On the Analysis of Locative Phrases with Graphs and Lexicon-Grammar: The Classifier/Proper Noun Pairing

Matthieu Constant

University of Marne-la-Vallée, 5 bld Descartes, Champs-sur-Marne,
77 454 Marne-la-Vallée Cedex 2, France
mconstant@iv-mlv.fr

Abstract. This paper analyses French locative prepositional phrases containing a location proper name *Npr* (e.g. *Méditerranée*) and its associated classifier *Nc* (e.g. *mer*). The (*Nc*, *Npr*) pairs are formally described with the aid of elementary sentences. We study their syntactic properties within adverbial support verb constructions and encode them in a Lexicon-Grammar Matrix. From this matrix, we build grammars in the form of graphs and evaluate their application to a journalistic corpus.

1 Introduction

The analysis of adverbials is an important issue in Natural Language Processing (NLP) due to the fact that adverbials can be inserted anywhere in sentences and thus make sentence analysis more difficult. The adverbials we are concerned with are of the form *Prep NP* (cf. Table 1), and more precisely with French locative *PP*s containing a location proper name *Npr* (e.g. *Paris*) and its associated classifier *Nc* (e.g. *ville*), where intuitively *Npr* names a place which belongs to a category *Nc*. Our study is domain-independent but is limited to *PP*s of the internal form:

Loc (Det1 + E) Nc [de (Det + E) + E] Npr (= Loc X')
= : dans la (ville de Paris+ mer Méditerranée)

They can be part of constructions of the form *N0 Vsup Loc N1*, e.g.

L'annonce de sa promotion a eu lieu dans le village de Nay

Where *Vsup =: avoir lieu* (*to take place*), *Npr =: Nay* and *Nc =: village*.

Our objective is to complete the lists of frozen and semi-frozen adverbials accumulated in the framework of the lexicon-grammar theory [7,9]. Those adverbials are encoded in electronic compound dictionaries (used in the linguistic platform INTEX [16,17]) and in Lexicon-Grammar Matrices (LGMs). Some complex adverbials (e.g. time adverbials [1,11]) have been described with Finite State Transducers (FSTs): they can be seen as an extension of electronic dictionaries [10]. These linguistic re-

E.M. Ranchhod and N.J. Mamede (Eds.): PorTAL 2002, LNAI 2389, pp. 33-42, 2002.

sources should be of great use in concrete applications such as information retrieval [4], phrase translation [5] and phrase segmentation for summarization [2,18].

In this paper, we not only automatically relate an *Npr* to its appropriate classifier (or semantic class) as many studies have already made [3,14,19], but we also provide, for each classifier *Nc* of a given list, a limited distribution of locative prepositions *Loc* that enter in the adverbial elementary sentence *N0 Vsup Loc X'*. This provides a solid basis for disambiguating *PP*s, e.g. distinguishing between arguments of predicates and adverbials. In the example below, as *à la ville de Paris* cannot be an adverbial (cf. section 3), it is necessarily an argument of the predicate *subvention* that enters in the elementary structure *N0 donner Det subvention à N1*.

Le gouvernement a donné une subvention à la dernière minute à la ville de Paris
(At the last minute, the government gave a subvention to the city of Paris)

First, we shall briefly study the construction *N0 Vsup Loc X'* from a linguistic point of view. Then, on the basis of this study, we shall describe how to build locative grammars with LGMs and graphs. Finally, we apply these grammars to journalistic corpora and evaluate the results.

Table 1. Notations

S	Sentence
NP	Noun Phrase
PP	Prepositional Phrase
Vsup	Support verb
Ni	*i*th nominal argument of a predicate (where *i* is an integer)
Prep	Preposition
Loc	Locative Preposition
Det, *Det1*	Determiners
E	Empty element
UN	The singular indefinite determiners *un* or *une* (a or an)
LE	*le*, *la* (the)
POSS	Possessive determiner

2 A Brief Linguistic Study

This linguistic study is limited to a few properties to clarify the paper. *Nc* and *Npr* can be related to each other in the sentence:

(1) (E + Det) Npr être (UN + des) Nc
:= *Paris est une ville* (Paris is a city)
La Méditerranée est une mer

Det is a definite determiner whose form only depends on *Npr* [12]. It is strictly limited to the set {*le*, *la*, *les*}. The sentence (1) can be reduced into three nominal constructions:

Det1 Nc Npr := *la mer Méditerranée* (the Mediterranean sea)
Det1 Nc de Npr := *la ville de Paris* (the city of Paris)
Det1 Nc de Det Npr := *les îles des Canaries* (the Canaries Islands)

2.1 Syntactic Properties of (*Nc*, *Npr*) Pairs in Noun Phrases

Det1 is LE or POSS. Note that '*un'* is strictly forbidden when the NP does not contain a modifier located at the end of it:

Une ville de Paris, embellie pour les fêtes

Det1 can also be in the plural for some (*Nc*, *Npr*) like (*île*,*Canaries*) because *Canaries* are islands [6]. *Det1* is also in the plural when several *Npr* with the same *Nc* are coordinated such as:

Les villes de Paris et (de+E) Lyon = la ville de Paris et la ville de Lyon

A given pair (*Nc*, *Npr*) does not always enter into all the constructions. Acceptability essentially depends on *Npr*:

L'île (de+E) Malte
*L'île (*E+de la+*de) Martinique*
*Les îles (E+de les+*de) Canaries*

Modifiers can also be inserted between the constituents. The markers *M1*, *M2* and *M3* show where modifiers can be placed into the *NP*s:

Det M1 Nc M2 de (E + Det) Npr M3
Det M1 Nc Npr M3

Moreover, some (*Nc*, *Npr*) only occur in the nominal forms; that is, they do not occur in the elementary sentence (1). They can be seen as compounds. Modifiers marked *M2* are forbidden in this case:

**Le Nord est une mer*
*La mer (E+ ?*déchainée) du Nord me fascine* (where *du* = *de le*)

2.2 Syntactic Properties of (*Nc*, *Npr*) Pairs in the Construction *N0 Vsup Loc N1*

Now, we examine the form *N0 Vsup Loc X'*. We study the distribution of its constituents. Let *Vsup* be *être* (to be). Let *Loc* be *dans*, *à*, *en* and *E*, very frequent locative prepositions in French. Our study is then reduced to:

(2) N0 être (dans + à + en + E) (Det1+E) Nc de (Det+E) Npr
(3) N0 être (dans + à + en + E) (Det1+E) Nc Npr

The sequence *Loc (Det1+E)* is limited to the set {*dans Det1*, *à det1*, *en*, *E*}, but each *Nc* has its own properties. Precise information on pairs cannot be encoded without at least an exhaustive study of all classifiers *Nc*, as illustrated in the following examples:

*Max est (*E + *en + dans la + *à la) ville de Paris*
*Max est (*E + en + dans la + *à la) mer (Méditerranée + du Nord)*
*Max est (E + *en + dans la + ?à la) rue (de la Paix + Censier)*
*Max est (E + *en +*dans le + à le) métro République*

We notice that preposition distribution is different for constructions (2) and (3):

*Max est (*E + en + dans la + *à la) région Ile-de-France*
*Max est (*E + *en + dans la + *à la) région de l'Ile-de-France*

The constructions with the prepositions *E* and *en* do not accept modifiers M1 (2.1):

**Belle rue de Rivoli, les passants se marchent les uns sur les autres.*
**En magnifique mer Méditerranée, la marée est peu marquée*

We notice that the adjective *plein* can be inserted in these cases; but it is part of the compound preposition *en plein*:

Max est égaré en pleine mer Méditerranée

3 Construction of Locative Grammars

3.1 NNPR Matrix Description

It has been shown in Section 2 that syntactic properties of *NP*s essentially depend on *Npr* (2.1) and on *Nc* for the prepositional distribution (2.2). A systematic study of *Npr* is necessary to encode the *NP*s precisely, which is time consuming due to number of *Npr*. In this section, our objective is to encode the syntactic properties of the form *N0 Vsup Loc X'*, which essentially depend on the classifiers (but not only [6]). To this end, we established a list of about one hundred *N*s. We encoded the properties of each element of the list into a Lexicon-Grammar Matrix (LGM) as in [8]. Each column corresponds to one property; that is, for example, to one *PP* form. Each row contains one lexical entry (one *Nc* in our case). At the intersection of a row and a column, a plus sign indicates that the corresponding *Nc* enters into the corresponding property, a minus sign, that it does not; and finally, a string indicates lexical information. Lexicon entries are displayed in rows and syntactic properties in columns. We show a sample of the LGM in Table 2.

Table 2. Lexicon-Grammar Matrix *PNNpr* where P1 =: *en Nc Npr*; P2= : *en Nc de (E+Det) Npr* ; P3 =: *dans LE Nc Npr* ; P4 =: *dans LE Nc de (E+Det) Npr* ; P5 =: *à LE Nc Npr ;* P6 =: *à LE Nc de (E+Det) Npr* ; P7 =: *Nc Npr* ; P8 =: *Nc de (E+Det) Npr*

(N0+Que S) avoir lieu Loc N1									
N	N plural	Loc N1							
		P1	P2	P3	P4	P5	P6	P7	P8
fleuve	-	-	-	+	+	-	-	-	-
Gare	-	+	+	+	+	+	+	+	+
Gave	-	-	-	-	+	-	-	-	-
ghetto	-	-	-	-	+	-	-	-	-
glacier	-	-	-	+	+	+	+	-	-
Golfe	-	-	-	+	+	-	-	-	-
Île	+	-	-	+	+	+	+	-	-

3.2 Reference Grammars

LGM is a simple and clear representation. Nevertheless, it cannot be directly applied to texts. By transforming the LGM into Finite State Graphs (FSG), the linguistic information encoded in the LGM can be applied. In this sub-section, we describe the process of transforming an LGM into an FSG. A simple way of doing this is to build a reference graph [13,15]. From this graph that represents the set of all possible forms, a graph will be created for each lexical entry. A transition *@i* (where *i* is an integer) is seen as a variable that refers to the *i*th column (or property) of the matrix. For each lexical entry (or each row), a new graph is automatically constructed from the reference graph by:

- removing the transition *@i* when the intersection of the *i*th column and the current row is '-'
- replacing *@i* by <*E*> (the empty element) when '+'
- replacing @i by the content of the intersection of the *i*th column and the current row, by default

The reference graph shown in Graph 1 (Fig. 1) represents a grammar that describes the forms *Loc N1* (presented above). Grey boxes represents sub-graphs. The graph **LE** describes the set {*le*, *la*, *l'*}; **Det**, the set {*le*, *la*, *les*, *l'*}; **POSS**, the possessive determiners like *mon*, *ma*, *ton*, *ta*, etc.; **Npr** describes the proper noun by the means of the tag <PRE> that stands for a word beginning with an uppercase letter. It recognises simple forms like *Paris* and complex ones like *Pyrénées-Atlantiques* and *La Havane*. <@1.N:p> stands for the plural form of the classifier symbolised by @1 (first column of the matrix): *Nc* for noun and *p* for plural. From Graph 1 (Fig. 1), we obtain Graph 2 (Fig. 2) for the lexical entry *mer*. All forbidden paths are automatically removed according to the LGM codification.

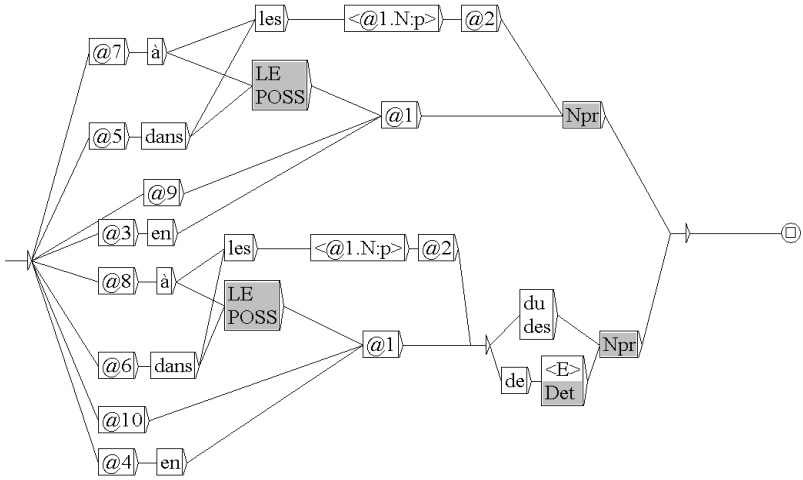

Fig. 1. Reference graph

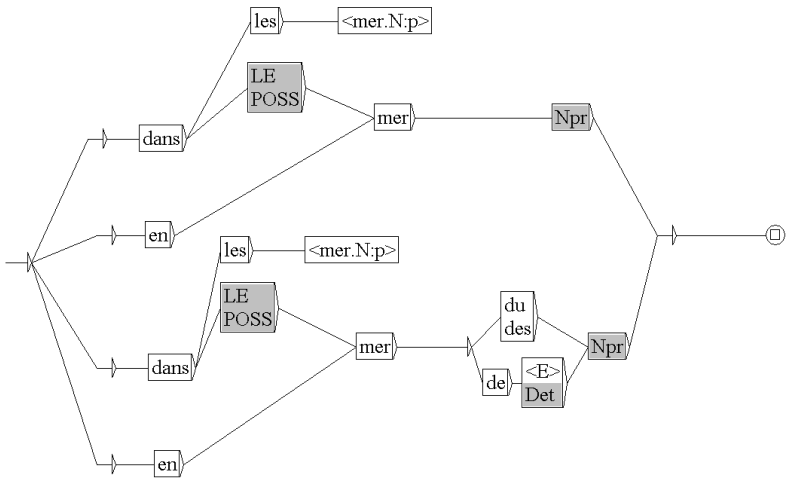

Fig. 2. Grammar for the lexical entry *mer*

4 Application of Locative Grammars: Some Experiments

4.1 Description of the Experiment

We apply locative grammars to a corpus. Our corpus consists of one year of the electronic version of the French newspaper *Le Monde* (year 1993). It comprises about ten million words. It is of great interest because of the quality of the writing and the large variety of themes that are treated. We undertook three experiments. The first one con-

sists in applying a basic grammar that describes the forms *Loc X'* without modifiers and with a small set of simple prepositions (Section 3). In the second experiment, we enlarge the set of prepositions with around three hundred compound locative prepositions. We manually included those prepositions into a graph (called **Loc**), after extracting them from the lists provided in [9] (e.g. *aux alentours de, en amont de*). We assume that these new prepositions have the same behaviour as the preposition *dans*. That means that if a construction allows the preposition *dans*, all the compounds are allowed. Finally, we add modifiers in our grammar by inserting a graph **Modif** at positions *M1* and *M2* as described in Section 2. **Modif** represents a general modifier of the form (<ADV>+<E>) <A>, where <ADV> stands for an adverb and <A> for an adjective.

4.2 Results of Experiments

The number of occurrences automatically found in our corpus for experiments 1, 2 and 3 are respectively *1,977*, *2,026* and *2,190*. We first observe that the *PP*s described are very rare: maximum 2,190 utterances in a 10-million word text. The classifier nouns occur 35,000 times on their own; that means that 6% of them are part of *PP*s recognized. Furthermore, we observe that 90% of the forms are base forms; that is, without modifiers and with a simple preposition (*dans*, *à*, *en* or *E*). Thus, the description of the modifiers and the compound locative prepositions is useful for only 10% of the forms. We provide below a selection of the concordances obtained:

.{S} Un peu plus tard, arrêt à la gare Saint-Lazare, que la patrouille parcourt en t
aires se compliquent aussi dans le déjà difficile département des Alpes-Maritimes.
ue Saddam, situé à une quarantaine de kilomètres à le nord de la ville de Mossoul
enseigne des Bonnes Choses, rue Falguière, où le cassoulet mijotait de éternité.{S}
ire et bancaire à le Trésor, rue de Bercy à Paris, qui vont avoir à émettre environ
ordement de marchandises dans le port de Rotterdam, le plus important du monde,
détruite par la guerre.{S} " Dans le village de Jocoaitique, qui se relève peu à peu
Yaoundé, ont été reconduits dans la province anglophone du Nord-Ouest.{S}
oirs ont été tués par balles dans le ghetto noir de Tokoza, à l'est de Johannesbourg
a foule de un policier noir, dans la cité de Evaton, la veille du Jour de l'an.{S} L
00 soldats français déployés dans la région de Hoddour.{S} M. Mellick a fait cette
piégée, mercredi 10 février, en plein centre du port pétrolier de Barrancabermeja,
[A] *son " bouchon " de la rue Vavin : " Là cuit un cassoulet de Castelnaudary.{S}*
[B] *eu plus tard dans une rue de Naplouse, en Cisjordanie occupée. {S} Ancien co*
[C] *usée de la SEITA, 12, rue Surcouf, 75007 Paris ;{S} tél.{S} : 45-66-60-17.{S}*
[D] *mois après la mise en route de Gabcikovo, la catastrophe annoncée ne a pas e*
[E] *-vous, c' était le 3 avril, dans le quartier sud de Manchester, à St Peter's Squar*
[G] *succède à un socialiste dans le plus petit canton de l'Hérault : Maurice Requi*
[H] *grand compagnon de route de Dave Holland et de Portal, présente un nouveau*

4.3 Problems Encountered

In this section, we briefly list some issues and ways to resolve them. The main problem encountered is noise: 434 utterances out of 2190 should not be recognized (i.e. around 20% of noise), which is not negligible. There are several reasons for this. The first is the use of the empty preposition and the ambiguity of the empty element. Many *Loc X'* where *Nc* accepts this preposition belong to a larger *NP*, as in [A] (in the concordances of 4.2). The application of a grammar representing an *NP* described in 2.1 with the longest match rule removes a large part of the noise. Therefore, [A] is now recognised as a NP (*la rue Vavin*). Using this method, we remove 382 utterances. Thus, we significantly reduce noise to 50 utterances i.e. about 2.3%. Errors like [B] should be removable through the application of a grammar of a general *NP*. Although [C] is well parsed, it could be part of a longer unit, a mail address (*12, rue Surcouf, 75007 Paris*). Note that this case appears 712 times in our corpus (28% of the total occurrences). Another problem comes from the use of modifiers in our grammars: *sud* in [E] (*Manchester* is not a *quartier*). The presence of superlatives creates noise such as in [G]: a way of removing it is to construct special superlative grammars. The generality of *Npr* is another issue. For example, a sequence like *la ville de Pagnol* (that is translated by *Pagnol's city*) is recognised: *Pagnol* is the name of a personality and not the name of a city. This problem can be solved by the use of dictionaries of proper nouns. In [G], the classifier *route* is part of the French compound *compagnon de route* (fellow traveller). The application of compound dictionaries should remove the wrong parsing.

5 Conclusion and Perspectives

This paper briefly studies the behaviour of the pair (*Nc*, *Npr*) within *NP*s and *PP*s. From the theory, we encode general syntactic rules into an LGM and then in the form of graphs. We then apply some of these grammars to a journalistic corpus. This method appears to be of great interest for Natural Language Processing even if, at first, it creates considerable noise (4.3). Furthermore, the linguistic study produces precise syntactic information on the sequences of the form *Loc X'*. It is also a source of semantic information, by associating an *Nc* to an unknown *Npr*. Moreover, we project to build dictionaries of locative *Npr* containing syntactic information or to complete existing ones.

6 References

1. Baptista, J.: Manhã, Tarde, Noite: analysis of temporal adverbs using local grammars. Seminarios de Linguistica 3, Faro, Universidade do Algarve (1999) 5–31

2. Boguraev, B., Neff, M.: Discourse Segmentation in aid of documentation summarization. In: Proceedings of Hawaii International Conference on System Sciences (HICSS-33), Minitrack on Digital Documents Understanding. Maui, Hawaii (2000)

3. Cucchiarelli, A., Luzi, D., Velardi, P.: Semantic tagging of unknown proper nouns. Natural Language Engineering, Vol. 5:2. Cambridge University Press (1999) 171–185

4. Domingues, C.: Etude d'outils informatiques et linguistiques pour l'aide à la recherche d'information dans un corpus documentaire. Thèse de doctorat en informatique, Université de Marne-la-Vallée (2001)

5. Fairon, C., Senellart, J.: Classes d'expressions bilingues gérées par des transducteurs finis, dates et titres de personnalité (anglais-français). Linguistique contrastive et traduction, Approches empiriques. Louvain-la-Neuve (1999)

6. Garrigues, M.: Prepositions and the names of countries and islands: a local grammar for the automatic analysis of texts. Language Research, Vol. 31:2. Seoul, Language Research Institute, Seoul National University (1995) 309–334

7. Gross, M.: Grammaire transformationnelle du français. Vol. 3, Syntaxe de l'adverbe. Paris, ASSTRIL, Université Paris 7 (1986)

8. Gross, M., Constructing Lexicon-Grammars. In: Atkins, B.T.S., Zampolli, A. (eds.): Computational Approaches to the Lexicon. Oxford University Press, Oxford (1994) 213–263

9. Gross, M.: Les formes *être Prép X* du français. Lingvisticae Investigationes, Vol. XX:2. John Benjamins, Amsterdam Philadelphia (1996) 217–270

10. Gross, M.: The Construction of Local Grammars. In: E. Roche and Y. Schabes (Eds.): Finite State Language Processing. The MIT Press, Cambridge, Mass. (1997) 329–352

11. Maurel, D.: Adverbes de date: étude préliminaire à leur traitement automatique. Lingvisticae Investigationes, Vol. XIV:1. John Benjamins, Amsterdam Philadelphia (1990) 31–63

12. Maurel D., Piton, O.: Un dictionnaire de noms propres pour *INTEX* : les noms propres géographiques. In : Fairon, C. (ed.): Analyse lexicale et syntaxique : le système INTEX. Lingvisticae Investigationes, Vol. XXII. John Benjamins, Amsterdam Philadelphia (1998) 279–289.

13. Roche, E.: Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire. Thèse de doctorat, Paris, Université Paris 7 (1993)

14. Senellart, J.: Locating noun phrases with finite state transducers. In: Proceedings of the 17th International Conference on Computational Linguistics (COLING98). Montreal (1998) 1212–1219

15. Senellart, J.: Reconnaissance automatique des entrées du lexique-grammaire des expressions figées. Travaux de linguistique. Bruxelles (1999)

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

17. Silberztein, M.: INTEX: a corpus processing system. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING94). Kyoto, Japan (1994) 579–582

18. Silberztein, M.: INTEX at IBM. In: Dister, A. (Ed.): Actes des Troisièmes Journées INTEX. Revue Informatique et Statistique dans les Sciences humaines. Liège (2000) 319–332

19. Wakao, T., Gaizauskas, R., Wilks, Y.: Evaluation of an algorithm for the recognition and classification of proper names. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING96), Copenhagen (1996) 418–423