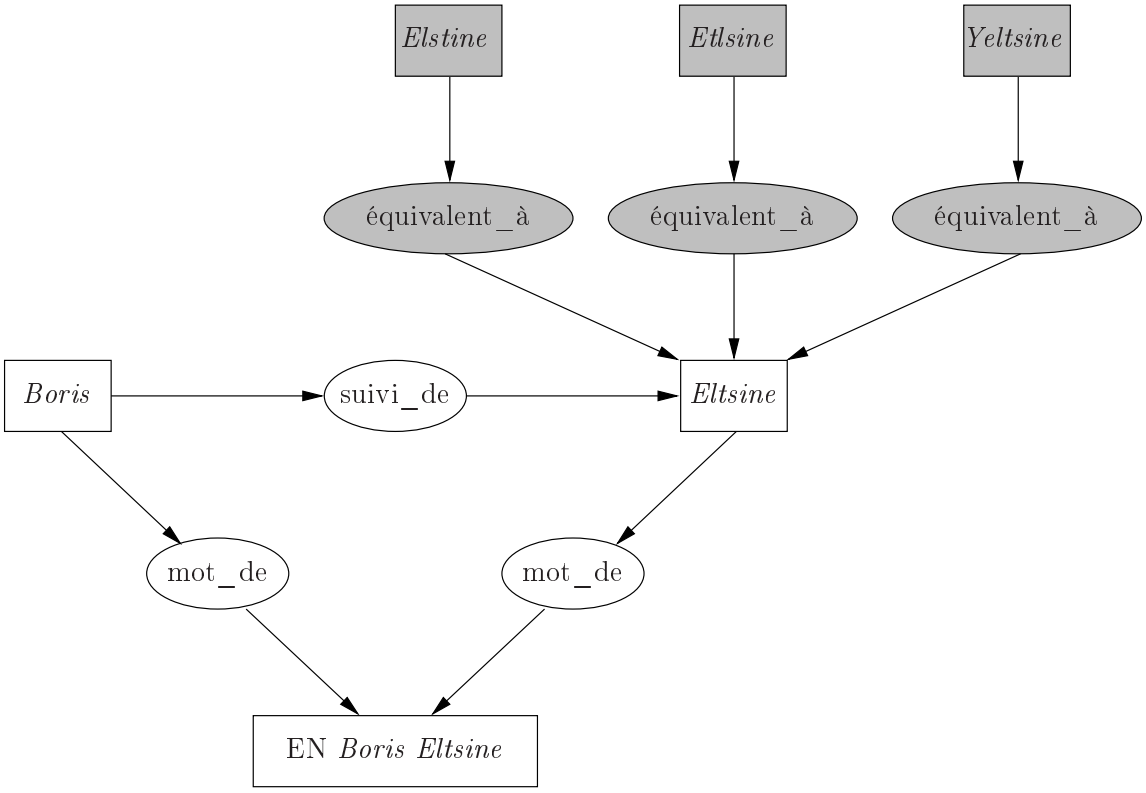


Figure 2.1 – Entités nommées « équivalentes »

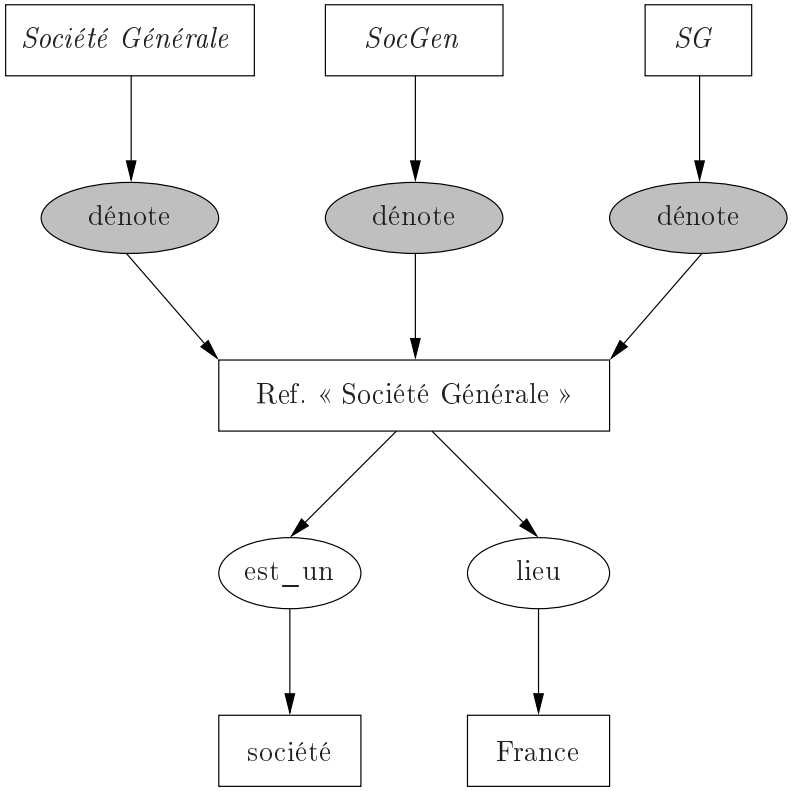


Figure 2.2 – Entités nommées synonymes

Nominator, le programme développé au Centre de Recherche T.J. Watson d'IBM par Wacholder et coll. [1997], se concentre uniquement sur la reconnaissance des entités nommées incluant les entités nommées mixtes et leurs variations. En créant **Nominator**, Wacholder et coll. [1997] ont essayé d'atteindre un compromis entre performance et rapidité d'exécution en adoptant un modèle qui nécessite un minimum de contexte et de connaissances linguistiques. Si une base de données d'entités nommées existe, **Nominator** peut l'utiliser ; cependant, le but ici est de réaliser un système qui fonctionne de façon optimale, sans avoir recours à ce genre de ressource.

Nominator, après avoir effectué la segmentation du texte, isole toutes les séquences textuelles caractéristiques des différentes catégories linguistiques d'entités nommées : les séquences comportant des mots en majuscules et éventuellement des conjonctions et des prépositions en minuscules comme *American Bar Association*, *Robert Jordan*, *ABA*, *Mr. Jordan of Steptoe & Johnson*, *the Victoria and Albert Museum*. **Nominator** décide ensuite, à l'aide d'heuristiques, d'un éventuel découpage des séquences textuelles comportant une préposition ou une coordination, en des séquences plus petites d'entités nommées indépendantes (p. ex. *Mr. Jordan of Steptoe & Johnson* → *Mr. Jordan* et *Steptoe & Johnson*). En revanche, la séquence *Victoria and Albert Museum* ne sera pas découpée, car le mot *Museum* apparaît dans la liste des noms acceptant une extension gauche. Par contre, le découpage sera possible, même pour un mot acceptant une extension, si un acronyme est détecté comme *IBM and Bell Laboratories* → *IBM*, *Bell Laboratories*. **Nominator** réalise également un traitement particulier pour les mots situés en début de phrase et qui comportent une majuscule à l'initiale, parce qu'ils peuvent faire partie d'une

entité nommée ou simplement débuter la phrase. Si un tel mot se retrouve dans le texte avec une majuscule, sans être en début de phrase, il est gardé ; autrement, il est rejeté. Par exemple, *White* sera gardé si *Mr. White* est retrouvé dans le texte.

Nominator n'utilise donc ni dictionnaire, ni véritables informations linguistiques. Il se base essentiellement sur la présence de la majuscule et sur des listes de composants associés à des propriétés linguistiques ou à des propriétés catégorisantes, afin d'assigner un type aux entités nommées (catégorisation à la MUC). Ces composants particuliers, qui composent l'évidence interne et dans une moindre mesure l'évidence externe, correspondent à la présence au sein d'une entité nommée complète ou mixte :

- d'une particule qui permet l'identification d'une personne ou d'une société ;
- d'un prénom pour l'identification d'une personne ;
- d'un sigle ou d'un acronyme pour les organisations ;
- d'un nom commun de type classifiant (mot déclencheur) précédant l'entité nommée, comme *l'avenue des Anglais*, *le docteur Knock*.

Nominator reconnaît également certaines variations des entités nommées (cf. section 2.2.4). L'évaluation du système donne 91 % d'identification correcte des frontières des entités nommées. Seules 71 % de ces entités nommées sont catégorisées, mais avec une précision de 99 %.

Cucchiarelli et coll. [1999] identifient les entités nommées pour l'italien et les catégorisent selon les classes MUC enrichies des noms de produits. Ce traitement combine approches symbolique et statistique.

Dans un premier temps, la reconnaissance s'effectue à l'aide d'un système proche de celui développé par Humphreys et coll. [1996] :

1. Les entités nommées les plus communes sont identifiées à l'aide de dictionnaires d'entités nommées, ainsi qu'une liste de mots déclencheurs par catégorie (p. ex. *Gulf* pour les noms de lieux ou *Association* pour les organisations).

2. une grammaire contextuelle, contenant environ 250 règles, est appliquée pour faire l'analyse grammaticale des entités nommées en contexte. La majorité de ces règles utilise l'évidence interne pour identifier et catégoriser les entités nommées composées.

Ensuite, pour pallier la difficulté liée à l'incomplétude des lexiques, un processus d'apprentissage est mis en place, afin d'améliorer les performances (cf. section 2.4.1.1).

Poibeau [1999] présente **SemTex**, un système de repérage et d'acquisition d'entités nommées pour le français et l'anglais qui est adapté à la prise de décision en veille économique ou stratégique. Ici, l'objectif est de pouvoir déterminer rapidement si un texte est intéressant ou non en repérant les noms d'entreprises, de dirigeants, de lieux et les dates figurant dans les textes pour suivre l'évolution du monde économique. L'architecture du système de Poibeau [1999] est héritée des systèmes d'extraction d'information ayant participé aux conférences MUC : il utilise le même type de catégorisation et procède par analyse locale progressivement étendue par un ensemble d'automates appliqués en cascade.

Dans un premier temps, le système de Poibeau [1999] procède à la fois à un découpage du texte en unités minimales et à un étiquetage lexical :

1. Les nombres sont analysés suivant des critères formels (un nombre est composé de chiffres et éventuellement d'un tiret, d'un point ou d'une virgule).

2. Un dictionnaire des entités nommées les plus courantes permet d'identifier et de catégoriser un certain nombre de noms de personnes et de lieux. Ce dictionnaire est particulièrement développé pour des listes semi-fermées comme les prénoms, beaucoup moins pour les noms.
3. Un dictionnaire d'amorces (mots déclencheurs) permet de reconnaître certaines séquences importantes pour la suite (p. ex. *M.*, *SA*, etc.).
4. un algorithme permet de traiter et de normaliser les sigles comme *I.B.M* ou *I.B.M.*, alors que le sigle *IBM*, écrit sans point, ne sera pas reconnu.
5. Les mots inconnus sont étiquetés en tant que tels au moyen d'un dictionnaire général de la langue. Ils reçoivent une étiquette différente selon qu'ils sont en majuscules ou en minuscules et qu'ils se situent en début de phrase ou non.
6. Parmi les mots figurant dans le dictionnaire général de la langue, ceux qui commencent par une majuscule sont étiquetés.

Cette analyse lexicale est réalisée au moyen d'un arbre de décision qui fournit un texte enrichi de balises entourant les mots repérés. Par exemple, dans la phrase suivante chaque mot reconnu est entouré d'une étiquette[2] :

```
Les <NUMBER> 10 </NUMBER> % restants appartiennent au fils adoptif d'
<PERSON> Enzo </PERSON> <PERSON><COMPANY> Ferrari </COMPANY></PERSON>,
<TR_PERSON> M. </TR_PERSON> <PERSON> Piero </PERSON>
<UFIRSTUNKNOWN> Lardi </UFIRSTUNKNOWN>.
```

À la suite de cette étape, le fonctionnement du système ne repose plus que sur l'analyse des suites d'étiquettes, indépendamment des formes effectivement présentes dans le texte.

Dans un second temps, une grammaire des entités nommées permet de reconnaître parmi les séquences précédentes celles qui sont susceptibles de former des entités (séquences pertinentes). À ce niveau, un ensemble de règles de réécriture est développé pour reconnaître et typer les entités nommées, mais aussi pour résoudre les ambiguïtés. Ces règles sont compilées et s'appliquent selon certaines heuristiques :

- les règles les plus longues (celles qui contiennent le plus d'éléments) s'appliquent les premières ;
- une règle ne peut s'appliquer à l'intérieur d'une séquence précédemment reconnue ;
- si deux règles de même longueur peuvent s'appliquer, le résultat est aléatoire. Poibeau [1999] justifie cette heuristique en affirmant que « Ce principe, en évitant d'avoir à gérer les conflits, permet d'assurer la robustesse du système. ».

Ainsi, la règle `TR_PERSON PERSON? UFIRSTUNKNOWN+ ==> PERSON` (c.-à-d. l'amorce d'un nom de personne comme M., Mme, etc., avec un nom ou un prénom optionnel et un ou plusieurs mots inconnus commençant par une majuscule) permet d'identifier la séquence *M. Piero Lardi* comme étant un nom de personne. Un système de préférence permet de résoudre les ambiguïtés comme celle liée au mot *Ferrari*. Dans ce cas, l'étiquette `<PERSON>` est privilégié, puisqu'une règle permet la reconnaissance de *Enzo Ferrari* comme un nom de personne. Cependant, lorsque le contexte ne permet pas de désambiguïser une expression, c'est l'étiquette la plus probable qui est choisie (*Ferrari* est classé comme étant préférentiellement un nom de société) ou, en l'absence d'une telle information, le système fait un choix aléatoire.

[2] La signification des étiquettes est la suivante : `<NUMBER>` pour un nombre, `<PERSON>` pour un nom de personne, `<COMPANY>` pour un nom de société, `<TR_PERSON>` pour l'amorce d'un nom de personne et `<UFIRSTUNKNOWN>` pour un mot inconnu commençant par une majuscule.

Dans un dernier temps, le système opère un marquage des entités nommées repérées au moyen de balises XML (une DTD est définie pour rendre directement compte du marquage vu précédemment).

Le développement du système a été réalisée sur un corpus de textes concernant l'actualité économique extraits du journal *Le Monde*. Dans un premier temps, le système procède à une adaptation sur ce corpus. Une évaluation manuelle révèle environ 20 % de candidats pertinents pour les entités nommées avec les règles et les lexiques initiaux. Puis, les lexiques sont complétés et les règles modifiées manuellement, afin d'améliorer le taux de reconnaissance. Par exemple, la règle :
`TR_PERSON PERSON? UFIRSTUNKNOWN+ ==> PERSON` ne permet de reconnaître que partiellement *M. Frederik de Klerk*. Poibeau [1999] crée alors une nouvelle règle :
`TR_PERSON PERSON? PREP UFIRSTUNKNOWN+ ==> PERSON`.
À son tour, cette règle ne permet de reconnaître que partiellement *M. Jean de la Guerivière*. Elle est donc modifiée pour donner : `TR_PERSON PERSON? PREP DET? UFIRSTUNKNOWN+ ==> PERSON`.

Après trois itérations, le système affiche un taux de reconnaissance (F-Mesure) de l'ordre de 90 %. Suite à cette phase de mise au point, le système est évalué sur un nouvel échantillon (un corpus de 25 textes[3] extraits du journal *Le Monde* contenant environ 20 000 mots) et affiche un taux de reconnaissance de l'ordre de 80 %. Pour l'anglais, les performances atteintes sur le corpus MUC-6 sont légèrement moins bonnes (90 %) que celles des meilleurs systèmes ayant participé à MUC-6 (p. ex. **BBN** [Miller et coll., 1998] obtient 98 %).

Poibeau [2002] évalue les performances de ce système sur des corpus non journalistiques :

- un corpus d'environ 2 200 mots issu de transcriptions manuelles de conversations téléphoniques en anglais. Il s'agit de transcriptions en casse mixte, ponctuées et contenant très peu d'erreur de reconnaissance ;
- un corpus composé d'un ensemble de courriers électroniques en anglais portant sur le domaine des télécommunications (50 000 mots environ).

Cette évaluation montre une chute considérable des performances de **SemTex** (-40 %). Un effondrement similaire, voire pire, est également constaté pour les autres systèmes testés que sont **Exibum** [Kosseim et Lapalme, 1998], **Alembic** [Aberdeen et coll., 1995], ou encore **TextPro** [Appelt et Martin, 1999].

Poibeau [2002] en conclut que seules des stratégies de typage dynamique mettant en jeu l'analyse d'autres occurrences peuvent amener à résoudre les cas difficiles que constituent les entités nommées trouvées sans contexte catégorisant. Dans cette optique, Poibeau [2001] reprend le système **SemTex** et y ajoute un regroupement des entités coréférentes, un traitement des structures énumératives et un mécanisme de révision (cf. section 2.2.4).

DR-LINK, le système développé par Paik et coll. [1996] est le seul à utiliser une catégorisation fine, neuf catégories et trente sous-catégories (cf. section 1.4.1) pour identifier les entités nommées de l'anglais. Le texte est tout d'abord traité par un étiqueteur probabiliste assignant les parties du discours. Puis, un identificateur des frontières de l'entité nommée utilise ces étiquettes pour regrouper les entités nommées adjacentes. Parallèlement, des heuristiques développées grâce à une analyse en corpus, sont appliquées pour regrouper les syntagmes des entités nommées avec des prépositions et des conjonctions enchâssées. Par exemple, une liste spécifiée d'entités nommées sera rattachée à d'autres entités nommées non-adjacentes si elles sont liées par la préposition

[3] Ces textes ont été sélectionnés en faisant une simple requête sur la base avec le mot-clé *Affaires internationales*.

of. Les entités nommées composant cette liste sont : *Council*, *Ministry*, *Secretary* et *University*. Le taux de précision de cette identification des frontières des entités nommées est de 96 %.

La catégorisation des entités identifiées utilise différentes méthodes. La première consiste à comparer les parties constitutives de l'entité nommée à une liste de tous les préfixes, infixes et suffixes identifiés pour établir la catégorie de celle-ci. Si aucun affixe n'est reconnu, l'entité nommée est projetée sur une base de données d'alias (de synonymes) pour déterminer si elle possède une autre forme. Si c'est le cas, elle est standardisée et catégorisée à ce stade. Sinon, elle est projetée sur une base de connaissances, construite à l'aide de ressources lexicales en ligne comme *the TRIPSTER Gazetteer*, *the 1992 CIA World Factbase* et *the Executive Desk Reference*. Jusqu'ici, Paik et coll. [1996] n'utilisent que des informations provenant des lexiques et de l'évidence interne. Cependant, si cette dernière étape est infructueuse, un ensemble d'heuristiques contextuelles, développées lors d'une analyse en corpus, suggère certaine catégories d'entités nommées. Par exemple, si une entité nommée est suivie d'une virgule et d'une autre entité nommée identifiée comme un état des États-Unis, elle sera étiquetée comme un nom de ville (p. ex. *Time, Illinois*). Enfin, si l'entité nommée n'est toujours pas catégorisée, elle est comparée à une liste de prénoms pour une dernière vérification de catégorisation en nom de personne.

L'évaluation de **DR-LINK** est effectuée sur un corpus regroupant 25 textes du *Wall Street Journal* sélectionnés aléatoirement et comportant 589 entités nommées. Le taux de précision obtenu est de 92 %, en comptant les dates (27) et les entités nommées correctement catégorisées comme diverses (64). Dans les même conditions, le taux de rappel atteint est de 91 %.

Proteus, le système de reconnaissance de Grishman [1995] procède tout d'abord à la segmentation du texte en phrases, puis en formes. Chaque forme est ensuite recherchée dans les différents lexiques (entités nommées et termes liés à des scénarios spécifiques). Un étiqueteur assignant les parties du discours (*BBN POST tagger*) est également appliqué au corpus. Les entités nommées sont alors reconnues à l'aide de motifs[4] limités à l'évidence interne. Si, subséquemment, une portion d'une de ces entités nommées est retrouvée, elle est marquée comme *alias* et reconnue comme entité nommée. La reconnaissance des scénarios spécifiques[5] (démission, succession, promotion, etc.), réalisée par ce système, permet également d'augmenter légèrement la reconnaissance des entités nommées. Après cette phase, Grishman [1995] accomplit une résolution des coréférences (cf. section 2.2.4). Les performances obtenues par ce système dans le cadre de MUC-6 sont de 88,19 % pour la F-Mesure.

L'analyse de l'évidence interne est donc indispensable à l'identification et les catégorisations des entités nommées et elle permet d'obtenir de bons taux de reconnaissance. Cependant, ces seuls indices ne sauraient être suffisant pour obtenir une couverture de l'ensemble des entités nommées. En effet, tous les systèmes de reconnaissance des entités nommées sont confrontés au problème posé par les entités nommées qui apparaissent sans élément d'évidence interne permettant leur catégorisation.

Wakao et coll. [1996] ; Trouilleux [1997] ; Wolinski et coll. [1995] ; Friburger [2002] ; Nenadić et Spacić [2000] utilisent également des lexiques d'entités nommées et de mots déclencheurs pour analyser l'évidence interne, mais étudient de surcroît le contexte plus général (évidence externe) des entités nommées pour parfaire leur identification et leur catégorisation.

[4]Il s'agit de motifs à états finis, traduits en LISP pour assurer une meilleure performance.

[5]Cette reconnaissance de scénarios peut être considérée comme utilisant l'évidence externe, mais nous ne la détaillerons pas, car elle dépasse le cadre de la simple reconnaissance des entités nommée.

2.2.3 Évidence externe

McDonald [1996] affirme que l'étude de l'évidence externe (*external evidence*), au même titre que celle de l'évidence interne, est essentielle pour obtenir de bons taux de rappel et de précision en reconnaissance des entités nommées. En se basant sur ces conclusions, de nombreux systèmes effectuant cette tâche examinent non seulement les mots composant l'entité nommée elle-même, mais également d'autres informations dans le texte.

Wakao et coll. [1996] sont du même avis que McDonald [1996] et montrent l'apport de l'analyse du contexte pour l'identification des entités nommées. Les systèmes **LaSIE** (Large Scale Information Extraction) et **LaSIE-II**, développés à l'Université de Scheffield [Gaizauskas et coll., 1995 ; Humphreys et coll., 1998] et utilisés dans la cadre des conférences MUC-6 et MUC-7 pour identifier les entités nommées dans le *Wall Street Journal*, mettent en œuvre trois niveaux de traitements :

Prétraitement lexical À ce niveau le texte est découpé en occurrences de phrases et en occurrences de formes, puis étiqueté par le catégoriseur de Brill [1994]. Une première série d'entités nommées est ensuite identifiée en s'appuyant, d'une part, sur des lexiques d'entités nommées (noms d'organisations, de lieux, de personnes, etc.), et d'autre part, sur des listes de mots déclencheurs (*trigger words*) qui indiquent que les mots qui entoure un mot déclencheur désignent probablement une entité nommée (par exemple, le mot *Gulf* est exploité pour identifier un nom de lieu).

Analyse grammaticale locale Deux grammaires locales (hors contexte) des entités nommées sont exploitées pour identifier les entités nommées, une grammaire spécifique qui a été écrite à la main et une grammaire générale qui a été automatiquement extraite d'un corpus de référence où les entités nommées sont identifiées. Ces deux grammaires sont utilisées en concurrence pour offrir la meilleure analyse. Au terme de cette analyse, effectuée par une unification à l'aide de *Prolog*, des types sémantiques sont attribués aux différentes entités nommées.

Analyse du discours Ce dernier niveau ne réalise pas de reconnaissance des entités nommées, mais propose un raffinement de la classification. Une analyse de la coréférence des entités nommées est également réalisée à ce stade (cf. section 2.2.4). Enfin, de nouveaux types sémantiques sont inférés pour des entités nommées identifiées mais non classifiées. Par exemple, pour l'apposition *Fort Lauderlade, Fla.*, si *Fla.* est classifié comme un nom de lieu, alors *Fort Lauderer* sera aussi classifié comme un nom de lieu.

Au cours de ce travail, Wakao et coll. [1996] montrent l'apport des différents niveaux de traitements pour identifier et classifier les entités nommées. Ils indiquent en particulier que l'étude du contexte (évidence externe) est indispensable pour l'identification d'un nom de personne ou d'organisation, alors que l'identification d'un nom de lieu peut se limiter au premier niveau de traitement (évidence interne).

Pour prendre en compte l'évidence externe, Trouilleux [1997] s'appuie sur un étiquetage morpho-syntaxique, alors que McDonald [1996] ; Wakao et coll. [1996] ; Wolinski et coll. [1995] utilisent une analyse syntaxique.

Trouilleux [1997] présente un système pour l'identification et le classement des entités nommées pour le français. Les entités nommées traitées par le système sont : les noms de personnes, les noms de lieux, les noms d'organisations, les noms de textes légaux, les noms d'évènements et

les autres noms qui désignent des entités nommées reconnues qui ne font pas partie des catégories précédentes. Dans le système développé par Trouilleux [1997], seules les entités nommées contenant au moins une lettre en majuscule sont identifiées. Le texte à analyser est d'abord étiqueté, puis des lexiques spécialisés (prénoms, noms de lieu, d'organisation et d'évènement, titres de civilité, militaires, civils, etc.) sont consultés pour modifier le précédent étiquetage. Ensuite, les contextes à gauche et à droite des mots commençant par une majuscule sont étudiés pour leur attacher des compléments. Le rattachement du contexte gauche est réalisé à partir de lexiques spécifiques qui permettent de qualifier le nom en majuscule : *physicien Eugene Wigner*, *empereur Shen Nung*, *océan Pacifique*, *volcan Tindaya*, etc. Le rattachement du contexte droit nécessite également des lexiques spécifiques, mais aussi un étiquetage morpho-syntaxique pour réaliser une extension sélective. Ainsi, Trouilleux [1997] décrit une grammaire du contexte droit en fonction du type du nom en majuscule :

Extension après un nom de personne Le rattachement à droite n'est pas réalisé.

Extension après un nom de lieu Les rattachements autorisés concernent les adjectifs et les syntagmes prépositionnels introduits par *de*, où la tête désigne un point cardinal.

Extension après un nom d'organisation ou d'évènement On distingue les extensions avec préposition des extensions sans préposition. Dans le premier cas, les prépositions valides sont *pour*, *sur*, *à*, *au*, *aux*, *contre* et *entre* utilisées avec des contraintes spécifiques. Dans le second cas, l'extension peut être une suite d'adjectifs sans préposition (p. ex. *Fonds monétaire international*).

Extension après une référence non identifiée Les rattachements autorisés sont :

1. Une séquence composée d'au moins un nom ou un adjectif pouvant comporter la conjonction *et*, la préposition *sans* ou la préposition *de* avec éventuellement le déterminant *le*.
2. Un chiffre ou un nombre (p. ex. *Century 21*).

Ainsi, un syntagme prépositionnel introduit par *de* à droite d'une entité nommée identifiée comme un nom de lieu n'est pas extrait par cette technique, comme pour le groupe prépositionnel *des nations* dans *championnat d'Europe des nations*.

Les résultats obtenus par ce type d'analyse, de l'ordre de 55 %, sont bien moins bons que ceux obtenus par les systèmes anglo-saxons. Cependant, ils laissent entrevoir des possibilités intéressantes si l'on considère la petitesse des différents lexiques utilisés qui n'offrent de fait qu'une couverture réduite.

Wolinski et coll. [1995] ont développé le système **Exoseme** qui analyse en continu les dépêches de l'Agence France Presse (AFP), afin de les classer en une cinquantaine de thèmes.

Le système **Exoseme** est constitué de plusieurs modules : un analyseur morphologique, un analyseur syntaxique, un module d'identification des entités nommées, un analyseur sémantique et un module de filtrage thématique. Le module d'identification des entités nommées travaille en deux temps : segmentation, puis catégorisation.

La segmentation des entités nommées repose sur un analyseur morphologique et un analyseur syntaxique. Ce dernier est nécessaire, en particulier dans le cas d'entités nommées contenant des marqueurs grammaticaux (p. ex. prépositions, conjonctions, virgules, points, etc.). En effet, pour un syntagme comme *Caisse du Crédit Agricole du Morbihan*, l'analyseur fournira deux interprétations, selon que *Morbihan* est attaché à *Crédit Agricole* ou à *Caisse*. De plus, Wolinski et coll.

[1995] remarquent que les entités nommées contiennent parfois des segments agrammaticaux, qui peuvent induire l'analyseur syntaxique en erreur. Par exemple, dans la phrase *The director of Dollfus, Mieg and Cie has announced positive results*, l'analyseur aura des difficultés à identifier *director* comme le sujet de *announce* s'il ne connaît pas l'entreprise *Dollfus, Mieg and Cie*. Pour traiter ce problème, l'analyseur syntaxique d'**Exoseme** délègue la segmentation de cette portion agrammaticale au module d'identification des entités nommées.

La catégorisation des entités nommées est nécessaire au bon fonctionnement de l'analyseur sémantique. Elle diffère selon que l'entité est connue ou non. Pour chaque entité nommée, sa catégorie et ses différents attributs sont directement représentés sous forme de graphe conceptuel.

Dans un premier temps, **Exoseme** reconnaît les entités nommées les plus fréquentes (p. ex. pays, villes, régions, hommes d'état, etc.), répertoriées par des dictionnaires, afin de les segmenter et les catégoriser correctement. Pour permettre le recensement de toutes les formes que peut prendre une entité nommée du fait des abréviations (*Société Générale* → *Soc. gen.*, *Sté générale*), des traductions (*Standard and Poor's* → *Standard and poors*, *Standard et Poor's*), des fautes communes (*Boris Eltsine* → *Boris Elstine*, *Boris Yeltsine*, *Boris Etlsine*), etc., certaines variations sont acceptées et regroupées en « termes équivalents ». Ceci permet de mettre en relation les formes les plus courantes d'une même entité nommée. Le même type de traitement est effectué avec les entités nommées synonymes : les formes synonymiques d'une même entité nommée (p. ex. *France* et *Hexagone* ou *Rue d'Antin* et *Paribas*) sont regroupées, la base de connaissance est modifiée en enrichissant le modèle de graphe conceptuel. Puis, à l'aide de ce que Wolinski et coll. [1995] appellent le « contexte local » et le « contexte global », la catégorie sémantique de l'entité nommée est désambiguïsée. Le « contexte local » s'apparente à l'évidence externe, alors que le « contexte global » repose sur la résolution de quelques coréférences : en général, la forme ambiguë d'une entité nommée (acronyme ou abréviation) n'apparaît pas seule ; elle est souvent accompagnée de la forme non-ambiguë ce qui permet de la catégoriser. Par exemple, *la Générale* peut référer *la Société Générale* ou *la Compagnie Générale des Eaux* ou encore *la Générale de Sucrière*. Wolinski et coll. [1995] recherchent donc à proximité de l'entité *la Générale* une forme non-ambiguë à laquelle elle pourrait référer.

Dans un second temps, **Exoseme** va essayer de traiter les entités nommées inconnues :

- en étudiant l'évidence interne à l'aide de lexiques et d'heuristiques
  (p. ex. `prénom + mot en majuscule → personne`
  `entreprise + "-" + nom de lieu → entreprise`) ;
- en étudiant le « contexte local » (évidence externe), essentiellement les appositions (p. ex. *Platon, le philosophe grec*, *Peskine, directeur du groupe*) et les compléments (p. ex. *le maire de Royan*, *les actionnaires de Fibaly*) ;
- en étudiant le « contexte global » : traitement de certaines coréférences (cf. section 2.2.4) ;
- en associant certains acronymes et leur forme étendue, afin de catégoriser ces premiers.

Le système **Exoseme** identifie 85 % des entités nommées présentes dans les dépêches de l'AFP et catégorise correctement 90 % d'entre elles.

McDonald [1996] utilise un analyseur à base de règles contextuelles de réécriture, le composant **PNF** (Proper Name Facility) de SPARSER, lorsque les marqueurs internes manquent ou sont ambigus. Les trois actions d'analyse de l'entité nommée, c'est-à-dire la délimitation, l'identification et son enregistrement, s'effectuent séparément à l'aide de trois modules indépendants.

L'algorithme de délimitation des entités nommées de **PNF** identifie simplement les séquences de mots commençant par une majuscule et les regroupe. La séquence s'arrête sur le premier mot

qui commence par une minuscule ou sur une virgule. Les autres cas sont traités individuellement : p. ex. *&* est considéré comme étendant la séquence, alors que les points sont pris comme terminant la séquence sans en faire partie, sauf pour les abréviations.

La classification se décompose en deux étapes :

1. L'analyse de la structure interne qui utilise l'information que la grammaire possède sur les entités nommées connues :
   - références à des villes ou des pays (p. ex. *Cambridge Savings Bank*) ;
   - mots déclencheurs comme *Church*, *Bank*, *Inc.*, *Mr.*, *Sir*, *Jr.*, etc. ;
   - éléments utilisés pour les heuristiques de classification comme *&* qui marque la présence d'une entreprise ou *II* qui peut-être utilisé pour les noms de personnes.
2. L'analyse du contexte lorsque l'étape précédente n'a pas réussi à assigner une catégorie non ambiguë à l'entité nommée. Cette analyse s'effectue à l'aide d'une grammaire intégrant la sémantique du domaine : les évènements types, les propriétés principales et les relations entre les différentes entités types du domaine.

L'enregistrement s'effectue dans un modèle sémantique où l'entité nommée est associée à un identificateur unique regroupant ses différentes interprétations. Il obtient 100 % pour la reconnaissance et la catégorisation des noms de personne et d'organisations pour un sous-domaine particulier.

NERC mis au point par Demiros et coll. [2000] est un système de reconnaissance des entités nommées pour le Grec. Ce système, élaboré dans le cadre des conférences MUC, est composé de quatre modules principaux, qui s'exécutent les uns à la suite des autres.

Le premier consiste dans un prétraitement lexical (type Multext) qui opère une séparation du texte en occurrences de phrases et de formes et identifie les abréviations, les chiffres et les dates simples. Les phrases et les formes sont identifiées avec une précision de 95 %.

Le second module de **NERC** effectue un étiquetage des parties du discours et une lemmatisation. Pour cela, Demiros et coll. [2000] utilisent l'étiqueteur de Brill [1993] entraîné sur un corpus grec de 250 000 mots annoté manuellement, dont la précision habituellement reportée descend ici à 90 %. Après la mise en place des étiquette des parties du discours, les lemmes sont retrouvés à partir d'un lexique morphologique du grec contenant 70 000 lemmes.

Le troisième module réalise la recherche des entités nommées à proprement parler. Pour cela, Demiros et coll. [2000] ont compilé des lexiques d'entités nommées à l'aide de différentes sources : *Pages Jaunes*, *Athens Stock Exchange*, *National Statistical Service of Greece*, ainsi que d'autres services et organisations. Ces lexiques ont été enrichis par des entités nommées extraites de textes[6] annotés manuellement. Finalement, ils regroupent 1 496 noms de personnes, 1 059 noms d'organisations et 793 noms de lieux. En plus de ces lexiques, ce module utilise des listes de mots et des expressions régulières indiquant la présence d'entités nommées. Ils sont obtenus sur un corpus d'entraînement, en extrayant les éléments du contexte dans une fenêtre de trois à cinq mots à droite et à gauche de l'entité nommée. Ces éléments sont ensuite manuellement regroupés en *clusters* selon leur usage et leur sémantique (57 *clusters* contenant 920 mots, groupes de mots et expressions régulières). Les mots d'un cluster se voient assigné une étiquette qui déclenchera la règle correspondante dans l'étape de balisage final des entités nommées.

Cette dernière constitue le module final de **NERC**. Durant cette phase, des règles sont appliquées pour entériner le marquage d'une entité nommée et en reconnaître de nouvelles. Ces

[6] 130 000 mots au total.

règles utilisent les entités nommées reconnues lors de l'étape précédente, les informations de casse, les étiquettes des parties du discours et les étiquettes correspondant aux *clusters* définis précédemment. Ces règles sont des expressions régulières compilées sous la forme de transducteurs d'états finis. Elles sont appliquées successivement avec priorité à la plus longue.

Le corpus d'évaluation est constitué de textes grecs téléchargés du *Web* (12 000 000 mots, dont 150 000 ont servi de corpus d'entraînement et de test). La densité d'entités nommées a été augmentée en ne gardant que les textes contenant le plus de mots commençant par une majuscule.

Les résultats obtenus par **NERC**, en terme de F-Mesure, sont de 71 % pour les noms de personnes, 83 % pour les noms de lieux et 76 % pour les noms d'organisations. Il s'agit de performances moyennes pour une langue que Demiros et coll. [2000] ne présentent pas comme comportant des difficultés supplémentaires à l'anglais, qui plus est dans le cadre des catégories MUC. On peut remarquer que la catégorie d'entités nommées qui pose le moins de problèmes à Demiros et coll. [2000] est celle des noms de lieux. Or Mikheev et coll. [1999] ont montré qu'il s'agissait de la catégorie qui souffrait le plus de l'absence de dictionnaire d'entités nommées et qui atteignait, au contraire, des performances intéressantes (comparables à celles de Demiros et coll. [2000]) avec leur seul usage. On peut donc penser que **NERC** possède un système de règles et d'heuristiques insuffisants.

Le système **ExtracNP** [Friburger, 2002] utilise des cascades de transducteurs[7] pour identifier et catégoriser les noms de personnes, les toponymes et les noms d'organisations du français.

Après un découpage en phrases [Friburger, 2000] et l'application des dictionnaires pour étiqueter le texte, la détection des entités nommées est réalisée par deux cascades de transducteurs[8] :

1. La première cascade de transducteurs permet l'extraction des entités nommées par la description de leurs grammaires locales grâce aux transducteurs (graphes d'**Intex**). Ces grammaires sont fortement lexicalisées et utilisent des preuves internes et externes[9], accompagnées d'indices morphologiques et syntaxiques.
2. La deuxième cascade de transducteurs est générée automatiquement à partir des entités nommées trouvées par la première. Elles sont utilisées directement, comme un dictionnaire, pour en retrouver les formes coréférentes qui seraient présentes dans le texte, mais qu'aucun autre indice n'aurait permis de reconnaître.

Parmi les noms de personnes, Friburger [2002] estime que 90 % sont accompagnés de preuves internes et externes permettant leur reconnaissance. Pour les autres, qui sont des noms de personnes très connues pour lesquelles l'auteur du texte n'estime pas nécessaire de préciser le prénom, elle crée un dictionnaire des célébrités. Les indices contextuels sont représentés par le contexte immédiat des entités nommées – dont le nombre de mots n'est pas limité – parmi lesquels : les civilités (*Mme*, *Monsieur*, etc.), les titres de toutes sortes (*président*, *évêque*, *lieutenant*, etc.), les professions (*juge*, *architecte*, etc.) et les adjectifs de nationalité (*japonais*, *allemand*, etc.). En plus de ce contexte immédiat propre à l'entité nommée, Friburger [2002] recherche les coordinations

[7]Une cascade de transducteurs est une succession de transducteurs appliqués sur un texte, dans un ordre précis, pour le transformer ou en extraire des motifs. Abney [1996] définit une cascade de transducteurs comme une séquence de couches (*sequence of strata*) qui décrivent les grammaires locales.

[8]Lors de l'application des transducteurs, les règles les plus longues sont appliquées les premières, indépendamment de la catégorie d'entités nommées qu'elles reconnaissent.

[9]Les « preuves internes » (resp. « externes »), sont des éléments de l'évidence interne (resp. externe) qui participent à la reconnaissance des entités nommées par un transducteur.

les plus fréquentes dans les textes journalistiques : *MM. Jacques Delors et Raymond Barre* ou *MM. Alyas (Le Faouët), Allain (Lorient), Bouet (Rennes)...* ou encore *MM. Maurice Girard, soixante-neuf ans, et Alain Mercadé, quarante-huit ans*. Les règles pour les noms de personnes sont au nombre de 43.

Pour les noms d'organisations, il en existe 35. Selon Friburger [2002], 50 % des noms d'organisations sont accompagnés d'une preuve interne car ce sont, pour la plupart, des entités nommées à base descriptive. Ils sont formés de noms communs qui permettent d'en deviner la nature. Par conséquent, ils sont rarement accompagnés d'un contexte gauche catégorisant (on ne dirait pas *la société Compagnie générale des eaux*). Pour déterminer la limite à droite d'un nom d'organisation, Friburger [2002] crée un dictionnaire d'adjectifs et de noms pouvant se retrouver dans une telle entité nommée, parmi lesquels se trouvent les noms et adjectifs toponymiques. Les noms d'organisations étrangers sont plus simples à reconnaître, car ils sont uniquement constitués de mots commençant par une majuscule et de mots outils, sans ambiguïté avec des mots français (*of*, *or*, *and*, etc.). De plus, ils sont le plus souvent composés d'une preuve interne à leur extrémité droite (*Ltd*, *Agency*, *Inc.*, etc.). Les preuves externes pour cette catégorie d'entités nommées sont essentiellement constituées d'un mot déclencheur (*groupe*, *agence*, *société*, etc.) et éventuellement un adjectif de nationalité ; environ 15 % des noms d'organisations sont accompagnés d'une telle preuve externe. Pour traiter les sigles, Friburger [2002] utilise, dans un contexte syntaxique très restreint, les étiquettes placées par le dictionnaire *Sigle-Prolex*[10]. 16 % des noms d'organisations sont ainsi reconnus, mais en dernier ressort, car cette technique est très imprécise. Les coordinations de noms d'organisations sont traités comme pour les noms de personnes avec des amorces au pluriel (p. ex. *les sociétés AXA, AGF et MAIF*).

Parmi les noms de lieux rencontrés par Friburger [2002], seuls 20 % possèdent un contexte gauche catégorisant ou plus rarement une preuve interne. Ceux qui ont une preuve interne sont les noms de villes ou de départements (*Chaumont-**sur**-Loire*) et les noms de lieux étrangers (*Trafalgar **Square***, *Main **Street***, *Hyde **Park***). Les preuves externes pour cette extraction sont par exemple les mots déclencheurs *mer*, *mont*, *département*, *estuaire*, etc. ou les points cardinaux (p. ex. *sud du sahara*, *Asie du sud-Est*). Il existe 31 graphes appelés par la cascade de transducteurs pour extraire les noms de lieux. Cependant, « Pour trouver la plus grande partie des noms de lieux mais aussi des gentilés » Friburger utilise principalement « le dictionnaire *Prolintex* de toponymes ». Ces graphes sont passés après ceux des noms de personnes et d'organisations pour éviter les ambiguïtés.

ExtracNP est évalué sur deux corpus : *Le Monde*[11] et *Ouest France*[12]. Ces corpus, bien que tous les deux journalistiques, ont une nature différente de par le fait que la rédaction du *Monde* obéit à des normes de présentation très strictes, ce qui est nettement moins le cas de *Ouest France*. Pour ce qui est du contenu, nous remarquons que les noms de personnes et d'organisation sont plus abondants dans *Le Monde* (resp. 1 936 et 1 588 contre 1 671 et 1 040), alors qu'il y a un plus grand nombre de noms de lieux dans *Ouest France* (3 627 contre 2 208). En terme de F-Mesure, les performances atteintes par **ExtracNP** pour les noms de personnes sont d'environ 93,5 % pour les deux corpus. Pour les noms de lieux, les résultats sont également très bon (95,5 % pour *Le Monde* et 94,5 % pour *Ouest France*). Selon Friburger [2002], ces derniers sont essentiellement dus aux travaux antérieurs du projet *Prolex* et aux dictionnaires *Prolintex* de toponymes ; la

[10] Il s'agit d'un dictionnaire de sigles et d'abréviations créé par Friburger [2002, chap. 4] et contenant environ 3 300 entrées.

[11] Deux numéros, soit 142 000 mots environ.

[12] Une série d'articles d'environ 152 000 mots.

seconde cascade de transducteurs ne permet une reconnaissance que marginale des noms de lieux (resp. 1,27 % et 0,14 % de rappel pour *Le Monde* et *Ouest France*, avec une précision très faible), alors qu'elle est plus intéressante pour les noms de personnes (resp. 9 % et 6,4 % de rappel avec des précisions de 84,5 % et 71,8 %, pour *Le Monde* et *Ouest France*). En ce qui concerne les noms d'organisations, dont près d'un quart sont trouvés grâce au dictionnaire des sigles, les performance sont un peu en dessous (90,1 % de F-Mesure pour *Le Monde* et 89,3 % de F-Mesure pour *Ouest France*). La seconde cascade de transducteurs est bien plus intéressante pour *Le Monde* (77,3 % de précision pour 10,1 % de rappel) que pour *Ouest France* (89,7 % de précision pour seulement 2,5 % de rappel). Sur l'ensemble des entités nommées reconnues dans les deux corpus, Friburger [2002] obtient un taux de F-Mesure d'environ 93,5 % et remarque que les grammaires permettant cette reconnaissance suivent la loi de Zipf, selon laquelle un petit nombre de règles s'appliquent fréquemment et assurent une bonne couverture, mais de nombreuses règles supplémentaires sont nécessaires à une amélioration du rappel.

Nenadić et Spacić [2000] ont mis au point un système de reconnaissance et d'acquisition de termes composés (unités lexicales comportant plusieurs formes) pour une langue fortement flexionnelle comme le Serbo-Croate. Ce système a été élaboré au LADL (Laboratoire d'Automatique Documentaire et Linguistique) et intégré au système *Intex*. Parmi ces termes composés, ils limitent leur étude aux entités nommées, car elles varient tant structurellement qu'au niveau lexical et qu'elles sont incessamment renouvelées.

L'étude de Nenadić et Spacić [2000] a été effectuée sur un corpus de textes issus de journaux yougoslaves présents sur le *Web*. Un prétraitement lexical est effectué sur ce corpus à l'aide d'un dictionnaire électronique (e-dictionnaire) qui sert de lemmatiseur : chaque entrée de ce dictionnaire est constituée d'un triplet (*forme*, *lemme*, *code morpho-syntaxique*) ; les paramètres de ces codes morpho-syntaxiques sont le genre, le nombre, le cas et le trait animé/inanimé. À la suite de cette lemmatisation, Nenadić et Spacić [2000] procèdent à d'autres traitements (désambiguïsation, reconnaissance des unités syntaxiques, etc.). Enfin, une grammaire locale (transducteur d'états finis), qui décrit des séquences de mots bien formées, est appliquée pour réduire encore les ambiguïtés lexicales.

Le traitement des entités nommées commence par la définition d'une grammaire locale décrivant la structure générale des entités nommées, en s'appuyant sur des connaissances théoriques. Cette grammaire est ensuite appliquée sur le corpus préalablement étiqueté, afin d'isoler les séquences correspondant à celle-ci. Cependant, ces séquences regroupent des syntagmes nominaux dont certains ne sont pas des entités nommées.

Ensuite, pour affiner la reconnaissance, Nenadić et Spacić [2000] intègrent, à la grammaire locale, des lexiques de mots-clés dénotant un type d'entités nommées : *fakultet* (lit. faculté), *institut*, *preduczecye* (lit. entreprise), etc. Ces lexiques sont obtenus en filtrant manuellement des listes recueillies en isolant la tête des syntagmes nominaux.

L'étape suivant consiste à générer les flexions de chaque mot-clé. En effet, les mots-clés peuvent prendre différentes formes fléchies, alors que le reste de l'entité nommée reste figé. Les éléments de chaque ensemble de flexions sont ensuite regroupés selon leurs caractéristiques lexicales (contexte gauche ou droit du mot-clé) et grammaticales. Un ensemble minimal de caractéristiques lexicales et morpho-syntaxiques inhérentes à chaque catégorie d'entité nommée[13] est ainsi créé, de sorte que toute autre étiquette peut être négligée dès lors que l'entité nommée

[13]Ces catégories sont plus fines que celles de MUC (p. ex. entités nommées ayant trait à l'éducation et aux sciences, noms d'entreprises et organisations du même type).

conserve une caractéristique de cet ensemble.

Enfin, de ces ensembles minimaux, sont générées des grammaires locales représentées par des graphes. Certains de ces graphes, similaires du point de vue de leurs constituants lexicaux, sont alors réunis pour former un graphe généralisé, afin de diminuer les redondances jusqu'à un certain niveau. Quelques autres heuristiques sont également utilisées :

- la première lettre d'une entité nommée doit être une majuscule, à moins que l'entité nommée soit entre guillemets ;
- chaque mot-clé dans le graphe doit être au singulier ;
- Nenadić et Spacić [2000] avance l'idée d'utiliser des anti-dictionnaires contenant les mots pouvant apparaître fréquemment dans le contexte d'un mot-clé sans faire partie de l'entité nommée.

Selon Nenadić et Spacić [2000] cette méthode est efficace pour acquérir des grammaires locales reconnaissant les termes composés (et particulièrement les entités nommées). Bien que Nenadić et Spacić [2000] affirment que leur méthode n'induit aucune fausse reconnaissance, il est regrettable qu'ils ne fournissent aucune évaluation de celle-ci.

2.2.4 Traitement des coréférences

La plupart des systèmes de reconnaissance des entités nommées procèdent à une recherche des formes coréférentes à certaines catégories d'entités nommées, essentiellement les noms de personnes et d'entreprises.

Nominator, le programme développé par Wacholder et coll. [1997] reconnaît certaines variations des entités nommées, comme les abréviations et certains types d'ellipses et effectue un regroupement sous une forme canonique. Par exemple, *American Bar Association* et *ABA* sont regroupés ainsi que *Robert Jordan* et *Mr. Jordan*.

Wolinski et coll. [1995] étudient également certaines variations en mettant en relation les formes les plus courantes d'une même entité nommée lorsque celle-ci est présente dans les dictionnaires ou, lorsque ce n'est pas le cas, en étudiant le « contexte global » de l'entité nommée : pour chaque entité nommée inconnue, Wolinski et coll. [1995] regardent dans le texte à la recherche d'une entité nommée catégorisée dans laquelle elle apparaît. Dans ce cas, ils établissent un lien entre ces deux entités nommées afin de transférer les attributs de celle qui a été préalablement catégorisée vers celle qui était inconnue. Cependant, certaines entités nommées peuvent partager un même radical : *Mr Mitterand* et *Mrs Mitterand* ou *Mr Bolloré* et *Bolloré Group*. Wolinski et coll. [1995] ne proposent pas de solution générale pour ce problème, mais affirme en résoudre les cas les plus fréquents.

Wakao et coll. [1996] étudient la résolution des coréférences pour les entités nommées afin de reconnaître les variations, essentiellement pour les noms d'entreprises. Pour Wakao et coll. [1996], lorsqu'un nom d'entreprise est utilisé pour la première fois comme *Ford Motor Co.* ou *Creative Artists Agency*, il le sera sous sa forme pleine, alors que dans la suite du texte l'auteur utilisera probablement une séquence abrégée : resp. *Ford* ou *CAA*. Pour déterminer quand associer deux entités nommées, 31 heuristiques ont été développées pour les noms d'organisations, onze pour les noms de personnes et trois pour les noms de lieux. Prenons deux noms *N1* et *N2*, il pourrait exister les heuristiques suivantes :

- si *N1* est une sous-séquence de *N2* dont les mots commencent par un majuscule, alors associer *N1* et *N2* (p. ex. *American Airlines Co.* et *American Airlines*) ;
- si *N1* est un nom de personne et *N2* est le prénom de *N1*, le patronyme de *N1* ou les deux, alors associer *N1* et *N2* (p. ex. *John J. Major Jr.* et *John Major*).

Le système réalisé par Poibeau [2002] repère de nouvelles entités nommées parmi les mots inconnus ou ceux qui n'ont pu être étiquetés par une des heuristiques mises au point manuellement (cf. section 2.2.2), ce, en s'appuyant sur les éléments déjà identifiés. Cette méthode de repérage consiste à étiqueter, le cas échéant, tout mot inconnu par le même type que celui d'une entité nommée déjà reconnue dont il est une sous-séquence. Pareillement, les mots inconnus appartenant à des structures énumératives contenant des entités nommées préalablement reconnues sont catégorisés de la façon identique. Ces techniques supposent que les textes analysés ont une certaine cohérence qui se manifeste par la reprise des mêmes éléments au sein du texte (hypothèse qui est souvent vérifiée dans les textes spécialisés).

En se fondant sur la théorie de Mooney [1993], Poibeau [2002] utilise également les coréférences pour réviser des étiquettes déjà posées. Les étiquettes posées jusqu'alors vont être modifiées si un contexte l'exige. Ainsi, les occurrences isolées du mot *Washington* sont étiquetées comme noms de lieu, mais si le système rencontre la séquence *Ms. Washington*, il va inférer que, dans le texte en question, *Washington* désigne plutôt une personne que la ville. Cette révision est limitée au texte en cours de traitement et non au corpus, afin de circonscrire les conflits de types.

L'intérêt d'un tel système est sa période d'adaptation qui permet de tenir compte du type de texte, et ce, à moindre coût (environ trois heures pour le corpus issu du journal *Le Monde*).

Après les différentes étapes de reconnaissance (entités nommées et scénarios), Grishman [1995] procède à une résolution des coréférences. Si un groupe nominal possède un déterminant ou un quantificateur indéfini (*un*, *quelques*, etc.), il est considéré comme information nouvelle. Autrement, Grishman [1995] cherche un antécédent possible, d'abord dans la phrase (de droite à gauche), puis dans la précédente (de gauche à droite), puis dans la précédente, etc. Un antécédent est accepté si :

- la classe du groupe nominal (dans la hiérarchie de Grishman [1995]) est égale ou plus générale que l'antécédent ;
- le groupe nominal et son antécédent sont de même nombre ;
- les modificateurs du groupe nominal trouvent une correspondance avec ceux de son antécédent.

Parmi les systèmes de reconnaissance des entités nommées que nous avons étudiés, la recherche des coréférences reste limitée à quelques catégories d'entités nommées et s'effectue exclusivement sur une base graphique. Les autres types de variations (métaphoriques, morpho-syntaxiques, elliptiques, anaphoriques, etc.) ne sont en aucun cas prises en compte. Néanmoins, il y a des hypothèses saillantes et pertinentes à retenir.

2.3 Méthodes d'apprentissage automatique (*machine learning*)

Les systèmes à base d'apprentissage automatique essaient généralement d'associer, à chaque forme, une des quatre étiquettes début, milieu, fin ou en dehors d'une entité nommée. Les mécanismes d'apprentissage sont le plus souvent les modèles de Markov cachés, les modèles de

maximisation de l'entropie et les arbres de décision. Parmi les différents systèmes de ce type, on peut distinguer deux catégories différentes :

Les systèmes avec corpus annotés présentent le désavantage de nécessiter de grands volumes de textes qui doivent être balisées par des experts. Souvent, ce coût dépasse celui de la réalisation manuelle ou semi-automatique de listes d'entités nommées [Poibeau, 2002]. L'absence de tels dictionnaires n'est donc pas un bon argument pour ce genre de systèmes.

Les systèmes sans corpus annotés sont rares. Les seuls que nous ayons trouvé nécessitent tout de même le recours à un analyseur grammatical et à un ensemble de règles initiales [Collins et Singer, 1999] ou d'amorces [Niu et coll., 2003]. Il est évident que sans corpus annotés, il faut trouver un autre moyen d'extraire l'information nécessaire à la phase d'apprentissage automatique. En outre, ces méthodes requièrent un volume de données d'apprentissage très élevé (971 746 phrases du *New York Times* pour Collins et Singer [1999] et huit millions de mots issus de textes journalistiques pour Niu et coll. [2003]).

Les méthodes à base d'apprentissage peuvent également être employées pour des modules plus spécifiques de la reconnaissance d'entités nommées (p. ex. détermination du caractère propre ou commun des mots ayant une majuscule à l'initiale après un point [Mikheev, 1999]).

2.3.1 Avec corpus annoté

Dans **Nymble**, Bikel et coll. [1997] utilisent un modèle classique de Markov à états cachés pour l'identification à partir d'un corpus d'apprentissage annoté des entités nommées MUC. Ce modèle comporte huit classes d'états ; chaque classe d'états correspondant à une classe d'entités MUC. À l'intérieur de chaque classe d'états, les auteurs utilisent un modèle bigramme. De manière à améliorer ce modèle ergodique, les auteurs associent à chacun des mots du corpus un trait dénotant sa graphie : par exemple, le mot *Sally* reçoit le trait `initCap`, le chiffre *1990* le trait `fourdigitNum`. Ces différents traits permettent de réduire considérablement le nombre de transitions et la taille du corpus d'apprentissage. Ils expérimentent leur modèle sur l'anglais et l'espagnol et annoncent une valeur de 90 % pour la F-Mesure dans la reconnaissance des entités nommées MUC. Ces résultats sont intéressants mais se limitent, en ce qui concerne les entités nommées, à ceux commençant par une majuscule et qui ne constituent qu'un sous-ensemble restreint de l'ensemble des entités nommées.

Il est à noter qu'**Nymble** est le système de reconnaissance des entités nommées utilisé par le centre de recherche américain BBN [Miller et coll., 1998].

En accord avec la définition de MUC, Palmer et Day [1997] évaluent la performance d'un programme de reconnaissance des entités nommées qui n'utilise *a priori* aucune connaissance linguistique, ni dictionnaire, ni liste de mots, n'effectue aucun traitement préliminaire sur le corpus, ni segmentation, ni pré-syntaxe, ni étiquetage morpho-syntaxique ou morphologique. À l'aide de simples expressions régulières n'utilisant pas la graphie, dépendantes toutefois de la langue, ils atteignent un score de 95 % pour la F-Mesure en ce qui concerne la reconnaissance des expressions numériques (NUMEX) et des expressions temporelles (TIMEX). La reconnaissance des entités nommées s'effectue à l'aide d'un corpus d'apprentissage annoté où 25 % des entités nommées sont manuellement identifiées. Le programme se charge juste de projeter dans le reste du corpus les entités nommées précédemment relevées, sans même prendre en compte leurs possibles variations. Leur expérience montre que dans un corpus homogène les entités nommées

se répètent, à l'inverse des corpus hétérogènes. Par exemple, pour les entités nommées pré-identifiées du français et pour un extrait du corpus Le Monde, la projection d'une partie des entités nommées rencontrées sur le reste du corpus couvre 23 % du total des entités ENAMEX à identifier. Palmer et Day [1997] obtiennent donc pour le français un score minimal de 34,5 %, toutes entités nommées confondues.

Le système **MENE** (*Maximum Entropy Named Entity*) de Borthwick [1999] se place dans le cadre de la théorie de la maximisation de l'entropie [Ratnaparkhi, 1997] et intègre diverses sources d'information : traits lexicaux, traits graphiques et traits indiquant la partie du texte (p. ex. titre ou corps du texte). Les lexiques[14] se composent de mots simples et de termes polylexicaux (noms de personnes, d'organisations, prénoms, suffixes d'entreprises, etc.) et sont de simples listes créées manuellement ou sont téléchargés automatiquement à partir du *Web*. La clef de la maximisation de l'entropie tient à ce qu'elle permet au concepteur de ne se focaliser que sur la recherche des traits caractérisant le problème, laissant à la routine d'estimation le soin d'assigner les poids relatifs à chaque trait.

Dans l'étiquetage des entités nommées réalisé par **MENE** dans le cadre de MUC-7, chaque forme se voit assigner une des $4n+1$ (ici 29 avec $n = 7$, le nombre de catégories) étiquettes parmi `x_start`, `x_continue`, `x_end`, `x_unique` (ou `x` est la catégorie de l'entité nommée) et `other` si la forme n'est pas une entité nommée.

La première étape de l'apprentissage[15] pour la modélisation par maximisation d'entropie consiste, pour le concepteur, à définir l'ensemble de traits. Après que le corpus d'entraînement a été segmenté et étiqueté manuellement, un fichier d' « évènements » est créé : il s'agit de paires $\langle h, f\rangle$ où f est un « futur » (une des étiquettes possibles) et h un « historique » (un contexte de un ou deux mots à droite et à gauche). Pour chaque évènement $\langle h, f\rangle$ et chaque trait g_i, $\#g_i$ recense le nombre de fois où $g_i(h, f)$ est déclenchée. Ensuite, seuls les traits tels que $\#g_i \geq 3$ sont conservés avec leur espérance ($K = \frac{\#g_i}{|C|}$). Enfin, dans une dernière étape, le programme d'estimation du maximum d'entropie (ME) de l'outil MEMT [Ristad, 1996] est appliqué afin de produire le modèle.

L'application du modèle à un nouveau texte se fait en deux étapes :

1. Calculer, pour chacune des formes, la probabilité conditionnelle des 29 futurs en fonction des différents traits du modèle.
2. Réaliser une recherche de Viterbi, afin de trouver la séquence valide de plus haute probabilité.

MENE est testé seul ou à la suite d'autres systèmes [Grishman, 1995 ; Lin, 1998b ; Krupka et Hausman, 1998] et avec différents volumes de données d'entraînement. Borthwick [1999] en conclut que le corpus d'entraînement doit comporter un minimum de 20 textes pour obtenir des performances acceptables (80 % de F-Mesure). D'autre part, sur les données d'entraînement, **MENE** améliore sensiblement les systèmes **Manitoba** et **Proteus** (environ 3 %). En revanche, sur les données inconnues, le gain est très faible, puisque **MENE** combiné aux trois autres systèmes en même temps ne surclasse **IsoQuest** que de 0,4 %.

Béchet et coll. [2000] présentent une méthode d'étiquetage sémantique des entités nommées fondée sur des arbres de décision. Ces derniers sont construits automatiquement sur un corpus

[14] Le nombre total d'entrées de ces lexiques est d'environ 23 000.

[15] Le corpus d'entraînement est composé de 350 articles de catastrophes aériennes, d'un total d'environ 270 000 mots.

d'apprentissage étiqueté, puis utilisés pour catégoriser les entités nommées d'un corpus de test. Les résultats de cet étiquetage sont utilisés pour enrichir un lexique des entités nommées qui, lui-même, peut être utilisé pour recalculer les paramètres d'un étiqueteur stochastique.

La classification des entités nommées s'effectue selon cinq catégories : les prénoms (`PRENOM`), les noms de famille (`FAMILLE`), les villes (`VILLE`), les pays (`PAYS`) et les organisations (`ORG`).

Les arbres de décisions utilisés sont des ACS (arbres de classification sémantiques). Ils ont été introduits par Kuhn et Mori [1995] et sont utilisés ici pour attribuer une étiquette sémantique à une entité nommée, en fonction des groupes nominaux dans lesquels elle apparaît. À chaque nœud de l'arbre, est associé une expression régulière construite sur un alphabet composé d'éléments lexicaux, d'étiquettes (parties du discours) et de symboles[16]. Chaque expression régulière est vue comme une question, celle de l'appartenance d'un groupe nominal au langage reconnu par cette expression. De plus, à chaque feuille de l'arbre est associée une distribution de probabilités sur les cinq étiquettes sémantiques qui permet d'estimer la probabilité qu'une entité nommée appartienne à chacune des classes. L'étiquette de probabilité maximale est appelée le *candidat* de la feuille.

La construction de l'ACS nécessite un corpus d'exemples, un ensemble de questions, un critère de division d'un nœud et une condition d'arrêt.

Le corpus d'exemples est constitué de groupes nominaux (entre un et douze mots) comprenant chacun une entité nommée appartenant à une classe sémantique. Il représente le corpus d'apprentissage sur lequel l'ACS sera construit.

Lors du processus de construction de l'ACS, chaque nœud est associé à une partie du corpus d'exemples, ainsi qu'à une expression régulière appelée la *structure connue* (SC). Au début de cette construction, la racine est associée à tous le corpus et à la SC `< + >`, qui reconnaît l'ensemble des groupes nominaux possibles. Les différentes expressions régulières pouvant être associé à un nœud sont construites à partir de la SC et de l'ensemble E composé du vocabulaire et des étiquettes morpho-syntaxiques. La construction consiste à remplacer successivement chaque symbole `+` de la SC par `i`, `+ i`, `i +` et `+ i +`, où `i` décrit l'ensemble E. Le nœud associé à la SC `< + président + >` générera, par exemple, lorsque `i` vaut *de* les huit expressions régulières suivantes :

`< de président + >`	`< + de président + >`	`< de + président + >`	`< + de + président + >`
`< + président de >`	`< + président + de >`	`< + président de + >`	`< + président + de + >`

Chaque question (expression régulière) partage donc en deux le corpus associé à un nœud. Il faut donc choisir à chaque fois la question la plus discriminante. Pour cela, les auteurs utilisent le critère d'impureté de Gini [Breiman et coll., 1984], qui est une mesure de l'homogénéité d'un ensemble. La question retenue est celle qui provoque la baisse maximale d'impureté entre un nœud et ses deux descendants directs. Lorsque tous les éléments d'un nœud ont la même étiquette l'impureté est minimale et vaut à zéro.

Lorsqu'un nœud ne contient plus qu'un groupe nominal ou que son impureté est inférieure à un seuil, il n'est plus scindé en deux.

L'arbre est donc ainsi créé et les groupes nominaux étiquetés contenus dans les feuilles vont permettre de calculer une distribution sur l'ensemble des cinq étiquettes. Plus cette distribution sera uniforme, moins l'expression associée à la feuille sera représentative d'une classe sémantique.

[16] Les symboles < et > matérialisent le début et la fin d'un groupe nominal, tandis que le symbole + correspond à une séquence quelconque composée d'au moins un mot.

Une feuille est appelée *discriminante* lorsque la probabilité de son candidat (*discriminance*) dépasse un certain seuil (*seuil de discriminance minimum*).

L'utilisation naturelle de cet ACS consiste simplement à présenter un groupe nominal à la racine de l'arbre et à parcourir l'arbre en fonction des réponses aux questions associés aux nœuds, jusqu'à atteindre une feuille ; l'entité nommée sera étiqueté par le candidat de la feuille si la discriminance de celle-là est supérieur au seuil de discriminance minimum (S_d).

Afin d'évaluer les capacités de l'ACS à étiqueter les entités nommées, une expérimentation a été réalisée sur deux corpus C_0[17] et C_1[18] extraits du *Monde* des années 91-92 et composés de groupes nominaux contenant des entités nommées. Les trois mesures utilisées sont les suivantes :

Précision : rapport du nombre d'entités nommées correctement étiquetées par l'ACS sur le nombre d'entités nommées étiquetées.

Couverture syntaxique : proportion des occurrences de groupes nominaux du corpus de test qui sont considérés comme discriminantes par l'ACS.

Couverture lexicale : proportion des entités nommées apparaissant au moins une fois dans un contexte discriminant.

Les résultats montrent qu'une bonne précision nécessite un seuil de discriminance élevé, correspondant à une faible couverture syntaxique. Par conséquent, l'ACS ne peut être utilisé directement comme classifieur. Béchet et coll. [2000] expliquent cela par l'ambiguïté de certains contextes et la limitation du contexte au seul groupe nominal[19]. Cependant, pour des niveaux élevés de discriminance, la précision est élevée : pour un seuil de 0,9 la précision atteint 95 %. De plus, à ce niveau de discriminance minimale, le rappel lexical reste élevé. C'est pour ces raisons que l'ACS va être utilisé *indirectement* pour la mise à jour de lexiques d'entités nommées (cf. section 2.4.1.1).

Sekine et coll. [1998] emploient également un arbre de décision pour identifier et catégoriser les entités nommées du japonais dans le cadre de la conférence MET-2[20] (en conjonction avec MUC-7).

Cet arbre de décision comporte trois types d'ensembles de caractéristiques :

- un ensemble de catégories des parties du discours obtenue grâce à l'étiqueteur JUMAN [Matsumoto et coll., 1997]. Cet ensemble regroupe les catégories mineur et majeurs de JUMAN ;
- des informations sur le type des caractères (Kanji, Hiragana, Katakana, alphabet, nombre symbole, etc.) ;
- des lexiques spécialisés (entités nommées et mots-clés) fondés sur les entrées des dictionnaires de JUMAN, ainsi que des listes trouvées sur le *Web* ou élaborées manuellement.

L'arbre de décision donne une information sur le début et la fin de chaque mot. Cette information peut prendre quatre valeurs : début d'une entité nommée (`OP`), continuation d'une entité nommée (`CN`), fin d'une entité nommée (`CL`) ou pas d'entité nommée (`none`). Par exemple, si un nom d'organisation couvre les mots A,B et C et que le mot suivant, D, n'est pas une entité nommée, on aura :

[17] 1,2 million de groupes nominaux contenant une entité nommée connue (dictionnaire de 265K mots).

[18] 695 groupes nominaux contenant une entité nommée qui apparaît au plus quatre fois (282 entités nommées différentes).

[19] Collins et Singer [1999] ont recours à un contexte plus large.

[20] En plus des catégories MUC, MET utilise la catégorie *Position* contenant différents titres liés à un nom de personne (*President*, *Professor*).

A : `org-OP-CN`
B : `org-CN-CN`
C : `org-CN-CL`
D : `none`

Pour éviter les séquences incohérentes, chaque feuille de l'arbre contient la probabilité de chaque étiquette. La sortie est obtenue à partir de la séquence cohérente de plus haute probabilité. L'algorithme de Viterbi est utilisé lors de la recherche de cette séquence (linéaire dans la taille des données). L'évaluation porte sur deux domaines journalistiques différents. Pour le premier (celui pour lequel le système était initialement créé), les performances atteignent une F-Mesure de 85 % pour un corpus d'apprentissage de 103 articles (2 368 entités nommées). Après quelques modifications (création d'un nouveau dictionnaire de *position* et modification du programme), Sekine et coll. [1998] obtiennent une F-Mesure de 84 %.

Renals et coll. [1999] comparent le comportement de deux systèmes de reconnaissance fondamentalement différents face à des données de mauvaises qualités (ici un corpus d'actualités radiodiffusées).

Le premier système (**Sprach-S**) consiste dans un modèle de langage (LM) statistique pour la reconnaissance des entités nommées. Ce modèle est obtenu à partir d'un corpus d'apprentissage étiqueté. Il s'agit d'un modèle n-gramme « *backed-off* » où le vocabulaire est l'ensemble des mots les plus fréquents annotés de l'information sur leur catégorie d'entité nommée. Les extensions unigrammes pour les entités de basse fréquence sont ajoutées, afin d'augmenter la taille du vocabulaire général. Pour l'évaluation, trois modèles de langage trigrammes (basés sur trois corpus d'entraînement différents[21]) sont utilisés, chacun avec un vocabulaire indépendant et des extensions unigrammes.

En combinant les trois modèles de langage, Renals et coll. [1999] présentent des résultats dont la F-Mesure est inférieure de 5 à 10 % à ceux de MITRE et BBN[22].

Le second système (**Sprach-R**) est lui un système fondé sur des règles écrites « à la main ». Il s'agit d'une version restreinte et légèrement modifiée du composant de reconnaissance des entités nommées de **LaSIE-II** Humphreys et coll. [1998], conçu pour des textes journalistiques correctement ponctués et de casse mixte. Les résultats obtenus en terme de F-Mesure sont 20 à 25 % moins bons que ceux rapportés aux conférences MUC-6 et MUC-7.

Les erreurs faites par **Sprach-R** présentent trois natures différentes :

1. **Erreurs de portage**. Elles surviennent à cause d'une adaptation trop rapide de **LaSIE-II** et d'une insuffisance des tests réalisés.
2. **Erreurs liés au genre de corpus**. En effet, contrairement aux textes journalistiques écrits, les corpus utilisés ici ne possèdent pas de segmentation du discours clairement identifiable, ce qui nuit au traitement des coréférences. De plus, les titres de personnes (p. ex. *Mr.*, *Dr.*) et les indicateurs d'entreprises (p. ex. *Inc.*, *Ltd.*) sont nettement moins présents. Enfin, les exemples d'entités nommées sont deux fois moins nombreux.
3. **Erreurs liés au format de corpus**. L'information perdu dans un texte au format SNOR (Speech Normalized Orthographic Representation) – telle la casse, la ponctuation ou les

[21] **H4** corpus de transcriptions d'actualités radiodiffusées (*Hub-4E*).
BN96 corpus de textes issus de BN 1996.
NA98 corpus de textes issus de journaux américains de 1996 à 1998.

[22] MITRE et BBN sont deux centres de recherche américains ayant obtenu de bons résultats aux conférences MUC [Miller et coll., 1998 ; Appelt et Israel, 1999 ; Boisen et coll., 2000].

nombres écrits en chiffres – est très dommageable pour la reconnaissance des entités nommées.

Ces deux derniers types d'erreurs sont étudiés plus précisément par Kubala et coll. [1998]. Les auteurs utilisent **Nymble** sur un corpus d'actualités radiodiffusées[23] et sur deux corpus de textes journalistiques écrits[24]. Sur ces derniers, le passage au format SNOR induit une perte de F-Mesure de 4,3 % pour NYT et 2,8 % pour WSJ (la différence entre ces deux corpus étant probablement due aux conventions de style plus strictes de WSJ).

Entre les corpus NYT et HUB-4[25], tous deux au format SNOR, la différence de F-Mesure n'est que de 4 % en faveur de NYT. Il est à noter que, sur des corpus oraux présentant un taux d'erreur (WER) de 20 % (HUB-4 possède un taux d'erreur de 0 %), cette perte s'élève à 14 %.

2.3.2 Sans corpus annoté

Collins et Singer [1999] proposent une étude sur l'utilisation des exemples non balisés pour la classification des entités nommées, dans le cadre des catégories MUC et pour l'anglais.

La tâche consiste à apprendre une fonction qui à partir d'une chaîne d'entrée (une entité nommée) donne son type. Leur approche considère deux types de règles : les règles *graphiques* et les règles *contextuelles*, utilisant implicitement les évidences internes et externes de McDonald [1996].

Collins et Singer [1999] montrent que l'utilisation d'exemples non balisés peut sensiblement réduire le besoin de supervision dans la tâche d'apprentissage. Avec environ 90 000 de ces exemples, les méthodes décrites atteignent plus de 91 % de précision. La seule supervision consiste en sept règles de bases. La clé de ces méthodes tient dans la redondance des données. En effet, pour de nombreuses entités nommées, la graphie ou le contexte seuls sont suffisants pour leur classification (p. ex. *Mr. Cooper, a vice president* peut être catégorisé grâce au contexte *president* ou à sa graphie et la chaîne *Mr.*).

Les données du problème sont constituées de 971 746 phrases du *New York Times* dont l'analyse grammaticale a été réalisée par Collins [1996]. Les séquences de mots qui correspondent aux critères suivants sont extraites comme exemples d'entités nommées :

- la séquence est une suite d'entités nommées à l'intérieur d'un syntagme nominal dont la tête est le dernier mot de cette séquence ;
- le syntagme nominal contenant la séquence apparaît dans l'un des deux contextes :
  1. Il y a un modificateur apposé au syntagme nominal, dont la tête est un nom singulier (p. ex. *Maury Cooper, a vice president*).
  2. Le syntagme nominal est complément d'une préposition qui est la tête d'un syntagme prépositionnel. Ce syntagme prépositionnel modifie lui-même un autre syntagme nominal dont la tête est un nom singulier (p. ex. *a plant in Georgia*).

En plus des entités nommées (*Maury Cooper* et *Georgia*), les indices contextuels sont également extraits (*president* et *plan_ in*). La chaîne de caractères correspondant à l'entité nommée est appelée *graphie* et l'indice contextuel est appelé *contexte*, formant ainsi une paire pour chaque entité nommée. Celle-là n'est pas l'unique élément composant une configuration (ensemble de

[23] Corpus HUB-4 distribuées par le LDC (Linguistic Data consortium) entre 1996 et 1997.
[24] New York Times News Service (NYT) utilisé à MUC-7 et Wall Street Journal (WSJ) utilisé à MUC-6.
[25] **Nymble** n'est entraîné que sur le corpus appropriée : NYT pour NYT et HUB-4 pour HUB-4.

traits) nécessaire à la représentation de chaque exemple pour l'algorithme d'apprentissage. Les traits utilisés sont : la chaîne complète (*Maury_ Cooper*), les différents termes contenus dans la chaîne (*Maury* et *Cooper*), des informations sur la casse et les caractères autres que les lettres, le contexte (*president*) et le type de ce contexte (apposition ou préposition).

Collins et Singer [1999] comparent différentes méthodes se fondant sur les résultats de Yarowsky [1995] (**Yarowsky-cautious**) et Blum et Mitchell [1998] (**DL-COTrain**) et présentent également une alternative qui construit deux classifieurs en parallèle qui satisfont au mieux deux conditions. En effet, cette méthode fait l'hypothèse forte que l'ensemble des traits peut être partitionné selon deux types, de sorte que chaque type indépendamment permette la classification. Chaque exemple x se décompose donc en deux ensembles de traits (x_1, x_2) et la fonction f de classification, en deux fonctions f_1 et f_2 telles que : $f(x) = f_1(x_1) = f_2(x_2)$. Prenons n paires $(x_{1,i}, x_{2,i})$ où les m premières ont l'étiquette y_i et les autres sont non balisés. Alors :

1. $f_1(x_{1,i}) = f_2(x_{2,i}) = y_i$ pour i de 1 à m.
2. Le choix de f_1 et f_2 doit minimiser le nombre d'exemples où $f_1(x_{1,i}) \neq f_2(x_{2,i})$.

Cet algorithme (**CoBoost**[26]) présente l'avantage d'être plus général et de s'adapter à tout algorithme d'apprentissage supervisé. Enfin, Collins et Singer [1999] décrivent une application de l'algorithme **EM** (maximisation de l'entropie) à la reconnaissance d'entités nommées, pour laquelle les données observés sont $(x_1, y_1)...(x_m, y_m), x_{m+1}...x_n)$ et les données cachées sont $y_{m+1}...y_n$.

L'évaluation porte sur 88 962 paires (*graphie*,*contexte*) extraites des données. 1 000 d'entre elles sont prises au hasard et étiquetées manuellement. Une balise `noise` est posée pour celles qui sont en dehors des trois catégories de bases et les expressions temporelles sont exclues de l'évaluation car elles peuvent être facilement identifiées par ailleurs. Il reste donc 962 paires dont 85 sont étiquetées `noise`. Les trois algorithmes **Yarowsky-cautious**, **DL-COTrain** et **CoBoost** obtiennent des résultats similaires avec une précision d'environ 83 %, alors que **EM** obtient 76 %. Cette évaluation ne rend pas compte du rappel du fait qu'elle repose sur les paires extraites : elle ignore les paires qui n'auraient pas été extraites.

Niu et coll. [2003] proposent un système à base d'amorces pour catégoriser les entités nommée de l'anglais, considérant que l'identification est préalablement réalisée par un analyseur grammatical (parties du discours). Pour pallier le problème de la propagation d'erreur dans les systèmes de co-apprentissage (la précision baisse itération après itération), Niu et coll. [2003] procèdent à une succession d'apprentissages pour fournir des données d'apprentissage (entités nommées balisées) plus nombreuses.

Dans un premier temps, plutôt que des lexiques spécialisés ou des règles créées manuellement [Collins et Singer, 1999 ; Cucerzan et Yarowsky, 2002 ; Kim et coll., 2002], Niu et coll. [2003] utilisent, comme amorces à leur système, quelques noms communs et pronoms qui représentent le concept de chaque catégorie d'entité nommée (p. ex. *il*, *elle*, *homme*, *femme* pour les noms de personnes). Ces amorces partagent des propriétés grammaticales (et non structurelles) avec la catégorie qu'elles représentent et sont beaucoup plus nombreuses que les entités nommées de cette catégorie : il y a donc plus d'information contextuelle disponible pour l'apprentissage. De plus, cette méthode permet des modifications relativement intuitives et rapides, ce qui est mis en valeur par l'ajout de la catégorie `PRO` (noms de produits), en plus des catégories MUC.

[26] Pour généraliser cet algorithme à plus de deux balises, Collins et Singer [1999] étendent l'algorithme AdaBoost.MH [Schapire et Singer, 1998] au co-apprentissage.

En se fondant sur le raisonnement de Lin [1998a] selon lequel les mots conceptuellement similaires apparaissent dans des contextes structurellement similaires, les structures grammaticales contenant les amorces sont donc extraites du corpus d'apprentissage[27] afin d'« entraîner » une liste de décisions pour la classification des entités nommées. Au total, cette liste comporte 1 290 règles d'une précision supérieure auxquelles sont ajoutées les quatre règles suivantes (en haut de la liste) :

`IsA`(man) → `PER`
`IsA`(city) → `LOC`
`IsA`(compagny) → `ORG`
`IsA`(software) → `PRO`

La taille du corpus d'apprentissage, environ huit millions de mots issus de textes journalistiques, compense le faible taux de rappel dont souffrent d'ordinaire les systèmes d'apprentissage fondés sur l'étiquetage grammatical : ici, 35-40 % des entités nommées sont associées à l'une des cinq relations de dépendance et seulement 5 % d'entre elles environ sont reconnues par la liste de décision. Finalement, sur l'ensemble du corpus d'apprentissage, la liste de décision permet tout de même l'extraction de 33 104 `PER`, 16 426 `LOC`, 11 908 `ORG` et 6 280 `PRO`.

En plus de la traditionnelle hypothèse de Gale et coll. [1992] selon laquelle une entité nommée ne peut avoir qu'un sens par document, Niu et coll. [2003] ajoutent l'heuristique qui veut qu'une entité nommée complexe n'appartienne qu'à une catégorie par domaine. Selon le même schéma de propagation/élimination que Yarowsky [1995], si une catégorie prédomine nettement pour un candidat, alors toutes ses occurrences sont étiquetées comme telles, sinon les étiquettes sont éliminées. Après cette phase, 134 722 `PER`, 186 488 `LOC`, 46 231 `ORG` et 19 173 `PRO` sont extraits avec une précision d'environ 90 %.

La dernière étape d'apprentissage consiste à entraîner un modèle de Markov à états cachés, similaire à celui de Bikel et coll. [1997], sur cet ensemble d'exemples d'entités nommées accompagnées de leur contexte (réduit à un mot de la même phrase).

Comparativement à leur système de reconnaissance à base d'apprentissage supervisé existant [Srihari et coll., 2000], Niu et coll. [2003] souffrent d'une dégradation de la F-Mesure de 5 % pour `PER`, 6 % pour `LOC` et 34 % pour `ORG` (pour des valeurs respectivement égales à 87,7 %, 82,3 % et 52,7 %). Cette dégradation plus importante pour les noms d'organisations est due à la diversité des concepts que regroupe cette catégorie et à la difficultés de les représenter par des amorces. Pour les noms de produits, les performances atteintes sont 69,9 % pour la F-Mesure (les raisons de ce taux moyennement satisfaisant sont les mêmes que pour la catégorie `ORG`).

2.3.3 Autres traitements

Mikheev [1999] présente une méthode nécessitant très peu de connaissances *a priori*, pour identifier le caractère propre ou commun des mots commençant par une majuscule et se trouvant en position qui rend cette tâche ambiguë. Le principe consiste à regarder l'usage non-ambigu qui est fait de ces mots dans le reste du texte.

Dans un premier temps, pour chaque séquence de deux termes ou plus, déjà reconnue de façon non-ambiguë comme étant une entité nommée, Mikheev [1999] considère les sous-ensembles de ces termes comme étant eux-mêmes des entités nommées : pour *Rocket Systems Development*

[27] Le corpus est étiqueté par le système **InfoXtract** [Srihari et coll., 2003] qui fournit cinq types de relation de dépendance : `Has_Predicate`, `Object_Of`, `Has_Amod`, `Possess` et `IsA`.

Co., ces sous-ensemble sont *Rocket Systems*, *Rocket Systems Co.*, *Rocket Systems Development Co.*, *Rocket Co.*, *Systems Development*, etc. Ce procédé s'accompagne d'une évidence négative constituée des bigrammes de noms communs du texte, qui va bloquer celui-ci.

Ensuite, Mikheev [1999] traite les mots simples : si un terme est présent avec une majuscule dans le document, mais pas en minuscules, celui-ci est considéré comme étant une entité nommée s'il est retrouvé à une position ambiguë dans le même document. Inversement, si un terme est utilisé uniquement en minuscules dans un document, ce ne sera jamais une entité nommée.

Enfin, Mikheev [1999] utilise une liste de noms communs les plus fréquemment retrouvés en début de phrase (cette méthode pouvant être alternée avec la précédente).

Sur 2 677 termes en position ambiguë, 2 363 sont correctement désambiguïsés et neuf le sont incorrectement, ce qui nous donne des taux de 99,62 % pour la précision et 88 7 % pour le rappel. Sur les 305 termes restant, 275 sont des noms communs. En considèrant tous les termes non désambiguïsés comme des noms communs, la précision est alors de 98,54 % et le rappel de 100 %.

Mikheev [1999] propose une amélioration en utilisant un lexique des séquences d'entités nommées identifiées dans différents textes (99,13 % de précision). Cela dit, cette amélioration nuit à l'intérêt principal de cette méthode qui est le peu de connaissances *a priori* nécessaire.

2.4 Méthodes mixtes

Parmi les systèmes utilisant à la fois des méthodes linguistiques (lexiques et règles de réécriture) et des méthodes d'apprentissage automatique, on peut distinguer deux façons de procéder :

1. Après un premier étiquetage obtenu par des méthodes linguistiques, un apprentissage automatique est réalisé sur les entités nommées catégorisées, afin de compléter les lexiques ou les règles de réécriture. Un nouvel étiquetage à base de méthodes linguistiques est ensuite réalisé.
2. Le passage aux méthodes d'apprentissage automatique se fait également après un premier étiquetage à base de lexiques et de règles, mais ces méthodes sont utilisées directement pour reconnaître de nouvelles entités nommées.

2.4.1 Mise à jour

La reconnaissance des entités nommées repose le plus souvent sur l'utilisation de lexiques et d'heuristiques qui permettent d'obtenir de bons résultats pour un domaine donné. Malheureusement, lors du passage à un nouveau domaine, les lexiques sont incomplets ou inexistants et les heuristiques ne sont plus toujours appropriées. Il est alors nécessaire d'adapter les lexiques et les heuristiques au domaine d'étude en limitant le recours à une analyse manuelle.

2.4.1.1 Mise à jour des lexiques

Face à l'incomplétude des lexiques lors du passage à un nouveau domaine ou une nouvelle langue, Cucchiarelli et coll. [1999] proposent une méthode contextuelle pour catégoriser les entités nommées inconnues en se basant sur une analyse syntaxique et sémantique des contextes similaires [Ide et Véronis, 1998].

Cette méthode est totalement automatique : pas de balisage manuel du corpus d'entraînement. En revanche, elle nécessite une connaissance *a priori*, qui se présente sous la forme de lexiques spécialisés et de règles contextuelles (cf. section 2.2.2). Elle est fondée sur l'hypothèse

de Gale et coll. [1992] selon laquelle un nom ne peut avoir qu'un sens par document ; cette hypothèse, particulièrement réaliste en ce qui concerne les entités nommées, rend possible l'analyse de tous les contextes dans lesquels apparaît une entité nommée.

Une étape préliminaire consiste à sélectionner les meilleurs catégories de *WordNet* pour un domaine donné, en sélectionnant automatiquement :

- un ensemble de catégories représentant le mieux la sémantique du domaine ;
- un « juste » niveau d'abstraction, pour concilier sur-ambiguïté et sous-généralisation ;
- un ensemble de catégories équilibrées : les mots doivent être distribués de façon homogène entre les catégories.

Cette méthode retient quatorze catégories *WordNet* pour le même domaine, dont cinq correspondent aux quatre catégories d'entités nommées de MUC-7 pour les domaines financiers.

Après la première étape de la reconnaissance (cf. section 2.2.2), Cucchiarelli et coll. [1999] obtiennent des performances de 70,5 % pour la précision et 67,3 % pour le rappel. 20 % des entités nommées sont alors identifiées mais pas catégorisées. Les entités nommées inconnues sont identifiées à l'aide de l'analyseur de Brill [1995] et les plus complexes le sont partiellement par des heuristiques simples.

Plutôt que d'étendre manuellement les lexiques et les règles, Cucchiarelli et coll. [1999] explorent la possibilité d'augmenter les performances en utilisant un algorithme explorant automatiquement l'évidence externe et mettant à jour les lexiques. Cette méthode nécessite les ressources suivantes : un corpus d'apprentissage dans le même domaine d'application, un analyseur syntaxique partiel, un lexique de base et un dictionnaire des synonymes. L'idée générale consiste à collecter tous les liens syntaxiques (*esls*) dans lesquels l'entité nommée inconnue (U_PN) apparaît. Pour chacun de ces contextes syntaxiques, des contextes similaires sont recherchés (même type de lien et même second argument ou un synonyme[28]) dans la base contenant les liens syntaxiques de toutes les entités nommées (PN_esls). Pour chaque entité nommée et chaque catégorie un score d'évidence est attribué ; ce score est une combinaison de la plausibilité [Basili et coll., 1994] des *esls* similaires à ceux de l'entité nommée et de l'inverse de l'ambiguïté du second argument de ces *esls*. Finalement, partant de l'hypothèse que, dans un domaine donné, une entité nommée a un unique sens, la catégorie dont le score d'évidence est le plus haut lui est assignée.

L'évaluation de cette méthode se fait en trois étape[29] :

1. En retirant les 35 entités nommées les plus fréquentes de chaque catégorie (140 au total), Cucchiarelli et coll. [1999] testent uniquement l'algorithme d'apprentissage. Les résultats semblent bons pour les organisations, les lieux et les personnes, mais moins pour les produits.

2. Cucchiarelli et coll. [1999] testent maintenant cet algorithme en complément d'une première reconnaissance. Cette étape d'apprentissage est répétée jusqu'à ne plus apporter d'information (concrètement trois fois). Cela nécessite une vérification manuelle des éléments à ajouter aux lexiques. Ils obtiennent une précision et un rappel respectivement de l'ordre de 88 % et 84 %.

3. Dans cette évaluation, Cucchiarelli et coll. [1999] montrent qu'une généralisation qui s'éten-

[28] Un poids plus grand est donné lorsque le second argument est le même ($\alpha = 0,7$) que lorsqu'il s'agit d'un synonyme ($\beta = 0,3$).

[29] Elle est effectué sur un corpus d'apprentissage d'un million de mots extrait du journal économique *Il Sole 24 Ore*.

drait au delà de la synonymie ferait baisser les performances, alors que cette simple généralisation les augmentent significativement.

Ces résultats restent tout de même moins performants que les systèmes anglo-saxons développés pour les conférences MUC. Cependant, Cucchiarelli et coll. [1999] arguent que ces derniers effectuent la désambiguïsation sémantique sur un plus petit nombre de classes et que la généralisation de l'évaluation en terme de gain d'information (*Information Gain*) permettrait une meilleur comparaison des systèmes de complexités différentes.

Malgré les résultats de cette troisième étape d'évaluation, Cucchiarelli et Velardi [2001] étudient plus précisément les effets d'une généralisation de cette méthode en relâchant progressivement les contraintes sur la notion de similarité. Cette expérience est réalisée sur un corpus anglais de textes du *Wall Street Journal*, avec le système VIE [Humphreys et coll., 1996] pour la première reconnaissance, car *WordNet* n'est pas disponible pour l'italien. Une méthode d'élagage est ensuite appliquée à *WordNet* [Cucchiarelli et coll., 1998], afin de réduire les ambiguïtés initiales du contexte (environ 27 % de réduction).

Dans cette expérience, les contextes similaires sont ceux qui ont le même type de lien, ainsi que le même second argument ou un argument qui possède un hyperonyme commun H dans *WordNet* (et non plus un simple synonyme). L'utilisation de telles ressources pour la généralisation de contexte a déjà été étudiée [Agirre et Rigau, 1996 ; Brill et Resnik, 1994], sans toutefois en démontrer l'utilité du point de vue des performances.

Le niveau L de généralisation pour H est le nombre de niveaux entre H et le second argument du contexte syntaxique de U_PN dans *WordNet* élagué (une valeur de zéro signifie une simple relation de synonymie). Les meilleurs résultats sont obtenus avec une valeur de L égale à deux (83 % de F-Mesure contre 79 % pour $L = 0$). Cette méthode peut donc présenter un apport concret dans l'efficacité d'un système de reconnaissance des entités nommées.

Béchet et coll. [2000] utilisent également les étiquettes attribué par l'ACS (cf. section 2.3.1) pour mettre à jour le lexique des entités nommées et ainsi recalculer les paramètres d'un étiqueteur stochastique. Le corpus utilisé pour cela est constitué de textes issus du *Monde* en 1987 et 1991 (107K groupes nominaux différents extraits d'environ 98M mots). L'arbre comprend 10,5K nœud et 10,5K feuilles.

La mise à jour du lexique des entités nommées s'effectue en recalculant, pour chaque groupe nominal de C_1 contenant une entité nommée m, les probabilités $P(m|c)$ dans le modèle de l'étiqueteur stochastique [Charniak et coll., 1993]. Le nouveau modèle est alors utilisé pour étiqueter C_1 (T_{ACS}) et les résultats sont comparés à ceux obtenus par la technique qui consiste à considérer que les mots inconnus ont la même probabilité d'appartenir à chaque classe [Weischedel et coll., 1993]. Ces résultats montrent une baisse de performance pour des valeur de S_d supérieur à 0,5. En effet, pour de telles valeurs, peu des 282 entités nommées ont été rencontrées dans des contextes discriminants et leur entrée lexicale n'a pas pu être mise à jour. Il est donc intéressant d'augmenter le nombre de contextes (et donc de contextes discriminants) dans lesquels apparaît une entité nommée.

Pour ce faire, Béchet et coll. [2000] utilisent le *Web* pour recueillir de nouveaux exemples d'occurrences pour les entités nommées. Ce processus est automatique et n'engendre pas de surcoût pour l'acquisition de corpus. Cependant, il est nécessaire, étant donné la quantité et l'hétérogénéité des données du *Web*, d'éliminer les données non pertinentes (listes, tableaux,

textes non francophones, etc.), par un filtrage des pages résultats de la requête[30]. Dans un premier temps, les parties non textuelles sont ôtées des données, puis ces dernières sont nettoyées (traitement des accents, segmentation en occurrences de phrases et de mots) et enfin, traitées par un analyseur morpho-syntaxique. Les entités nommées inconnues sont étiquetées `INC`, les groupes nominaux sont extraits comme précédemment (avec `INC` au lieu des étiquettes sémantiques), puis sont analysés par l'ACS et ceux tombant dans les feuilles ayant un seuil supérieur à S_d sont conservés.

Sur les 282 entités nommées de C_1, la taille moyenne des textes *html* collectés est de 3Mo. À l'issue du nettoyage, seulement 110Ko de texte sont conservés pour chaque entité nommée, qui se retrouve en moyenne 14,5 fois dans ces données[31]. Les 5 000 groupes nominaux ainsi extraits sont ajoutés aux 695 de C_1 pour constituer C_2.

Après avoir recalculé les paramètres de l'étiqueteur stochastique grâce à C_2, le corpus C_1 est réétiqueté (T_{Web}). Bien que les résultats obtenus soient légèrement meilleurs que sur $T_{groupesnominaux}$, le gain reste marginal sur l'ensemble des 282 entrées (+9,8 % contre +5,2 % pour un S_d de 0,8). En revanche, si en considérant uniquement les entités nommées ayant été vues dans un contexte discriminant, l'apport du *Web* devient significatif : sur 50 % des entrées, le taux d'étiquetage passe de 40 % sur T_{ACS} à 90 % sur T_{Web}. Ce manque de performances est attribué au fort bruit qui subsiste dans ces textes, même après nettoyage, ainsi qu'au style et au domaine sémantique des corpus collectés qui diffèrent très souvent de ceux des corpus initiaux.

2.4.1.2 Inférence d'heuristiques

Gallippi [1996] propose une technique d'apprentissage qui, à partir de quelques heuristiques valides quelle que soit la langue, infère de nouvelles heuristiques caractéristiques du domaine d'étude. L'approche utilisée ici repose sur une stratégie d'acquisition à base d'arbres de décision qui utilisent des informations contextuelles. Dans un premier temps, chaque mot est étiqueté avec un ensemble de traits : morphologique, syntaxique, sémantique, partie du discours, *designators*[32], etc.

Dans un second temps, les entités nommées sont délimités grâce à l'utilisation de patrons basés sur les parties du discours qui ont été créés manuellement. Puis, pour chaque type d'entité nommée, les traits sont sélectionnés et organisés en un arbre de décision. Le résultat est une collection hiérarchique de caractéristiques cooccurrentes qui permettent de prédire l'inclusion ou l'exclusion par rapport à un type d'entité nommée. Enfin, les arbres générés sont appliqués sur le corpus et évalués. Une analyse manuelle conduit à la découverte de nouveaux traits qui sont ajoutés aux précédents et le processus est réitéré.

Cette technique qui a été évaluée sur différentes langues obtient des résultats satisfaisants[33] : 94,4 % pour l'anglais, 89,2 % pour l'espagnol et 83,1 % pour le japonais.

[30] La requête est effectué à l'aide d'un moteur de recherche, en spécifiant la langue désirée et un mot-clé : l'entité nommée.

[31] Il faut noter que 87,6 % des formes se retrouvent dans les textes collectés sur le *Web*.

[32] Seuls, ces traits fournissent une forte présomption sur l'appartenance ou la non appartenance d'un syntagme à une catégorie d'entité nommée.

[33] Nous indiquons ici la valeur moyenne de la *F-Mesure* pour les différentes catégories (type MUC) d'entités nommées.

2.4.2 Utilisation de méthodes à base d'apprentissage après un premier étiquetage linguistique

Mikheev et coll. [1999] étudient un système combinant règles et apprentissage automatique. Ce système, élaboré pour MUC-7 utilise les évidences interne et externe définies par McDonald [1996].

Lors d'une première étape, ce système déclenche des règles estimées sûres. Ces règles associent évidences interne et externe et nécessitent un étiquetage assignant les parties du discours et un étiquetage sémantique simple (adjectifs, professions, relatifs). À ce stade, les noms de lieux appartenant aux lexiques sont étiquetés comme tels, uniquement en présence d'un contexte favorable.

Après que ces premières règles ont été appliquées, le système effectue une association partielle des entités déjà reconnues : pour chacune d'entre elles, il génère toutes les combinaisons des formes les composant, en en préservant l'ordre, puis les marquent, si elles sont retrouvées ailleurs dans le texte, comme possibles entités nommées. Cette information est fournie à un modèle de maximisation de l'entropie pré-entraîné [Mikheev, 1998] qui prend en compte les informations contextuelles sur l'entité nommée (p. ex. sa position dans la phrase, son existence en minuscule dans le texte, en général, etc.). À ce stade, si le modèle produit une réponse positive, le système effectue une assignation ferme.

Une fois cette étape effectuée, le système applique de nouveaux les règles de grammaire, mais avec des contraintes contextuelles plus lâches. Notons que c'est seulement à ce stade que le lexique des noms de personnes est utilisé et que Mikheev et coll. [1999] ne se préoccupe plus du fait qu'un nom de personne peut désigner une organisation (ce problème devrait déjà avoir été résolu). C'est également durant cette phase que le système tente de résoudre le problème des conjonctions et des majuscules en début de phrase et qu'il marque les entités nommées présentes dans les lexiques sans vérifier le contexte dans lequel elles apparaissent.

Le système ayant épuisé ses ressources (lexiques et règles), une autre association partielle fondée de nouveau sur le modèle d'entropie maximum est effectuée.

Dans une dernière phase, Mikheev et coll. [1999] considèrent les entités nommées présentes dans les titres entièrement écrits en majuscules. Pour ce faire, ils utilisent uniquement la méthode à base d'apprentissage : association totale ou partielle des formes du titre avec des entités nommées reconnues dans le texte (avec variation de la casse), puis utilisation d'un modèle d'entropie maximum entraîné sur les titres du document.

Les résultats obtenus par ce système à la compétition MUC, étaient de 93,39 % (rappel et précision combinés).

Quasthoff et coll. [2002] présentent une méthode de reconnaissance des noms de personnes allemands nécessitant peu de connaissance *a priori* et réalisant un apprentissage non-supervisé sur des textes bruts.

Dans un premier temps, le texte est balisé selon deux types de balises différents, chaque mot pouvant en avoir plusieurs :

1. Balises dépendantes du problème. Pour les noms de personnes :
   - titre ou profession (TI) ;
   - prénom (FN) ;
   - nom (LN).
2. Balises indépendantes du problème, mais dépendantes du langage :

- mot commençant par une minuscule (LC) ;
- mot commençant par une majuscule (UC) ;
- marque de ponctuation (PM) ;
- déterminant (DET).

L'algorithme commence avec un ensemble de règles (quatorze) basées sur ces différentes balises et créées manuellement, ainsi qu'une base d'éléments de noms de personnes (neuf prénoms et dix noms) et comporte deux phases :

1. **Recherche de candidats** : à l'aide de ces données appliquées sur le corpus[34], une liste de candidats est générée.
2. **Vérification** : chaque candidat est testé avant d'être validé. Quasthoff et coll. [2002] vérifient, dans le corpus, si le candidat apparaît assez souvent en compagnie d'éléments déjà validés.

Ce cycle est réitéré tant que de nouveaux éléments de noms de personnes ont été découverts, en intégrants ces derniers dans la base. Quasthoff et coll. [2002] discriminent deux phases durant le processus :

1. Une phase d'expansion, durant laquelle le nombre d'éléments nouveaux augmente.
2. Une phase d'épuisement, durant laquelle le nombre d'éléments nouveaux diminue jusqu'à zéro.

La phase d'expansion dure jusqu'au cinquième cycle, puis cède la place à la phase d'épuisement jusqu'au onzième cycle. À la fin, on dénombre 25 000 éléments de noms de personnes, permettant la reconnaissance de 150 000 noms de personnes.

Cet algorithme fonctionne sous certaine conditions :

- taille du corpus suffisante ;
- fréquence des éléments de départ suffisante ;
- relation appropriée (p. ex. noms de personnes).

Après avoir éliminé les prénoms de plus de dix caractères, le rappel et la précision atteignent respectivement 97,5 % et 71,4 %. En Allemand, la plupart des mots mal catégorisés comme prénoms sont des titres ou des professions. Or, s'ils ne peuvent pas être différenciés par les règles utilisées, ils diffèrent fortement d'un point de vue morphologique (les titres sont généralement plus longs, car composés et certains composants sont utilisés très fréquemment). Par conséquent Quasthoff et coll. [2002] ont mis en place une méthode pour distinguer les prénoms des titres et des professions, en utilisant le fait que la formation des mots suit des règles dépendantes du langage. Pour cela trois classifieurs sont implémentés : un qui étudie la fin des mots, l'autre qui en examine le début et enfin le dernier qui en analyse chaque sous-mot. Cette distinction atteint une précision de 94,7 % et un rappel de 94,5 %.

Pour augmenter le taux de rappel, deux solutions sont possibles : ajouter des règles manuellement ou mettre en place un apprentissage de règles. Pour cette dernière, plus intéressante, Quasthoff et coll. [2002] proposent une méthode qui vise à générer des règles qui trouvent un maximum de candidats plutôt que des règles ayant une précision de 100 %. Le seul critère qu'elles doivent remplir, pour éviter d'être trop générales, est de contenir au moins l'une des balises spécifiques (c.-à-d. `FN`, `LN`, `TI`). Les règles obtenant une précision suffisante (ici 0,7) sont gardées pour les cycles suivants. La seule information chiffrée fournie par Quasthoff et coll. [2002] porte

[34] Il s'agit d'un corpus de 10 millions de phrases, tiré de journaux des dix dernières années, qui est utilisé à la fois pour l'identification des candidats et pour vérifier que ceux-ci sont bien des noms de personnes.

sur le fait que la précision des règle décroît sur la fin du processus du fait de la génération de règles moins strictes. En revanche, ils affirment que cet apprentissage de règle est positif en terme de rappel et de précision, mais ils ne donnent pas de résultats concrets.

2.5 Synthèse

Les systèmes de reconnaissance et de catégorisation des entités nommées pour l'indexation, que ce soit pour l'anglais ou pour le français, présentent de très bons taux de précision et de rappel. Cependant, ces systèmes, qui fonctionnent sur des domaines de spécialité, fournissent une catégorisation très générale limitée aux noms de personnes, d'organisations et de lieux, et n'identifient que les entités nommées simples ou composées marquées graphiquement par une majuscule. Ces systèmes examinent l'évidence interne et le contexte gauche (resp. droit) immédiat pour le français (resp. pour l'anglais), qui est comparé à des listes prédéfinies de mots déclencheurs comme les particules, les prénoms, etc. Le contexte droit (resp. gauche) est lui généralement ignoré pour le français (resp. pour l'anglais), ce qui implique que la délimitation à droite des entités nommées complexes s'arrête au premier mot plein sans majuscule. Les variations des entités nommées se limitent quant à elles à l'identification des sigles et des abréviations. L'évidence externe n'est donc pas beaucoup exploitée. Or, Wakao et coll. [1996] ont montrés que si l'identification d'un nom de lieu peut se limiter au traitement de l'évidence interne, l'analyse de l'évidence externe est en revanche indispensable pour l'identification d'un nom de personne ou d'organisation.

Les difficultés de la reconnaissance des entités nommées apparaissent dès lors que l'on cherche à obtenir une catégorisation plus précise par l'analyse du contexte gauche (la catégorie `ARTIFACT` nouvellement recherchée par les différents systèmes obtient de loin les moins bons résultats), que l'on veut améliorer la délimitation des entités nommées composées ou mixtes par l'analyse du contexte droit, ou encore que l'on souhaite prendre en compte toutes les variations des entités nommées. Une catégorisation précise nécessite de nombreuses listes de mots déclencheurs (liste de particules), des lexiques recensant des noms communs et des entités nommées catégorisants (listes de prénoms, de noms de métiers, etc.). Cette tâche se heurte à l'incomplétude de ces listes qu'il faudra mettre à jour à l'aide de méthodes automatiques d'acquisition, et à la poly-catégorisation de certains mots : par exemple, *France* est un prénom dans *France Roche* mais fait partie intégrante de l'entité nommée de société dans *France Telecom*. L'amélioration de la délimitation des entités nommées composées ou mixtes par l'analyse du contexte droit se heurte aux problèmes de la modification, de la résolution de l'attachement prépositionnel et de la portée de la coordination. Cette tâche a été effectuée pour l'anglais, à l'aide d'heuristiques et pour le français, à l'aide de grammaires décrivant les contextes droits autorisés pour une catégorie. Quant à la variation des entités nommées, les systèmes se limitent aux variations les plus faciles à identifier comme les variations graphiques, certaines ellipses et les sigles ou abréviations. Il reste donc à raffiner et à étendre à d'autres catégories les grammaires locales permettant de décrire les contextes droits autorisés des entités nommées et aussi à définir des méthodes d'apprentissage d'acquisition de ces contextes. Enfin, les variations morpho-syntaxiques ou métaphoriques n'ont jamais été considérées par les systèmes existants alors qu'elles sont très productives dans la langue.

Dans ce contexte, quelles méthodes adopter ? Malgré toutes les études menées sur les avantages et les inconvénients des différents types de systèmes, il ne paraît pas évident de déterminer

quel en est le meilleur. En effet, Sekine et Eriguchi [2000] constatent que « les trois premiers systèmes proviennent chacun d'une catégorie différente ; le meilleur système était fondé sur des règles écrites à la main, le second sur des règles acquises automatiquement et le troisième était totalement automatique. » [35] Si l'incomplétude des lexiques et la mise à jour des systèmes à bases de règles créées manuellement entravent le passage à de nouveaux corpus, il en va de même pour les systèmes à base d'apprentissage sur corpus annoté. En effet, un avantage proclamé de ces derniers consiste dans leur plus grande adaptabilité. Or, pour traiter des données issues d'une nouvelle source ou d'un nouveau domaine, il est nécessaire de ré-entraîner le système. Il faut donc de nouveau disposer d'un corpus annoté dont Borthwick [1999] évalue l'élaboration à trois semaine pour 300 000 mots, qui est la taille qu'il juge nécessaire pour l'apprentissage de son système[36].

Cependant, Poibeau [2002] montrent que les systèmes hybrides, utilisant différentes sources de connaissances, obtiennent les meilleurs résultats. Selon lui, « un système à base de règles peut aisément être adapté pour peu que l'analyste ait une bonne connaissance du système initial. Une bonne structuration des règles et une bonne ergonomie facilitent aussi la maintenance et l'adaptabilité d'un tel système ». Il met également en évidence la nécessité, pour les systèmes à base d'apprentissage sans corpus annoté, de disposer de « bons » exemples initiaux. De plus, le fait que ces systèmes ne disposent pas de dictionnaires initiaux reste un lourd handicap pour leurs performances globales (le rappel reste très faible), de même que le problème général de « sur-apprentissage », qui reste posé pour tout système fondé uniquement sur l'apprentissage. Cependant, cette technique paraît intéressante, non pas directement en extraction, mais pour améliorer un système qui effectue déjà une reconnaissance des entités nommées (pour mettre à jour les lexiques, pour inférer de nouvelles règles ou pour réviser le précédent étiquetage).

Nous avons donc décidé d'élaborer un système se fondant sur des règles de grammaire élaborées manuellement, intégrant les évidences internes et externes, exploitant des lexiques spécialisés de taille raisonnable et comportant des mécanismes d'apprentissage et de révision. Pour ne pas multiplier les traitements et donc le temps d'exécution, nous avons également choisi de ne pas utiliser de prétraitement grammatical (étiquetage syntaxique, des parties du discours, lemmatisation, etc.).

Quelles que soient les méthodes mises en place, les performances atteintes par le système dépendent de la catégorisation retenue. En effet, pour les catégories ENAMEX de MUC (noms de personnes, de lieux et d'organisations), les résultats obtenus sont meilleurs que pour une « nouvelle » catégorie comme les noms de produits, car elles ont été largement étudiées et sont relativement bien connues à présent : il existe donc de nombreuses ressources pour les traiter (dictionnaires, grammaires locales, etc.).

Pour choisir la typologie dans laquelle notre système devra catégoriser les entités nommées, nous avons décidé de partir d'une base typologique existante – celle de Grass [2000] (cf. section 1.4.1) – et de l'enrichir en s'appuyant sur une étude en corpus.

[35] the top three systems came from each category; the best system was a hand created pattern based system, the second system was an automatically created pattern based system and the third system was a fully automatic system.

[36] Bien que certains systèmes obtiennent des performances acceptables avec des corpus d'apprentissage plus petits, les performances continuent à croître avec la taille de ceux-ci [Kubala et coll., 1998].

Chapitre 3

Catégorisation des noms propres : étude préliminaire

Après avoir réalisé l'étude des différents travaux concernant la reconnaissance et la catégorisation des entités nommées, nous avons pu constater que les résultats obtenus pour le français, si l'on cherche à obtenir une reconnaissance exhaustive dans une catégorisation fine, ne sont pas satisfaisants [Trouilleux, 1997]. Nous avons donc opté pour une typologie des entités nommées basée sur des critères référentiels, qui tiennent d'une nature double : sémantique par le critère d'unicité du référent et pragmatique en ce qu'elle s'organise autour d'une évaluation des composantes de la réalité objective composant le référent du nom [Grass, 1999]. Nous avons également voulu cette typologie très détaillée et la plus complète possible. Ces paramètres semblent être les plus appropriés dans une optique de traduction ou de recherche d'information. Il nous sera toujours possible de réduire cette typologie en regroupant des catégories, afin de réaliser des tâches comme la veille économique ou l'indexation documentaire. L'extension d'une catégorisation trop générale serait très difficile et bien moins prévoyante.

D'autre part, les entités nommées possèdent, en français, certaines caractéristiques graphiques (voire lexicales ou syntaxiques) qui peuvent les distinguer des noms communs (cf. chapitre 1) : présence de la majuscule, absence de flexion morphologique (p. ex. absence de morphème flexionnel de nombre pour les sigles et les acronymes) ou d'article. En fonction de la présence ou de l'absence de ces caractéristiques, la reconnaissance d'une entité nommée s'avèrera plus ou moins facile à réaliser. Nous avons donc décidé d'établir également une typologie graphique et d'étudier la distribution des entités nommées selon les différentes classes de celle-là. De plus, certaines catégories référentielles peuvent avoir un lien fort avec des catégories graphiques : les sigles, par exemple, sont majoritairement liés à des noms d'organisations (bien que pouvant également être des pays comme *la RFA* ou *l'URSS* ou des noms d'institut de recherche comme *l'IRIN* ou *le CNRS*).

Pour l'un et l'autre de nos types de critères (graphiques et référentielles) et après avoir étudié différents travaux, nous avons retenu une typologie existante, afin de constituer le point de départ de notre propre catégorisation. Ensuite, nous avons effectué un étude en corpus afin de compléter et affiner ces deux typologies. Enfin, nous avons réalisé une analyse, sur ce même corpus, des distributions des entités nommées selon les deux catégorisations finales.

3.1 Bases typologiques

3.1.1 Base référentielle

Nous avons choisi d'adopter la typologie de Grass [1999] – inspirée de Bauer [1985] – comme base pour notre classification des entités nommées. Elle est fondée sur des critères pragmatiques et possède l'avantage d'être relativement complète et détaillée : elle constitue donc un bon point de départ pour notre étude.

En effet, une catégorisation comme celle proposée dans le cadre des projets MUC, MET ou encore IREX, est beaucoup trop pauvre, notamment pour la traduction ou la recherche d'information.

Prenons les catégories de la classe ENAMEX de MUC (noms de personnes, de lieux et d'organisations).

Dans le cadre d'une traduction de l'allemand vers le français, par exemple [Grass, 1999], cette classe regroupe à elle seule différentes entités nommées qui ne seront pas traduites de la même manière :

- parmi les noms de lieux, les installations militaires ou les pays se traduisent, alors que les villes (sauf les plus connues qui peuvent être facilement énumérées) ou les microtoponymes ne se traduisent pas ;
- parmi les noms de personnes, les ethnonymes se traduisent, tandis que les prénoms et les patronymes ne se traduisent pas (ou se transcrivent) ;
- parmi les noms d'organisations, les partis politiques et les organisations sont traduits, tandis que les entreprises industrielles ne le sont pas (sauf éléments distincts de la base dans les formes polylexicales).

Il n'est donc pas judicieux de les classer ensembles dans une optique de traduction.

D'autre part, certaines entités nommées ne trouvent aucunement place dans cette classification, comme les noms d'ensembles artistiques, d'évènements, d'œuvres, etc. Or, en recherche d'information notamment, il peut être très intéressant de reconnaître ces entités nommées, tout comme différencier les noms de villes, des noms de cours d'eaux ou d'édifices, afin d'extraire un maximum de quantité d'information.

En revanche, la typologie de Grass [1999] peut être comparée à celle de Paik et coll. [1996], car cette dernière est également détaillée et basée sur des critères pragmatiques. Les arguments qui nous ont fait pencher en faveur de celle de Grass [1999] sont les suivants :

- dans la catégorisation de Paik et coll. [1996], il y a certaines classes entières de Grass [1999] qui n'apparaissent pas, comme les *praxonymes* ou les *phénonymes*, alors qu'à l'inverse toutes celles présentes dans Paik et coll. [1996] peuvent trouver une place dans les classes ou les catégories de Grass [1999], hormis les catégories *drugs* et *chemicals* ;
- certaine classes de Paik et coll. [1996], comme la classe *organization*, sont divisées en catégories dont nous voyons mal sur quels critères elles ont été différenciées (*company*, *company type*, *government*, *U.S. government*, *organization*). Or, il ne nous semble pas nécessaire de créer une catégorie spécifique pour les États-Unis ;
- la typologie de Paik et coll. [1996] n'est motivée que par le taux de couverture que ses catégories atteignent sur l'ensemble des entités nommées d'un corpus, il n'est absolument pas fait état de justifications autres (sémantiques notamment). En effet, certaines classes pourraient, peut-être, être regroupées selon des critères sémantiques : par exemple, les

classes *Affiliation* et *Human* seraient regroupées dans la classe *anthroponymes* par Grass [1999]. Or, Paik et coll. [1996] ont séparé ces deux classes, mais on ne connaît pas les raisons et les choix qui ont amenés les auteurs à le faire. Cela rend également plus difficile l'extension de certaines classes ou catégories, car leurs fondements sémantiques restent flous.

Ce sont toutes ces raisons qui nous font opter pour la typologie de Grass [1999].

Grass [1999] énumère donc ce qui, par convention, constitue un nom propre, il prend en considération des éléments extra-linguistiques propres au référent. Il reprend l'idée de Bauer [1985] selon laquelle :

> « *La classification des noms propres s'organise autour d'une évaluation des composantes de la réalité objective composant le référent du nom.* » [1]

Cette classification, orientée vers le référent, est organisée autour de cinq classes principales : *anthroponymes* (noms individuels et collectifs), *toponymes* (nom de lieux au sens général), *ergonymes* (produits manufacturés, entreprises, œuvres intellectuelles), *praxonymes* (faits et évènements) et *phénonymes* (catastrophes naturelles, astres et comètes). Elles contiennent chacune des sous-catégories. Ces dernières, ainsi que des exemples les illustrant, sont explicitées au tableau 3.1.

3.1.2 Base graphique

Jonasson [1994] propose une classification pour le français, qui distingue trois types morphologiques et lexicaux pour les noms propres :

les noms propres « purs » sont composés par une forme lexicale spécialisée dans le nom propre (p. ex. *Paris*, *Valérie*, *Aristote*, *Majorque*, *la France*, *l'Atlantique*, *la Seine*) ;

les noms propres à base descriptive peuvent être constitués d'un ou plusieurs noms communs, éventuellement accompagnés de modificateurs adjectivaux ou prépositionnels (p. ex. *le Jardin des Plantes*, *l'Académie Française*, *le Grand Palais*, *le Centre national de la recherche scientifique*) ;

les noms propres à base mixte contiennent des noms propres purs ainsi que des noms communs et/ou des adjectifs (p. ex. *le Collège de France*, *le palais de Chaillot*, *la Nouvelle-Orléans*).

Allerton [1987] présente une typologie des noms propres de l'anglais, basée sur des critères morpho-syntaxiques, qui est très similaire à celle de Jonasson [1994]. Selon cette typologie, les noms propres sont réparties en quatre classes :

les noms propres purs ont la même définition que chez Jonasson (p. ex. *Los Angeles*, *France*, *Jordan*) ;

les noms propres basés sur des noms communs sont équivalents aux noms propres à base descriptive de Jonasson (p. ex. *Hyde Park*, *Trafalgar Square*) ;

les noms propres mixtes ont la même définition que chez Jonasson (p. ex. *Central Park*, *Main Street*) ;

[1] Traduction de : « Die Einteilung der Eigennamen richtet sich nach der Bewertung der den Namen als Referenten zugrunde liegenden Bestandteile der objektiven Realität. », réalisée par Grass [1999].

Tableau 3.1 – Typologie pragmatique des noms propres proposée par Grass [1999].

Anthroponymes	
Patronymes	*Delanoë*, *Zidane*
Prénoms	*Alexandre*, *Mohand Areski*
Ethnonymes	*le Serbo-Croate*, *les Italiens*
Partis et autres organisations	*la FFF*, *ATTAC*, *le Parti socialiste*
Ensembles artistiques	*Muse*, *l'Orchestre National de Barbès*
Pseudonymes	*l'Aigle des Açores*, *le Petit Père des Peuples*
Zoonymes	*Félix*, *Titi*
Toponymes	
Pays	*la France*, *le Sahara occidental*
Villes	*Paris*, *Belo Horizonte*
Microtoponymes	*le Quartier Latin*, *Prenzlauerberg*
Hydronymes	*la Seine*, *le lac Ontario*
Oronymes	*les Alpes*, *le Kilimandjaro*
Déserts	*le Sahara*, *la Puna d'Atacama*
Rues	*le Faubourg Saint-Antoine*, *le bd Ménilmontant*
Édifices	*la gare Montparnasse*, *l'Hôtel de Ville*
Ergonymes	
Sites de production	*Renault Wilword*
Entreprises industrielles	*Sud-Marine industrie*, *la Générale des eaux*
Coopératives	*Semences de Provence*, *Chlorophylle*
Marques et produits	*Volkswagen*, *Passat*
Établissements d'ens. et de rech.	*l'Université de Nantes*, *le CNAM*
Installations militaires	*la caserne Salianski*, *la Grande muraille de Chine*
Livres, publications et œuvres d'art	*Matrix*, *les Princes d'Ambre*
Praxonymes	
Faits historiques	*la Guerre d'Algérie*, *la guerre de Cent Ans*
Maladies	*la maladie de Parkinson*
Évènements culturels	*le Festival de Cannes*, *la Love Parade*
Phénonymes	
Catastrophes naturelles	*l'ouragan Juan*, *le typhon Linda*
Zones de haute et basse pressions	*la Tramontane*, *l'anticyclone des Açores*
Astres et comètes	*la Terre*, *la comète de Faye*

les noms propres codés correspondent aux acronymes, sigles ainsi que toute combinaison de lettres et de chiffres (p. ex. *UK*, *NATO*, *M16*, *L7*, *Boeing 747*).

Les différences essentielles avec la classification de Jonasson résident dans cette dernière catégorie, ainsi que dans le fait que les entités nommées à base descriptive ou mixte de l'anglais seront considérées comme pures en français et que les entités nommées du français ne portent pas nécessairement une majuscule sur tous les mots qui les composent.

Dans le cadre de notre étude, ces typologies ne nous conviennent pas. En effet, elles intègrent des critères morphologiques, mais également lexicaux et syntaxiques. Or, nous avons fait le choix de ne pas avoir recours à des prétraitements linguistiques. Par conséquent, ne voulons prendre en compte que les critères graphiques.

Dans cette optique, Daille et Morin [2000] introduisent une terminologie des entités nommées inspirée de celle de Jonasson [1994], non plus fondée sur des critères linguistiques mais uniquement graphiques. Cette terminologie distingue trois types d'entités nommées : les simples, les composées et les mixtes (cf. section 2.1.2).

Nous avons finalement décidé de retenir cette dernière classification comme base pour notre étude graphique des entités nommées.

3.2 Étude en corpus

Cette partie consiste en une étude sur la présence des entités nommées dans des textes. Le but de cette étude est principalement de concevoir une typologie des entités nommées évolutive et la plus fine possible, en prenant comme base celle de Grass [1999] et en l'étendant de façon à obtenir une typologie la plus exhaustive possible. Parallèlement, nous souhaitons étudier la répartition des entités nommées en corpus, en fonction de leurs caractéristiques graphiques et référentielles.

Le corpus constitué pour cette tâche se compose d'échantillons de deux corpus différents : *La Recherche*[2] (17 067 mots) et *Le Monde*[3] (20 866 mots). Pour *La Recherche*, ces échantillons consistent en trois textes :

1. un article de neurobiologie ;
2. un article d'astronautique ;
3. un article traitant de la quête de la langue originelle.

En revanche, les échantillons du *Monde* regroupent un bien plus grand nombre d'articles, de taille plus petite et au contenu très varié. Il est important de le préciser, car en ce qui concerne *La Recherche*, le niveau de représentation des différentes catégories d'entités nommées sera conditionné par le type de l'article. En effet, celui qui traite de la quête de la langue originelle comportera beaucoup de toponymes, mais très peu d'anthroponymes, alors que l'article sur l'astronautique comportera moins d'anthroponymes et plus d'ergonymes (produits manufacturés) ou de phénonymes (astres). Nous avons choisi ces deux échantillons afin d'avoir des textes de presse quotidienne (*Le Monde*) qui sont plus généralistes et des textes de presse scientifique (*La Recherche*), plus spécialisés. Cependant, ce choix comporte des limites (pas de textes littéraires par exemple),

[2]Corpus *La Recherche* - année 1998 - distribué par ELRA (`http://www.icp.inpg.fr/ELRA`).
[3]Corpus de textes *Le Monde* - année 1997 - European Corpus Initiative (ECI) distribué par ELRA.

mais il nous paraît difficile de traiter tous les types de corpus avec une masse suffisamment représentative. Nous nous sommes donc concentrés sur des articles de presse écrite, car ce sont ceux que l'on peut se procurer le plus facilement et sur lesquels sont effectués le plus grand nombre de traitements automatiques.

Cette étude est constituée de deux parties : la première porte sur la catégorisation référentielle et va nous permettre de fixer notre typologie des entités nommées, tandis que la seconde se base sur la graphie des entités nommées.

3.2.1 Étude référentielle

Ici, l'objectif est d'établir une catégorisation référentielle pour les entités nommées du français, la plus fine et la plus exhaustive possible.

Pour juger du caractère nommé d'une entité rencontrée, nous nous appuyons essentiellement sur l'intuition de l'opposition entre entité nommée et nom commun, que nous avons détaillée à la section 1.4 et qui se fonde essentiellement sur le critère graphique de la majuscule à l'initiale et le critère sémantique de la référence unique.

À mesure que nous essayons de classer les entités nommées identifiées dans nos échantillons de corpus, selon les catégories de la typologie de Grass [1999], nous nous apercevons qu'il s'en trouvent certaines qui n'ont leur place dans aucune de ces catégories. Il va donc nous falloir en étendre certaines et en créer d'autres.

Il est important de noter que l'extension ou la création de catégories se fait uniquement sur des critères sémantiques/référentiels ; en aucun cas nous n'avons tenu compte de la traduction ou de l'application dans laquelle pourrait intervenir notre système de reconnaissance. En effet, il faudrait réaliser une étude pour savoir comment distinguer les entités nommées d'une catégorie, selon des critères de traduction. De plus, cette étude ne donnerait pas les mêmes résultats selon les langues mises en jeux. Cependant, nous espérons que le fait de nous baser sur des tels critères et de chercher à obtenir une catégorisation fin et évolutive favorisera l'adaptation de notre système aux différents domaines du TALN.

Les anthroponymes

Cette classe regroupe donc les entités nommées qui caractérisent une ou plusieurs personnes.

Chez Grass [1999], elle comporte une catégorie constituée par les hypocoristiques définis comme étant des « termes qui expriment une intention caressante, affectueuse » (p. ex. *mon amour*, *mon/ma chéri(e)*). S'il peut être intéressant, dans une optique de traduction, d'identifier ces derniers car leur traduction ne se fait pas littéralement (p. ex. *Bärchen* ou *Tiger* ne se traduisent pas en *petit ours* ou *tigre*), il ne s'agit pas vraiment d'entités nommées. Nous avons donc décidé de retirer cette catégorie de notre typologie.

Parmi les anthroponymes définis par Grass [1999], il existe déjà une catégorie contenant les noms de partis et autres organisations. Nous l'étendons aux noms d'institutions publiques (p. ex. *le Parlement*, *la Banque de France*).

La catégorie des ethnonymes contient les noms d'habitants de pays. Nous y ajoutons naturellement les noms d'habitants de villes (gentilés) ou d'autres toponymes[4] (p. ex. *les Franciliens*, *les Nantais*). Par extension, nous y adjoignons les noms regroupant les groupes de personnes appartenant à une même période historique (p. ex. *les Néendertaliens*), un même club de sports

[4]Nous ne retenons que entités nommées spécifiques et non les syntagmes de la forme *les habitant de...*

(p. ex. *les Ajaïstes*), un même mouvement artistique (p. ex. *les Surréalistes*), politique (p. ex. *les Socialistes*), religieux (p. ex. *les Musulmans*), etc. En définitive, cette catégorie regroupe tous les noms « collectifs ».

La catégorie des ensembles artistiques, groupes musicaux et troupes de théâtres est enrichie des noms de médias d'une part (télévision : *Canal+*, radio : *France Info*, presse écrite : *France Football*, etc.) et des équipes sportives d'autres part (p. ex. *le Paris-SG*, *l'équipe de France de football*). Il faut noter que, pour les noms de médias, il ne pouvait être fait d'assimilation avec les noms d'organisations. En effet, si certains noms de médias correspondent directement à un nom d'organisation, d'autres comme le nom d'un journal, par exemple, peuvent être différents des sociétés qui les éditent (*l'Equipe Magazine*). Il faut donc séparer ces deux concepts, d'où la nécessité d'ajouter les noms de médias aux ensembles artistiques et de ne pas les réduire aux organisations.

Les toponymes

Pour les toponymes, la tâche est un peu plus ardue, notamment pour les zones géographiques. En effet, Grass [1999] limite la catégorisation aux villes, pays et microtoponymes. Or, il existe nombre d'entités nommées de zones géographiques qui n'appartiennent à aucune de ces catégories (p. ex. *l'Afrique*, *la Bretagne*, *le Monde*). Il est donc nécessaire d'y ajouter des catégories pour couvrir tous les types de taille de zones géographiques, tout en prenant gare à ce qu'il soit ensuite possible de trouver des moyens à mettre en place pour les différencier. Nous avons donc décidé d'ajouter les zones plus vastes que les pays (p. ex. *l'Europe*, *Le Royaume Uni*, *l'Amérique Latine*, *le Proche-Orient*), ainsi que celles dont la taille est comprise entre celle d'une ville et d'un pays telles que les régions, les départements, les états américains, les provinces des différents pays (p. ex. *la Californie*, *l'Île de France*, *la Corse*, *la Côte d'Azur*). Malgré tout, il reste certaines zones géographiques comme les noms de péninsules, de plages, de côtes, d'îles, archipels, etc. qui n'appartiennent à aucune catégorie, car leur taille peut varier : p. ex. une île peut avoir une dimension comprise entre une ville et un pays comme *l'Île Maurice* ou *Belle Île en Mer*, elle peut être plus petite qu'une ville comme *l'Île Versailles* ou *l'Île de la Cité*, il peut s'agir d'un pays comme *l'Île de Madagascar*, elle peut être partagée entre plusieurs pays comme *l'Île de Bornéo*, contenir plusieurs pays comme *la Grande Bretagne*, etc. Pour ces différentes zones géographiques, nous avons décidé de ne pas attribuer de catégories, mais uniquement la classe toponyme, car le nombre de nouvelles catégories à créer auraient été trop important.

Quid des noms de forêts ?

Les ergonymes

La catégorie qui va être la plus étendue est celle qui regroupe les œuvres. À la base, Grass [1999] incluait les titres de livres, les noms de publications et d'œuvres d'art. Dans un premier temps, comme Grass [1999] ne le fait pas explicitement (il distingue même les titres de livres de œuvres d'art), nous avons inclus les œuvres de toutes les formes d'art : architecture, sculpture, peinture, dessin, danse, chant, poésie, musique, théâtre, littérature, cinéma, photographie, télévision, etc. Dans un second temps, le problème s'est posé de catégoriser un grand nombre d'entités nommées qui avaient, pour seul point commun, d'être des productions intellectuelles, généralement créées par une seule personne dont le patronyme a donné nom à cette production (projets, plans, théorèmes, lois, prix, etc.). Nous avons donc décidé de regrouper ces entités nommées ainsi que les noms de publications et les œuvres d'art sous la catégorie œuvres intellectuelles.

Les praxonymes

Une conséquence directe de la création de la catégorie œuvres intellectuelles est l'assimilation des noms de maladies dans celle-là et donc leur disparition de la classe des praxonymes. Nous verrons dans la section 3.5 que ce choix n'est pas judicieux.

La dernière extension nécessaire concerne la catégorie évènements culturels. En effet, un meeting politique (*le congrès de Rennes*), une compétition sportive (*la Coupe du Monde de football*), un salon (*le salon de l'Automobile*), une foire (*la foire de Paris*), etc. ne peuvent pas être considérés comme des évènements culturels. Ces types d'entités nommées n'ont donc pas de place dans la typologie de Grass [1999]. Or, il définit les praxonymes comme étant « tous les noms utilisés pour désigner des faits et évènements dont les déclencheurs, les responsables, les participants et les patients sont des être humains » [5]. Ces type d'entités nommées peuvent donc être assimilées à cette catégorie, car ils constituent, eux aussi, des évènements. Nous les inclurons donc dans cette catégorie, qui devient alors simplement la catégorie évènements.

Au sein de la classe des praxonymes, nous avons ajouté une nouvelle catégorie regroupant les différentes périodes historiques comme *le Paléolitique*, *la Renaissance* ou encore *le siècle des Lumières*. Cette catégorie est absente chez Grass [1999] et trouve, par définition, sa place parmi les praxonymes. Or, il nous semble que ces entités nommées ne constituent ni vraiment des faits, ni vraiment des évènements, qui sont plus ponctuels. Nous avons donc décidé de créer cette catégorie.

Les phénonymes

Les phénonymes regroupent, pour Grass [1999], les catastrophes naturelles, les zones de haute et basse pressions, les astres et les comètes. Nous avons décidé de séparer cette classe en deux catégories : les entités astrales (astres, galaxies, etc.) d'une part et le reste de l'autre dans la catégorie phénomènes naturels (cyclones, typhons, courants aériens et marins, etc.).

Les zoonymes

Par rapport à la typologie de Bauer [1985], Grass [1999] ajoute une classe pour les noms d'animaux de compagnie (les zoonymes), parmi lesquels certains sont presque lexicalisés (p. ex. *Félix*, *Médor*). Plutôt que de créer une classe ne contenant qu'une seule catégorie, il assimile celle-là aux anthroponymes. Bien que cette assimilation ne soit pas étymologiquement réalisable, Grass [1999] argue qu'elle peut être réalisée car il pourrait s'agir d'une « humanisation de l'animal avec pour corollaire un changement de classe sémantique ».

Cette catégorie ne se trouve pas représentée dans les textes sur lesquels nous avons fait notre étude, mais nous l'inclurons tout de même dans notre typologie.

3.2.2 Étude graphique

La distinction des entités nommées suivant des critères uniquement graphiques est très importante en français. En effet, une langue comme l'anglais est uniquement confrontée à la difficulté d'identifier les entités nommées pures et composées, alors que le français est confronté à la fois

[5] Traduction de la définition de Bauer [1985] : « alle Namen, die zur Bezeichung von Ereignissen und Geschehnissen benutzt werden, als deren Auslöser, Träger, Teilnehmer und Betroffene Menschen gelten können ».

aux entités nommées pures et composées, mais aussi aux entités nommées mixtes. De plus, suivant la graphie des entités nommées, leur identification et leur classification seront traitées de manières différentes. C'est pour cela qu'il nous faut distinguer un maximum de catégories graphiques : nous créerons une nouvelle catégorie graphique dès qu'il nous sera possible de prévoir qu'elle posera des problèmes différents de celles déjà prises en compte.

Nous reprenons donc les critères uniquement morphologiques de la classification de Daille et Morin [2000], tout en en modifiant la terminologie, en la complétant et en l'affinant. Nous reprenons la notion de « pureté » de Jonasson [1994] et Allerton [1987], en l'accordant, non plus à la présence de noms communs ou non dans l'entité nommée, mais à la présence de majuscule en première lettre des mots la composant. Nous introduisons en outre un critère de « complexité » pour différencier les entités nommées pures constituées d'une ou de plusieurs formes. Enfin, nous séparons les entités nommées mixtes en deux catégories et en ajoutons une nouvelle (la catégorie `sigles`), afin d'établir un maximum de différenciations.

En résumé :

les `entités nommées pures simples` sont constituées d'une seule forme commençant par une majuscule comme *France* ou *Aristote* ;

les `entités nommées pures composées` sont constituées de plusieurs formes commençant par une majuscule comme *Conflans Saint-Honorine*. Parmi celles-ci, nous distinguons la sous-catégorie `Prénom Nom`[6] : entités nommées constituées d'un ou plusieurs prénoms et d'une forme commençant par une majuscule référant à un nom de personne comme *Paul Valéry* ;

les `entités nommées faiblement mixtes` sont constituées de plusieurs mots commençant par une majuscule et contenant des entités fonctionnelles en minuscule comme *le Jardin des Plantes*. Cette liste d'entités fonctionnelles est fermée et comprend les prépositions, les articles, etc. ;

les `entités nommées mixtes` sont constituées de plusieurs formes dont au moins une commence par une majuscule comme *Comité international de la Croix-Rouge*, *Mouvement contre le racisme et pour l'amitié entre les peuples* ;

les `sigles` sont constituées d'une seule forme comportant plusieurs majuscules – éventuellement suivies d'un point – qui réfèrent elles-mêmes à une autre forme comme *USA* ou *U.R.S.S.*. Les entités nommées appartenant à cette catégorie, qu'il est important de distinguer au niveau graphique, réfèrent à des entités nommées pures composées et des entités nommées mixtes (faiblement ou non).

La différence entre un sigle et un acronyme tient du fait qu'un sigle est une suite de lettres initiales constituant l'abréviation de plusieurs termes formant un terme unique prononcé avec les noms des lettres, alors qu'un acronyme est un groupe d'initiales abréviatives plus ou moins lexicalisé qui se prononce comme s'il s'agissait d'un nouveau mot. Cette distinction n'étant pas nécessaire dans le cadres de nos travaux, nous emploieront dorénavant, indifféremment *sigle* et *acronyme*.

Tableau 3.2 – Distribution des entités nommées en fonction de leur longueur

	Longueur						**Longueur**
	1	2	3	4	5	6+	**moyenne**
Anthroponymes							
Patronymes	126	162	38	5	1		1,8
Prénoms	58	2					1
Ethnonymes	46			1			1
Organisations	98	18	13	19	11	9	2,3
Ensembles artistiques	44	26	12	1	1		1,7
Divers Anthroponymes	2	1					1,3
Toponymes							
Toponymes > Pays	30	9	3	2			1,5
Pays	154	1	5				1
Ville < Toponymes < Pays	34	3	9	7	2	1	2
Villes	96	3					1
Microtoponymes	1	1					1,5
Hydronymes	16	1	1				1,2
Oronymes	27	19	10	5	2	1	2,1
Rues		1	1				2,5
Désert	1		1				2
Édifices	6	8	8	4			2,4
Divers toponymes	6	1	3				1,7
Ergonymes							
Marques et produits	93	25		2			1,25
Entreprises industrielles	3	1					1,25
Établissements d'ens. et de rech.	1		3		2	2	4,1
Œuvres	2		2	2		1	3,3
Praxonymes							
Faits historiques		1	5				2,8
Évènements	5		8	1	7		3,2
Périodes historiques	1						1
Total	**850**	**283**	**122**	**49**	**26**	**14**	**1,55**

3.3 Analyse quantitative des résultats

En préambule à l'analyse quantitative des résultats des études graphiques et référentielles, nous avons effectué une autre étude en corpus portant sur la répartition des entités nommées à travers leur catégorie référentielle et en fonction de leur longueur (cf. tableau 3.2).

Il faut avant tout remarquer qu'il est difficile de calculer la proportion d'entités nommées dans des textes non balisés car, s'il nous est possible de connaître le total des mots et des entités nommées présents dans ces textes, il ne nous paraît pas envisageable de calculer la proportion exacte des mots qui font partie d'entités nommées, car il nous faudrait recompter, pour chaque entité nommée, le nombre de mots qui la composent.

Pour obtenir une évaluation satisfaisante, nous avons donc observé les entités nommées d'un corpus regroupant des textes du *Monde* (13 459 mots), une page *Web* de la *FAO* (6 993 mots), une page *Web* du *Monde diplomatique* ayant pour titre « Les armes biologiques de la guerre de Corée » (2 598 mots) et une page *Web* traitant de la contrefaçon. Le but de cette observation est d'établir une valeur moyenne du nombre de formes que contiennent les entités nommées du français. Pour ce faire, nous avons étiqueté manuellement ces textes, puis nous avons compté, à l'aide d'un programme, le nombre d'entités nommées de chaque catégorie et leur longueur.

Ainsi, nous pouvons en conclure que notre corpus contient 1 344 entités nommées au total, qui ont une longueur moyenne de 1,55 formes. Par conséquent, à défaut d'une statistique précise de la proportion d'entités nommées présentes dans nos corpus (*La Recherche* et *Le Monde*), nous pouvons utiliser cet indice pour en donner une bonne approximation (cf. section 3.3.2).

Les résultats quantitatifs, exposés aux tableaux 3.3 et 3.4, ont été obtenus grâce à une étude et un comptage manuel. Dans un premier temps, toutes les entités nommées ont été identifiées et sur-lignées. Puis, selon le critère retenu (graphique ou référentiel), nous les avons comptées.

3.3.1 Résultats de l'étude référentielle

Nous avons donc créé, à l'aide d'une étude en corpus, une catégorisation des entités nommées basée sur des critères référentiels (cf. section 3.2.1), qui doit désormais être validée par une étude numérique de la distribution des entités nommées, en fonction de leur appartenance aux différentes catégories de cette typologie : les nouvelles entités nommées rencontrées dans les textes devront y trouver place. Cette étude est présentée au tableau 3.3[7].

Pour cela, nous évaluons, pour chaque catégorie de notre typologie, le nombre d'entités nommées qui prennent place dans celle-là. Les entités nommées qui ne rentrent pas dans ces catégories sont placées dans les catégories « fourre-tout ». Ces catégories sont au nombre de six :

- une par classe, réunissant les entités nommées dont nous avons identifié l'appartenance à celles-ci (grâce au contexte de leurs apparitions ou à nos connaissances à priori), mais dont la catégorie est incertaine voire inexistante ;
- une dernière, dans laquelle se trouvent toutes les entités nommées sur lesquelles nous n'avons aucune information dans le texte, mais qui paraissent être des entités nommées.

Les cinq premières portent le nom *divers* suivi du nom de la classe à laquelle elles font référence et la dernière s'appelle « autres ».

[6]Nous introduisons cette sous-catégorie en raison de la facilité à identifier et classifier les prénoms par des dictionnaires.

[7]Les catégories étendues ou ajoutées par rapport à la typologie de Grass [1999] apparaissent précédées d'un astérisque.

Tableau 3.3 – Distribution des entités nommées en fonction de leurs caractéristiques référentielles

	La Recherche		*Le Monde*	
	# Occ.	**Proportion**	**# Occ.**	**Proportion**
Anthroponymes	**194**	**52,0 %**	**1066**	**73,8 %**
Patronymes	97		437	
Prénoms	66		310	
Ethnonymes	15		37	
∗ Organisations	16		194	
∗ Ensembles artistiques	0		87	
Pseudonymes	0		1	
∗ Zoonymes	0		0	
∗ Divers anthroponymes	0		0	
Toponymes	**107**	**28,7 %**	**271**	**18,7 %**
∗ Toponymes > Pays	53		17	
Pays	22		73	
∗ Villes < Toponymes < Pays	17		33	
Villes	10		108	
Microtoponymes	0		16	
Hydronymes	4		9	
Oronymes	0		0	
Rues	0		4	
Déserts	1		0	
Édifices	0		15	
∗ Divers toponymes	0		0	
Ergonymes	**64**	**17,2 %**	**92**	**6,4 %**
Marques et produits	31		37	
Entreprises industrielles	0		4	
Sites de production	0		0	
Coopératives	0		0	
Établissements d'ens. et de rech.	27		7	
Installations militaires	0		0	
∗ Œuvres	6		44	
∗ Divers ergonymes	0		0	
Praxonymes	**3**	**0,8 %**	**16**	**1,1 %**
Faits historiques	0		0	
∗ Évènements	0		15	
∗ Périodes historiques	3		1	
∗ Divers praxonymes	0		0	
Phénonymes	**5**	**1,3 %**	**0**	**0 %**
Phénomènes naturels	0		0	
Entités astrales	5		0	
∗ Divers phénonymes	0		0	
Autres	**0**		**0**	
Total	**373**		**1 445**	

Pour le comptage, il faut tout d'abord noter qu'une entité nommée constituée d'un ethnonyme, d'un prénom et d'un nom incrémentera chacune de ces trois catégories : p. ex. en rencontrant l'entité nommée *Michel Platini*, nous ajoutons une unité aux prénoms (*Michel*), ainsi qu'aux patronymes (*Platini*).

De plus, certaines catégories référentielles ne sont que peu ou pas représentées dans ces deux échantillons de corpus (zoonymes, oronymes, sites de production, coopératives, maladies, faits historiques, phénomènes naturels, etc.). Pour autant, elles ne sont pas à négliger, car il est facile de se rendre compte qu'elles peuvent être présentes dans d'autres textes et qu'elles ne pourraient être catégorisées à l'aide de catégories autres que celles que nous avons prévues.

Après une étude *a priori* et une expérimentation en corpus, notre catégorisation a atteint une très haute stabilité : quasiment toutes les nouvelles entités nommées identifiées trouvent une place dans notre typologie.

Les classes les plus volumineuses sont les anthroponymes et les toponymes : une proportion des entités nommées respectivement égale à 52 % et 29 % pour *La Recherche* et 74 % et 12 % pour *Le Monde*, soit plus de 80 % des entités nommées de ces deux échantillons de corpus. À l'intérieur de ces classes, ce sont les catégories patronymes (50 % et 41 %), prénoms (34 % et 29 %) et organisations (8 % et 18 %) qui regroupent près de 90 % des anthroponymes. Quant à la classe des toponymes elle est composée environ aux trois quarts de toponymes > pays (50 % et 4 %), de pays (21 % et 27 %) et de villes (10 % et 40 %). À elles seules, ces six catégories représentent pas loin de 77 % de toutes les entités nommées que nous avons identifiées manuellement.

3.3.2 Résultats de l'étude graphique

Grâce à l'indice de la longueur moyenne d'une entité nommée établi lors d'une étude en corpus (cf. section 3.3), nous avons pu établir la proportion des formes de nos corpus qui font partie d'une entité nommée[8]. Nous constatons qu'il y a nettement plus d'entités nommées dans l'échantillon du corpus *Le Monde* que dans celui de *La Recherche* (resp. 7,2 % et 2,2 %), cela, toutes catégories graphiques confondues (cf. tableau 3.4).

Tableau 3.4 – Distribution des entités nommées en fonction de leurs caractéristiques graphiques

	La Recherche	***Le Monde***
`EN pures simples`	145	313
`EN pures composées`	25	89
`Prénom Nom`	68	299
`EN faiblement mixtes`	21	35
`EN mixtes`	44	144
`sigles`	15	127
Total	318	1 007
Proportion	2,9 %	7,5 %

Les `EN pures simples` sont les plus présentes dans les deux corpus (46 % des entités nommées pour *La Recherche* et 31 % *Le Monde*). Elles sont faciles à identifier du fait qu'elles sont

[8]Cette proportion correspond au rapport entre nombre d'entités nommées multiplié par la longueur moyenne d'une entité nommée, et le nombre total de formes du corpus.

constituées d'une seule forme. Cependant, elles peuvent présenter une difficulté lors de la catégorisation. En effet, si l'une des entités nommées n'est pas recensée dans un des lexiques, il faut faire appel au contexte pour la catégoriser.

Les `EN pures composées` sont moins présentes que les simples (7,8 % et 8,8 %). Elles présentent les mêmes caractéristiques, mais sont toutefois plus facile à catégoriser, car il suffit que l'une des formes composant l'entité nommée soit présente dans un lexique pour pouvoir y parvenir. Pour la catégorie `Prénom Nom`, très présente dans nos échantillons (21,4 % et 29,7 %), la catégorisation devient plus facile, dans la mesure où il est aisé de trouver un lexique de prénom relativement complet. De plus, les entités nommées de cette catégorie sont présentes dans des schémas lexicaux repérables à l'aide de mots déclencheurs.

Les `EN faiblement mixtes` sont un peu moins présentes que les `EN pures composées` (6,6 % et 3,5 %), mais présentent les mêmes problèmes. La difficulté de catégorisation sera peut-être un peu plus élevée du fait des entités fonctionnelles en minuscule.

Les `EN mixtes` sont loin d'être à négliger du fait de leur nombre (13,8 % et 14,3 %). En revanche, elles sont très compliquées à délimiter du fait de la difficulté à identifier leur contexte droit. Cependant, le problème de leur délimitation réglée, leur catégorisation peut s'avérer assez simple, notamment grâce aux mots communs les composant qui peuvent dénoter leur catégorie.

Les `sigles` posent le problème inverse à celui posé par les `EN mixtes` : l'identification est quasiment immédiate (elles sont constituées uniquement de majuscules et le plus souvent composées d'un seul mot), alors que la catégorisation est presque impossible si l'on ne connaît pas les mots auxquels chaque lettre du sigle fait référence. La présence de ces sigles est moins importante dans l'échantillon de *La Recherche* que dans celui du journal *Le Monde* (4,7 % et 12,6 %).

3.4 Synthèse

Hormis ces résultats quantitatifs, des remarques qualitatives peuvent être émises sur les liens entre les catégories référentielles et graphiques.

Les patronymes et les prénoms composent la catégorie `Prénom Nom`. Les ethnonymes, l'ensemble des toponymes, les maladies, les périodes historiques, les phénomènes naturels et les entités astrales sont essentiellement des `EN pures simples` (*les Français*, *le Parisien*, *la France*, *les Alpes*, *la Renaissance*, *le Paléolitique*, *le cyclone Hugo*, etc.). Cependant, les toponymes, par exemple, peuvent être des `EN pures composées` ou des `EN faiblement mixtes` (*l'Europe de l'Ouest*, *l'Océan Indien*, etc.), voir même des `sigles` (*la RFA*, *l'URSS*, *USA*, etc.). Les organisations sont composées de `sigles`, d'`EN Pures Composées pures`, d'`EN faiblement mixtes` et d'`EN mixtes` (*la CEE*, *la Communauté Économique Européenne*, *l'Association of Ceramic Industry*, *le Centre national des lettres*, etc.). Ces trois dernières catégories regroupent également les ensembles artistiques, les sites de production, les entreprises industrielles, les coopératives, les établissements d'enseignement et de recherche, les installations militaires, les œuvres, les faits historiques et les évènements.

Nous avons donc mis au point une typologie qui semble constituer un bon moyen pour permettre la reconnaissance exhaustive et automatique des entités nommées pour le français, dans une catégorisation fine. Elle s'organise autour de cinq classes principales.

La première classe regroupe les **ANTHROPONYMES**, qui caractérisent une ou plusieurs personnes, parmi lesquels on trouve les noms individuels (**patronymes**, **prénoms**, **pseudonymes**, **zoonymes**), ainsi que les noms collectifs comme les **ethnonymes** (étendus aux gentilés et

aux groupements de personnes d'une même période historique, d'un même mouvement idéologique, etc.), mais aussi les noms de partis, d'institutions et autres **organisations**. Les noms d'**ensembles artistiques**, de médias et d'équipes sportives font également partie de cette classe.

Les TOPONYMES, noms de lieux au sens général, comprennent donc les noms de **villes** et de **pays**, mais aussi de zones géographiques de toutes les tailles : **toponymes** > **pays**, **villes** < **toponymes** < **pays** et **microtoponymes**. Les **hydronymes** (noms de lieux en rapport avec l'eau), les **oronymes** (noms de lieux en rapport avec la montagne), les noms de **rues**, de **déserts** et d'**édifices** sont aussi présents parmi cette classe.

Les ERGONYMES (du grec *ergon* : travail, force) regroupent les objets et produits manufacturés. On y trouve donc les noms de **marques et produits**, de **sites de production** et d'**entreprises industrielles**. Les **coopératives**, les **établissements d'enseignement et de recherche**, ainsi que les **installations militaires** y sont également inclus. À cette classe, appartiennent aussi les titres de livres, les noms de publications et d'œuvres d'art (**œuvres**).

Les PRAXONYMES (du grec *praxis* : action) sont les faits et les évènements dont les acteurs sont des êtres humains. Il s'agit de syntagmes représentant des **faits historiques**, des **évènement culturels**, sportifs, politiques, etc. On y adjoint également les **périodes historiques**.

La dernière classe est constituée par les PHÉNONYMES (de grec *phainómenon* : ce qui apparaît) qui comprennent les **phénomènes naturels**, ainsi que les **entités astrales**.

3.5 Évolution de la typologie des entités nommées et modularité

Tout au long de la construction de notre système de reconnaissance, nous avons été amenés à opérer quelques évolutions de notre catégorisation des entités nommées. Il est intéressant de voir que Grass [2000] apporte également quelques modifications à sa typologie. Nous avons donc décidé de comparer les modifications de Grass [2000] avec celles que nous avons effectuées lors de notre étude en corpus et tout au long de l'élaboration de notre système.

Parmi les patronymes, nous avons simplement inclus explicitement, les noms de dieux (p. ex. *Zeus*, *Ra*), de héros de la mythologie (p. ex. *Pandore*, *Anhur*) et d'autres personnages fictifs (p. ex. *Tintin*, *Corwin d'Ambre*). Si dans notre typologie, la classe des anthroponymes est restée stable, Grass [2000] a apporté deux modifications importantes à la sienne. La première consiste dans l'ajout des clubs sportifs à la catégorie contenant déjà les partis politiques et les autres organisations. Le choix de rapprocher les clubs sportifs des organisations ou des ensembles artistiques est discutable[9] : un club de sport n'est pas plus proche d'un parti politique ou d'une organisation que d'un ensemble artistique. Néanmoins, nous avons choisi cette dernière solution de façon arbitraire. La séparation des groupes musicaux « modernes » et des ensembles artistiques et orchestres classiques, en deux catégories, constitue la deuxième modification. Grass [2000] ne donne pas explicitement de justification à cette différenciation, mais pour une traduction de l'allemand vers le français, il préconise de ne pas traduire les premiers, mais de traduire les autres. Il aurait été intéressant, dans une optique de traduction, de distinguer ces deux catégories, mais nous n'avons utilisé que des critères sémantiques/référentiels pour établir notre typologie. De plus, cette distinction aurait probablement été difficile à obtenir automatiquement par notre système.

La seule modification que nous avons apportée aux toponymes concerne les catégories des

[9]Notons que dans une traduction de l'allemand vers le français, Grass [2000] déduit les mêmes règles pour les deux catégories.

édifices (monuments et autres constructions célèbres) et des installations militaires. En effet, si dans un premier temps nous avions simplement fait migré cette dernière catégorie des ergonymes vers les toponymes, nous l'avons finalement intégrée parmi les édifices, car il nous apparaît que la différence entre *le Pentagone* et *la Maison Blanche*, par exemple, n'est pas assez importante pour faire deux catégories, d'autant que les installations militaires sont très peu nombreuses dans les textes français (il n'y en a aucune parmi tous nos corpus). Grass [2000] s'est arrêté à la première évolution et a donc fait changé de classe la catégorie des installations militaires qui appartient maintenant aux toponymes.

Parmi les ergonymes, nous avons rapproché les sites de productions et les entreprises industrielles, ce que semble également faire Grass [2000]. En revanche, si les coopératives semblent toujours appartenir aux ergonymes, Grass [2000] ne précise pas dans quelle catégorie elles se trouvent. Pour notre part, nous avons choisi de les déplacer parmi les organisations pour les mêmes raisons qui nous ont poussés à déplacer les installations militaires parmi les édifices. Grass [2000] ajoute aux ergonymes une catégorie contenant les objets issus des mythologies, des comtes, des légendes, des fictions, etc. (p. ex. *Dard*, *le cheval de Troie*, *la Batmobile*). Nous avons décidé de reprendre cette catégorie pour parfaire notre typologie, car, bien que nous n'ayons pas rencontré de telles entités nommées, elles sont susceptibles d'apparaître, auquel cas il faudra les catégoriser.

La plus grande évolution que nous avons apporté à notre typologie concerne la catégorie des œuvres. En effet, nos y avions inclus les productions intellectuelles, généralement créées par une seule personne dont le patronyme a donné nom à cette production. Or, en revoyant la définition des praxonymes, il nous est apparu qu'il était plus logique de placer ces entités nommées dans cette classe. Cette remise en cause nous a été confirmée par le fait que Grass [2000] introduit ces entités nommées, mais les classes avec les maladies qui appartenaient déjà aux praxonymes. Nous avons donc scindé les œuvres en deux catégories :

- les œuvres « artistiques » (cinématographiques, littéraires, picturales, etc.), qui restent parmi les ergonymes ;
- les œuvres « abstraites » (lois, maladies, théorèmes, projets, plans, etc.), que nous plaçons dans les praxonymes.

Toutes ces modifications mettent en lumière le côté évolutif et modulaire de notre typologie des entités nommées. En effet, il nous est très facile, à partir d'une classification fine, de regrouper certaines catégories, simplement en modifiant la partie droite des règles de réécriture concernées (cf. section 4.2.2.2). Pour scinder une catégorie en plusieurs, ce n'est pas beaucoup plus difficile : il faut modifier la partie droite de certaines règles et éventuellement scinder le lexique des mots déclencheurs[10] (cf. section 4.2.2.1), ainsi que multiplier les règles associées (une par catégorie référentielle).

Nous avons donc constitué une typologie des entités nommées du français comportant 27 catégories réparties en cinq classes (cf. tableau 3.5). Cette typologie va nous servir de base pour la catégorisation automatique des entités nommées.

[10]Nous verrons qu'il s'agit là du seul lexique nécessitant une telle scission, car les autres possèdent une unité sémantique qu'il ne paraît pas possible de briser (une ville est une ville, de même pour un parti politique, un prénom, etc.).

Tableau 3.5 – Notre typologie référentielle des entités nommées

Anthroponymes	
Patronymes	*Hadès*, *Zidane*, *Corwin d'Ambre*
Prénoms	*Alexandre*, *Mohand Areski*
Ethnonymes	*le Serbo-Croate*, *les Franciliens*
Organisations	*la FFF*, *ATTAC*, *le Parti socialiste*, *Chlorophylle*
Ensembles artistiques	*le Paris-SG*, *Muse*, *l'Orchestre National de Barbès*
Pseudonymes	*l'Aigle des Açores*, *le Petit Père des Peuples*
Zoonymes	*Félix*, *Titi*
Toponymes	
Toponymes > Pays	*l'Europe*, *les Balkans*
Pays	*la France*, *l'Algérie*
Pays > Toponymes > Villes	*l'Île de France*, *la Californie*
Villes	*Paris*, *Belo Horizonte*
Microtoponymes	*Prenzlauerberg*, *le Quartier Latin*
Hydronymes	*l'océan Atlantique*, *la Seine*, *les chutes du Niagara*
Oronymes	*les Alpes*, *le Kilimandjaro*
Déserts	*le Sahara*, *la Puna d'Atacama*
Rues	*le Faubourg Saint-Antoine*, *le bd Ménilmontant*
Édifices	*la gare Montparnasse*, *l'Hôtel de Ville*, *la Grande muraille de Chine*
Ergonymes	
Entreprises industrielles	*Sud-Marine industrie*, *Renault Wilword*
Marques et produits	*Volkswagen*, *Passat*
Établissements d'ens. et de rech.	*l'Université de Nantes*, *le CNAM*
Œuvres matérielles	*Matrix*, *Les Princes d'Ambre*, *la Joconde*
Praxonymes	
Faits historiques	*la Guerre d'Algérie*, *le Traité de Versailles*
Évènements	*la Fête de la musique*, *la Coupe du Monde*
Périodes historiques	*le Paléolithique*, *le Moyen-Âge*, *la 5ème République*
Œuvres abstraites	*la loi de Moore*, *la maladie de Creuzfeld-Jacob*, *le Projet Albion*
Phénonymes	
Phénomènes naturels	*l'ouragan Juan*, *le Gulf Stream*, *la Tramontane*
Astres et comètes	*la Terre*, *la comète de Faye*

Chapitre 4

Nemesis : un système de reconnaissance des entités nommées du Français

Dans ce chapitre, nous présentons **Nemesis**, notre système d'identification et de catégorisation automatiques des entités nommées dans les textes français. Nous exposons tout d'abord la méthodologie de la reconnaissance des entités nommées, puis l'architecture logicielle de **Nemesis**.

4.1 Méthodologie de la reconnaissance

À la suite de notre étude sur les différentes typologies de noms propres et d'entités nommées (cf. sections 1.2.1 et 1.4.1), ainsi que de notre étude en corpus (cf. section 3.2), nous pouvons augurer que notre typologie couvrira un maximum d'entités nommées, tout en en laissant un minimum dans les catégories « fourre-tout », pour les raisons suivantes :

- sur le nombre relativement conséquent d'entités nommées repérées, toutes ont trouvé place dans l'une des 27 catégories de notre typologie ;
- toutes les catégories de noms propres et d'entités nommées des différentes classifications rencontrées possèdent une équivalence dans notre typologie, par l'intermédiaire d'une ou plusieurs catégories.

Malgré tout, nous ne perdrons pas de vu, durant l'élaboration de notre système de reconnaissance, que notre typologie peut évoluer et être adaptée selon l'application qui pourra en être faite, comme en témoigne la section 3.5.

Pour parvenir à l'identification et la catégorisation des entités nommées, nombre de système de reconnaissance des entités nommées s'appuient sur un étiquetage morpho-syntaxique [Paik et coll., 1996 ; Trouilleux, 1997] ou procèdent à une analyse syntaxique [McDonald, 1996 ; Wakao et coll., 1996 ; Wolinski et coll., 1995]. Contrairement à ces systèmes, **Nemesis** n'utilise que des lexiques et des règles de réécritures (lexico-sémantiques, graphiques et morphologiques) à partir de textes bruts, sans faire appel au moindre étiquetage linguistique.

En effet, à l'instar de Wakao et coll. [1996], notre but consiste à mettre en place différents modules (analyse de la structure interne, du contexte immédiat, prétraitement sur les sigles, apprentissage, analyse du contexte large, etc.) en les ajoutant les uns à la suite des autres et en évaluant le coût et l'apport de chacun. Afin de minimiser les prétraitements et donc les sources de bruit, nous avons décidé d'étudier quels résultats nous pouvions obtenir sans requérir à un étiquetage linguistique.

En outre, nous avons décidé d'effectuer la catégorisation des entités nommées en même temps que leur identification, car il nous paraît que ces deux tâches sont très fortement liées. En effet,

l'identification des limites de l'entité nommée va aider à sa catégorisation. D'autre part, la découverte d'un mot-clé permettant la catégorisation d'une entité nommée permet également d'en déterminer la limite à gauche.

Nous nous attachons maintenant à déterminer les éléments des entités nommées que nous allons baliser et plus précisément à résoudre le cas des mots déclencheurs.

4.1.1 Quelle partie de l'entité nommée retenir ?

La première question à nous poser pour effectuer la reconnaissance automatique des entités nommées, concerne leur délimitation et plus précisément les différents éléments à retenir pour le balisage : p. ex. dans *le député européen Daniel Cohn-Bendit* (resp. *le mont Everest*) faut-il retenir *député européen* (resp. *mont*) qui fait partie du syntagme nominal, mais pas réellement de l'entité nommée ?

Ce type de problème intervient avec les mots déclencheurs, pour lesquels nous devons déterminer l'appartenance à l'entité nommée dont ils permettent la reconnaissance et avec laquelle ils forment un même syntagme nominal. Bien que le choix d'écarter ou de conserver certains éléments, comme les titres patronymiques (civils, militaires, religieux, etc.), puisse être fait arbitrairement sans que cela ne pose problème, il n'en est pas moins d'autres circonstances dans lesquelles ce choix est délicat. Dans ce choix, plusieurs paramètres peuvent entrer en considération : le nombre des mots déclencheurs (singulier ou pluriel), la casse de leur initiale ou la présence d'une préposition entre les mots déclencheurs et l'entité nommée.

En étudiant la catégorie des oronymes, nous avons tenté d'établir des généralités, en fonction de la casse de l'initiale des mots déclencheurs :

- si les mots déclencheurs possèdent une majuscule à l'initiale, ils font partie de l'entité nommée (p. ex. *le Pic du Midi*, *le Massif Central*) ;
- sinon, on peut les écarter tout en conservant une cohérence à l'entité nommée (p. ex. *le mont Everest*, *les montagnes de l'Himalaya*, *la chaîne des Alpes*).

Cependant, nous avons trouvé de nombreux contre-exemples à cette dernière assertion : *la vallée du Yagnob* (le *Yagnob* étant une rivière, le mot-clé *vallée* ne peut être écarté), *la montagne Pelée*, *le dôme du Gouter*, etc. En combinant les différents paramètres énoncés précédemment, nous ne sommes pas parvenu à établir de comportement commun à tous les éléments d'une même catégorie et encore moins à toutes les entités nommées, notamment à cause des variations graphiques d'une même entité nommée (*montagne Pelée*, *Montagne Pelée*).

Finalement, il nous paraît très compliqué de trouver un ou plusieurs critères qui permettent de décider automatiquement de l'inclusion d'un mot-clé dans une entité nommée. Malgré tout, il ne faut pas perdre de vue le but d'un système de reconnaissance des entités nommées, qui est principalement l'aide à la recherche d'information. Il vaut donc mieux obtenir un surplus d'information plutôt que d'en perdre, car s'il est envisageable de filtrer ce surplus, il est en revanche beaucoup plus compliquer de retrouver l'information manquante *a posteriori*. Or, si l'on prend les oronymes *la montagne Pelée* et *le massif du Mont-Blanc*, il n'est pas très gênant d'extraire *le massif du Mont-Blanc* plutôt que *Mont-Blanc*, alors qu'il serait très dommageable de n'extraire que *Pelée* au lieu de *montagne Pelée*.

Malgré tout, nous avons mis en place une solution qui consiste à scinder certains lexiques de mots déclencheurs en deux groupes :

1. Les éléments qui peuvent faire partie de l'entité nommée dont ils permettent la reconnaissance.
2. Les éléments qui ne font jamais partie de l'entité nommée dont ils permettent la reconnaissance.

Cette scission induit la multiplication d'un certains nombres de lexiques et de règles de réécritures. Cependant, nous ne tiendrons pas compte de cette multiplication dans le calcul du nombre d'éléments des lexiques et du nombre de grammaire, car cette séparation a été faite *a posteriori* et ne présente pas d'intérêt autre que la précision du balisage.

4.1.2 Faisabilité et méthodologie générale de la reconnaissance des entités nommées en une phase

Un fois établie la partie des entités nommées à retenir pour le balisage, nous devons nous assurer de la faisabilité de leur identification et leur catégorisation selon notre typologie (cf. tableau 3.5). En effet, il nous faut concevoir, en nous basant sur les études précédentes (cf. chapitres 2 et 3), les méthodes, les algorithmes à mettre en place pour parvenir à cette identification et cette catégorisation. Différents problèmes sont alors à résoudre.

L'identification

À l'intérieur du problème général que constitue l'identification automatique se trouve une difficulté particulière : la délimitation des entités nommées (quand s'arrête l'une, quand commence l'autre, problèmes de recouvrement, etc.).

La délimitation de la frontière gauche de l'entité nommée pose essentiellement le problème d'inclure ou non les mots déclencheurs dans le candidat retenu. En effet, dès lors que l'on a fait un choix (cf. section 4.1.1), l'identification de cette frontière gauche est quasi immédiate car constituée de mots déclencheurs (référencés par nos lexiques) ou de formes possédant une majuscule à l'initiale.

La détermination de la frontière droite de l'entité nommée se heurte à un plus grand nombre de difficultés : modification, résolution de l'attachement prépositionnel et de la porté de la coordination.

En plus de cette difficulté de délimitation, se pose le problème de la sur-composition : une entité nommée mixte peut contenir une entité nommée d'une autre catégorie (p. ex. *Guerre d'Algérie*, *Université de Nantes*). Dans certains cas, cette sur-composition peut être multiple (p. ex. *le Festival de l'Université de Natal*, *le gardien de l'équipe de France de football*). Nous avons donc décidé de ne retenir que la forme la plus complète constituant une entité nommée (*le Festival de l'Université de Natal*, *l'équipe de France de football*).

La catégorisation

Là aussi, en plus de la difficulté que pose la catégorisation automatique à proprement parler, il y a un problème supplémentaire. Ce problème réside dans les ambiguïtés portant sur les entités nommées pures (cf. section 2.1.2.1). En effet, une même entité nommée, prise hors contexte, peut appartenir à différentes catégories référentielles : *Paris* peut référer à une ville, un prénom, un nom patronymique, mais peut également composer une partie d'un nom d'organisation (*Paris International*), un nom d'équipe sportive (*Paris Saint-Germain*), un parfum (*Paris d'Yves Saint-Laurent*), etc.

La gestion des coréférences

La coréférence est un problème récurrent en traitement automatique des entités nommées :

Jack Lang, *le ministre J. Lang*, *J. Lang*, *Lang* sont autant de façons différentes de désigner la personne de *Jack Lang*. Ce problème ne se limite pas aux noms de personnes, il touche également les noms d'organisations comme *Ligue des communistes de Yougoslavie*, *LCY*, *Ligue*, etc.

Au vu des résultats des différents systèmes de reconnaissance des entités nommées, il nous apparaît que la solution la plus efficace pour parvenir à la gestion de tous ces paramètres consiste en une approche mixte. Nous avons donc décidé de mettre en place un système fondé initialement sur des méthodes linguistiques (lexiques et règles élaborés manuellement) utilisant les évidences interne et externe. Après que ce « noyau » a atteint des performances difficilement améliorables par ces seules techniques, nous mettrons en place des mécanismes complémentaires à base d'apprentissage, afin d'améliorer les performances de **Nemesis**. En effet, face à l'incomplétude des lexiques et au passage à de nouveaux corpus, les méthodes linguistiques ont montré leurs limites. Nous avons donc choisi d'utiliser ces techniques pour obtenir une première reconnaissance des entités nommées en privilégiant la précision sur le rappel, afin que l'ensemble des exemples ainsi obtenu soit le plus fiable possible pour l'apprentissage.

4.2 Architecture logicielle

L'architecture de **Nemesis**, présentée à la figure 4.1, se compose principalement de quatre modules qui effectuent un traitement séquentiel immédiat des données.

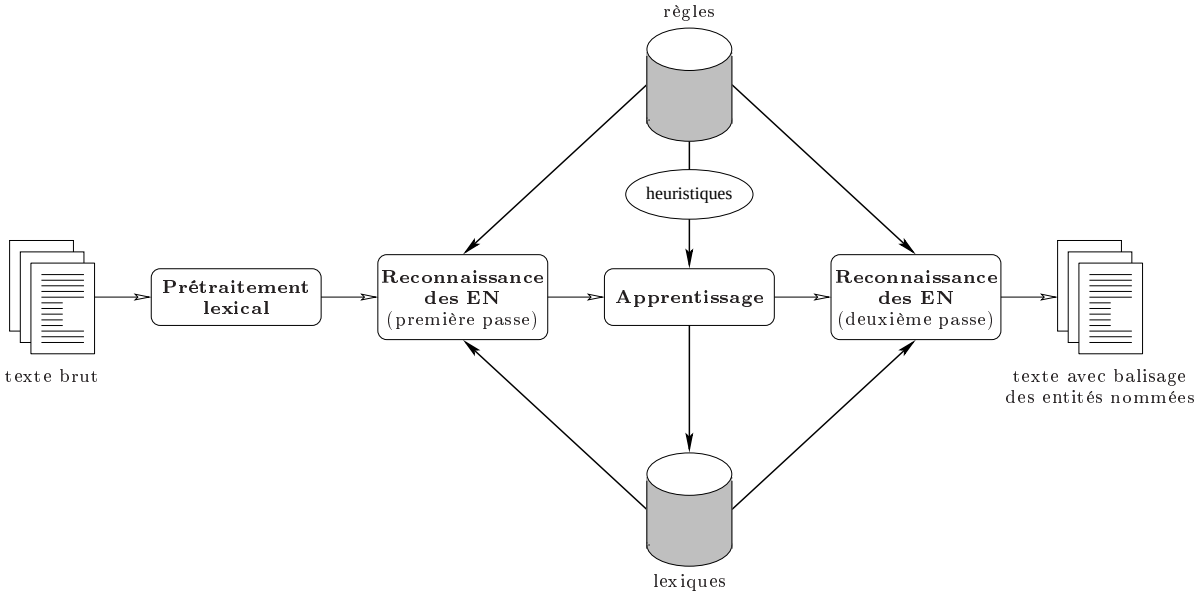


Figure 4.1 – Architecture générale de **Nemesis**

Le premier module constitue une phase de prétraitement lexical qui s'effectue en deux étapes : une segmentation du texte brut en occurrences de formes et de phrases, puis un traitement particulier pour les sigles, qui associe les sigles à leurs formes étendues et balise ces deux éléments. La

segmentation permet la projection des lexiques et le traitement sur les sigles facilite l'identification et la catégorisation d'entités nommées parmi celles qui posent le plus de problèmes comme les sigles et les entités nommées mixtes (cf. section 3.2.2).

Le second module réalise une première reconnaissance des entités nommées. Lors de cette reconnaissance, l'identification et la catégorisation automatiques sont effectuées parallèlement en analysant la structure interne des entités nommées et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que des règles lexico-sémantiques, graphiques et morphologiques. Durant cette première reconnaissance, les sigles et leurs formes étendues sont également catégorisés grâce à des règles examinant cette dernière. Ce module ne traite pas les entités nommées nécessitant l'étude d'un contexte plus large pour être identifiées et catégorisées.

En parallèle à cette première reconnaissance, le module d'apprentissage crée des listes temporaires à l'aide des entités nommées déjà reconnues, en appliquant un ensemble d'heuristiques.

Ces listes d'entités nommées sont utilisées par le dernier module qui procède à une seconde reconnaissance des entités nommées. Ils sont d'abord projetés sur le corpus, avant qu'un certain nombre de règles soient appliquées. Ces règles sont composées de celles appliquées pendant la première phase, plus certaines nouvelles règles qui font intervenir des étiquettes de listes qui n'existaient pas auparavant (noms patronymiques, noms d'évènements, noms d'édifices, etc.). Les règles déclenchées durant cette seconde passe sont prioritaires sur celles de la première : ce module effectue par-là un mécanisme de révision. Cette deuxième passe permet donc d'identifier et de catégoriser les différentes coréférences des entités nommées reconnues lors de la première passe, dont le contexte ne permet pas l'identification ou la catégorisation correcte.

Tout au long des deux phases de reconnaissance des entités nommées, toutes les formes coréférant à une même entité nommée se voient attribuer un même identifiant.

À la suite de ces quatre modules principaux, un filtrage des entités reconnues en début de phrase permet d'écarter celles qui ne seraient pas des entités nommées. Enfin, le fichier, balisé par les deux passes de reconnaissance des entités nommées, est « nettoyé » et formaté en *XML*.

Nous ne présentons, dans cette section, qu'une description des différents modules de **Nemesis**, sans en fournir l'impact ou la précision. L'évaluation de chaque module est présentée à la section 5.1.

4.2.1 Prétraitement lexical

La grande majorité des systèmes effectuant un traitement linguistique sur des textes ont recours à cette phase de prétraitement. Ces prétraitements peuvent aller de la simple segmentation jusqu'à des techniques beaucoup plus complexes et coûteuses (analyses morphologique, syntaxique, sémantique, morpho-syntaxique, etc.). Pour ne pas multiplier les traitements et donc les sources d'erreurs, nous avons choisi de ne pas utiliser d'étiquetage linguistique, mais de nous limiter à un prétraitement lexical qui s'effectue en deux étapes : segmentation du texte, puis association des sigles et de leur forme étendue (cf. figure 4.2.1).

Nous avons choisi d'effectuer cette dernière phase le plus tôt possible et de ne pas en modifier les résultats, car ces derniers sont très fiables (cf. section 5.1.2).

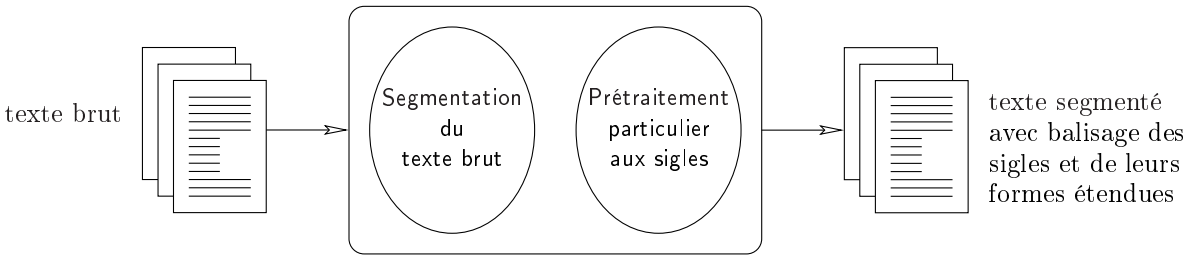


Figure 4.2 – Module de prétraitement lexical

4.2.1.1 Segmentation

Ce prétraitement consiste à isoler chaque forme et chaque phrase du texte. Pour cette segmentation, nous avons repris des scripts *perl* existants, développés dans le cadre du projet ELRA[1] et par Morin [1999], puis nous les avons corrigés et enrichis.

Dans nos corpus de textes bruts, certaines formes sont directement suivies de virgules, de points, d'apostrophes, de parenthèses, etc. Or, pour projeter nos lexiques, il nous faut absolument séparer ces formes des autres éléments du texte de façon à pouvoir effectuer directement une comparaison entre les unités lexicales contenues par nos lexiques et les formes du texte sur lequel nous voulons les projeter. Pour ce faire, nous avons inséré un caractère « blanc » entre les formes du texte et les signes les suivant ou les précédant directement.

Cette phase de la segmentation comportent des règles dont certaines tiennent compte d'une mauvaise typographie. Ces règles mettent en jeu les symboles suivants :

.	?	!	,	;	:	..	...	–	/	(	)	[	]	{	}	«	»	"	’

Nous ne détaillons pas dans quelles mesures ces symboles sont « détachées » des autres formes lexicales, et par quels moyens. En effet, il s'agit là d'un prétraitement indispensable à tout traitement automatique sur des textes électroniques. De fait, c'est une tâche qui ne pose plus de réels problèmes et qui a largement été étudiée [Friburger, 2000 ; Grefenstette et Tapanainen, 1994 ; Aït-Mokhtar, 1997].

Il est à noter que cette opération est effectuée une seule fois par corpus ; cependant, une phase de formatage, qui reprend un processus presque similaire, mais inverse, sera nécessaire à la visualisation (cf. section 4.2.5.2) et à la comparaison entre les résultats obtenus et les résultats attendus (cf. section 5.1).

Lors de cette phase, **Nemesis** traite également le corpus, afin qu'il soit composer d'une et une seule phrase par ligne. La difficulté réside dans :

[1]Constitution d'un corpus scientifique du français moderne (1999-2000)
Projet industrielle avec la revue *La Recherche* (resp. Béatrice Daille, LINA)
Ce projet avait pour objectif la réalisation d'un corpus scientifique du français moderne à partir d'articles de la revue *La Recherche* dans le cadre d'une convention avec ELRA (European Language Resources Association). Il s'agissait de créer un corpus monolingue qui puisse donner une vision des usages des différentes disciplines scientifiques. Le corpus représente une ressource d'environ un million de mots constituée de 180 articles de la revue *La Recherche* couvrant une trentaine de thèmes entre les années 1998 et 1999.

- l'ambiguïté du point, qui peut composer un sigle, un acronyme ou une abréviation, ou marquer la fin de la phrase, voire les deux en même temps [Dister, 1997] ;
- l'ambiguïté de la majuscule à l'initiale, qui peut composer une entité nommée ou un acronyme, ou marquer le début d'une phrase, voire même aucun des deux (partie d'un texte entièrement rédigée en majuscule) ou les deux à la fois.

Notre script commence par ajouter un retour à la ligne (`RC`) s'il rencontre un caractère autre qu'une majuscule, suivie d'un caractère de fin de phrase ([.] [?] [!]), d'un blanc et d'une majuscule, ce qui permet de composer avec la majorité des cas (notamment les sigles en fin de phrase). Ensuite, à l'aide d'une liste des abréviations les plus courantes se terminant par un point ([*cf.*] [*e.g.*] [ex.] [*i.e.*] [*vs.*] [M.] [*Mr.*] [Dr.] [J.-C.] [c.-à-d.] [tél.] , etc.), il supprime les `RC` se situant immédiatement après ces abréviations.

La difficulté principale de la segmentation en phrases réside dans le traitement des citations. En effet, une citation peut constituer une nouvelle phrase ou bien être intégrée dans la phrase courante, et une même citation peut contenir plusieurs phrases. Or, le schéma de citation perturbe les patrons utilisés préalablement (présence des guillemets, fin de la phrase précédente par [:] , enchaînement de plusieurs citations constituant des phrases différentes et séparées par [,] , etc.). Nous avons donc ajouté des règles qui placent sur une nouvelle ligne chaque citation constituant une nouvelle phrase, uniquement si la phrase dans laquelle la citation est insérée ne se poursuit pas derrière. En effet, cela nuirait à l'unité de la phrase initiale et nous estimons qu'il vaut mieux y avoir plusieurs phrases sur la même ligne qu'une même phrase sur plusieurs lignes. Enfin, et pour les mêmes raisons, nous avons ajouté une règle qui « rassemble » deux lignes si la première comporte une parenthèse qui a été ouverte mais pas fermée.

Nous n'avons pas évalué directement la précision de cette segmentation, car notre but n'était pas ici de réaliser un module de segmentation. En revanche, des erreurs dans ce prétraitement peuvent dégrader les résultats de la reconnaissance des entités nommées et c'est cette dégradation qu'il nous intéresse d'évaluer. Or, nous n'avons trouvé aucune erreur induite par ce prétraitement lors de l'évaluation de **Nemesis** (cf. section 5.1). Nous pouvons donc considérer que la segmentation en occurrences de formes et de phrases est correctement effectuée dans le cadre de notre système.

4.2.1.2 Association des sigles et de leur forme étendue

Les sigles sont très présents dans les corpus que nous avons étudiés (cf. tableau 3.4) et font le plus souvent référence à des entités nommées. Cependant, parmi les différents dispositifs de reconnaissance des noms propres que nous avons étudiés (cf. chapitre 2), seuls Wolinski et coll. [1995] et Wacholder et coll. [1997] utilisent l'association entre les sigles et leur forme étendue, mais uniquement pour découvrir des coréférences en ce qui concerne Wacholder et coll. [1997], et non pour aider à la reconnaissance des entités nommées. Pourtant, pouvoir associer ces sigles et leur forme étendue présente un intérêt multiple.

En premier lieu, cela permet l'identification de la forme étendue – dont il est difficile de fixer avec précision les limites, à droite notamment – ainsi que la catégorisation du sigle, lorsque celle de la forme étendue est accomplie (cf. section 4.2.2). Ces deux opérations sont réalisées uniquement en étudiant les structures locales. En effet, dans les corpus que nous avons étudiés, lorsqu'un sigle apparaît pour la première fois dans un texte, il est souvent accompagné de sa forme étendue dans une fenêtre textuelle de taille réduite, sous différents schémas graphiques

qui permettent leur association : `(SIGLE) Forme étendue`, `Forme étendue (SIGLE)`, etc. (p. ex. *la FFF (Fédération française de football)*, *la Fédération nationale des agences d'urbanisme (FNAU)*).

Le sigle ainsi que sa forme étendue peuvent également apparaître isolément dans le texte, éventuellement avant, mais plus souvent après avoir été rencontrés dans un tel schéma. Il va donc être également possible, grâce à ce traitement sur les sigles, d'identifier et de catégoriser ces occurrences isolées lors de la seconde reconnaissance des entités nommées, par la mise à jour des lexiques qui sera effectuée par le module d'apprentissage (cf. section 4.2.3).

Il est à noter que les catégories graphiques représentées par les sigles et leurs formes étendues (respectivement `Sigles` et `EN mixtes`) sont celles qui sont les plus difficiles à catégoriser pour les `Sigles` et à identifier pour les `EN mixtes`. Un tel traitement sur les sigles permet de résoudre, en même temps, ces deux problèmes, pour un nombre non négligeable d'entités nommées.

Le traitement des sigles que nous avons implémenté se fonde sur un outil d'extraction de sigles réalisé par Morin [1999], que nous avons modifié car les sigles qu'il traite ne sont pas uniquement des entités nommées.

L'analyse de Morin [1999] permet de dégager les points suivants :

- un sigle correspond à une seule unité lexicale ;
- le premier élément d'un sigle ou d'un acronyme correspond à la première lettre de sa forme étendue ;
- un sigle est composé de lettres majuscules, de chiffres et de symboles (p. ex. « *tiret* » et « *et commercial* ») ;
- un sigle est composé d'au moins deux symboles et comprend toujours une lettre ;
- les éléments d'un sigle peuvent être séparés par un point ;
- un sigle peut se terminer par la lettre « *s* » minuscule pour désigner la marque du pluriel en anglais.

Partant du fait que, dans les textes, le sigle apparaît une première fois avec sa forme étendue avant d'être utilisé seul, Morin [1999] a mis en évidence différentes constructions qui rendent compte de la création d'un sigle. La construction la plus classique consiste encore à faire suivre la forme étendue par le sigle entre parenthèses et inversement :

Délégation à l'aménagement du territoire et à l'action régionale (DATAR) (4.1)
CNT (Caisse nationale des télécommunications) (4.2)

Nous retiendrons essentiellement ces deux formes, car les autres formes possibles pour les acronymes, proposées par Morin [1999], font référence à des formes étendues qui ne sont pas des entités nommées, mais des unité lexicales complexes renvoyant généralement à des termes scientifiques (*« rice hoja blanca virus » (RHBV)*, *virus de la mosaique africaine du manioc (« ACMV »)*, *protéines de transfert de lipides (ou LTP)*, *(RML, reste de masse des litières)*, etc.).

À partir des remarques précédentes, Morin [1999] a défini la méthodologie suivante pour extraire automatiquement un sigle et sa forme étendue :

1. identifier en corpus les différentes constructions ;
2. extraire de la construction un sigle candidat et une forme étendue candidate[2] ;
3. associer le sigle et la forme étendue s'il y a adéquation entre les deux.

[2]Dans le cas des constructions où la forme étendue candidate n'est pas contrainte sur sa partie gauche, Morin [1999] remonte jusqu'au début de la phrase.

```
% Définition de la construction (1) %
% Exemple : Délégation à l'aménagement du territoire et à l'action régionale ( DATAR ) %
Construction_1 ← "[^(]+ ( [^ ]+ )"
...
Indice ← 1
Tant que ( pas fin de corpus ) Faire
    Phrase ← LirePhraseCorpus ()
    Pour ( toutes les constructions ) Faire
        Tant que ( Instanciation (Phrase, Construction_i) ) Faire
            FormeEtendueCandidate ← ExtraireFormeEtendue (Phrase)
            % Exemple : Délégation à l'aménagement du territoire et à l'action régionale %
            SigleCandidate ← ExtraireSigle (Phrase)
            % Exemple : DATAR %
            SiglePatron ← SigleCandidat
            Pour ( chaque lettre de SiglePatron ) Faire
                Substituer ("A", "[AÁÂaàâ].*")
                Substituer ("B", "[B].*")
                Substituer ("C", "[C].*")
                Substituer ("D", "[D].*")
                Substituer ("E", "[EÉÈÊËeéèêë].*")
                ...
            Fin Pour
            SiglePatron ← " ([1-9]+-)?" + SiglePatron
            % Exemple : ␣([1-9]+-)?[D].*[AÁÂaàâ].*[T].*[AÁÂaàâ].*[R].* %
            SiglePatronInverse ← InverserSigle (SiglePatron)
            % Exemple : ␣([1-9]+-)?[R].*[AÁÂaàâ].*[T].*[AÁÂaàâ].*[D].* %
            Si ( Instanciation (FormeEtendueCandidate, SiglePatron) Et
            LimiteTailleSigle (SigleCandidat) Et LimiteDébutSigle (SigleCandidat) ) Alors
                Baliser (Phrase, SigleCandidat, SIGLE, Indice)
                Baliser (Phrase, FormeEtendueCandidate, DEF, Indice)
                Écrire (Phrase, FichierResultat)
                Incrémenter (Indice)
            Sinon
                Si ( Instanciation (FormeEtendueCandidate, SiglePatronInverse) Et
                LimiteTailleSigle (SigleCandidat) Et LimiteDébutSigle (SigleCandidat) ) Alors
                    Baliser (Phrase, SigleCandidat, SIGLE, Indice)
                    Baliser (Phrase, FormeEtendueCandidate, DEF, Indice)
                    Écrire (Phrase, FichierResultat)
                    Incrémenter (Indice)
                Fin Si
            Fin Si
        Fin Tant que
    Fin Pour
Fin Tant que
```

Algorithme 4.1 – Extraction d'un sigle et de sa forme étendue

La première étape, qui doit être effectuée manuellement, ne fait pas partie du programme. Elle peut être enrichie de nouvelles constructions, si nous étions amenés à en trouver. Après une évaluation des sigles extraits et une légère adaptation de l'algorithme de base [Morin, 1999], nous obtenons l'algorithme 4.1[3].

Une fois les sigles et leur formes étendue identifiés et associés, il nous faut adapter et intégrer les résultats de ce traitement dans notre système. Plutôt que de créer une liste des sigles et de leurs formes étendues pour nos textes, nous avons choisi de les baliser en affectant un numéro à chacun des couples pour pouvoir associer le bon sigle et la bonne forme étendue. Si un sigle apparaît deux fois avec sa forme étendue, nous affecterons un numéro différent à chaque couple.

Nous aurons, par exemple :

Fédération nationale des agences d' urbanisme (FNAU)
⟶ **<DEF 1> Fédération nationale des agences d' urbanisme </DEF 1>**
(<SIGLE 1> FNAU </SIGLE 1>)

Syndicat d' études et de programmation de l' agglomération lyonnaise (SEPAL)
⟶ **<DEF 2> Syndicat d' études et de programmation de l'agglomération lyonnaise </DEF 2>**
(<SIGLE 2> SEPAL </SIGLE 2>)

Une fois les sigles et leur formes étendue associés et balisés, il nous faudra les catégoriser lors de la première reconnaissance des entités nommées (cf. section 4.2.2).

4.2.2 Première reconnaissance

La première reconnaissance des entités nommées analyse la structure « interne » de celles-ci. Cette structure, appelée « évidence interne » (de l'anglais *internal evidence*) et définie par McDonald [1996], est constituée des différentes formes composant l'entité nommée.

Outre l'évidence interne, cette première reconnaissance considère également une partie du contexte des entités nommées, autrement appelé « évidence externe » (de l'anglais *external evidence*) et défini là encore par McDonald [1996]. Les éléments de l'évidence externe que nous étudions sont les contextes immédiats – gauche et droit – de l'entité nommée. Par contexte immédiat, nous entendons les mots ne faisant pas partie intégrante de l'entité nommée, mais se trouvant immédiatement à sa gauche – dans le même syntagme nominal – et à sa droite – en apposition – et permettant de l'identifier et de la catégoriser (p. ex. *Mr* ***EN*** ou ***EN,*** *le philosophe grec* ou *le mont* ***EN*** ou ***EN,*** *capitale...* ou *le festival de* ***EN*** ou ***EN,*** *le cyclone...*, etc.).

Pour parvenir à cette première reconnaissance par l'analyse de ces différents indices, nous utilisons d'une part des dictionnaires électroniques d'entités nommées et des lexiques de mots déclencheurs (cf. section 4.2.2.1) – servant comme indices pour l'identification et la catégorisation des entités nommées – et d'autre part des règles de réécriture (cf. section 4.2.2.2). Ces règles sont appliquées après la projection des lexiques et se fondent sur ces lexiques, ainsi que sur des indices graphiques et morphologiques, pour repérer la présence d'entités nommées, les identifier et enfin les catégoriser (cf. figure 4.3).

4.2.2.1 Création et projection des lexiques

L'utilisation de dictionnaires électroniques d'entités nommées est courante dans les systèmes déjà développés en reconnaissance des noms propres (cf. chapitre 2). De plus, Mikheev et coll.

[3]Les modifications apportées par rapport à l'algorithme de Morin [1999] apparaissent en gris

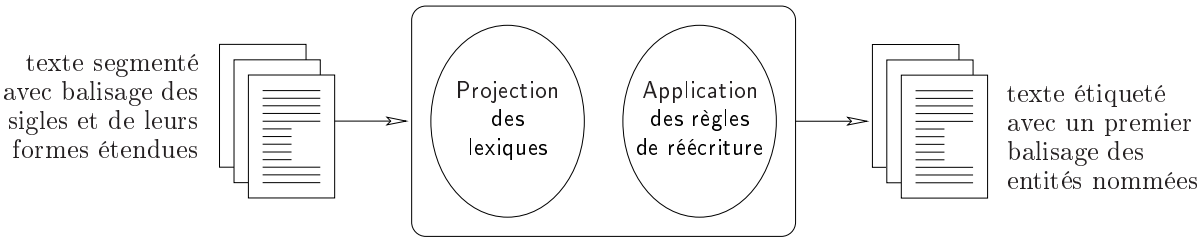


Figure 4.3 – Module effectuant la première reconnaissance des entités nommées

[1999] démontrent le caractère indispensable, mais insuffisant, du recours à des listes minimales d'entités nommées et suggèrent de combiner leur utilisation à une analyse des évidences internes et externes pour obtenir de bons taux de rappel et de précision. Nous avons donc choisi d'exploiter des dictionnaires électroniques d'entités nommées, mais également des lexiques de mots déclencheurs qui nous permettent cette analyse.

Nos lexiques sont au nombre de 61 et regroupent un total de 77526 unités lexicales. Ils constituent des fichiers contenant une ou plusieurs formes par ligne représentant une seule unité lexicale de ce lexique (la liste de tous les lexiques utilisés par **Nemesis**, ainsi que leur taille, sont présentés à l'annexe A). Ces unités lexicales peuvent tenir un ou plusieurs rôles, selon les catégories d'entités nommées que les règles dans lesquelles elles sont utilisées permettent d'identifier[4] :

`EN` : l'unité lexicale est une entité nommée connue (p. ex. *OMS*, *Alexandre*, *Canal+*) ;

`mot déclencheur` : l'unité lexicale fait partie de l'entité nommée (p. ex. *Fédération*, *Boulevard*) ;

`contexte` : l'unité lexicale appartient au contexte immédiat de l'entité nommée, mais ne fait pas partie de celle-ci (p. ex. *philosophe*, *français*) ;

`fin d'EN` : l'unité lexicale est la dernière forme composant l'entité nommée (p. ex. *football*, *régional*) ;

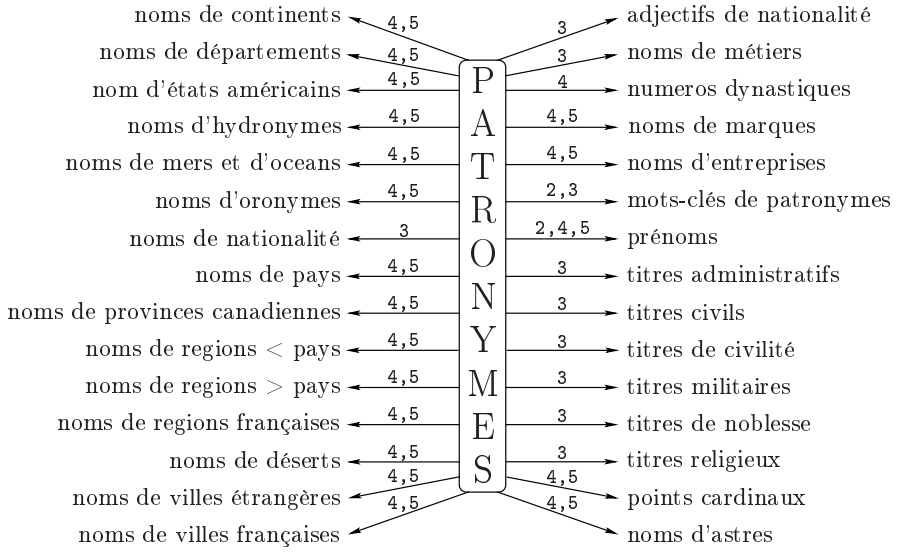
`élément d'EN` : il s'agit de toutes les unités lexicales pouvant faire partie de l'entité nommée, mais sans en permettre la délimitation ou la catégorisation.

Nous pouvons donc assigner des rôles à nos lexiques, en fonction des catégories référentielles pour la reconnaissance desquelles ils sont utilisés. Cette assignation peut être visualisée sous deux angles :

1. en prenant comme point central une catégorie référentielle : p. ex. la reconnaissance des patronymes nécessite l'utilisation des unités lexicales du dictionnaire des noms de pays comme `fin d'EN` ou `élément d'EN` (cf. figure 4.4) ;
2. en prenant comme référent un lexique : p. ex. les unités lexicales du dictionnaire des noms de pays sont utilisées uniquement comme `fin d'EN` pour la reconnaissance des ensembles artistiques (cf. figure 4.5).

Chaque catégorie référentielle utilise un nombre réduit de lexiques (cf. tableau 4.1). Sur nos 27 catégories d'entités nommées, seuls cinq d'entre elles (les patronymes, les organisations, les ensembles artistiques, les entreprises industrielles et les établissements d'enseignement et de

[4]Les unités lexicales d'un même lexique peuvent avoir des fonctionnements hétérogènes, même au sein d'une catégorie.



LÉGENDE : `1 EN - 2 Mot déclencheur - 3 Contexte - 4 Fin d'EN - 5 Élément d'EN`

Figure 4.4 – Rôles des lexiques pour la reconnaissance des patronymes

recherche) utilisent plus de dix de nos 61 lexiques. Ce nombre plus important s'explique par la grande variété de mots pouvant composer les entités nommées de ces deux catégories. Malgré tout, nous ne tenons pas compte, dans ce calcul, de l'utilisation indirecte des lexiques. En effet, dans nos règles, nous voulons parfois rechercher n'importe quelle forme avec une majuscule à l'initiale, ce qui inclut fatalement de nombreuses unités lexicales de nos lexiques ; on ne peut pas considérer que ces dernières soient réellement utilisées.

Création Pour obtenir nos lexiques, nous avons procédé de trois façons différentes :

1. En adaptant des lexiques existants.
2. En les créant manuellement.
3. En les créant automatiquement à partir du *Web*.

Nous nous sommes donc d'abord fondé sur les nombreux lexiques existants, qu'il est aisé de se procurer (prénoms, noms de villes, de pays, adjectifs de nationalité, etc.). Malheureusement, ces lexiques sont loin d'être parfaits : ils sont le plus souvent trop réduits – auquel cas il faut les étendre – ou trop volumineux – ce qui implique de les filtrer ou d'en trouver d'autres pour les remplacer.

Pour le lexique des prénoms, par exemple, nous avons commencé avec un lexique comportant plus de 313 000 entrées. La projection de ce dernier engendrait beaucoup de bruit – car trop volumineux – et provoquait l'étiquetage de nombreux mots qui ne sont pas des prénoms dans nos textes, ni même parfois des entités nommées. Par conséquent, nous avons trouvé et testé un lexique comportant un peu plus de 3 400 prénoms. Là, le problème inverse – moins important de notre point de vue – s'est posé : certains prénoms, présents dans nos échantillons de corpus, ne

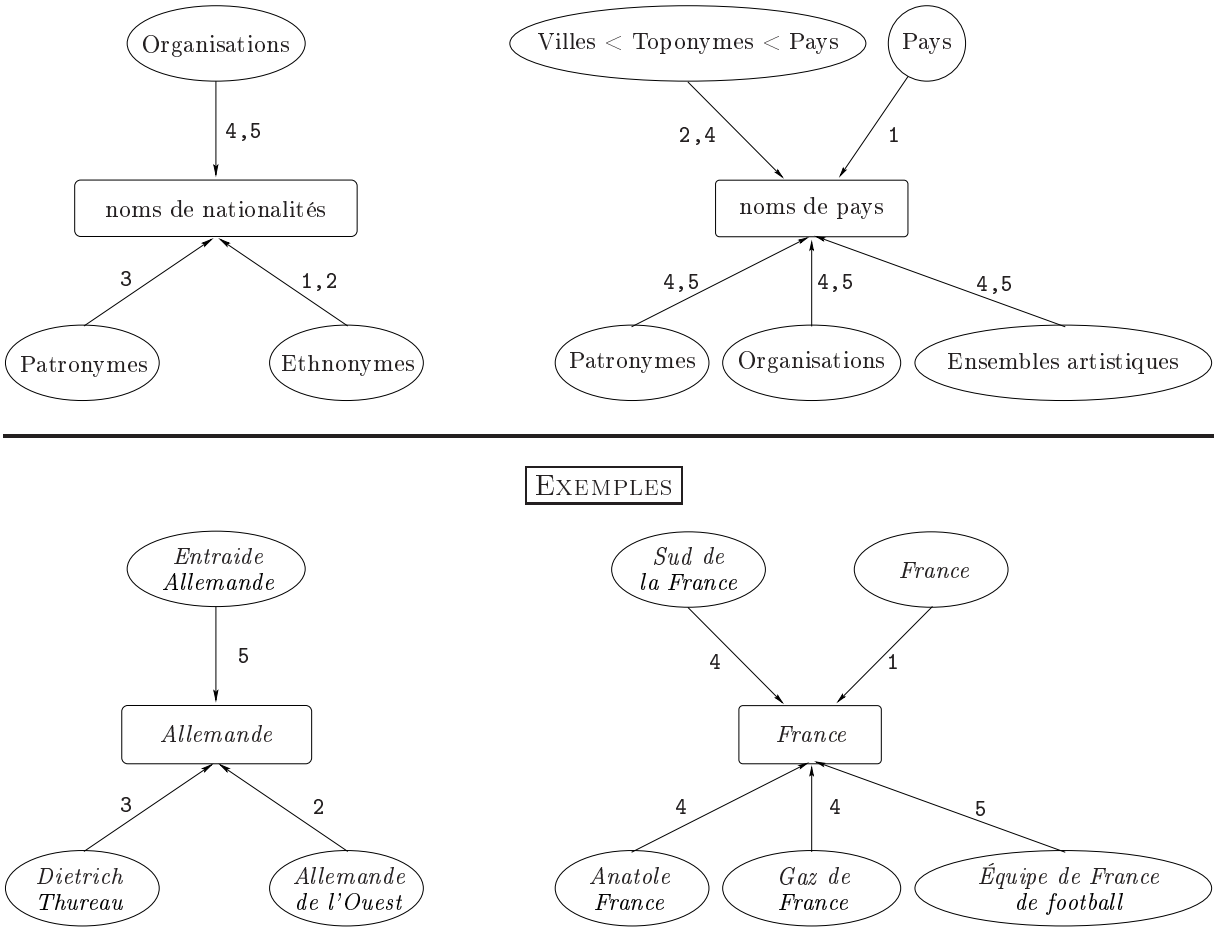


Figure 4.5 – Deux lexiques et leur rôles selon les catégories référentielles

l'étaient pas dans notre lexique (*Joël*, *Démis*, *Hannu*, *Juha*, etc.). Nous avons finalement trouvé et opté pour un lexique intermédiaire de 7 668 prénoms.

Une fois les lexiques existants exploités, nous avons du en créer d'autres de toute pièce. Pour cela, deux méthodes ont été retenues. Pour les lexiques de faible volume, il s'agissait de créer manuellement une liste, la plus exhaustive possible, des mots composant ces lexiques, comme par exemple pour :

- les sports : *basket-ball*, *cyclisme*, *football*, *handball*, *rugby*, *tennis*, etc. ;
- les titres religieux : *archevêque*, *cardinal*, *évêque*, *frère*, *mère*, *mère supérieur*, *moine*, *pape*, *père*, *prêtre*, *révérend*, *révérend père*, *sœur* ;

Bien que la plupart des lexiques ainsi créés ne soient pas exhaustifs (p. ex. les sports), ils contiennent les unités lexicales les plus couramment usités.

Tableau 4.1 – Nombre de lexiques utilisés par catégorie référentielle

Anthroponymes					
Ensembles artistiques	30	Patronymes	30	Prénoms	1
Organisations	25	Ethnonymes	2		
Toponymes					
Villes < Toponymes < Pays	7	Édifices	5	Pays	1
Toponymes > Pays	6	Hydronymes	4	Villes	1
Oronymes	6	Microtoponymes	2	Rues	1
Ergonymes					
Entreprises industrielles	30	Marques et produits	2		
Établissements d'ens. et de rech.	26	Œuvres matérielles	1		
Praxonymes					
Évènements	2	Faits historiques	1		
Périodes historiques	2	Œuvres abstraites	1		
Phénonymes					
Astres et comètes	2	Phénomènes naturels	1		

Concernant les lexiques pour lesquels il n'était pas envisageable de créer manuellement de telles listes, la technique employée consiste à récupérer sur le *Web* des pages qui listent des noms propres contenant les mots que nous recherchons. Ces fichiers contiennent souvent des entités nommées mixtes, dont seule une partie nous intéresse. De plus, les pages *Web* sont « polluées » par des balises ou d'autres informations qui ne nous intéressent pas. Nous filtrons donc ces fichier, afin d'obtenir un fichier de la même forme que nos autres lexiques (une unité lexicale par ligne). Malgré tout, il reste souvent quelques mots qu'il faut retirer et cela ne peut être fait que manuellement. Prenons l'exemple du lexique répertoriant les mots clés d'organisations.

En premier lieu, les balises, les sigles, la ponctuation, etc. sont retirés du fichier *html*. À partir du fichier ainsi obtenu, nous ne gardons que le premier mot de chaque ligne, à condition que celui-ci commence par une majuscule. Ensuite, nous ne conservons qu'une occurrence par unité lexicale. Enfin, nous effectuons manuellement un dernier filtrage du lexique, afin d'éliminer les unités lexicales qui ne conviennent pas.

Une fois nos lexiques obtenus par l'une de ces trois méthodes (récupération, création manuelle ou création automatique), nous avons effectué quelques enrichissements :

- la plupart des unités de nos lexiques ont un genre et un nombre unique. Or, pour certains d'entre eux, nous souhaitons avoir les formes fléchies ; nous les avons donc ajoutées ;
- pour les lexiques de noms communs (les différents mots clés, les adjectifs, les métiers, les sports et les titres), nous avons utilisé la base de données terminologiques du français *EuroWordNet* et le dictionnaire des synonymes en ligne du laboratoire de linguistique CRISCO[5], afin de réunir les synonymes de ces noms communs, que nous avons filtrés manuellement et ajouté aux lexiques concernés.

Quelle que soit la méthode employée pour les obtenir, ces lexiques ne peuvent pas contenir tous les éléments nécessaires et il nous faudra une solution face à cette incomplétude. Il est également à noter qu'en plus de ces lexiques, nous utilisons deux variables regroupant les entités fonctionnelles et les particules pouvant intervenir dans les noms de personnes. En effet, la projection de lexiques

[5] http://elsap1.unicaen.fr/cherches.html

contenant ces deux types d'unités lexicales est beaucoup trop longue car ces dernières sont très fréquents dans les textes français.

Projection Une fois les lexiques créés, il nous faut maintenant les projeter sur le corpus à traiter. Pour cela, le corpus est tout d'abord transformé en un fichier inverse [Salton et McGill, 1983], afin de limiter les accès disques[6]. Ce fichier inverse est une table d'association, dans laquelle chaque clé est un mot et la valeur associée représente la liste de toutes les coordonnées où l'on retrouve ce mot dans le corpus.

Ensuite, les lexiques sont projetés unité lexicale par unité lexicale sur le corpus en mémorisant les formes qui correspondent à l'aide d'étiquettes lexico-sémantiques attachées à chaque lexique (une étiquette par lexique correspondant à la nature des éléments qui le composent). Ces unités lexicales peuvent comporter une ou plusieurs formes et le traitement, dans ce dernier cas, est bien plus complexe. En effet, si pour les unités lexicales à une seule forme, il suffit de regarder chaque clé du fichier inverse et lui affecter une étiquette s'il correspond, le traitement pour les unités lexicales à plusieurs formes est plus complexe : il faut comparer la première forme à chaque clé du fichier, puis vérifier que les formes suivantes se trouvent bien présentes dedans, à des coordonnées qui suivent celles la première clé trouvée à l'intérieur du corpus (cf. algorithme 4.2[7]).

```
Pour ( chaque unité lexicale du lexique ) Faire
    Si ( c'est une unité lexicale avec une seule forme ) Alors
        Marquer cette forme avec l'étiquette appropriée dans la variable %newtable
        Ajouter cette forme dans la variable %savetable
    Sinon
        Séparer les formes de l'élément
        Mettre ces formes dans une liste
        Tant que ( la liste n'est pas vide  Et
        la forme de la liste est présente dans la variable %table  ) Faire
            Extraire les coordonnées de la forme
        Fin Tant que
        Si ( toutes les formes de l'unité lexicale existent dans %table ) Alors
            Pour ( chaque coordonnée de la première forme ) Faire
                Vérifier pour les autres formes de l'unité lexicale
                l'existence de coordonnées qui suivent celles de la première
                Si ( les formes d'une unité lexicale sont retrouvées comme étant adjacentes ) Alors
                    Ajouter l'étiquette dans le tableau %multipletable, aux indices correspondant
                    aux coordonnées de toutes les formes de l'unité lexicale
                Fin Si
            Fin Pour
        Fin Si
    Fin Si
Fin Pour
```

Algorithme 4.2 – Projection d'un lexique

Enfin, les formes possédant une majuscule à l'initiale sont marquées comme étant potentiel-

[6] Du fait de la grande taille des données, il est important de tenir compte des temps d'exécution.

[7] La variable `%table` contient le fichier inverse correspondant au corpus à traiter.

lement des entités nommées (par l'étiquette NP), et le fichier inverse est repassé, avec les clés étiquetées, sous forme de corpus, pour pouvoir poursuivre le traitement.

Voici un exemple de texte étiqueté par la projection des lexiques sur un fichier de corpus :

> Elle/PRENOM fabriquera , dans un premier temps , le produit/CLE-MARQUE liquide qui entre dans le processus des photocopies ainsi que des pièces détachées pour la filiale/CLE-ENTREPRISE de Minolta/MARQUE en RFA/PAYS _._ Le/NP Français/NOM-NATIONALITE Didier/PRENOM Auriol/NOM/VILLE (Lancia/MARQUE intégrale 16 S/NP) a pris la tête du rallye/CLE-EVENEMENT de Monte-Carlo/VILLE à l' issue de l' étape de classement (560 km dont six spéciales chronométrées de 124 km) disputée dimanche 21 janvier entre Monaco/VILLE et Aubenas/VILLE _._

4.2.2.2 Formalisme et application des règles

Une fois la projection des lexiques réalisée, les balises posées lors du prétraitement sur les sigles (cf. section 4.2.1.2) sont exploitées et les règles de réécriture sont appliquées. Ces règles poursuivent un objectif double :

1. Elles vont permettre d'identifier et de catégoriser automatiquement certaines entités nommées du textes, qui correspondent à une partie ou l'ensemble d'un motif, en les délimitant à l'aide de balises correspondant à leur catégorie.
2. Lorsqu'un motif est reconnu dans le texte, elles vont extraire, de ce motif, certaines informations nécessaires à la création de nouveaux lexiques, qui seront utilisés lors de la seconde passe.

Formalisme Les règles de réécriture s'appuient essentiellement sur les évidences internes et externes définies par McDonald [1996] et utilisent des patrons basés principalement sur les étiquettes lexico-sémantiques correspondant aux lexiques, mais aussi sur des indices morphologiques ou graphiques. Voici le formalisme retenu pour la conception de ces règles :

- X, une règle de réécriture ;
- $P_X \rightarrow A_X$, la forme générale de X ;
- P_X, le patron de X ;
- A_X, les actions à effectuer sur les éléments des « regroupements[8] » de P_X ;
- $[w_X^1, w_X^n]$ (n *entier* ≥ 1), un intervalle discret, représentant P_X, avec w_X^1 et w_X^n respectivement les éléments de début et de fin du patron. Un ou plusieurs éléments de P_X peuvent être « regroupés » entre crochets ;
- w_X^i, le $i^{ème}$ élément de P_X, qui peut être :

  Forme une forme quelconque (p. ex. la forme *et* peut être directement recherchée),

  Morphème une forme possédant un caractère morphologique particulier (p. ex. un suffixe caractéristique des noms d'habitants comme *-ais*, *-in*, *-ois*, *-on*, etc.),

  Balise une étiquette lexico-sémantique référant à une forme appartenant à un lexique (un nom de pays, un métier, un mot clé d'organisation, etc.),

  NP une étiquette référant à une forme n'appartenant pas à un lexique, mais comportant une majuscule à l'initiale ;

- $[A_X^1, A_X^m]$ (m *entier* ≥ 1), un intervalle discret, représentant A_X ;

[8] Nous entendons ici par le mot regroupement le sens qui lui est donné dans le cadre des expressions régulières.

- A_X^j, l'action à effectuer sur les éléments du $j^{ème}$ regroupement de P_X. A_X^j peut être :
  $Catégorie_X^j$ le nom de la balise à apposer autour des éléments du regroupement, correspondant à une catégorie d'entités nommées,
  $Fichier_X^j$ le nom du fichier temporaire correspondant au lexique à mettre à jour (un regroupement pour un élément du lexique).

À chaque w_X^i peut être associé un quantificateur : `?` (zéro ou une fois), `+` (une fois ou plus), `*` (zéro fois ou plus). Au niveau de l'implémentation (sous forme d'expressions régulières), nous distinguons les **formes** des **balises**, des **morphèmes** et des **NP**, en faisant précéder les trois derniers d'un dollar :

```
[ [ $Suffixe_habitant ] ] [ $Prenom+ $NP* [ $Particule? $NP+ $Numéro?] ]
```
→ **Ethnonyme Patronyme** (p. ex. *le castelroussin Jimmy Algérino*)

```
[ $Clé_équipe $article_min $pays $article_min $sport ]
```
→ **Ensemble_artistique** (p. ex. *l'équipe de France de football*)

Dans A_X, les noms de catégories sont différenciés des noms de fichiers en mettant ces derniers entre les caractères « inférieur » et « supérieur » :

```
[ $Clé_astre+ $Article? [ $NP ] ]
```
→ **Astre <noms-astres>** (p. ex. *la planète Mars*)

```
[ $Titre_de_civilité $Prenom* $NP* [ $Particule? $NP+ $Numero? ] ]
```
→ **Patronyme <noms-patronymiques>** (p. ex. *Don Diego de la Vega*)

Certaines règles sont « factorisées » : si plusieurs règles ont exactement la même forme à un w_X^i prêt, elles sont unifiées en remplaçant le w_X^i différent par une variable représentant l'un ou l'autre ($A\ B\ C \rightarrow Catégorie$ et $A\ B\ D \rightarrow Catégorie$ donnent $A\ B\ CouD \rightarrow Catégorie$). Nous avons, par exemple, la variable `$Maj` qui représente toute forme possédant une majuscule à l'initiale (appartenant à nos lexiques ou non), la variable `$Fin_org` qui contient toutes les formes susceptibles de clôturer un nom d'organisation (les mêmes éléments que `$Maj`, mais aussi les adjectifs géographiques, les noms de sports, etc.), ou encore la variable `$Tous` qui regroupe tous les éléments balisés, les formes prenant une majuscule ou un chiffre à l'initiale, et les entités fonctionnelles :

```
[$Clé_hydro $Article_min+ $MAJ+]
```
→ **Hydronyme** (p. ex. *rives de la Kamogawa*)

```
$Clé_organisation $Tous* $Fin_org ]
```
→ **Organisation** (p. ex. *Association des sylviculteurs du Sud-Ouest*)

À chaque w_X^i peut être associé un rôle. Les premiers éléments des patrons regroupent les formes possédant les rôles de `contexte` ou de `mot déclencheur`, alors que les derniers ont plutôt pour rôle `fin d'EN`. Ces derniers sont donc moins « fiables », car ils ne permettent pas de catégoriser les entités nommées, ni d'en identifier la présence, mais simplement d'en définir la limite à droite. Quant aux éléments de type `élément d'EN`, ils sont encore moins fiables de par leur nature. Nous avons conçu 95 règles de cette forme pour la première passe de reconnaissance des entités nommées : 25 pour les anthroponymes, 49 pour les toponymes, 12 règles pour les ergonymes, 6 règles pour les praxonymes et 3 règles pour les phénonymes.

Application Avant l'application proprement dite des règles de réécriture, il nous faut trouver un moyen d'exploiter l'information que procurent les balises posées autour des sigles et de leur forme étendue (cf. section 4.2.1.2). Pour y parvenir, nous procédons comme suit :

1. Pour chaque ligne la présence d'un sigle contenant un mot déclencheur de certaines catégories est détectée [9].
2. Le sigle, ainsi que sa forme étendue sont balisés suivant le type correspondant au mot déclencheur.
3. Les lexiques sont mis à jour avec les entités nommées identifiées et catégorisées.
4. Le sigle et sa forme étendue sont retirés de la ligne à traiter par l'application des règles, afin de ne rien ajouter qui pourrait bruiter l'identification et la catégorisation de ces deux entités nommées.

Ensuite, nous procédons à la compilation des règles pour vérifier leur validité et préparer à leur application. Cette compilation commence par une vérification simple de la syntaxe de ces règles, à l'aide des listes contenant les variables et les balises valides :

- chaque règle doit posséder une partie gauche et une partie droite séparée par une flèche ;
- la partie gauche doit être composée de variable valides ou de mots simples ;
- la partie droite doit être composée de balises valides ou d'un nom de fichier entre les caractères « inférieur » et « supérieur ».

Si, ne serait-ce qu'une seule de ces clauses n'est pas valide, le traitement s'arrête en indiquant le numéro de la règle défaillante, ainsi que la raison de sa non validité.

Ensuite, chaque règle est transformée en une expression régulière correspondant.

Enfin, chaque ligne du corpus est traitée de façon à y appliquer chaque règle (cf. algorithme 4.3). Les balises étant ajoutées à l'intérieur des autres balises déjà présentes, l'ordre d'application de ces règles peut avoir son importance, car lors de la procédure de passage au format *XML* (cf. section 4.2.5.2), ce sont les balises les plus extérieures qui sont conservées et remplacées :

<PATRONYME> <PRÉNOM> Ricardo/PRENOM </PRÉNOM> <PATRONYME>
Bofill/NOM-PATRONYMIQUE/NP </PATRONYME> </PATRONYME>
→ **<NP Catégorie=Patronyme Classe=Anthroponyme> Ricardo Bofill </NP>**

Nous pouvons donc affecter des niveaux de priorité à nos règles, en jouant sur l'ordre de leur application. Nous avons alors choisi :

1. De favoriser les règles balisant des entités nommées à l'aide d'un contexte lexico-sémantique plutôt que celles qui ne font intervenir qu'une étiquette correspondant directement à une catégorie d'entité nommée.
2. De privilégier une catégorie par rapport à une autre en cas d'ambiguïté liée à l'absence d'un contexte catégorisant. Nous avons par exemple choisi de résoudre les ambiguïtés prénom/ville par le choix de la ville, simplement en plaçant la règle utilisant seul le lexique des noms de villes avant celle utilisant seul le lexique des prénoms. Cela signifie que si une entité nommée est étiquetée comme ville et prénom par les lexiques, elle sera catégorisée comme ville si elle ne rentre pas dans le schéma d'une règle plus complexe.

[9] Cela permet à la fois de catégoriser le sigle et sa forme étendue, mais aussi de ne traiter que ceux référant à des entités nommées.

```
Pour ( chaque ligne du corpus à traiter ) Faire
    Pour ( chaque règle ) Faire
        Tant que ( la règle peut s'appliquer à la ligne ) Faire
            appliquer la règle sur une copie de la ligne
            supprimer de la ligne les formes sur lesquelles l'application s'est faite
        Fin Tant que
    Fin Pour
    ligne ← copie de la ligne
Fin Pour
```

Algorithme 4.3 – Application des règles

4.2.3 Apprentissage automatique et seconde reconnaissance

Pour améliorer les performances de notre système, nous avons mis au point une méthode d'apprentissage automatique basée sur des heuristiques[10], afin de créer de nouveaux lexiques. Contrairement à ceux de Poibeau [1999], ces lexiques sont obtenus automatiquement. Cependant, nous n'utilisons pas directement les informations fournies par le module d'apprentissage pour enrichir nos lexiques de base, car nous ne voulons pas risquer de bruiter ces derniers, mais cela pourra être fait après une validation manuelle des unités lexicales des nouveaux lexiques (apprentissage supervisé). La phase d'apprentissage s'effectue en réalité lors de la première reconnaissance et les informations recueillies vont être utilisées lors de la seconde reconnaissance. En effet, certaines de nos règles permettent la mise à jour de lexiques en associant le nom du fichier temporaire du lexique et le regroupement de formes correspondant à l'unité lexicale à ajouter (cf. section 4.2.2.2). Ces unités lexicales peuvent être des entités nommées à identifier et à catégoriser directement, mais aussi de nouveaux mots déclencheurs.

Nemesis comporte 17 heuristiques issues de règles portant sur les anthroponymes, 13 de règles sur les toponymes, 7 de règles sur les ergonymes, 5 de règles sur les praxonymes, et 2 de règles sur les phénonymes. Voici quelques exemples d'heuristiques permettant cet apprentissage, où C représente une forme candidate au statut de nom propre :

- soit le schéma $C_1\ C_2$, si l'entité $C_1\ C_2$ est catégorisée en tant que patronyme avec C_2 un nom patronymique inconnu, alors C_2 est ajoutée au lexique des noms patronymiques (p. ex. *Lang* dans *Jack Lang*) ;
- prenons le schéma $C_1\ C_2\ C_3$, où C_2 est un prénom, C_3 une forme quelconque commençant par une majuscule et C_1 possède un des suffixes caractéristiques des adjectifs de nationalité (*ois*, *ais*, *and*, etc.). C_1 est alors ajoutée au lexique des ethnonymes (p. ex. *Marseillais* à partir de *Marseillais Robin Huc*) ;
- le plus souvent, lorsqu'un sigle apparaît dans un texte, il est lié à sa forme étendue lors du prétraitement lexical (cf. section 4.2.1). Lorsque la première forme de la forme étendue s'avère être un mot clé pour les noms d'organisations, le sigle ainsi que sa forme étendue sont ajoutés au lexique des noms d'organisations (p. ex. *FFF* et *Fédération française de football*).

Durant la seconde phase de reconnaissance des entités nommées, ces lexiques sont de nouveau projetés sur le corpus (cf. section 4.2.2.1). La nouvelle information apportée par cette projection va être utilisée :

[10] Ces heuristiques permettent également le résolution de certaines coréférences (cf. section 4.2.4).

1. En appliquant de nouveau les règles de la première passe mettant en jeux ces lexiques.
2. En créant de nouvelles règles se servant des lexiques qui n'existaient pas préalablement.

Dans tous les cas, les règles déclenchées lors de cette seconde phase de reconnaissance sont prioritaires sur les règles déclenchées précédemment, de telle sorte que **Nemesis** effectue un mécanisme de révision sur les entités nommées déjà étiquetées.

Nous avons choisi de n'effectuer que deux passes et non une multiplicité de réitération du processus jusqu'à obtention de performances acceptables. En effet, nous avons pu remarquer que le gain d'une seconde passe était important (cf. section 5.1), mais que l'apport d'autres passes n'auraient pas un impact aussi intéressant, surtout par rapport au coût d'une telle opération.

4.2.4 Gestion des coréférences

Durant l'exécution des deux phases de reconnaissance des entités nommées, les formes coréférant à une même entité nommée se voient attribuer un même identifiant dans un tableau d'associations. Pour une gestion complète des coréférences, il faut tenir compte des nombreuses variations qui peuvent opérer sur les entités nommées : graphiques (p. ex. *Parti Socialiste* → *Parti socialiste*), morpho-syntaxiques (p. ex. *les habitants de Nantes* → *les Nantais*), les sigles (p. ex. *École Normale Supérieur* → *ENS*), les coordinations (p. ex. *le Grand et le Petit Palais*), les ellipses (p. ex. *École Normale Supérieur* → *Normale sup*) et les métaphores (p. ex. *l'Everest* → *le toit du monde*). Il s'agit d'une problématique très délicate et nous ne prétendons pas la résoudre ici. Malgré tout, **Nemesis** prend en compte certaines de ces variations :

- les sigles et leurs forme étendue associée, ainsi que leurs autres occurrences se voient assignés le même identifiant ;
- les différentes façons de désigner une personne sont également reliées. En effet, un certain nombre d'heuristiques de la phase d'apprentissage (cf. section 4.2.3) permettent d'identifier le nom de famille de cette personne, ce qui nous permet d'en retrouver toutes les occurrences. Malheureusement, si un même texte parle de plusieurs personnes d'une même famille, elles seront regroupées sous le même identifiant ;
- d'une façon plus générale, nous traitons un certain nombre de variations graphiques et d'ellipses en mémorisant les entités nommées reconnues et leur catégorie. Ainsi, si deux entités nommées ont la même catégorie et que l'une est sous-chaîne de l'autre ou diffère simplement par sa graphie, elles sont associées.

4.2.5 Post-traitements

À la suite des quatre module principaux composant **Nemesis**, nous appliquons deux post-traitements. Le premier vise à éliminer, parmi les candidats au statut d'entité nommée se trouvant en début de phrase, ceux qui sont en réalité des noms communs. Le second post-traitement constitue un module de nettoyage du texte et de présentation des résultats sous la forme d'un fichier *XML*.

4.2.5.1 Traitement des entités nommées en début de phrase

Une fois les phases de reconnaissance achevées, il demeure de nombreuses entités auxquelles nous avons attribué, à tort, le statut d'entité nommée, comme pour *Elle* dans l'exemple suivant :

<PRENOM> Elle/PRENOM </PRENOM> fabriquera , dans un
premier temps , le produit/CLE-MARQUE liquide qui entre dans le processus des photocopies ainsi que des pièces détachées pour la filiale/CLE-ENTREPRISE de <MARQUE> Minolta/MARQUE </MARQUE>
en <PAYS> RFA/PAYS </PAYS> _._

Pour pallier ce problème, nous avons mis en place un post-traitement fondé sur un anti-dictionnaire. Cet anti-dictionnaire regroupe les unités lexicales pouvant se trouver en début de phrase, mais n'étant pas des entités nommées, comme les entités fonctionnelles, les pronoms, les adverbes, etc. Ainsi, chaque entité nommée supposée, trouvée en début de phrase, est comparée aux éléments de l'anti-dictionnaire : si un de ces éléments correspond à cette entité, les balises sont retirées.

4.2.5.2 Nettoyage et présentation au format *XML*

Conséquemment aux différents traitements permettant l'identification et la catégorisation des entités nommées, le texte brut originel est accompagné des étiquettes correspondant à nos lexiques et des balises indiquant la catégorie référentielle des entités qu'elles entourent :

Le/NP <ETHNONYME> <ETHNONYME> <ETHNONYME> Français/NOM-NATIONALITE
</ETHNONYME> </ETHNONYME> </ETHNONYME> <PATRONYME> <PATRONYME>
<PATRONYME> <PRENOM> Didier/PRENOM </PRENOM> <PATRONYME> Auriol/NOM/VILLE
</PATRONYME> </PATRONYME> </PATRONYME> </PATRONYME> (<MARQUE>
Lancia/MARQUE </MARQUE> intégrale 16 S/NP) a pris la tête du <EVENEMENT>
rallye/CLE-EVENEMENT de <VILLE> Monte-Carlo/VILLE </VILLE> </EVENEMENT> à l' issue de
l' étape de classement (560 km dont six spéciales chronométrées de 124 km) disputée dimanche
21 janvier entre <VILLE> Monaco/VILLE </VILLE> et <VILLE> Aubenas/VILLE </VILLE> _._

Le premier nettoyage que nous effectuons consiste à retirer les étiquettes lexico-sémantiques posées lors de la projection des lexiques :

Le <ETHNONYME> <ETHNONYME> <ETHNONYME> Français </ETHNONYME>
</ETHNONYME> </ETHNONYME> <PATRONYME> <PATRONYME> <PATRONYME>
<PRENOM> Didier </PRENOM> <PATRONYME> Auriol </PATRONYME> </PATRONYME>
</PATRONYME> </PATRONYME> (<MARQUE> Lancia </MARQUE> intégrale 16 S) a pris la
tête du <EVENEMENT> rallye de <VILLE> Monte-Carlo </VILLE> </EVENEMENT> à l' issue de l'
étape de classement (560 km dont six spéciales chronométrées de 124 km) disputée dimanche 21 janvier
entre <VILLE> Monaco </VILLE> et <VILLE> Aubenas </VILLE> _._

Certaines entités nommées possèdent plusieurs balises en raison de leur polysémie ou simplement parce qu'elles ont été balisées par le déclenchement de plusieurs règles différentes concluant à la même catégorie. C'est le cas des entités *Vosges* et *Ricardo Bofill* dans les phrases suivantes :

L' usine , qui devrait être implantée à <VILLE> Eloyes </VILLE> (<TOPONYME_M> <ORONYME>
Vosges </ORONYME> </TOPONYME_M>) représente un investissement d' environ 3,7 milliards de
yens (148 milliards de francs) _._
[...]
La somptueuse maquette de <PATRONYME> <PRENOM> Ricardo </PRENOM> <PATRONYME>
Bofill </PATRONYME> </PATRONYME> exposée auparavant dans le grand hall de la mairie a été
rangée au placard.

Dans ces cas là, le filtrage consiste à ne retenir que la balise la plus extérieure pour la catégorisation :

L' usine , qui devrait être implantée à <VILLE> Eloyes </VILLE> (<TOPONYME_M>
Vosges </TOPONYME_M>)
représente un investissement d' environ 3,7 milliards de yens (148 milliards de francs) _._
[...]
La somptueuse maquette de <PATRONYME> Ricardo Bofill </PATRONYME>
exposée auparavant dans le grand hall de la mairie a été rangée au placard.

Si les effets de la segmentation en occurrence de phrases ne sont gênants ni pour des traitements ultérieurs, ni pour la lisibilité du texte, en revanche la segmentation en occurrences de formes pose ce type de problème. Il nous faut donc en annuler les effets, simplement en appliquant une procédure inverse à cette segmentation.

Une fois ces différents nettoyages effectués, nous remplaçons les balises par des balises *XML* équivalentes qui comportent les attributs `Classe` et `Catégorie` de l'entité nommée. Outre ces deux attributs indiquant la nature de l'entité nommée, un troisième est ajouté qui indique l'identifiant de celle-là. Cet identifiant est obtenu par l'extraction des informations contenues dans le tableau d'associations alimenté durant tout le processus de reconnaissance des entités nommées. Le fichier *XML* final, résultant de l'exécution de **Nemesis**, a la forme suivante :

L'usine, qui devrait être implantée à <NP Catégorie=Ville Classe=Toponyme Id=8> Eloyes </NP> (<NP Catégorie=Toponyme_moyen Classe=Toponyme Id=9> Vosges </NP>) représente un investissement d'environ 3,7 milliards de yens (148 milliards de francs).
Elle fabriquera, dans un premier temps, le produit liquide qui entre dans le processus des photocopies ainsi que des pièces détachées pour la filiale de <NP Catégorie=Marque_ou_produit Classe=Ergonyme Id=10> Minolta </NP> en <NP Catégorie=Pays Classe=Toponyme Id=11> RFA </NP>.
[...]
La somptueuse maquette de <NP Catégorie=Patronyme Classe=Anthroponyme Id=18> Ricardo Bofill </NP> exposée auparavant dans le grand hall de la mairie a été rangée au placard.
[...]
Le <NP Catégorie=Ethnonyme Classe=Anthroponyme Id=124> Français </NP> <NP Catégorie=Patronyme Classe=Anthroponyme Id=177> Didier Auriol </NP> (<NP Catégorie=Marque_ou_produit Classe=Ergonyme Id=178> Lancia </NP> intégrale 16 S) a pris la tête du <NP Catégorie=Évènement Classe=Praxonyme Id=179> rallye de Monte-Carlo </NP> à l'issue de l'étape de classement (560 km dont six spéciales chronométrées de 124 km) disputée dimanche 21 janvier entre <NP Catégorie=Ville Classe=Toponyme Id=180> Monaco </NP> et <NP Catégorie=Ville Classe=Toponyme Id=181> Aubenas </NP>.
[...]
<NP Catégorie=Patronyme Classe=Anthroponyme Id=209> Bruno Saby </NP> dispose d'une intégrale 16 S sortie des mêmes ateliers que celles d'<NP Catégorie=Patronyme Classe=Anthroponyme Id=177> Auriol </NP> et de <NP Catégorie=Patronyme Classe=Anthroponyme Id=185> Biasion </NP>, mais il doit, en revanche, assurer son assistance avec sa propre équipe.
[...]
Les deux premières journées du quatorzième congrès (extraordinaire) de la <NP Catégorie=Organisation Classe=Anthroponyme Id=6> Ligue des communistes de Yougoslavie </NP> (<NP Catégorie=Organisation Classe=Anthroponyme Id=6> LCY </NP>) ont illustré l'état de délabrement dans lequel se trouve le parti depuis déjà quelques années.
Dans un discours-fleuve qui constituait le plus petit dénominateur commun des positions respectives des partis des six républiques et des deux provinces autonomes, <NP Catégorie=Patronyme Classe=Anthroponyme Id=218> Mr Milan Pancevski </NP> s'est prononcé pour la liberté d'association politique (et donc l'abandon du monopole de la <NP Catégorie=Organisation Classe=Anthroponyme Id=6> Ligue </NP>), pour la réforme du système économique et politique, ainsi que du fonctionnement de la <NP Catégorie=Organisation Classe=Anthroponyme Id=6> LCY </NP>.
[...]

4.3 Conclusion

Nous avons réalisé dans un premier temps le « noyau » de **Nemesis**, un système de reconnaissance incrémentielle des entités nommées du français. Ce « noyau » est fondé sur des méthodes linguistiques (lexiques et règles élaborées manuellement) utilisant l'évidence interne et dans une moindre mesure l'évidence externe. Il s'applique à des textes bruts et ne nécessite aucun étiquetage linguistique.

Dans un second temps, nous avons adjoint à ce « noyau » un module d'apprentissage basé sur des heuristiques qui effectue une mise à jour des lexiques, avant de procéder à une seconde phase de reconnaissance des entités nommées.

Finalement, **Nemesis** est un système incrémentielle fondé sur des méthodes mixtes – linguistiques et à base d'apprentissage – qui réalise une identification des entités nommées du français et leur reconnaissance selon une typologie la plus fine et la plus exhaustive possible. Il effectue également un traitement de certaines coréférences et présente les résultats sous la forme d'un

fichier *XML*.

Dans le chapitre suivant, nous évaluons tout d'abord les performances de **Nemesis** et étudions l'apport des différents modules dans le processus de reconnaissance des entités nommées.

Ensuite, nous proposons quelques améliorations de **Nemesis** face à l'incomplétude des lexiques et au passage à de nouveaux corpus, reposant sur une plus large utilisation de l'évidence externe, l'apprentissage de règles de réécritures et la recherche de nouveaux contextes via le *Web*.

Chapitre 5

Évaluation et améliorations

5.1 Évaluation de Nemesis

Nemesis est un système d'identification et de catégorisation qui effectue un traitement séquentiel immédiat des données par l'application, en cascade, de différents modules. Nous présentons maintenant une évaluation des performances globales de **Nemesis** sur l'ensemble des 27 catégories d'entités nommées, ainsi qu'une évaluation de l'apport de chaque module sur la classe des anthroponymes qui regroupe 73,8 % des entités nommées de notre échantillon du corpus *Le Monde* et 52 % de celles de l'échantillon du corpus *La Recherche* (cf. 3.3). De plus, les différentes catégories de la classe des anthroponymes présentent des caractéristiques morphologiques et lexico-sémantiques différentes, et donc des difficultés variées. En effet, si les patronymes, les prénoms et les ethnonymes sont relativement faciles à identifier et à catégoriser à partir du moment où nous possédons les lexiques nécessaires, en revanche, les organisations, comme les ensembles artistiques, présentent de grandes difficultés quant à l'identification de leurs limites à droite, et les lexiques seront d'une faible utilité pour ce problème.

5.1.1 Méthodologie de l'évaluation

Afin de pouvoir évaluer autrement que manuellement et intuitivement nos résultats, il est important de mettre en place une procédure d'évaluation automatique.

5.1.1.1 Les Corpus

Pour appliquer cette procédure d'évaluation, il nous faut tout d'abord créer manuellement un fichier de référence (corpus de validation), afin de pouvoir le comparer au fichier résultant de l'application de **Nemesis**. Nous avons décidé d'effectuer cette comparaison au niveau des résultats finals au format *XML*. Comme cette phase est longue et fastidieuse, nous avons limité la taille des échantillons formant le corpus d'évaluation. Ce corpus d'évaluation est composé :

- du corpus de test utilisé lors de la réalisation de **Nemesis** (environ 7 000 mots issus du corpus *Le Monde*[1]) ;
- d'autres articles du corpus *Le Monde* (environ 6 400 mots) ;
- un article de la revue *Unasylva*[2] (Revue internationale des forêts et des industries forestières), pris sur le site internet de la FAO et portant sur les écosystèmes de montagne et leur mise en valeur (environ 7 000 mots) ;

[1] Corpus de textes *Le Monde* - année 1997 - European Corpus Initiative (ECI) distribué par ELRA.
[2] `http://www.fao.org/forestry/site/8572/fr`

- un article du *Monde Diplomatique*[3] portant sur les armes biologiques de la guerre de Corée (environ 2 600 mots) ;
- une page *Web* du site de la Direction générale des douanes et droits indirects[4] et portant sur la contrefaçon (environ 3 500 mots) ;

Nous avons donc réuni des échantillons de différentes natures formant un corpus de 26 580 mots et 1 284 entités nommées, puis nous y avons poser manuellement les balises *XML* correspondant aux catégories de chacune de ces entités nommées.

5.1.1.2 Procédure d'évaluation

Après avoir créé ce fichier de validation, il faut mettre en place une procédure de comparaison : elle consiste à prendre chaque ligne de chacun des deux fichiers (corpus de validation et fichier contenant les résultats de l'application de **Nemesis**) et de les comparer. Ici quatre possibilités et donc quatre types de résultats sont à envisager :

1. Une entité nommée est correctement identifiée et catégorisée → écrire dans le fichier « résultats » : une indication qui signifie le succès, puis l'entité nommée elle-même.
2. Une entité nommée est correctement identifiée mais mal catégorisée → écrire dans le fichier « résultats » : la catégorie attendue et l'entité nommée, puis la catégorie attribuée automatiquement.
3. Une entité nommée est mal identifiée (trop ou pas assez d'éléments rattachés) → écrire dans le fichier « résultats » : la catégorie, l'entité nommée attendue, puis l'élément identifié.
4. Un nom commun est identifié comme étant une entité nommée → écrire dans le fichier « résultats » : l'élément identifié et sa catégorie.

Dans chacun des cas où l'entité est mal identifiée ou mal catégorisée, la catégorie attendue est écrite en premier, afin de pouvoir trier le fichier selon celle-ci, et ainsi permettre une meilleur visualisation des problèmes posés par chaque catégorie.

Voici quelques exemples de lignes présentes dans le fichier « résultats » :

- `Ok! : Ricardo Bofill` → *Ricardo Bofill* a été correctement identifiée et catégorisée ;
- `Catégorie=Évènement Classe=praxonyme : Camel Trophy -> Catégorie=Patronyme Classe=Anthroponyme` → l'entité nommée *Caml Trophy* a été correctement identifiée, mais catégorisée en tant que patronyme et non évènement ;
- `Catégorie=Marque_ou_produit Classe=Ergonyme : Lancia intégrale 16 S -> Lancia` → l'entité nommée *Lancia intégrale 16 S* a été correctement catégorisée, mais insuffisamment identifiée ;
- `RIEN -> Vue : Catégorie=Ville Classe=Toponyme` → le nom commun *Vue* a été identifié et catégorisée à tort comme ville.

5.1.2 Évaluation de l'apport de chaque module

Cette évaluation porte uniquement sur les sept catégories de la classe des anthroponymes et a été réalisée sur notre corpus de test (environ 7 000 mots issus du corpus *Le Monde*). Pour évaluer l'impact de chaque partie de **Nemesis**, nous avons procédé à une évaluation en quatre parties : nous avons commencé par évaluer le système avec le traitement le plus simple, puis

[3] `http://www.monde-diplomatique.fr`
[4] `http://www.douane.minefi.gouv.fr`

nous l'avons fait après l'ajout de chaque module supplémentaire (prétraitement sur les sigles et seconde passe) et enfin après quelques améliorations apportées à la première passe.

5.1.2.1 Évaluation du système avec découpage du texte et première reconnaissance

Dans un premier temps, nous nous attardons sur les résultats qui mettent en jeu un système composé uniquement de la projection de lexiques de bases et l'utilisation de règles lexico-sémantiques, graphiques et morphologiques de réécriture portant sur la structure interne des entités nommées (cf. tableau 5.1).

Tableau 5.1 – Évaluation du système de base

	Rappel 56,8 %	**Précision 87,8 %**
EN correctement identifiées et catégorisées		137
EN identifiées mais mal catégorisées		2
EN mal identifiées		102
Noms communs identifiées comme EN		2

Il faut tout d'abord noter que, sur les 241 anthroponymes présents dans cet échantillon, 137 ont été correctement identifiés et catégorisés par **Nemesis**, soit un taux de rappel de 56,8 %. Parmi les 104 entités nommées restantes, 15 ont été insuffisamment identifiées (en général, le rattachement à droite de certaines formes n'est pas opéré) et 87 ne l'ont pas du tout été (pauvreté des lexiques, entités nommées dont le premier mot ne commence pas par une majuscule, etc.). Enfin, l'entité nommée *Hermès Parfums* apparaît deux fois en tant que patronyme, alors que c'est en fait un nom de marque. D'autre part, l'unité lexicale *Elle* apparaît deux fois en tant que prénom, alors qu'il s'agit du pronom personnel. Ce type d'erreur sera corrigé en mettant en place, juste avant de passer le corpus sous un format *XML*, un post-traitement basé sur un anti-dictionnaire regroupant les unités lexicales pouvant se trouver en début de phrase, mais n'étant pas des entités nommées, comme les entités fonctionnelles, les pronoms, les adverbes, etc.

Le taux de rappel dépassant les 50 % est plutôt encourageant, à la vue de la simplicité des méthodes mises en œuvre. Il pourra nettement être amélioré par la suite. Quant au taux de précision, il est naturellement élevé (87,8 %), car nous avons fait le choix de prendre des lexiques et des règles qui induisent un minimum de bruit. En effet, il nous paraît plus facile d'étendre ces règles et ce lexiques pour augmenter le taux de rappel, que de les réduire en filtrant ceux qui produisent des résultats erronés.

5.1.2.2 Évaluation du système avec prétraitement et première passe

Le prétraitement associant les sigles et leur forme étendue, ainsi que le post-traitement à base d'anti-dictionnaire (cf. tableau 5.2) ont été naturellement fait augmenter les taux de rappel et de précision – respectivement +5 % et +5,3 %.

Le prétraitement sur les sigles et leurs formes étendues procure un gain intéressant sur l'identification et la catégorisation des entités nommées. En effet, sur notre échantillon de corpus de validation, il nous permet d'identifier et de catégoriser six sigles et six entités nommées mixtes lors du premier passage, sans induire le moindre bruit. D'une part, ce sont des entités nommées

Tableau 5.2 – Évaluation du système enrichi du prétraitement sur les sigles

	Rappel 61,8 %	**Précision 93,1 %**
EN correctement identifiées et catégorisées		149
EN identifiées mais mal catégorisées		2
EN mal identifiées		90
Noms communs identifiées comme EN		0

qu'il aurait été difficile de traiter correctement (cf. section 3.2.2), d'autre part, le gain sera augmenter lors de la deuxième passe, car les sigles et leur forme étendue sont de nouveau présents dans le reste du texte. L'utilisation d'un anti-dictionnaire a permis l'élimination des deux noms communs identifiées comme entités nommées

5.1.2.3 Évaluation du système avec prétraitement, première et seconde passe

La dernière étape du système réside dans l'exécution de la seconde passe. Nous allons donc étudier l'apport que celle-ci procure (cf. tableau 5.3), et ainsi faire une évaluation globale de notre système.

Les nouveaux lexiques induits lors de la première passe sont au nombre de trois pour les anthroponymes : les noms patronymiques, les noms de nationalité et les noms d'organisations (sigles et formes étendus).

Par conséquent, les nouvelles entités nommées identifiées et catégorisées par la deuxième passe seront essentiellement des patronymes, des ethnonymes et des organisations. Pour notre corpus de test, 19 sont des patronymes et trois des organisations. Cette grande majorité vient du fait que le nom patronymique est souvent retrouvé seul après l'avoir été avec un prénom : il est donc reconnu avec le prénom à la première passe, puis seul à la seconde.

Tableau 5.3 – Évaluation du système avec seconde passe

	Rappel 71 %	**Précision 94 %**
EN correctement identifiées et catégorisées		171
EN identifiées mais mal catégorisées		2
EN mal identifiées		68
Noms communs identifiées comme EN		0

5.1.2.4 Évaluation finale

Une fois ces différents modules mis en place et leur impact étudié, nous avons amélioré la première reconnaissance des entités nommées par l'adjonction de quatre lexiques, quelques nouvelles règles et le remplacement du lexiques des prénoms. Les nouveaux lexiques sont des lexiques de partis politiques, de noms de médias (télévision, presse, etc.) ou encore des noms

d'institutions françaises et internationales. Ces quelques ajouts nous apporte un gain important, notamment les trois lexiques précédant. L'apport de ces modifications est détaillé au tableau 5.4.

Tableau 5.4 – Évaluation finale

	Rappel 85,9 %	**Précision 95 %**
EN correctement identifiées et catégorisées		207
EN identifiées mais mal catégorisées		2
EN mal identifiées		34
Noms communs identifiées comme EN		0

Bien qu'ayant été obtenus sur notre corpus de test et limités aux différentes catégories d'anthroponymes, ces résultats sont intéressants et permettent de juger de l'apport des différents traitements. Cependant, il nous faut maintenant effectuer une évaluation générale de **Nemesis** portant sur toutes les catégories d'entités nommées et sur des données plus vastes, plus diverses et différentes de celles sur lesquelles nous nous sommes basé pour réaliser **Nemesis**.

5.1.3 Évaluation générale de Nemesis

Cette évaluation, présentée au tableau 5.5, a été réalisée sur un corpus composé d'un total d'environ 26 600 mots pour 1 284 entités nommées (cf. corpus d'évaluation, section 5.1.1.1).

Tableau 5.5 – Résultats de l'évaluation générale de **Nemesis**

	EN correctement identifiées et catégorisées	EN identifiées mais mal catégorisées	EN mal identifiées	EN non reconnues
Anthroponymes	509	8	24	94
Toponymes	402	18	9	53
Ergonymes	84	18	8	29
Praxonymes	20	3	2	3
Total	**1015**	**47**	**43**	**179**

	Taux de rappel	Taux de précision
Anthroponymes	80,2 %	93,6 %
Toponymes	83,4 %	92,6 %
Ergonymes	60,4 %	74,3 %
Praxonymes	71,4 %	80 %
Total	79 %	91 %

Nous pouvons remarquer que la classe des phénonymes n'est pas représentée dans ce corpus d'évaluation. En effet les entités nommées de cette classe (phénomènes naturels, astres et comètes) sont en général peu fréquentes (cf. tableau 3.3). D'autre part, seuls onze noms communs ont été étiquetés, à tort, comme entités nommées.

Comparativement aux résultats obtenus sur notre corpus de test (cf. tableau 5.4), la reconnaissance des anthroponymes et des toponymes est moins efficace sur ce corpus d'évaluation. Cette perte de performance est naturelle du fait que **Nemesis** a été modifié au vu des résultats obtenus sur le corpus de test. Par conséquent, les résultats sur ce dernier sont inévitablement flatteurs. D'autre part, la nature même de certains des textes qui composent notre corpus d'évaluation (les pages *Web*) perturbe la reconnaissance des entités nommées. En effet, ces textes sont rédigés avec des conventions moins strictes (règles typographiques, structure, etc.) qui dégradent les traitements :

- les largesses dans la formation des sigles altèrent leur identification et l'association avec leur forme étendue ;
- la présence de titre comportant des entités nommées dépourvues de contexte empêche l'utilisation de l'évidence externe ;
- les portions de textes écrites entièrement en majuscules rendent inopérante la projection de nos lexiques ;
- etc.

Malgré tout, cette baisse des performances reste limitée (-1,9 % en précision et -4,3 % en rappel).

Les résultats pour ces deux classes sont bien meilleurs que ceux obtenus sur les ergonymes et, dans une moindre mesure, les praxonymes. Il existe plusieurs explications à ce phénomène :

1. Les anthroponymes et les toponymes regroupent toutes les entités nommées des classes MUC (personnes, lieux, organisations), qui ont été largement étudiées. Par conséquent, les lexiques développés pour celles-là sont plus nombreux et de meilleure qualité. De plus, de nombreuses grammaires ont été crées pour extraire ces entités nommées [Trouilleux, 1997 ; Friburger, 2002 ; Poibeau, 2002].

2. Les ergonymes sont probablement les entités nommées les plus difficiles à reconnaître, ce qui explique le déficit sur les taux de leur reconnaissance (environ -21 % pour la précision et -19 % pour le rappel par rapport aux anthroponymes et aux toponymes). En effet, si les établissements d'enseignement et de recherche peuvent être reconnus grâce à leur évidence interne – tout comme les patronymes, les organisations et les ensembles artistiques, qui composent la très grande majorité des anthroponymes – il n'en est rien pour les entreprises industrielles, les marques, les produits et les œuvres matérielles. De plus, il n'existe pas de lexiques suffisamment larges et fiables pour ces catégories d'entités nommées, à l'inverse de la plupart des toponymes. Cela est notamment du au fait que ces catégories sont très ouvertes – beaucoup plus que les toponymes – et que toute entité nommée pourrait en faire partie (p. ex. pour les marques et les produits ou encore les noms d'œuvres matérielles). Enfin, leurs contextes gauche et droit immédiats ne permettent pas, dans la plupart des cas, la reconnaissance de ces entités nommées, alors que les toponymes qui ne figurent pas dans nos lexiques sont souvent accompagnés d'un contexte immédiat permettant leur catégorisation.

3. Parmi les praxonymes, la catégories des périodes historiques est légèrement à part, car il est possible d'en construire un lexique relativement exhaustif, du moins à un instant donné. Un tel lexique est d'ailleurs indispensable, car le contexte dans lequel apparaissent ces entités nommées est très peu catégorisant. Concernant les autres entités nommées de cette classe (les évènements, les faits historiques et les œuvres abstraites), la création de pareils lexiques n'est pas envisageable, pour les mêmes raisons que celles précédemment évoquées pour les ergonymes. Malgré tout, les résultats obtenus sur cette classe sont moins bons que ceux

obtenus sur les anthroponymes et les toponymes (environ -10 % pour le rappel et -13 % pour la précision). Cela est dû à un certain nombres d'évènements qui sont suffisamment connus pour apparaître sans aucun élément de contexte permettant leur catégorisation (*le Monte-Carlo* à trois reprises, *le RAC*, *Roland-Garros*). Comme nous n'utilisons pas de lexiques contenant des noms d'évènements, nous ne pouvons pas obtenir leur reconnaissance.

Nous obtenons donc, sur l'ensemble de nos textes et des entités nommées, des taux de reconnaissance qui avoisinent 80 % pour le rappel et 90 % pour la précision. Ces résultats sont satisfaisants, mais nous ne saurions nous en satisfaire. C'est pourquoi nous avons choisi d'étudier des voies d'amélioration de **Nemesis** à deux niveaux différents :

1. En améliorant les traitements existant par une étude des conflits qui peuvent survenir entre les règles et des possibilités d'inférence de nouvelles règles.
2. En effectuant de nouveaux traitements comme la gestion des sur-compositions référentielles et l'analyse des structures énumératives).

5.2 Vers un module de désambiguïsation et d'apprentissage de règles

Les problèmes de conflits engendrés par la reconnaissance des entités nommées sont récurrents dans les systèmes que nous avons rencontrés. Poibeau [1999], s'il définit trois heuristiques pour l'application de ses règles de réécriture, ne gère pas la résolution des conflits. Trouilleux [1997] lui gère, *a priori*, quelques cas d'ambiguïté (noms de lieux qui peuvent entrer dans la composition de patronymes ou d'organisations) et de sur-compositions (pour les organisations). Friburger [2002] définit, *a priori*, un ordre à l'application de ses transducteurs, afin de traiter le cas de l'ambiguïté de sur-composition entre les noms de personnes et les noms d'organisations.

À partir des règles lexico-sémantiques de réécritures élaborées, nous avons étudié les problèmes que leur application peut engendrer : au niveau conceptuel tout d'abord (sur les patrons de règles), puis, au niveau expérimental (sur les entités nommées reconnues). Cette étude ayant été réalisée à une période où **Nemesis** n'était pas abouti, elle ne porte que sur les classes des anthroponymes et des toponymes et se limite à 53 règles de réécritures.

5.2.1 Étude conceptuelle

Nous avons analysé l'ensemble de nos règles, afin d'identifier les différentes façons dont elles pouvaient entrer en conflit et comment réduire *a priori* la réalisation de ces conflits sur les entités nommées reconnues (ambiguïtés sur l'identification ou la catégorisation).

Prenons comme exemple les deux règles suivantes :

$$[w_1^1, w_1^{n_1}] \rightarrow catégorie_1 \tag{5.1}$$

$$[w_2^1, w_2^{n_2}] \rightarrow catégorie_2 \tag{5.2}$$

5.2.1.1 Chevauchement de patrons

Définition 1. *Il y a un conflit de chevauchement de patrons de la règle 5.2 sur la règle 5.1* $\Longleftrightarrow$ $\exists\ i,j \in \mathbb{N}^2$ *(*$1 < i \leq n_1$ *et* $1 \leq j < n_2$*) tels que* $[w_1^i, w_1^{n_1}] = [w_2^1, w_2^j]$.

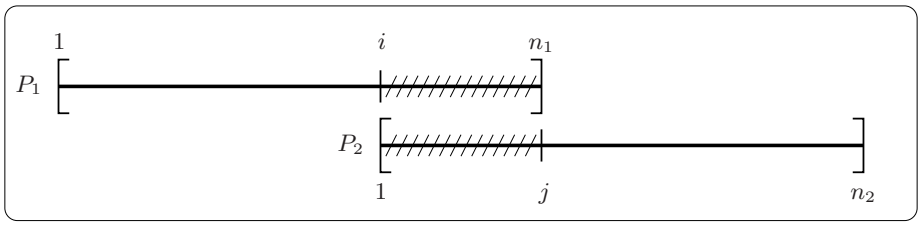


Figure 5.1 – Chevauchement de la règle 5.2 sur la règle 5.1

Les conflits de ce type (cf. figure 5.1) apparaissent à 43 occasions (moins de 1,6 % des conflits possibles par calcul combinatoire) et concernent 26 règles différentes (sur 53 règles). Pour chacune de ces 43 occurrences, le chevauchement n'a lieu que sur un seul élément des patrons (cas où $i = n_1$ et $j = 1$) qui est de type `mot déclencheur` ou `EN` pour la règle 5.2 et `fin d'EN` ou `EN` pour la règle 5.1. Par conséquent, les conflits de chevauchement de patrons n'ont que peu de risques de se concrétiser sur les entités nommées reconnues lors de l'application des règles de réécriture. En effet, il paraît peu probable de rencontrer en corpus une suite de formes qui corresponde à la suite d'éléments de patron $(w_1^1, \ldots, w_1^{n_1} = w_2^1, \ldots, w_2^{n_2})$: ${}_1(\textit{Europe de l'}\ {}_2(\textit{Est})_1\ \textit{de l'Afrique})_2$ ou ${}_1(\textit{XV de}\ {}_2(\textit{France})_1\ \textit{méridionale})_2$.

Au niveau de la conception de ces règles, nous ne gérons pas encore la résolution de conflit, mais pour limiter leur apparition nous réordonnons les règles en terme de priorité : lorsqu'un conflit de chevauchement de patrons est identifié entre la règle 5.1 et la règle 5.2, nous choisirons d'appliquer en priorité la règle qui possède le plus grand nombre d'éléments de types `contexte` et `mot déclencheur`, car ils en font *a priori* la règle la plus fiable[5] (cf. section 4.2.2.2). Poibeau [1999] donne la priorité aux règles les plus longues, sans se soucier des éléments qui les composent, et, dans le cas de deux règles de même longueur, produit un résultat aléatoire.

5.2.1.2 Inclusion de patrons

Définition 2. *Il y a un conflit d'inclusion de patrons de la règle 5.2 dans la règle 5.1* $\Longleftrightarrow$ $[w_2^1, w_2^{n_2}] \subseteq [w_1^1, w_1^{n_1}]$.

De tels conflits, dont les différents cas sont schématisés à la figure 5.2, se présentent à 35 reprises (moins de 1,3 % des conflits possibles) et concernent 28 règles différentes. Ils correspondent au cas où le patron d'une règle est totalement inclus dans celui d'une autre règle. Contrairement aux conflits de chevauchement, les conflits d'inclusion vont certainement se concrétiser sur les entités nommées reconnues lors de l'application des règles de réécriture sur corpus. En effet, si P_1 est associé à une série de formes, P_2, qui est inclus dans P_1, sera forcément associé à une série de formes incluse dans la première.

Là encore, la constatation de tels conflits peut nous permettre de réordonner les règles en terme de priorité : s'il existe un conflit d'inclusion de patrons de la règle 5.2 dans la règle 5.1, nous choisirons de privilégier la règle 5.1 sur la règle 5.2. En effet, comme cette règle a un patron qui possède plus d'éléments, notamment ceux de types `contexte` et `mot déclencheur`, la règle 5.1 est la plus spécifique et possède un niveau de fiabilité plus grand. Prenons les patrons ${}_2(\texttt{\$Prénom* \$NP+})_2$ et ${}_1(\texttt{\$Métier \$Adj\_nationalité? [\$Prénom* \$NP+]})_1$ qui correspon-

[5] Un poids plus fort pourrait également être donné aux lexiques utilisés pour une unique catégorie.

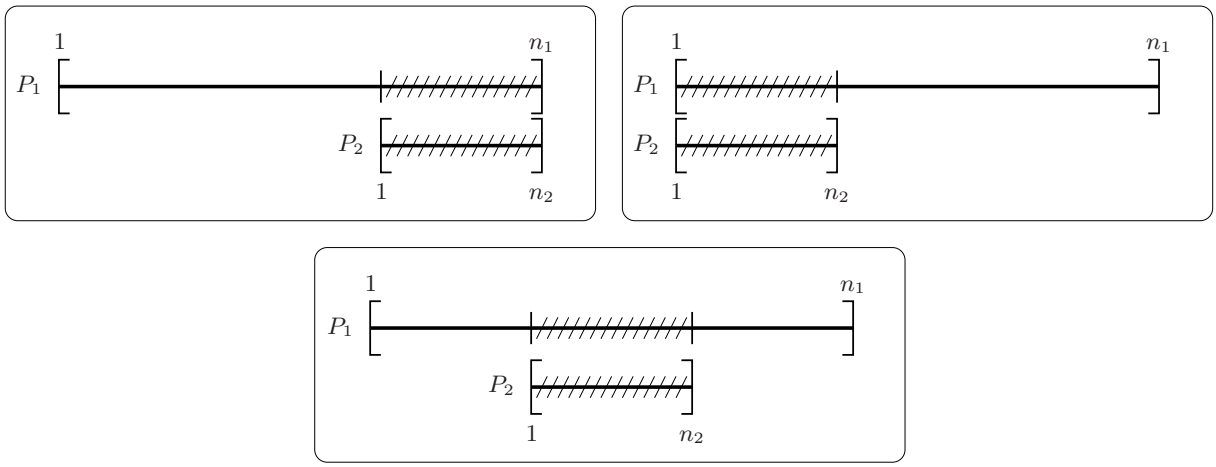


Figure 5.2 – Inclusion de la règle 5.2 dans la règle 5.1

draient, par exemple, à la suite de formes $_1$(*philosophe français* $_2$(*René Descartes*)$_2$)$_1$. Le patron de la règle 5.1 est le plus fiable du fait des éléments de type `contexte` (*philosophe* et *français*). Nous privilégions donc par défaut les entités nommées maximales, tout comme Friburger [2002].

5.2.2 Étude expérimentale

Cette étude porte sur les anthroponymes et les toponymes, dont l'ensemble des catégories représente plus de 80 % des entités nommées présentes dans nos corpus (cf. section 3.3.1). Elle a été réalisée sur un corpus composé des textes de notre corpus de test, plus les autres textes extraits du *Monde* (environ 13 400 mots, 250 anthroponymes et 80 toponymes). L'observation des conflits qui apparaissent lors de la reconnaissance des entités nommées en corpus nous permet de vérifier si nos règles sont correctement ordonnées. Elle peut éventuellement permettre la mise au point d'heuristiques pour la désambiguïsation des conflits. Parmi ces conflits, nous introduisons un nouveau type (accolement), qui n'existe pas au niveau des règles, mais s'observe lors de leur application sur le corpus.

5.2.2.1 Chevauchement

Nous n'avons relevé aucun conflit de chevauchement sur les anthroponymes et les toponymes reconnus. Si ce type de conflit apparaissait, il témoignerait très probablement d'un problème lors de la conception d'une, voire de deux règles. Dans ce cas, différentes solutions s'offriraient alors à nous :

1. Éliminer l'une des deux règles voire les deux.
2. Inverser la priorité sur ces deux règles.
3. Fusionner les deux règles.
4. Créer une troisième règle, prioritaire sur les deux premières.

5.2.2.2 Inclusion

Les conflits d'inclusion de patrons, eux, se concrétisent pendant la phase d'application des règles de réécriture sur un corpus : $_1$(*Université de* $_2$(*Nantes*)$_2$)$_1$, $_1$(*Guerre d'*$_2$(*Algérie*)$_2$)$_1$, $_1$(*Anatole* $_2$(*France*)$_2$)$_1$. Dans ce cas, nous gardons de préférence l'entité nommée maximale, car c'est la plus « fiable ». En effet, les formes qui la composent ont nécessité une correspondance avec un plus grand nombre d'éléments lexicaux de type `mot déclencheur` (*Université*, *Guerre*, *Anatole*).

Par conséquent, la présence de tels conflits parmi les entités nommées d'un corpus ne pose pas de problème d'ambiguïté, mais peut indiquer la présence de sur-compositions référentielles, comme dans *Université de Nantes* ou *Guerre d'Algérie*.

5.2.2.3 Accolement

Ce type de conflit n'a de sens qu'en regard des entités nommées reconnues lors de l'application des règles. Il correspond au cas où deux entités nommées sont identifiées, l'une se situant immédiatement à la suite de l'autre. Il ne s'agit pas réellement d'un conflit dans la mesure où les deux entités nommées sont distinctes. Cependant, il peut être intéressant de repérer ces cas, afin de voir si nous n'avons pas commis une erreur. Jusqu'à présent, dans nos corpus, ce cas n'apparaît que pour dix paires d'anthroponymes, soit 3,8 % des entités nommées de cette classe. Il s'agit à chaque fois d'un ethnonyme suivi immédiatement d'un patronyme comme dans *les* $_1$(*Français*)$_1$ $_2$(*Voltaire*)$_2$, *Rousseau et Baudelaire* ou *le* $_1$(*Brésilien*)$_1$ $_2$(*Ricardo Bofill*)$_2$. Dans cette circonstance, il n'y a pas d'erreur. Cependant, nous pourrions avoir affaire à une entité nommée plus complète dont nous aurions mal identifié les limites[6] : $_1$(*Fédération*)$_1$ $_2$(*Française de football*)$_2$ qui devrait être une seule entité nommée, $_1$(*Caisse Primaire d'*)$_1$ $_2$(*Assurance Maladie dans le rouge*)$_2$ qui devrait être regroupée en une seule entité nommée privée de la partie *dans le rouge*[7]. Dans de tels cas, il faudrait étudier la présence de chacune des deux entités nommées séparément dans le texte (cf. tableau 5.6). Les solutions à apporter peuvent être affinées en ne retenant pas systématiquement l'entité nommée la plus complète, mais éventuellement une sous-partie de celle-ci (comme pour satisfaire à l'exemple précédent).

Tableau 5.6 – Désambiguïsation de l'accolement d'entités nommées

		Présence de l'EN 1 seule	
		oui	**non**
Présence de l'EN 2 seule	**oui**	prendre les deux séparément	prendre l'EN 2 ou l'EN complète
	non	prendre l'EN 1 ou l'EN complète	prendre l'EN complète

La découverte de ce type de conflits et les solutions retenues pour les résoudre pourront naturellement nous conduire à modifier nos règles dans les mêmes mesures que pour les conflits de chevauchement de patrons (cf. section 5.2.2.1) : si l'entité nommée complète est retenue, nous pourrons éventuellement généraliser cette décision et fusionner les deux règles ou en créer une troisième prioritaire sur les deux autres.

[6] Il s'agit là d'une pure conjecture : ce n'est absolument pas le cas dans nos résultats, mais il se pourrait que cela survienne sur d'autres corpus ou avec d'autres règles.

[7] Ces exemples restent hypothétiques, car il n'est notamment pas possible à **Nemesis** de reconnaître une entité nommée terminant par un déterminant (*d'*).

5.2.3 Possibilités d'inférer de nouvelles règles

Dès lors que des situations de conflits sont identifiées entre les entités nommées reconnues après une première application des règles de réécriture, il nous est possible d'en inférer de nouvelles :

Par un processus automatique en créant une règle résultant de la fusion de deux règles. Nous introduisons un opérateur de fusion $\oplus$, inspiré de l'opérateur de fusion de règles morphologiques de Mikheev [1997] :

$$(1) \oplus (2) = (w_1^1, \ldots, w_1^i, w_2^j, \ldots, w_2^{n2}) \rightarrow \mathit{catégorie}_1$$

$$\text{avec les contraintes } (w_1^{i+1}, \ldots, w_1^{n1}) = (w_2^1, \ldots, w_2^{j-1}) \text{ ou } (i = n_1 \text{ et } j = 1)$$

Cet opérateur peut être utilisé dans le cas d'un chevauchement ou d'un accolement d'entités nommées dont nous ne retrouverions pas de présence individuellement.

Par un processus supervisé en pointant les situations de conflit et en proposant un certain nombre de règles qui pourraient permettre de les résoudre. Dans ce cas, différentes solutions pourraient être proposées pour chaque conflit et une généralisation de la solution pourrait être retenue par l'utilisateur.

Cette deuxième méthode paraît être la plus efficace et comporter le moins de risques, dans la mesure où nous ne sommes pas du tout sûrs des règles inférées par la première méthode et que la deuxième nous permet d'en proposer un plus grand nombre. De plus, l'apport de cette approche est de laisser une plus grande marge de manœuvre à l'utilisateur et de l'intégrer dans le processus de reconnaissance : il n'a plus l'impression d'avoir affaire à une boîte noire, il peut interagir avec le système.

Toutes les remarques faites en ce qui concerne la réévaluation des priorités sur les règles, suivant les formes de conflits qu'elles peuvent engendrer dans leur conception ou leur application, pourront nous permettre de dégager un algorithme d'insertion pour les nouvelles règles induites par un processus d'apprentissage.

5.3 Les limites de Nemesis

Bien que les résultats de l'évaluation de **Nemesis** soient intéressants, nous constatons que la cause de la grande majorité des mauvaises reconnaissances est double. La non reconnaissance d'une entité nommée est d'abord due à l'incomplétude de nos lexiques. En effet, sur 269 entités nommées mal reconnues, 222 ne sont pas reconnues ou mal identifiées, ce qui témoigne de l'absence de leur forme complète dans nos lexiques. Or, nous nous refusons à travailler à l'aide de grandes bases d'entités nommées – ce qui serait une solution palliative – pour différentes raisons :

- il n'est pas possible de créer – et d'utiliser – des listes exhaustives d'entités nommées, en témoignent les 1,5 million de prénoms pour les seuls États-Unis ;
- ils seraient immédiatement surannés. En effet, des organisations, des marques, etc. se créent en permanence, ce qui rend une liste non exhaustive dès qu'elle est créée ;
- toutes les variations devraient y figurer (p. ex. *Ligue des communistes de Yougoslavie*, *LCY*, *Ligue*, etc.) ;
- de tels lexiques ne sont ni nécessaires [Mikheev et coll., 1999] ni suffisants (p. ex. la surcomposition référentielle ou la polysémie ne sauraient être résolues avec des lexiques).

Dès lors qu'une entité nommée est absente de nos lexiques, il faut nous en remettre au contexte pour l'identifier et la catégoriser. Or, sur 222 entités nommées pas reconnues ou mal identifiées, 179 sont des entités nommées pour lesquelles nous n'avons trouvé aucun mot déclencheur, aucun élément du contexte immédiat permettant de les catégoriser. Il nous faut donc trouver d'autres moyens pour permettre la reconnaissance de ces entités nommées.

5.4 Utilisation plus large de l'évidence externe et révision des catégories précédemment attribuées

Pour réaliser la reconnaissance des entités nommées jusqu'alors non ou mal reconnues, nous avons adjoint à **Nemesis** une troisième phase de reconnaissance, fondée sur une analyse plus large de l'évidence externe.

5.4.1 Le problème de la sur-composition référentielle

On parle de sur-composition référentielle lorsqu'une entité nommée mixte contient une entité nommée d'une autre catégorie référentielle. Nous avons vu que dans de tels cas, nous cherchons à étiqueter uniquement la forme la plus complète constituant une entité nommée : p. ex. *le festival de l'Université de Natal*, *le congrès du Parti Socialiste*.

Lors des deux premières phases de reconnaissance, seules sont considérées les sur-compositions mettant en jeu uniquement un mot déclencheur – en plus d'une entité nommée pure : p. ex. *la loi de Moore*, *l'ouragan Mitch*, *l'Université de Nantes*, *le mur de Berlin*. Cependant, **Nemesis** n'effectue pas de traitement spécifique de la sur-composition ; ces entités nommées sont reconnues par de simples règles du type : `[ $Clé_catégorieX $Maj+ ]` → **`CatégorieX`**. Ces règles ne font pas la distinction entre une entité nommée déjà reconnue et une suite de forme possédant une majuscule à l'initiale (variable `Maj`).

Partant de l'observation selon laquelle deux noms propres ne peuvent pas apparaître de façon contiguë sans ponctuation entre eux [Niu et coll., 2003], nous avons donc décidé d'étudier les sur-compositions référentielles les plus fréquentes, afin d'en induire des règles de révision sur les catégories d'entités nommées impliquées.

Tout d'abord, nous avons constaté que, eu égard à nos catégories référentielles, il existe quelques exceptions à la thèse de Niu et coll. [2003]. En effet, lorsque l'on étudie les différentes combinaisons de couples de catégories référentielles, nous avons relevé six cas où deux entités nommées peuvent se jouxter sans qu'il s'agisse d'une sur-composition :

- `Ethnonyme Patronyme` (p. ex. *le Brésilien Januy Santos Reis*) ;
- `Ethnonyme Organisation` (p. ex. *le Français ATTAC*) ;
- `Ethnonyme Entreprise industrielle` (p. ex. *le Français Sud-Marine Industrie*) ;
- `Ethnonyme Marque` (p. ex. *le Suédois Ikéa*) ;
- `Marque ou produit Marque ou produit` (p. ex. *AMD Athlon*, *Volkswagen Passat*) ;
- `Pays Patronyme` (p. ex. *le champion de France Éric Caritoux*) ;

Nous ne modifions donc pas les catégories affectées à ces entités nommées lorsque nous les rencontrons.

En revanche, nous avons discerné de réels cas de sur-compositions. Dans cette optique, nous avons regroupé certaines catégories référentielles de la façon suivante :

1. Les personnes physiques : patronymes et prénoms.
2. Les personnes morales : organisations, ensembles artistiques, entreprises industrielles, marques et établissements d'enseignement et de recherche.
3. Les toponymes.

Dès lors, nous pouvons identifier lesquelles de ces trois groupes de catégories peuvent entrer dans une surcomposition, et ce, pour chaque catégorie (cf. tableau 5.7).

Tableau 5.7 – Sur-compositions possibles

	Groupes entrant en surcomposition	**Exemples**
Faits historiques	Toponymes	*la chute du Mur de Berlin*
Évènements	Toponymes Personnes morales	*le festival de l'Université de Natal* *le Congrès des Verts*, *l'AG de l'ATALA*
Œuvres abstraites	Personnes physiques	*la loi de Murphy*, *le plan Marshall*
Édifices	Toponymes	*la Muraille de Chine*, *la tour de Pise*
Établissements d'ens. et de rech.	Toponymes Personnes physiques	*l'Université de Nantes* *le Lycée Nicolas Appert*
Organisations	Toponymes Personnes physiques	*l'Orchestre National de Barbès* *l'Association Bernard Grégory*
Phénomènes naturels	Personnes physiques	*l'ouragan Mitch*, *le cyclone Hugo*
Astres et comètes	Personnes physiques	*la comète de Halley*, *le Nuage d'Oort*
Entreprises indus.	Toponymes	*Renault Wilword*, *Nokia France*

À partir de ces résultats, nous avons construit des règles de réécritures basées sur les étiquettes posées lors de la projection des lexiques de mots déclencheurs, et sur les catégories d'entités nommées trouvées lors des deux premières passes. Ces règles sont au nombre de 20 et sont de la forme :

```
[ $Cat_organisation $Cat_patronyme ] → Organisation
[ $Cat_organisation $Article? $Classe_toponyme ] → Organisation
[ $Cle_œuvre_abs $Article? $Cat_patronyme ] → Œuvre
[ $Cat_entreprise $Classe_toponyme ] → Entreprise
[ $Cat_établissement $Cat_patronyme ] → Établissement
```

5.4.2 Traitement des structures énumératives

L'idée de départ de ce traitement est que les éléments d'une énumération sont de même nature. Cela est d'autant plus vrai pour les entités nommées. En effet, il est très rare d'énumérer des entités nommées de catégories référentielles différentes. Le seul cas que nous ayons relevé dans nos corpus concerne une énumération mélangeant des toponymes de différentes catégories (p. ex. *de l'Inde* (Pays), *des Philippine* (Pays) *et du Royaume-Uni* (Toponyme>Pays)). Cependant, cela reste quelque chose de marginal et l'on reste malgré tout dans la même classe.

Ce traitement s'effectue en trois étapes :

1. Identification des entités nommées pures et faiblement mixtes.
2. Analyse des différents éléments composant l'énumération (identification et association avec la catégorie référentielle le cas échéant).
3. Catégorisation des items et choix des heuristiques en cas d'ambiguïté.

5.4.2.1 Identification des entités nommées pures et faiblement mixtes

Parmi les 179 entités nommées non reconnues lors des deux premières étapes de reconnaissance, 166 sont des entités nommées pures (*Hindû-Kûsh*, *Mésopotamie*, *Rakesh Agrawal*), faiblement mixtes (*Socialisme et République*, *Îles de Beauté*) ou des sigles (*RAC*, *FAO*). Or, pour ces catégories graphiques (cf. section 3.2.2), l'identification des limites de l'entité nommée est immédiate. Nous identifions donc, dans un premier temps, les entités nommées non reconnues durant des deux premières passes de **Nemesis**. Cette procédure, si elle permet d'identifier les sigles et les entités nommées pures ou faiblement mixtes (100 % de rappel), accorde également à un nombre important de noms communs le statut de potentielle entité nommée (seulement 60 % de précision environ). Cependant, dans notre démarche, nous privilégions fortement le rappel, car nous ne cherchons pas à identifier les entités nommées, mais plutôt celles qui pourraient potentiellement l'être. En effet, une fois les candidats au statut d'entité nommée identifiés, nous les étiquetons uniquement lorsqu'ils apparaissent dans une structure énumérative avec des entités nommées déjà reconnues.

5.4.2.2 Analyse des différents éléments composant l'énumération

Nous recherchons donc, dans un deuxième temps, les structures énumératives mettant en jeu des entités nommées. Pour cela, nous avons créé 11 règles avec les variables suivantes :

- `$CEN_EN` : une entité nommée déjà reconnue ou un candidat entité nommée ;
- `$Virg`, `$Parent_O` et `$Parent_F` : respectivement une virgule, une parenthèse ouvrante et une parenthèse fermante ;
- `$Fin_enum` : une unité lexicale qui introduit les dernier élément d'une énumération (p. ex. *et*, *ainsi que*, *ou encore*) ;
- `$Last_item` : un élément qui clos une énumération (p. ex. | *etc.* | *...* | *, etc.* | *(...)* |).

Ces règles sont similaires à la partie gauche des règles des deux premières passes (cf. 4.2.2.2), auxquelles s'ajoute la possibilité de poser un quantificateur (`?`, `+` et `*`) sur plusieurs éléments, et non plus un seul. Pour cela, il suffit de regrouper ces différents éléments à l'intérieur d'accolades et de mettre le quantificateur à la suite. Voici quelques exemples de règles d'identification des structures énumératives :

`{ $CEN_EN $Virg }+ $CEN_EN $Fin_enum` (p. ex. *Paris, Londres, Berlin, etc.*)
`{ $CEN_EN $Virg }* $CEN_EN $Last_item $CEN_EN`
(p. ex. *Paris, Londres et Berlin*)
`{ à $CEN_EN $Virg }+ à $CEN_EN $Fin_enum`
(p. ex. *à Paris, à Londres, à Berlin, etc.*)
`{ pour $CEN_EN $Virg }* pour $CEN_EN $Last_item pour $CEN_EN`
(p. ex. *pour Paris, pour Londres et pour Berlin*)

À chaque fois qu'une telle règle est déclenchée, l'énumération identifiée est analysée : chaque élément est extrait et associé à sa catégorie référentielle le cas échéant. Sa position dans l'énumération est également relevée.

5.4.2.3 Catégorisation des éléments et choix des heuristiques en cas d'ambiguïté

À partir de là, il existe trois configurations concernant les catégories référentielles des entités nommées qui composent l'énumération et qui ont déjà été reconnues :

1. Il n'existe qu'une unique catégorie.
2. Il y a plusieurs catégories différentes, mais une seule catégorie dominante (la plus représentée).
3. Il y a plusieurs catégories différentes et plusieurs catégories dominantes.

Dans le premier cas, il n'y a pas d'ambiguïté et la solution est immédiate : nous balisons toutes les entités nommées de l'énumération avec cette unique catégorie.

Pour les deux autres cas, il peut y avoir une ambiguïté et nous avons mis en place plusieurs heuristiques. En effet, la question se pose tout d'abord de savoir si les catégories référentielles des entités nommées qui ont été précédemment reconnues doivent être révisées ou si seules les entités nommées non reconnues doivent être balisées. Suivant notre idée de départ, nous avons décidé d'affecter la même catégorie référentielle à toutes les entités nommées de l'énumération. À présent, il nous faut déterminer cette unique catégorie.

Dans le cas où il n'y aurait qu'une seule catégorie dominante, nous avons finalement choisi de prendre cette catégorie comme étant l'unique catégorie des entités nommées de l'énumération.

En revanche, s'il existe plusieurs catégories dominantes, il nous faut trouver un autre critère de désambiguïsation. Partant du constat que seule la première entité nommée de l'énumération pouvait avoir été catégorisée par une règle faisant intervenir le contexte gauche immédiat[8], nous avons décidé de retenir sa catégorie pour baliser toutes les entités nommées de l'énumération. Dans le cas où cette catégorie manquerait, nous n'avons pas opté pour la catégorie de la dernière entité nommée (dont la reconnaissance pourrait utiliser le contexte droit immédiat), mais plutôt en fonction de la catégorie graphique : nous privilégions les entités nommées des plus complexes aux plus simples (mixtes, sigles, faiblement mixtes, pures complexes, pures simples), car il est plus probable qu'elles aient été reconnues grâce à l'évidence interne qui est un indice plus fiable que leur simple présence dans un lexique. Les catégories référentielles des sigles ont une priorité élevée, car ils n'ont pu être reconnus que par association avec leur forme étendue qui est le plus souvent une entité nommée mixte.

La dernière de nos heuristiques concerne la différentiation entre les énumérations et les simples coordinations. En effet, la règle `{ $CEN_EN $Virg }* $CEN_EN $Last_item $CEN_EN` reconnaît une simple coordination comme une énumération, et cela peut poser problème. Pour les deux premières configurations, où un seul élément de la coordination aurait été préalablement reconnu, il nous paraît intéressant de baliser le second élément avec cette même catégorie, dans la mesure où il s'agit d'une entité nommée que nous n'arriverions pas à catégoriser autrement. En revanche, dans la troisième configuration, où les deux éléments de la coordination auraient été préalablement reconnus, nous avons décidé de ne pas effectuer de révision, car cela amenait plus de bruit que de corrections.

5.4.3 Évaluation

Une fois cette troisième phase de reconnaissance mise en place, nous avons évalué l'impact de ces deux traitements (sur-compositions référentielles et structures énumératives). À priori, nous

[8] Il s'agit de nos règles les plus fiables.

pouvons déjà constater que le traitement des sur-compositions ne permet pas la reconnaissance de nouvelles entités nommées, mais une révision des catégories précédemment affectées. L'apport en sera donc proportionnel sur les taux de rappel et de précision. En revanche, le traitement des structures énumératives (et des coordinations), s'il effectue également un travail de révision, permet principalement de catégoriser des entités nommées jusqu'alors non reconnues, car privées de contexte catégorisant. Ce traitement devrait permettre principalement l'augmentation du taux de rappel.

Les résultats présentés au tableau 5.8 confirment cette hypothèse. En effet, l'augmentation du taux de précision est plus faible (+0,7 %) que celle du taux de rappel (+2,2 %). En réalité, le traitement des sur-compositions a permis la révision de sept entités nommées (cinq organisations, une marque et un toponyme), alors que l'utilisation des structures énumératives a donné lieu a la reconnaissance de 20 entités nommées jusqu'alors non reconnues (six anthroponymes, dix toponymes et quatre ergonymes).

Tableau 5.8 – Résultats de l'évaluation de **Nemesis** avec une troisième passe

	Taux de rappel	Taux de précision
Anthroponymes	81,9 %	94,5 %
Toponymes	85,7 %	93 %
Ergonymes	64 %	76,1 %
Praxonymes	71,4 %	80 %
Total	81,2 %	91,7 %

Si les résultats ne sont pas spectaculaires, ils ont le mérite de n'avoir induit qu'une seule erreur (la catégorisation de *Royaume-Uni* en tant que pays) pour 27 entités nommées correctement balisées. Il s'agit donc d'un module très fiable et qui pourrait s'avérer d'une plus grande efficacité sur des textes où les deux premières passes auraient donné de moins bons résultats. En effet, si l'on prend par exemple les praxonymes : pour nos corpus la troisième passe n'est d'aucune utilité, alors que le traitement des sur-compositions – au contraire de celui sur les structures énumratives – est particulièrement intéressant pour cette classe au vu des nombreux toponymes qui peuvent en composer les entités nommées. Le fait qu'aucune révision n'ait été opérée sur ces entités nommées lors de la troisième passe s'explique par le fait qu'elles ont avaient été reconnues lors de la première passe, car les toponymes qui les composent faisaient parti de nos lexiques.

Une autre qualité de ce module consiste dans la complémentaire des traitements mis en place quant aux catégories graphiques dont ils permettent la reconnaissance. En effet, le traitement des sur-compositions ne permet que la révision des entités nommées mixtes ou pures complexes, alors que l'utilisation des structures énumératives ne donne lieu qu'à la reconnaissance de sigles, d'entités nommées pures et dans une moindre mesure d'entités nommées faiblement mixtes.

L'apport procuré par ce module est en réalité plus important qu'il n'y paraît en ce qui concerne les entités nommées reconnues grâce à l'utilisation des structures énumératives. En effet, ces entités nommées peuvent se retrouver ailleurs dans le texte et ce module pourrait être pleinement exploité après une autre passe qui rechercherait d'autres occurrences des entités nommées nouvellement reconnues. Dans notre corpus, nous avons identifié huit entités nommées qui pourraient être reconnues de cette façon, ce qui augmenterait de 40 % les performances de ce traitement.

5.5 L'utilisation du *Web* dans la reconnaissance des entités nommées

Pour identifier et catégoriser les entités nommées, **Nemesis** utilise largement les informations présentes dans le corpus : structure interne des entités nommées, contextes gauche et droit immédiats, contexte plus large à l'intérieur de la phrase (traitement des sur-compositions et des structures énumératives) ou plus loin dans le texte (apprentissage et gestion des coréférences). Même si ces traitements pourraient être plus performants, ils ne sauraient être suffisants. En effet, sur les 18,8 % de taux de silence[9], 65,7 % viennent d'entités nommées qui ne sont pas encore reconnues, car la plupart sont suffisamment connues pour que la compréhension du lecteur ne nécessite pas d'indices contextuelles permettant leur catégorisation (noms sans prénom, sigles sans leur forme étendue, marques, etc.). Par conséquent, ces entités nommées – tout comme une partie des 34,3 % restants – ne pourraient être reconnues par les actuels modules de **Nemesis**.

Nous avons donc décidé de rechercher la présence de ces entités nommées dans des contextes extérieurs aux textes que nous souhaitons traiter, afin de parvenir à leur catégorisation. Pour cela, nous avons étudié la possibilité d'utiliser le *Web*.

5.5.1 Recherche de nouveaux contextes via le *Web*

Nous avons vu que l'identification de la plupart des entités nommées non reconnues jusqu'alors (environ 93 %) est immédiate (cf. section 5.4.2.1), du fait de la catégorie graphique à laquelle ils appartiennent (sigles, entités nommées pures, ou faiblement mixtes).

Une fois les limites de ces entités nommées identifiées, nous en érigeons une liste. Nous lançons alors, pour chaque élément de cette liste, un processus (*thread*) qui émet une requête *html* sur *www.google.fr* avec les paramètres suivants :

- *catégorie* : pages francophones ;
- *nombre de résultats* : 20 par pages ;
- *recherche* : l'entité nommée entre guillemets.

Ensuite, pour chaque page donnée en réponse par *google*, le processus émet une nouvelle requête sur l'*url* de cette page, afin d'en récupérer le contenu et l'enregistrer dans un fichier.

Lorsque l'ensemble des processus sont terminés, un traitement est effectué pour chaque entité nommée et sur chaque fichier correspondant à la recherche lancée sur cette entité nommée. Pour les entités nommées pures ou faiblement mixtes, la méthode consiste à rechercher un élément catégorisant dans le contexte immédiat (droit ou gauche) de l'entité nommée. Pour les sigles, il s'agit de retrouver un schéma classique entre ce sigle et une éventuelle forme étendue, puis de déterminer la catégorie référentielle de cette dernière. Quelle que soit la catégorie graphique, il est très probable que l'entité nommée soit polysémique, surtout pour les sigles.

Pour les entités nommées pures ou faiblement mixtes, chaque fichier est traité ligne par ligne de la façon suivante :

1. Les balises *html* sont supprimées et la ligne est passée au format texte.
2. Si l'entité nommée est présente dans la ligne, son contexte est étudié pour y retrouver un élément de type `mot déclencheur` ou `contexte`.
3. Le cas échéant, la catégorie référentielle associée à cet élément est recensée.

[9] $\text{Silence} = \frac{\text{Réponses incorrectes ou oubliées}}{\text{Réponses attendues}} = 1 - \text{Rappel}$

À la fin, la catégorie référentielle la plus souvent retrouvée est associée à l'entité nommée.

Pour les sigles, le procédé est similaire :

1. Les balises *html* sont supprimées et la ligne est passée au format texte.
2. Si l'entité nommée est présente dans la ligne, une éventuelle forme étendue, correspondant à celle-ci selon certains schémas (cf. section 4.2.1.2), est recherchée.
3. Le cas échéant, cette forme étendue est recensée.

À la fin, l'entité nommée se voit attribuée la catégorie référentielle de la forme étendue la plus souvent associée au sigle, si celle-ci est une entité nommée.

Un tel traitement est excessivement lourd, dans la mesure où les lexiques contenant des éléments de type `mot déclencheur` ou `contexte` sont relativement volumineux (11 342 éléments au total) notamment pour les prénoms (9 216 entrées). Partant d'un tel constat, et de l'hypothèse de Gale et coll. [1992], selon laquelle un nom ne peut avoir qu'un sens par document (hypothèse particulièrement réaliste en ce qui concerne les noms propres), nous avons décidé d'arrêter le traitement d'un document `html`, dès lors que l'entité nommée y a été trouvée dans un contexte catégorisant.

5.5.2 Évaluation de l'apport du *Web* et induction d'heuristiques

Nous avons évalué l'apport que pouvait amener l'utilisation du *Web* dans la reconnaissance des entités nommées sur un corpus regroupant les textes de notre corpus de test (environ 7 000 mots issues du *Monde*) et l'article de la revue *Unasylva* appartenant à notre corpus d'évaluation (environ 7 000 mots), pour un total de 649 entités nommées.

À priori, cette méthode paraît permettre la catégorisation de certaines entités nommées : *Hindû-Kûsh* est ainsi retrouvé dans quatre contextes différents (*monts*, *chaîne*, *massif*, *montagnes*) que nous savons appartenir à la même catégorie référentielle. On trouve également de nombreuses fois *Rakesh Agrawal* avec le contexte *Dr* ou encore *Sagarmatha* avec *parc national* ou *parc*. Ces informations devraient nous permettre de reconnaître correctement ces entités nommées. Cependant, lorsque le traitement est automatisé, différents problèmes se posent.

Tout d'abord, durant la phase d'identification des entités nommées pures ou faiblement mixtes et des sigles, un certain nombre d'éléments sont retenus, à tort, comme potentiellement entité nommée (environ 40 % des éléments identifiés). Au total, 118 formes (44 pour *Le Monde* et 74 pour la page *Web* de la FAO), dont 20 sigles, sont traitées pour 72 entités nommées, dont 8 sigles. Cela pose deux problèmes :

1. Augmentation du temps de traitement.
2. Reconnaissance de noms communs comme entités nommées.

La reconnaissance d'un nom commun comme entité nommée est rare dans les résultats que nous avons obtenus (quatre occurrences). De plus, il n'y a, à chaque fois, qu'une très faible présomption (une seule catégorie référentielle retrouvée une seule fois). Par conséquent, un seuil devrait être suffisant pour réduire ce type de problèmes (voir la suite de cette section). Pour ce qui est de l'augmentation du temps de traitement, la seule possibilité que nous envisagions pour pouvoir ne traiter que les entités nommées et pas les noms communs, consiste à ne pas essayer de reconnaître, par cette méthode, les entités nommées présentes en début de phrase.

D'autre part, le fait de travailler sur des pages *Web* pose problème : de nombreuses *url* sont introuvables, une même phrase peut se trouver sur plusieurs lignes, etc. Cela induit une baisse du taux de rappel. Dans le premier cas, nous pourrions prendre les dix ou vingt premières pages *Web* d'une taille suffisant, pour être sûrs qu'elles contiennent un minimum d'information. Dans le second cas, nous pourrions éventuellement travailler sur l'ensemble du fichier au lieu de le faire ligne par ligne, mais cela ralentirait sensiblement le traitement.

L'utilisation du *Web* dans **Nemesis** peut apporter un gain important sur le taux de rappel, car elle permet la catégorisation d'entités nommées non reconnues jusqu'alors. En revanche, elle peut dans le même temps faire baisser le taux de précision si nous ne nous assurons pas d'un minimum de présomption lors de la catégorisation. En effet, il est fort probable que le fait de retrouver sur le *Web* une entité nommée dans un unique contexte catégorisant n'est pas suffisant pour prendre une décision. Cependant, il est délicat de définir un seuil d'occurrences à partir duquel une entité nommée peut être catégorisée. Cela reste donc un paramètre à préciser selon que l'on veut privilégier le rappel ou la précision.

De plus, une même forme peut tenir le rôle de `contexte` ou de `mot déclencheur` pour différente catégories référentielles (p. ex. *Général* peut être un `mot déclencheur` pour une entreprise ou un `contexte` pour un patronyme, *groupe* peut être un `contexte` pour une entreprise, une organisation ou un ensemble artistique). Sachant que nous avons décidé de ne garder qu'une catégorie référentielle par texte pour une même entité nommée, l'ordre dans lequel nous allons explorer nos lexiques va être déterminant. Pour cela, nous nous basons sur un indice de confiance établi manuellement pour chacun de nos lexiques (p. ex. le lexique des titres militaires possède un indice plus élevé que celui des mots-clés d'organisation, lui-même plus élevé que celui des mots-clés d'entreprises).

Enfin, la forte polysémie liée aux entités nommées pose le problème le plus important pour la mise en place d'un module de reconnaissance à partir du *Web*. En effet, une entité nommée polysémique a de forts risques de voir ses différentes catégories retrouvées sur le *Web*. Pour cela, nous disposons de différents indices :

- le nombre d'occurrences, trouvées via le *Web*, de chaque catégorie pour une entité nommée donnée ;
- le nombre de contextes différents, trouvées via le *Web*, de chaque catégorie pour une entité nommée donnée ;
- la distribution moyenne des entités nommées en fonction de leur catégorie référentielle (cf. tableau 3.3) ;
- la distribution des entités nommées en fonction de leur catégorie référentielle dans le corpus en cours de traitement (nous nous basons ici via les entités nommées déjà reconnues).

Là encore, il est très difficile de dire comment prendre en compte tous ces critères pour donner le plus sûrement la catégorie correcte. Sur les 72 entités nommées traitées, 41 ont été trouvées en présence d'un contexte catégorisant sur le *Web* (cf. tableau 5.9). Sur ces 44 entités nommées, 22 sont largement bien catégorisées et cinq mal catégorisées (p. ex. *Schutzenberger* comme patronyme alors qu'il s'agit d'une brasserie, *Cyrnos* comme édifice alors qu'il s'agit d'un ferry). Dix de ces entités nommées n'ont été catégorisées que par un seul contexte (quatre correctement et six incorrectement). Enfin, sept entités nommées ont été retrouvées dans différents contextes contradictoires (*Massif du Vercors* et *tour du Vercors* ou *Parc de Sagarmatha*, *Poste Sagarmatha* et *Radio Sagarmatha*). Or, sur ces deux exemples appartenant au corpus *Unasylva*, la distribution des entités nommées indiquent clairement que c'est un texte qui traite de géographie (plus

de 31 % des entités nommées contre moins de 20 % en général). Par conséquent, il serait légitime de privilégier les catégories de cette classe. En prenant ce seul critère de décision, nous obtenons une catégorisation correcte pour six de ces sept entités nommées.

Tableau 5.9 – Résultats de l'évaluation de l'utilisation du *Web*

EN bien catégorisées en majorité avec un seuil supérieur à 1	22
EN bien catégorisées en majorité avec un seuil égal à 1	4
EN mal catégorisées en majorité avec un seuil supérieur à 1	5
EN mal catégorisées en majorité avec un seuil égal à 1	6
EN pouvant être bien catégorisée selon le critère de décision	7

Par conséquent, nous pouvons estimer – avec un seuil entre deux et quatre, et en tenant compte de la distribution des entités nommées en fonction de leur catégorie référentielle dans le corpus en cours de traitement – que le taux de précision de ce module est d'environ 80 %. La plupart des entités nommées catégorisées apparaissant plusieurs fois dans le texte, ce module augmente actuellement le taux de rappel d'un peu plus de 5 % (81,2 % avant, 86,5 % après), tandis qu'il fait baisser la précision d'environ 2 % (91,7 % avant, 90 % après).

En plus du seuil au dessous duquel nous choisissons de ne pas catégoriser une entité nommée et des critères de décision, il reste d'autres paramètres qui peuvent faire varier les résultats : le nombre de pages explorées, le moteur de recherche utilisé ou encore la catégorie linguistique de la recherche (pages francophones ou pages « France » pour *google*). Par conséquent, ce module n'est probablement pas encore optimisé.

Malgré le faible gain, ces premiers résultats restent encourageants dans la voie d'une reconnaissance des entités nommées inconnues à l'aide du *Web*, dans la mesure où un tel traitement permet la catégorisation d'entités nommées pour lesquelles nous n'avions aucune information quant à leur catégorie référentielle. En ce sens, même si le taux général de précision a chuté, notre module n'induit pas de réel bruit[10], car il ne modifie pas les résultats préalablement corrects.

Certaines améliorations peuvent encore être apportées à notre module : prise en compte uniquement des pages ne résultant pas d'une erreur *http*, mise au point de nouvelles heuristiques, utilisation d'autres moteurs de recherche (notamment les encyclopédies et atlas en lignes), traitement permettant de « raccrocher » les phrases se trouvant sur plusieurs lignes, etc.

5.6 Conclusion

Les résultats obtenus, sur l'ensemble de nos textes et des entités nommées, sont d'environ 80 % pour le rappel et 90 % pour la précision.

Nous avons tout d'abord effectué une étude des conflits liés à l'application des règles de réécritures. Cette étude nous a permis de réordonner nos règles et ainsi modifier leurs priorités d'application pour limiter *a priori* les conflits qui pourraient survenir lors de leur application. Cette étude pourra être d'avantage exploitée afin de mettre en place un module de désambiguïsation et d'inférence de nouvelles règles accompagné d'un algorithme d'insertion des nouvelles

[10] $\text{Bruit} = \frac{\text{Réponses incorrectes}}{\text{Réponses apportées}} = 1 - \text{Précision}$

règles parmi celles déjà existantes. Pour cela, il nous faudra approfondir cette étude en l'étendant à toutes les classes d'entités nommées et avec une version complète de **Nemesis**.

Ensuite, nous avons mis en place une troisième étape de reconnaissance des entités nommées, basée sur une plus large utilisation de l'évidence externe (traitement des sur-compositions référentielles et utilisation des structures énumératives), afin de palier l'incomplétude des lexiques et le manque d'informations permettant la catégorisation dans le contexte immédiat. Ce module permet un gain de 0,7 % en précision et 2,2 % en rappel. Cependant, ce gain pourra être augmenté par une dernière passe qui n'effectuerait que de la recherche de coréférences.

Enfin, nous avons étudié les possibilités d'utilisation du *Web* pour la reconnaissance automatique d'entités nommées, en y recherchant de nouveaux contextes catégorisant. Dans ce cadre, nous avons donc implémenté un module qui apporte une amélioration de 5 % sur le taux de rappel, avec une perte de 2 % sur la précision. Ce module pourra également être perfectionné, tant les paramètres dont il dépend sont nombreux.

Chapitre 6

Conclusion et perspectives

Ce dernier chapitre a pour but de faire le bilan du travail réalisé durant notre thèse. Pour ce faire, nous rappelons dans un premier temps les enjeux et la problématique, puis nous résumons les travaux effectués et enfin nous dégageons les perspectives qu'ouvre cette thèse.

6.1 Conclusion

Notre travail s'inscrit dans le domaine du Traitement Automatique des Langues Naturelles (TALN) et plus précisément en reconnaissance des entités nommées du français. Cette tâche, qui consiste à identifier et à catégoriser automatiquement les entités nommées présentes dans des textes électroniques écrits en français, constitue un enjeu considérable tant sa réalisation est nécessaire à de nombreuses applications du TALN (veille technologique, indexation de documents, extraction d'information, traduction automatique, etc.).

Si cette reconnaissance est réalisée de façon satisfaisante en extraction d'information, pour des textes journalistiques anglais, elle reste en revanche insuffisante, dès lors qu'elle porte sur des textes français, en particulier lorsque l'on souhaite obtenir une catégorisation fine. Nous avons donc réalisé un système d'identification et de catégorisation automatiques des entités nommées du français : **Nemesis**. Ce système s'appuie sur une typologie des entités nommées que nous avons voulu évolutive et la plus fine possible, afin de permettre son utilisation dans toutes les applications du TALN qui la requièrent.

Avant toutes choses, nous avons interrogé dans ce manuscrit les concepts de « nom propre » et d'« entité nommée », afin de faire émerger des paramètres définitoires opérationnels de ce dernier. Pour cela, nous avons étudié les fondements typologiques et linguistiques de l'intuition de « nom propre » afin d'en restituer la nature et les origines, puis nous avons situé les entités nommées dans le continuum noms propres/noms communs, avant de dégager les paramètres linguistiques caractérisant les entités nommées et de définir un principe de sélection de celles-ci en corpus.

Nous avons ensuite présenté un état de l'art des différents travaux réalisés en reconnaissance des entités nommées et avons montré que les meilleurs performances sont atteintes par des systèmes hybrides qui allient méthodes linguistiques et méthodes à base d'apprentissage. Par ailleurs, nous avons constaté – et ce, quelles que soient les méthodes mises en place – que les performances diminuent fortement dès lors que la catégorisation recherchée est plus fine que les simples classes MUC.

À la suite de cet état de l'art, une double étude en corpus – graphique et référentielle – a été présentée. L'étude référentielle a conduit à l'élaboration de notre typologie des entités nommées du français, comportant 27 catégories réparties en cinq classes, et sur laquelle sera fondée notre

système de reconnaissance automatique des entités nommées. L'étude graphique quant à elle nous a donné à voir les problèmes posés par l'identification et la catégorisation des entités nommées selon leur graphie.

Conséquemment à ces trois études, nous avons réalisé **Nemesis**, un système d'identification et de catégorisation automatiques des entités nommées dans les textes français. Son architecture logicielle se compose principalement de quatre modules qui effectuent un traitement séquentiel immédiat des données à partir de textes bruts. Le premier module constitue une phase de prétraitement lexical (segmentation du texte en formes et en phrases et association des sigles à leurs formes étendues). Le second module procède à une première reconnaissance des entités nommées en analysant leur structure interne et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles lexico-sémantiques, graphiques et morphologiques. En parallèle à cette première reconnaissance, un module d'apprentissage crée des listes temporaires à l'aide des entités nommées déjà reconnues, en appliquant un ensemble d'heuristiques. Ces listes d'entités nommées sont alors utilisées par le dernier module qui ré-applique des règles utilisées lors de la première reconnaissance et en applique de nouvelles, afin d'effectuer une seconde reconnaissance. Parallèlement aux deux phases de reconnaissance des entités nommées, toutes les formes co-référant à une même entité nommée se voient attribuer un même identifiant.

Après avoir évalué l'apport de chacun de ces modules sur les anthroponymes et les toponymes, nous avons effectué une évaluation de l'ensemble de **Nemesis** sur toutes les catégories d'entités nommées. Les performances atteintes sont satisfaisantes, mais elles mettent en relief les limites d'un tel système. Si le taux de précision obtenu est très satisfaisant (environ 90 %), un gain substantiel du taux de rappel (d'environ 79 %) ne saurait faire l'économie de la mise en place d'autres techniques. Nous avons donc ajouté différents modules permettant d'améliorer le rappel : examen d'un contexte encore plus large (traitement des sur-compositions référentielles et analyse des structures énumératives) et utilisation du Web comme source de nouveaux contextes. Nous avons également proposé une étude des conflits engendrés par les règles de réécritures qui nous a permis de réordonner les priorités d'application de nos règles, afin de limiter ces conflits *a priori* et de faciliter l'ajout de nouvelles règles créées manuellement.

6.2 Perspectives

Les premières perspectives concernent les améliorations qui peuvent être apportées à **Nemesis**. La première concerne l'identification de la frontière droite des entités nommées. Nous constatons que les entités nommées mal identifiées sont des entités nommées mixtes appartenant en grande majorité aux noms d'organisations, mais également aux ensembles artistiques, aux entreprises industrielles, aux établissements d'enseignement et de recherche, aux évènements et aux œuvres (cf. section 3.4). Elles sont correctement catégorisées par **Nemesis**, mais il manque des éléments à rattacher à droite. Une analyse locale plus fine du contexte droit des entités nommées de ces catégories pourrait permettre de rattacher la partie droite manquante et d'augmenter ainsi la précision de la reconnaissance.

Nous avons vu dans le chapitre 5 que notre étude sur les conflits pouvait nous mener à la réalisation d'un module de désambiguïsation et d'apprentissage de nouvelles règles. Ce module pourrait permettre une amélioration des performances de **Nemesis** face au passage à de nouveaux corpus en repérant automatiquement les situations de conflit sur les entités nommées reconnues, puis en proposant à l'utilisateur, par un processus supervisé, des solutions pour résoudre ces

conflits, notamment par l'inférence de nouvelles règles accompagnée d'un algorithme d'insertion automatique de celles-ci.

Dans le même chapitre, nous avons montré que les résultats de notre module de traitement des sur-compositions référentielles et d'analyse des structures énumératives n'étaient pas totalement exploités. En effet, ce module permet la reconnaissance de nouvelles entités nommées qui peuvent se retrouver ailleurs dans le texte, mais toujours pas catégorisées. Il serait donc intéressant de mettre en place une dernière étape de reconnaissance qui ne chercherait que les autres occurrences des entités nommées reconnues par ce module, afin de les catégoriser.

Une autre voie d'amélioration qui pourrait être explorée concerne la mise en place d'un module à intégrer à la dernière passe. Ce module effectuerait deux traitements différents.

Le premier consisterait en une validation des entités nommées présentes en début de phrases, en fonction de la présence, ailleurs dans le texte, de leur première forme. En effet, si cette première forme se retrouve, toujours avec une majuscule à l'initiale, au milieu d'une phrase, il y a de fortes présomptions sur son statut d'entité nommées ; à l'inverse, si cette forme se retrouve en minuscule dans le texte, il y a des chances que le sujet ne soit pas une entité nommée [cf. Wacholder et coll., 1997].

Le second traitement réaliserait la catégorisation des entités nommées identifiées mais pas catégorisées (les entités nommées pures et faiblement mixtes). Pour cela, nous envisageons deux méthodes complémentaires. La première consiste à fusionner une entité nommée déjà reconnue et un candidat entité nommée qui la jouxterait, en partant de l'observation selon laquelle deux noms propres ne peuvent pas apparaître de façon contiguë sans ponctuation entre eux [Niu et coll., 2003]. La deuxième méthode consiste à rechercher des sous-chaînes, non catégorisées, d'entités nommées déjà reconnues et de les traiter comme des coréférences de cette entité nommée si elles appartiennent à une catégorie graphique des entités nommées.

Ce dernier traitement ne serait certainement pas très précis (identification trop large et noms communs reconnus comme entités nommées), mais permettrait de couvrir un maximum d'entités nommées. Il faudrait donc que son intégration soit optionnelle en fonction de la tâche à réaliser. Malgré tout, ce traitement nous paraît intéressant car il est plus facile de filtrer l'information extraite que de la retrouver dans le corpus.

Les perspectives les plus intéressantes de notre travail concernent l'intégration de **Nemesis** dans des applications de traitement automatique des langues. En effet, l'élaboration d'un système de reconnaissance des entités nommées n'est pas une fin en soi. À ce titre, il nous paraît primordial d'intégrer **Nemesis** dans de telles applications et d'évaluer son apport, notamment par rapport aux systèmes de reconnaissance des entités nommées existants. *A priori*, cet apport possède une nature double liée à notre typologie des entités nommées : la finesse des catégories référentielles et l'exhaustivité de la reconnaissance. En effet, la plupart des systèmes actuels ne traitent que les noms de personnes, d'organisations et de lieux. Par conséquent, ils omettent la reconnaissance de plus de 20 % des entités nommées, principalement les praxonymes, les phénonymes, les œuvres, les établissements d'enseignement et de recherche, les marques et produits et les ensembles artistiques (cf. tableau 3.3). De fait, même avec un taux de reconnaissance de 100 %, la quantité d'information extraite ne devrait pas être aussi importante qu'avec **Nemesis**, dont le taux de reconnaissance (P&R) n'atteint que 84,7 %. Cette quantité d'information extraite par **Nemesis** peut être encore augmentée selon la tâche à réaliser. En effet, si pour de la veille technologique une typologie de faible granularité est suffisant, en revanche, pour des systèmes de

Question/Réponse (Q/R) ou de traduction automatique, une granularité plus importante sera pertinente et par conséquent la quantité d'information augmentée.

Si l'intégration de **Nemesis** peut se faire dans tous les domaines nécessitant un module de reconnaissance des entités nommées en adaptant cette granularité, elle sera plus profitable à certaines applications. Parmi celles-là, nous en avons envisagées plusieurs.

Tout d'abord la traduction automatique pour laquelle il faut étudier le comportement de chaque catégorie référentielle face à la traduction (traduire, ne pas traduire, transcrire) et ce pour chaque langue concernée [Grass, 2000].

Ensuite, les systèmes de Question/Réponse et les moteurs de recherche fondés sur une approche linguistique (cf. *www.synomia.fr*) pour lesquels un module de reconnaissance des entités nommées est indispensable. Dans ce cadre, la reconnaissance des praxonymes ou des marques notamment peut s'avérer très pertinente dans des questions ou des requêtes comme *En quelle année la France a-t-elle remporté la Coupe du Monde de football ?* ou *Quelle marque commercialise le Nutella ?*

Enfin, les systèmes de courriels proposant un transfert dans la prise de décision, qui nécessitent une analyse surfacique associée à un module de reconnaissance des entités nommées, afin de réacheminer les courriels vers le destinataire le plus compétent pour y répondre.

L'intégration de **Nemesis** à l'une de ces applications reste donc la principale perspective de cette thèse.

Bibliographie

J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson et M. Vilain. Mitre: Description of the alembic system as used for muc-6. Dans *Proceedings of the 6th Message Understanding Conference*, Columbia, États-Unis, nov 1995. Morgan Kaufmann.

S. Abney. Partial parsing via finite-state cascades. Dans *Proceedings of the 8th European Summer School in Logic, Language and Information (ESSLLI'96) Robust Parsing Workshop*, pages 8–15, Prague, République Tchèque, 1996.

E. Agirre et G. Rigau. Word sense disambiguation using conceptual density. Dans *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

D. J. Allerton. The linguistics and socialinguistic status of proper names. *Journal of Pragmatics*, 11 : 61–92, 1987.

C. Aone et M. Ramos-Santacruz. Rees: A large-scale relation and evaluation system. Dans *Proceedings of ANLP/NAACL-2000*, 1998.

D. Appelt et D. Martin. Named entity recognition in speech: Approach and results using the textpro system. Dans *Proceedings of the DARPA Broadcast News Workshop*, pages 51–54, Herndon, États Unis, 1999.

D. E. Appelt et D. J. Israel. Introduction to information extraction technology. Tutorial of the International Joint Conference on Artificial Intelligence (IJCAI'99), 1999.

S. Aït-Mokhtar. Du texte ASCII au texte lemmatisé : la présyntaxe en une seule étape. Dans *Actes, Quatrième conférence sur Traitement Automatique du Langage Naturel (TALN'97)*, pages 60–69, Grenoble, France, juin 1997.

R. Basili, A. Marziali et M. T. Pazienza. Modelling syntactic uncertainty in lexical acquisition from texts. *Journal of Quantitative Linguistics 1*, 1(1) : 62–81, 1994.

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

F. Béchet, A. Nasr et F. Genet. Tagging unknown proper names using decision trees. Dans *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, pages 77–84, Hong-Kong, octobre 2000.

F. Béchet et F. Yvon. Les noms propres en traitement automatique de la parole. *Traitement automatique des langues*, 41(3) : 671–708, 2000.

C. Belleil. *Traitement informatique des noms propres en lien avec la géographie.* Thèse en informatique, Université de Nantes, 1997.

D. M. Bikel, S. Miller, R. Schwartz et R. Weischedel. Nymble: A high-performance learning name finder. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington DC, États-Unis, 1997.

A. Blum et T. Mitchell. Combining labaled and unlabeled data with co-training. Dans *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*, 1998.

S. Boisen, M. Crystal, R. Schwartz, R. Stone et R. Weischedel. Annotating resources for information extraction. Dans *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, pages 1211–1214, Athènes, Grèce, mai 2000.

A. Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. Thèse de doctorat, Université de New York, New York, États-Unis, septembre 1999.

L. Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, 1984.

E. Brill. *A Corpus-Based Approach to Language Learning*. Thèse de doctorat, Université de Pennsylvanie, juin 1993.

E. Brill. Some advances in transformation-based part of speech tagging. Dans *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, pages 722–727, Seattle, États-Unis, 1994.

E. Brill. Transformation-based error-driven learning and natural language processing: a case study part of speech tagging. *Computational Linguistics*, 21(24), 1995.

E. Brill et P. Resnik. A rule-based approach to prepositional phrase attachment disambiguation. Dans *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Tokyo, Japon, 1994.

E. Charniak, C. Hendrickson, N. Jacobson et M. Perkowitz. Equations for part-of-speech tagging. Dans *Proceedings of the 11th National Conference on Artificial Intelligence*, pages 784–789, Menlo Park, États-Unis, juillet 1993.

S. Coates-Stephens. The analysis and acquisition of proper names for the understanding of free text. Dans *Computers and the Humanities*, volume 26, pages 441–456. Kluwer Academic Publishers, Hingham, États-Unis, 1993.

M. Collins. A new statistical parser based on bigram lexical dependencies. Dans *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, 1996.

M. Collins et Y. Singer. Unsupervised models for named entity classification. Dans *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-WVLC'99)*, pages 100–110, College Park, États-Unis, 1999.

A. Cucchiarelli, D. Luzi et P. Velardi. Automatic semantic tagging of unknown proper names. Dans C. Boitet et P. Whitelock : rédacteurs, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, volume 1, pages 286–292, Montréal, Canada, 1998. Morgan Kaufmann Publishers.

A. Cucchiarelli, D. Luzi et P. Velardi. Semantic tagging of unknown proper nouns. *Natural Language Engineering*, 5(2) : 171–185, 1999.

A. Cucchiarelli et P. Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1) : 123–131, mars 2001.

S. Cucerzan et D. Yarowsky. Language independent named entity recognition using a unified model of internal and contextual evidence. Dans *Proceedings of CoNLL-2002*, pages 171–175, 2002.

B. Daille et E. Morin. Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. *Traitement automatique des langues*, 41(3) : 601–621, 2000.

I. Demiros, S. Boutsis, V. Giouli, M. Liakata, H. Papageorgiou et S. Piperidis. Named entity recognition in greek texts. Dans *Proceedings of LREC 2000*, 2000.

A. Dister. Problématique des fins de phrase en traitement automatique du français. Dans L. Rosier, F. Tilkin et J.-M. Defays : rédacteurs, *Champs linguistiques*, page 470. Duculot, Paris, 1997.

E. Eggert, D. Maurel et C. Belleil. Allomorphies et supplétions dans la formation des gentilés. application au traitement informatique. *Cahiers de Lexicologie*, 73 : 167–179, 1998.

N. Friburger. Pré-traitement pour une fouille de textes basée sur les noms propres. Dans *Actes, TALN-Récital 2000*, pages 471–476, Lausanne, Suisse, juin 2000.

N. Friburger. *Reconnaissance automatique des noms propres. Application à la classification automatique de textes journalistiques.* Thèse de doctorat, Université de Tours, décembre 2002.

R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham et Y. Wilks. Description of the lasie system as used for muc-6. Dans *Proceedings of the 6th Message Understanding Conference*, pages 207–220, Columbia, États-Unis, novembre 1995. Morgan Kaufmann.

W. Gale, K. Church et D. Yarowsky. One sense per discourse. Dans *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 1992.

A. F. Gallippi. Learning to recognize names across languages. Dans *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 424–429, Copenhague, Danemark, August 1996.

M.-N. Gary-Prieur. Le nom propre constitue-t-il une catégorie linguistique ? Dans *Syntaxe et sémantique des noms propres*, numéro 92 de Langue française, pages 4–25. Larousse, décembre 1991a.

M.-N. Gary-Prieur : rédacteur. *Syntaxe et sémantique des noms propres.* Numéro 92 de Langue française. Larousse, décembre 1991b.

M.-N. Gary-Prieur. *Grammaire du nom propre.* Linguistique nouvelle. Presses Universitaires de France - PUF, Paris, 1994.

C. Girardin. Contenu, usage social et interdits dans le dictionnaire. *Langue Française*, (43) : 84–99, 1979.

T. Grass. Typologie et traductibilité des noms propres de l'allemand vers le français à partir d'un corpus journalistique. Dans *Journée d'Étude de l'ATALA « Le traitement automatique des noms propres »*, Université Paris 7, France, mai 1999. ATALA.

T. Grass. Typologie et traductibilité des noms propres de l'allemand vers le français. *Traitement automatique des langues*, 41(3) : 643–670, 2000.

G. Grefenstette et P. Tapanainen. What is a Word, What is a Sentence? Problems of Tokenization. Dans *Actes, 3rd International Conference on Computational Lexicography (COMPLEX'94)*, pages 79–87, Budapest, July 1994.

M. Grevisse. *Le bon usage. Grammaire française.* Duculot, 13ème édition refondue par André Goosse, Paris, juillet 1993.

R. Grishman. The nyu system for muc-6 or where's the syntax? Dans *Proceedings of the 6th Message Understanding Conference*, Columbia, États-Unis, novembre 1995. Morgan Kauffmann.

P. Hayes. Namefinder - software that finds names in text. Dans *Proceedings of the fourth RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO'94)*, pages 762–774, New York, oct 1994.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham et Y. Wilks. University of sheffield: Description of the lasie-ii system as used for muc-7. Dans *Proceedings of the 7th Message Understanding Conference*, Fairfax, États-Unis, mai 1998.

K. Humphreys, R. Gaizauskas, H. Cunningham et S. Azzam. Vie technical specifications. ILASH, Université de Sheffield, 1996.

N. Ide et J. Véronis. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1) : 1–40, 1998.

K. Jonasson. *Le Nom Propre. Constructions et interprétations.* Champs linguistiques. Duculot, 1994.

J.-H. Kim, I.-H. Kang et K.-S. Choi. Unsupervised named entity classification models and their ensembles. Dans *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, Taipei, Taiwan, 2002.

G. Kleiber. *Problèmes de référence : descriptions définies et noms propres.* Klincksiek, Paris, France, 1981.

G. Kleiber. *La sémantique du prototype. Catégories et sens lexical.* Linguistique nouvelle. Presses Universitaires de France - PUF, Paris, 1990.

L. Kosseim et G. Lapalme. Exibum : Un système expérimental d'extraction d'information bilingue. Dans *Actes de la Rencontre Internationale sur l'extraction, le filtrage et le résumé automatiques (RIFRA-98)*, pages 129–140, Sfax, Tunisie, novembre 1998.

G. R. Krupka et K. Hausman. Isoquest: Description of the netowl^{TM} extractor system as used for muc-7. Dans *Proceedings of the 7th Message Understanding Conference*, Fairfax, États-Unis, mai 1998.

F. Kubala, R. Schwartz, R. Stone et R. Weischedel. Named entity extraction from speech. Dans *Proceedings of the DARPA Workshop on Broadcast News Understanding Systems*, Lansdowne, États-Unis, 1998.

R. Kuhn et R. D. Mori. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5) : 449–460, mai 1995.

M. Le Bihan. *Le Nom Propre. Étude de grammaire et de réthorique.* Thèse de doctorat, Université de Bretagne Occidentale, 1974.

D. Lin. Automatic retrieval and clustering of similar words. Dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 768–774, 1998a.

D. Lin. Using collocation statistics in information extraction. Dans *Proceedings of the 7th Message Understanding Conference*, Fairfax, États-Unis, mai 1998b.

Y. Matsumoto, S. Kurohashi, O. Yamaji, Y. Taeki et M. Nagao. Japanese morphological analysing system: Juman. Université de Kyoto et Institut de Science et Technologie de Nara (NAIST), 1997.

D. D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. Dans B. Boguraev et J. Pustejovsky : rédacteurs, *Corpus Processing for Lexical Acquisition*, Language, Speech and Communications, chapitre 2, pages 21–40. MIT Press, 1996.

A. Mikheev. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3) : 405–423, 1997.

A. Mikheev. Feature lattices for maximum entropy modelling. Dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics*, pages 848–854, Montréal, Québec, August 1998.

A. Mikheev. A knowledge-free method for capitalized word disambiguation. Dans *Proceedings of the 37th Annual Meeting of the ACL*, pages 159–166, Université de Californie, États-Unis, 1999.

A. Mikheev, M. Moens et C. Grover. Named entity recognition without gazetteers. Dans *Proceedings of the Ninth International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, Norvège, juin 1999.

S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel et the Annotation Group. Algorithms that learn to extract information – bbn: Description of the sift system as used for muc-7. Dans *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, États-Unis, mai 1998.

J. Molino : rédacteur. *Le nom propre*. Langages. Larousse, Paris, juin 1982a.

J. Molino. Le nom propre dans la langue. Dans *Le nom propre*, numéro 66 de Langages, pages 5–20. Larousse, Paris, juin 1982b.

R. J. Mooney. Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning*, 10 : 79–110, 1993.

E. Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse en informatique, Institut de Recherche en Informatique de Nantes, 1999.

MUC-3. *Proceedings of the 3rd Message Understanding Conference*, San Diego, États-Unis, 1991. Morgan Kauffmann.

MUC-4. *Proceedings of the 4th Message Understanding Conference*, San Mateo, États-Unis, 1992. Morgan Kauffmann.

MUC-5. *Proceedings of the 5th Message Understanding Conference*, San Mateo, États-Unis, 1993. Morgan Kauffmann.

MUC-6. *Proceedings of the 6th Message Understanding Conference*, Columbia, États-Unis, novembre 1995. Morgan Kauffmann.

MUC-7. *Proceedings of the 7th Message Understanding Conference*, Fairfax, Virginie, États-Unis, 1998. Morgan Kauffmann.

G. Nenadić et I. Spacić. Recognition and acquisition of compound names from corpora. Dans *Natural Language Processing - NLP 2000, Second International Conference*, pages 38–48, Patras, Greece, 2000.

C. Niu, W. Li, J. Ding et R. K. Srihari. A bootstrapping approach to named entity classification using successive learners. Dans *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'2003)*, pages 335–342, Sapporo, Japon, juillet 2003.

M. Noailly : rédacteur. *Nom propre et nomination, Actes du colloque de Brest*, avril 1994.

W. Paik, E. D. Liddy, E. Yu et M. McKenna. Categorizing and standardizing proper nouns for efficient information retrieval. Dans B. Boguraev et J. Pustejovsky : rédacteurs, *Corpus Processing for Lexical Acquisition*, Language, Speech and Communications, chapitre 4, pages 61–76. MIT Press, 1996.

D. D. Palmer et D. S. Day. A statistical profile for the named entity task. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 190–193, Washington DC, États-Unis, 1997.

T. Poibeau. Repérage des entités nommées : un enjeu pour les systèmes de veille. Dans *Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA'99)*, volume 19, pages 43–51. Terminologies nouvelles, Nantes, France, 1999.

T. Poibeau. 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*, 2001.

T. Poibeau. *Extraction d'information à base de connaissances hybrides.* Thèse de doctorat, Université Paris-Nord, mars 2002.

A. Popescu-Belis. Évaluation numérique de la résolution de la référence : critiques et propositions. *Traitement automatique de la langue*, 40(2) : 117–146, 2000.

U. Quasthoff, C. Biemann et C. Wolff. Named entity learning and verification: Expectation maximization in large corpora. Dans *Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 8–14, 2002.

A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Rapport technique 97-08, Institute for Research in Cognitive Science, Université de Pennsylvanie, 1997.

S. Renals, Y. Gotoh, R. Gaizauskas et M. Stevenson. Baseline ie-ne experiments using the sprach/lasie system. Dans *Proceedings of the DARPA Broadcast News Workshop*, pages 47–50, Herndon, États-Unis, 1999.

A. Rey. *Le Petit Robert des noms propres*, chapitre Préface, pages IX–XIX. Le Robert, 2003.

J. Rey-Debove. Nom propre, lexique et dictionnaires de langue. Dans M. Noailly : rédacteur, *Nom propre et nomination, Actes du colloque de Brest*, pages 107–122, avril 1994.

E. S. Ristad. Maximum entropy modeling toolkit. The Computation and Language E-Print Archive, décembre 1996.

E. Rosch. Natural categories. *Cognitive Psychology*, (4) : 328–350, 1973.

G. Salton et M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

R. E. Schapire et Y. Singer. Improved boosting algorithms using confidence-rated predictions. Dans *Proceedings of the 8th Annual Conference on Computational Learning Theory (COLT'98)*, pages 80–91, 1998.

S. Sekine et Y. Eriguchi. Japanese named entity extraction evaluation - analysis of results. Dans *Proceedings of the 18th International Conference on Computational Linguistices (COLING'2000)*, pages 25–30, Saarbrücken, Allemagne, 2000.

S. Sekine, R. Grishman et H. Shinnou. A decision tree method for finding and classifying names in japanese texts. In Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

S. Sekine et H. Isahara. Irex project overview. Proceedings of the IREX Workshop, 1999.

S. Sekine, K. sudo et C. Nobata. Extended named entity hierarchy. Dans *Proceedings of the third International Conference on Language Resources and Evaluation (LREC'2002)*, volume 5, pages 1818–1824, Îles Canaries, Espagne, 2002.

J. Senellart. Tools for locating noun phrases with finite state transducers. Dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 1212–1217, Montréal, Canada, 1998.

T. Spriet, F. Béchet, M. El-Bèze, C. de Loupy et L. Khouri. Traitement automatique des noms inconnus. Dans *Actes de la troisième conférence sur le Traitement Automatique du Langage Naturel (TALN'96)*, pages 170–179, Marseille, France, 1996.

R. K. Srihari, W. Li, C. Niu et T. Cornell. Infoxtract: An information discovery engine supported by new levels of information extraction. Dans *HLT-NAACL'2003 Workshop on Software Engineering and Architecture of Language Technology Systems*, Edmonton, Canada, 2003.

R. K. Srihari, C. Niu et W. Li. A hybrid approach for named entity and sub-type tagging. Dans *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, États-Unis, 2000.

M. Stevenson et R. Gaizauskas. Using corpus-derived name lists for named entity recognition. Dans *Proceedings of the 6th. Applied Natural Processing Conference*, pages 290–294, 2000.

F. Trouilleux. Identification et classement automatique des noms propres dans des textes français. Dea linguistique, logique et informatique, Université Blaise-Pascal Clermont II, septembre 1997.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

E. M. Voorhees. Overview of the trec 2002 question answering track. Dans E. M. Voorhees et D. K. Harman : rédacteurs, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*. NIST Special Publication: SP 500-251, 2003.

N. Wacholder, Y. Ravin et M. Choi. Disambiguisation of proper names in text. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 202–208, Washington DC, États-Unis, 1997.

T. Wakao, R. Gaizauskas et Y. Wilks. Evaluation of an algorithm for the recognition and classification of proper names. Dans *Proceedings of COLING'96*, volume 1, pages 418–423, Copenhague, Danemark, August 1996.

R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw et J. Palmucci. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19 (2) : 359–382, juin 1993.

F. Wolinski, F. Vichot et B. Dillet. Automatic processing of proper names in texts. Dans *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, pages 23–30, Dublin, Irelande, March 1995.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. Dans *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, États-Unis, 1995.

F. Zabeeh. *What is in a name? : an inquiry into the semantics and pragmatics of proper names.* The Hague: Martinus Nijhoff, 1968.

Table des matières

Annexes

Annexe A

Les lexiques utilisés par Nemesis

Nom	Contenu du lexique	# élém.
adj-géographiques	adjectifs géographiques	108
adj-nationalite	adjectifs de nationalité	448
cles-astres	mots clés d'astres	30
cles-catastrophes	mots clés de catastrophes	30
cles-deserts	mots clés de déserts	6
cles-edifices-ang	mots clés d'édifices en anglais	31
cles-edifices	mots clés d'édifices	290
cles-ensembles	mots clés d'ensembles artistiques	4
cles-entreprises	mots clés d'entreprises	148
cles-entreprises-fin	mots clés de fin d'entreprises	65
cles-etablissements	mots clés d'établissements d'ens. et de rech.	57
cles-evenements	mots clés d'évènements	56
cles-faits	mots clés de faits historiques	18
cles-hydro	mots clés de hydronymes	106
cles-marques	mots clés de marques	6
cles-micro	mots clés de microtoponymes	12
cles-oeuvres-mat	mots clés d'œuvres matérielles	52
cles-oeuvres-abs	mots clés d'œuvres abstraites	30
cles-organisations	mots clés d'organisations	124
cles-org-min	mots clés d'organisation en minuscules	122
cles-oro	mots clés d'oronymes	68
cles-patronymes	mots clés de patronymes	12
cles-periodes	mots clés de périodes historiques	10
cles-rues	mots clés de rues	16
cles-topo	mots clés de toponymes divers	8
cles-topo-moyens	mots clés de villes $<$ toponymes $<$ pays	9
metiers	métiers	162
noms-astres	noms d'astres	15
noms-continents	noms de continents	20
noms-departements	noms de départements	95
noms-deserts	noms de déserts	33
noms-entreprises	noms d'entreprises	596
noms-etats-USA	noms d'états américains	54

suite à la page suivante

Nom	Contenu du lexique	# élém.
noms-hydronymes	noms d'hydronymes	240
noms-institutions	noms d'institutions	33
noms-marques	noms de marques	15 515
noms-medias	noms de médias	110
noms-mers-et-oceans	noms de mers et d'océans	49
noms-monuments	noms de monuments	16
noms-nationalite	noms de nationalité	418
noms-oronymes	noms d'oronymes	89
noms-partis-politiques	noms de partis politiques	25
noms-pays	noms de pays	280
noms-periodes-historiques	noms de périodes historiques	39
noms-provinces-Canada	noms de provinces canadiennes	14
noms-regions	noms de régions françaises	24
noms-regions-	noms de régions plus petites qu'un pays	117
noms-regions+	noms de régions plus grandes qu'un pays	30
noms-villes-etrangeres	noms de villes étrangères	1 795
noms-villes-francaises	noms de villes françaises	46 481
numeros-dynastiques	numeros dynastiques	21
points-cardinaux-maj	points cardinaux avec une majuscule à l'initiale	16
points-cardinaux-min	points cardinaux en minuscules	16
prenoms	prénoms	9 216
sports	noms de sports	7
titres-administratifs	titres administratifs	11
titres-civils	titres civils	10
titres-de-civilite	titres de civilité	31
titres-militaires	titres militaires	45
titres-noblesses	titres de noblesse	23
titres-religieux	titres religieux	14

Tableau A.1 – Liste des lexiques et leur taille

Identification et catégorisation automatiques des entités nommées dans les textes français

Nordine FOUROUR

Résumé

La reconnaissance des entités nommées (EN) reste un problème pour de nombreuses applications de Traitement Automatique des Langues Naturelles (TALN). Si cette reconnaissance est réalisée de façon satisfaisante en extraction d'information, pour des textes journalistiques anglais, elle reste en revanche insuffisante, dès lors qu'elle porte sur des textes français, en particulier lorsque l'on souhaite obtenir une catégorisation fine. Conséquemment à une étude linguistique permettant l'émergence de paramètres définitoires opérationnels liés au concept d' « entité nommée », un état de l'art des différents travaux réalisés dans ce domaine et une étude en corpus portant sur la distribution des EN en fonction de leurs caractéristiques graphiques et référentielles, nous présentons **Nemesis**, un système d'identification et de catégorisation des EN pour le français. Ce système s'appuie sur une typologie des EN évolutive, la plus exhaustive et la plus fine possible. Son architecture logicielle se compose principalement de quatre modules (prétraitements, première reconnaissance des EN, apprentissage, seconde reconnaissance) qui effectuent un traitement séquentiel immédiat des données à partir de textes bruts. La reconnaissance des EN est réalisée en analysant leur structure interne et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles de réécriture. Dans cette version minimale, **Nemesis** atteint environ 90 % en précision et 80 % en rappel. Un gain en rappel ne pouvant faire l'économie de la mise en place d'autres techniques, nous proposons donc différents modules optionnels pour faire face à l'incomplétude des lexiques et au passage à de nouveaux corpus : examen d'un contexte encore plus large et utilisation du *Web* comme source de nouveaux contextes. Nous proposons également une étude des conflits engendrés par les règles de réécritures en vu de l'établissement d'un module de désambiguïsation et d'apprentissage de règles.

Mots-clés : noms propres, entités nommées, corpus, identification, catégorisation, mots déclencheurs, règles de réécriture, évidence interne, évidence externe, apprentissage automatique, révision, surcomposition référentielle