Les mots inconnus sont-ils des noms propres ?

Denis Maurel

LI (Université de Tours) – 64, avenue Jean-Portalis – 37200 Tours – France
denis.maurel@univ-tours.fr

Abstract

The voluminous textual data parsing is confronted with a specific class of words, the proper names, above all in case of business and strategic intelligence or information retrieval. The proper names are not often found in electronic dictionaries and one makes no distinction between them and unknown words. Meanwhile, instead of just declare to lump together proper names and unknown words, we present in this paper the result of a specific tagging experiment in every proper name and every unknown word of a newspaper issue.

Résumé

L'analyse de données textuelles volumineuses, surtout dans le cadre de la veille stratégique ou de la recherche d'information, est confrontée à une catégorie spécifique de mots, les *noms propres*, souvent absents des dictionnaires électroniques et, de ce fait, assimilés à des « Mots inconnus ». Cependant, au lieu de se contenter d'affirmations communes sur l'amalgame entre mots inconnus et noms propres, nous présentons ici les résultats d'une expérience de catégorisation de tous les mots inconnus et tous les noms propres d'un numéro du journal *Le Monde*.

Keywords: NLP, proper names, unknown words, tagging.

Mots-clés : TAL, noms propres, mots inconnus, étiquetage.

1. Introduction

L'analyse de données textuelles volumineuses, surtout dans le cadre de la veille stratégique ou de la recherche d'information, est confrontée à une catégorie spécifique de mots, les *noms propres*. Ceux-ci sont souvent absents des dictionnaires électroniques, alors qu'ils sont porteurs d'une riche sémantique tout en constituant, à eux seuls, plus de 10% des textes journalistiques (Coates-Stephens, 1993). En général donc, les noms propres rentrent dans la catégorie des « mots inconnus », c'est-à-dire non répertoriés par les différents dictionnaires de langues consultés, où se retrouvent les fautes d'orthographe, les abréviations, les dérivations peu communes, etc.

Par exemple, pour Ren et Perrault (1992), un mot inconnu commençant par une majuscule est considéré comme nom propre. Certains, moins expéditifs, utilisent des méthodes plus fines pour repérer les noms propres, souvent à l'aider de grammaires locales (Mac Donald, 1996 ; Mani et Richard MacMillan, 1996 ; Coates-Stephens, 1993) et proposent même parfois de classer les noms propres identifiés (Paik, *et al.*, 1996 ; Friburger et Maurel, 2001). En effet, la recherche et l'extraction d'informations ou l'aide à la traduction nécessitent de délimiter précisément les noms propres, de les catégoriser et même, parfois, de les relier entre eux, comme cela est proposé dans le projet *Prolex* (Maurel *et al*., 1996) sur le traitement automatique des noms propres. Dans le cadre de ce projet, aujourd'hui multilingue, nous avons défini une typologie des noms propres et de leur relations (Grass *et al.*, 2002), que nous utiliserons ici pour présenter nos résultats.

Cependant, au lieu de se contenter d'affirmations communes sur l'amalgame entre mots inconnus et noms propres, nous avons tenté une expérience et nous avons catégorisé tous les mots inconnus et tous les noms propres d'un journal (*Le Monde*, daté du 15 janvier 1999), après l'utilisation d'un logiciel d'analyse lexicale, le logiciel *Intex* (Silberztein, 1993), dans sa version pour le français de septembre 2001.

2. Définir les noms propres

Avant de chercher à reconnaître des *noms propres*, il est nécessaire de définir cette catégorie linguistique. Mais la définition des noms propres est loin de faire l'unanimité parmi les linguistes. La définition classique est celle que donne, par exemple, le *Bon Usage* (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. » Elle est relayée par les propos de Gary-Prieur (1994 : 7) : « Alors que l'interprétation d'un nom commun ne met en jeu que la compétence lexicale, celle du nom propre requiert presque toujours une mise en relation avec le référent initial, qui mobilise des connaissances discursives. »

Cette définition correspond bien à des noms de personne (*Chirac*) ou de lieu (*Belgique*, *Bruxelles*). Mais elle nécessite d'être complétée afin de rendre compte de la réalité. C'est (Jonasson, 1994) qui distingue les noms propres « purs » (correspondant la définition de Grevisse) et les noms propres « descriptifs » qui résultent souvent de la composition d'un nom propre avec une expansion, comme *Jardin des Plantes* ou *Organisation mondiale de la Santé*. Certains semblent être des descriptions définies figées ou en cours de figement. La classe des noms propres purs est relativement fermée, alors que celle des noms propres descriptifs est ouverte à la création lexicale.

Nous prendrons donc comme définition celle de Jonasson (1994 : 21) qui déclare *nom propre* : « Toute expression associée dans la mémoire à long terme à un particulier en vertu d'un lien dénominatif conventionnel stable. » Cette définition, plus souple que la précédente, permet d'inclure les noms propres descriptifs, qu'il est convenu dans le monde du traitement automatique des langues, depuis les conférences *MUC*, d'appeler *entités nommées*[1] (Chinchor, 1997).

Souvent les linguistes mettent en avant le caractère de « désignateur rigide » du nom propre. Ainsi, selon Kripke (1980), un nom propre doit « désigner le même particulier dans tous les mondes possibles » et, selon Kleiber (1981 : 316), il n'est pas lié « aux situations passagères et aux propriétés accidentelles que peut connaître un particulier ». Cela devrait permettre de distinguer le nom propre (notamment descriptif) de la description définie, puisque le référent d'un nom propre peut évoluer sans que le nom propre utilisé ne soit changé.

Il est clair que *le président Chirac* n'est pas un nom propre, mais une description de la fonction politique exercée par l'homme qui a pour nom propre *Chirac*. Cependant, cette distinction n'est plus vraie pour les formes longues des noms de pays où la description est figée par des considérations politiques (Piton et Maurel 1997). Le dernier exemple en date est celui du *Zaïre* qui est devenu, suite à un coup d'état, la *République démocratique du Congo*. Peut-on dire que le référent (le pays africain désigné) n'est plus le même qu'avant ? Cela semble difficile à admettre. Et, dans une perspective de recherche d'information, il faudrait sans doute associer les deux termes.

[1] Pour être exact, les entités nommées comprennent non seulement les noms propres, mais aussi les dates et les mesures.

Enfin, il nous faut signaler l'importance de la métonymie des noms propres dans les textes journalistiques : *Paris*, *France*, *Français* peuvent désigner le même référent, le *gouvernement français*, et non leur propre référent. C'est ce qui nous a conduit, dans le projet *Prolex*, à envisager le nom propre comme une entité relationnelle (Maurel *et al*., 2000).

Enfin, en français, le critère formel le plus remarquable du nom propre, qui le distingue du nom commun ou de la description définie, est la majuscule. Celle-ci « indique le début d'une expression linguistique qui, d'un certain point de vue, constitue un tout » (Gary-Prieur, 1991 : 22). Mais cette majuscule n'apparaît pas sur tous les termes d'une entité nommée, ce qui pose le problème de la reconnaissance de sa limite droite (Friburger, 2002).

3. Le logiciel *Intex*

Le logiciel *Intex* est basé sur plusieurs bases de connaissance, hiérarchisées en quatre niveaux :

1. Les caractères, qui sont regroupés en trois classes : les chiffres, les lettres (définies dans un fichier d'alphabet) et les séparateurs (les autres caractères).
2. Les phrases, qui sont décrites par un transducteur à nombre fini d'états, chargé de découper le texte.
3. Les séquences de lettres (ou *mots simples*), qui sont répertoriés dans un dictionnaire de formes fléchies comportant 675 249 entrées, le *Delaf* (Courtois et Silberztein, 1990).
4. Les séquences de mots (ou *mots composés*), qui sont aussi répertoriés dans des dictionnaires, mais que nous n'utiliserons pas dans cette étude.

Après l'analyse lexicale par *Intex*, avec les bases 1, 2 et 3 ci-dessus, du journal *Le Monde* daté du 15 janvier 1999, nous avons obtenu les résultats suivants :

- 3 945 phrases.
- 114 217 éléments (dont 14 672 différents), répartis en séquences de lettres (86 409, dont 14 633 différentes), en chiffres (4 875) et en séquences de séparateurs (22 933, dont 29 différentes).
- 12 716 mots simples (présents dans le dictionnaire Delaf) et 1 917 mots inconnus.

Précisons ce que le logiciel *Intex* désigne par *mot inconnu* : il s'agit d'un mot formellement non présent dans le dictionnaire *Delaf*. Par exemple, la consultation du dictionnaire pour la phrase : « J'ai aperçu Jean, Pierre, Denis et Agnès. » donnera les résultats de la *Figure 1*.

Mots simples[2]	Mots inconnus
ai,avoir.V+z1:P1s aperçu,apercevoir.V+z1:Kms aperçu,aperçu.N+z1:ms et,et.CONJC+z1 jean,jean.N+z1:ms pierre,pierre.N+z1:fs pierre,pierrer.V+z2:P1s:P3s:S1s:S3s:Y2s	Agnès Denis

Figure 1. Les listes de mots après la consultation du dictionnaire

[2] Le *j'* a été reconnu lors de la préanalyse comme une élision non ambiguë (*j',je.PRO+PPV:1ms:1fs*), c'est pourquoi il n'apparaît pas ici.

Ainsi, à cette étape, *Pierre* ne sera pas un mot inconnu, à cause de *la pierre* et du verbe (rare[3]) *pierrer*, *Jean* non plus, à cause du pantalon, *le jean*, alors que *Agnès* et *Denis* le seront.

4. Les mots inconnus

Les mots inconnus de ce journal représentent 4 % de l'ensemble des mots simples du texte et 13 % des mots distincts (*Figure 2*).

Mots simples	Tous les mots		Les mots distincts	
Reconnus	82 524	96%	12 716	87%
Inconnus	3 885	4%	1 917	13%
Total	86 409	100%	14 633	100%

Figure 2. La répartition des mots simples du journal

Sur les 1 917 mots inconnus distincts, 243 ne sont pas des noms propres. Ceux-ci représentent donc 87 % des mots inconnus distincts. Cependant, sur les 243 mots inconnus distincts qui ne sont pas des noms propres, seul 23 commencent par une majuscule. En assimilant les 1 940 mots inconnus commençant par une majuscule à des noms propres, on ne se trompe que d'un peu plus de 1 %.

La principale source de mots inconnus (108, soit plus des deux cinquièmes) provient de la création lexicale :

- 21 abréviations : *Apocal*, *convalo*, *Corresp*, *éd*, *ème*, *gouv*, *MM*, *Mme*, *NDLR*, *rens*, *Tél*, etc.
- 11 chiffres romains[4] : *III*, *VIIIe*, *XIIe*, *XIXe*, *XVe*, *XVIe*, *XVIIe*, *XVIIIe*, *XXe*, *XXIe*, *XXs*.
- 15 dérivés de noms propres[5] : *ajaccien*, *belgitude*, *braudélienne*, *brejnévisme*, *cantabrique*, *delorien*, *khyal*, *lommoise*, *Newtonianisme*, *pivertisme*, *venézuélienne*, *Villeneuvoises*, *villiériste*, *wallisien*, *wolfienne*.
- 40 néologismes : *anthropogammamétrique*, *athéologique*, *autofiction*, *cacateux*, *europhorie*, *europrogressistes*, *germanité*, *médusante*, *métropolisation*, *milléniale*, *parentalité*, *pornocrate*, *rapporteure*, *remixage*, *surdemandés*, *surinterprétation*, etc.
- 5 onomatopées : *Epopopoï*, *Io*, *io*, *ito*, *ouba*.
- 14 sigles : *BB*, *CAFB*, *CMU*, *DG*, *HLM*, *NC*, *PDG*, *PIB*, *Ran*, *SBK*, *SDF*, *SMIC*, *TBP*, *ZEP*.
- 1 terme : *Erasmus*.

On totalise plus de la moitié des mots inconnus, si on ajoute à cette liste 31 emprunts :

- 4 mots latins : *exempla*, *Librium*, *tempora*, *vademecum*.
- 27 mots étrangers : *cyborgs*, *daddy*, *Euroland*, *joke*, *moderno*, *private*, *reals*, *sanpei*, *Seises*, *that*, *toros*, *völkisch*, etc.

Ensuite viennent les erreurs (53, un peu plus d'un cinquième) qui se répartissent entre :

[3] Le dictionnaire *Delaf* est divisé en trois couches, les mots courants (+z1), les mots rares (+z2) et les mots techniques (+z3).

[4] Le logiciel *Intex* permet l'utilisation d'un transducteur étiquetant les chiffres romains, mais nous n'avions pas sélectionner cette option, pour nous limiter aux seuls mots du dictionnaire *Delaf*.

[5] Dans le cadre du projet *Prolex*, nous avons étudié en détail la formation des gentilés (noms d'habitants), à partir d'un toponyme (Eggert *et al*., 1998 ; Eggert, 2002).

- 5 accents incorrects : *Nevadá*, *García*, *Martín*, *Vázquez*, *Louýs*.
- 9 fautes de frappe : *ccomptes*, *cespremiers*, *Cest*, *Cressson*, *Ilse*, *lerapproche*, *nérations*, etc.
- 38 fautes d'orthographe : *agrave*, *aseptie*, *automonie*, *cathéchiste*, *élevent*, *milllions*, *mutelles*, *secretaire*, *télévison*, *viendont*, etc.

Enfin, les 52 mots inconnus restants sont le résultat d'un mauvais paramétrage de la version d'*Intex* utilisée : les lettres majuscules accentuées ne figuraient pas dans le fichier d'alphabet et étaient donc assimilées à des séparateurs. La version actuelle du logiciel a corrigé ce défaut. La *Figure 3* présente cette répartition, secteur par secteur.

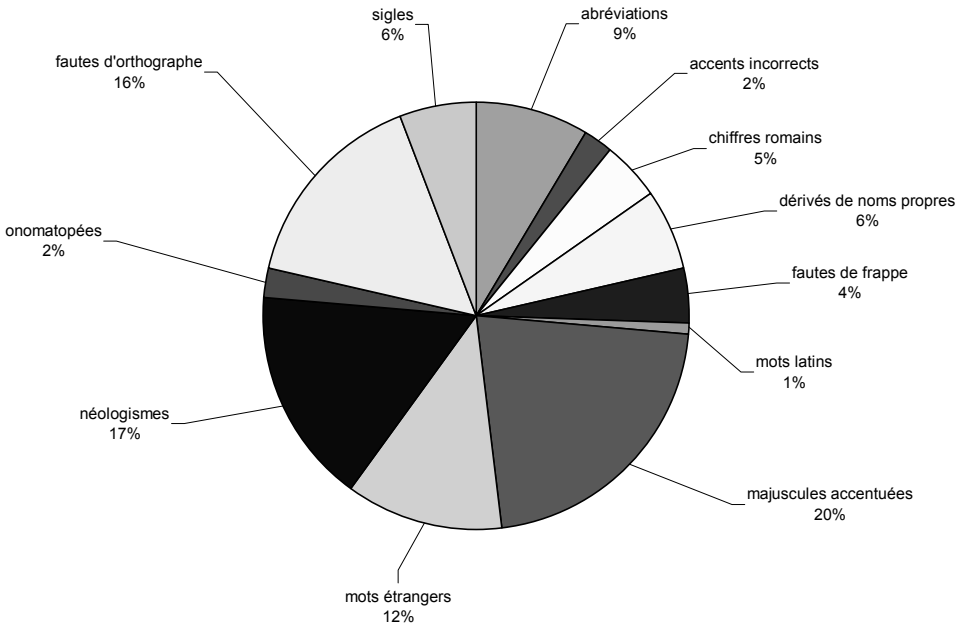

Figure 3. Les mots inconnus qui ne sont pas des noms propres

5. Les noms propres

Nous avons vu à la section 3 que les noms propres forment un ensemble de mots distinct de celui des mots inconnus, à cause de l'homographie de certains d'entre eux avec des noms communs. Mais, si nous suivons la définition de Jonasson (voir section 2), il y a plus : les noms propres sont souvent des mots composés, alors que les mots inconnus sont des mots simples. Le décompte des uns et des autres est donc radicalement différent, comme le montrent les exemples de la *Figure 4*.

Noms propres	Nombre de mots inconnus
Banque Centrale Européenne	0
Banco Real	1
Banca commerciale italiana	2[6]
Banca di Roma	3

Figure 4. Des noms propres et des mots inconnus

[6] Évidemment, les deux mots *Banco* et *commerciale* ne sont pas reconnus pour eux-même, mais ils possèdent des homographes en français.

Précisons le résultat donné à la section précédente : 243 mots inconnus distincts (sur 1 917) ne sont pas des noms propres, mais les 1 674 mots inconnus restant ne sont pas tous des noms propres purs, mais peuvent être simplement des parties de noms propres descriptifs. Pour connaître le nombre exact de noms propres du journal *Le Monde* du 15 janvier 1999, nous les avons effectivement dénombrés, par supervision humaine[7], pour arriver au chiffre de 1 419. Remarquons que, parmi ceux-ci, seuls 1 033 sont des mots simples. On constate donc qu'assimiler les 1 940 mots inconnus commençant par une majuscule avec des noms propres se révèle être très approximatif… Dans l'exemple de *Banco Real* (*Figure 4*), *Banco* n'est pas un mot inconnu et *Real* n'est pas un nom propre. C'est le mot composé *Banco Real* qui est intéressant à catégoriser.

Nous les avons ensuite classés en suivant notre typologie à deux niveaux (Grass *et al.*, 2002).

Après avoir étudié différentes classifications[8] concernant les noms propres, trop *ad hoc*, puis avoir envisagé d'utiliser le système des classes d'objets (Le Pesant et Mathieu-Colas, 1998), trop étendu pour être utilisé pour les seuls noms propres, nous avons préféré créer notre propre typologie, plus adaptée au traitement linguistique et notamment lexico-sémantique. Nous avons donc choisi de définir quatre *hypertypes*, qui reprennent en fait les traits classiques : les anthroponymes (trait *humain*), les toponymes (trait *locatif*), les ergonymes (trait *inanimé*) et les pragmonymes (trait *événement*) ; et, ensuite, vingt-neuf *types*, qui sont plus des *commentaires* portés sur les hypertypes, commentaires destinés principalement à l'aide à la traduction ou à la recherche d'information.

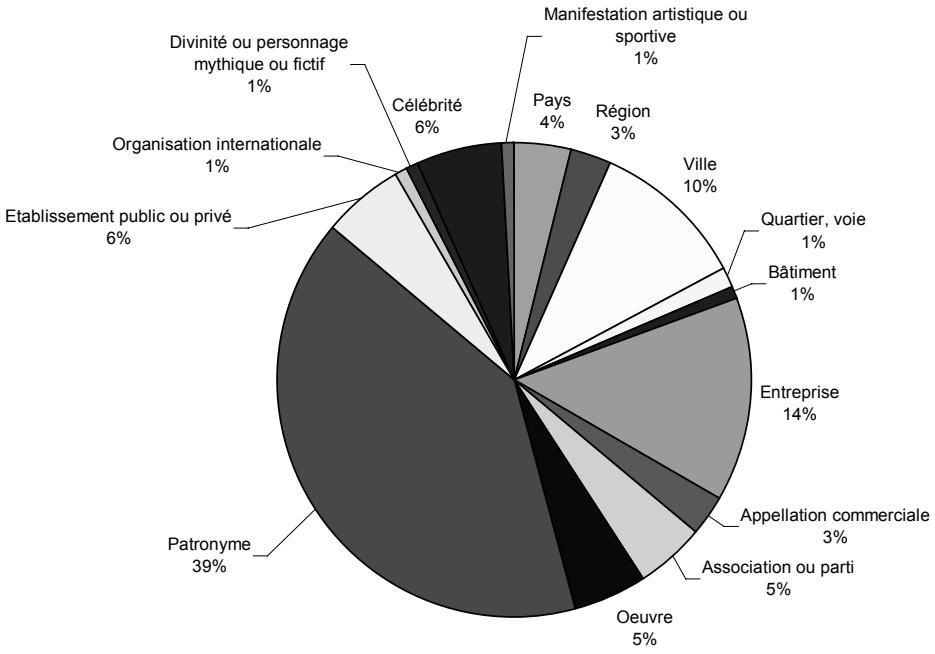

Figure 5. Les types des noms propres rencontrés

Parmi ces noms propres, comme on peut s'y attendre dans un quotidien, on trouve essentiellement des anthroponymes (1 010, presque les trois quarts), dont deux tiers de noms de personne et un tiers de noms collectifs. Viennent ensuite 284 toponymes (un cinquième de l'en-

[7] L'auteur remercie Caroline Ferré et Cécile Romanet pour leur aide dans le dépouillement des données.

[8] Issues de l'onomastique (Bauer, 1985), de MUC (Chinchor, 1997 ; Paik *et al.*, 1996), de l'économie (Zabeeh, 1968) et de la traduction (Ballard, 2001).

semble) et 108 ergonymes (un dixième), puis, très minoritaires, 17 pragmonymes. Le détail des types rencontrés est présenté sur la *Figure 5*, pour autant qu'ils représentent au moins 1 % du total ; de ce fait, quinze types seulement y sont présents (huit types sont sous-représentés et six sont absents du journal). Bien sûr, ces chiffres souffrent de la petitesse de notre corpus qui, dépouillé manuellement, ne pouvait être trop important.

Cette présence de nombreux noms propres composés (27%) justifie pleinement la mise en œuvre de techniques de reconnaissance adaptées à cette classe de mots. Dans le cadre du traitement automatique des langues, nous nous sommes attachés à développer des outils permettant de délimiter correctement les noms propres descriptifs.

Nous utilisons pour cela des cascades de transducteurs à nombre fini d'états (voir un exemple sur la *Figure 6*). Les résultats obtenus, cette fois-ci sur un corpus plus important (1 Mo de corpus journalistique) sont présentés dans (Friburger, 2002) et résumés sur la *Figure 7*.

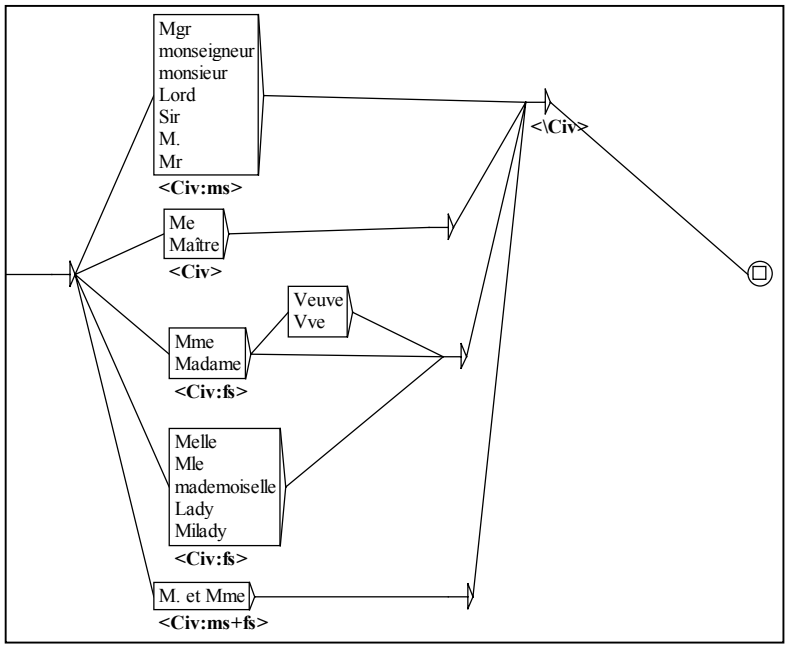

Figure 6. Un exemple de transducteur de la cascade : la reconnaissance des civilités

Reconnaissance automatique des noms propres Résultats obtenus sur 1 Mo de corpus journalistique	
Rappel	93,2
Précision	94,4

Figure 7. Les résultats obtenus par notre cascade

6. Conclusion

Notre étude, d'abord automatique, à l'aide du logiciel *Intex*, puis manuelle, d'un numéro du journal *Le Monde*, permet de vérifier précisément sur un exemple réel un certain nombre

d'affirmations souvent rencontrées, mais jamais chiffrées, concernant les mots inconnus, les noms propres et le rapport entre les deux.

Nous avons vu que, si les mots inconnus capitalisés sont effectivement presque toujours des parties de noms propres (à 1,2 % près), le décompte de ces derniers n'a aucun lien véritable avec les premiers, du fait que les noms propres descriptifs sont souvent composés en partie de noms communs. L'importance de la présence des noms propres a aussi été chiffrée précisément. Il n'est pas sûr qu'il soit pertinent de comparer ce nombre, qui est constitué, pour une grande part, de mots composés, à celui des mots simples du texte, mais, ce faisant, on obtiendrait alors 11 % de noms propres sur l'ensemble des mots distincts du journal. L'assimilation des mots inconnus commençant par une majuscule avec des noms propres est, elle aussi, très sujette à caution. La seule approximation acceptable est de les reconnaître comme partie de nom propre.

Références

Ballard M. (2001). *Le nom propre en traduction*. Ophrys.

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

Chinchor N. (1997). *Muc-7 Named Entity Task Definition*. consultable sur le site http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendices.

Coates-Stephens S. (1993). *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*. Kluwer Academic Publishers.

Courtois B., Silberztein M. (1990). Dictionnaires électroniques du français. *Langues française*, vol. (87) : 11-22.

Eggert E. (2002). *La dérivation toponymes-gentilés en français : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique*. Thèse de doctorat en linguistique.

Eggert E., Maurel D. et Belleil C. (1998). Allomorphies et supplétions dans la formation des gentilés. Application au traitement informatique. *Cahiers de Lexicologie*, vol. (73) : 167-179.

Friburger N. (2002), *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*. Thèse de doctorat en informatique, Université de Tours.

Friburger N. et Maurel D. (2001). Élaboration d'une cascade de transducteurs pour l'extraction de motifs : l'exemple des noms de personnes. In *Actes de la Huitième conférence annuelle sur le traitement automatique des langues naturelles (TALN 2001) :* 183-192.

Gary-Prieur M.-N. (1991). Le nom propre constitue-t-il une catégorie linguistique ? *Langue française*, vol. (92) : 4-25.

Gary-Prieur M.-N. (1994). *Grammaire du nom propre*. PUF.

Grass T., Maurel D. et Piton O. (2002). Description of a multilingual database of proper names. In *Proceedings of PorTal 2002*, LNCS.

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

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

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Klincksieck.

Kripke S. (1980). *Naming and Necessity.* Harvard University Press.

Le Pesant D. et Mathieu-Colas M. (1998). Introduction aux classes d'objets. *Langages*, vol. (131) : 6-33.

MacDonald D. (1996). Internal and external evidence in the identification and semantic categorisation of Proper Names. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 21-39

Mani I. et Richard MacMillan T. (1996). Indentifying Unknown Proper Names in Newswire Text. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 41-59.

Maurel D., Belleil C., Eggert E. et Piton O. (1996). Le projet PROLEX. In Séminaire *Représentations et Outils pour les Bases Lexicales, Morphologie Robuste de l'action Lexique du GDR-PRC CHM* : 164-175.

Maurel D., Piton O. et Eggert E. (2000). Les relations entre noms propres : lieux et habitants dans le projet *Prolex*. *Traitement automatique des langues*, vol. (41/1) : 623-641.

Paik W., Liddy E. D., Yu E. et McKenna M. (1996). Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 61-73

Piton O. et Maurel D. (1997). Le traitement informatique de la géographie politique internationale. In *Colloque Franche-Comté Traitement automatique des langues (FRACTAL 97)*, in *Bulag* (numéro spécial) : 321-328.

Ren X. et Perrault F. (1992). The typology of Unknown Words : An Experimental Study of Two Corpora. In *Proceedings of COLING 92*.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX*. Masson.

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