View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Érudit

Article

« Apport du Web dans la reconnaissance des entités nommées »

Nordine Fourour et Emmanuel Morin

Revue québécoise de linguistique, vol. 32, n° 1, 2003, p. 41-60.

Pour citer cet article, utiliser l'information suivante :

URI: http://id.erudit.org/iderudit/012243ar

DOI: 10.7202/012243ar

Note : les règles d'écriture des références bibliographiques peuvent varier selon les différents domaines du savoir.

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter à l'URI https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche. *Érudit* offre des services d'édition numérique de documents scientifiques depuis 1998.

Pour communiquer avec les responsables d'Érudit : info@erudit.org

Document téléchargé le 9 février 2017 06:23

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal
Reproduction interdite sans autorisation de l'éditeur

APPORT DU WEB DANS LA RECONNAISSANCE DES ENTITÉS NOMMÉES

Nordine Fourour
Emmanuel Morin
Université de Nantes

1. Introduction

La reconnaissance des entités nommées pour le français est un problème qui se pose dans les différents domaines du traitement automatique de la langue naturelle (TALN) :veille technologique, indexation de textes ou traduction (Daille et Morin 2000). Cette reconnaissance a été convenablement réalisée en extraction d'information (EI) pour des textes journalistiques anglais (précision et rappel supérieurs à 90 %) (MUC-7 1998).

Le terme entité nommée (EN) regroupe les noms propres communément reconnus comme tels (la classe ENAMEX des conférences MUC), ainsi qu'un certain nombre d'entités qui ne sont pas toujours considérées comme noms propres : les noms collectifs (*les Français*, *les néandertaliens*, etc.), les maladies ou encore les noms de personnages mythiques ou fictifs (*Hercule*, *Colombo*, etc.).

En EI, les entités nommées sont généralement séparées en quatre classes : personnes, lieux, organisations et expressions temporelles. Bien que cette catégorisation regroupe une grande partie des entités nommées présentes dans les textes journalistiques, elle est limitée et inadaptée à la traduction, car elle reste insuffisamment exhaustive et peu fine. C'est pour cela que nous proposons une typologie générale la plus complète possible. Cette typologie, indépendante du domaine, est vérifiée expérimentalement par une étude de corpus.

Pour le français, comme pour l'anglais (Wacholder et coll. 1997), cette reconnaissance se heurte aux problèmes d'ambiguïté liés aux majuscules en début de phrase (Mikheev 1999) et à la localisation des limites à droite du nom propre : modification adjectivale et attachement des prépositions et des coordinations, possibilité que certaines entités nommées soient composées en quasi

totalité de mots en minuscules. Pour aider à cette délimitation, nous proposons des critères graphiques qui sont également validés sur corpus. En plus de cette ambiguïté se posent les problèmes de surcomposition : une EN complexe peut contenir une EN d'une autre catégorie référentielle (p. ex. *Guerre d'Algérie*, *Université de Nantes)*.

La reconnaissance des entités nommées est donc une tâche complexe, qui nécessite le recours à de nombreuses ressources lexicales (liste de prénoms, d'entreprises, de régions, de fleuves, de groupes de musique, etc.). Les lexiques que nous exploitons sont le plus souvent incomplets. En effet, il est difficile de créer des lexiques exhaustifs : recenser l'ensemble des cours d'eau de la planète serait une tâche presque impossible. D'autre part, la maintenance de ces lexiques est une activité très lourde, comme pour les noms d'entreprises, par exemple. Face à ces différentes difficultés, nous étudions de quelles façons le Web peut être exploité comme source de connaissance pour le TALN, à l'instar des travaux de Grefenstette 1999 pour la traduction à bases d'exemples ou de Jacquemin et Bush 2000 pour la collecte d'entités nommées. À la différence de Jacquemin et Bush 2000, nous ne cherchons pas à collecter des identités nommées à partir du Web, mais uniquement à catégoriser des entités nommées déjà identifiées par notre système. En ce sens, nous considérons le Web comme une source de connaissance externe apte à proposer différents exemples linguistiques de l'usage d'une identité nommée.

Après la présentation des catégorisations graphique et référentielle proposées (section 2), nous décrivons le système permettant l'identification et la catégorisation des entités nommées (section 3). Puis, nous évaluons les performances de ce système (section 4) et pointons les limites de ce dernier (section 5). Ensuite, nous proposons un couplage de notre système avec un module de reconnaissance à partir du Web (section 6), ainsi qu'une évaluation de l'impact de ce dernier dans Némésis (section 7). Enfin, nous présentons nos conclusions et les perspectives qu'ouvre notre travail (section 8).

2. Catégorisations

Nous présentons successivement les résultats d'une étude portant sur la distribution des différentes catégories référentielles des entités nommées, puis ceux de l'étude graphique. Les résultats quantitatifs que nous présentons ont été obtenus manuellement. Toutes les entités nommées ont été identifiées, catégorisées et comptées. Ces études ont été réalisées sur un corpus regroupant des échantillons de deux périodiques dont les textes sont disponibles sous format

électronique : *La Recherche*[1] (17 067 mots) et *Le Monde*[2] (20 866 mots). Nous concluons cette étude par quelques remarques sur les liens mis au jour entre catégories graphiques et référentielles.

2.1 Catégorisation référentielle

Notre objectif est d'établir une catégorisation référentielle fine et stable pour les entités nommées : les nouvelles EN rencontrées dans les textes devront y trouver place. Cependant, cette typologie pourra être évolutive : ajout d'un niveau de profondeur supplémentaire pour raffiner des catégories qui s'avèreraient trop vastes. Dans le cadre de la traduction automatique ou humaine assistée par ordinateur, une catégorisation précise du nom propre est utile pour décider de son traitement. Selon sa catégorie référentielle, il devra être traduit, transposé ou non traduit.

Parmi les catégorisations des entités nommées, la plus couramment répandue est celle utilisée pour les conférences MUC. Les informations à identifier au cours de ces conférences sont divisées en trois catégories :

1° **ENAMEX** : noms propres référant à des noms de personnes, lieux ou organisations;
2° **TIMEX** : expressions temporelles divisées en dates et heures;
3° **NUMEX** : expressions numériques (pourcentages ou des valeurs monétaires).

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 en TALN : les noms de médias, d'évènements, de documents, etc. n'y sont pas représentés. Paik et coll. 1996 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, dont les 8 premières couvrent 89 % des EN du corpus d'étude de Paik et coll. 1996 :

Géographique : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques;
Appartenance : religions, nationalités;
Organisation : entreprises, types d'entreprises, institutions, organisations;
Humain : personnes, fonctions;
Document : documents;
Équipement : logiciels, matériels, machines;

1 Corpus *La Recherche* – année 1998 – distribué par ELRA (http://www.icp.inpg.fr/ELRA).

2 Corpus *Le Monde* – année 1997 – European Corpus Initiative (ECI) distribué par ELRA.

Scientifique : maladies, drogues, médicaments;
Temporelle : dates et heures;
Divers : autres noms d'entités nommées.

Quant à eux, Wolinski et coll. 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 seule classification existante pour la traduction, à notre connaissance, est celle réalisée par Grass 2000, inspirée de la typologie du linguiste germanophone Bauer 1985. Il énumère ce qui, par convention, constitue un nom propre et prend en considération des éléments extralinguistiques propres au référent. Cette typologie comporte cinq classes :

Anthroponymes : noms de personnes individuelles et de groupes;
Toponymes : noms de lieux;
Ergonymes : objets et produits manufacturés;
Praxonymes : faits historiques, maladies, évènements culturels;
Phénonymes : ouragans, zones de pressions, astres et comètes.

Hormis la classe des entités temporelles, il existe de nombreuses similitudes entre la catégorisation de Paik et coll. 1996 et celle de Grass 2000. Néanmoins, certaines classes de Grass 2000, comme les praxonymes ou les phénonymes, n'apparaissent pas chez Paik et coll. 1996. Inversement, toutes les classes présentes dans Paik et coll. 1996 peuvent s'insérer dans les classes de la typologie de Grass 2000. De plus, cette dernière a été construite indépendamment d'un corpus et apparaît comme l'une des catégorisations existantes les plus complètes.

Nous avons donc adopté la typologie proposée par Grass 2000, comme base pour notre catégorisation. Toutes les entités nommées rencontrées dans nos corpus trouvent place dans les cinq classes, et une majorité s'inscrit dans les catégories. Néanmoins, il est nécessaire d'étendre certaines catégories et d'en créer de nouvelles. La distribution des entités nommées en fonction de leur catégorie dans la typologie ainsi obtenue est présentée au Tableau 1, où les catégories étendues ou créées apparaissent précédées d'un astérisque.

Tableau 1
Distribution des entités nommées en fonction de leur catégorie référentielle

	La Recherche		*Le Monde*	
	# Occ.	**Proportion**	**# Occ.**	**Proportion**
ANTHROPONYMES	**194**	**52,0 %**	**1 066**	**73,8 %**
Patronymes	97		437	
Prénoms	66		310	
Ethnonymes	15		37	
* Organisations	16		194	
* Ensembles artistiques	0		87	
Pseudonymes	0		1	
* Zoonymes	0		0	
TOPONYMES	**107**	**28,7 %**	**271**	**18,7 %**
* Toponymes > Pays	53		17	
Pays	22		73	
* Pays > Toponymes > Villes	17		33	
Villes	10		108	
Microtoponymes	0		16	
Hydronymes	4		9	
Oronymes	0		0	
Rues	0		4	
Déserts	1		0	
Édifices	0		15	
ERGONYMES	**64**	**17,2 %**	**92**	**6,4 %**
Entreprises industrielles	0		4	
Marques et produits	31		37	
Établissements d'enseignement et de recherche	27		7	
* Oeuvres	6		44	
PRAXONYMES	**3**	**0,8 %**	**16**	**1,1 %**
Faits historiques	0		0	
* Évènements culturels, sportifs, politiques	0		15	
* Périodes historiques	3		1	
PHÉNONYMES	**5**	**1,3 %**	**0**	**0 %**
Catastrophes naturelles	0		0	
Astres et comètes	0		0	
TOTAL	**373**		**1 445**	

2.2 Catégorisation graphique

La distinction des entités nommées suivant des critères graphiques est intéressante dans une optique de reconnaissance automatique. En effet, selon 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 : constituées d'une seule unité lexicale commençant par une majuscule, comme *France* ou *Aristote*;

EN pures complexes : constituées de plusieurs unités lexicales commençant par une majuscule, comme *Conflans Sainte-Honorine*. Nous introduisons la sous-catégorie Prénom Nom : entités nommées constituées d'un ou plusieurs prénoms et d'une unité lexicale commençant par une majuscule référant à un nom de personne, comme *Paul Valéry*;

EN faiblement mixtes : constituées de plusieurs mots commençant par une majuscule et contenant des mots de liaison en minuscules, comme *Jardin des Plantes*. Cette liste de mots de liaison est fermée et comprend des prépositions, des articles, etc.;

EN mixtes : constituées de plusieurs unités lexicales 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*;

Sigles : entités nommées constituées d'une seule unité lexicale comportant plusieurs majuscules qui réfèrent elles-mêmes à une autre unité lexicale, comme *USA*. 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).

Tableau 2
Présence d'entités nommées en fonction de leurs caractéristiques graphiques

	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**	**1 007**

Le Tableau 2 présente les résultats de l'étude quantitative de la présence des entités nommées selon leurs caractéristiques graphiques. Il montre qu'il y a plus d'entités nommées dans l'échantillon du corpus du *Monde* que dans celui de *La Recherche* (resp. 1 007 et 318) 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 % pour *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 des sigles est moins importante dans l'échantillon de *La Recherche* que dans celui du *Monde* (4,7 % et 12,6 %).

Il est intéressant d'établir les liens entre catégories référentielles et graphiques, afin de concevoir les traitements à effectuer (lexiques, règles, etc.) pour chaque classe. Les patronymes et les prénoms forment des EN complexes appartenant à la sous-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 (*Parisien*, *France*, *Alpes*, *Renaissance*, le cyclone *Hugo*). Cependant, les toponymes, par exemple, peuvent être des EN pures complexes ou des EN mixtes (*Europe de l'ouest*, *Océan Indien*), voire même des sigles (*RFA*, *URSS*, USA). Les organisations sont composées de sigles, d'EN pures complexes, faiblement mixtes et mixtes (*CEE*, *Communauté économique Européenne*, *Association of Ceramic Industry).* 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. Ces liens pourront être exprimés sous forme de règles pondérées.

Ces deux études ainsi que notre typologie vont servir de base à la mise en place de notre système de reconnaissance des entités nommées.

3. Architecture logicielle

Némésis, élaboré conséquemment à cette étude, est un système qui permet l'identification des bornes des EN, ainsi que leur catégorisation selon cette typologie. Son architecture, présentée à la Figure 1, se compose de quatre modules qui effectuent un traitement séquentiel immédiat des données : prétraitement lexical, première reconnaissance des entités nommées, apprentissage et seconde reconnaissance des entités nommées.

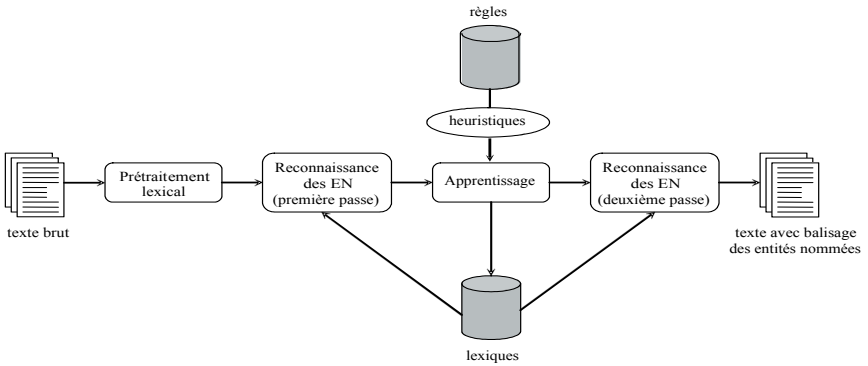

Fig. 1 : Architecture du système Némésis

3.1 Prétraitement lexical

Le prétraitement lexical s'effectue en deux étapes : segmentation du texte en phrases et en mots, puis association des sigles et de leur forme étendue. Cette dernière phase est réalisée uniquement en étudiant les structures locales. Par exemple, lorsqu'un sigle apparaît pour la première fois dans un texte, il est souvent accompagné de sa forme étendue (p. ex. *la FFF (Fédération française de football)*). L'association des sigles et de leur forme étendue permet donc d'identifier des EN mixtes et de catégoriser des sigles. Parmi les différents dispositifs de reconnaissance des noms propres que nous avons étudiés, seuls Wolinski et coll. 1995 et Wacholder et coll. 1997 utilisent l'association entre les sigles et leur forme étendue, mais uniquement pour les coréférences en ce qui concerne les derniers.

3.1.1 Projection des lexiques

La projection a lieu en trois étapes :

1° passage du texte en fichier inverse (Salton et McGill 1983) pour limiter les accès disque;
2° projection : les étiquettes sémantiques liées aux lexiques sont associées aux différentes formes du texte;
3° étiquetage des mots commençant par une majuscule et absents des lexiques par NP.

Il a été démontré que l'utilisation de lexiques spécialisés était la base de tout système de reconnaissance des noms propres (McDonald 1994, Wakao et

coll. 1996). Nos lexiques ont été construits soit manuellement, soit automatiquement, en exploitant des ressources textuelles (pages Web, etc.). Les éléments composant ces lexiques peuvent tenir un ou plusieurs rôles :

- `EN` : l'élément est une entité nommée connue (*OMS*, *Alexandre*, *Canal+*);
- `mot déclencheur` : l'élément fait partie de l'entité nommée (*Fédération*, *Boulevard*);
- `contexte` : l'élément appartient au contexte gauche immédiat de l'EN, mais ne fait pas partie de celle-ci (*philosophe*, *français*);
- `fin d'EN` : l'élément est la dernière forme composant l'entité nommée (*football*, *régional*);
- `élément d'EN` : il s'agit de tous les éléments lexicaux pouvant faire partie de l'EN, mais sans en permettre la délimitation ou la catégorisation.

Nous pouvons donc assigner des rôles à nos lexiques en fonction des catégories référentielles dans lesquelles ils sont utilisés. Cette assignation peut être visualisée sous deux angles : en prenant comme point central soit une catégorie référentielle (p.ex. la reconnaissance des patronymes utilise les éléments du lexique des noms de pays comme `fin d'EN` ou `élément d'EN` (cf. Figure 3), soit un lexique (les éléments du lexique des noms de pays sont utilisés uniquement comme `fin d'EN` pour la reconnaissance des ensembles artistiques (cf. Figure 2).

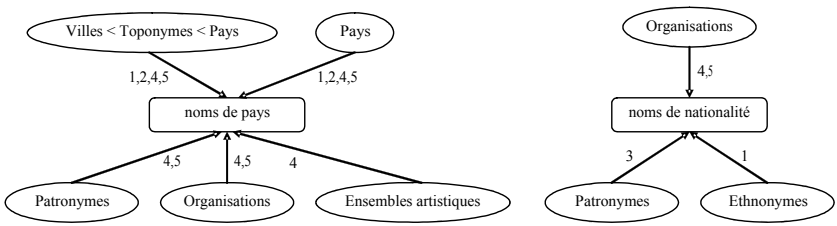

Fig. 2 : Deux lexiques et leur rôle selon les catégories référentielles

Chaque catégorie référentielle utilise un nombre réduit de lexiques (cf. Tableau 3). Sur nos 45 lexiques, regroupant 60 761 éléments, seuls les patronymes et les organisations utilisent plus de 10 lexiques. Ce nombre plus important s'explique par la grande variété de mots pouvant composer les EN de ces deux catégories.

Tableau 3
Nombre de lexiques utilisés pour les anthroponymes et les toponymes

ANTHROPONYMES					
Patronymes	29	Prénoms	1	Ethnonymes	3
Organisations	45	Ensembles artistiques	6		
TOPONYMES					
Toponymes > Pays	5	Pays	1	Villes < Toponymes < Pays	10
Villes	2	Microtoponymes	2	Hydronymes	4
Oronymes	3	Rues	2	Édifices	5

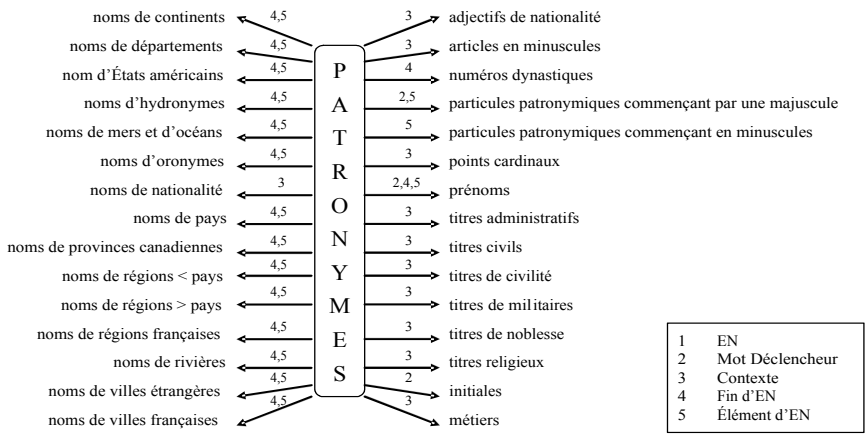

Fig. 3 : Rôles des lexiques pour la reconnaissance des patronymes

3.1.2 Application des règles

Une fois la projection des lexiques réalisée, les règles lexicosémantiques de réécriture sont appliquées, afin de permettre une première reconnaissance des entités nommées. Ces règles s'appuient essentiellement sur l'évidence interne définie par McDonald 1994 et utilisent des patrons basés sur les étiquettes sémantiques correspondant aux lexiques. Voici le formalisme retenu pour la conception de ces règles :

- X, une règle de réécriture;
- $P_X \rightarrow catégorie_X$, la forme générale de X ;

- P_X, le patron de X ;
- $catégorie_X$, la catégorie référentielle de la balise à poser pour X (Patronyme, Pays, etc.);
- $[\, w_X^1 , w_X^n \,]$, un intervalle discret, représentant P_X, avec w_X^1 et w_X^n les éléments de début et de fin respectives du patron;
- w_X^i, le $i^{ème}$ élément de P_X, pouvant être :

forme : une forme quelconque (p.ex. la forme *équipe* peut être directement recherchée);
balise : une étiquette sémantique référant à une forme appartenant à un lexique (p.ex. 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 commençant par une majuscule.

Au niveau de l'implémentation (sous forme d'expressions régulières), nous distinguons les **balises** et les **np** des **formes** en faisant précéder les premiers d'un signe de dollar. À chaque w_X^i peut être associé un quantificateur : `?` (0 ou 1 fois), + (1 fois ou plus), `*` (0 fois ou plus). La partie de P_X à baliser est mise entre crochets. Certaines règles sont «factorisées» : si plusieurs règles ont exactement la même forme à un w_X^i près, elles sont unifiées en remplaçant le w_X^i différent par une variable représentant l'un ou l'autre (*A B C* → Cat_1 et *A B D* → Cat_1 donnent *A B C ou D* → Cat_1).

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 EN, 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 94 règles de cette forme, comme :

```
$Clé_hydro $Article_min+ [$NP+] → Hydronyme
```
(*rives de la Kamogawa* ou *eaux du Yangtsé Kiang*).

```
$Titre_militaire $Adj_nationalité? [$Prénom* $NP] → Patronyme
```
(*commandant Massoud* ou *général bosno-serbe Radko Mladic*).

3.2 Apprentissage et seconde reconnaissance

La mise à jour des lexiques a été abordée dans certains systèmes (Cucchiarelli et coll. 1998, Poibeau 1999), mais pose encore un grand nombre de problèmes. La mise en place d'un tel module vise deux objectifs : la résolution de certaines coréférences et l'identification de nouvelles entités.

La coréférence est un problème récurrent dans le traitement automatique des noms propres : *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.

Pour améliorer les performances de notre système, nous avons mis au point une méthode basée sur des heuristiques, afin de créer de nouveaux lexiques. Contrairement à ceux de Poibeau 1999, ces lexiques sont obtenus automatiquement et pourront enrichir nos lexiques de base après vérification manuelle; en effet, nous ne voulons pas risquer de bruiter ces derniers. Voici quelques exemples d'heuristiques permettant cet apprentissage, où *C* représente un candidat nom propre (Némésis comporte un total de 32 heuristiques comme celles-ci) :

- soit la forme C_1 C_2, si C_1 C_2 est catégorisé en tant que patronyme avec C_2 un nom patronymique inconnu, alors C_2 est ajouté au lexique des noms patronymiques (p. ex. *Lang* dans *Jack Lang*). Cette heuristique intervient dans la résolution des coréférences de noms de personnes;
- prenons la forme 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…*). C_1 est alors ajouté 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 tendue lors du prétraitement lexical (cf. section 3.1). Lorsque le premier élément 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*). Cette heuristique permet le traitement de coréférences portant sur les noms d'organisations.

Après avoir été ainsi obtenus, ces lexiques sont de nouveau projetés sur le corpus (cf. section précédente).

La mise en place de ce processus apporte une amélioration non négligeable aux performances de Némésis (+0,6 % pour la précision et +5,3 % pour le rappel) tout en étant très fiable (Fourour 2001).

4. Évaluation

Pour réaliser l'évaluation de notre système, nous avons balisé manuellement les EN de chaque catégorie présentes dans nos corpus et les avons com-

parées automatiquement à celles trouvées par notre système. Cette évaluation, présentée au Tableau 4, a été réalisée sur un corpus composé de textes extraits du journal français *Le Monde* et de pages Web, ce qui représente un total d'environ 31 000 mots et de 1 284 EN.

Tableau 4
Résultats de l'évaluation de Némésis

	EN correctement identifiées et catégorisées	EN identifiées mais mal catégorisées	EN mal identifiées	EN non identifiées
Anthroponymes	509	8	24	94
Toponymes	402	18	9	53
Ergonymes	84	18	8	29
Praxonymes	20	3	2	3
Total	**1 015**	**47**	**43**	**179**

Nous pouvons remarquer que la classe des phénonymes n'est pas représentée dans ce corpus d'évaluation; en effet les EN de cette classe sont en général peu fréquentes (cf. Tableau 1).

Finalement, les performances atteintes par Némésis sur l'ensemble des entités nommées sont de 79 %[3] pour le rappel et 91 %[4] pour la précision.

5. Limites de Némésis

Bien que les résultats de cette évaluation soient intéressants, nous avons constaté que la grande majorité des EN non reconnues l'étaient à cause d'un manque d'information contextuelle. En effet, sur 269 EN mal reconnues, 179 sont des EN pour lesquelles nous n'avons trouvé aucun mot déclencheur, aucun élément du contexte de gauche permettant de les catégoriser. Or, nous nous refusons à travailler à l'aide de grandes bases de noms propres (ce qui serait une solution) pour différentes raisons :

- il n'est pas possible de créer (et d'utiliser) des listes exhaustives d'EN (p. ex. 1,5 million de prénoms pour les seuls États-Unis);

3 (1 015 / 1 284)

4 (1 015+47+43+12) / 1 284, où 12 représente le nombre de noms communs reconnus comme EN.

- ils seraient immédiatement surannés (p. ex. des organisations se créent en permanence);
- 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. surcomposition référentielle, polysémie).

Il nous faut donc trouver un autre moyen pour permettre la reconnaissance de ces entités nommées. Pour cela, nous avons étudié la possibilité d'utiliser le Web pour retrouver d'autres contextes dans lesquels ces entités nommées apparaissent, afin de les catégoriser.

6. Reconnaissance d'EN à partir du Web

Notons tout d'abord que les entités nommées non reconnues sont toutes des EN 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 2.2), l'identification des limites de l'EN est immédiate. Nous avons donc, dans un premier temps, recensé toutes les entités nommées non reconnues durant des deux premières passes de Némésis.

Une fois cette liste constituée, nous lançons, pour chaque EN, un processus (*thread*) qui émet une requête *http* 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 de l'enregistrer dans un fichier.

Lorsque l'ensemble des processus sont terminés, un traitement est effectué pour chaque EN et sur chaque fichier correspondant à la recherche lancée sur cette EN. Pour les EN pures ou faiblement mixtes, la méthode consiste à rechercher un élément catégorisant dans le contexte immédiat (droit ou gauche) de l'EN. 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.

Pour les EN 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'EN 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.

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'EN est présente dans la ligne, une éventuelle forme étendue, correspondant à celle-ci selon certains schémas (cf. section 3.1) 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 EN.

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.

7. 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 des textes du journal français *Le Monde* (8 154 mots) et une page Web de la FAO (8 007 mots), contenant 649 entités nommées au total.

A priori, cette méthode paraît permettre la catégorisation de certaines EN : *Hindû-Kûsh* est ainsi retrouvé dans 4 contextes différents (*monts*, *chaîne*, *massif*, *montagnes*) que nous savons appartenir à la même catégorie référentielle. Nous trouvons 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 EN. Cependant, lorsqu'on automatise le traitement, différents problèmes se posent.

Tout d'abord, durant la phase d'identification des EN pures ou faiblement mixtes et des sigles, un certain nombre de formes sont retenues, à tort, comme potentiellement EN. Au total, 118 formes (44 pour *Le Monde* et 74 pour la page Web de la FAO), dont 20 sigles, sont traitées pour 72 EN, 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 (4 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 plus loin). Pour ce qui est de l'augmentation du temps de traitement, nous n'envisageons pas d'autre solution pour pouvoir ne traiter que les entités nommées et non pas les noms communs.

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 de taille suffisante 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 Némésis peut permettre un gain important sur le taux de rappel, car il permet la catégorisation d'EN non reconnues jusqu'alors. En revanche, il 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 qu'on veut privilégier le rappel ou la précision.

De plus, une même forme peut tenir le rôle de contexte ou de déclencheur pour différentes 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 clefs d'organisation, lui-même plus élevé que celui des mots clefs 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, sur le Web, de chaque catégorie pour une EN;
- le nombre de contextes différents, sur le Web, de chaque catégorie pour une EN;
- la distribution moyenne des EN en fonction de leur catégorie référentielle (cf. Tableau 1);
- la distribution des EN en fonction de leur catégorie référentielle dans le corpus en cours de traitement (nous nous basons ici sur les EN 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 (Tableau 5). Sur ces 44 EN, 22 sont largement bien catégorisées, et 5 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). 10 de ces EN n'ont été catégorisées que par un seul contexte (4 correctement, et 6 incorrectement). Enfin, 7 EN 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 de la FAO, la distribution des EN indique clairement que c'est un texte qui traite de géographie (plus de 31 % des EN 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, on obtient une catégorisation correcte pour 6 de ces 7 EN.

Tableau 5
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 catégorisées selon le critère de décision	7

Par conséquent, nous pouvons estimer (avec un seuil entre 2 et 4, et en tenant compte de la distribution des EN 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 EN catégorisées apparaissant plusieurs fois dans le texte, ce module augmente actuellement le taux de rappel d'un peu plus de 5 % (79 % avant, 84,5 % après), tandis qu'il fait baisser la précision d'environ 2 % (91 % avant, 89 % 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.

8. Conclusions et perspectives

Nous avons présenté l'architecture logicielle de Némésis, un système de reconnaissance des entités nommées pour le français se basant sur des catégorisations graphique et référentielle dont les critères ont été vérifiés dans le corpus. Partant de ce système, nous avons étudié les possibilités d'utiliser le Web dans cette tâche et avons implémenté un module de reconnaissance des entités nommées à partir du Web. Ce module permet un gain de 5 % sur le taux de rappel, avec une perte de 2 % sur la précision.

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 la catégorie référentielle. En ce sens, même si le taux général de précision a chuté, notre module de reconnaissance à partir du Web n'induit pas de «bruit», car il ne modifie pas les résultats préalablement corrects.

Certaines améliorations peuvent encore être apportées à notre module : prise en compte exclusive 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.

Références

BAUER, G. 1985 *Namenkunde des Deutschen*, coll. Germanistische Lehrbuchsammlung, vol. 21.

CUCCHIARELLI, A., D. LUZI et V. PAOLA 1998 «Using Corpus Evidence for Automatic Gazetteer Extension», dans *Proceedings of LREC 98*, p. 83–89.

DAILLE, B. et E. MORIN 2002 «Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations», *Traitement Automatique des Langues* 41-3 : 601–621.

FOUROUR, N. 2001 «Identification et catégorisation automatiques des anthroponymes du Français», dans *Actes de la 8ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles,* vol 1, p. 441–450.

GALE, W., K. CHURCH et D. YAROWSKY 1992 «One Sense Per Discourse», dans *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, p. 233–237.

GRASS, T. 2000 «Typologie et traductibilité des noms propres de l'allemand vers le français», *Traitement Automatique des Langues* 41-3 : 643–670.

GREFENSTETTE, G. 1999 «The WWW as a Resource for Example-Based MT Tasks», dans *Proceedings of ASLIB Translating and the Computer 21 Conference*, Londres. Texte disponible à l'adresse **http://www.xcre.xerox.com/Publications/Attachments/1999-004/99_aslib.pdf**.

JACQUEMIN, C. et C. BUSH 2000 «Fouille du Web pour la collecte d'Entités Nommées», dans *Actes de la 7ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2000)*, p. 187–196.

JONASSON, K. 1994 *Le Nom Propre. Constructions et interprétations*, coll. Champs linguistiques, Gembloux et Paris, Duculot.

MCDONALD, D. D. 1994 «Internal and External Evidence in the Identification and Semantic Categorization of Proper Names», dans *Corpus Processing for Lexical Acquisition*, chapitre 2, p. 61–76.

MIKHEEV, A. 1999 «A Knowledge-free Method for Capitalized Word Disambiguation», dans *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99)*, p. 159–166, College Park, University of Maryland.

MIKHEEV, A., M. MOENS et C. GROVER 1999 «Named Entity Recognition without Gazetteers», dans *Proceedings of the 9th International Conference of the European Chapter of the Association for Computational Linguistics (EACL 99)*, p. 1–8, Bergen.

MUC-7 1998 *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

PAIK, W., E. D. LIDDY, E. YU et M. MCKENNA 1996 «Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval», dans B. Boguraev, J. Pustejovsky et coll., *Corpus Processing for Lexical Acquisition*, Language, Speech and Communications, chapitre 4, Cambridge (Mass.), MIT Press.

Poibeau, T. 1999 «Repérage des entités nommées : un enjeu pour les systèmes de veille», dans *Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA 99)*, vol. 19, p. 43–51.

Salton, G. et M. McGill 1983 *Introduction to Modern Information Retrieval*, New-York, McGraw-Hill.

Wacholder, N., Y. Ravin et M. Choi 1997 «Disambiguation of Proper Names in Texts», dans *Procedings of the 5th Conference on Applied Natural Language Processing (ANLP 97)*, p. 202–208.

Wakao, T., R. Gaizauskas et Y. Wilks 1996 «Evaluation of an Algorithm for the Recognition and Classification of Proper Names» dans *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, vol. 1, p. 418–423.

Wolinski, F., F. Vichot et B. Dillet 1995 «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)*, p. 23–30.