Meta
Journal des traducteurs
Translators' Journal

META

Linguistique et reconnaissance automatique des noms propres

Nathalie Friburger

Volume 51, numéro 4, décembre 2006

La traduction des noms propres (1) et Langue, traduction et mondialisation : interactions d'hier, interactions d'aujourd'hui
Language, Translation and Globalization: Interactions from Yesterday, Interactions from Today (2)

URI : https://id.erudit.org/iderudit/014331ar
DOI : https://doi.org/10.7202/014331ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Friburger, N. (2006). Linguistique et reconnaissance automatique des noms propres. *Meta*, *51*(4), 637–650. https://doi.org/10.7202/014331ar

Résumé de l'article

Cet article présente les aspects linguistiques utilisés par les informaticiens pour créer des systèmes de reconnaissance automatique de noms propres. Ces systèmes doivent non seulement repérer correctement les noms propres dans les textes mais aussi leur donner une catégorie (lieux, personnes, organisations…). Nous montrerons les différents indices utilisés ainsi que les difficultés liées à cette tâche.

Linguistique et reconnaissance automatique des noms propres

NATHALIE FRIBURGER
Université de Tours, Tours, France
nathalie.friburger@univ-tours.fr

RÉSUMÉ

Cet article présente les aspects linguistiques utilisés par les informaticiens pour créer des systèmes de reconnaissance automatique de noms propres. Ces systèmes doivent non seulement repérer correctement les noms propres dans les textes mais aussi leur donner une catégorie (lieux, personnes, organisations...). Nous montrerons les différents indices utilisés ainsi que les difficultés liées à cette tâche.

ABSTRACT

This article presents the linguistic aspects used by computer scientists to create systems to automatically recognize proper names. Those systems must locate correctly proper names but, moreover, they must give a categorization (places, persons, organisations...). We will show the different clues and difficulties linked to this task.

MOTS-CLÉS/KEYWORDS

catégorie, nom propre, Prolex, reconnaissance automatique, traitement des ambiguïtés

Introduction

Avec la très grande quantité d'informations textuelles disponibles sur Internet ou, de manière plus générale, sur support informatique, créer des outils qui automatisent l'exploration ou l'extraction d'informations pertinentes, qui facilitent l'accès aux informations et minimisent le travail humain est crucial. Ces systèmes doivent faire face aux difficultés propres à l'écrit : les informations contenues dans les textes sont non structurées et les constructions langagières sont en partie imprévisibles.

Le travail présenté ici s'inscrit dans le projet Prolex[1]. Ce projet rassemble des travaux informatiques et linguistiques pour l'élaboration de ressources électroniques autour des noms propres (dictionnaires). Afin d'enrichir ces ressources, nous avons créé un système de reconnaissance automatique de noms propres et travaillé sur des textes journalistiques. Ce type de texte permet un enrichissement rapide des dictionnaires car les noms propres y sont très fréquents. Reconnaître et donner une catégorie (lieu, personne…) à un nom propre de manière automatique est un enjeu important pour aller vers des systèmes de traduction automatique des noms propres.

Ce travail s'inscrit aussi dans le cadre de la linguistique harissienne, telle qu'elle a été mise en application par Maurice Gross à travers le système Intex[2] de Silberztein (1993).

Le but de cet article est de présenter les aspects linguistiques utilisés par les informaticiens pour créer des outils qui reconnaissent automatiquement les noms propres. Nous rappellerons tout d'abord comment les noms propres sont considérés en linguistique ainsi que les situations dans lesquelles ils peuvent apparaître en corpus, ceci

afin de clarifier les problèmes auxquels devront faire face les informaticiens pour automatiser la reconnaissance automatique des noms propres. Ensuite, nous présenterons les différentes manières d'automatiser la reconnaissance des noms propres ainsi que les indices sur lesquelles cette reconnaissance s'appuie. Nous présenterons aussi quelques résultats chiffrés d'une étude en corpus afin de confronter nos idées à la réalité des textes journalistiques.

1. Les noms propres en français

Il existe plusieurs manières de définir un nom propre mais aucune ne fait l'unanimité auprès des linguistes ; citons, par exemple, la définition du nom propre que donne *Le Bon Usage* de Grevisse et Goosse (1986 : 751) : « Le nom propre n'a pas de signification véritable, de définition ; il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière ». D'autre part, Gary-Prieur (1994 : 7) dit que l'interprétation du nom propre « requiert presque toujours une mise en relation avec le référent initial » et « mobilise des connaissances discursives ». Le nom propre se situe dans l'espace et le temps ; il renvoie au domaine de la description dont parle Molino (1982) sous le nom de deixis.

Sur le plan sémantique, il existe trois grandes approches linguistiques que résume Jonasson (1994 : 114) :

- Soit le nom propre est *vide de sens* (il réfère sans désigner).
- Soit le sens du nom propre est *une description du référent* (on considère qu'il a un sens fort et qu'il identifie de manière univoque un référent, ou qu'il a un sens réduit à des traits sémantiques généraux : trait féminin / masculin, humain / non-humain, etc.).
- Soit le sens du nom propre est *un prédicat de dénomination* : il ne décrit pas l'objet dénoté mais lui donne un nom, par exemple tel homme « est appelé » *Alexandre*.

Les noms propres n'ont donc pas de signification dans le sens où on l'entend pour un nom commun. Mais peut-on vraiment classer les noms propres et les noms communs en deux catégories bien distinctes ? *Organisation des Nations Unies*, *Jardin des Plantes*, *Parisien*, *Mérovingien*, *Vivendi*, *EDF*, *Renault 5* ou *14 juillet 1789* sont-ils des noms propres ?

En fait, il semble y avoir un continuum entre les noms propres et les noms communs. Selon Grevisse et l'acception commune, les « véritables noms propres » sont les noms de lieux (villes, monuments, régions, pays, îles, montagnes…) et les noms de personnes. Il semble que les noms de société (ex. : *Vivendi, EDF*) soient des noms propres acceptables : en effet, ces noms ont les propriétés de noms propres classiques ; ils désignent une entité dont nous avons une image mentale bien précise mais qui ne peut être définie comme on le ferait pour un nom commun. *Organisation des Nations Unies* désigne aussi une organisation unique et bien connue : ce nom propre, composé de noms communs et d'adjectifs ayant tous individuellement une signification qui peut aider à la compréhension de l'entité qu'ils désignent, semble être lui aussi un nom propre. Les noms tels que *Révolution française* réfèrent aussi à un contenu précis, il ne s'agit pas seulement ici d'une révolution qui aurait eu lieu en France mais d'un événement important de notre histoire, situé dans l'espace (*en France*) et le temps (*1789-1799, Ancien Régime*), à la manière d'un nom propre de ville ou de personne. *Renault 5* est le nom d'une marque de voiture reproduite à des milliers d'exemplaires

mais ce terme désigne uniquement la voiture *Renault 5* connue pour ces caractéristiques particulières. Rey-Debove (1994) ajoute qu'un nom de marque désigne une classe engagée dans une hyperonymie mais considère ces noms comme de faux noms propres : du point de vue du TAL (Traitement Automatique des Langues), on les considérera avec autant d'intérêt que des noms propres. Les dérivés de noms propres (gentilés, ethnonymes, périodes historiques, etc.), bien qu'ils aient une définition (ex. : *Parisien* = habitant de Paris, *Mérovingien* = descendant de Mérovée), sont souvent considérés comme des noms propres ; ils ne désignent pas un individu, mais un groupe qui a une certaine individualité.

Ces exemples montrent combien la limite entre noms propres et noms communs n'est pas claire. Les informaticiens qui travaillent dans le domaine de l'extraction d'information, ont abordé le problème de manière pragmatique ; ils ont défini la notion d'**entités nommées**[3] pour regrouper tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantité. La suite de cet article parlera des noms propres au sens large.

1.1. La productivité des noms propres

Comme les autres mots, les noms propres participent à la création morpho-syntaxique des locuteurs du français. Lexicalisation, détermination et dérivation les rendent particulièrement productifs.

Les noms propres lexicalisés ne sont pas considérés dans notre travail de reconnaissance automatique de noms propres ; en effet, un *frigidaire* et un *bic* ne sont plus des noms propres (ces noms sont utilisés comme synonymes de *réfrigérateur* et *stylo-bille*).

Beaucoup de noms propres sont utilisés avec un article défini (ex. : *la France, la Seine*). L'article défini peut aussi être intégré au nom propre : l'article appartient à la morphologie du nom propre, il ne dispose d'aucune autonomie (ex. : *Le Corbusier, Le Mans, La Fontaine, Les Seychelles*, etc.). Les noms propres utilisés dans un emploi métaphorique, dénominatif ou fractionné/multiplié (Garric et Maurel 2000) sont souvent accompagnés de l'article.

Les dérivés de noms propres proviennent principalement de noms de personnes (ex : *chiraquien, pasteuriser, homérique*) et de noms de lieux géographiques (ex : *italien, italo, italianisant*) (voir Eggert *et al.* 1998).

1.2. Typologies de noms propres

Les différentes typologies que nous exposons dans la suite ont été proposées par des linguistes et des informaticiens et éclairent les travaux sur la reconnaissance automatique des noms propres.

Typologies morpho-syntaxiques des noms propres

Les noms propres font partie de la catégorie syntaxique des noms. Les noms propres ont, en français, certaines caractéristiques qui les distinguent des noms communs « la plupart du temps » : absence d'article, absence de flexion morphologique, présence d'une majuscule, mais ces caractéristiques ne sont pas absolues ; il existe des noms

propres employés avec des articles (ex: *la Suisse*), d'autres ont une marque de flexion (ex.: *des Allemands, les deux Corées*) et ils ne se résument pas à des mots portant forcément une majuscule initiale.

Jonasson (1994) propose pour le français deux types de noms propres:

- Les noms propres purs: ce sont des « noms propres véritables » (ex.: *Jean-Pierre Papin, Boulogne-Billancourt*); Jonasson remarque qu'ils ne renseignent pas sur les propriétés de l'objet qu'ils désignent. Ce sont des noms de lieux ou de personnes que l'on peut repérer à l'aide de la majuscule.
- Les noms propres mixtes ou à base descriptive: les noms propres mixtes contiennent des noms propres purs et des noms communs (ex.: *le Collège de France, la tour Eiffel, le golfe Juan*) mais aussi des adjectifs (ex.: *La Nouvelle-Orléans*). Les noms propres à base descriptive sont composés d'un ou plusieurs noms communs éventuellement accompagnés d'adjectifs ou de prépositions (ex.: *le Massif central*, *la Banque centrale européenne, la Grande Barrière de Corail*). Les noms propres à base descriptive ou mixte sont des lieux, rues, places, parcs, bâtiments, organisations de toutes sortes.

Jonasson ajoute que « si on considère un trait comme la monoréférentialité, il est bien plus caractéristique des Npr[4] descriptifs ou mixtes que des Npr purs. Les premiers sont en général forgés expressément pour convenir à un seul particulier, qu'ils désignent en le décrivant, et ne sont normalement pas utilisés associés à d'autres particuliers. »

Daille et Morin (2000) introduisent une typologie basée sur des critères graphiques plutôt que sur la présence de noms communs ou non dans le nom propre:

- Les noms propres simples: un seul mot commençant par une majuscule (ex.: *Marseille, France*).
- Les noms propres complexes: ceux-ci sont composés de plusieurs unités lexicales pleines comportant toutes une majuscule (ex.: *Quai d'Orsay*, *Grand Palais*) mais ils peuvent contenir indifféremment des noms communs et des noms propres.
- Les noms propres mixtes: ils sont constitués de plusieurs unités lexicales comportant ou non des majuscules comme le *palais de Chaillot* ou *le Front populaire.*

Typologies sémantiques des noms propres

Zabeeh (1968), Bauer (1985), Grass (2000) proposent des classifications propres à l'onomastique. Grass *et al.* (2002) définissent une classification basée sur deux niveaux hiérarchiques (comme Paik *et al.* 1996) avec une couverture des noms propres très complète. Le premier niveau est celui des hypertypes: un hypertype correspond aux traits sémantiques primitifs (anthroponymes, toponymes, ergonymes et pragmonymes). Le second niveau est celui des types: il comprend des champs lexicaux relativement homogènes, en relation d'hyponymie avec les **hypertypes**[5].

Du côté du traitement automatique des langues, les travaux sur l'extraction des noms propres ont conduit les informaticiens à proposer des typologies plus simples et d'usage pratique pour le traitement informatique mais qui tiennent suffisamment compte de la réalité des noms propres.

2. Reconnaissance automatique des noms propres

Si un lecteur ne connaît pas un nom propre, le discours général lui fera reconnaître le nom propre mais aussi comprendre de quel type il est: lieu, personne ou autre. Pour le lecteur humain, il y a deux niveaux de reconnaissance du nom propre, qui ne

sont pas exclusifs l'un de l'autre : soit le nom propre est reconnu parce qu'il est connu, et il appartient à l'univers commun des connaissances (ex. : *La Loire*, *Paris*, *Sartre*), soit ce sont la graphie (présence de majuscule) et la sémantique des prédicats qui induisent le type du nom propre ou le précisent en cas d'ambiguïté.

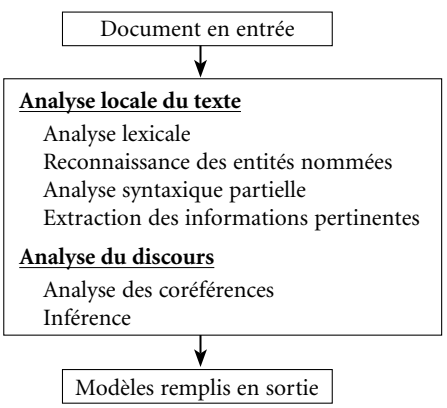
2.1. Les recherches dans le domaine de l'extraction automatique d'entités nommées

En fait, peu de recherches en informatique ont porté sur les noms propres avant la création du programme de recherche **MUC**[6] en 1987. Le but de ce programme est de promouvoir la recherche en compréhension automatique des textes. MUC prend la forme d'un concours par lequel les systèmes participants sont évalués. La tâche principale proposée par MUC est l'extraction des informations contenues dans les textes pour répondre aux questions : qui ?, quand ?, où ?, quoi ?, comment ? À partir de la sixième conférence MUC, des sous-tâches telles que la recherche de coréférences, la désambiguïsation du sens des mots, la recherche des entités nommées, etc. ont été proposées aux informaticiens participants.

Pour comprendre l'imbrication des tâches proposées par MUC, la Figure 1 présente l'architecture générale d'un système d'extraction d'informations. L'extraction est réalisée en deux étapes : on procède d'abord à une analyse locale permettant de trouver des faits simples, puis on analyse le discours afin d'inférer des faits plus complexes. La tâche d'extraction des entités nommées a lieu pendant l'analyse locale, après une analyse lexicale du texte, et simplifie le reste des traitements.

FIGURE 1

Description générale d'un système d'extraction d'informations



La sous-tâche d'extraction des entités nommées propose de distinguer trois types d'entités à reconnaître et à catégoriser : ENAMEX, TIMEX et NUMEX. TIMEX contient les expressions de temps et de dates, NUMEX rassemble les nombres et pourcentages, ainsi que les quantités monétaires, ENAMEX est composé des noms propres ou assimilés et des sigles.

Il existe trois principaux types de systèmes pour extraire les noms propres :

Les systèmes à base de règles[7] : La majorité des systèmes utilise cette approche. Les systèmes typiques à base de règles utilisent des descriptions linguistiques et des

indices permettant de repérer les noms propres (majuscule, présence d'un mot particulier…), ainsi que des dictionnaires de noms propres déjà connus. Les règles sont écrites à la main. Ces systèmes obtiennent de très bons résultats mais ils demandent un investissement humain conséquent. Ce type de stratégies n'est pas idéal pour des textes ne répondant pas à des critères rédactionnels stricts (par exemple, les e-mails). Dans cet article, tous les indices linguistiques utilisés par ce type de textes seront décrits.

Les systèmes à apprentissage[8] : Ils construisent leurs connaissances des noms propres automatiquement grâce à un apprentissage sur un corpus d'entraînement. Ces systèmes peuvent être très vite adaptés à tout type de textes mais donnent des résultats moins précis que les systèmes à base de règles. Les systèmes d'apprentissage minimisent le travail de description mais ont des résultats plus faibles.

Les systèmes hybrides[9] : Ils utilisent des règles écrites à la main mais construisent aussi une partie de leurs règles à l'aide d'informations syntaxiques et d'informations sur le discours tirées de données d'entraînement grâce à des algorithmes d'apprentissage.

Deux mesures sont utilisées pour évaluer les résultats d'un système d'extraction automatique de noms propres. Le **rappel** mesure la quantité de réponses correctes d'un système par rapport au nombre de réponses idéales. La **précision** est la quantité de réponses correctes du système parmi l'ensemble des réponses qu'il a fournies (correctes et incorrectes).

Les systèmes d'extraction des entités nommées ont obtenus très rapidement de très bons résultats sur l'anglais. Grishman et Sundheim (1996) notent que la plupart des participants arrivent à plus de 90 % de rappel et de précision, le meilleur score étant de 96 % avec une précision de 97 %. Sundheim (1995) dit que le rappel pour la tâche d'extraction des entités nommées réalisée par un humain est de 97 % seulement. Les résultats affichés par les différents systèmes à MUC sont très bons mais il faut rappeler qu'ils traitent de textes très homogènes limités à un domaine assez restreint (ex. : des dépêches AFP).

Testé sur un corpus de journaux *Le Monde*, notre système ExtracNP, créé pour le projet Prolex, obtient les meilleurs résultats pour l'instant sur le français : 93,2 % des noms propres du *Monde* avec une précision de 94,4 %. Il s'agit d'un système à base de règles qui utilise le formalisme des transducteurs du système Intex. Pour des informations sur les mécanismes utilisés dans ExtracNP, il est intéressant de lire Friburger (2002) et Friburger, Maurel (2001).

2.2. Comment reconnaître des noms propres avec un système à base de règles ?

La reconnaissance et le typage des noms propres sont deux problèmes croisés. En effet, pour extraire un nom propre, on utilise des indices qui permettent de le repérer, mais aussi de le catégoriser. Un système d'extraction, n'utilisant que la syntaxe, ne peut faire de distinction entre un nom propre et un nom commun, et ne pourra affecter une catégorie au nom propre. Les noms propres ont un aspect systématique et une structure qu'il est possible de décrire à l'aide d'informations souvent plus lexicales que syntaxiques.

Le premier indice naïf pour extraire les noms propres est la majuscule : il est très insuffisant car un nom propre peut être composé de plusieurs mots dont certains ne

portent pas de majuscule. De plus, la majuscule qui se trouve sur le premier mot d'une phrase est ambiguë : s'agit-il d'un nom propre ou, simplement, d'un mot banal portant une majuscule parce qu'il est au début d'une phrase ?

En fait, les indices les plus sûrs pour détecter et catégoriser les noms propres sont leurs contextes d'apparition droits ou gauches et/ou leur composition interne.

Preuve interne et externe

McDonald (1996) propose un outil de reconnaissance et de classification des noms propres fondé sur les notions de **preuve interne** et **preuve externe**. La plupart des outils informatiques de reconnaissance de noms propres utilisent ces preuves sans les nommer ainsi.

Les **preuves internes** se trouvent à l'intérieur même du nom propre. Ce sont des mots qui permettent de le repérer à coup sûr et, éventuellement, de le typer. Les preuves internes peuvent prendre la forme d'un ou plusieurs mots ou d'une abréviation connue pour faire partie d'un nom propre (ex. : ***Organisation*** *des Nations Unies, le* ***Mont*** *Blanc, Wall Street* ***Journal***). De tels mots se trouvent en début ou fin de noms propres (surtout dans les noms d'organisation). Un prénom peut aussi être utilisé comme preuve interne (Ex : **George** Sand).

La **preuve externe** est le contexte d'apparition des noms propres dans la phrase. Les noms propres sont une manière de référer à des individus d'un type spécifique. Dans le discours, surtout journalistique, l'auteur donne aux lecteurs des informations complémentaires sur les personnes, lieux, organisations qu'il cite : ces informations peuvent aider, dans un processus automatique, à déterminer le type d'un nom propre. La preuve externe sera aussi appelée **contexte droit** ou **contexte gauche** selon que le contexte se trouve à la droite ou à la gauche du nom propre dans la phrase (ex. : *la* ***ville de*** *Marseille, le* ***professeur*** *Tournesol, le* ***groupe*** *Vivendi, Derrick,* ***l'inspecteur allemand***).

Structure syntaxique des noms propres accompagnés d'une preuve externe

Noailly (1991), Gary-Prieur (1994), Forsgren (1994) détaillent les constructions dans lesquelles peuvent intervenir des noms propres ; le nom propre peut être épithète, attribut, sujet, objet, en apposition. Les noms propres apparaissent donc dans des constructions complexes. Leurs contextes peuvent contenir simplement un adjectif (ex. : *l'****anglais*** *Tony Blair),* ou prendre une forme plus complexe (ex. : *le* ***chef du gouvernement français,*** *Lionel Jospin*). Une incise peut permettre d'exprimer une relation entre noms propres (ex. : *Frédéric Mitterrand,* ***neveu de*** *François Mitterrand* ou *Canal Plus,* ***filiale de*** *Vivendi*). Ces structures peuvent être composées pour donner des formes plus complexes (ex. : ***la société française*** *Canal Plus,* ***filiale de*** *Vivendi*).

Variation des noms propres

Pour reconnaître les noms propres, il faudra prendre en compte leurs variations que Daille et Morin (2000) listent de la manière suivante : les variantes graphiques (ex. : *Parti Socialiste* ou *Parti socialiste*), les variantes telles que les sigles ou abréviations, certaines coordinations (ex. : *le grand palais et le petit palais* → *le grand et le petit palais*), les ellipses (ex. : *école normale sup* → *normale sup* → *normale*).

2.3. Le traitement des ambiguïtés

Résolution des ambiguïtés structurelles : la délimitation des noms propres

Jacquemin et Bush (2000) ont défini les problèmes d'extraction partielle (les erreurs liées aux mauvaises délimitations des entités nommées) comme suit :

- La sur-reconnaissance : la séquence reconnue contient l'entité nommée mais est trop longue.
- La sous-reconnaissance décrit le fait que l'entité reconnue est contenue dans l'entité initiale. Par exemple, dans la phrase *L'ancien président Valéry Giscard d'Estaing a visité Vulcania*, si on ne repère que *Valéry Giscard*, l'entité est sous-reconnue car on aurait dû trouver *Valéry Giscard d'Estaing*.

La sur-reconnaissance et la sous-reconnaissance se manifestent surtout à la droite des noms propres.

On trouve assez simplement le début d'un nom propre (présence d'une majuscule) mais les mots qui suivent n'en portent pas forcément ; par conséquent, la limite droite est difficile à trouver (ex. : *La Fédération nationale de la Mutualité française*). Wolinski *et al.* (1995) résolvent en partie ce problème en segmentant les noms propres grâce aux marqueurs grammaticaux (prépositions, conjonctions, virgules, points), mais cette segmentation est insuffisante puisque des noms propres peuvent contenir des conjonctions ou des prépositions. Trouilleux (1997) décrit une grammaire du contexte droit pour le français. L'extension à droite d'un nom propre peut contenir des adjectifs, noms, prépositions, déterminants, coordinations qui dépendent du type du nom propre, et se termine nécessairement par un nom ou un adjectif. Les possibilités d'extension à droite dépendent du type de nom propre considéré. L'idée de décrire une grammaire des extensions possibles des noms propres selon leurs types est intéressante. Néanmoins, même si un adjectif est autorisé après tel ou tel type de noms propres, cela pose problème : *l'Europe centrale* désigne bien une entité, mais dans *l'Europe riche*, seul *Europe* est un nom propre, *riche* n'en fait pas partie. En anglais, il y a beaucoup moins de problèmes de limites droites ; en effet, les noms propres portent sur tous les mots qui les composent une majuscule et se terminent souvent par un mot indiquant leur catégorie (ex. : *Central Park*, *National Security Agency*).

Une autre ambiguïté structurelle tient à la présence de la majuscule en début de phrase et après un point : cette majuscule marque-t-elle un début de phrase ? un nom propre ? les deux ? ou ni l'un ni l'autre ? Il faut donc segmenter les textes en phrases afin de désambiguïser la majuscule et le point et de connaître ainsi les début et fin de phrase (lire Silberztein 1993, Dister 1997, Friburger *et al.* 2000). Le point est ambigu en présence de majuscules ou de chiffres. Hormis les débuts de phrase, les motifs contenant à la fois des majuscules et des points sont de quatre types :

- Les noms de personnes lorsqu'ils sont précédés de titres ou civilités abrégés (ex. : ***M.*** *Dupont,* ***Mme*** *Durand*), ou lorsqu'ils sont précédés d'un prénom abrégé (ex. : ***J.*** *Dupont*).
- Les sigles (ex. : *La* ***S.N.C.F.*** *gère les chemins de fer*).
- Les mots composés se terminant par une lettre majuscule et les symboles (ex. : *Ce timbre coûte 20* ***F****. Il a été acheté chez un philatéliste*)
- Les abréviations diverses (ex. : ***éd.*** *Gallimard,* ***Cf.*** *France-Italie en juin 2000*).

Les symboles composés d'une seule lettre majuscule : abréviations d'unités de mesure (ex. : *V* = volt, *W* = watt, etc.) et symboles monétaires (ex. : *F* = Franc, etc.) ne posent pas de problème à l'intérieur d'une phrase, car ils ne sont pas suivis d'un point (ex. : *Un Magritte de 14 000* ***F*** *a été volé au Centre Pompidou*). Mais lorsqu'ils sont en fin de phrase, les symboles sont suivis d'un point qui rend la séquence ambiguë : un symbole suivi d'un point pourrait être analysé comme l'initiale d'un prénom, et le mot en majuscules qui commence la phrase suivante comme un nom de famille (ex. : *C'est un Magritte de 14 000* ***F. Volé*** *au Centre Pompidou, il ne sera sans doute jamais retrouvé*). Ces ambiguités peuvent être la plupart du temps résolues de manière automatique grâce à des règles ; une fois ces règles appliquées, la reconnaissance des noms propres se fera sans se soucier de l'ambiguïté des points finaux de phrases avec ceux qui sont contenus dans les noms propres.

Résolution des ambiguïtés sémantiques

Wolinski *et al.* (1995) proposent de désambiguïser le type d'un nom propre par un contexte local et une base de connaissances : si on rencontre *Saint-Louis* et *États-Unis* dans le même texte, on parle certainement de la capitale du Missouri. Le contexte global permet, lui aussi, la désambiguïsation d'un nom propre, si une partie de nom propre, déjà trouvé et catégorisé, se retrouve quelque part dans le texte.

La preuve externe est nécessaire pour obtenir des performances élevées dans l'extraction automatique des noms propres. Si on ne prend en compte que la preuve interne, on peut aboutir à des erreurs de classification. Par exemple, le nom propre contenu dans l'expression « *la société Hugues Aircraft* » pose un problème : *Hugues* est un prénom. La seule preuve interne apportée par ce prénom fait penser que *Hugues Aircraft* est un nom de personne, ce qui est contredit par la preuve externe. Ce type d'erreur de catégorisation est très fréquent entre noms de personnes et noms d'organisations.

Une heuristique de désambiguïsation :
Les mots ont un seul sens par discours

Les entités nommées sont introduites dans les textes une première fois avec leur forme la plus explicite ou la plus complète. Ensuite on y réfère par des raccourcis et variantes plus informels. Cette heuristique est utilisée par presque tous les systèmes d'extraction d'entités nommées. Gale *et al.* (1992), à travers un travail sur la désambiguïsation du sens, observent qu'en anglais, si un nom polysémique apparaît deux fois ou plus dans un discours[10], le plus souvent toutes ces occurrences partagent le même sens. Cette tendance est très forte puisque, selon eux, 98 % des noms polysémiques respectent cette loi dans les discours « bien écrits ». On peut suspecter des résultats très similaires pour le français.

3. Étude en corpus

Afin de mieux nous rendre compte de la quantité de noms propres que l'on peut repérer et catégoriser grâce à des indices linguistiques, nous avons réalisé une étude en corpus. Dans cette étude, portant sur un numéro du journal *Le Monde*, nous avons dénombré les preuves externes et internes suivant le type des noms propres qu'elles accompagnent[11]. Nous avons trouvé au total 3 755 noms propres faisant partie des

catégories suivantes : personnes (27 %), lieux (35,2 %) et gentilés (8,4 %), organisations (27 %) ; les objets et marques, phénomènes et catastrophes, manifestations et événements représentent le pourcentage restant.

Les résultats du travail sur *Le Monde* montrent que 50,4 % de tous les noms propres de ce journal sont accompagnés de preuves : c'est assez peu. 93 % des noms de personnes et 65 % des noms d'organisations sont accompagnés d'une preuve. Les lieux, eux, sont plus rarement accompagnés d'une preuve (20 % d'entre eux seulement). Le reste des noms propres (sans indices clairs permettant de les repérer) devra donc être trouvé par des moyens autres tels que l'apprentissage.

De manière plus détaillée, voyons quels seront les moyens de reconnaître chacun de ces types d'entités nommées (résultats obtenus sur des journaux *Le Monde*).

Noms de personnes

Les **anthroponymes** (prénoms et patronymes) sont très nombreux. Malgré cela, les noms de personnes sont les entités les plus faciles à extraire : les indices pour les repérer sont très nombreux. Comme l'ont déjà remarqué Kim et Evens (1996), l'auteur d'un article de journal donne en général une première fois la forme complète du nom de personne accompagnée d'informations, puis des formes abrégées.

19,4 % de noms de personnes sont détectables à l'aide d'un contexte gauche (le plus souvent une civilité, une fonction ou un métier[12]), mais la majorité (60,1 %) le sont par une preuve interne qui est un prénom ou une particule de patronyme non ambiguë en français (ex. : *von, di,* etc.). On remarque que 8,6 % des noms de personnes sont accompagnés d'une double preuve : un contexte gauche et une preuve interne (ex : *la* ***danseuse Marie-Claude*** *Pietragala*).

45 % des noms de personnes[13] sont précédés d'un contexte contenant une civilité, un titre ou un nom de profession, suivis du patronyme. S'y ajoute éventuellement un prénom (ex : *le* ***président péruvien Alberto Fujimori***).

Le dictionnaire des noms de professions de Fairon (2000) nous sera d'une grande utilité pour l'extraction des noms de personnes. Ce dictionnaire nous servira de preuve externe pour les noms de personnes. Dans le cadre du projet Prolex, un dictionnaire des prénoms français et étrangers[14] (nommé ***prénom-prolex***) a été élaboré. Ce dictionnaire nous servira de preuve interne pour trouver les noms de personnes.

45 % des noms de personnes[15] n'ont pas de contextes descriptibles mais contiennent la preuve interne apportée par un prénom suivi du patronyme (ex. : **Adrien** Friez). La reconnaissance des prénoms se fait sur la base d'indices morphologiques et grâce à notre dictionnaire. Nous attachons une importance toute particulière à l'extraction des prénoms car générer une variante de nom de personne sera plus simple si on sait différencier le prénom du patronyme. Nous distinguons les prénoms simples (ex. : *Danièle*, *Louis*), les prénoms abrégés simples (ex. : *E.* pour *Emmanuel*), les prénoms composés (ex. : *Jean-Pierre*, *Charles Edouard*), ou en partie abrégés (ex. : *Pierre-J*), les prénoms composés abrégés (ex. : *J.P.*, *J.-P.*, *J-P*, *J-P*).

Les patronymes prennent les formes suivantes : les patronymes « simples » (ex. : *Dupont, Durand-Pérec*), les patronymes « composés » d'une particule ; ce sont surtout des noms d'origine étrangère (ex. : *Mac Donnell-Douglas, O'Ryan, von Bulow, El Amra,* etc.) mais aussi des formes françaises (ex. : *L'Huillier, Le Falch'un*), les patronymes français à particules (ex : *Dupont de Nemours, de Neuville, de la Fontaine*). Nous avons distingué les patronymes français à particule des patronymes composés car la parti-

cule française (*de*, *du*) est très ambiguë en français (avec la préposition *de*), ce qui n'est pas le cas pour les particules étrangères.

5 % des noms de personnes sont trouvés par : un contexte droit (beaucoup plus rare qu'un contexte gauche) ou grâce à la présence d'un verbe utilisé pour désigner une action mettant en jeu une personne (ex. : dire, expliquer, etc.) ; cependant, ces verbes peuvent être employés avec un sujet non humain (nom d'organisation) : cet indice est donc difficilement exploitable.

Les derniers noms de personnes (5 %) n'ont aucun contexte, même complexe, qui puisse les distinguer à coup sûr d'autres noms propres. Ces noms de personnes sans contexte sont principalement ceux de personnes très connues pour lesquelles l'auteur du texte estime qu'il n'est pas nécessaire de préciser le prénom, ni le titre ou la profession, ou ceux de personnes déjà citées dans l'article (ex. : ***Picasso*** *n'est pas le premier à passer à la postérité commerciale*). Nous envisageons pour les traiter de créer un dictionnaire de célébrités.

Noms d'organisations

Enfin, les **noms d'organisations** sont très nombreux et possèdent de nombreuses variantes (ex. : *Organisation des Nations Unies, Nations Unies, ONU*). De plus, ces noms apparaissent et disparaissent au gré de l'économie.

51,2 % des noms d'organisations commencent par une preuve interne : un premier mot capitalisé qui dénote d'un nom d'organisation (ex. : ***Fonds*** *Monétaire International*). Ce sont, pour la plupart, des noms propres à base descriptive. Ils sont formés de noms communs qui permettent de deviner qu'il s'agit d'une entreprise (ex. : *Société européenne des satellites*), d'une organisation (ex. : *Organisation mondiale de la santé*), d'une banque (ex. : *Banque de France*) ou autre (ex. : *Front populaire biélorusse, Union cycliste internationale*).

D'ailleurs, 7 % des noms d'organisations contenant une preuve interne sont en fait des noms étrangers (ex. : *Mellon* ***Foundation***, ***Bank*** *of America*), assez simples à trouver : ils portent des majuscules au début de chacun de leurs mots, ils peuvent aussi contenir des mots tels que *of, and* (ces mots ne sont absolument pas ambigus avec des mots français). À leur extrémité droite, ces noms propres ont assez souvent une preuve interne du type *Ltd, Research* etc. Les noms propres étrangers ne posent pas de problème de limite droite (ex. : Defense Intelligence **Agency**, European Language Resources **Association**).

1 % peuvent être trouvés grâce à leur morphologie, par exemple la présence d'une esperluette (ex. : *AT&T*).

La preuve externe gauche représente 12,7 % des noms d'organisation (ex. : ***filiale de*** *Vivendi*) et seulement 1,2 % pour le contexte droit. Les noms d'organisation annoncés par une preuve externe prennent surtout la forme de noms propres purs : ce sont principalement des noms d'entreprises. Leur preuve externe est en général constituée d'un mot tel que *groupe, agence, société*, etc. et éventuellement d'un adjectif toponymique (ex. : **groupe britannique** Cable & Wireless, **société** Suez-Lyonnaise des eaux, **compagnie** Airbus).

Noms de lieux

Les **toponymes** sont des noms propres relativement stables c'est-à-dire que le nom donné à tel ou tel toponyme change rarement (Piton et Maurel 1997), mais des modifications

se font au gré de l'histoire (*Châlons-sur-Marne* est devenu récemment *Châlons-en-Champagne)*. Les toponymes sont, de plus, en nombre assez limité.

Seuls 20 % des noms de lieux ont un contexte gauche et parfois droit (ex. : ***l'estuaire de la** Seine, **la mer** Baltique*) et quelques-uns une preuve interne. Ceux qui ont une preuve interne sont surtout des noms de ville ou de département (ex. : *Chaumont-**sur**-Loire* = les tirets et « *sur* » sont typiques d'un nom de ville, *Asie **du Sud-Est***) ou des noms de lieux étrangers (ex : *Trafalgar **Square**, Main **Street**, Yosemite National **Park***).

Pour trouver la plus grande partie des noms de lieux mais aussi des gentilés, nous allons principalement utiliser le dictionnaire *Prolintex*[16] de toponymes (Maurel et Piton 1999) réalisé dans le cadre du projet Prolex. Ce dictionnaire permettra de reconnaître de nombreux noms de lieux car ceux-ci sont difficiles à repérer par manque de preuve.

Tous les noms propres confondus

Par rapport à l'ensemble des résultats connus, notre système obtient les meilleurs résultats actuels sur le français (sur *Le Monde*, rappel de 93,2 % et précision de 94,4 %) et va pouvoir contribuer à l'extension des dictionnaires du projet Prolex de traitement automatique des noms propres, développé à l'université de Tours. Les noms propres, qui ne peuvent être catégorisés, peuvent au moins être repérés à l'aide d'indices syntaxiques et de la présence de lettres capitales. Il faudra utiliser d'autres moyens pour les typer ; par exemple, leur affecter le même type que leurs homonymes trouvés dans un même texte.

L'étude en corpus nous montre que les différents types de noms propres ne sont pas égaux en ce qui a trait à leur découverte car leurs contextes d'apparition et leur nombre dans les articles de journaux sont très différents. Il faut donc adapter les moyens d'identifier les noms propres en fonction de ces différences.

4. Conclusion

Les méthodes à base de règles permettent de trouver une grande partie des noms propres mais elles ont leurs limites : l'inévitable incomplétude de la grammaire explique une partie des erreurs ou des réponses manquantes. Les ambiguïtés ne sont pas toutes résolues, l'absence de contextes autour des noms propres empêche leurs repérage et catégorisation.

Ces travaux nous permettent d'envisager des recherches sur le traitement automatique des coréférences et anaphores. Mais nous prévoyons surtout des travaux autour de dictionnaires multilingues de noms propres ainsi qu'une application à la traduction automatique : en effet, la reconnaissance des noms propres et de leurs types est d'un très grand intérêt pour envisager de traduire automatiquement et correctement son contexte.

NOTES

1. Initié et coordonné par D. Maurel au Laboratoire d'Informatique de Tours.
2. Nous nous servons des ressources logicielles et linguistiques d'Intex pour développer notre système.
3. Dans la suite, nous utiliserons indifféremment les termes de **noms propres** ou **entités nommées** pour désigner les noms propres au sens large du terme.

4. *Npr* est l'abréviation du mot « nom propre » pour les linguistes.
5. **Anthroponymes** : patronyme, prénom, célébrité, divinité ou personnage, entreprise, association ou parti, établissement public ou privé, organisation internationale, gentilé ou ethnonyme.
   Toponymes : région, ville, groupe de pays, bâtiment, hydronyme, objet céleste, lieux divers.
   Ergonymes : appellation commerciale, entreprise (aussi des anthroponymes), œuvre, objet.
   Pragmonymes : phénomène météorologique, catastrophe, manifestation artistique ou sportive, fête, événement.
6. Message Understanding Conference, <http: //www.muc.saic.org>.
7. FUNES de Coates-Stephens (1993), FASTUS, Exoseme de Wolinski *et al.* (1995), etc.
8. Alembic, BBN IdentiFinder, etc.
9. LTG system, SemTex, Nemesis, etc.
10. Les discours testés sont des articles de l'encyclopédie américaine Grolier, du thésaurus Roget et du corpus Brown (Brown Corpus of Standard American English).
11. Ce travail a été réalisé sur un journal *Le Monde* complet, daté du 12 janvier 1999.
12. *Mme, Monsieur, président, ministre, général, lieutenant, cardinal, évêque, le juge, l'architecte, etc.*
13. Remarquons que cette proportion tombe à 33 % dans le journal *Ouest France* : les journalistes de ce journal ajoutent moins de détails sur la fonction des personnes citées.
14. La couverture actuelle de notre dictionnaire des noms propres sur le journal *Le Monde* représente 96 % des prénoms.
15. Dans le journal *Ouest France*, 59 % des noms de personnes sont accompagnés d'un prénom seul.
16. Ce dictionnaire contient plus de 100 000 entrées.

RÉFÉRENCES

Bauer, G. (1998) : *Deutsche Namenkunde*, Berlin, Weidler Buchverlag.

Daille, B. et E. Morin (2000) : « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », *Traitement Automatique des Langues* 41-3, p. 601-621.

Dister, A. (1997) : « Problématique des fins de phrase en traitement automatique du français », *in* Dehays, J., Rosier, L. et F. Tilkin (eds) *Champs Linguistiques*, Paris, Duculot.

Eggert, E., Maurel, D. et Belleil, C. (1998) : « Allomorphies et supplétions dans la formation des gentilés : application au traitement informatique », *Cahiers de lexicologie* 73-2, p. 167-179.

Fairon, C. (2000) : *Structures non-connexes. Grammaire des incises en français : description linguistique et outils informatiques*, Thèse de doctorat en informatique, Université Paris 7.

Forsgren, M. (1994) : « Nom propre, référence, prédication et fonction grammaticale », *in* Noailly, M. (éd.), *Nom propre et nomination* (Actes du colloque de Brest), p. 95-106.

Fourour, N. (2002) : « Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français », *Actes de TALN'2002*, Nancy, p. 265-274.

Friburger, N. (2002) : « Cascade de transducteurs pour INTEX : Un nouvel outil », In *5es Journées Intex*, Marseille, 11-12 juin 2002.

Friburger, N. et D. Maurel (2001) : « Finite-State Transducer Cascade To Extract Proper Nouns in Texts », In *Proceedings of 2nd Conference on Implementing and Application of Automata (CIAA'2001)*, Pretoria, 23-25 juillet 2001. (à paraître dans *LNCS*).

Friburger, N., A. Dister et D. Maurel (2000), « Améliorer le découpage des phrases sous Intex », Revue *Informatique et Statistique dans les Sciences Humaines (RISSH)*, 36-1/4, p. 181-199.

Gale, W. K., Church, K. and D. Yarowsky (1992) : « One sense per discourse », Dans *Proceedings of the DARPA Speech and Natural Language Workshop*, New York, Harriman, p. 233-237.

Garric, N. et D. Maurel (2000) : « Désambiguïsation des noms propres déterminés par l'utilisation des grammaires locales », *Revue française de Linguistique appliquée* 5-2, p. 85-100.

Gary-Prieur, M. N. (1994) : *Grammaire du nom propre*, Paris, Presse universitaire de France.

Grass, T. (2000) : « Typologie et traductibilité des noms propres de l'allemand vers le français à partir d'un corpus journalistique », TAL 41-3, p. 643-669.

Grass, T., Maurel, D. et O. Piton (2002) : « Description of a Multilingual Database of Proper Names », Dans *Proc. of Portal 2002, LNCS*, 23-26 juillet 2002, Faro, p. 137-150.

Grevisse, M. et A. Goosse (1986) : *Le Bon Usage*, Gembloux, Duculot.

Grishman, R. et B. Sundheim (1996) : « Message Understanding Conference – 6 : a brief history », Dans *Proc. of 16th International Conference on Computational Linguistics (COLING-96)*, Californie, Morgan Kaufmann, p. 466-471.

Jacquemin, C. et C. Bush (2000) : « Combining Lexical and Formatting Cues for Named Entity Acquisition from the Web », Dans *Proc. Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, Hong Kong, p. 181-189.

Jonasson, K. (1994) : *Le nom propre. Constructions et interprétations*, Paris, Duculot.

Maurel, D. et O. Piton (1999) : « Un dictionnaire de noms propres pour Intex : Les noms propres géographiques », *Linguisticae Investigationes* 22, p. 277-287.

McDonald, D. D. (1996) : « Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names », *in* Boguraev, B. and J. Pustejovsky (eds.) *Corpus Processing for Lexical Acquisition*, Cambridge, MIT, p. 32-43.

Molino, J. (1982) : « Le nom propre dans la langue », *Langages* 66, p. 5-21.

Noailly, M. (1991) : « « L'énigmatique Tombouctou » : nom propre et position de l'épithète », *Langue française* 92.

Paik, W., Liddy, E. D., Yu, E. et M. Mckenna (1996) : « Categorizing and Standardizing Proper Nouns for efficient Information Retrieval », *in* Boguraev, B. and J. Pustejovsky (eds.) *Corpus processing for lexical acquisition*, Cambridge, MIT, p. 61-73.

Piton, O. et D. Maurel (1997) : « Le traitement informatique de la géographie politique internationale », *Bulag*, numéro spécial, p. 321-328.

Rey-Debove, J. (1994) : « Nom propre, lexique et dictionnaires de langue », *in* Noailly, M. (ed.) *Nom propre et nomination* (Actes du colloque de Brest), p. 107-122.

Silberztein, M. (1993) : *Dictionnaires électroniques et analyse automatique de textes – Le système INTEX*, Paris, Masson.

Sundheim, B. M. (1995) : « Overview of Results of the MUC-6 Evaluation », *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, p. 13-31.

Trouilleux, F. (1997) : *Identification et classement automatique des noms propres en français*, Rapport de DEA, Clermont-Ferrrand.

Wolinski, F., Vichot, F. et B. Dillet (1995) : « Automatic Processing of Proper Names in Texts », *Proceedings of the Seventh Conference of the European Chapter of the Association for Computatinal Linguistics (EACL'95)*, Dublin, University College of Dublin, p. 23-30.

Zabeeh, F. (1968) : *What's in a Name? An Inquiry into the Semantics and Pragmatics of Proper Names*, La Haye, Martinus Nijhoff.