Catégorisation des noms propres : une étude en corpus

Béatrice Daille, Nordine Fourour & Emmanuel Morin*

Cet article présente l'étude préliminaire que nous avons menée en corpus sur la catégorisation graphique et référentielle des noms propres du français dans le but d'établir les spécifications linguistiques d'un système pour leur reconnaissance et leur catégorisation automatiques. Après la présentation des catégorisations des noms propres existantes, nous confrontons celle que nous avons retenue aux noms propres apparaissant dans notre corpus et montrons son adéquation à rendre compte de toute la palette des catégories apparaissant dans les textes même s'il est nécessaire de l'étendre pour quelques catégories. Nous réalisons ensuite une étude numérique sur la présence des noms propres en fonction des catégories graphiques et référentielles. En guise de synthèse et de conclusion, nous résumons les points importants révélés par cette étude et leur intérêt pour la réalisation d'un systéme.

This article presents the feasibility study which we have realised in corpus on the graphic and referential categorisation of the proper names of French. We have the purpose of establishing the linguistics specifications of a system for their recognition and their automatic categorisation. After the presentation of existing categorisations of proper names, we compare the one we retained with the proper names appearing in our corpus and show its adequacy even if it is necessary to extend it to new categories appearing in the texts. We then realise a numerical study on the presence of proper names according to the graphic and semantic categories. Finally, we summarise the significant points revealed by this study and their interest for the realisation of an automatic system.

* IRIN (Institut de recherche en informatique de Nantes), Université de Nantes.

Cahiers de Grammaire 25 (2000), « Sémantique et Corpus », pp. 115-129

1. Introduction

La reconnaissance des noms propres est un problème récurrent dans le traitement automatique de la langue naturelle (TALN), pour l'indexation de textes, la veille technologique ou encore la traduction. Notre objectif à terme est de développer un système de reconnaissance et de catégorisation automatique des noms propres du français s'appuyant sur une typologie générale indépendante du corpus. Cette typologie qui se veut la plus complète possible pourra être utilisée tant dans des tâches d'indexation automatique qu'en traduction. Dans le but d'établir les spécifications linguistiques d'un tel système, nous présentons l'étude préliminaire que nous avons menée sur corpus. Après la définition et la présentation des catégorisations des noms propres existantes pour différentes taches du TALN (cf. section 2), nous confrontons celle que nous avons retenue aux noms propres apparaissant dans notre corpus et montrons son adéquation à rendre compte de toute la palette des catégories apparaissant dans les textes même s'il est nécessaire de l'étendre pour quelques catégories (cf. section 3). Nous réalisons ensuite une étude numérique sur la présence des noms propres en fonction des catégories graphiques et référentielles (cf. section 4). En guise de conclusion, nous résumons les points importants révélés par cette étude qui vont servir de spécifications à notre système de reconnaissance et de catégorisation des noms propres.

2. Identification et catégorisation du nom propre

De nombreux systèmes ont été développés, principalement sur l'anglais, pour reconnaître et catégoriser les noms propres apparaissant dans les textes (cf. Daille et Morin 2000). Ces systèmes, pour la plupart, s'appuient pour leur reconnaissance sur la présence de majuscule et pour leur catégorisation sur des listes de noms propres et sur des listes d'unités lexicales possédant une propriété catégorisatrice. Ainsi, une particule, comme un prénom, si elle apparaît dans le voisinage d'un mot commençant par une majuscule permet de le catégoriser comme personne. Ces unités lexicales sont appelées « évidence interne » par McDonald, D. (1994). Le contexte est généralement ignoré : la limitation à droite du nom propre s'arrête généralement au premier mot plein sans majuscule. Les noms propres identifiés par ces systèmes possèdent donc tous la propriété de commencer par une majuscule et doivent appartenir à une liste fermée de catégories référentielles. Selon les besoins applicatifs, différentes granularités de catégorisations sont adoptées. Nous présentons successivement les catégorisations adoptées en extraction d'information puis celles utilisées en traduction.

2.1. Catégorisation pour l'extraction d'information

Les systèmes d'extraction d'information cherchent à extraire automatiquement à partir d'un corpus spécialisé des relations propres au domaine reflété par le corpus. Les conférences MUC (Message

Understanding Conference) ont pour objectif d'évaluer les différents systèmes d'extraction d'information sur des domaines comme le terrorisme en Amérique Latine (MUC-3 1991 ; MUC-4 1992), la fusion d'entreprises internationales et la fabrication de circuits électroniques (MUC-5 1993) ou les changements de dirigeants des entreprises (MUC-6 1995). 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 (1992) et MUC-5 (1993) 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 telles que le nom du groupe terroriste, le nom de la victime, le type d'agression, la date de l'agression.

Les informations à identifier au cours des conférences MUC sont divisées en trois catégories :

1. **ENAMEX** désigne des noms propres qui peuvent faire référence à des noms de personne, de lieu ou d'organisation.
2. **TIMEX** désigne des expressions temporelles divisées en dates et heures.
3. **NUMEX** désigne des expressions numériques qui font référence à des pourcentages ou à des valeurs monétaires.

La classe ENAMEX (Entity Name Extraction) qui regroupent les noms propres est communément appelée la classe des entités nommées.

Cependant, les entités prises en compte par les systèmes de reconnaissance développés dans le cadre des conférences MUC ne considèrent pas toute la palette des entités intéressantes d'un point de vue extraction d'information. Paik *et al.* (1994) présentent une autre classification des entités, regroupant entités nommées et entités temporelles, réalisée à partir d'une étude du *Wall Street Journal* qui comporte 30 catégories divisées en 9 classes :

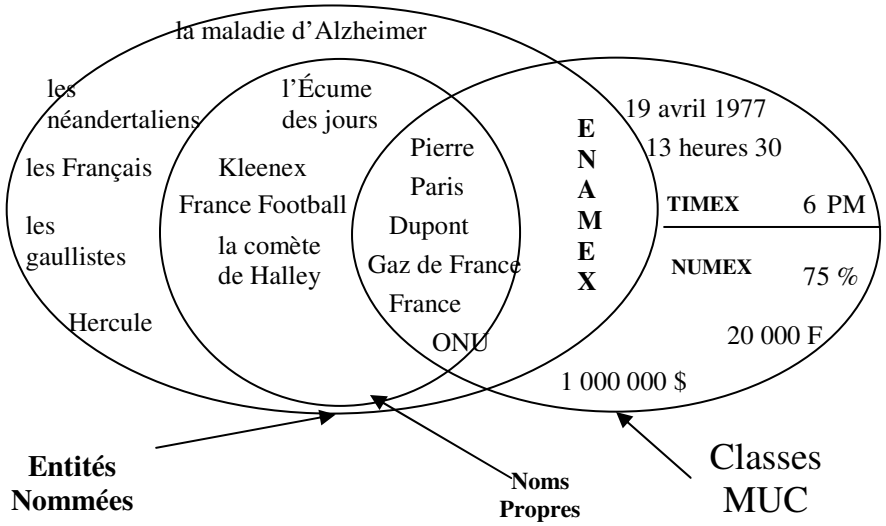
1. **Géographique** : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques.
2. **Affiliation** : religions, nationalités.
3. **Organisation** : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations.
4. **Humain** : personnes, fonctions.
5. **Document** : documents.
6. **Equipement** : logiciels, matériels, machines.

7. **Scientifique** : maladies, drogues, médicaments.
8. **Temporelle** : dates et heures.
9. **Divers** : autres noms d'entités nommées.

Les 8 premières classes couvrent 89 % des entités présentes dans le corpus d'étude de Paik *et al.* (1994).

Wolinski *et al.* (1995) ont défini une catégorisation comprenant une cinquantaine de thèmes pour permettre le classement automatique des dépêches de l'Agence France Presse. Cette catégorisation n'est malheureusement pas détaillée dans leur article.

La notion d'entité nommée inclut donc les noms propres, mais aussi les gentilés, les personnages de légendes, les maladies ou les drogues qui ne sont pas toujours considérés comme des noms propres. La figure ci-dessous illustre la notion d'entité nommée.



2.2. Catégorisation pour la traduction

Dans le cadre de la traduction automatique ou la traduction humaine assistée par ordinateur, une catégorisation précise du nom propre est utile pour décider de son traitement applicatif. Selon sa catégorie référentielle, il devra être traduit, transposé ou non traduit. La seule classification existante pour la traduction, à notre connaissance, est celle réalisée par le linguiste germanophone Bauer (1985). Il énumère ce qui, par convention, constitue un nom propre, il prend en considération des éléments extra-linguistiques propres au référent :

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

Sa typologie est constituée de six classes principales, avec pour chacune, plusieurs catégories :

1. **Anthroponymes** : les personnes individuelles ou les groupes : patronymes, prénoms, pseudonymes, gentilés, hypocoristes, ethnonymes, groupes musicaux modernes, ensembles artistiques et orchestres classiques, partis et organisations.
2. **Toponymes** : les noms de lieux : pays, villes, microtoponymes, hydronymes, oronymes, installations militaires.
3. **Ergonymes** : les objets et les produits manufacturés et par extension les marques, entreprises, établissements d'enseignement et de recherche, titres de livres, de films, de publications, d'œuvre d'art.
4. **Praxonymes** : les faits historiques, les maladies, les événements culturels.
5. **Phénonymes** : les ouragans, les zones de haute et de basse pressions, les astres et les comètes.
6. **Zoonymes** : les noms d'animaux familiers.

Hormis la classe des entités temporelles, il existe de nombreuses similitudes entre la catégorisation de Paik *et al.* (1994) et celle de Bauer (1985). Néanmoins, certaines classes de Bauer (1985) comme les praxonymes ou les phénonymes n'apparaissent ni dans les classes, ni dans les catégories de Paik *et al.* (1994). Inversement, toutes les catégories présentes dans Paik *et al.* (1994) peuvent s'insérer dans les classes de Bauer (1985).

3. Etude référentielle en corpus

Nous avons constitué un corpus regroupant des échantillons de deux périodiques dont les textes sont disponibles sous format électronique : *La Recherche*[2] (17 067 mots) et *Le Monde*[3] (20 866 mots). Ces échantillons ont été extraits aléatoirement. Pour *La Recherche*, cet échantillon consiste en trois textes : le premier est un article de neurobiologie, le deuxième d'astronautique et le dernier traite de la quête de la langue originelle.

---

1 Traduction de : *« Die Einteilung der Eigennamen richtet sich nach der Bewertung der den Namen als Referenten zugrunde liegenden Bestandteile der objektiven Realität. »* faite par Grass T. (1999) lors de la journée d'études de l'ATALA sur Le traitement automatique des noms propres.

2 Corpus de textes *La Recherche* - année 1998 - distribué par ELRA (http://www.icp.inpg.fr/ELRA).

3 Corpus de textes *Le Monde* - année 1987 - European Corpus Initiative (ECI) distribué par ELRA (http://www.icp.inpg.fr/ELRA).

L'échantillon du *Monde* regroupe environ une cinquantaine d'articles abordant divers sujets de société.

Notre objectif est d'établir une catégorisation référentielle stable pour les entités nommées rencontrées dans les textes. Nous adoptons la typologie des noms propres proposée par Bauer (1985) et présentée en section 2.2. Cette catégorisation a été construite indépendamment d'un corpus et apparaît comme l'une des catégorisations existantes les plus complètes. Nous confrontons donc les entités nommées rencontrées dans nos corpus aux classes et catégories de Bauer (1985). Toutes les entités nommées trouvent place dans les 5 premières classes. Une majorité s'inscrivent dans les catégories. Néanmoins, il est nécessaire d'étendre certaines catégories et d'en créer de nouvelles. Nous détaillons ci-dessous les catégories étendues ou nouvellement créées en fonction des classes.

Les anthroponymes

Les noms d'institutions, comme *le Parlement*, *la Banque de France,* sont ajoutés à la catégorie contenant déjà les noms de partis et autres organisations. Aux ethnonymes sont ajoutés les noms regroupant les personnes appartenant à une même période historique comme *les Néandertaliens*, un même mouvement idéologique comme *les Communistes*.
La catégorie des ensembles artistiques est étendue aux :

- noms de média comme la télévision *Canal+,* la radio *France Info*, la presse écrite *France Football* ;
- équipes sportives : *le Paris-SG*, *l'équipe de France de football*.

Il faut noter que, pour les médias, il ne pouvait être fait assimilation avec les noms d'organisations. En effet, si certains noms de média correspondent directement à un nom d'organisation, d'autres comme le nom d'un journal, par exemple *l'Equipe Magazine*, peuvent être différents des sociétés qui les éditent ; ces deux concepts doivent donc être différenciés. D'où la nécessité d'ajouter les noms de médias aux ensembles artistiques et de ne pas simplement les réduire à des organisations.

Les toponymes

Pour les toponymes, la tâche est plus difficile notamment pour les zones géographiques. En effet, Bauer (1985) limite la catégorisation aux villes, pays et microtoponymes. Il est donc nécessaire d'ajouter des catégories présentes chez Paik *et al.* (1994) pour couvrir toutes les taille de zones géographiques. Nous ajoutons les zones plus vastes que les pays : les continents comme *l'Europ*e, ainsi que celles dont la taille est comprise entre ville et pays comme les régions administratives ou non *l'Ile de France*, les départements *la Savoie*, les états des pays fédéraux *la Californie*, les provinces *la Corse*.

Les ergonymes

La catégorie que nous étendons le plus est celle qui regroupe les œuvres. Bauer (1985) incluait les titres de livres, les noms de publications et d'œuvres d'art. Le problème s'est alors posé de catégoriser un grand nombre d'entités nommées qui avaient pour seul point commun de n'être que des productions intellectuelles, généralement créées par une seule personne dont le patronyme a donné nom à cette production comme les projets, les plans, les théorèmes, les lois, les allégories comme *le mythe de la Caverne*. Nous avons donc décidé de les regrouper dans une seule et même catégorie : œuvres intellectuelles.

Les praxonymes

La dernière extension nécessaire concerne les événements culturels. Un meeting politique comme *le congrès de Rennes*, une compétition sportive comme *la Coupe du Monde de football*, un salon comme *le salon de l'Automobile*, une foire comme *la foire de Paris*, sont des événements mais que l'on peut difficilement qualifier de culturels. Ces derniers types d'entités nommées n'ont donc pas été inclus dans la typologie de Bauer (1985). Nous étendons donc la catégorie événements culturels à la catégorie événements culturels, sportifs, politiques, commerciaux, etc. Nous avons ajouté une nouvelle catégorie regroupant les différentes périodes historiques comme *le Paléolithique* ou *la Renaissance*.

Les zoonymes

Les zoonymes chez Bauer (1985) formaient une classe à part. Nous avons décidé de l'inclure en tant que catégorie dans la classe des anthroponymes. Nous n'avons rencontré aucun représentant de cette catégorie dans nos corpus.

Après l'examen du premier quart des entités nommées de chacun des échantillons de notre corpus, notre catégorisation a atteint une très haute stabilité. Toutes les autres entités nommées identifiées dans le reste du corpus trouvent leur place dans notre typologie. Le tableau 1 illustre à l'aide d'exemples les classes et catégories instanciées dans notre corpus. Les catégories étendues ou créées apparaissent précédées d'un astérisque.

ANTHROPONYMES	
Patronymes	*Dupont, Durant*
Prénoms	*Alexandre, Jean-Paul*
Ethnonymes	*L'Italien, les Français*
*Partis et autres organisations	*Le PCF, l'ONU, l'Union européenne*
*Ensembles artistiques, groupes musicaux et troupes de théâtre	*Les Toten Hosen, l'orchestre philharmonique de New York*
Pseudonymes	*L'Ange vert*
Zoonymes	*Médor*
TOPONYMES	
*Toponymes **>** Pays	*Europe*
Pays	*France, Sahara occidental*
*Pays **<** Toponymes **>** Villes	*l'Ile de France*
Villes	*Paris, Belo Horizonte*
Microtoponymes	*Le Quartier Latin, Prenzlauerberg*
Hydronymes	*La Manche, la Seine, le lac Ontario*
Oronymes	*Les Andes, les Alpes*
Rues	*La rue de la Paix, le Faubourg Saint Honoré*
Déserts	*Le Sahara, le désert de Gobi*
Édifices	*La Maison Blanche, la gare Montparnasse*
ERGONYMES	
Sites de production	*Renault Wilword*
Marques de produits	*Coca, Kleenex, Scotch*
Entreprises industrielles	*Microsoft Corporation, Sud-Marine industrie*
Coopératives	*Semences de Provence*
Établissement d'enseignement et de recherche	*Université de Nantes*
Installations militaires	*la ligne Maginot*
*Œuvres intellectuelles	*Matrix, l'Écume des jours*
PRAXONYMES	
Faits historiques	*La Guerre de Cent Ans*
Maladies	*La maladie d'Alzheimer*
*Événements culturels, sportifs, commerciaux, etc.	*Le Festival du film de Berlin*
*Période historique	*le Paléolitique*
PHENONYMES	
Catastrophes naturelles	*Le cyclone Mitch*
Astres et comètes	*La comète de Halley*

Tableau 1 – Typologie des noms propres selon Bauer (1985) étendue

Il reste cependant un certain nombre de syntagmes (une vingtaine rencontrée dans le corpus *Le Monde*) pour lesquels il est difficile de trancher sur leur statut ou non d'entité nommée. Pour *le peloton voltigeur motocycliste (PVM)*, nous avons répondu affirmativement car il se place naturellement dans la classe des anthroponymes, catégorie des organisations. En revanche, nous avons décidé de ne pas considérer comme entité nommée *le système Ariane pour passagers auxiliaires (ASAP)*. Ce syntagme pourrait s'inscrire dans la classe des ergonymes, catégorie œuvres, si son auteur était connu ou s'il était issu d'une production intellectuelle. De même, nous ne considérons pas *le Conseiller régional*, à l'inverse *du Président de la République*, comme une entité nommée car il est difficile d'identifier quelle est la personne à laquelle il est fait référence par ce titre.

Tous les syntagmes que nous considérons comme des entités nommées s'intègrent aisément dans l'une des classes et l'une des catégories introduites. Le problème porte plus sur la composition de certaines d'entre-elles. Ainsi avec *le festival de l'Université de Natal*, faut-il considérer *le festival de l'Université de Natal* ou *l'Université de Natal* ? De même, *avec le forum FR3-RMC*, faut-il considérer *le forum FR3-RMC* ou uniquement *FR3* et *RMC* ? Cette propriété de composition des entités nommées en terme de classes et catégories référentielles nécessitera d'être plus étudiée de manière à permettre l'introduction éventuelle de granularité lors de leur identification et catégorisation automatiques.

Par ailleurs, certaines catégories qui ont été étendues comme par exemple la catégorie œuvre pourront éventuellement elles-mêmes être sous-catégorisées en différents types si nécessaire. Par exemple, si nous examinons les entités nommées de la terminologie médicale recensées par Bodenreider, O. et Zweigenbaum, P. (2000), celles-ci s'inscrivent aisément dans notre typologie :

1. Les maladies, syndromes, etc. (*maladie de Parkinson*, *souffle de Graham Steel*, *anémie de Cooley*…) sont catégorisés comme praxonymes, catégorie maladies.
2. Les partis du corps auxquelles un scientifique célèbre a laissé son nom (*tubercule de Lisfranc*, *ganglion de Gasser*…) sont classées dans les ergonymes, catégorie œuvres du fait que leur découverte résulte d'un travail scientifique.
3. Les procédés auxquels l'inventeur a laissé son nom sont aussi catégorisés comme ergonymes, catégorie œuvres (*technique de Kenneth Jones*, *colostomy de Hartmann*, *manœuvre de Heimlich*…).

La catégorie œuvre regroupe donc à la fois des partie du corps et des procédés qu'il pourrait être intéressant de distinguer.

4. Analyse quantitative

Les résultats quantitatifs que nous présentons ont été obtenus manuellement. Toutes les entités nommées ont été identifiées, catégorisées et comptées. Nous présentons successivement les résultats d'une étude portant sur la représentation numérique des différentes graphies des entités nommées, puis ceux de l'étude référentielle. Nous concluons cette étude par quelques remarques sur les liens mis à jour entre catégories graphiques et référentielles.

4.1. Classification graphique

La distinction des entités nommées suivant des critères graphiques est intéressante dans une optique de reconnaissance automatique. Suivant la graphie, l'identification et la classification des entités nommées entraîneront des traitements différents. Nous distinguons les catégories suivantes inspirées de la terminologie de Jonasson (1994) :

- **EN pures simples** : entités nommées constituées d'une seule unité lexicale commençant par une majuscule comme *France*, *Aristote* ;
- **EN pures complexes** : entités nommées constituées de plusieurs unités lexicales commençant par une majuscule comme *Conflans Saint-Honorine*. Nous introduisons la sous-catégorie Prénom Nom : entités nommées constituées d'un prénom (ou plusieurs prénoms) et d'une unité lexicale commençant par une majuscule référant à un nom de personne comme *Paul Valéry* ;
- **EN faiblement mixtes** : entités nommées constituées de plusieurs mots commençant par une majuscule et contenant des mots de liaison en minuscule comme *le Jardin des Plantes*. Cette liste de mots de liaison est fermée et comprend les articles, des prépositions et des conjonctions de coordination, etc. ;
- **EN mixtes** : entités nommées constituées de plusieurs unités lexicales dont au moins une commence par une majuscule comme le *Comité international de la Croix-Rouge*, *le Mouvement contre le racisme et pour l'amitié entre les peuples* ;
- **Sigles** : entités nommées constituées d'une seule unité lexicale comportant plus d'une majuscule et dont chaque lettre en majuscule réfère elle-même à une autre unité lexicale comme *USA*. Il est à noter que les entités nommées appartenant à cette catégorie, qu'il est important de distinguer au niveau graphique, réfèrent à des EN pures complexes et à des EN mixtes (faibles ou non).

Le tableau 2 résume le nombre d'entités nommées rencontrées dans notre corpus classées suivant leur caractéristique graphique.

	La Recherche	Le Monde
EN pures simples	145	313
EN pures complexes	25	89
Prénom Nom	68	299
EN faiblement mixtes	21	35
EN mixtes	44	144
Sigles	15	127
Total	318	1007
Proportion d'EN dans l'échantillon	2,2 %	7,2 %

Tableau 2 – Présence des EN en fonction de leurs caractéristiques graphiques

Nous pouvons constater, tout d'abord, qu'il y a plus d'entités nommées dans l'échantillon du corpus Le Monde que dans celui de La Recherche (respectivement 7,2 % et 2,2 %) et ceci toutes catégories graphiques confondues. Les EN pures simples sont les plus présentes dans les deux corpus (46 % des entités nommées pour La Recherche et 31 % Le Monde). Les EN pures complexes sont moins présentes que les simples (7,8 % et 8,8 %). Les EN faiblement mixtes sont un peu moins présentes que les EN pures complexes (6,6 % et 3,5 %). Les EN mixtes sont loin d'être négligeables (13,8 % et 14,3 %). 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 %).

4.2. Classification référentielle

Le tableau 3 résume le nombre d'entités nommées rencontrées dans notre corpus classées en fonction de notre typologie des noms propres établie à partir de celle de Bauer (1985).

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 : en rencontrant l'entité complète le *Français Michel Platini*, on ajoutera une unité aux ethnonymes pour *Français*, une aux prénoms pour *Michel*, ainsi qu'une aux Patronymes pour *Platini*.

	La Recherche	Le Monde
ANTHROPONYMES	**194**	**1066**
Patronymes	97	437
Prénoms	66	310
Ethonymes	15	37
Organisations	16	194
Ensembles artistiques	0	87
Pseudonymes	0	1
Zoonymes	0	0
TOPONYMES	**107**	**270**
Toponymes > Pays	53	13
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
Edifices	0	14
ERGONYMES	**64**	**93**
Sites de production	0	0
Marques et produits	31	37
Entreprises industrielles	0	4
Coopératives	0	0
Etablissements d'enseignement et de recherche	27	7
Installations militaires	0	1
Œuvres intellectuelles	6	44
PRAXONYMES	**3**	**16**
Faits historiques	0	0
Maladies	0	0
Evénements culturels, sportifs, politiques	0	15
Périodes historiques	3	1
PHENONYMES	**5**	**0**
Catastrophes naturelles	0	0
Astres et comètes	5	0
TOTAL	**373**	**1445**

Tableau 3 – Présence des entités nommées en fonction de leur catégorie référentielle

Certaines classes et catégories référentielles ne sont que peu ou pas représentées dans ces deux échantillons de corpus (zoonymes, oronymes,

sites de production, coopératives, maladies, faits historiques, catastrophes naturelles, etc.). La nature de nos échantillons textuels influe sur la représentativité des classes et catégories des entités nommées rencontrées.

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 environ 77 % de toutes les entités nommées que nous avons identifiées manuellement.

5 . Synthèse et conclusion

Cette étude en corpus permet de dégager les points suivants pour la réalisation d'un système de reconnaissance et de catégorisation des entités nommées.

Catégorisation graphique

Les classes graphiques d'entités nommées mettant en jeu des unités complexes représentent plus de 50 % des entités nommées présentent dans notre corpus. Il est donc essentiel qu'un système d'identification et de catégorisation des entités nommées les prennent en compte.

L'identification des EN pures complexes, comme d'ailleurs les EN pures et les sigles, est triviale dans la plus part des cas puisqu'il s'agit de vérifier la seule présence de majuscules. Elle pose problème en début de phrase et dans les titres. L'identification des EN faiblement mixtes nécessite en plus l'utilisation de listes prédéfinies de mots de liaison. Ces mots fonctionnels peuvent introduire des problèmes de résolution de l'attachement prépositionnel et de portée de la coordination. La reconnaissance des EN mixte nécessite au minimum un étiquetage grammatical pour établir les limites du syntagme nominal dans son expansion à droite. Néanmoins, même avec la connaissance des parties du discours, l'identification des EN mixtes se heurte aux problèmes de la résolution de l'attachement prépositionnel, de la portée de la coordination mais aussi de la modification. L'adjectif modifiant un nom propre comme *fédéral* dans *Allemagne fédérale* constitue une EN mixte, ce qui n'est pas le cas avec l'adjectif *lointain* dans *Chine lointaine*. Certaines de ces EN mixtes peuvent être identifiées grâce à la présence d'un sigle. Il reste à évaluer pour les EN mixtes non accompagnées de leur sigle l'intérêt de disposer d'informations grammaticales permettant de décrire les expansions droites potentielles de chaque classe ou catégorie.

Catégorisation référentielle

La classification originale de Bauer (1985) pour catégoriser les entités nommées est suffisamment générale pour être utilisable sur différents types de corpus. Cependant, pour une application particulière spécifique à un domaine, il sera nécessaire d'étendre ou de créer de nouvelles catégories ou même d'en distinguer des types différents. De plus, nous avons vu qu'il était parfois difficile de trancher entre plusieurs niveaux référentiels pour une entité nommée. Un système d'identification et de catégorisation des entités nommées indépendant d'une application et d'un type de corpus devra donc permettre à l'utilisateur de pouvoir spécifier le degré de catégorisation qu'il désire et aussi de lui permettre de créer de nouvelles sous-catégories. La catégorisation automatique s'effectuera à l'aide de patrons qui s'appuient sur des lexiques spécifiques à chaque classe ou catégorie. Ces lexiques seront de deux sortes :

- Entités nommées connues ;
- Unités lexicales à fonction catégorisatrice pouvant ou non faire partie de l'entité nommée selon sa catégorie graphique.

Ces lexiques devront être mis à jour automatiquement à chaque application des règles du système.

Liens entre catégorisation graphique et référentielle

Des remarques peuvent être émises sur les liens entre les catégories graphiques et référentielles. Les patronymes et les prénoms forment des EN complexes appartenant à la catégorie Prénom Nom. Les ethnonymes, l'ensemble des toponymes, les maladies, les périodes historiques, les catastrophes naturelles, les astres et les comètes sont essentiellement des EN pures simples (*Français*, *Parisien*, *la France*, *les Alpes*, *la Renaissance, le Paléolithique*, *le cyclone Hugo*). Cependant, les toponymes, par exemple, peuvent être des EN pures complexes ou des EN mixtes (*Europe de l'Ouest*, *l'Océan indien*), voire même des sigles (*RFA*, *URSS*, *USA*). Les organisations sont composées de Sigles, d'EN pures complexes, d'EN faiblement mixtes et d'EN mixtes (*la CEE*, *la Communauté économique Européenne*, *Association of Ceramic Industry*, *le Centre national des lettres*). Ces trois dernières catégories graphiques 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 événements. Ces liens entre graphie et référence pourront être exprimés sous forme de règles pondérées.

Nous avons présenté une étude préliminaire en corpus sur la catégorisation graphique et référentielle des noms propres. Cette étude nous a permis d'établir un classement graphique et une catégorisation référentielle des noms propres. Ces deux catégorisations, essentielles pour la reconnaissance et la catégorisation automatiques des noms propres, n'ont

jamais été clairement établies par les travaux précédents (cf. Daille et Morin 2000). Elle nous a aussi permis de montrer la nécessité d'une catégorisation évolutive en fonction du corpus et la prise en compte d'une granularité dans la catégorisation.

Références bibliographiques

Bauer, G. (1985), *Namenkunde des Deuschen*, Germanistische Lehrbuch-sammlung Band 21.

Bodenreider, O. & Zweigenbaum, P. (2000), « Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles », *Traitement Automatique des Langues (T.A.L),* 41.3, Paris, Hermès.

Daille, B. & Morin, E. (2000), « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », *Traitement Automatique des Langues (T.A.L),* 41.3, Paris, Hermès.

Jonasson, K. (1994), *Le Nom Propre, Constructions et interprétations*, Duculot, Champs linguistiques.

McDonald, D. (1994), « Internal and External Evidence in the Identification and Semantic Categorization of Proper Nouns » , in B. Boguraev & J. Pustejovsky (eds), *Corpus Processing for Lexical Aquisitio*, MIT Press, chap. 2.

MUC-3 (1991), *Proceedings of the 3rd Message Understanding Conference*, San Diego, CA, Morgan Kauffmann.

MUC-4 (1992), *Proceedings of the 4th Message Understanding Conference*, San Mateo, CA, Morgan Kauffmann.

MUC-5 (1993), *Proceedings of the 5th Message Understanding Conference*, San Mateo, CA, Morgan Kauffmann.

MUC-6 (1995), *Proceedings of the 6th Message Understanding Conference*, Columbia, Maryland, Morgan Kauffmann.

Paik, W., Liddy, E.D., Yu, E. & McKenna, M. (1994), « Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval », in , B. Boguraev, & J. Pustejovsly (eds), *Corpus Processing for Lexical Acquisition*, MIT Press, chap. 4.

Wolinski, F., Vichot, F. & Dillet, B. (1995), « Automatic Processing of Proper Names in Texts », in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, Dublin, Ireland, pp. 23-30.