Récital 2001, Tours, 2-5 juillet 2001

Identification et catégorisation automatiques des anthroponymes du Français

Nordine Fourour
Institut de Recherche en Informatique de Nantes - Université de Nantes
2, chemin de la Houssinière - BP 92208 - 44322 Nantes Cedex 3, France
fourour@irin.univ-nantes.fr

Résumé - Abstract

Cet article présente un système de reconnaissance des noms propres pour le Français. Les spécifications de ce système ont été réalisées à la suite d'une étude en corpus et s'appuient sur des critères graphiques et référentiels. Les critères graphiques permettent de concevoir les traitements à mettre en place pour la délimitation des noms propres et la catégorisation repose sur les critères référentiels. Le système se base sur des règles de grammaire, exploite des lexiques spécialisés et comporte un module d'apprentissage. Les performances atteintes par le système, sur les anthroponymes, sont de 89,4 % pour le rappel et 94,6 % pour la précision.

Mots-clefs : Entités nommées, reconnaissance automatique, procédure incrémentielle.

This paper presents a French proper name recognizer. The specifications of this system have been elaborated through corpus investigation upon graphical and semantic criteria. The graphical criteria allow to presuppose some processes to identify proper names boundaries and the semantic classification is used to categorize them. The system is grammar-rule based, uses specialized lexicons, and includes a learning processing. The system performance evaluated on the categories composing anthroponym class achieves 94.6% of precision and 89.4% of recall.

Keywords : Name entities, automatic recognition, incremental process.

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 la traduction (Daille et Morin, 2001). Cette reconnaissance a été réalisée de façon satisfaisante en extraction d'information (EI), pour des textes journalistiques anglais : la majorité des systèmes de reconnaissance des noms propres en compétition aux récentes conférences MUC (MUC-7, 1998) ont une précision et un rappel supérieurs à 90 %.

La reconnaissance des noms propres pour le français, comme pour l'anglais (Wacholder et al., 1997), se heurte au problème de l'ambiguïté levée par la détermination des limites à droite du nom propre. D'une part, il faut résoudre des problèmes de modification adjectivale et d'attachement des prépositions et des coordinations. D'autre part, certains noms propres peuvent

être composés en quasi totalité de mots en minuscules. Pour aider à cette délimitation, nous proposons des critères graphiques qui sont vérifiés expérimentalement par une étude en corpus.

En EI, les noms propres sont généralement séparés en 4 classes : personnes, lieux, organisations et expressions temporelles. Bien que cette catégorisation regroupe une grande partie des noms propres présents 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 également validée en corpus.

Après la présentation des catégorisations graphique et référentielle proposées (cf. section 2), nous décrivons le système permettant l'identification et la catégorisation des entités nommées[1] (cf. section 3). Ensuite, nous évaluons les performances de ce système (cf. section 4). Enfin, nous présentons nos conclusions et les perspectives qu'ouvre notre travail.

2 Catégorisations

Nous présentons successivement les résultats d'une étude portant sur la représentation numérique des différentes catégories référentielles des entités nommées, puis ceux de l'étude graphique[2]. Ces études ont été réalisées sur un corpus regroupant des échantillons de deux périodiques dont les textes sont disponibles sous format électronique : La Recherche[3] (17 067 mots) et Le Monde[4] (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'avéreraient 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.

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

ENAMEX noms propres faisant référence à des noms de personnes, lieux ou organisations ;

TIMEX expressions temporelles divisées en dates et heures ;

NUMEX expressions numériques référant à des 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

[1] Ce terme 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.)

[2] Les résultats quantitatifs que nous présentons ont été obtenus manuellement. Toutes les entités nommées ont été identifiées, catégorisées et comptées.

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

[4] Corpus de textes *Le Monde* - année 1997 - European Corpus Initiative (ECI) distribué par ELRA

	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	
Toponymes	**107**	**28,7 %**	**270**	**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		14	
Ergonymes	**64**	**17,2 %**	**93**	**6,4 %**
Sites de production	0		0	
Marques et produits	31		37	
Entreprises industrielles	0		4	
Coopératives	0		0	
Établissements d'enseignement et de recherche	27		7	
Installations militaires	0		1	
∗ Œuvres intellectuelles	6		44	
Praxonymes	**3**	**0,8 %**	**16**	**1,1 %**
Faits historiques	0		0	
Maladies	0		0	
∗ Évènements culturels, sportifs, politiques	0		0	
∗ Périodes historiques	3		1	
Phénonymes	**5**	**1,3 %**	**0**	**0 %**
Catastrophes naturelles	0		0	
Astres et comètes	5		0	
Total	**373**		**1 445**	

TAB. 1 – Distribution des entités nommées en fonction de leur catégorie référentielle

noms de médias, d'évènements, de documents, etc. n'y sont pas représentés. Paik et al. (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 al. (1996) :

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 (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.

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 seule classification existante pour la traduction, à notre connaissance, est celle réalisée par le linguiste germanophone Bauer. Il énumère ce qui, par convention, constitue un nom propre et prend en considération des éléments extra-linguistiques propres au référent. Cette typologie comporte 5 classes :

Anthroponymes noms de personnes individuelles et 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 al. (1996) et celle de Bauer (1985). Néanmoins, certaines classes de Bauer (1985), comme les praxonymes ou les phénonymes n'apparaissent pas chez Paik et al. (1996). Inversement, toutes les classes présentes dans Paik et al. (1996) peuvent s'insérer dans les classes de la typologie de Bauer (1985). 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 Bauer (1985), comme base pour notre catégorisation. Toutes les entités nommées rencontrées dans nos corpus trouvent place dans les 5 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[5].

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 Saint-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 minuscule comme *le 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* ;

[5]Les catégories étendues ou créées apparaissent précédées d'un astérisque.

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).

	La Recherche	Le Monde
EN Pures Simples	145	313
EN Pures Complexes	25	89
Prénom+Nom	68	299
EN Mixtes Simples	21	35
EN Mixtes Complexes	44	144
Sigles	15	127
Total	318	1 007

TAB. 2 – Présence d'entités nommées en fonction de leurs caractéristiques graphiques

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 Le 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 événements. 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 Présentation du système

Ici, nous présentons le système mis en place pour identifier et catégoriser les entités nommées, dont l'architecture est illustrée à la figure 1. Il s'agit d'un système composé de quatre modules qui s'appliquent les uns à la suite des autres : prétraitement lexical, première reconnaissance des entités nommées, apprentissage et seconde reconnaissance des entités nommées.

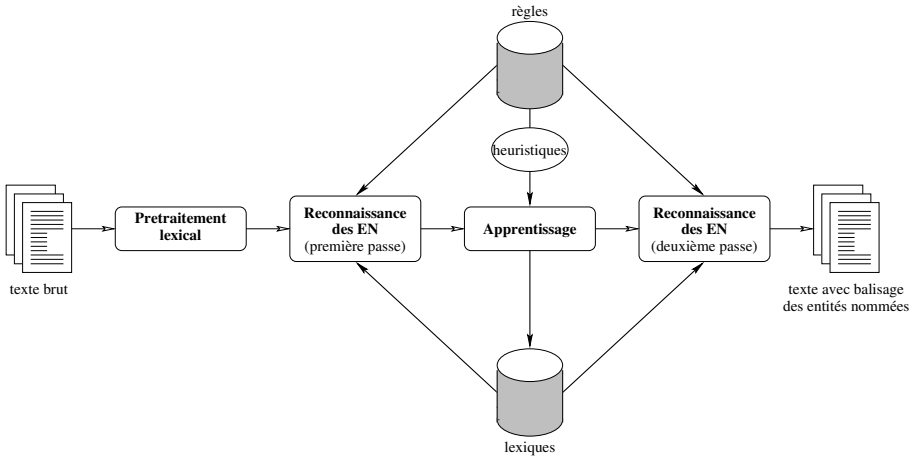

FIG. 1 – Architecture du système

3.1 Prétraitement lexical

Le prétraitement lexical s'effectue en deux étapes : segmentation du texte en phrases et 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 (e.g. *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 al. (1995) et Wacholder et al. (1997) utilisent l'association entre les sigles et leur forme étendue, mais uniquement pour les coréférences en ce qui concerne Wacholder et al. (1997).

3.2 Première reconnaissance

Ce module est composé de deux parties : projection de lexiques spécialisés et application de règles de grammaire. La projection s'effectue comme suit :

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

L'utilisation de lexiques spécialisés est à la base de tout système de reconnaissance des noms propres, comme le montrent McDonald (1994) et Wakao et al. (1996). Ces lexiques ont été obtenus en réalisant manuellement des listes les plus exhaustives possibles ou en appliquant des filtres sur des ressources issues du Web. Nous avons également enrichi ces lexiques en faisant varier le genre, le nombre et la casse lorsque ceci était nécessaire. Les éléments composant ces lexiques peuvent tenir un ou plusieurs rôles, selon les catégories d'entités nommées que les règles dans lesquelles ils sont utilisés permettent d'identifier[6] :

EN l'élément est une entité nommée connue (*OMS*, *Alexandre*, *Canal +*) ;

contexte l'élément appartient au contexte gauche immédiat, mais ne fait pas partie de l'entité nommée (*philosophe*, *français*). Il permet de la catégoriser et d'en identifier le début ;

[6]Les mots d'un même lexique peuvent avoir des fonctionnements hétérogènes, même au sein d'une catégorie.

mot déclencheur ces éléments possèdent le même rôle que les éléments de type contexte, à ceci près qu'ils font partie de l'entité nommée (*Fédération*, *Boulevard*) ;

élément de l'EN[7] l'élément appartient à l'entité nommée, mais ne permet pas de la catégoriser (*football*, *régional*). Ce type d'élément permet de définir les limites à droite des entités nommées.

Contenu du lexique	Rôles	Nombre d'éléments
adjectifs géographiques	élément de l'EN	52
adjectifs de nationalité	contexte, élément de l'EN	444
clés de microtoponymes	mot déclencheur	2
clés de noms d'organisation (majuscule)	mot déclencheur, élément de l'EN	119
clés de noms d'organisation (minuscule)	élément de l'EN	119
initiales	mot déclencheur	26
noms d'institutions	EN, élément de l'EN	34
noms de médias	EN	111
métiers	contexte, élément de l'EN	162
noms de continents	EN, mot déclencheur, élément de l'EN	6
noms de nationalité	EN, mot déclencheur, contexte	416
noms de pays	EN, mot déclencheur, élément de l'EN	258
noms de régions	EN, élément de l'EN	24
noms de villes étrangères	EN, élément de l'EN	1 790
noms de villes françaises	EN, élément de l'EN	35 956
numéros dynastiques	mot déclencheur, élément de l'EN	21
particules (majuscule)	élément de l'EN	36
particules (minuscule)	élément de l'EN	29
partis politiques	EN, élément de l'EN	25
points cardinaux	élément de l'EN	16
prénoms	EN, mot déclencheur, élément de l'EN	7 668
sports	élément de l'EN	7
titres administratifs	contexte	2
titres civils	contexte	9
titres de civilité	contexte	31
titres militaires	contexte	45
titres de noblesse	contexte	23
titres religieux	contexte	13

TAB. 3 – Liste des lexiques utilisés avec taille et rôles.

Le tableau 3 recense les lexiques utilisés pour le traitement des anthroponymes, ainsi que les différents rôles que leurs éléments tiennent dans les règles[7]. Dans la plupart des cas, pour une catégorie et un lexique donné, il n'y a qu'un rôle possible : seule la combinaison *EN-mot déclencheur* est possible. En revanche, le même lexique peut jouer n'importe quel rôle pour une autre catégorie : les éléments du lexique des noms de nationalité sont utilisés comme entités nommées et mots déclencheurs pour les ethnonymes ([*Français*], [*Allemande* de l'ouest]) et comme contexte pour les patronymes (*Français* [René Descartes]) ; les métiers servent de contexte pour les patronymes (*philosophe* [Denis Diderot]) et, pour les noms d'organisations, font partie de l'entité nommée ([Conseil National de l'Ordre des *médecins*]).

Une fois la projection des lexiques réalisée, les règles lexico-sémantiques 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. Nous distinguons principalement deux types de règles : contextuelles et hors contexte. Les premières ont recours au contexte gauche et ne balisent qu'une partie du patron `TITRE_RELIGIEUX ADJ_NATIONALITÉ? [NP]` → `PATRONYME` : l'*évêque français Gaillot*, alors que les secondes ne tiennent compte

[7] Pour le moment, toute forme étiquetée par la projection des lexiques peut faire partie d'un nom d'organisation. Nous ne considérerons donc comme éléments d'organisations que ceux qui terminent un nom d'organisation.

que de la structure propre de l'entité nommée et balisent le patron entier [NOM_NATIONALITÉ PARTICULE_MIN$^+$ POINT_CARDINAL] $\rightarrow$ ETHNONYME : l'*Allemande de l'ouest*.

En sus des précédentes règles, nous avons des règles qui traitent les sigles et leur forme étendue. La catégorie à laquelle appartient la forme étendue est déterminée, puis elle est balisée, ainsi que son sigle : dans *Association des joueurs professionnels (ATP)*, le terme *Association* est étiqueté comme clé d'organisation et permet de déduire que l'ensemble de la forme étendue et son sigle sont des organisations.

L'utilisation de lexiques et de patrons lexico-sémantiques est classique pour ce type de traitement (McDonald, 1994). En revanche, les rôles de ces éléments lexicaux n'ont pas été clairement définis auparavant.

3.3 Apprentissage et seconde reconnaissance

La mise à jour des lexiques a été abordée dans certains systèmes (Poibeau, 1999; Cucchiarelli et al., 1998), mais elle pose encore de nombreux problèmes. L'intérêt d'un tel module est double : résoudre certaines coréférences et identifier de nouvelles entités (cf. section 4).

La coréférence est un problème récurrent en 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 et 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 (Il y a 22 heuristiques comme celles-ci pour les patronymes, 1 pour les ethnonymes et 1 pour les organisations) :

– 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 (e.g. *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 (e.g. *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 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 (e.g. *FFF* et *Fédération française de football*). Cette heuristique permet le traitement de coréférences portant sur les noms d'organisations.

Ces lexiques sont de nouveau projetés sur le corpus (c.f. section précédente). La nouvelle information apportée par cette projection va être utilisée :

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.

4 Évaluation

L'évaluation de notre système est restreinte aux catégories composant les anthroponymes. Pour la réaliser, nous avons balisé manuellement les patronymes, prénoms, etc. présents dans nos corpus et les avons comparés à ceux trouvés par notre système[8]. Un premier échantillon[9] nous a servi de corpus d'étude pour l'évaluation de l'apport de chacun de nos modules.

		EN correctement identifiées et catégorisées	EN identifiées mais mal catégorisées	EN non ou mal identifiées
Étape 1	Précision 90,6 %	174	4	71
	Rappel 70,7 %			
Étape 2	Précision 94 %	186	4	59
	Rappel 79,7 %			
Étape 3	Précision 94,6 %	209	4	36
	Rappel 85 %			
Corpus de test	Précision 95,3 %	263	2	19
	Rappel 93,3 %			

TAB. 4 – Résultats des différentes étapes de l'évaluation du système

Dans un premier temps, nous avons étudié les performances du programme n'utilisant que la segmentation du texte et la première passe de reconnaissance des entités nommées (cf. tableau 4, étape 1). Nous obtenons déjà des taux de rappel et de précision intéressants (resp. 70,7 % et 90,6 %). La plupart des anthroponymes non reconnus n'ont pas du tout été identifiés comme entités nommées (pauvreté des lexiques, entités nommées dont le premier élément ne commence pas par une majuscule, etc.), ce qui explique une perte sur le rappel plutôt que sur la précision.

Ensuite, le prétraitement sur les sigles a été ajouté au programme. Ceci nous permet d'augmenter la couverture sans induire de bruit (6 sigles et 6 formes étendues, tous correctement reconnus) : le rappel et la précision en sont donc sensiblement augmentés (resp. 9 % et 3,4 %). Ces résultats (étape 2) n'illustrent qu'une partie du gain obtenu avec ce module. En effet, lors de la deuxième passe, les coréférences aux sigles et aux formes étendues seront identifiées et catégorisées grâce à la conjugaison de ces deux derniers modules. Ces résultats (étape 3) montrent que 23 anthroponymes supplémentaires ont été correctement identifiés et catégorisés. Cela a n'a qu'une faible incidence sur la précision au contraire du rappel (resp. +0,6 % et +5,3 %), car cette première était déjà élevée.[10]

Enfin, une fois l'impact de chaque module étudié sur le corpus d'étude, il nous paraissait important de tester l'ensemble de notre programme à l'aide d'un autre échantillon[11], sur lequel nous n'avons pas travaillé. En effet, le fait d'avoir construit les lexiques, les règles, etc. en utilisant le premier échantillon fausse les résultats et tend à les embellir. Pourtant, les performances restent très élevés (cf. tableau 4, étape 4) : +8,3 % pour le rappel et +0,7 % pour la précision. Ceci est lié à la qualité du texte qui, dans une grande partie, traite de l'audiovisuel, domaine dans lequel nous possédons un lexique performant des noms de médias.

[8] Nous ne détaillons pas les résultats de chaque catégorie, mais évaluons l'ensemble des catégories composant les anthroponymes.

[9] Extrait du Monde composé de 6 987 mots et 246 anthroponymes

[10] Pour le moment, les informations obtenues par apprentissage ne comportaient pas d'erreur.

[11] Extrait du Monde composé de 6 381 mots et 282 anthroponymes

5 Conclusions et perspectives

Cet article a décrit le développement d'un système de reconnaissance des entités nommées pour le français, se basant sur des catégorisations graphique et référentielle dont les critères ont été vérifiés en corpus. Ce système comporte un traitement lexical et grammatical, un traitement des sigles et une phase de mise à jour des lexiques spécifiques. La version actuelle du système atteint 95 % en précision et 89,4 % en rappel (ces résultats font la moyenne sur les deux corpus).

La première amélioration que nous pourrons apporter au système sera d'exploiter plus largement la catégorisation graphique en l'intégrant dans la conception des règles. Deuxièmement, certaines catégories de notre typologie devront être raffinées, se scindant en plusieurs : des œuvres intellectuelles paraissent déjà pouvoir se dégager différentes sous-catégories (œuvres artistiques, maladies, partis du corps, etc. qui portent le nom de leur découvreur…). Cette typologie n'est donc pas figée et pourrait devenir modulaire : si l'on prend *le festival de l'Université de Natal*, il pourrait être intéressant, selon le domaine d'application, de ne reconnaître que *l'Université de Natal* ou même simplement *Natal*. Troisièmement, il faudra étendre la reconnaissance à toutes les classes de notre typologie. Cela va poser de nouveaux problèmes : pour les toponymes, les informations apportées par le traitement de l'évidence interne ne sont pas suffisantes ; il faudra donc trouver des techniques supplémentaires pour leur reconnaissance. De plus, plus le nombre de catégories à reconnaître augmente, plus il va être difficile de gérer les ambiguïtés. Enfin, il nous paraît essentiel de parvenir à induire automatiquement des règles pour reconnaître les entités nommées et de pouvoir les insérer parmi les règles préexistantes.

Références

Bauer, G. (1985). *Namenkunde des Deuschen*. Germanistische Lehrbuchsammlung Band 21.

Cucchiarelli, A., Luzi, D., et Paola, V. (1998). Using corpus evidence for automatic gazetteer extension. In *Proceedings of LREC'98*.

Daille, B. et Morin, E. (2001). Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. *t.a.l.*, **41**(3).

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

McDonald, D. D. (1994). Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus Processing for Lexical Acquisition*, chapter 2.

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

Paik, W., Liddy, E. D., Yu, E., et McKenna, M. (1996). Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition*.

Poibeau, T. (1999). Repérage des entités nommées : un enjeu pour les systèmes de veille. In *Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA'99)*.

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

Wacholder, N., Ravin, Y., et Choi, M. (1997). Disambiguation of proper names in text. In *Procedings of ANLP'97*, pages 202–208.

Wakao, T., Gaizauskas, R., et Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of COLING'96*, volume 1, pages 418–423.

Wolinski, F., Vichot, F., et Dillet, B. (1995). Automatic processing of proper names in texts. In *Proceedings of EACL'95*, pages 23–30.