# Incremental Recognition and Referential Categorization of French Proper Names

# Nordine Fourour, Emmanuel Morin and Béatrice Daille

Institut de Recherche en Informatique de Nantes 2, chemin de la Houssinière - BP 92208 44322 Nantes Cedex 3, France {fourour,morin,daille}@irin.univ-nantes.fr

#### Abstract

This paper presents Nemesis, a French proper name (PN) recognizer for Large-scale Information Extraction (IE), whose specifications have been elaborated through corpus investigation both in terms of referential categories and graphical structures. The graphical criteria are used to identify proper names and the referential classification to categorize them. The system is a classical one: it is rule-based and uses specialized lexicons without any linguistic preprocessing. Its originality consists on a modular architecture which includes a learning process. The system up to now recognizes anthroponyms and toponyms with performance achieving 95 % of precision and 90 % of recall.

### 1. Introduction

The automating of French proper name recognition is a recurring problem in the different domain of Natural Language Processing (NLP): Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT) (Daille and Morin, 2000). This challenge has been successfully taken up in the area of IE from newswire English texts. Thus, proper name recognizers reported in recent Message Understanding Conference (MUC-7, 1998) achieved respectable performance (more than 90% in both precision and recall).

In French proper name recognition, the delimitation of the boundaries of the proper name exhibits structural ambiguities (in the same way as in English (Wacholder et al., 1997)). On one hand, adjectival modification, prepositional attachment and coordination scope must be resolved. For instance, the prepositional phrase des nations in Le Championnat d'Europe des nations a lieu... (lit. European Championship of the nations takes place...) is included into the proper name, whereas it is not included in Les Premiers Ministres des nations se sont rencontrés... (lit. Prime Ministers of the nations met themselves...). On the other hand, proper names composed of several lexical units can include no capitalized words. In addition to these ambiguities, arises the problem of the composition: a complex PN can be composed of one or more common nouns and an already identified simple PN, belonging to another referential category such as a historical event, a cultural, political, or sports event (e.g. Guerre d'Algérie (lit. War of Algeria)), an educational and research institution (e.g. Université de Nantes (lit. University of Nantes)).

The remainder of this article is organized as follows. In section 2., we propose a referential categorization of the proper names based on the classification put forward by Grass (2000), and then we introduce a graphical categorization of the proper names adapted to their delimitation. Section 3. describes the system for French anthroponyms and toponyms recognition. Section 4. gives the contribution to proper name recognition of each of the components of the Nemesis system. In section 5., the extension of the recognition to other categories thanks to referential composition is presented. Section 6. presents our conclusion and future work.

# 2. Corpus investigation

We present the results of an investigation on the numerical representation of the proper names according to their referential and graphical categories<sup>1</sup>. This investigation was carried out on a corpus consisting of a set of randomly selected documents from *La Recherche*<sup>2</sup> (17,067 words) and *Le Monde*<sup>3</sup> (20,866 words).

#### 2.1. Referential categorization

The goal of this section is to draw up a fine, domain independent, and stable categorization of proper names.

In IE, PN's are usually classified as organization, person, location and time names (MUC-7, 1998). Although this semantic categorization covers a large part of proper names present in newswire texts, it is rather limited and not really adapted for Large-scale IE. For translation, an exhaustive classification, based on referential criteria, is proposed by Grass (2000) to determine how the proper names should be handled through the translation process. Depending on the referential category, the proper name will be translated, not-translated or transposed. This classification is composed of 5 classes and 28 categories:

Anthroponyms names of persons or groups;

Toponyms names of location;

Ergonyms manufactured objects and products;

Praxonyms historical events and diseases;

Phenonyms natural disasters, stars and comets.

<sup>&</sup>lt;sup>1</sup>The numerical results were achieved manually. All PN's were identified, categorized and counted.

<sup>&</sup>lt;sup>2</sup>Corpus of texts *La Recherche* - year 1998 - distributed by ELRA (http://www.icp.inpg.fr/ELRA)

<sup>&</sup>lt;sup>3</sup>Corpus of texts *Le Monde* - year 1997 - European Corpus Initiative (ECI) distributed by ELRA

	La Recherche		Le Monde	
	# Occ.	Proportion	# Occ.	Proportion
Anthroponyms	194	52.0 %	1,066	73.8 %
Surname	97		437	
First name	66		310	
Ethnonyms	15		37	
* Organizations	16		194	
<ul> <li>Artistic groups</li> </ul>	0		87	
Pseudonyms	0		1	
* Names of pets	0		0	
Toponyms	107	28.7 %	270	18.7 %
* Toponyms > Countries	53		17	
Countries	22		73	
* Countries < Toponyms > Cities	17		33	
Cities	10		108	
Microtoponyms	0		16	
Hydronyms	4		9	
Oronyms	0		0	
Streets	0		4	
Deserts	1		0	
Edifices	0		14	
Ergonyms	64	17.2 %	93	6.4 %
Sites of production	0		0	
Trademarks and products	31		37	
Industrial companies	0		4	
Cooperatives	0		0	
Educational and research institution	27		7	
Military installations	0		1	
* Intellectual works	6		44	
Praxonyms	3	0.8 %	16	1.1 %
Historical events	0		0	
Diseases	0		0	
* Cultural, sports, and political events	0		15	
* Historical period	3		1	
Phenonyms	5	1.3%	0	0.0 %
Natural disaster	0		0	
Starts and comets	5		0	
Total	373		1,445	

Table 1: Distribution of Proper Names according to referential categorization

Grass' classification has been verified and enhanced through corpus analysis of French corpora. Table 1 shows the distribution of proper names according to the referential criteria, where asterisked categories have been modified. For instance, we create the category intellectual works, which groups title of books, movies, and publications and other creations such as theorems, laws, plans, etc. We found that the above six categories correctly account for about 80 % of all PN's present in our corpus: the more voluminous classes are anthroponyms and toponyms (64 % and 20 % of all PN's). Moreover, we noticed that some PN's belonging to other referential categories are built from these two classes. Thus, we decided to concentrate in a fi rst step on the processing of the anthroponym and toponym classes.

#### 2.2. Graphical categorization

As outlined in the introduction, we introduce a graphical categorization in order to estimate the proportion of complex PN's composed of several words with or without capital letters. This categorization is useful in a perspective of automatic recognition of PN's. Indeed, according to the written form of the PN's, different processing will be set up for their identification and categorization.

This categorization is based on the terminology of (Jonasson, 1994). The 5 following structures should be met:

- **pure simple proper names** Proper names composed of one lexical unit beginning with capital letter such as *France*, *Aristote*.
- **pure complex proper names** Proper names composed of several lexical units beginning with capital letter such as *Conflans Saint-Honorine*. In this category, we distinguish the form **First name Surname** composed of one ore more first names and one lexical unit beginning with a capital letter. This form refers to person

	La Recherche		Le Monde	
	# Occ.	Proportion	# Occ.	Proportion
pure simple proper names	145	45.6 %	313	31.1 %
pure complex proper names	25	7.9 %	89	8.8 %
first name surname	68	21.4 %	299	29.7 %
slightly mixed proper names	21	6.6%	35	3.5 %
mixed proper names	44	13.8 %	144	14.3 %
acronyms	15	4.7 %	127	12.6 %
Total	318		1,007	

Table 2: Distribution of Proper Names according to graphical criteria

names such as Alexandre Jardin.

- **slightly mixed proper names** Proper names composed of several words beginning with a capital letter and including grammatical words (conjunction, preposition, etc.) such as *Jardin des Plantes* (lit. Botanical Garden).
- **mixed proper names** Proper names composed of several lexical units where at least one of them begins with a capital letter such as *Comité international de la Croix-Rouge* (lit. International Red Cross Committee).
- **acronyms** Proper names composed of one uppercase lexical unit where each letter corresponds to a lexical unit.

Table 2 presents the distribution of our proper names according to the graphical criteria. This distribution shows three levels of dispatching: 60% (pure simple proper names and fi rst name surname), 15% (mixed proper names), and 15% (pure complex and slightly mixed proper names). From these categories, pure simple and complex proper names are easy to identify whereas mixed (slightly or not) proper names are more difficult to delimit. Acronyms are more present in *Le Monde* (12.6%) than in *La Recherche* (4.7%). Since acronyms are composed of capital letters, they are easy to identify but difficult to classify without knowing the extended form.

Nemesis, our PN recognizer, is based on this corpus investigation.

# 3. System for French anthroponyms and toponyms recognition

In order to delimit and classify PN's belonging to the anthroponym and toponym classes, we developed Nemesis a system which consists of a pipelined architecture and is made up of four modules: a lexical preprocessor, a grammar rule processor (fi rst pass), a learning processor, and another grammar rule processor (second pass). The global structure of this system is illustrated in fi gure 1.

#### 3.1. Lexical preprocessor

Lexical preprocessing is a two-step process. First, text is segmented into occurrences of sentences and forms. Second, acronyms are detected and associated to their extended form by exploring local structures (the detail algorithms are presented in (Morin, 1999)). For instance, when an acronym is introduced in a text, its extended form is given using parentheses: *CJCN (Confédération des Jeunes Chercheurs Nantais)* (lit. Union of Young Researchers from Nantes). Extended forms correspond to mixed proper names which are useful to classify acronyms. Therefore, the recognition of acronyms and their extended form allows to identify mixed proper names and to classify the acronyms.

Among the different PN recognizers we study, only Wolinski et al. (1995) and Wacholder et al. (1997) use this association, but only to solve some coreferences as far as Wacholder et al. (1997) are concerned.

#### **3.2.** Rewriting rule processor (first pass)

The grammar rule process is also a two-step process: first, a projection of the specialized dictionaries, and second, a recognition and a categorization of the proper names through the rewriting rules.

#### 3.2.1. Lexicons projection

As outline by McDonald (1994) and Wakao et al. (1996), the use of specialized lexicons is the base of every PN recognizer. The projection of our lexicons is divided into three parts:

- 1. conversion of the text into invert file (Salton and McGill, 1983);
- 2. projection: the forms receive the semantical tags linked to the dictionaries (60,530 words divided into 42 fi les);
- forms not tagged by the lexicons and whose initial letter is capitalized are tagged with NP.

Our lexicons were constructed either manually or automatically by exploiting textual resources (*Web* pages, etc.). The items of the dictionaries are of several types and have different functions:

- PN refers to known proper names such as *OMS* (lit. WHO), Alexandre. It allows the recognition and the categorization of the PN;
- Context belongs to the immediate left context, but not to the PN (e.g. *philosophe*, *français*). It allows the delimitation of the left boundary and the categorization. Thus the proper name *Descartes* within the nominal phrase *le philosophe français Descartes* (lit. the



Figure 1: Global structure of the Nemesis system

Content of the levicon	Functions	Number of elements
Content of the textcon	Tunctions	Number of ciefficities
geographical adjectives	end of the PN, element of the PN	100
nationality adjectives	context, end of the PN, element of the PN	438
microtoponym keys	trigger word	6
organization name keys (capitalized)	trigger word, end of the PN, element of the PN	119
organization name keys (lower case)	end of the PN, element of the PN	117
building name keys	trigger word	21
hydronym keys	context	35
oronym keys	trigger word	18
street name keys	trigger word	8
miscellaneous toponym keys	trigger word	8
initials	trigger word	26
institution names	PN, end of the PN, element of the PN	37
media names	PN end of the PN, element of the PN	111
profession names	context, end of the PN, element of the PN	162
continent names	PN, trigger word, end of the PN, element of the PN	20
département names	end of the PN, element of the PN	134
US state names	PN, trigger word, end of the PN, element of the PN	53
Canadian province names	PN, trigger word, end of the PN, element of the PN	14
hydronym names	PN, trigger word, end of the PN, element of the PN	250
sea and ocean names	PN, trigger word, end of the PN, element of the PN	91
river names	PN, trigger word, end of the PN, element of the PN	110
oronym names	PN, trigger word, end of the PN, element of the PN	35
nationality names	PN, trigger word, context, end of the PN	416
country names	PN, trigger word, end of the PN, element of the PN	287
région names	PN, trigger word, end of the PN, element of the PN	25
names of toponyms > countries	PN, trigger word, end of the PN, element of the PN	23
names of towns < toponyms < countries	PN, trigger word, end of the PN, element of the PN	119
foreign town names	PN, end of the PN, element of the PN	1,790
French town names	PN, end of the PN, element of the PN	46,484
dynastic numbers	trigger word, end of the PN, element of the PN	21
handles (capitalized)	element of the PN	36
handles (lower case)	element of the PN	29
political parties	PN, end of the PN, element of the PN	25
cardinal points	end of the PN, element of the PN	16
fi rst names	PN trigger word end of the PN element of the PN	9 216
sport names	end of the PN element of the PN	7
administrative titles	context	2
civil titles	context	9
titles of civility	context	31
military titles	context	45
titles of nobility	context	
religious titles	context	13

Table 3: List of the lexicons used with size and functions

French philosopher Descartes) is captured thanks to the contexts *philosophe français*;

- Trigger Word is used in the same way as context but belongs to the PN (e.g. F'ed'eration). Thus the proper name *Fédération francaise de football* (lit. French Football Federation) is captured thanks to the trigger word *Fédération*;
- End of the PN is the last form belonging to the PN (e.g. *football, régional*) and only allows the delimitation of its left boundary;
- Element of the PN gathers all lexical forms able to belong to the PN, but without allowing its delimitation nor its categorization.

Table 3 compiles a register of the lexicons used for the recognition of the anthroponym and toponym classes, with

their size and functions. The same lexicon can have various functions, even for the same category: the elements of the lexicon containing nationality names are used as PN's (*Français* (lit. French)), and trigger words (*Allemande de l'Ouest* (lit. West German)) for the ethnonyms, and as context (*Français René Descartes* (lit. French Ren 'e Descartes)) for the patronyms.

#### 3.2.2. Rewriting rule application

After the projection of the lexicons, the rewriting rules, which match the proper names only using internal evidence (McDonald, 1994), parses the tagged text.

There are 38 rules out of 53 using contexts or trigger words:

 $Hydro_key$  Article\_min+ [NP+]  $\rightarrow$  Hydronym (e.g. *rives de la Kamogawa* (lit. banks of Kamogawa)).

Other rules only use PN: [\$Media+] → Artistic\_group (e.g. France Football or Sports Illustrated).

#### 3.3. Learnig processor and second pass

The update of the lexicons was tackled by some systems (Poibeau, 1999; Cucchiarelli et al., 1998), but still raises many problems. The first contribution of this module is to resolve some coreferences. Indeed, coreferences are a recurrent problem for proper name processing: for example *M. Lang, Jack Lang, le ministre J. Lang* (lit. the minister J. Lang), *Lang*, etc. are different ways of naming the same person *Jack Lang*. The same problem affects the organization names: *Ligue des communistes de Yougoslavie* (lit. Communist League of Yugoslavia), *LCY, Ligue*, etc.

To carry out the recognition and categorization of the forms that have no contextual elements, we set up a heuristic-based method to update lexicons.

Example of such heuristics are:

- Given the form  $N_1 N_2 N_3$ , if  $N_2$  is a first name,  $N_3$  is either a surname or an uppercase initial form, and  $N_1$  ends with a suffix such as *ois*, *ais*, etc., then  $N_1$  is added to ethnonyms lexicon: *Marseillais* from the form *le Marseillais Robin Huc* (lit. Robin Huc from Marseilles).
- Given the form  $N_1 N_2$ , if  $N_1 N_2$  is classified as person name, and  $N_2$  is an unknown surname, then  $N_2$  is added to surname lexicon (e.g. *Lang* from the form *Jack Lang*). This heuristic performs coreferences resolution for person names.
- Generally, when an acronym is introduced in a text, its extended form is identified during the lexical preprocessing (see section 3.1.). When the first item of the extended form is an organization trigger word then both the acronyms and its extended form are added to the organization lexicon: *FFR* and *Fédération française de rugby* (lit. French Federation of Rugby). This heuristic performs coreferences resolution for the organization names.

Contrary to those of Poibeau (1999), these lexicons are obtained through an automatic process. There are 8 such heuristics for the patronyms, 1 for the ethnonyms, 1 for the organization names, and 3 for the toponyms.

The second pass is identical to the first pass of the grammar rule processing using the new items added to the dictionaries and new rules deduced by the new trigger words obtained through the preceding step.

The setting up of this process brings an amelioration of the performance of Nemesis: +0.6% of precision and +5.3% of recall.

### 4. Experiments

In this section, we evaluate how much each component of the Nemesis system contributes to the proper name recognition. This experimentation has been done on a study corpus consisting of 6,987 words (including 246 anthroponyms) from *Le Monde* and restricted to the anthroponym class.

For this evaluation, we distinguish three different settings:

**Setting 1** only the tokenized corpus and the first pass of grammar rules are used;

Setting 2 the acronym processing is added to 1;

Setting 3 the learning process and the second pass of rewriting rules are added to 2.

Table 4 shows the contribution of each setting of the proper name recognizer. The second column indicates the recall and precision of the proper name recognition. The last three columns indicate the number of proper names correctly identified and classified, the number of proper names correctly identified but wrongly classified, and the number of proper names not or wrongly identified (not properly de-limited).

With the setting 1, Nemesis reach interesting precision and recall (up to 70% and 90% respectively). Most of the anthroponyms wrongly recognized are not identified at all, explaining the more important loss for the recall than for the precision.

The contribution of the acronym processing (setting 2) is noticeable: +9% for the recall and +3.4% for the precision. Moreover, this contribution is larger. Indeed, during the second pass (setting 3), the coreferences to the acronyms and their extended form will be recognized thanks to the both modules (acronym processing and second pass of rewriting rules). Unlike the recall, the precision is not considerably increased, because of its already high rate: +5.3% and +0.6% respectively.

After evaluating how much each component of the Nemesis system contributes to the proper name recognition, it seemed important to us to evaluate the whole system on a blind test corpus (6,381 words (including 282 anthroponyms) from *Le Monde*). Indeed, the fact of setting up our lexicons and our rewriting rules thanks to the first corpus distorts the results and tend to increased the performance. However, the performance remains considerably high on the test corpus (see table 4). It is due to quality of the texts,

			PN's correctly identified and classified	PN's correctly identified but wrongly classified	PN not or wrongly identified
Sotting 1	Precision	90.6%	174	1	71
Setting 1	Recall	70.7 %	174	4	/1
Sotting 1	Precision	94 %	196	1	50
Setting 2	Recall	79.7 %	180	4	59
Setting 3	Precision	94.6 %	209	1	26
	Recall	85 %		4	
Test	Precision	95.3 %	263	2	10
corpus	Recall	93.3 %	205	2	19

Table 4: Contribution of each component of the Nemesis PN recognizer

which are, in a big part, about media, radio, and television, domains in which we have effective lexicons.

Currently, the system performance, evaluated on the categories composing the anthroponym and the toponym classes, achieves 95 % of precision and 90 % of recall.

# 5. Extension to other categories thanks to referential composition

Anthroponyms and toponyms correctly account for about 84% of all PN's present in our corpus and some PN's belonging to other referential categories are built from these two classes. Thus, the identification of the toponyms and the anthroponyms can allow us to recognize other referential categories of Grass' classification such as some subclasses of ergonyms and praxonyms. We first concentrate on classes which are composed of toponyms such as a town or a country. These complex PN's refer either to a political event (Congrès de Rennes (lit. Conference of Rennes)), a sports event (Internationaux des États-Unis (lit. US Open)), an educational and research institution (Université de Nantes), etc. This referential composition is expressed in the rules thanks to trigger words and referential categories and takes place in the second pass (see section 3.3.):

[\$Event\_key \$Article+ \$Toponym]  $\rightarrow$  Event (e.g. *Rallye de Monte-Carlo* (lit. Monte Carlo Rally), *Internationaux des États-Unis*).

[\$Research\_inst.\_key \$Article+ \$Toponym] → Research Institution

(e.g. Université de Nantes)

The evaluation of the contribution and the quality of these rules was carried out on the fi rst corpus (6,987 words). 1 of the 4 educational and research institutions, and 7 of the 13 sports events are built from the toponym class. On one side, our referential rules allow the recognition of all these composed PN's except the sports event *championnat d'Europe des nations* (lit. European Nations Cup), where only *championnat d'Europe* is identified. On the other side, no PN is wrongly recognize by these rules.

Although reduced, the contribution of such rules is interesting, because they seem not to produce any noise. Moreover, the improvement of the performance of the system in terms of referential coverage is to be evaluated on more voluminous categories such as trademarks or products and intellectual works. In this perspective, we noted some patterns to design such rules:

- products are often present with the name of its trademark (e.g. *Volkswagen Passat*, *Ximian Evolution*, *AMD Athlon*);
- intellectual works are often composed of a trigger word and the patronym of the person who create this work (e.g. *théorème de Pythagore* (lit. Pythagoras' Theorem), *loi Evin* (lit. Evin's Law)).

# 6. Conclusion and Future Work

This paper described the development of Nemesis, a French proper name recognizer which structure is based on a corpus investigation. This system includes lexical and grammatical processing and a learning processing to update the specialized lexicons. With the current version of the system, we reach 90 % of recall and 95 % of precision for the anthroponym and toponym classes. This paper also describe an investigation into the extension of the recognition to other categories thanks to referential composition.

The first improvement will be to extend the set up of referential rules to other categories. A second improvement will be to automatically build new rules from the first pass of the rewriting rule processor. Finally, it seems necessary to conceive an algorithm allowing the automatic insertion of rules. Currently, the rules are manually ordered in order to avoid conflicts and an investigation of the conflicts appearing during rewriting rule processor is done. The automatic insertion of a rule to a suitable rank could be driven by a confi dence score based on the size of the context appearing in the rule and the possible conflict occurring between rules.

# 7. References

- Alessandro Cucchiarelli, Danilo Luzi, and Velardi Paola. 1998. Using corpus evidence for automatic gazetteer extension. In *Proceedings of LREC*'98, pages 83–89.
- B'eatrice Daille and Emmanuel Morin. 2000. Reconnaissance automatique des noms propres de la langue 'ecrite : Les r'ecentes r'ealisations. *Traitement automatique des langues*, 41(3):601–621.

- Thierry Grass. 2000. Typologie et traductabilit é des noms propres de l'allemand vers le français. *Traitement automatique des langues*, 41(3):643–670.
- Kerstin Jonasson. 1994. Le Nom Propre. Constructions et interprétations. Champs linguistiques. Duculot.
- David D. McDonald. 1994. Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus Processing for Lexical Acquisition*, chapter 2.
- Emmanuel Morin. 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse en informatique, Institut de Recherche en Informatique de Nantes.
- MUC-7. 1998. Proceedings of the 7th seven Message Understanding Conference (MUC-7).
- Thierry Poibeau. 1999. Rep´erage des entit´es nomm´ees : un enjeu pour les systèmes de veille. In Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA'99), volume 19, pages 43–51.
- Gerard Salton and Michael McGill. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.
- Nina Wacholder, Yael Ravin, and Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of ANLP'97*, pages 202–208.
- Takahiro Wakao, Robert Gaizauskas, and Yorick Wilks. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of COL-ING'96*, volume 1, pages 418–423.
- Francis Wolinski, Frantz Vichot, and Bruno Dillet. 1995. Automatic processing of proper names in texts. In *Proceedings of EACL'95*, pages 23–30.