Analyse et représentation d'adverbes locatifs

4.1 Introduction

L'analyse automatique des groupes prépositionnels (*Prep GN*) dans les textes est un problème bien connu et difficile du domaine du TAL. L'une des principales difficultés consiste à distinguer les groupes prépositionnels arguments des compléments circonstanciels (ou adverbes) :

Le gouvernement donne une subvention à Rome (à Rome = complément essentiel)
Luc se repose (à Rome = complément circonstanciel)

Les principales méthodes de résolution sont statistiques (D. Hindle et M. Rooth, 1994 ; E. Brill et P. Resnik, 1994 ; M. Collins et J. Brooks, 1995 ; J. Zavrel et al, 1997 ; etc.). Il existe des méthodes utilisant des indices linguistiques (C. Fabre et al, 2002). Une autre difficulté consiste à repérer la classe sémantique à laquelle ils appartiennent (temps, lieu, manière, etc.). Les chercheurs lexique-grammairiens ont montré que certaines classes de compléments circonstanciels sont facilement représentables à l'aide d'automates finis lexicalisés comme les dates et les durées (D. Maurel,1990 ; M. Gross, 2002 pour le français ; J. Baptista, 2002, 2003 pour le portugais). Dans ce chapitre, nous décrivons un type particulier de compléments prépositionnels : les compléments locatifs en français. Il existe déjà de nombreux travaux linguistiques généraux sur les constructions locatives (C. Vandeloise, 1986 ; A. Borillo, 1998). Dans le cadre du lexique-grammaire, nous citons A. Guillet et C. Leclère (1992) et J.P. Boons (1985). La plupart des travaux de TAL réalisés sur le sujet cherchent avant tout à décrire des contraintes sémantiques dans ces compléments. Par exemple, certains cherchent à construire des modèles géométriques dans l'espace utilisant notamment des déplacements élémentaires pour décrire les mouvements (Y. Mathet, 2002). Ce sujet est sensible dans le domaine du TAL et quelques projets ont été mis en place pour traiter ces objets linguistiques, en particulier le projet GeoSem (laboratoires GREYC, ESO, ERSS et MEDIA/EPFL) dont une des composantes consiste à repérer des séquences locatives géographiques et à leur assigner un marquage sémantique fin.

La forme générale d'un complément locatif peut être décrite comme la forme nominale prépositionnelle *Loc GN* où *Loc* correspond à une préposition locative et *GN* à un groupe nominal. Notre objectif est de trouver un certain nombre de contraintes locales dépendant du lexique permettant d'améliorer la description de tels compléments locatifs. Nous considérons que nos compléments rentrent dans la construction à verbe support *N*0 *Vsup Loc Det N Modif*. Des études sur les structures *être Prep X* (L. Danlos, 1980 ; M. Gross, 1996) ont montré les fortes contraintes qui existent entre les différents constituants. Nous décidons d'examiner un ensemble limité de noms *N* et particulièrement de séquences nominales formées de noms propres de lieux géographiques et/ou de leurs classifieurs locatifs associés (*Paris* a pour classifieur locatif *ville*). Nous montrerons qu'il existe un certain nombre de contraintes qui peuvent être décrites dans des graphes ou des tables syntaxiques (ensuite transformées en graphes).
Nous ferons d'abord quelques rappels sur les adverbes et les groupes prépositionnels locatifs afin de rendre notre argumentation plus claire. Nous étudierons également les prépositions locatives simples et composées dont nous construirons des grammaires locales. L'application de ces grammaires pointant clairement l'ambiguïté naturelle générée par la reconnaissance locale de telles structures, nous nous consacrons à la description de contraintes locales entre les constituants d'un groupe prépositionnel ayant pour nom tête un nom propre (simple ou composé) de lieu géographique. Dans un premier temps, nous regardons le comportement du couple (*Npr*, *Nc*) dans un groupe nominal où *Npr* est un nom d'un lieu géographique (ex : *Pas-de-Calais*) et *Nc* est son classifieur locatif associé (ex : *région*). Ce couple forme un nom propre composé *Nprc* :

pic du Midi (*Npr* =: Midi ; *Nc* =: *pic*)
mer Méditerranée (*Npr* =: *Méditerranée* ; Nc =: *mer*)
région Pas-de-Calais (*Npr* =: *Pas-de-Calais* ; *Nc* =: *région*)
île de Malte (*Npr* =: *Malte* ; *Nc* =: *île*)

Le degré de figement est variable selon ses élements lexicaux. Par exemple, la séquence *mer de Glace* est plus figée que *mer de Norvège* car la première structure désigne un glacier et pas une mer.

Dans le même temps, nous étudions certaines caractéristiques linguistiques des classifieurs et des noms propres utilisés : déterminants, modifieurs, etc. Enfin, nous décrivons la distribution prépositionnelle des groupes prépositionnels locatifs rentrant dans la structure *N*0 *Vsup Loc Det N Modif* lorsque *N* prend trois formes :

- *N* = : *Nprc*
- *N* = : *Npr*
- *N* = : *Nc*

Nos descriptions sous la forme de tables syntaxiques vont aboutir à un système relationnel de tables syntaxiques, non compatible avec la méthode de conversion des tables en grammaires locales d'E. Roche (1993). Pour confronter nos représentations linguistiques à des textes, nous implantons de nouveaux formalismes et algorithmes de conversion.

4.2 Préliminaires linguistiques

4.2.1 Adverbes généralisés

4.2.1.1 Notions d'adverbes généralisés et d'objets

Un adverbe dans la littérature traditionnelle est un mot invariable élémentaire (ex. *hier*) ou dérivé (ex. *doucement*). Dans le cadre du lexique-grammaire, C. Molinier (1990) a réalisé une classification des adverbes en *–ment* au moyen de critères formels. M. Gross (1986) va plus loin et définit la notion d'adverbe généralisé qui est :

- soit un adverbe traditionnel : *ici, couramment*
- soit un groupe nominal prépositionnel où la préposition est parfois zéro : *dans trois jours, sur l'étagère, en train*
- soit une subordonnée composée d'une conjonction de subordination suivie d'une phrase: *dès que la pluie cessera*

Ainsi, il regroupe sous un même terme trois catégories formelles bien distinctes dans la grammaire traditionnelle. L'un de ses arguments est que ces trois formes répondent en général aux questions en *où, quand, comment, pourquoi,* etc. souvent associées aux compléments circonstanciels. Il définit même la structure globale des adverbes par la formule classique d'un groupe nominal prépositionnel :

Prép Dét N Modif

où chaque élément peut être absent ou contracté. Il rappelle qu'un modifieur peut prendre la forme complétive *qu P* et que la conjonction de subordination est souvent de la forme *Conjs = : Prép (E + ce) que*. Désormais, nous adoptons ce principe et abrégeons le terme adverbe généralisé en adverbe.

Un complément essentiel, de même structure globale qu'un adverbe, est un argument essentiel d'un prédicat. Par exemple, le verbe *donner* possède deux compléments essentiels (un objet direct avec la préposition zéro et un datif avec la préposition *à*) :

Max donne sa clé à la gardienne

4.2.1.2 Distinction entre adverbes et objets

La distinction entre adverbe et complément essentiel (ou objet) est un problème important parce que ces deux éléments ont la même structure de surface et que la détermination des arguments des prédicats (compléments essentiels) est une étape fondamentale de l'analyse syntaxique. Il est généralement admis que, dans un complément essentiel, le choix de la préposition dépend en large partie du prédicat et que, dans un adverbe, la préposition dépend avant tout du groupe nominal. Les adverbes et les objets des verbes sont généralement distingués à l'aide de quelques critères traditionnels. D'abord, les compléments essentiels répondent aux questions en *Prep (que + qui + quoi)*, ce qui n'est pas le cas des compléments circonstanciels qui répondent aux questions en *quand, où, comment,* etc. Les adverbes sont mobiles dans la phrase, ce qui est moins vrai avec les objets. M. Gross (1986) montre à l'aide de quelques exemples que ces critères ne sont pas toujours valables. Ils ne sont ni nécessaires ni suffisants. M. Gross est sceptique quant à leur utilisation car cela « fait perdre toute cohérence au domaine complexe des adverbes » (M. Gross, 1986, p. 22). Il faut traiter les adverbes au cas par cas.

4.2.1.3 La portée des adverbes

Il est très souvent possible d'expliciter la relation entre un adverbe *Adv* et une phrase *P* dans laquelle il peut être inséré. Elle s'exprime à l'aide de phrases simples qui mettent en évidence la portée de l'adverbe. Ces constructions utilisent des verbes-supports *Vsup* comme *être*, *avoir lieu*, *se passer*, etc. Il existe deux cas bien distincts :

- Le cas où l'adverbe porte sur la phrase[55]. Dans cette situation, la relation entre *Adv* et *P* peut s'expliciter à l'aide de la construction

  (E + Le fait) que P Vsup Adv

Par exemple, la phrase :

Max boit du champagne régulièrement

peut être interprétée comme :

Max boit du champagne # Que Max boive du champagne se passe régulièrement

Cette construction est souvent considérée comme théorique car stylistiquement difficile. Il est souvent plus naturel d'utiliser le pronom *cela* portant sur *P* :

P # cela Vsup Adv
= : Max boit du champagne # cela se passe régulièrement

La nominalisation de *P* est parfois plus naturelle comme dans

Max arrive lundi prochain
= Max arrive # l'arrivée de Max a lieu lundi prochain

La relation entre *Adv* et *P* n'est pas toujours facile à mettre en évidence, et même pas toujours possible. L'exemple suivant est tiré de M. Gross (1986) :

A l'étonnement de Paul, Max a réussi son examen

La relation entre l'adverbe *à l'étonnement de Paul* et la phrase simple *Max a réussi son examen* s'explicite à l'aide de la phrase ci-dessous où l'adverbe est présent sous la forme d'une phrase à verbe :

Que Max a réussi son examen étonne Paul

Car la structure combinant *Adv* et *P* n'est pas très naturelle :

? Que Max a réussi son examen se passe à l'étonnement de Paul*

- Le cas où l'adverbe porte sur un argument de la phrase

[55] A. Guillet et C. Leclère (1992) préfèrent utiliser le terme « sous-phrase ».

Parfois, un adverbe ne porte pas sur la phrase mais simplement sur un argument N_i de cette phrase. Dans ce cas-là, la relation s'exprime par la construction suivante plus naturelle que la précédente:

N_i Vsup Adv

Par exemple, dans la phrase ci-dessous :

Max a demandé de bonne foi le résultat du match de football
= Max demande le résultat du match de football # Max est de bonne foi

l'adverbe *de bonne foi* apparaît comme un modifieur de *Max* permutable dans la phrase.

Ainsi, pour résumer, la relation entre un *Adv* et une phrase $P = N_0\ V\ Prep_1\ N_1 \ldots Prep_j\ N_j$ peut s'expliciter par la construction :

(Que P + N_i) Vsup Adv

Ces constructions ont été abondamment étudiées dans le cadre du lexique-grammaire : essentiellement lorsqu'il s'agit de formes figées ou semi-figées, non seulement en français (L. Danlos, 1980 ; M. Gross, 1996), mais aussi dans d'autres langues (portugais : E. Ranchhod, 1989).

4.2.2 Les compléments prépositionnels locatifs

4.2.2.1 Notion de complément locatif

Nous rappelons la structure des compléments locatifs :

Prep Det N Modif (*Prep* peut être zéro)

Intuitivement, pour qu'il soit considéré comme locatif, le nom-tête *N* doit dénoter un lieu. Cependant, il existe des cas où *N* est un lieu mais le complément n'est pas locatif comme dans la deuxième phrase des exemples tirés de A. Guillet et C. Leclère (1992) :

Les Gaulois ont envahi Rome (*Rome* est un lieu)
Les Gaulois ont vaincu Rome (*Rome* est pris pour les Romains et donc non locatif)

Un critère formel pour distinguer les compléments locatifs des autres compléments pourrait être le choix de la préposition.En effet, certaines prépositions comme *dans, à, sur, contre, de, par* ont un caractère locatif et leur présence dans un complément pourrait justifier l'utilisation du qualificatif « locatif ». Par exemple, dans les phrases ci-dessous, les compléments commençant par les prépositions citées ci-dessus sont tous locatifs.

Le loup se cache dans la forêt
Max part à Paris
Léa pose le stylo sur la table
Les affiches sont collées sur le mur
Luc revient de la plage
La balle passe par la fenêtre

Cependant, ces prépositions sont ambiguës et ont des emplois non locatifs. Ainsi, leur présence dans un complément ne garantit pas le caractère locatif de celui-ci. Par exemple, dans les phrases suivantes, les compléments utilisés ne sont pas locatifs.

La température de l'eau est dans les vingt degrés celsius
Max m'a rendu visite à l'improviste
Max porte son choix sur Léa
Georges veut faire la guerre contre tout le monde

Par ailleurs, intuitivement, certaines prépositions telles que *avec* sont considérées comme non locatives. Cependant, il existe des cas où, malgré la présence de ces prépositions, les compléments sont locatifs comme dans :

Max range les fourchettes avec les couteaux

Traditionnellement, on associe les questions en *où* et *Prép où* à la notion de lieu. En effet, cette propriété formelle est valable dans un grand nombre de cas :

Où part Max ? à Paris
Où sont collées les affiches ? sur le mur
D'où revient Luc ? de la plage
Par où passe la balle ? par la fenêtre

Il existe cependant des contre-exemples qui rendent ce critère non valable pour tous les compléments locatifs. J. P. Boons (1985) estime que la question en *où* est un critère suffisant, mais pas nécessaire. En effet, les phrases suivantes sont interdites :

** Où revient Luc ? de la plage*

Où Hannibal marche-t-il ? (J.P. Boons, 1985)
? Sur Rome

On peut même se demander si c'est un critère suffisant. M. Gross (1975) avait déjà entamé la discussion en montrant que certaines infinitives dont il n'est pas clair qu'elles soient locatives répondaient à la question en *où* :

Où va Paul ? Acheter du pain

Certains diront que la réponse à cette question n'est pas *acheter du pain*, mais le locatif *à la boulangerie* qui est le lieu sous-entendu par le procès d'aller acheter du pain (C. Leclère, 2002) :

Où va Paul ? A la boulangerie, pour acheter du pain.

La question en *Prep où* est, quant à elle, clairement non suffisante, comme le montre la phrase :

Par où Max va-t-il commencer ? Par le montage des roues.
(où *par le montage des roues* ne dénote pas un lieu)

Pour conclure sur ce sujet, J.P. Boons (1985) estime que l'on ne peut pas couvrir exactement le champ sémantique du lieu à l'aide de critères formels. On peut seulement s'en approcher.

4.2.2.2 Compléments locatifs et verbes supports

Précédemment, nous avons tenté de distinguer adverbes et compléments essentiels. Pour les locatifs, un même complément peut être soit un adverbe soit un argument. Prenons les exemples suivants :

Max chante dans sa chambre
Max va dans sa chambre

En essayant de relier le complément au reste de la phrase, ces deux constructions réagissent différemment :

Que Max chante se passe dans sa chambre
** Que Max aille se passe dans sa chambre*

Bien que la première construction soit d'une acceptabilité difficile, nous l'acceptons, alors que la deuxième est clairement inacceptable. Nous en concluons que, dans la première phrase, *dans sa chambre* est un adverbe. Il est clair que, dans le deuxième exemple, *dans sa chambre* est un complément essentiel et même obligatoire, la phrase suivante étant interdite :

** Max va*

Soit la phrase suivante :

Marie a sauté dans un lac gelé en Suisse

Elle possède deux compléments locatifs (*dans un lac* et *en Suisse*). Nous montrons qu'ils n'ont pas le même rôle dans cette phrase : *en Suisse* joue le rôle d'un adverbe portant sur la phrase élémentaire (c'est le lieu du procès) alors que *dans un lac* est un complément essentiel du verbe *sauter*.

Que Marie (a + ait) sauté dans un lac s'est passé en Suisse
** Que Marie (a + ait) sauté (E+en Suisse) s'est passé dans un lac*

J.P. Boons (1985) et A. Guillet et C. Leclère (1992) montrent que, à l'instar des phrases contenant un adverbe, les phrases à constructions locatives (i.e. dans lesquelles il existe un complément essentiel locatif) peuvent aussi être analysées à l'aide de constructions à verbes-supports. Il est notamment possible d'employer les arguments dans des phrases qui expriment une localisation, avant et/ou après le procès. Le verbe-support neutre *être* est parfaitement adapté dans ce cas-là. Nous utilisons les notations E_i et E_f pour représenter respectivement l'état initial et l'état final d'un procès. Les exemples suivants n'ont pas besoin d'être expliqués.

Max met l'assiette dans l'armoire
E_f *: l'assiette est dans l'armoire*

Max revient de la plage
E_i *: Max est à la plage*

Max va de Paris à Marseille en voiture
E_i : Max est à Paris
E_f : Max est à Marseille

Ainsi, dans l'exemple précédent contenant deux locatifs, on a aussi l'interprétation :

Marie a sauté dans un lac gelé en Suisse
E_f : Marie est dans un lac gelé.

Les mouvements des différents arguments nominaux peuvent aussi être explicités à l'aide de verbes-supports de mouvement comme *aller*, *venir*, *passer*, etc. (il en existe quelques autres).

Max plonge le savon dans l'évier
Procès : *Le savon va dans l'évier*
E_f : le savon est dans l'évier

Marie prend un fruit (de + dans + de dedans) la corbeille
*E_i : le fruit est (*de + dans + dedans) la corbeille*[56]
Procès : *le fruit vient de la corbeille*

Luc jette une balle par la fenêtre (J.P. Boons, 1985)
E_i : la balle est de ce côté-ci de la fenêtre
Procès : *la balle passe par la fenêtre*
E_f : la balle est de ce côté-là de la fenêtre

Notons que J. P. Boons (1985) va plus loin dans l'interprétation sémantique. Il estime qu'il peut exister plusieurs états finaux dans le procès. Dans la phrase suivante,

Les Russes lancent une bombe sur Paris

Le premier état final pourrait être que la bombe atteint sa trajectoire. Il peut être explicité par la phrase :

La bombe est lancée sur Paris

Ce premier état final induit lui-même un deuxième état final :

La bombe est sur Paris

4.2.3 Les prépositions locatives

4.2.3.1 Prépositions simples

L'une des caractéristiques d'un complément locatif est la présence d'une préposition locative avant le groupe nominal[57]. Pour certains linguistes dont D. Le Pesant (2003), la préposition locative est un prédicat. Cette discussion n'est pas pertinente à notre propos.

[56] cf. C. Leclère et A. Guillet, pour plus de détails.
[57] Même si cela n'est pas toujours le cas : *Luc met les couteaux avec les fourchettes* où *avec les fourchettes* est locatif (*Où Luc met-il les couteaux ? avec les fourchettes.*) alors que la préposition *avec* ne l'est pas.

Il est facile de faire une liste exhaustive des prépositions locatives simples : *dans*, *avant*, *devant*, *sur*, *à*, *contre*, *après*, *derrière*, *sous*, etc. Cependant, comme le montre la section précédente, chacune d'elles a des emplois non locatifs. L'ambiguïté avec le temps, par exemple, est récurrente comme le montrent les ensembles de phrases suivants, avec les prépositions *avant*, *dans*, *à* et *après* :

Max s'arrête avant la forêt (lieu)
Max s'arrête avant la nuit (temps)

Dans la ville, la température a augmenté (lieu)
Dans l'après-midi, la température a augmenté (temps)

A Paris, Marie rencontrera du monde (lieu)
A huit heures, Marie rencontrera du monde (temps)

Max et Marie ont disparu après l'étang (lieu)
Max et Marie ont disparu après leur rencontre (temps)

Lorsque la préposition fait partie d'une expression figée ou semi-figée qui peut être décrite sous la forme de grammaire locale, le problème est résolu rapidement : c'est le cas de *dans l'après-midi* et *à huit heures* qui sont reconnues par les grammaires de dates (D. Maurel, 1990 ; M. Gross, 2002). Dans les autres cas, il faut regarder le groupe nominal qui suit la préposition. En général, si le nom-tête de ce groupe nominal est concret, alors on a très souvent affaire à une préposition locative : *dans la forêt*, *après l'étang*, etc. Cependant, ces constatations ne sont que des tendances. Il existe de nombreux contre-exemples. Des noms concrets peuvent être noms-têtes de compléments de temps :

Max partira (après + dans) deux verres
= Max partira après qu'il aura bu deux verres

Mais le comportement du nom *verre* n'est pas exclusif et ce dernier peut très bien appartenir à un complément locatif :

La fourmi a bu dans deux verres

De même, comment déterminer localement la classe à laquelle appartient l'adverbe comprenant le nom *facteur* dans les phrases suivantes :

Max est passé après le facteur = Max est passé après que le facteur soit passé (temps)
? Sur cette image, Max est assis après le facteur [58] (lieu)

Il existe également un problème avec les noms humains collectifs qui désignent à la fois un lieu et un ensemble de personnes :

A la course de voile, Centrale a coupé la ligne avant cette école (temps)
Pour aller à la maison, tu dois tourner avant cette école (lieu)

Les noms prédicatifs désignant un événement se retrouvent plus facilement dans des adverbes de temps : *après leur rencontre*. Cependant, rien n'est moins sûr avec la phrase suivante où

[58] on considère le *facteur* comme un objet inerte.

l'on est incapable de dire si l'on a un adverbe de temps ou un adverbe de lieu (sans doute les deux).

Au concert de Johnny, Marie a rencontré Luc.
Où Marie a-t-elle rencontré Luc ? au concert de Johnny
(= *dans le lieu où Johnny a fait son concert*)
Quand Marie a-t-elle rencontré Luc ? au concert de Johnny

Par ailleurs, pour une même préposition locative, il existe plusieurs emplois locatifs. C'est le cas de la préposition *sur*. L'interprétation sémantique dépend de la nature (aspect, géométrie, etc.) du nom *N* suivant la préposition. (cf. C. Vandeloise, 1986)

Le plateau est sur la table.
= Le plateau est posé sur la table

La branche est sur l'eau
= La branche flotte sur l'eau

L'alpiniste est sur la falaise
= Luc est accroché à la falaise

Luc est actuellement sur Paris
= Luc est actuellement à Paris

On distingue différents sens pour ces quatre phrases. La première phrase indique un contact de surface entre le plateau et la table dans une position horizontale, le plateau étant au-dessus de la table (cf. dessin ci-dessous).

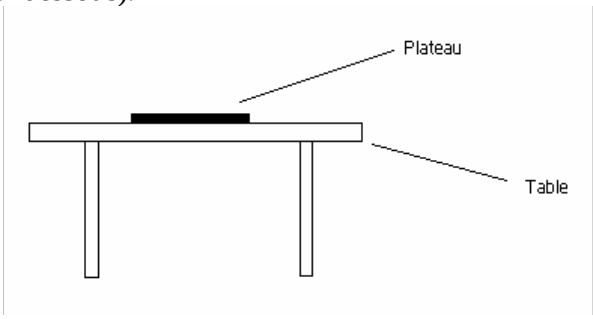

Figure 70 : sur (*plateau*, *table*)

Pour la deuxième phrase, le contact est différent. En effet, la branche est partiellement immergée dans l'eau.

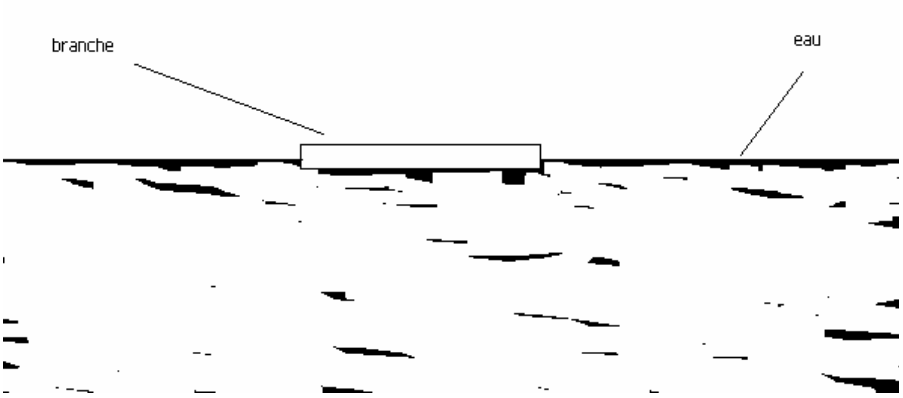

Figure 71 : sur (*branche, eau*)

La troisième phrase indique un contact de surface entre la falaise et l'alpiniste, dans une position verticale (l'alpiniste est accroché).

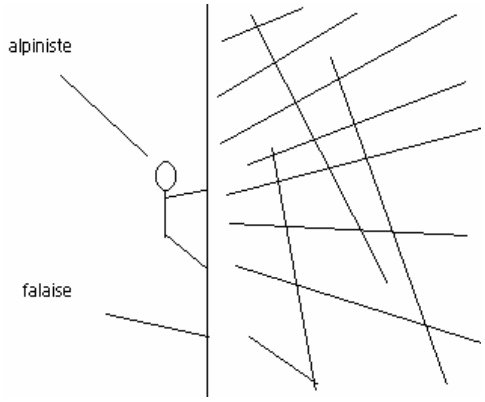

Figure 72 : sur (*alpiniste, falaise*)

En l'absence d'un modèle formel du sens, ces distinctions ne sont pas systématisables. Il existe d'ailleurs des intermédiaires entre ces trois situations.
Enfin, la quatrième phrase est équivalente à *Luc est à Paris*, mais avec une nuance sémantique.

Les prépositions locatives peuvent aussi appartenir à des adverbes figés locatifs (cf. L. Danlos, 1980 ; M. Gross, 1986, 1996). Dans ce cas-là, l'analyse est plus simple car le sens est rattachable à la séquence entière et non représentable par la composition des différents sens des constituants de l'adverbe :

Max est sur la route
Où est Max ? sur la route

Max est à l'asile
Où est Max ? à l'asile

Marie est à l'air du grand large (M. Gross, 1996)
Où est Marie ?A l'air du grand large

En général, les compléments locatifs sont libres et la distribution des prépositions dépend en gros de la géométrie et des propriétés physiques des arguments (cf. C. Vandeloise, 1985).

Cependant, comment peut-on expliquer les différences de distribution des prépositions dans les phrases suivantes où les arguments sont des pièces d'un bâtiment ? (M. Garrigues, 1995).

Luc est (à la + dans la + en) cuisine.
*Luc est (*à la + dans la + *en) chambre*
*Luc est (à la + dans la + *en) salle de bains*
Luc est (à la + dans la + en) salle d'opération

L'utilisation de la préposition *en* peut s'expliquer pour *cuisine* et *salle d'opération* car les phrases impliquent que Luc est l'acteur d'un procès : Luc travaille dans la cuisine et Luc subit ou pratique une opération dans la salle d'opération. Par contre, la distribution de *à* est difficilement explicable. Il en est de même pour les noms *rue* et *place*. Pourquoi observe-t-on une différence de distribution entre ces deux noms pourtant proches ?

*Luc est (dans + *sur) la rue*
*Luc est (*dans + sur) la place*[59]

On peut imaginer que l'utilisation de la préposition *dans* avec le nom *rue* s'expliquerait par le fait qu'une rue peut être vue comme une boîte ouverte, les côtés fermés étant la route et les bâtiments longeant cette route. Il ne semble pas exister d'explication pour *place* (la ressemblance avec un plateau serait une explication bien fantaisiste). Ainsi, il n'est pas évident de prédire exactement la distribution prépositionnelle à partir d'une analyse sémantique basée sur la géométrie et différentes propriétés physiques des arguments. Une solution consisterait à étudier systématiquement le comportement de chaque nom (M. Garrigues, 1995) et de coder pour chacun leur distribution prépositionnelle.

4.2.3.2 Prépositions composées

Nous nous appuyons ici sur les travaux d'A. Borillo (1989) et M. Gross (1996). Nous regardons de plus près les prépositions locatives composées dont la grande majorité ont la forme interne *Prep Det N de* : *à l'intérieur de*, *en amont de*, etc. On travaille ainsi sur la construction *N*0 *Vsup Prep Det N de N*1. A. Borillo (1989) explique que le nom *N* sert à localiser *N*0 par rapport à *N*1. Il ajoute une précision supplémentaire par rapport aux prépositions simples. Il exprime soit une localisation interne soit une localisation externe. Les noms de localisation interne indiquent que *N*0 est localisé au contact ou à l'intérieur de *N*1 : ex. *à l'extrémité de*, *en haut de*, *sur le coin de*, etc. Les noms de localisation externe permettent de localiser *N*0 par rapport à *N*1, mais il n'y a ni contact (*à l'extrémité de la tige*) ni inclusion (*à l'intérieur du combiné*) entre les deux arguments : *à côté de*, *à droite de*, etc. A. Borillo note également la combinaison d'adjectifs de localisation interne (*central*, *supérieur*, etc.) avec des noms qui désignent des partitions de l'objet *N*1 (*partie*, *zone*, etc.) :

La fourmi se trouve sur la zone extérieure du mur
Marie est dans la partie nord de l'île

Elle a systématiquement examiné chaque nom et chaque adjectif, puis codé leurs distributions lexicales dans des tables[60] dont nous nous sommes servi pour construire manuellement des graphes décrivant des prépositions locatives composées. Nous complétons ces listes par celles de M. Gross (1996). Nous prenons notamment les prépositions composées répertoriées dans la

[59] *Luc est dans la place* a un autre sens.
[60] E. Laporte (2002) a aussi réalisé une étude sur ces adjectifs en se servant de cette liste.

table **EPCDN** de M. Gross qui traite des noms (C) figés dans la construction : *N*0 *être Prep C de N*1. Ces expressions ne sont pas uniquement locatives, celles qui le sont sont même minoritaires : environ 36% des entrées ont un emploi locatif (337 sur 933 entrées). L'extraction des expressions locatives est facilitée par le fait qu'une colonne de la table indique si l'expression répond à la question en *où*. Les entrées candidates sont donc celles qui ont un signe '+' à cette colonne. Par ailleurs, nous ajoutons les prépositions ne rentrant pas dans la construction *Prep Det de* comme *face à*. Les graphes construits dans la section sur les expressions de mesure et représentant des prépositions locatives sont également intégrés à l'ensemble des graphes des prépositions composées.
Par ailleurs, une bonne proportion des prépositions composées extraites de ces listes peuvent être transformées en adverbes figés par effacement du groupe nominal libre :

Max est à l'ouest (de Paris + E)
Luc est à (l'intérieur + côté) (de la maison + E)

Certaines n'acceptent pas cette transformation d'effacement :

*Max est en bord (de fleuve + *E)*

Cette propriété a été codée par M. Gross dans la table **EPCDN** dans une colonne. Comme précédemment, nous extrayons manuellement les entrées candidates. Nous complétons nos listes par celles données par D. Le Pesant (2002).

Les prépositions composées permettent de localiser précisément *N*0 par rapport à *N*1, mais cela n'empêche pas que certaines soient ambigues :

Max est au bord du (suicide + fleuve)
*Où est Max ? Au bord du (*suicide + fleuve)*

Cette scène est à la limite du supportable
*Où est cette scène ? *à la limite du supportable*

On note que l'ambiguïté avec les dates est toujours aussi récurrente :

Notre empereur préféré est né à la fin du 18ème siècle. (date)
Luc a épousé Léa au début de l'année dernière. (date)

Nous organisons nos graphes de la manière suivante. Nous représentons 8 formes de prépositions composées :

- *à Det N de*[61] = : *à l'ouest de, à la frontière de*
- *en N de* [62]=: *en amont de*
- *sur Det N de* =: *sur le bord de*
- *dans Det N de* =: *dans le centre de*
- *dans la zone Adj de*[63] = : *dans la zone (ouest + supérieure) de*

[61] Le graphe associé est **ADetNDe**. Pour clarifier le graphe, nous représentons les directions du type *à l'ouest de* dans un graphe différent **ALeOuestDe**.
[62] Le graphe associé est **EnNDe_Loc**.
[63] Le graphe associé est **DansLaZoneAdjDe**. Les adjectifs répertoriés ont des caractéristiques purement géométriques dans l'espace (*supérieur, gauche, extérieur,* etc.). Le sous-graphe **Adj-direction** décrit les adjectifs

- *à Dnum Metre de =: à 10 km (E + à vol d'oiseau) de Paris, à 20 kilomètres au nord de*
- *à un N de Dnum Metre de = : à une hauteur de 30 m de*
- résiduels =: *face à, hors de*

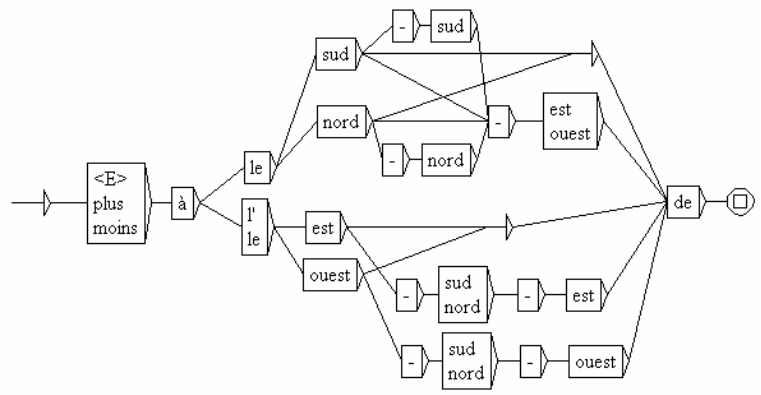

Figure 73 : ALeOuestDe

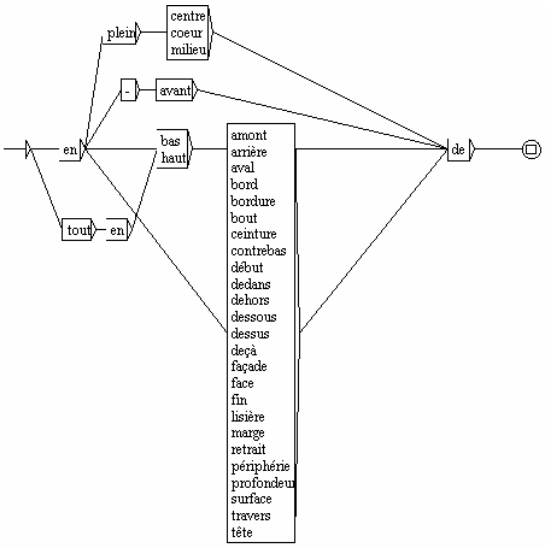

Figure 74 : EnNDe_Loc

nord, est, sud-ouest, etc. Notre grammaire ne reconnaît pas des expressions telles que *dans la zone néerlandaise de la ville* où *néerlandaise* n'est pas une caractéristique spatiale mais une caractéristique démographique de la ville.

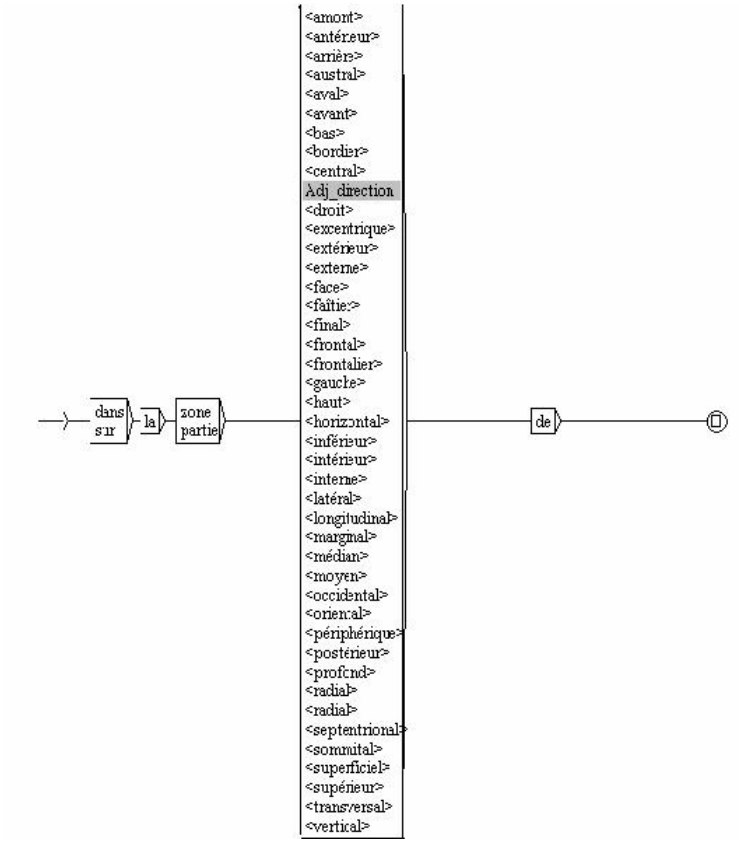

Figure 75 : DansLaZoneAdjDe

4.2.3.3 Quelques statistiques

Nous avons montré dans l'absolu que la localisation d'un complément locatif ne peut se résumer au repérage d'une préposition locative et d'un groupe nominal du fait de l'ambiguïté des prépositions. Dans cette section, nous mesurons quantitativement le taux d'erreur qu'un tel processus introduirait en supposant que la reconnaissance du groupe nominal soit parfaite[64]. Nous prenons une année du journal *Le Monde* (1994) et regardons manuellement le comportement des 1 000 premières prépositions appartenant à l'ensemble {*dans*, *avant*, *devant*, *sur*, *contre*, *après*, *derrière*, *sous*}. Au total, 413 occurrences de ces prépositions ont un emploi locatif, soit 41,3%. La proportion d'emplois locatifs la plus élevée est de 81% avec la préposition *dans*, ce qui très insuffisant. Par ailleurs, nous constatons que certaines prépositions comme *avant* ou *après* appartiennent quasiment toujours à des adverbes de temps (respectivement 90% et 86% des cas). Sur l'ensemble des prépositions trouvées, nous n'avons que 31 prépositions appartenant à des adverbes figés ou semi-figés de temps facilement repérables par des grammaires locales de temps existantes (soit 3,1% de l'ensemble des prépositions) comme

dans l'après-midi
dans la soirée de jeudi
avant le 18 mars

Ceci montre bien la difficulté qu'il y aura à distinguer emplois locatifs et temporels de manière fiable. A terme, le meilleur moyen sera de regarder le groupe nominal du

[64] Ce qui est clairement impossible dans l'état actuel des recherches en linguistique appliquée.

complément locatif. Par exemple, si le nom-tête est un prédicat, la probabilité d'avoir un adverbe de temps sera très grande ; par ailleurs, si c'est un nom concret, la probabilité d'avoir un adverbe locatif sera aussi très grande.

Certaines prépositions comme *contre* ou *sur* sont souvent des prépositions attachées à des prédicats du type

Max proteste contre l'inflation
Luc fait une enquête sur la vie mouvementée de Max

Dans ces cas-là, l'ambiguïté est résolue naturellement par l'analyse syntaxique de la phrase.
Quelques résultats détaillés sont donnés dans la table ci-dessous qui utilise les notations suivantes :

NO : Nombre d'occurrences
NOL : Nombre d'occurrences à emploi locatif
%NOL : pourcentage d'occurrences à emploi locatif

Loc simple	NO	NOL	%NOL
dans	426	345	81
avant	51	0	0
devant	18	9	50
sur	294	76	26
contre	67	0	0
après	96	2	2
sous	45	3	7
derrière	3	2	67
Total	1000	413	41

Table 9 : proportion d'emplois locatifs des prépositions {*dans, avant, devant, sur, contre, après, derrière, sous*} dans notre corpus

Ci-dessous, nous donnons un ensemble d'exemples trouvés dans les textes où la préposition reconnue n'a pas un emploi locatif :

[Noms composés]

après-guerre
après-midi

[adverbes figés ou semi-figés]

avant tout
avant la fin de l'année
dans le même temps
dans les mois à venir
dans ces conditions

[adverbes libres]

dans un hébreu limpide

[prédéterminants]

dans la limite mensuelle de 1 000 francs

[prépositions d'un argument d'un prédicat]

Furieux devant l'entêtement de le jeune douanier (…)
un recours hiérarchique devant le garde de les sceaux contre le juge
(…) a lancé une croisade nationale contre la violence à la télévision

[conjonctions de coordination]

sept à huit heures de avion de Tokyo contre onze de Los Angeles

[conjonctions de subordination figées]

dans la mesure où des enfants meurent, les (…)

[phrases figées]

Sommée de balayer devant sa porte, l'industrie (…)
Max a du pain sur la planche
La jeune indienne a été tuée sur le coup

Parmi les phrases figées données ci-dessus, aucune ne répond à la question en *où* :

*Où l'industrie est-elle sommée de balayer ? *devant sa porte*
*Où Max a-t-il du pain ? * sur la planche*
Où la jeune indienne a-t-elle été tuée ? sur le coup*

Notons que nous avons traité les prépositions *à* et *en* séparément car elles sont très fréquentes par rapport aux autres. Nous avons trouvé 30% de prépositions *à* qui ont un emploi locatif dans le texte (40 % pour *en*)[65]. Pour les emplois locatifs, la très grande majorité de ces prépositions appartiennent à des adverbes contenant un nom propre de lieu (*à Paris*, *en France*). Ainsi, un bon moyen d'améliorer le repérage local des adverbes locatifs est d'étudier le comportement des noms propres de lieu dans ces adverbes (cf. sections suivantes).

D'autre part, nous avons voulu mesurer quantitativement l'apport de la reconnaissance des prépositions locatives composées dans le repérage des compléments locatifs. Nous regardons les prépositions du type *Loc Det N de*. Nous avons appliqué les grammaires de ce type de préposition et examiné manuellement les 1 000 premières occurrences. Nous constatons que 529 d'entre elles sont correctement analysées et ont un emploi locatif (soit environ 53%). Nous donnons ci-dessous quelques exemples d'emplois non locatifs trouvés :

[65] C'est un calcul assez approximatif car nous avons regardé les 200 premières occurrences de chacune des prépositions. Mais notre but est de donner une vague tendance.

[prépositions non locatives]

Luc a été gentil à l'endroit de Marie (*à l'endroit de = envers*)
L'argent est à la base des malheurs de Paul (?à la base du fait que P)
A côté de la crise de septembre, celle-ci est mineure (*à côté du fait que* P)

[adverbes figés non locatifs]

Au bout du compte, Léa est très accueillante

[phrases figées non locatives]

Max est passé à côté de quelque chose de grand

[modifieurs du nom]

L'engagement à droite de cet historien n'est pas passé inaperçu

D'autres prépositions ont un emploi locatif mais sont mal analysées du fait de l'ambiguïté du nom *N* :

Le lieutenant Martin a été affecté à la base de Pau (*N = : base (E + militaire)*)

Dans une perspective d'une analyse sémantique, la reconnaissance des prépositions composées augmente la précision de la localisation car le nom *N* dans ces prépositions indique une information supplémentaire de localisation. La reconnaissance des compléments locatifs avec des prépositions composées est un peu plus facile que celle des compléments locatifs avec des prépositions simples (écart approximatif de 10%). Le comportement des prépositions est très hétérogène. En effet, dans beaucoup de cas, soit toutes les occurrences sont locatives (*à l'ouest de*, *au cœur de*, etc.), soit aucune occurrence (ou presque) n'est locative (*au début de*, *à la fin de*, etc.). En utilisant une approche statistique, ce résultat revient à considérer que les prépositions du deuxième cas (ex : *au début de*) ne peuvent être locatives, ce qui n'est pas vrai dans l'absolu. Par contre, certaines ont un comportement relativement équilibré : *au milieu de* avec 51% de locatifs. Des formules statistiques simples pourraient largement améliorer les résultats grâce aux comportements extrêmes. Cependant, les prépositions composées sont très peu fréquentes par rapport aux prépositions simples comme le montre le tableau ci-dessous. Ainsi, il n'est pas possible d'améliorer les résultats généraux de manière significative en améliorant l'analyse des compléments locatifs à préposition composée.

NO : nombre d'occurrences de *Loc*
NCL : nombre d'occurrences de *Loc* dans une préposition composée locative de la forme *Loc (E+Det) N de*
%NCL1 : pourcentage de prépositions composées locatives (*Loc Det N de*) par rapport au nombre de *Loc* à emploi locatif ou pas dans le texte
%NCL2 : pourcentage de prépositions composées locatives (*Loc Det N de*) par rapport au nombre de *Loc* à emploi locatif dans le texte

Loc simple	NO	NCL	%NCL1	%NCL2[66]
à	3059	56	1,8	6,1
sur	486	2	0,4	1,6
en	1455	4	0,3	0,7
Total	5000			

Table 10 : proportion de prépositions composées par rapport aux prépositions simples

4.3 Grammaires locales de noms propres composés de lieu

4.3.1 Remarques préliminaires

L'objectif principal de ce chapitre est d'étudier le comportement syntaxique d'un ensemble de noms *N* au sein de la construction locative *N*0 *Vsup Loc Det N Modif*. Nous limitons cet ensemble aux noms propres de lieux géographiques :

La fête se passe (à Paris + sur la Seine)
Marie se trouve en Californie
Luc est dans la mer Méditerranée

Cette étude présente beaucoup d'intérêt car elle complète les travaux effectués sur les expressions en *être Prep X*. Elle est également un sérieux apport pour l'analyse automatique de textes du fait de la haute fréquence des noms propres de lieu dans certains types de corpus tels que les textes journalistiques. Avant toute étude sur ces constructions, il est nécessaire de comprendre le comportement interne des noms propres de lieu. Nous proposons une description entièrement lexicale et syntaxique, basée sur la méthodologie du lexique-grammaire.

Des études linguistiques ont été consacrées aux noms propres depuis longtemps comme le montre le volume de *Langages* (J. Molino, 1982) consacré à ce sujet. Mais, depuis quelques années, elles ont pris une nouvelle ampleur du fait du développement du TAL et de la fréquence des noms propres dans les textes. Pour toutes les références sur ce sujet, nous conseillons au lecteur de se référer à K. Jonasson (1995). De manière générale, les études sur ce sujet ont cherché à donner un sens aux noms propres car leur syntaxe paraît limitée. Dans cette étude, nous regardons les noms propres de lieu d'un autre point de vue. En effet, nous considérons ces objets linguistiques comme des formes composées comprenant un classifieur de lieu et nous remettons la syntaxe au centre de la discussion. Par exemple, *Méditerranée* est la forme réduite de *mer Méditerranée* et a un comportement syntaxique différent de *mer du Nord*. Par une étude systématique, nous montrons que la méthodologie du lexique-grammaire est parfaitement adaptée au traitement de ces objets.

Dans le domaine du TAL, les principaux travaux réalisés jusqu'à présent consistent à localiser les noms propres dans les textes et à leur assigner des classes sémantiques : T. Wakao et al. (1996), J. Senellart (1998), A. Cucchiarelli et al. (1999), N. Fourour et al. (2002), N. Friburger (2002), etc. Les méthodes utilisées consistent essentiellement à exploiter la présence de classifieurs : par exemple, un nom propre précédé du nom *Monsieur* (*Monsieur Chirac*) est classé comme un nom de personne. Les systèmes de réponses automatiques à des questions sont une des applications de ces classifications automatiques : O. Ferret et al. (2001), C. Fairon et P. Watrin (2003), etc.

[66] Calculé par la formule suivante %NCL2 = NCL /(%NOL*NO) avec %NOL étant le pourcentage des prépositions simples *Loc* ayant un emploi locatif. (exemple pour *à* : %NCL2 = 56/(30/100*3059).

Du fait que les noms propres en général sont en nombre « quasi-infini » et varient énormément au cours du temps, il est impossible de tous les répertorier dans des dictionnaires, d'où l'intérêt d'outils automatiques d'extraction et de classification. Certaines classes sont un peu plus fermées et plus stables, c'est le cas des toponymes. Ainsi, D. Maurel et O. Piton (1998) se sont lancés dans la construction d'un dictionnaire de toponymes[67] (précisément les noms de régions, de pays et de villes) et d'hydronymes : le dictionnaire électronique *Prolintex*. Ce projet est clairement gigantesque mais son intérêt est indéniable pour le TAL. Chaque entrée du dictionnaire contient plusieurs types d'informations :

- d'ordre linguistique : identification de la classe, information flexionnelle, information syntaxique sur les déterminants pour certains toponymes comme les villes ;
- d'ordre extra-linguistique (positionnement géographique : par exemple, *Paris* est une ville de France)[68]

Par exemple, l'entrée *Tours,.N+PR+DetZ+Toponyme+Ville:ms:fs* (provenant de la deuxième version sans information d'ordre extra-linguistique) signifie que *Tours* a les caractéristiques suivantes : c'est un nom propre (*N+PR*), toponymique (*+Toponyme*), désignant une ville (*+Ville*), qui n'a pas de déterminant obligatoire (*+DetZ*), qui peut s'accorder au masculin singulier et au féminin singulier (*:ms:fs*). Le premier intérêt de ce dictionnaire est de reconnaître avec une grande précision les noms propres de lieu. Ensuite, il est une première base solide pour un système de traduction des toponymes (T. Grass et al., 2002). Les outils d'extraction automatique des noms propres peuvent servir à trouver de nouvelles entrées candidates à ce dictionnaire.

Dans cette partie, nous étudions le comportement syntaxique de noms propres composés notés *Nprc*. En général, les noms propres de lieu que nous utilisons sont des formes réduites (ou courtes) de séquences composées. En effet, chaque lieu a un nom (*Npr*) qui peut être complexe (*Pas-de-Calais, Los Angeles, etc.*) et appartient à une classe de lieu (définie par un ou plusieurs classifieurs *Nc*) : *région Pas-de-Calais, ville de Los Angeles*. Dans certains cas, le nom du lieu est suffisant pour désigner ce lieu comme *Pyrénées-Atlantique* ou *Méditerranée*. Ce sont en fait des formes réduites de noms propres composés comprenant un classifieur *Nc* et un nom *Npr* :

> *Pyrénées-Atlantiques* est la forme courte de *département de les Pyrénées-Atlantiques*
> *Méditerranée* est la forme courte de *mer Méditerranée*

Désormais, nous désignons ces noms propres composés par le couple (*Npr, Nc*) : nous notons la forme longue *Nprc* et la forme courte *Npr*. Notre approche consiste à regarder la composition interne de tels noms composés et à constituer des lexiques en s'appuyant sur les travaux de D. Maurel et O. Piton. Notre travail se distingue de ce dernier par le fait, notamment, que nous ne nous intéressons pas aux propriétés extra-linguistiques et nous codons dans nos lexiques de nouvelles contraintes lexicales et syntaxiques qui foisonnent dans ce type de noms. Ce travail sert de base à notre étude sur la distribution prépositionnelle dans les adverbes locatifs. Dans un premier temps, nous justifions le terme de nom composé utilisé pour le couple (*Npr, Nc*) à l'aide d'arguments linguistiques. Dans ce cas, nous parlerons plutôt de noms propres composés étendus. Ensuite, nous effectuons une classification formelle des noms propres composés de lieu. Puis, nous examinons les contraintes syntaxiques internes auxquels ils sont soumis. Par la suite, nous regardons leur comportement dans des groupes nominaux simples : déterminants et modifieurs. Enfin, nous

[67] cf. aussi D. Maurel et al. (1995).
[68] La deuxième version sortie en 2003 ne contient plus ce genre d'information et se trouve à l'URL suivante : http://www.li.univ-tours.fr/BdTln/Prolintex.html.

codons toutes les contraintes trouvées sous la forme de tables syntaxiques et de grammaires locales. Notons que nous n'avons pas la prétention de couvrir tous les noms propres composés de lieu existants (et loin de là). Il existe trois raisons à cela :

- la liste exhaustive de tous les noms propres de lieu est immense (cf. la construction de *Prolintex* qui dure depuis huit ans).
- le comportement syntaxique est souvent flou pour les noms de lieu peu connus : nous avons consitué nos lexiques à partir de nos connaissances géographiques qui relèvent de la culture générale et nous verrons, malgré tout, que les données accumulées comprennent un grand nombre de contraintes syntaxiques.
- notre objectif est surtout méthodologique : montrer une méthode claire et rigoureuse d'accumulation de noms propres de lieu.

Ce sujet présente une autre difficulté. Un nom propre peut admettre plusieurs classifieurs sans être ambigu pour autant : ainsi, une ville peut aussi être une station de ski.

la (station de ski + ville) de Courchevel
*la (*station de ski + ville) de Paris*

Pour certains noms propres, le classifieur n'est pas clair comme pour les petites villes : est-ce un bourg, un village, une bourgade ? Il existe un autre cas où le classifieur n'est pas clair : les noms de massifs montagneux comme les *Alpes*. En effet, il est possible de dire *(massif + chaîne) des Alpes*. S'il n'existe pas de classifieur clair dans la forme longue d'un nom propre répertorié, nous ne tenons pas compte de ce nom propre. Pour les villes, nous ne traitons que les grosses villes qui sont clairement des villes (ex : *Paris, Bordeaux, le Pirée, La Havane*).

4.3.2 Critères de définition des Nprc

Des critères formels sont nécessaires pour distinguer les différents types de séquences de la forme *Detc Nc de (E + Det) Npr* :

(a) *Une ville de France*
(b) *La ville de Pagnol*
(c) *La ville de Paris*[69]

Nous distinguons formellement ces trois cas en examinant les phrases élémentaires de base à partir desquels sont formés les groupes nominaux :

(a) N0 être situé Loc N1 (*Nc* nom tête de *N*0 et *Npr* celui de *N*1)
UN Nc qui être situé Loc (E + Det) Npr =: *une ville qui est située en France*
= * *UN Nc de Loc (E + Det) Npr* =: *une ville de en France* (M. Gross, 1986)
= *UN Nc de (E + Det) Npr =: une ville de France*

(b) N0 vivre Loc N1 =: Pagnol vit dans une ville
la ville (où vit + de) Pagnol[70]

(c) (E + Det) Npr être UN Nc =: Paris est une ville
la ville (qu'est + E) Paris

[69] On suppose que *Paris* est un nom de ville et non un nom de personne par exemple.
[70] Il existe bien d'autres interprétations pour ce type de groupe nominal.

Dorénavant, la structure que nous étudions est la structure (c).

4.3.3 Statut syntaxique des Nprc

Quel est le statut syntaxique des séquences que nous traitons ? Une première approche consisterait à considérer ces séquences comme des appositions (M. Rothenberg, non daté). Les séquences *la mer Méditerranée, la ville de Paris, l'état de Californie* et *l'île d'Ouessant* peuvent être mises sur le même plan que *le roi Louis XIV* où la séquence *Louis XIV* est traditionnellement considérée comme une apposition. La préposition *de* se trouvant entre *Nc* et *Npr* est le principal obstacle à cette solution même si elle est parfois considérée comme une préposition neutre d'apposition[71]. La séquence *le roi Louis XIV* est à raprocher de la phrase classificatrice réflexive :

Louis XIV est un roi / Ce roi est Louis XIV

ce qui revient à admettre l'hypothèse d'une transformation

le roi qu'est Louis XIV
= le roi Louis XIV

On observe le même comportement pour *la ville de Paris* et *l'île d'Ouessant.*

Paris est une ville/ Cette ville est Paris
Ouessant est une île / Cette île est Ouessant

La ville qu'est Paris = la ville de Paris

La séquence *la mer méditerranée* se comporte différemment alors qu'elle a la même forme de base (sans *de*) que *le roi Louis XIV* : le déterminant *la* précédant *Méditerranée* est obligatoire dans la phrase classificatrice.

*(*E + la) Méditerranée est une mer / Cette mer est (*E + la) Méditerranée*

Il en est de même pour *l'état de Californie* qui a obligatoirement besoin du déterminant *la* dans la phrase classificatrice :

*(*E + la) Californie est un état / Cet état est (*E + la) Californie*

L'utilisation du déterminant indéfini *un* sans modifieur est interdite comme dans les appositions :

*G. Bush n'apprécie pas (le + *un) président Chirac*
*Les français n'appréciaient pas (le + *un) roi Louis XIV*

*Les touristes apprécient (l' + *une) île de Ouessant*
*Luc contemple (la + *une) mer Méditerranée*

[71] Les cas d'alternance entre la forme en *de* et la forme sans *de* sont rarissimes :

la région (E + du) Nord-Pas-de-Calais

Dans les séquences figées telles que *mer du Nord, île du Diable, vallée de la Mort* ou *mer Noire*, les phrases classificatrices dissociant *Nc* et *Npr* ne sont pas autorisées :

**Le Nord est une mer*
**Le Diable est une île*
**La Mort est une vallée*
**(E + La) Noire est une mer*

Ainsi, les groupes nominaux qui nous intéressent présentent à la fois des ressemblances et des différences avec les appositions. Il en est de même de certaines formes à appositions sans *de* : la séquence *le bâtiment A* ne peut être rapprochée d'une phrase classificatrice telle que *A est un bâtiment*.

Dans l'une des dernières conversations que nous avons eue avec lui, M. Gross parlait de déterminant nominal[72] pour la séquence *la ville de*, ce qui pourrait sembler exact à première vue :

Les touristes apprécient (la ville de + E) (Toulouse + New York + La Havane)

La possible insertion d'adjectifs fragilise cette analyse. En effet, les déterminants nominaux acceptent à peu près uniquement des adjectifs intensifs :

Cette entreprise a embauché une (E + belle + grosse) fournée d'ingénieurs
Cette entreprise a embauché une fournée (E+ hallucinante + importante) d'ingénieurs
Cette entreprise a embauché (des + trois cents) ingénieurs

La variation lexicale des modifieurs des déterminants nominaux est bien plus étendue avec *la ville de* :

La ville (surpeuplée + polluée + américaine + …) de Mexico

G. Gross (1991) parle de construction inverse et place les constructions du type *la ville de Paris* sur le même plan qu'une séquence du type *ce salaud de*[73] :

(*Ce + Le) salaud de Luc m'a craché à la figure*

En effet, il montre que les séquences *ce salaud de* et *la ville de* proviennent d'un même type de phrase classificatrice :

Boston est une ville
Max est un salaud

Notons que la réduction de la phrase classificatrice contenant le nom *salaud* interdit la présence du déterminant du sujet :

Le facteur est un salaud ; il a couché avec ma femme
*= Ce salaud de (E + *le) facteur a couché avec ma femme*

[72] Un déterminant nominal très particulier qui pourrait s'appeler déterminant nominal locatif.
[73] La classe des noms appartenant à cette séquence est facilement listable : *connard, ignorant, énergumène*, etc.

Ce n'est pas le cas avec tous les noms propres de lieu :

Le Nord est un département ; il est connu pour son climat froid
*= le département de (*E + le) Nord est connu pour son pétrole*

On peut également se demander si les objets que nous étudions sont des noms composés. D'après G. Gross (1996), un nom composé a généralement la constitution syntaxique d'un groupe nominal libre : par exemple, *Det N Adj* = : *un pantalon bleu* (libre) + *un cordon bleu* (nom composé). Nos couples (*Nc, Npr*) ont clairement la même constitution qu'un groupe nominal libre :

- *Det N de Npr* =: *la vallée d'Aspe* [composé] + *la maison de Luc* [libre]
- *Det N de Det Npr* =: *le col du Somport* [composé] + *la pente du Vignemale* [libre]
- *Det N Npr* =: *la mer Méditerranée* [composé] + *le colonel Dupond* [libre]
- etc.

Par contre, les constituants syntaxiques d'un nom composé présentent un certain figement. Dans le meilleur des cas, le sens global du mot composé n'est pas compositionnel : il ne peut pas être calculé simplement à l'aide du sens des différents constituants. Cela apparaît très clairement avec le nom *cordon bleu* (excellent cuisinier). Mais ce n'est pas toujours le cas : *vin blanc* peut notamment être partiellement analysé à l'aide de la phrase classificatrice suivante :

Un vin blanc est un vin

Cette interprétation n'est pas valable pour *cordon bleu* (excellent cuisinier) ou *panier percé* (dépensier) :

** Un cordon bleu est un cordon*
** Un panier percé est un panier*

Traditionnellement, les noms composés du type *vin blanc* sont appelés noms composés endocentriques et ceux du type *cordon bleu* sont appelés noms composés exocentriques. Les couples (*Npr, Nc*) entrent clairement dans la première catégorie :

La mer (du Nord + Méditerranée) est une mer
Le mont (des Oliviers + Ventoux) est un mont

Dans certains cas, la phrase classificatrice n'est pas valide :

**la mer de Glace est une mer*

En effet, l'utilisation du classifieur *mer* est ici une métaphore lexicalisée. On peut également noter le cas d'*Ile de France* dont le classifieur est *région* et non *île* :

**l'Ile de France est une île*
l'Ile de France est une région

Le figement de certaines séquences est clair comme pour les expressions *mer du Nord* et *île du Diable*. Par ailleurs, la composition interne dépend de chaque couple et n'est pas toujours prédictible renforçant alors l'impression de figement de ces séquences :

*la mer (E + *d') Adriatique + la mer(*E + d')Aral*
=> l'(Aral + Adriatique) est une mer

*l'île (*E +de la) Réunion + l'île (E + *de) Maurice)*
=> (la Réunion + Maurice) est une île

D'autre part, certains couples (*Npr, Nc*) sont obligatoirement au pluriel[74] :

*(Les îles + *L'île) de les Canaries*
=> Les Canaries sont des îles.

Pour certains classifieurs *Nc*, le figement des couples (*Npr, Nc*) est moins net. Par exemple, les couples ayant pour classifieur *ville* ont un comportement prédictible si l'on connaît les propriétés syntaxiques propres à *Npr*. En effet, certains prennent des déterminants dont la première lettre est une majuscule[75] comme *Le Havre*, *La Havane* ou *Les Saisies*. Ces informations sont codées dans *Prolintex* sous la forme d'un trait syntaxique (+*DetLe* pour *Le Havre* ; +*DetLa* pour *La Havane* ; +*DetLes* pour *Les Saisies*). Les noms de villes ne prenant pas de déterminant comme *Toulouse* ont le trait syntaxique +*DetZ*.

la ville de (E + Det) Npr = : *la ville de (La Havane + Toulouse + Les Saisies)*

Dans ces exemples, *Npr* sélectionne un type de classifieur (*ville*) et non un autre (*mer* est interdit). Mais cette sélection lexicale est-elle suffisante pour considérer ces séquences comme des noms composés ? Le fait que l'on puisse insérer des modifieurs entre le nom *ville* et la préposition *de* fragilise cette thèse :

La ville musulmane de Téhéran
La ville francophone de Québec

De plus, ce type de comportement est habituellement décrit comme un figement mais avec la notion de nom approprié.

Pour finir cette discussion, nous évoquons un dernier point pouvant laisser penser que les objets que l'on traite sont des noms composés. En effet, il est bien connu que la présence d'un modifeur est interdite dans la séquence suivante :

*Max est en mer (E + *déchaînée)*

Par contre, la phrase suivante est parfaitement naturelle :

Max est en mer (Méditerranée + du Nord)

[74] Pour les îles, ces séquences au pluriel désignent des ensembles d'îles.
[75] On trouve des textes où ces déterminants n'ont pas de majuscule.

Cela montre que *Méditerranée* et *du Nord* ne sont pas des modifieurs classiques et que *mer Méditerranée* et *mer du Nord* pourraient former une unité. Par ailleurs, si l'on prend la *vallée d'Aspe*, on observe les situations suivantes :

** Luc est en vallée (E + fertile)*
Luc est en vallée de (Aspe + la Tarentaise)
** Luc est en vallée de (Aspe + la Tarentaise) fertile*

On a le même phénomène avec *école d'ingénieur* qui est clairement un nom composé.

**Max est en école (E + célèbre + mixte)*
Max est en école d'ingénieur
**Max est en école d'ingénieur célèbre*

Cette discussion est moins intéressante par son côté terminologique que par le fait qu'elle montre les particularités linguistiques des expressions que l'on traite. Il faut retenir de cette section l'existence claire de figements à différents niveaux : composition interne et détermination. Nous parlerons dorénavant de noms propres composés étendus que nous abrégerons en noms propres composés.

4.3.4 Composition syntaxique des formes longues et classification

Nous proposons maintenant d'examiner la structure interne des noms propres composés. De manière générale, la composition interne des noms composés est extrêmement variée :

- Adjectif Nom (AN) : *faux-cul, rouge gorge*, etc.
- Nom Adjectif (NA) : *cordon bleu, carte bleue*, etc.
- Nom de Nom (NDN) : *acte de foi, gardien de but*, etc.
- Nom à Nom (NAN): *panier à pain, moulin à vent*, etc.
- Nom Adjectif de Nom (NADN) : *offre publique d'achat*, etc.
- Etc.

Il en est de même pour les noms propres composés pour lesquels nous utilisons une notation similaire : *N* pour nom, *A* pour adjectif, *P* pour préposition, *D* pour déterminant, *Npr* pour nom propre (forme courte), *Npr-a* pour un adjectif morphologiquement dérivé du nom propre *Npr*. Jusqu'à présent, nous avons répertorié les formes longues suivantes :

- *LE Nc Npr* =: *la mer Méditerranée* (NNpr)
- *LE Nc Adj Npr* =: *l'océan glacial Arctique* (NANpr)
- *LE Nc Prep Npr* =: *l'état de Californie* (NPNpr)
- *LE Nc Prep Det Npr* =: *l'île à les Moines + le pic de le Midi* (NPDNpr)
- *LE Nc Npr-a*[76] =: *la République française* (NNpr-a)
- *Le Nc Adj de Npr* =: *la République arabe d'Égypte* (NAPNpr)
- *LE Nc Adj de Det Npr* =: la *République démocratique de le Congo* (NAPDNpr)

Il existe des noms officiels de pays ayant des structures syntaxiques internes encore plus complexes :

- *LE Nc Adj et Adj de Npr* = : *la République populaire et démocratique de Corée*

[76] *Npr-a* est l'adjectif morphologiquement lié à *Npr* (ex : pour *Npr* = : *France*, *Npr-a* = : *française*).

- *LE Nc Adj Npr-a Adj et Adj= : la Jamahiriya arabe libyenne populaire et socialiste*

La très grande majorité des noms propres de lieu rentrent dans au moins une des trois structures suivantes : NNpr, NPNpr ou NPDNpr. Les structures contenant un adjectif concernent essentiellement les noms de pays dont les formes officielles courtes et longues sont répertoriées à partir de la liste diffusée par la Délégation Générale à la Langue Française[77]. Ils ont également été étudiés par O. Piton et al. (1997) qui, à partir de cette liste, ont systématiquement regardé leur comportement dans un texte journalistique. Nous examinons, pour l'instant, les formes *LE Nc ((de) Det) Npr* où les séquences entre parenthèses sont optionnelles. Ces séquences sont très largement majoritaires dans l'ensemble des noms propres de lieu. Comme nous l'avons mentionné précédemment, la plupart des noms propres rentrent dans au moins l'une de ces structures :

mont Ventoux, mer Méditerranée, île Maurice (NPDNpr)
île de Malte, col de Splandelle, mer de Barentz (NPNpr)
département de la Gironde, col de le Tourmalet (NPDNpr)

Les déterminants *Det* de la forme longue de type NPDNpr sont limités aux déterminants définis suivants : *le*, *la* et *les*. Nous donnons ci-dessous quelques exemples :

l'île de (la Barbade + le Diable)
l'état de (le Texas + la Californie)
le col de (le Tourmalet + la Colombière)
le mont de les Oliviers

Les autres déterminants sont exclus. Par exemple, les séquences suivantes sont clairement interdites :

** L'île de (cette + sa + une) Barbade*
** l'état de (ce + son + un) Texas*
** le mont de (ces + ses + des) Oliviers*

Certains couples rentrent dans plusieurs structures équivalentes. Dans plusieurs noms propres dont la forme longue comprend la préposition *de*, le déterminant *Det* est optionnel :

La principauté de (E + le) Liechtenstein
L'état de (le + E) Vermont[78]
L'état de (le + E) Washington[79]

Parfois, c'est la préposition *de* qui est facultative :

Les îles (E + de les) Canaries
Le désert (de + ?E)Mojave[80]

Dans certains cas comme pour le couple (*Ile-de-France, région*), les trois structures internes sont possibles :

[77] http://www.culture.fr/culture/dglf/ressources/pays/pays.htm.
[78] La forme *état de Vermont* trouvée dans le Monde 1994 (1 occurrence).
[79] La forme surprenante *état du Washington* aussi trouvée dans le Monde 1994 (2 occurrences).
[80] La forme *désert Mojave* a été trouvée dans le Monde 1994 (1 occurrence).

La région (de la + de + E) Ile-de-France

Certains noms propres n'ont pas de forme longue associée à *Npr*, c'est le cas de *Manche* et *Canada* :

La Manche est une mer
** la mer (E+ de + de la) Manche*

Le Canada est (un pays + un état)
** l'état du Canada*

Pour certains, on peut facilement associer un classifieur : la *Manche* est clairement une mer ; le *Canada* est à la fois un pays, un état, etc. et il est très difficile de faire un choix clair.

Ces premières constatations générales convergent déjà vers la difficulté (voire l'impossibilité) de prévoir la composition syntaxique de la forme longue d'un nom propre, étant donné son couple (*Npr*, *Nc*). Une étude systématique est donc nécessaire. Etant donné la quantité astronomique de noms propres de lieu à répertorier, il est nécessaire de les classer. Une première solution consiste à regrouper les noms propres selon leur structure interne et plus exactement leur structure interne la plus longue. Par exemple, les couples (*Réunion*, *île*), (*Ile-de-France*, *région*), (*Vermont*, *état*) et (*Pirée*, *ville*) seraient insérés dans la classe NPDNpr car leurs formes les plus longues sont de ce type :

L'île de la Réunion
La région (de l' + de + E) Ile-de-France
L'état (de le + de) Vermont
La ville de le Pirée

Le couple (*Iran*, *république*) serait intégré à la classe NAPNpr car sa forme la plus longue contient l'adjectif approprié *islamique* : *la république islamique d'Iran*. Le gros avantage de cette classification est qu'elle est très facile à mettre en place : construction d'une table par structure syntaxique. Ainsi, la classification de formes rares comme *Jamahiriya arabe libyenne populaire et socialiste* ne pose pas de problème. Etant donné un couple et sa forme la plus longue, l'ajout est immédiat. Cependant, cette méthode présente de multiples inconvénients. Tout d'abord, l'intuition linguistique sur les noms propres de lieu n'est pas toujours très fiable. L'acceptabilité de certaines structures peu connues est surtout basée sur les tendances générales, sur des attestations trouvées dans les corpus de travail[81] ou sur des intuitions phonologiques. Ainsi, la classification pour beaucoup de noms propres serait basée sur une sorte d' « approximation »[82]. D'autre part, certaines structures dans lesquelles rentrent certains noms propres sont plutôt marginales dans l'ensemble des noms propres, mais spécifiques de certains classifieurs. Tout d'abord, les noms des océans ont tous la structure NNpr (*océan Atlantique*, *océan Pacifique*, *océan Indien*), sauf un qui comprend un adjectif approprié facultatif entre le classifieur *océan* et *Npr*[83] : *océan (glacial + E) Arctique*. Ensuite, certains classifieurs comme *département*, *république* ou *état* sont très enclins à accepter des

[81] Une occurrence d'une expression dans un texte peut être une erreur de l'auteur ; l'absence d'une expression dans un texte ne prouve pas son inacceptabilité.
[82] On peut remarquer que ces problèmes épistémologiques sont généraux et non spécifiques à cette classification. Cependant, ils sont amplifiés par la spécificité du domaine des noms propres.
[83] Pour les océans, le choix de catégoriser Atlantique comme *Adj* ou *Npr* est arbitraire.

structures adjectivales, ce qui n'est pas le cas de la très grande majorité des classifieurs :

Le département (landais + des Landes)
L'état (de Californie + californien)
La région (de l'Ile-de-France + francilienne)
La république (française + ?de France[84]*)*

Puis, un classifieur tel que *île* possède une particularité : un ensemble d'îles est souvent considéré comme un archipel et l'on observe des formes telles que

L'archipel de (*les îles de + E*) *les Açores*

Les observations générales précédentes mettent en évidence la grande hétérogénéité du comportement des noms propres, même ceux possédant le même classifieur. Cette constatation est vraie pour un certain nombre de classifieurs comme *mer*, *île* ou *république* :

*La mer (de le + *de + *E) Nord*
*La mer (*de le + de + *E) Barentz*
*La mer (*de la + *de + E) Méditerranée*

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

*Le mont (*de + *de le + E)Ventoux*
*Le mont (*de + de les + *E) Oliviers*

*La république (française + ?*de France)*
La république islamique de Iran
La république démocratique et populaire de Corée

Cependant, excepté pour quelques classifieurs comme *état* (au sens de pays) ou *république*, l'ensemble des structures possibles des formes longues est limité à NNpr, NPNpr et NPDNpr (voire NNpr-a), ce qui ne pose pas de problème de représentation dans une table syntaxique. Par ailleurs, pour quelques classifieurs tels que *ville*, *département*, *état* (partie administrative d'un pays), le nombre de structures est clairement limité. Par exemple, le classifieur *ville* n'accepte que les formes longues comprenant la préposition *de* (NPNpr et NPDNpr) :

?**la ville (Paris + Le Havre) est un haut-lieu touristique*
La ville de (Paris + Le Havre) est un haut-lieu touristique
? la ville (parisienne + havraise) est un haut-lieu touristique*

Pour le type NNpr-a, il semble que si l'on remplace le nom *ville* par *cité*, alors la dernière phrase devient plus naturelle :

La cité parisienne est un haut-lieu touristique

[84] Le nom propre *République de France*, bien qu'il nous paraisse interdit, a été trouvé sur Internet via le moteur de recherche *Google*.

Nous donnons ci-dessous un tableau montrant les restrictions sur les structures syntaxiques dans lesquelles peuvent apparaître les classifieurs *ville*, *état* et *département* :

Nc	NNpr	NPNpr	NPDNpr	NNpr-a
ville	-	+	+	-
état	-	+	+	+
département	-	+	+	+

Table 11 : comportement de classifieurs

Nous proposons donc de classer les noms propres selon leur nom classifieur. Cette méthode présente de nombreux avantages. La classification d'un nom propre est immédiate et n'est pas sujette à un risque d'erreur. La distribution des noms propres est mieux répartie, même s'il existe des classes comme les noms d'océans qui sont très petites par rapport à d'autres telles que celle des villes[85]. Le nombre de colonnes des tables sera exactement adapté aux besoins, évitant ainsi d'avoir des parties creuses dans les tables, comme cela aurait pu être le cas dans des tables classées selon la structure syntaxique interne des noms propres donc sémantiquement moins spécifiques. Jusqu'à présent, nous avons travaillé sur une soixantaine de classifieurs. On pourrait reprocher à cette méthode de générer un très grand nombre de tables. Mais, vu le nombre d'entrées potentielles, ce n'est pas un problème, plutôt un avantage. Nous avons dénombré deux gros défauts à cette approche. Tout d'abord, il existe des noms propres qui n'ont pas de formes longues et dont le classifieur n'est pas clair (*Canada*, *Roumanie*, etc.). La solution est de construire une table résiduelle contenant ce genre d'entrées. Ensuite, quelques noms propres comme *mer de Glace* ne rentre pas dans la phrase classificatrice *Detc Nprc être (UN + des) Nc* comme le font la quasi-totalité des noms propres de lieu :

** La mer de Glace est une mer*
La mer de Corée est une mer

Dans cet exemple, la *mer de Glace* n'est pas une mer, mais plutôt un glacier (dans les Alpes). Cet emploi est uniquement métaphorique. Si l'on place cette entrée dans la même classe, il y a une hétérogénéité sémantique. Mais, cela n'est pas très dérangeant car il suffit de rajouter une colonne dans la table des noms de mers désignant cette déviation sémantique.

De manière générale, le processus de classification et d'étude systématique d'objets linguistiques est freiné par les différents emplois que peuvent avoir certains mots de la langue. Notre étude ne fait pas exception à la règle. D'abord, il existe différents emplois pour certains noms propres *Npr*[86]. Par exemple, *Nord* est soit une mer soit un département français. Dans ce cas-là, la distinction est évidente car ils n'ont pas le même classifieur (*mer* et *département*) et sont classés dans deux tables différentes. Il existe un cas très difficile à traiter : les lieux qui ont le même nom et le même classifieur comme *Paris* (*ville*) qui est à la fois une ville de France et une ville des Etats-Unis (et peut-être même ailleurs). Il n'est malheureusement pas possible de distinguer ces deux entrées dans nos tables, sauf si l'on ajoute des propriétés extra-linguistiques reliées à la position géographique des lieux que l'on traite (et l'on sort de notre cadre de travail). Il existe également différents emplois pour les noms classifieurs. Certains d'entre eux sont ambigus. Il existe deux types d'ambiguïtés : le classifieur ambigu peut avoir

- deux emplois du type *Nc* faisant partie d'un nom propre *Nprc* (par exemple, *état* [pays

[85] Il existe des centaines de milliers de ville dans le monde et seulement quatre océans.
[86] Cf. également O. Piton et D. Maurel (2001).

ou région administrative] ou *côte* [bord de mer ou pente])
- un emploi de type *Nc* et un emploi n'entrant pas dans notre étude (ex : *région* [zone administrative ou zone approximative]).

Les premiers cas sont distingués par des critères syntaxiques (ex : *côte* cf. section sur la distribution prépositionnelle) ou simplement par notre connaissance du monde (ex : *état*).
Examinons maintenant le deuxième cas et prenons l'exemple de *région* :

la région du Havre
la région du Nord-Pas-de-Calais

Le premier exemple désigne la région autour du Havre et ne rentre pas dans notre étude; la seconde désigne la région administrative du Nord-Pas-de-Calais et rentre dans notre étude. Un premier moyen de les distinguer syntaxiquement est d'utiliser la phrase classificatrice :

** Le Havre est une région*
Le Nord-Pas-de-Calais est une région

Le nom *région* du premier cas pourrait être rattaché aux noms de localisation spatiale au sens d'A. Borillo (1989) (cf. section sur les prépositions locatives composées). En effet, l'expression *la région de* peut être remplacée par *les alentours de* où *alentour* est un nom de localisation externe :

Max a bien connu la région du Havre
?= Max a bien connu les alentours du Havre

Ainsi, si l'on pousse l'analyse plus loin, on peut considérer l'expression *dans la région de* comme une préposition locative dans la phrase suivante car elle peut également être remplacée par la préposition locative composée *près de* :

Marie habite (*dans la région de + dans les alentours de + près de) (la ville du + Le) Havre*

L'analyse précédente n'est clairement pas valable pour le deuxième cas :

Marie a bien connu la région du Limousin
≠ Marie a bien connu les alentours du Limousin

4.3.5 Réduction des formes longues et figement

Les formes longues des noms propres (*Nprc*) peuvent être réduites en formes courtes (*Npr*) par le processus suivant. Tout d'abord, la séquence *LE Adj* Nc Adj* (de)*[87] est effacée dans *Nprc*. Puis, s'il ne se trouve pas explicitement dans la forme longue, le déterminant *Det* est ajouté à gauche de cette nouvelle séquence. Ce déterminant peut être vide (*Det* = : *E*).

le département de les Pyrénées-Atlantiques
= les Pyrénées-Atlantiques (déterminant *les* explicite)

[87] La séquence *(de)* signifie que *de* est optionnel. Le symbole * est le symbole de Kleene : *a** désigne l'expression régulière *(E + a + aa +aaa + …)*.

la ville de Paris
= Paris

la république de Hongrie
*= *Hongrie*
= la Hongrie

le Mont Ventoux
*= *Ventoux*
= le Ventoux

? le fleuve Seine
*= *Seine*
= la Seine

Les jugements d'acceptabilité dans les exemples ci-dessus se réfèrent à des phrases du type

Ceci est Nprc
=: Ceci est (la ville de Paris + Paris)

Les déterminants définis ne sont pas les seuls à être autorisés avec les formes courtes. On trouve également des déterminants possessifs avec un sens affectif très expressif :

sa république du Mali => son Mali
sa mer Méditerranée => sa Méditerranée
son mont Ventoux => son Ventoux
son fleuve Seine => sa Seine

Les noms propres *Npr* ayant le déterminant *Det* vide acceptent sans difficulté particulière les déterminants possessifs (K. Jonasson, 1995 ; D. Maurel et O. Piton, 1998) :

sa ville de Paris => son Paris
son île de Tahiti => son Tahiti

De même, ces séquences acceptent des modifieurs : *sa belle ville de Paris* devient *son beau Paris*. Ces variantes sont très connues (voir K. Jonasson, 1995).

La réduction des formes longues en formes courtes n'est pas régulière. Il arrive que la forme courte de certains couples soit différente de *Npr*. Par exemple, la forme courte de *république démocratique populaire de Corée* est *Corée du Nord* et non *Corée*. Ce phénomène est limité car il ne touche quasiment que les noms de pays. Ensuite et surtout, certaines formes longues très figées ne peuvent être réduites. Nous donnons ci-dessous quelques d'exemples :

*(la mer de + *E) le Nord*
*(la mer de + *la) Corée*
*(la mer + *la) Noire*
(*le pic de + *E) le Midi*
*(l'île de + *E) la Tortue*

Notons que, dans certains textes, on trouve le classifieur de certaines formes longues figées,

avec la première lettre en majuscule : *la Vallée de la Mort*.

Le figement peut servir notamment à distinguer des emplois ambigus de *Npr*. Par exemple, *Nord* est soit un département soit une mer. Ces deux entrées se distinguent non seulement par leur classifieur (donc ils sont dans deux tables différentes), mais aussi par leur comportement syntaxique. En effet, *la mer du Nord* n'a pas de forme courte alors que *le département du Nord* est clairement réductible à la forme courte *le Nord*.

Nous nous attachons maintenant à montrer les différents degrés de figement dans *Nprc*. D'abord, le figement peut être distributionnel comme pour le nom *ville*. Tous les noms de villes entrent dans la construction nominale suivante :

La ville de Npr =: la ville de (Paris + Le Havre)

Cette analyse est possible, même si certains noms de villes possèdent des déterminants (ex : *Le Havre*, *La Havane*, *Les Saisies*) ; ces derniers ayant souvent leur première lettre en majuscule, l'ensemble peut être vu comme un *Npr*. Ainsi, la reconnaissance de la forme longue composée requiert seulement une bonne grammaire de *Npr* (cf. applications). On retrouve aussi ces déterminants écrits sans majuscule. Lorsque le déterminant *Le* est précédé de la préposition *à* ou *de*, la séquence est contractée en *au* ou en *du* (équivalent avec *à le* ou *de le*) : *à Le Havre = au Havre = à le Havre*. Par conséquent, il est préférable de considérer le déterminant en dehors de *Npr*. Ainsi, nous avons la structure

La ville de (Det+E) Npr

Linguistiquement, le choix de considérer *Det* comme inclus ou non dans *Npr* est arbitraire ; ce choix peut être guidé par des considérations sur les applications.

Il est alors absolument nécessaire de regarder chaque nom de ville et lui assigner, quand cela est nécessaire, un déterminant. C'est ce qui a été réalisé dans *Prolintex*.

Pour certains classifieurs comme *état* (état américain), les couples (*Nc*, *Npr*) semblent obéir à quelques règles approximatives. Etant donné le déterminant *Det* utilisé, il est possible de produire la forme longue associée :

- si le déterminant *Det* est *le*, le couple (*Npr*, *Nc*) accepte la construction nominale

  l'état de Det Npr =: *l'état de le Texas*

- Si le déterminant *Det* est *la* ou s'il est vide, le couple forme un nom composé de la forme :

  l'état de Npr =: l'état de (New York + Californie)

- si le déterminant Det est l', les deux sont possibles :

  l'état de (E + l') Npr =: l'état de (E + l') Oregon

Comme précédemment, la génération de telles formes composées nécessite l'association systématique d'un déterminant défini à chaque nom d'état :

Californie => la
Orégon => l'
Texas => le
New-York => E

Ces règles sont approximatives : il existe des exceptions. Par exemple, la forme *état de le Vermont*, bien qu'il sélectionne le déterminant *le*, possède une autre variante : *l'état de Vermont*.
Par ailleurs, un classifieur comme *département* restreint la structure des noms composés à *LE Nc de Det Npr*. La seule variation vient du déterminant *Det* qui dépend de *Npr* :

le département de (le Nord + la Picardie + les Landes)

Il existe cependant une exception avec le *Territoire de Belfort* dont la forme composée paraît douteuse :

? Lundi dernier, j'ai visité le département du Territoire de Belfort*

4.3.6 Les noms propres composés dans les groupes nominaux

Nous examinons comment se comportent les noms propres composés locatifs dans les groupes nominaux. Nous regardons d'abord la détermination en distinguant les emplois singuliers des emplois pluriels, puis nous étudions la distribution des modifieurs.

4.3.6.1 Détermination : les emplois singuliers

Lorsqu'ils possèdent une forme courte, les couples ayant un emploi singulier entrent dans la construction

*(E + Det) Npr être (UN + *des) Nc*

*Paris est une ville / * Paris (sont + E) des villes*
*La Réunion est une île / * la Réunion (sont + E) des îles*

Dans le groupe nominal issu de la phrase, on notera Det_c le déterminant de *Nc* : Det_c *Nc Npr*, Det_c *Nc de Npr*, Det_c *Nc de Det Npr*, etc.

La distribution de Det_c dans les constructions nominales dérivées est complexe. Le déterminant le plus naturel est le déterminant défini LE (*le*, *la*) :

la ville de Pau m'a toujours fasciné
la région (E + de le) Nord-Pas-de-Calais attire beaucoup de touristes belges

Le déterminant indéfini UN et le déterminant démonstratif *ce* sont quant à eux interdits :

** Marie apprécie de revoir (une + cette) ville de Paris*
** Marie adore (un + cet) état de la Californie*

Les déterminants démonstratifs *un* et *ce* avec un modifieur sont acceptables :

*Max a pu observer une ville de Paris (*E + dévastée par les bombes)*

*J'apprécie de revoir cette (bonne vieille + ?*E) ville de Paris (E + que j'aime tant)*

Le démonstratif *ce* sans modifieur semble acceptable dans quelques rares cas mais produit l'impression d'une ellipse :

Marie me casse les pieds avec cette (?E + satanée) avenue des Champs-Elysées

Notons que le déterminant possessif est parfaitement acceptable mais avec une interprétation particulière :

Max aime parler de son île de la Guadeloupe
≠ Max aime parler de sa maison

Cette distribution est similaire pour les formes figées :

*Sophie aime parler de (la + sa + *cette + *une) mer (du Nord + Morte)*
Sophie a vu une mer Morte déchaînée par les vents tourbillonnants
Sophie déplore une triste mer du Nord

4.3.6.2 Détermination : les emplois pluriels

Il existe deux cas d'emplois pluriels. Tout d'abord, nous examinons celui où le couple (*Nc, Npr*) rentre dans la construction

(E + Det) Npr être (UN + des) Nc*
Les Shetland sont (une + des) îles*

Nous retrouvons alors la même distribution de Det_c (transposées au pluriel) dans les groupes nominaux :

*Max décrit (les + ?ses + ?*ces + *des) îles Canaries*
*Luc me casse les pieds avec (les + ses + ? ces + *des) îles Shetland*

Dans ces exemples, nous avons affaire à des regroupements d'îles qui s'identifient par leur emploi pluriel. Les noms composés tels que *les îles Marshall* fonctionnent de la même manière.

Le deuxième cas correspond à la coordination de couples (*Nc, Npr*). Le groupe nominal pluriel

les villes de Paris, (de + E) Lyon et (de + E) Marseille

est en fait la factorisation de trois groupes nominaux singuliers :

la ville de Paris ; la ville de Lyon ; la ville de Marseille

Par ailleurs, nous constatons l'apparition dans les textes de groupes nominaux pluriels comme

les villes Paris, Lyon et Marseille ont conclu un accord commercial

alors que l'emploi singulier est exclu :

? La ville Paris est candidate à l'organisation des JO*

Ces coordinations sont très naturelles dans les constructions nominales contenant la préposition *de* :

Les royaumes de Belgique et du Danemark ne sont pas très éloignés l'un de l'autre.
Les îles de Sardaigne et de Sicile attirent beaucoup de touristes.

Les coordinations de formes nominales sans préposition *de* sont aussi possible :

Max aime survoler les mers Méditerranée et Adriatique

Par contre, la coordination entre constructions de différentes structures est plus difficile :

? Max déteste les îles Maurice et de la Réunion*

La factorisation du classifieur dans une coordination de noms propres composés figés est acceptable, même si elle n'est pas très naturelle :

?Luc a déjà affronté les mers de Corée et du Nord
?Luc a déjà affronté les mers Morte et Noire
?Marie a traversé les vallées de la Mort et des Rois

Ces constructions sont plus naturelles avec des noms propres moins figés :

Marie a traversé les vallées de Chevreuse et de la Maurienne.

4.3.6.3 Les modifieurs autour de *Nc*

Nous examinons la distribution des modifieurs dans les groupes nominaux étudiés précédemment. Les groupes nominaux peuvent avoir les formes suivantes où *M1*, *M2* et *M3* sont trois positions de modifieurs :

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

Le département très montagneux des Pyrénées-Atlantiques
La grande gare Montparnasse

La position M2 dans le groupe nominal sans préposition *de* est interdite :

La mer très polluée qu'est la Méditerranée
la mer très polluée Méditerranée
la très polluée mer Méditerranée

La position M2 est stylistiquement réservée à des modifieurs très courts :

La ville où Marie a vécu toute son enfance de Venise
La ville inoubliable de Venise

On retrouve les modifieurs longs à la position M3 :

La ville de Venise, où Marie a vécu toute son enfance

Le cas des noms propres composés figés est quasiment similaire. La seule différence réside dans le fait que l'insertion d'un modifieur en position M2 est la plupart du temps totalement interdite :

*La mer (E + *agitée) du Nord attire beaucoup de touristes*
*Max escalade le pic (E + *dangereux) du Midi*

4.3.7 Codage des contraintes internes

4.3.7.1 Représentation sous la forme de tables

Dans les sections précédentes, nous avons montré un ensemble de contraintes au sein des noms propres composés. Nous les codons maintenant dans des tables syntaxiques où chaque ligne correspond à une entrée lexicale (i.e. un nom propre composé *Nprc*). Nous construisons une table pour chaque classe de *Nprc*, c'est-à-dire pour chaque classifieur. Soit un classifieur *X*, alors sa table syntaxique sera nommée **NNpr-X**. Par exemple, la table associée au classifieur *mer* est la table **NNpr-mer**. Lorsque le classifieur est ambigu comme *état*, nous précisons l'indice associé à cet emploi. Par exemple, nous aurons la table **NNpr-état-69** pour *état* au sens de « région administrative » (69 est l'indice associé à cet emploi). La table **Npr** est la table décrivant l'ensemble des noms propres de lieu n'ayant pas de forme longue et dont le classifieur ne peut être exactement précisé. Jusqu'à présent, nous avons entamé le codage d'une soixantaine de tables. Certaines comme les départements (français), les états américains, les provinces canadiennes ou les régions (françaises) sont complètes. Dans la suite, nous expliquons le contenu de quelques tables codées.
Avant tout, nous expliquons les intitulés des colonnes utilisés. Les noms ont été choisis de telle manière qu'ils soient explicites pour le lecteur. Les principaux intitulés sont synthétisés et expliqués dans le tableau ci-dessous :

Intitulé	**Propriété ou information lexicale**
Nc	Classifieur
Npr	Nom propre associé au classifieur
Detc pluriel	Le nom propre composé est au pluriel (ex : *les îles Marshall*)
Prep	Préposition
Det	Déterminant défini associé à *Npr*
Npr-a	Adjectif morphologiquement lié à *Npr*
Adj1, *Adj2*, *Adj3*	Adjectifs de (*Npr*, *Nc*)

Certains intitulés sont des structures dans lesquels les noms propres peuvent rentrer : *LE Nc de Npr*, *LE Nc Npr-a*, etc. Par ailleurs, à chaque entrée lexicale, nous associons un numéro dans la colonne intitulée *Index Npr*. De même, chaque classifieur possède un numéro unique (colonne *Index Nc*). Dans la suite, nous décrivons quelques tables qui sont ordonnées selon le degré de complexité. Nous donnons des extraits de chacune d'elles. Dans les extraits de tables que nous proposons, les classifieurs peuvent apparaître redondants mais leur présence rend la lecture plus facile.

4.3.7.2 Table Npr

La table *Npr* est la plus simple et elle contient les noms propres qui ne possèdent pas de forme longue et de classifieur clair tels que le *Canada*. Cette table comprend surtout des noms de pays.

A	B	C	D
Indcx Npr	Dct	Npr	Variantes Npr
8	<E>	Antigua-et-Barbuda	-
6	la+l'	Australie	-
9	la	Barbade	-
10	le	Belize	-
11	le	Burkina Faso	Burkina
12	le	Canada	-
28	la	Grande-Bretagne	-
14	la	Grenade	-
15	la+l'	Irlande	-
29	la+l'	Irlande de le Nord	-
16	la	Jamaïque	-
17	le	Japon	-
18	<E>	Kiribati	-
34	le	Labrador	-
19	la	Malaisie	-
20	la	Mongolie	-
31	le	Negara Brunei Darussalam	Brunéi Darussalam+Brunei
21	la	Nouvelle-Zélande	-
22	la	Papouasie-Nouvelle-Guinée	Papouasie
35	le	Pays basque	-
30	la	Roumanie	-
23	<E>	Sainte-Lucie	-
24	<E>	Saint-Vincent-et-les Grenadines	-
33	<E>	Terre-Neuve	-
25	le	Turkménistan	-
26	<E>	Tuvalu	-
27	la+l'	Ukraine	-

Table 12 : échantillon de la table Npr

4.3.7.3 Table NNPr-département

A	B	C	D	E	F	G	H	I
Index Npr	Index Nc	Nc	Prep	Det	Npr	LE Nc de Npr	Npr-a	LE Nc Npr-a
1	8	département	de	le+l'	Ain	-	-	-
2	8	département	de	le+l'	Aisne	-	-	-
3	8	département	de	le+l'	Allier	-	-	-
4	8	département	de	les	Alpes-de-Haute-Provence	-	-	-
5	8	département	de	les	Hautes-Alpes	-	-	-
6	8	département	de	les	Alpes-Maritimes	-	-	-
7	8	département	de	le+l'	Ardèche	+	ardéchois	+
8	8	département	de	les	Ardennes	-	ardennais	+
9	8	département	de	le+l'	Ariège	-	ariégeois	+
10	8	département	de	le+l'	Aube	-	aubois	+
11	8	département	de	la+l'	Aude	-	audois	+
12	8	département	de	le+l'	Aveyron	-	aveyronnais	+
13	8	département	de	les	Bouches-de le-Rhône	-	-	-
14	8	département	de	le	Calvados	-	-	-
15	8	département	de	le	Cantal	-	cantalien+cantalou	+
16	8	département	de	la	Charente	+	charentais	+
17	8	département	de	la	Charente-Maritime	+	-	-
18	8	département	de	le	Cher	-	-	-
19	8	département	de	la	Corrèze	+	corrézien	+
20	8	département	de	la	Corse-de le-Sud	+	-	-
21	8	département	de	la	Haute-Corse	+	-	-
22	8	département	de	la	Côte-de or	-	-	-

Table 13 : échantillon de la table NNpr-département

4.3.7.4 Table NNpr-mer

A	B	C	D	E	F	G	H	I	J	K	L	M
Index Npr	Index Nc	Nc	Modif M2 optionnel	Prep	Det (forme longue)	Npr	Variante Npr	LE Nc de Npr	LE Nc Npr	(Detc Nprc+Det Npr) être un Nc	Det (forme courte)	Det Npr
1	19	mer	-	-	-	Adriatique	-	-	+	+	l'+la	+
21	19	mer	+	de	les	Antilles	-	-	-	+	-	-
2	19	mer	-	-	-	Baltique	-	-	+	+	la	+
11	19	mer	+	de	-	Barentz	-	+	-	+	-	-
3	19	mer	-	-	-	Blanche	-	-	+	+	-	-
22	19	mer	+	de	les	Caraïbes	-	-	-	+	-	-
12	19	mer	+	de	-	Chine	-	+	-	+	-	-
13	19	mer	+	de	-	Chine Méridionale	Chine de le Sud	+	-	+	-	-
14	19	mer	+	de	-	Corée	-	+	-	+	-	-
15	19	mer	+	de	-	Crète	-	+	-	+	-	-
4	19	mer	-	-	-	Egée	-	-	+	+	-	-
23	19	mer	-	de	le+l'	Est	-	-	-	+	-	-
16	19	mer	-	de	-	Glace	-	+	-	-	-	-
5	19	mer	-	-	-	Intérieure	-	-	+	+	-	-
17	19	mer	+	de	-	Irlande	-	+	-	+	-	-
24	19	mer	+	de	le	Japon	-	-	-	+	-	-
6	19	mer	-	-	-	Jaune	-	-	+	+	-	-
18	19	mer	+	de	-	Java	-	+	-	+	-	-
19	19	mer	+	de	-	Kara	-	+	-	+	-	-
27	19	mer	-	-	-	Manche	-	-	-	+	la	+
7	19	mer	-	-	-	Méditerranée	-	-	+	+	la	+
8	19	mer	-	-	-	Morte	-	-	+	+	-	-
9	19	mer	-	-	-	Noire	-	-	+	+	-	-
25	19	mer	-	de	le	Nord	-	-	-	+	-	-

Table 14 : échantillon de la table NNpr-mer

Quelques remarques de lecture de la table :
Lorsque la case correspondant à Prep est -, cela signifie que la structure LE Nc Prep Det Npr est interdite. Il en est de même avec les éléments de la colonnes Det (forme longue).

4.3.7.5 Table *NNpr-île*

La table **NNpr-île** est un peu plus compliquée. La colonne *Detc pluriel* indique si le nom propre composé est au pluriel : *les îles Bahamas* / * *l'île Bahamas*. La colonne intitulée *LE archipel de Det Npr* indique si le nom propre composé a une variante de cette forme :

Les îles Bermudes
= l'archipel des Bermudes

Cette propriété est valable dans les cas où le classifieur est *île* et où le nom propre est obligatoirement au pluriel, ce qui n'est pas une surprise car un archipel est un ensemble de plusieurs îles.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Index Npr	Detc pluriel	Index Nc	Nc	Modif M2	Prep	Det	Npr	LE Nc de Npr	LE Nc Npr	LE archipel de Det Npr	Npr-a	LE Nc Npr-a	Det (forme courte)	Det Npr
1	+	17	île	-	-	-	Aléoutiennes	-	+	+	-	-	les	+
56	-	17	île	-	de	la	Ascension	-	-	+	-	-	la	-
2	+	17	île	-	-	-	Bahamas	-	+	+	bahamien	+	les	+
3	+	17	île	+	de	les	Baléares	-	+	+	-	-	les	+
57	-	17	île	+	de	la	Barbade	-	-	-	barbadien	+	la	+
34	-	17	île	-	de	-	Beauté	+	-	-	-	-	-	-
4	+	17	île	-	-	-	Bermudes	-	+	+	-	-	les	+
35	-	17	île	+	de	-	Bornéo	+	-	-	-	-	<E>	+
5	+	17	île	+	de	les	Canaries	-	+	+	-	-	les	+
58	+	17	île	+	de	le	Cap-Vert	-	-	-	cap-verdien	+	le	+
6	+	17	île	-	-	-	Caraïbes	-	+	+	-	-	les	+
7	+	17	île	-	-	-	Célèbes	-	+	+	-	-	les	+
36	-	17	île	+	de	-	Ceylan	+	-	-	-	-	<E>	+
37	-	17	île	+	de	-	Chypre	+	-	-	chypriote	+	<E>	+
38	-	17	île	+	de	-	Corse	+	-	-	corse	+	la	+
39	-	17	île	+	de	-	Crète	+	-	-	crétois	+	la	+
40	-	17	île	+	de	-	Cuba	+	-	-	cubain	+	<E>	+
59	-	17	île	+	de	le	Diable	-	-	-	-	-	le	-
41	-	17	île	+	de	-	Elbe	+	-	-	-	-	<E>	+
8	+	17	île	-	-	-	Féroé	-	+	-	-	-	les	+

Table 15 : échantillon de la table NNpr-île

4.3.7.6 Table *NNpr-république*

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Index Npr	Index Nc	Nc	Adj1	Prep	Det	Npr	LE Nc Adj1 de Det Npr	LE Nc Adj1 de Npr	LE Nc de Det Npr	LE Nc de Npr	Npr-a	LE Nc Npr-a	Det	Variante Det	Det Npr	Variante Npr
106	30	république	-	-	-	-	-	-	-	-	dominicaine	+	-	-	-	-
124	30	république	socialiste	de	le	:Vietnam	+	-	+	-	vietnamien	-	le	-	+	-
1	30	république	-	de	-	Afrique de le Sud	-	-	-	+	sud-africain	+	la+l'	-	+	-
2	30	république	-	de	-	Albanie	-	-	-	+	albanais	+	l'+la	-	+	-
107	30	république	fédérale	de	-	Allemagne	-	+	+	-	allemand	-	l'+la	-	+	-
4	30	république	-	de	-	Angola	-	-	-	+	angolais	+	le+l'	-	+	-
95	30	république	-	-	-	Argentine	-	-	-	-	argentin	+	la+l'	-	+	-
5	30	république	-	de	-	Arménie	-	-	-	+	arménien	+	la+l'	-	+	-
6	30	république	-	de	-	Autriche	-	-	-	+	autrichien	+	la+l'	-	+	-
3	30	république	-	de	-	Azerbaïdjan	-	-	-	+	-	-	le+l'	-	+	-
117	30	république	populaire	de	le	Bangladesh	+	-	+	-	-	-	le	-	+	-
53	30	république	-	de	le	Bélarus+Belarus	-	-	+	-	-	-	le	-	+	-
7	30	république	-	de	-	Belau	-	-	-	+	-	-	<E>	-	+	-
51	30	république	-	de	le	Bénin	-	-	+	-	-	-	le	-	+	-
52	30	république	-	de	le	Bhoutan	-	-	+	-	-	-	le	-	+	-
8	30	république	-	de	-	Biélorussie	-	-	-	+	biélorusse	+	la	-	+	-
9	30	république	-	de	-	Bolivie	-	-	-	+	bolivien	+	la	-	+	-
10	30	république	-	de	-	Bosnie-Herzégovine	-	-	-	+	bosniaque	-	la	-	+	-
54	30	république	-	de	le	Botswana	-	-	+	-	-	-	le	-	+	-
118	30	république	fédérative	de	le	Brésil	+	-	+	-	brésilien	-	le	-	+	-
11	30	république	-	de	-	Bulgarie	-	-	-	+	bulgare	+	la	-	+	-
55	30	république	-	de	le	Burundi	-	-	+	-	-	-	le	-	+	-
56	30	république	-	de	le	Cameroun	-	-	+	-	camerounais	-	le	-	+	-
57	30	république	-	de	le	Cap-Vert	-	-	+	-	cap-verdien	-	le	-	+	-
102	30	république	-	-	-	Centrafrique	-	-	-	-	centrafricaine	+	le	-	+	-
58	30	république	-	de	le	Chili	-	-	+	-	chilien	-	le	-	+	-
108	30	république	populaire	de	-	Chine	-	+	+	-	chinoise	-	la	-	+	-
12	30	république	-	de	-	Chypre	-	-	-	+	chypriote	-	<E>	-	+	-
13	30	république	-	de	-	Colombie	-	-	-	+	colombien	+	la	-	+	-
59	30	république	-	de	le	Congo	-	-	+	-	congolais	-	le	-	+	-
119	30	république	démocratique	de	le	Congo	+	-	+	-	-	-	le	-	+	Zaïre
14	30	république	-	de	-	Corée	-	-	-	+	sud-coréen	-	la	-	-	Corée du Sud

Table 16 : échantillon de la table NNpr-république

On peut noter que *:Vietnam* fait appel à un graphe du même nom qui décrit toutes les variantes orthographiques.

4.3.7.7 Vers une reconnaissance automatique de groupes nominaux

Le lexique accumulé jusqu'à présent est de taille modeste : au total, nous avons codé manuellement environ 650 entrées, réparties en 50 tables. Nous avons d'abord repris les résultats de O. Piton et al. (1997) pour les noms de pays et constitué une liste des noms officiels des pays à l'aide de la liste diffusée par la Délégation Française à la langue française[88]. Le travail réalisé sur les noms d'îles par M. Garrigues (1995) nous a permis de répertorier les contraintes syntaxiques auxquelles ils sont soumis. Nous avons également listé tous les départements français, les régions françaises, les états américains et les provinces canadiennes. Par ailleurs, nous avons complété nos listes à l'aide de savants dosages d'autres types de lieu géophysiques : *mer*, *pic*, *vallée*, etc. Le lexique accumulé est donc voué à être largement complété. Cependant, d'un point de vue linguistique, ce petit ensemble nous a permis de mettre en lumière un vaste ensemble de phénomènes linguistiques facilement représentables dans des tables syntaxiques. Il est probable, dans le futur, que d'autres contraintes soient découvertes et ajoutées aux tables existantes. En fait, cette étude avait surtout pour but d'ouvrir une voie dans le traitement linguistique des noms propres de lieu et de montrer que la méthodologie du lexique-grammaire pouvait être appliquée avec succès à leur analyse syntaxique.

Un premier moyen de compléter automatiquement nos tables de manière spectaculaire, serait d'utiliser comme ressource le dictionnaire *Prolintex* qui contient pour certains types de noms

[88] http://www.culture.fr/culture/dglf/ressources/pays/FRANCAIS.HTM

de lieu (les villes par exemple) toutes les informations syntaxiques nécessaires. En effet, les noms de ville ont un comportement syntaxique stable : ils ont tous pour forme *la ville de Det Npr* (avec *Det = le, la, les, E*). La seule information nécessaire pour les coder est de connaître le déterminant associé au nom propre et cette information est codée dans *Prolintex* comme le montrent les les exemples ci-dessous :

Laval,Laval.N+PR+Top+Ville+DetZ:ms:fs
Pirée,Pirée.N+PR+Top+Ville+DetLe:ms

A chaque nom de ville est associé un certain nombre d'informations : la catégorie grammaticale (*N*), le type (nom propre : *+PR* ; toponyme : *+Top*), la classe locative (*+Ville*), un trait syntaxique (déterminant *+DetZ*, *+DetLe*, etc.) et des informations flexionnelles (*ms* pour masculin singulier). Pour ajouter dans nos tables les noms de villes, il suffit d'extraire dans *Prolintex* les entrées candidates : par exemple, à l'aide de l'expression régulière :

N\+PR\+Top\+Ville\+(DetZ+DetLe+DetLa+DetLes+DetL)[89]

Puis, connaissant le format de la table des noms de ville, on ajoute l'entrée en complétant automatiquement ses colonnes. Ce travail est difficilement réalisable par un linguiste car il faut des notions informatiques.

La première application du codage de nos tables consiste à reconnaître des groupes nominaux ayant pour tête un nom propre composé de lieu géographique. Il suffit de convertir les tables en graphes de la même manière qu'avec les expressions de mesure. Pour chaque table, il convient de construire un graphe patron (ou graphe paramétré). Mais, étant donné le nombre important de tables, une telle démarche serait fastidieuse. Une solution plus pratique est de construire automatiquement l'ensemble des graphes patrons à partir d'une table générique (ou méta-table) et d'un graphe générique (ou un méta-graphe patron), comme l'a fait S. Paumier (2003) pour transformer les tables syntaxiques des verbes du français en graphes. Dans la table générique, chaque ligne correspond à une des tables syntaxiques. Les colonnes correspondent à l'ensemble des propriétés syntaxiques répertoriées dans notre étude. Etant donné une ligne (soit une table), chaque colonne indique si la propriété associée à celle-ci est codée dans la table. Si c'est le cas, il suffit d'indiquer dans quelle colonne de la table sélectionnée se trouve la propriété : par exemple, si la propriété se trouve dans la colonne *D*, on inscrit *@D* dans la colonne correspondant à cette propriété dans la méta-table. Si cette propriété n'est pas codée dans la table, cela signifie que pour toutes les entrées de la table, cette propriété est implicitement vraie ou fausse. Si la propriété est vrai, on insère le signe +, sinon on insère le signe -. La méta-table contient donc des booléens ou des indices de colonnes des entrées (des tables).
Nous avons répertorié l'ensemble des propriétés. Puis pour chacune des tables, nous avons établi la correspondance entre ses propres propriétés et l'ensemble des propriétés de la méta-table. Nous donnons un extraît d'une méta-table de notre ensemble de tables de type *NNpr* (**Npr-île**, **Npr-département**, etc.). Nous n'avons pas répertorié toutes les propriétés afin d'obtenir une figure plus claire. Par exemple, la table **NNpr-republique** est décrite à la cinquième ligne. Si on la lit de gauche à droite, cette ligne indique que les formes longues sont implicitement au singulier (et pas au pluriel). Le classifieur *Nc* est codé dans la colonne C de **NNpr-republique**. Implicitement, on ne peut pas insérer de modifieur en position M2 (**la république surpeuplée de Croatie*), etc.

[89] Le symbole \+ reconnaît le caractère + ; le symbole + est l'opérateur d'union des expressions régulières ; le symbole * est l'opérateur de Kleene.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Table	Dctc singulier	Dctc Pluriel	Nc	Modif M2	Adj1	Prep	Dct (forme longue)	Npr	Variante Npr (forme longue)	LE Nc Adj1 de Dct Npr	LE Nc Adj1 de Npr	LE Nc Adj1 Npr	LE Nc de Dct Npr	LE Nc de Npr	LE Nc Npr	Npr-a	LE Nc Npr-a	(Dctc Nprc+Dct Npr) être un Nc	LE archipel de Dct Npr	Dct (forme courte)	Variante Dct (forme courte)	Dct Npr	Variante Npr (forme courte)
NNpr-ile	I@B	@B	@D	@E	-	@F	@G	@H	-	-	-	-	+	@I	@J	@L	@M	+	@K	@N	-	@O	-
NNpr-departement	+	-	@C	+	-	@D	@E	@F	-	-	-	-	+	@G	-	@H	@I	+	-	@E	-	@J	-
NNpr-mer	+	-	@C	@D	-	@E	@F	@G	@H	-	-	-	+	@I	@J	-	-	@K	-	@L	-	@M	-
NNpr-ocean	+	-	@C	-	@D	-	-	@E	-	-	-	+	-	-	+	-	-	+	-	@F	-	@G	-
NNpr-republique	+	-	@C	-	@D	@E	@F	@G	-	@H	@I	-	@J	@K	-	@L	@M	+	-	@N	@O	@P	@Q
NNpr-region	+	-	@C	+	-	@D	@E	@F	-	-	-	-	+	@G	@H	@I	@J	+	-	@K	-	@L	-
NNpr-lac	+	-	@C	+	-	@D	@E	@F	-	-	-	-	+	@G	@H	-	-	+	-	@I	-	@J	-

Table 17 : méta-table NNpr

Nous construisons ensuite le graphe générique qui est le graphe paramétré de toutes les tables, avec toutes les propriétés décrites dans les colonnes B à X de la méta-table. Nous donnons ce graphe ci-dessous. Chaque variable *@i* où *i* est l'identifiant d'une colonne de la méta-table fait référence à la propriété correspondante. L'application du méta-graphe patron à la méta-table génère les graphes patrons de nos tables de type NNpr. Pour chaque ligne, un graphe dont le nom est l'élément de la colonne A, est généré. On utilise la méthode de conversion montrée dans le chapitre sur les expressions de mesure. Cette méthode élimine les chemins représentant des propriétés non acceptées par l'entrée (signe -), garde ceux qui décrivent des propriétés autorisées et ajoute les informations que l'on utilise dans les propriétés. Par exemple, pour la première ligne, le graphe patron associé à la table **NNpr-île** sera automatiquement généré en élaguant et en ajoutant les informations nécessaires à une copie du méta-graphe patron. La variable *@I* sera remplacée par l'information '*@H*' qui, dans le graphe patron associé à **NNpr-île** identifie la colonne *Npr* de **NNpr-île**. La variable @L (qui symbolise la structure *LE Nc Adj de Npr*) sera supprimée du graphe patron. @N (qui symbolise la structure *LE Nc de Npr*) sera, quant à elle, remplacée par l'étiquette vide, ce qui autorisera automatiquement l'utilisation de la structure associée à cette variable. Pour cette entrée, nous obtenons le graphe patron ci-dessous (sous le méta-graphe patron) :

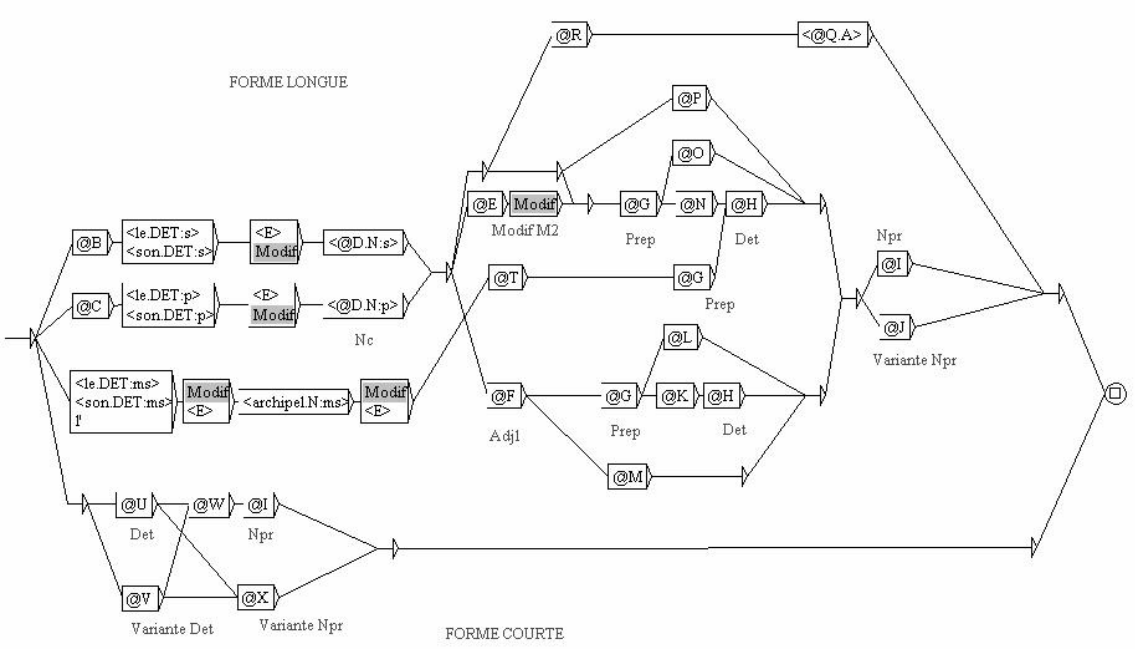

Figure 76 : méta graphe patron

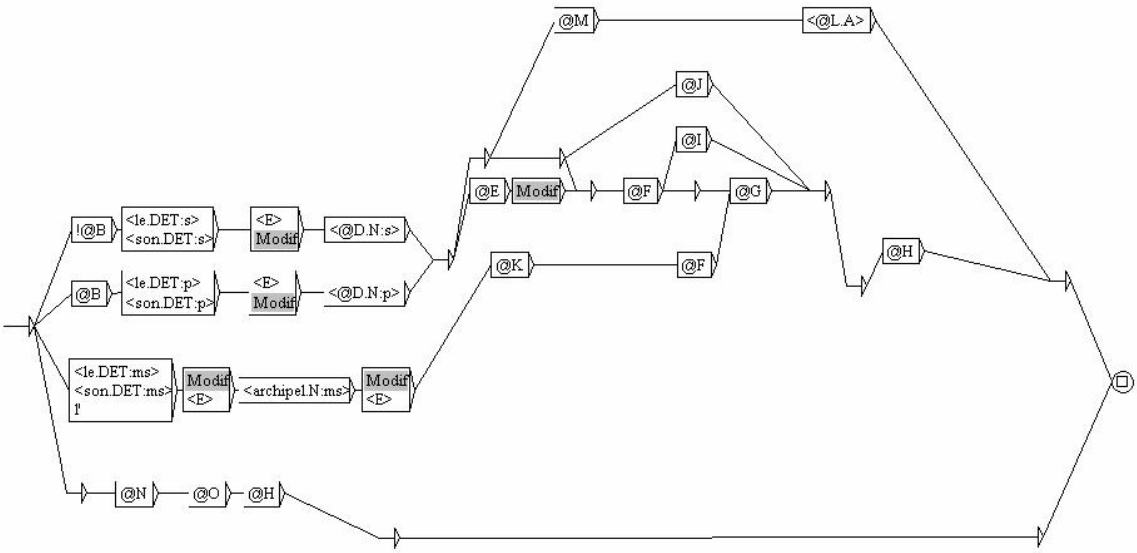

Figure 77 : graphe patron de la table NNpr-île

Le graphe fait référence aux colonnes par l'intermédiaire des variables *@B*, *@D*, *@E*, … correspondant respectivement aux colonnes *B*, *D*, *E*, … de la table **NNpr-île**. Le symbole *!@B* est la négation logique de *@B*. Ce symbole (utilisé dans Unitex) permet de décrire de manière simple une instruction du type *if … then … else…*. En effet, la colonne *B* indique si le classifieur est au pluriel (signe +) ou au singulier (signe -). Ainsi, le graphe patron doit contenir les deux possibilités. Les symboles (*@B* et *!@B*) permettent de sélectionner une des deux. L'étiquette *<@D.N:s>* indique que le classifieur sous forme lemmatisée, symbolisé par *@D* est un nom au singulier. *<@D.N:p>* indique qu'il est au pluriel. Le sous-graphe **Modif** est un graphe décrivant un modifieur adjectival quelconque, équivalent à l'expression régulière (<*ADV*> + <*E*>) <*A*> avec <*A*> adjectif et <*ADV*>

adverbe. En résolvant les références du graphe patron ci-dessus à la table **NNpr-île** pour l'entrée *île de Bornéo*, on obtient par exemple, le graphe ci-dessous.

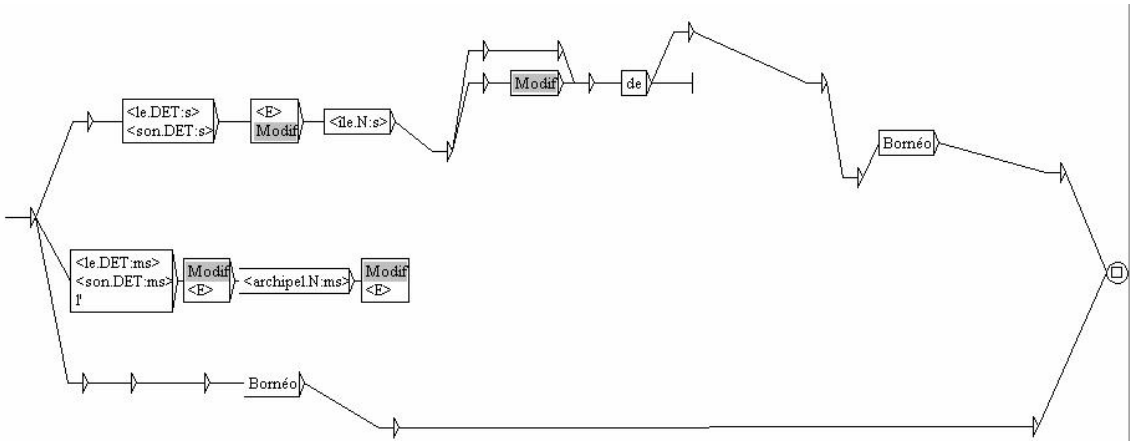

Figure 78 : entrée *île de Bornéo*

A partir de chacun des graphes patrons générés, nous générons les graphes associées à leurs entrées lexicales. Après union de ces graphes, nous obtenons un automate fini. Cet automate est en quelque sorte un mini-dictionnaire syntaxique de groupes nominaux simples. Nous avons confronté nos résultats à des textes en appliquant cet automate à une année du journal *Le Monde* (1994). Bien que nous ne puissions obtenir des résultats satisfaisant du fait de la non-exhaustivité de nos lexiques, ce test permet surtout de mettre en évidence les erreurs de codage : par exemple, une restriction trop forte ou simplement un oubli ou une erreur d'étourderie, etc. Par ce moyen, nous avons pu corriger un certain nombre de fautes.

Cette méthode de reconnaissance des groupes nominaux par transformation des tables en graphes pose des problèmes pour certains types de phénomènes linguistiques comme les coordonnées :

les villes de Paris et de Lyon
les états de Californie et du Texas

Ces expressions ne peuvent pas être représentées à l'aide de notre méthode car elles mettent en jeu différentes entrées lexicales dans un même graphe (supposé correspondre à une seule entrée). Ce problème est résolu si l'on change de stratégie en agissant en deux étapes :

(a) on transforme les tables en dictionnaire (type *Prolintex*)
(b) on construit des grammaires à l'aide de graphes utilisant les informations codées dans le dictionnaire

Par exemple, si l'on veut construire des grammaires reconnaissant des groupes nominaux avec des noms de villes coordonnés, il suffit de construire le graphe ci-dessous où l'on combine chaque classe de noms de villes qui prend un déterminant donné *Det* (ex : <N+Top+Pvil+DetLa>) avec ce même déterminant *Det* (ex : *la*). Il est évident que les noms de villes ont un graphe simple du fait de leur comportement syntaxique homogène. Pour les autres classifieurs, les grammaires construites deviennent rapidement complexes.

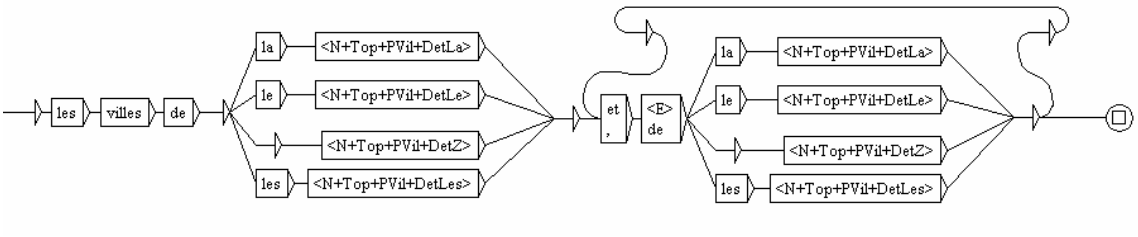

Figure 79 : coordination de villes

La difficulté de cette approche réside dans l'étape (a) que nous n'avons pas implémentée. Pour chaque ligne d'une table, il faut générer une entrée de dictionnaire contenant les informations suivantes : une forme fléchie, forme canonique, catégorie, traits syntaxiques et sémantiques [nom propre, locatif, classe de lieu, déterminants, etc.], informations flexionnelles. Les formes fléchies et canoniques correspondent à *Npr*, il suffit donc d'indiquer dans quelle colonne il est codé. La catégorie est clairement un nom (*N*). La difficulté réside dans l'extraction automatique des traits syntaxiques et lexicaux des tables pour les insérer dans l'entrée du dictionnaire. Il faut d'abord associer un nom compact, cohérent et unique à chacun de ces traits, puis trouver des procédures d'extraction. Comme ce type de travail n'est pas facilement accessible aux linguistes, il est nécessaire de prédéfinir des procédures d'extraction et donc de normaliser le codage des tables, ce qui n'entre pas dans le cadre de notre travail. Par ailleurs, pour implémenter cette approche, il faudrait aussi coder des informations flexionnelles dans les tables. Le gros inconvénient de cette approche est que les entrées du dictionnaire risquent de devenir illisibles du fait du grand nombre de propriétés. Par ailleurs, le phénomène de la coordination est général. Il existe le même problème avec les verbes par exemple :

Marie chante et danse.

Comme la représentation par grammaires des phrases simples au sein du lexique-grammaire est toujours en cours et loin d'être achevée, le traitement des coordonnées n'est pas d'actualité.

4.4 Description de groupes prépositionnels

4.4.1 Formes longues et variation prépositionnelle

Nous étudions la distribution prépositionnelle des noms propres composés *Nprc* = (*Npr*,*Nc*) qui entrent dans la construction à verbe support :

(Que P + N0) Vsup Loc Detc Nprc

Nous limitons notre ensemble des verbes supports au verbe *être*. Le verbe support *être* est un verbe neutre statique. Parfois, pour améliorer certaines acceptabilités, nous utilisons à la place les verbes *se passer*, *se dérouler* ou *se trouver* ayant un sens plus marqué. De même, nous limitons l'ensemble des prépositions *Loc* aux prépositions *à*, *dans*, *en* et *E* (préposition vide). Plus exactement, nous regardons les constructions suivantes :

(Que P + N0) être ((à + dans) le + en + E) Nprc

Nous ne regardons pas la préposition *sur* car son interprétation sémantique dépend trop du classifieur utilisé et du contexte.

Contrairement aux groupes nominaux, le comportement syntaxique de ces constructions ne dépend pas des deux éléments du couple (*Npr, Nc*), mais uniquement du classifieur *Nc*. Cela ne parait pas illogique car la nature et la forme du lieu sont explicitées par *Nc*. Par contre, là encore, un examen systématique est nécessaire car le comportement distributionnel des classifieurs est difficilement prédictible :

*Max est (dans l' + *à l + *en + *E) état de (le Texas + la Californie + Oregon)*
*Luc est (dans la + ?à la + en + *E) mer (du Nord + Noire+ Méditerranée)*
*Marie est (dans la + *à la + ?* en + *E) ville de (Paris + Le Havre + La Havane)*
*Léa est (dans la + à la + *en + E) rue (de la Paix + Monge)*
*Le randonneur est (?dans le + à le + *en + *E) pic (du Midi + du Vignemale)*

Il existe pourtant un cas particulier avec le classifieur *île* (M. Garrigues, 1995). En effet, l'emploi des prépositions *à et dans* n'est pas très naturel avec (*Corse, île*) et (*Sardaigne, île*) alors qu'il l'est avec les autres couples :

? Max est (à + dans) l'île de (Corse + Sardaigne)
Max est (à + dans) l'île de (Crête + Ré + la Martinique + la Guadeloupe)
Max est (à + dans) l'île Maurice

Cependant, l'interdiction n'est pas nette et nous considérons tous ces cas acceptables.

Le nom *région* a un comportement très spécifique. En effet, la préposition *en* ne peut s'employer avec toutes les structures nominales. Les structures nominales comprenant la préposition *de* ne peuvent pas se combiner avec cette préposition locative, alors qu'elles autorisent la préposition *dans* :

Léa est en région (Nord-Pas-de-Calais + Midi-Pyrénées + Ile-de-France)
** Léa est en région de (le Nord-Pas-de-Calais + le Midi-Pyrénées + Ile-de-France)*

Cette restriction n'est pas valable pour les autres classifieurs acceptant la préposition *en* :

Luc est en vallée d'Aspe
Luc est en mer de Corée

Concrètement, nous avons extrait de nos tables de noms propres tous les classifeurs : nous en comptons environ 70[90] en ajoutant quelques variations lexicales[91]. Pour chacun d'entre eux, nous avons systématiquement regardé et codé sa distribution prépositionnelle dans la table syntaxique **PNNpr** dont nous donnons un échantillon ci-dessous :

[90] A terme, le nombre de classifieurs va augmenter au fur et à mesure que le nombre de noms propres augmentera.
[91] Par exemple, la classe des villes est aussi sélectionnée par les classifieurs *village, commune, station balnéaire*, etc.

Loc Detc Nprc							
Loc = :à	Loc =: dans	Loc =:sur	Loc =: en	Loc =:E	Nc	Index Nc	en Nc...de ...
+	+	+	-	-	aiguille	1	-
-	+	+	-	-	archipel	2	-
+	+	+	-	+	avenue	3	-
-	+	+	+	-	baie	43	+
+	-	+	-	+	boulevard	68	-
-	+	+	-	-	bourg	44	-
-	+	+	-	-	bourgade	67	-
+	-	+	-	-	butte	45	-
-	+	+	-	-	canton	46	-
+	+	+	-	-	cité	47	-
-	+	+	-	-	cité	48	-
+	+	+	-	-	col	4	-
-	-	+	-	-	colline	49	-
+	+	+	-	-	Commonwealth	5	-
-	+	+	-	-	commune	50	-
-	+	+	-	-	comté	51	-
-	+	+	-	-	confédération	6	-
-	+	+	-	-	Cordillère	7	-
+	+	+	-	-	côte	52	-
-	-	+	-	-	côte	53	-
-	+	+	-	-	département	8	-
-	+	+	-	-	département d'outre-mer	71	-
-	+	+	-	-	désert	9	-
+	+	+	-	-	émirat	10	-
-	+	+	-	-	étang	54	-
-	+	+	-	-	état	11	-
-	+	+	-	-	état	69	-
+	+	+	-	+	États-unis	12	-
-	-	+	-	-	étoile	13	-
-	+	+	-	-	fédération	14	-
-	+	+	-	-	fleuve	15	-
-	+	+	-	-	gave	55	-
-	+	+	-	-	ghetto	56	-
+	+	+	-	-	glacier	57	-
-	+	+	-	-	golfe	58	-
+	+	+	-	-	grand-duché	16	-
+	+	+	-	-	île	17	-
+	+	+	-	-	impasse	59	-
+	+	+	-	-	lac	18	-

Table 18 : échantillon de la table PNNpr

On remarquera que *Commonwealth* et *Etats-Unis* sont des classifieurs car on a :

Le Commonwealth de (les Bahamas + la Dominique)
Les Etats-Unis de (Amérique + le Mexique)

La préposition *à*[92] pose souvent des problèmes d'acceptabilité car elle est très fréquemment utilisée dans le discours populaire et il est difficile de trouver une limite entre ce qui est acceptable ou pas. Dans notre cas, les phrases suivantes sont peu naturelles :

? Le rendez-vous est à la rue (Monge + de la Paix)
? Le rendez-vous est à l'avenue des Champs-Élysées

[92] La préposition *à* est un gros sujet d'étude (cf. B. Lamiroy, 2002).

Nous sommes confrontés à un autre cas litigieux lorsque la préposition *à* est combinée avec le classifieur *côte* (au sens de pente). En effet, la phrase suivante parait peu naturelle sans contexte :

? Le cycliste est à la côte Saint-Martin

Elle l'est plus si la côte Saint-Martin est considérée comme un point parmi un itinéraire de difficultés pré-établies.

On peut noter, d'autre part, que la préposition *dans* est très peu naturelle avec le classifieur *avenue* ou *boulevard*, alors qu'elle l'est avec le nom *rue*, alors qu'ils appartiennent tous trois à la même notion sémantique.

? La course a lieu dans l'avenue (des Champs-Elysées + Montmarte)*
? La course a lieu dans le boulevard Haussmann*
La course a lieu dans la rue Monge

Ces classifieurs sélectionnent préférentiellement la préposition *sur* :

Luc est sur l'avenue (des Champs-Elysées + Montmarte)
Luc est sur le boulevard Haussmann

L'étude de la distribution prépositionnelle des classifieurs permet de distinguer clairement plusieurs emplois ambigus. D'abord, le nom *cité* au sens de « quartier » se combine naturellement avec la préposition *à* alors que ce n'est pas vrai pour le nom *cité* au sens de « ville » :

Max (habite + se trouve) à la Cité des (Ulis + Sapins) (« quartier »)
** Luc (habite + se trouve) à la cité (E + médiévale) de Carcassonne* (« ville »)

De même, le nom *côte* au sens « pente » accepte sans difficulté la préposition *dans*, ce qui n'est pas le cas du nom *côte* au sens « bord de mer ».

Max (habite + se trouve) dans la côte Saint Martin (« pente »)
** Max (habite + se trouve) dans la côte d'Azur* (« bord de mer »)

Remarque :
Dans le cas hypothétique où la distribution prépositionnelle dépendrait des deux éléments du couple (*Npr*, *Nc*), il suffit de décrire ces couples séparément des autres dans des tables décrivant à la fois le comportement interne et la distribution prépositionnelle de la forme longue.

4.4.2 Formes courtes et variation prépositionnelle

Pour l'instant, nous avons étudié le comportement distributionnel des formes longues des noms propres dans des groupes prépositionnels. Il est intéressant de le comparer avec celui de leurs formes courtes associées. Les formes courtes sont des formes réduites des formes longues. Il est donc logique que leur distributions prépositionnelles soient identiques comme dans :

La croisière se déroule (dans la + sur la + en) (mer + E) Méditerranée
Max est perdu dans (le désert de + E) le Sahara
La rencontre a lieu (dans + à + sur) les (îles + E) Maldives
Max est (?à + sur + ?dans) (l'avenue de + E) les Champs-Élysées
Le coureur cycliste se trouve (à + dans + sur) (le col de + E) le Tourmalet

Ce phénomène naturel et logique n'est pas une généralité. Il existe de nombreux exemples montrant une disparité dans les distributions entre les formes longues et leurs formes courtes associées, comme le montrent les quelques exemples ci-dessous :

*Paul est à (*la ville de + E) Paris*
*Marie est à (l'île de + *la + *E) Crète*
Léa est en (**état de + E*) *Californie*
*Max est en (principauté de + *E) Monaco*
Le président est (*avenue de les + *les + *E*) *Champs-Élysées*

Pour certains classifieurs comme *département, île, état, province*, etc., la séquence *dans la* a parfois tendance à se contracter dans la préposition *en*, mais seulement lorsque le classifieur est absent :

Luc est dans (le département de + E) la Meurthe-et-Moselle
Luc est en Meurthe-et-Moselle
** Luc est en département de la Meurthe-et-Moselle*

L'argent se trouve dans (l' île de + ?E) la Martinique
L'argent se trouve en Martinique
** L'argent se trouve en île de la Martinique*

Marie est dans (l'état de + ?la) Caroline du Nord
Marie est en Caroline du Nord
**Marie est en état de Caroline du Nord*

Mais, ce phénomène n'est pas une règle générale car, pour certains noms propres, la préposition *en* est interdite :

*L'argent se trouve dans (l' île de + ?*E) la Réunion*
** L'argent est en Réunion*

Luc est dans (le département de + ?E) la Côte d'Or
*?*Luc est en Côte D'or*

Cette contraction est également observée pour les noms propres *Npr* commençant par une voyelle :

Luc est dans (le département de + ?E) l'Isère
Luc est en Isère

Marie est dans (la province de + E) l' Alberta
Marie est en Alberta

Là encore, il existe des contre-exemples comme pour (*Yonne,département*) :

Luc est dans (le département de + E) l'Yonne
**Luc est en Yonne*

De même, du fait de la présence du nom *île* à l'intérieur de *Npr* la préposition *en* ne se combine pas avec le couple (*Ile-du-Prince-Edouard, province*).

Mon fils est retenu dans la province de l'Ile-du-Prince-Edouard
*Mon fils est retenu (à l'+*en) Ile-du-Prince-Edouard*

On notera, cependant, que *Ile-de-France* se combine avec la préposition *en* :

Mon fils est retenu en Ile-de-France

Par ailleurs, les noms propres *Npr* prenant pour déterminant *le* et ne commençant pas par une voyelle peuvent ne pas autoriser cette contraction :

Luc est dans (le département de + E) le Cher
**Luc est en Cher*

Léa est dans (l'état de + E) le Texas
**Léa est en Texas*

Cependant, ce n'est pas toujours vrai :

L'élection est dans (le département de + E) le (Loir-et-Cher + Val-de-Marne + Indre-et-Loire)
L'élection est en (Loir-et-Cher + Val-de-Marne + Indre-et-Loire)

Léa est dans (l'état de + E) le Wyoming
Léa est en Wyoming

Pour les noms de provinces canadiennes et d'états américains, on observe que la combinaison des formes courtes avec la préposition *à* est généralement possible lorsque le *Npr* prend *le* comme *Det* :

Cette insurrection populaire a eu lieu au (Texas + Wyoming) [*Nc =*: *état*]
Max se trouve au Québec [*Nc =*: *province*]

Pourtant, la combinaison entre *à* et *Vermont* (réduction d'*état du Vermont*), n'est pas acceptée :

** Cette insurrection populaire a eu lieu au Vermont* [*Nc =*: *état*]

Ainsi, bien que l'on constate certaines tendances générales pour chaque classifieur, ces quelques exemples montrent clairement que la distribution prépositionnelle des formes courtes des noms propres composés n'est pas toujours prédictible.

Les noms propres composés *Nprc* peuvent être divisés en deux catégories disjointes : ceux dont la distribution prépositionnelle des formes courtes dépend uniquement de leur classifieur (catégorie 1) et ceux dont la distribution prépositionnelle des formes courtes n'est pas prédictible à partir du classifieur (catégorie 2).
Nous reprenons chaque table de noms propres. Toutes les tables qui comportent des noms propres de la deuxième catégorie sont reprises. On y ajoute des colonnes représentant la distribution prépositionnelle de notre ensemble de prépositions locatives. C'est le cas pour la table **NNpr-île** pour laquelle nous ajoutons trois colonnes : la première correspondant à la préposition *en* (colonne P), la deuxième à la préposition *dans* (colonne Q) et la troisième à la préposition *à* (colonne R). En effet, le comportement prépositionnel des *Nprc* ayant le nom *île* pour classifieur est difficilement prédictible :

*Max est (*à la + en) (Crète + Corse)* (*Det =*: *la*)
*Max est (à la + *en) Réunion* (*Det =: la*)
Léa est (en + à) Haïti (*Det =*: *E*)
*Léa est (*en + à) (Bornéo + Maurice)* (*Det =*: *E*)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Index Npr	Detc pluriel	Index Nc	Nc	Modif M2	Prep	Det	Npr	LE Nc de Npr	LE Nc Npr	LE archipel de Det Npr	Npr-a	LE Nc Npr-a	Det (forme courte)	Det Npr	en Npr	dans Det Npr	à Det Npr
1	+	17	île	-	-	-	Aléoutiennes	-	+	+	-	-	les	+	-	+	+
56	-	17	île	-	de	la	Ascension	-	-	+	-	-	la	-	-	-	-
2	+	17	île	-	-	-	Bahamas	-	+	+	bahamien	+	les	+	-	+	+
3	+	17	île	+	de	les	Baléares	-	+	+	-	-	les	+	-	+	+
57	-	17	île	+	de	la	Barbade	-	-	-	barbadien	+	la	+	-	-	+
34	-	17	île	-	de	-	Beauté	+	-	-	-	-	-	-	-	-	-
4	+	17	île	-	-	-	Bermudes	-	+	+	-	-	les	+	-	+	+
35	-	17	île	+	de	-	Bornéo	+	-	-	-	-	<E>	+	-	-	+
5	+	17	île	+	de	les	Canaries	-	+	+	-	-	les	+	-	+	+
58	+	17	île	+	de	le	Cap-Vert	-	-	-	cap-verdien	+	le	+	-	-	+
6	+	17	île	-	-	-	Caraïbes	-	+	+	-	-	les	+	-	+	+
7	+	17	île	-	-	-	Célèbes	-	+	+	-	-	les	+	-	+	+
36	-	17	île	+	de	-	Ceylan	+	-	-	-	-	<E>	+	-	-	+
37	-	17	île	+	de	-	Chypre	+	-	-	chypriote	+	<E>	+	-	-	+
38	-	17	île	+	de	-	Corse	+	-	-	corse	+	la	+	+	-	-
39	-	17	île	+	de	-	Crète	+	-	-	crétois	+	la	+	+	-	-
40	-	17	île	+	de	-	Cuba	+	-	-	cubain	+	<E>	+	-	-	+
59	-	17	île	+	de	le	Diable	-	-	-	-	-	le	-	-	-	-
41	-	17	île	+	de	-	Elbe	+	-	-	-	-	<E>	+	-	-	+
8	+	17	île	-	-	-	Féroé	-	+	-	-	-	les	+	-	+	+
9	+	17	île	-	-	-	Fidji	-	+	-	fidjien	+	les	+	-	-	+
60	-	17	île	+	de	la	Grenade	-	-	-	grenadien	+	la	+	-	-	-
61	-	17	île	+	de	la	Guadeloupe	-	-	-	guadeloupée	+	la	+	+	-	+
42	-	17	île	+	de	-	Guernesey	+	-	-	-	-	<E>	+	-	-	+
43	-	17	île	+	de	-	Haïti	+	-	-	haïtien	-	<E>	+	+	-	+
10	+	17	île	-	-	-	Hawaï	-	+	-	-	-	<E>	+	-	-	+

Table 19 : reprise de NNpr-île

Pour la table **NNpr-république**, nous ajoutons même une quatrième colonne représentant la séquence *à Npr* (différente de *à Det Npr*). Cette dernière colonne est nécessaire pour coder les propriétés du nom propre *république du Panama* car ce nom apparaît dans deux types de groupes prépositionnels :

Max se trouve à (E + le) Panama

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Index Npr	Index Nc	Nc	Adj1	Prep	Det	Npr	LE Nc Adj1 de Det Npr	LE Nc Adj1 de Npr	LE Nc de Det Npr	LE Nc de Npr	Npr-a	LE Nc Npr-a	Det	Variante Det	Det Npr	Variante Npr	en Npr	dans Det Npr	à Det Npr	à Npr
106	30	république	-	-	-	-	-	-	-	-	dominicaine	+	-	-	-	-	-	-	-	-
124	30	république	socialiste	de	le	:Vietnam	+	-	+	-	vietnamien	-	le	-	+	-	-	-	+	-
1	30	république	-	de	-	Afrique de le Sud	-	-	-	+	sud-africain	+	la+l'	-	+	-	+	-	-	-
2	30	république	-	de	-	Albanie	-	-	-	+	albanais	+	l'+la	-	+	-	+	-	-	-
107	30	république	fédérale	de	-	Allemagne	-	+	+	-	allemand	-	l'+la	-	+	-	+	-	-	-
4	30	république	-	de	-	Angola	-	-	-	+	angolais	+	le+l'	-	+	-	+	-	-	-
95	30	république	-	-	-	Argentine	-	-	-	-	argentin	+	la+l'	-	+	-	+	-	-	-
5	30	république	-	de	-	Arménie	-	-	-	+	arménien	+	la+l'	-	+	-	+	-	-	-
6	30	république	-	de	-	Autriche	-	-	-	+	autrichien	+	la+l'	-	+	-	+	-	-	-
3	30	république	-	de	-	Azerbaïdjan	-	-	-	+	-	-	le+l'	-	+	-	+	-	-	-
117	30	république	populaire	de	le	Bangladesh	+	-	+	-	-	-	le	-	+	-	-	-	+	-
53	30	république	-	de	le	Bélarus+Belarus	-	-	+	-	-	-	le	-	+	-	-	+	+	-
7	30	république	-	de	-	Belau	-	-	-	+	-	-	<E>	-	+	-	-	-	+	-
51	30	république	-	de	le	Bénin	-	-	+	-	-	-	le	-	+	-	-	-	+	-
52	30	république	-	de	le	Bhoutan	-	-	+	-	-	-	le	-	+	-	-	-	+	-
8	30	république	-	de	-	Biélorussie	-	-	-	+	biélorusse	+	la	-	+	-	+	-	-	-
9	30	république	-	de	-	Bolivie	-	-	-	+	bolivien	+	la	-	+	-	+	-	-	-
10	30	république	-	de	-	Bosnie-Herzégovine	-	-	-	+	bosniaque	-	la	-	+	-	+	-	-	-
54	30	république	-	de	le	Botswana	-	-	+	-	-	-	le	-	+	-	-	-	+	-
118	30	république	fédérative	de	le	Brésil	+	-	+	-	brésilien	-	le	-	+	-	-	-	+	-
11	30	république	-	de	-	Bulgarie	-	-	-	+	bulgare	+	la	-	+	-	+	-	-	-
55	30	république	-	de	le	Burundi	-	-	+	-	-	-	le	-	+	-	-	-	+	-
56	30	république	-	de	le	Cameroun	-	-	+	-	camerounais	-	le	-	+	-	-	-	+	-
57	30	république	-	de	le	Cap-Vert	-	-	+	-	cap-verdien	-	le	-	+	-	-	-	+	-
102	30	république	-	-	-	Centrafrique	-	-	-	-	centrafricaine	+	le	-	+	-	-	+	+	-
58	30	république	-	de	le	Chili	-	-	+	-	chilien	-	le	-	+	-	-	-	+	-
108	30	république	populaire	de	-	Chine	-	+	+	-	chinoise	-	la	-	+	-	+	-	-	-
12	30	république	-	de	-	Chypre	-	-	-	+	chypriote	-	<E>	-	+	-	-	-	+	-
13	30	république	-	de	-	Colombie	-	-	-	+	colombien	+	la	-	+	-	+	-	-	-
59	30	république	-	de	le	Congo	-	-	+	-	congolais	-	le	-	+	-	-	-	+	-
119	30	république	démocratique	de	le	Congo	+	-	+	-	-	-	le	-	+	Zaïre	-	-	+	-
14	30	république	-	de	-	Corée	-	-	-	+	sud-coréen	-	la	-	-	Corée du Sud	+	-	-	-

Table 20 : reprise de NNpr-république

Les tables dont les noms propres font partie de la catégorie 1 ne sont pas modifiées. Les rares exceptions sont enlevées de la table et codées dans une table séparée. En l'état actuel des travaux, nous n'avons pas rencontré de tels cas. Nous ne codons pas dans la table **NNpr-ville** le fait que la préposition *à* ne soit acceptée qu'avec les formes courtes des noms de ville[93] : nous le ferons lors de la transformation des tables en graphes (cf. dernière section du chapitre).

4.4.3 Des adverbes figés locatifs

Les compléments prépositionnels examinés jusqu'à présent comprennent obligatoirement le constituant *Npr*. Mais que se passerait-il si l'on effaçait *Npr* ? Regardons la phrase ci-dessous :

> *Max est dans la vallée de la Maurienne*
> *= Max est dans la vallée*

Il peut s'agir d'un effacement avec coréférence :

> *Max habite dans la vallée de la Maurienne. Marie espère un jour lui rendre visite dans la vallée*

Cependant, intuitivement, *dans la vallée* peut aussi référer à une vallée non nommée, identifiable de façon unique grâce au contexte :

[93] Pour quelques noms de ville comme *Avignon* ou *Arles*, la présence de la préposition *en* est autorisée : *en Avignon*. Ce phénomène est difficilement explicable : peut-être le fait que ces villes étaient des états pontificaux ; certains parlent même de snobisme (notre source a souhaité garder l'anonymat).

Luc a skié toute la journée à Courchevel ; il va redescendre dans la vallée en voiture

Dans cette phrase, la vallée dont on parle est celle qui se trouve en contrebas de Courchevel. Ce phénomène est très courant comme le montrent les exemples ci-dessous :

Paul est dans la maison. Par la fenêtre, il voit que Paul est dans la rue

Devant cette difficulté d'interprétation sémantique, nous ne traitons pas ces cas et nous regardons plutôt les constructions figées locatives de la forme :

N0 être Loc Det Nc =: *Paul est en montagne*

Ces constructions ont été très peu étudiées dans le cadre des études sur les constructions *être Prep X* au sein du lexique-grammaire. La préposition *en* est corrélée à un fort degré de figement. Par exemple, elle ne peut pas se combiner avec les formes longues des noms propres de villes, alors que la séquence sans *Npr* est tout à fait naturelle :

Marie est en ville
*?*Marie est en ville de Paris*

Lorsque la forme longue accepte la préposition *en* comme pour les noms de mers, la séquence sans *Npr* a un sens plus générique :

Luc est en mer Méditerranée
Luc est en mer (sens générique)

Par ailleurs, la règle de « pseudo-effacement » du *Npr* est loin d'être régulière, c'est plutôt une exception, ce qui renforce le caractère figé de *en mer* :

Luc est en vallée d'Aspe
**Luc est en vallée*

Max est en baie des Anges
**Max est en baie*

La présence de la préposition *en* est naturelle devant le nom *région* (sans *Npr*), mais ce dernier est au pluriel :

Les ministres sont en régions

Dans cette dernière phrase, le nom *région* n'est pas au sens de région administrative (ex : *Aquitaine*) ou un nom de localisation autour d'un lieu (*la région de Lyon*) car la phrase signifie que les ministres se trouvent en province. D'ailleurs, si l'on prend le nom *province*, on observe un phénomène similaire :

Luc est en province

Cette phrase signifie que Luc est en France mais n'est pas à Paris. Le nom *province* n'a donc pas le sens d'une partie administrative d'un pays (*province du Nouveau-Brunswick*).

Notons que le nom *principauté* rentre dans un adverbe locatif utilisant la préposition *en :*

Luc est en principauté

Cette expression est une réduction de *en principauté de Monaco*.

Si l'on regarde le nom *cité* au sens de « quartier », il est relativement naturel de dire la phrase suivante :

Ma famille (habite + ?est) en cité
Où habite ma famille ? en cité

Par contre, on n'observe pas la même construction avec *cité* au sens de « ville » :

**Luc (habite + est) en cité*
Luc est en ville

Pour d'autres classifieurs, on observe également un figement avec la préposition *en*, mais l'emploi est non locatif car les phrases ne répondent pas à la question en *où* :

La course est en côte
Comment est la course ? en côte
*Où est la course ? *en côte*

Ce figement est quand même un moyen de différencier les deux emplois de *côte* car on ne peut avoir ce comportement avec le nom *côte* au sens de « bord de mer ».

Le classifieur *république* rentre aussi dans un adverbe figé (*en république*) qui n'a pas un sens locatif :

Nous sommes en république
*Où est-ce que nous sommes ? *en république*

Le nom *butte* entre dans une construction non locative semi figée du type *être Prep C Prep* où *C* est un nom figé avec la préposition :

Max est en butte à des problèmes
**Max est en butte*

Là encore, la construction ne répond pas à la question en *où* :

A quoi Max est-il en butte ? à des problèmes
*Où Max est-il en butte ? *à des problèmes*

Le figement est possible avec d'autres prépositions que *en*, comme *à*. Il existe par exemple plusieurs variantes figées ayant le même sens qui utilisent le classifieur *montagne* :

Marie est (en + à la) montagne

Nous décidons d'élargir le champ d'investigation à d'autres classifieurs de lieu tels que *campagne* qui est ambigu : *campagne* opposé à « ville » (sens locatif) et *campagne* comme « campagne électorale, publicitaire, etc. » (sens non locatif). Ces deux emplois se distinguent par leur distribution prépositionnelle : l'emploi locatif interdit la préposition *en* mais accepte la préposition *à* ; l'emploi non locatif ne se combine pas avec *être à* mais avec *être en*.

*Marie est (*en + à la) campagne*
*Luc est (en + *à la) campagne (E + électorale)*

Regardons maintenant le classifieur *territoire*. On observe, dans ce cas-là, un phénomène particulier : la préposition *en* n'est acceptée que si *territoire* est suivi d'un adjectif :

**Léa est en territoire*
Léa est en territoire (français + contaminé + ennemi)
**Léa est en territoire (de la France + qui appartient à la France)*

Par contre, l'utilisation de la préposition *sur* n'est pas contrainte même si certains cas sont limites :

Luc est sur le territoire (E + français + ?de la France + ?qui appartient à la France)

Le nom classifieur *banlieue* accepte aussi la présence de la préposition *en* au sein d'une expression figée :

Les troubles sont en banlieue

Ce nom peut aussi apparaître comme un nom de localisation spécialisé à la ville : dans la phrase ci-dessous, le sujet Luc est situé par rapport à *Paris*.

Luc est dans la banlieue (de Paris + parisienne)

L'emploi de la préposition *en* est autorisée même si elle est plus naturelle avec l'adjectif *parisienne* que le complément de nom *de Paris* :

Luc est en banlieue (?de Paris + parisienne)

Il est facile de faire le parallèle avec *région* dont nous avons montré précédemment qu'il pouvait être considéré comme un nom de localisation spatiale dans certains cas :

Luc est dans la région (de Lyon + lyonnaise)

L'emploi de la préposition *en* est difficile si le nom *région* n'est pas suivi d'un adjectif :

*Luc est en région (*de Paris + parisienne)*

Nous avons répertorié manuellement les adverbes figés locatifs. Ils pourraient très bien être intégrés à la table EPC de M. Gross, mais, pour une meilleure lisibilité du lecteur, nous les représentons dans un graphe car ils sont en petit nombre. Notons que nous avons inséré des expressions utilisant la préposition *dans* comme *dans les îles*, *dans la rue*, *dans le désert* ou *dans la région* car elles paraissent relativement figées.

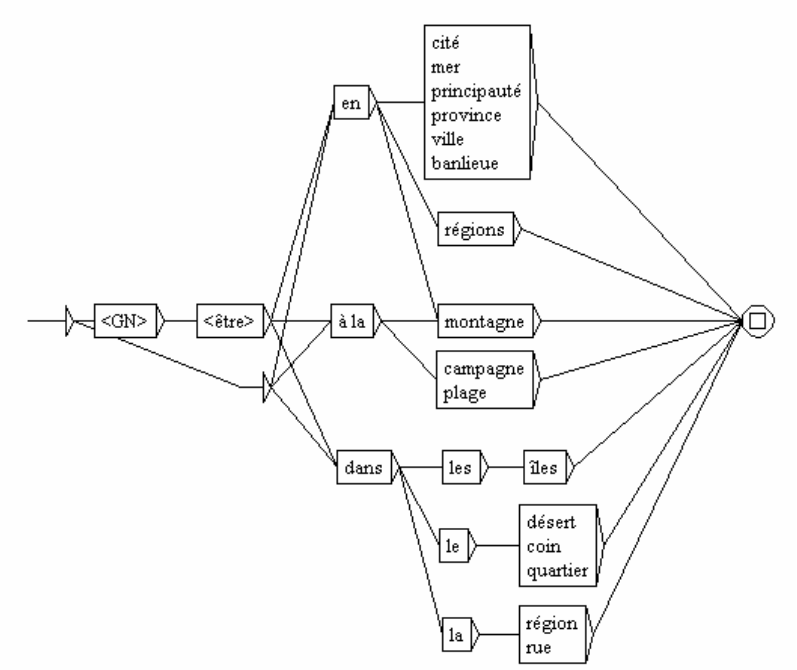

Figure 80 : EPC-locatif

Si l'on applique ce graphe (ou du moins la partie à droite du verbe *être*), on constate que le bruit est très important car le contexte d'analyse est très local. Nous donnons ci-dessous quelques exemples d'erreurs que l'on retrouve dans les textes :

- prépositions sélectionnées par le verbe de la phrase :

```
entreprises avaient contribué en 1993 à la campagne électorale.
party de adieu dans une cabine adossée à la montagne(…)
```

- analyse trop courte du groupe nominal

```
un séjour dans l'Ouest canadien, dans la région de Cariboo-Chilcotin,(…)
jeudi 17 mars à le petit matin, dans la région toulonnaise, ont été (…)
nombre de personnes inscrites à l'ANPE dans la région Ile-de-France.(…)
ne collision sur le Rhin, un naufrage en mer de le Nord et de les (…)
```

- autre :

```
les enseignants, à propos ?{S} De ville en ville et de chaîne en chaîne,(…)
```

4.5 Un nouveau système de conversion des tables en graphes

4.5.1 Préliminaires

4.5.1.1 Rappels sur la méthode standard de conversion

Nous rappelons brièvement la méthode standard de conversion. Comme nous l'avons déjà utilisée précédemment sous la forme d'exemples, nous la décrivons de manière formelle pour être cohérent avec la suite. L'objectif est de convertir une table en automates finis (usuellement appelés graphes dans la communauté RELEX) à l'aide d'un automate (ou graphe) de référence (ou automate paramétré). Formellement, un automate fini est un 4-uplet $<Q, I, F, \Sigma, \delta>$ où Q est un ensemble d'états, I est l'ensemble des états initiaux, F est l'ensemble des états finaux, Σ est l'alphabet et $\delta \subseteq Q \times \Sigma^* \times Q$ est l'ensemble des transitions. Pratiquement, les automates finis sont représentés à l'aide de graphes où les transitions sont des boites et les états ne sont pas représentés (sauf l'état initial et l'état final).

On suppose que Σ est un alphabet quelconque disjoint de l'alphabet des booléens ({+,-}). Soit M une table avec n colonnes qui contient des booléens et des éléments de Σ^*. Le graphe de référence associé à M est un automate fini dont l'alphabet est $\Sigma \cup \Delta$: Δ est un alphabet auxiliaire disjoint de Σ. Chaque élément de Δ est appelé variable et correspond à une colonne de M. Pratiquement, cet automate contient des variables $@j$ où l'entier $j \in [1, n]$[94] correspond à la colonne j de M. Le symbole *!* représente le NON logique et peut parfois être placé avant la variable.

Nous montrons maintenant l'utilisation de ce graphe de référence pour convertir la table M en graphes. Soit $T_i = < Q_i, I_i, F_i, \Sigma_i, \delta_i>$ une collection de graphes. T_0 est le graphe de référence (ou paramétré) associé à la table M. Pour tout $i > 0$, T_i sera le graphe associé à la ligne i de M et sera automatiquement construit à l'aide de l'algorithme suivant :

[94] Nous supposons ici que le symbole *j* est un entier pour des raisons de facilité. Nous avons vu dans les exemples que cela pouvait être un autre type d'identifiant : *A, B, …, Z, AA, AB,* etc.

```
pour chaque ligne i de M
        Ti ← T0.copie()
        pour chaque transition t = (q,a,q')∈ δi
                si a ∈ Δ
                // a = [ !]@j[95]
                        val ← M(i,j)
                        si a = !@j alors
                                si M(i,j) = + alors
                                        val ← -
                                finSi
                                si M(i,j) = - alors
                                        val ← +
                                finSi
                        finSi
                        si val = + alors
                                Ti.modifierEtiquette(t,ε)[96]
                        sinon
                                si val = - alors
                                        Ti.supprimerTransition(t)
                                sinon
                                        Ti.modifierEtiquette(t, val)
                                finSi
                        finSi
                finSi
        finPour
    finPour
```

Nous avons utilisé cette méthode pour convertir nos tables de groupes prépositionnels locatifs de manière approximative. L'approximation est dû au fait que nous ne tenons pas compte des tables des noms propres car nous réalisons une description générale de ces derniers dans le graphe-patron[97]. Le constituant *Npr* est décrit « graphiquement » dans le graphe **Npr** à l'aide des méta-étiquettes <*PRE*> et <*MOT*> qui désignent respectivement tout mot commençant par une majuscule et tout mot graphique (séquence de caractères). Le graphe-patron **PNNprApprox** représente l'ensemble des structures de la table **PNNpr** avec toutes les prépositions. Il est paramétré par les variables *@B*, *@C*, @D, @E qui correspondent respectivement aux prépositions *à* (colonne B), *dans* (colonne C), *en* (colonne D) et *E* (colonne E) et qui prennent une valeur lorsqu'on choisit une entrée de la table. La variable *@F* désigne le classifieur (colonne F). Pour chaque entrée de **PNNpr**, on peut alors construire automatiquement une version du graphe reconnaissant les groupes prépositionnels locatifs dans lesquels elle peut rentrer. Pour l'entrée *île* par exemple, les variables *@B* et *@C* sont remplacées par le mot vide <*E*>. Les boîtes contenant *@D* et *@E* sont supprimées. Enfin, *@F* est remplacée par l'entrée lexicale *île* se trouvant à la colonne F.

[95] [!] signifie que le symbole ! est optionnel.
[96] ε est le symbole vide ; la procédure *modifierEtiquette* modifie l'étiquette de la transition *t* de T_i par ε.
[97] Ceci est une reprise de M. Constant (2002).

A	B	C	D	E	F
Indcx Nc	Loc=: à	Loc =: dans	Loc =: cn	Loc =: E	Nc
1	-	+	-	-	département
2	+	+	+	-	mer
3	-	+	+	-	région
4	+	+	-	-	île
5	-	+	-	-	océan
6	-	+	-	-	état
7	-	+	-	-	état

Table 21 : table PNNpr

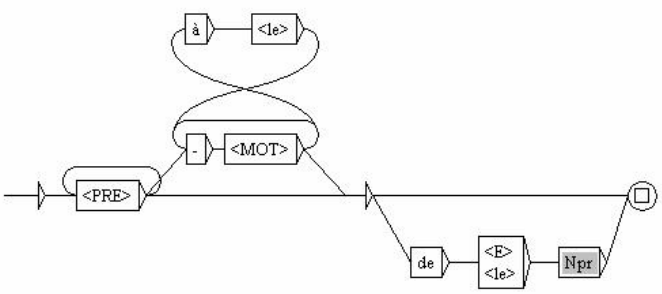

Figure 81 : Npr (générique)

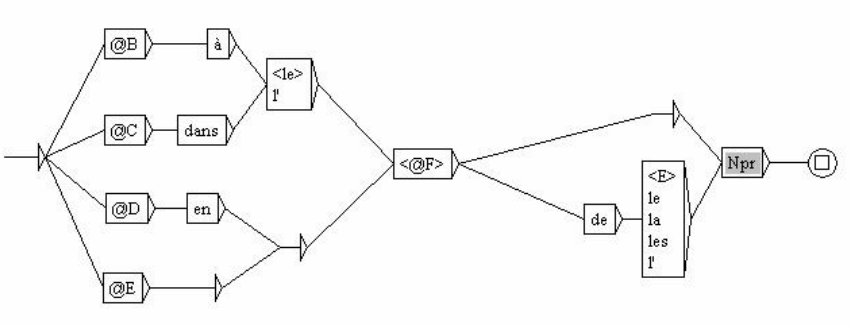

Figure 82 : graphe patron de PNNpr (approximation)

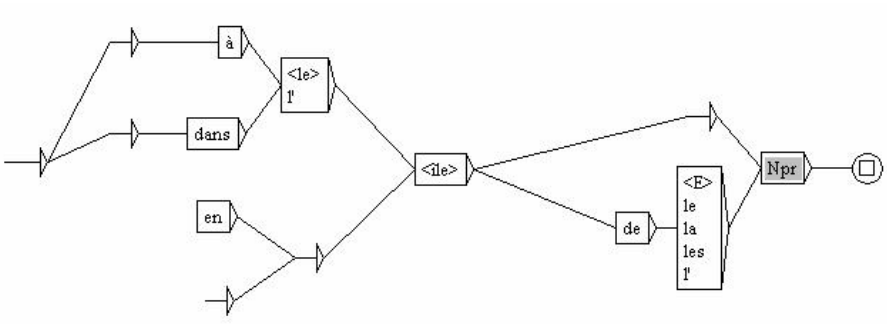

Figure 83 : entrée *île*

4.5.1.2 Un système relationnel

Dans la section précédente (cf. 4.4), nous avons construit un ensemble conséquent de tables. Les tables élaborées sont différentes de celles que l'on a l'habitude de construire. En effet, il

existe des relations entre elles : particulièrement entre la table des classifieurs et chacune des tables du type NNpr. Pour chaque table de noms propres, chaque ligne ayant pour classifieur *Nc* pointe sur la ligne de **PNNpr** qui décrit l'entrée de *Nc*. En fait, comme dans chaque table, tous les noms propres ont le même classifieur, toutes ses lignes pointent vers la même entrée de **PNNpr**. Mais cela est dû à notre choix de classification pour les noms propres. Si nous les avions classés selon leur structure interne, toutes les lignes de la table de type NNpr ne pointeraient pas sur la même ligne de **PNNpr** car il y aurait plusieurs classifieurs dans une même table. Ainsi, on peut dire que nous avons construit plusieurs systèmes relationnels de tables syntaxiques, composés chacun de deux tables (une table de type **NNpr** et la table **PNNpr**). Il est donc nécessaire de ré-examiner la méthode standard de conversion des tables en graphes, qui n'est plus valable dans ce type de système. Ce processus requiert des informations contenues dans les deus tables du système, ce qui n'est pas réalisable avec l'approche traditionnelle. Dans la suite, nous développons un modèle et un algorithme prenant en compte ces exigences.

Pour simplifier la compréhension du lecteur, nous supposons que la table de type NNpr contient différents classifieurs. Cet exemple sera utilisé tout au long de cette partie.

A	B	C	D	E	F	G	H	I	J
Indcx Npr	Indcx Nc	Nc	Prcp	Dct	Npr	LE Nc Prcp Dct Npr	LE Nc Prcp Npr	LE Nc Npr	Dct Npr
1	1	département	de	le	Nord	+	-	-	+
2	2	mer	de	le	Nord	+	-	-	-
3	3	région	de	le	Nord-Pas-de-Calais	+	-	+	+
4	1	département	de	l'	Oise	+	+	-	+
5	4	île	de	<E>	Oléron	-	+	-	+
6	5	océan	-	le	Pacifique	-	-	+	+
7	6	état	de	<E>	Israël	-	+	-	+
8	7	état	de	la	Californie	-	+	-	+
9	2	mer	-	la	Méditerranée	-	-	+	+
10	3	région	de	les	Pays de la Loire	+	-	-	+

Table 22 : NNpr

A	B	C	D	E	F
Indcx Nc	Loc=: à	Loc =: dans	Loc =: en	Loc =: E	Nc
1	-	+	-	-	département
2	+	+	+	-	mer
3	-	+	+	-	région
4	+	+	-	-	île
5	-	+	-	-	océan
6	-	+	-	-	état
7	-	+	-	-	état

Table 23 : PNNpr

Soit *M* une table qui contient des éléments lexicaux et des booléens (+ pour vrai et – pour faux). Chaque entrée lexicale est une clé primaire. La colonne contenant les entrées lexicales est appelée colonne primaire de la table. On appelle éléments secondaires les autres éléments lexicaux de la table : par exemple, dans **NNpr**, à la ligne 4, *département*, *de*, *le*, etc. sont des éléments secondaires alors que *Oise* est l'entrée lexicale. Le comportement syntaxique de certains éléments secondaires est parfois représenté indépendamment dans d'autres tables. Dans notre exemple, c'est le cas des *Nc* dont la distribution des prépositions est représentée dans la table **PNNpr**, lorsqu'ils sont associés aux noms propres *Npr*. Pour chaque ligne d'une table *M*, de tels éléments sélectionnent une ligne d'une autre table *M'*. Informellement, on peut considérer cela comme un appel à une sous-ligne qui contient les informations syntaxiques sur ces éléments. Par exemple, le classifieur *île* de l'entrée *île d'Oléron* de **NNpr** sélectionne la ligne 4 de **PNNpr** où *île* est la clé primaire. De tels éléments sont appelés clés secondaires. Les colonnes les contenant sont des colonnes secondaires. Une colonne clé est soit une colonne secondaire soit une colonne primaire.

Quelques détails techniques :

Notons qu'une clé est supposée unique dans la tradition des systèmes relationnels de bases de données (J.D. Ullman, 1979 ; G. Gardarin, 1999). Or, en réalité, nous avons vu que les entrées lexicales sont ambiguës. Par exemple, le nom *état* est ambigu dans la table **PNNpr** : c'est soit un classifieur de pays (*l'état d'Israël*), soit une partition administrative d'un pays (*l'état de Californie*)[98]. Dans la table **NNpr**, l'entrée lexicale *Nord* désigne soit une mer, soit un département français. Il est donc impossible que ces éléments utilisés tels quels soient des clés primaires. Pour régler ce problème, nous ajoutons, dans nos tables, une colonne dans laquelle, pour chaque entrée lexicale, nous associons un entier unique. Ainsi, chaque entrée a un identifiant unique qui permet de la différencier des autres : 6 est l'identifiant du classifieur *état* désignant un pays et 7 est celui correspondant au classifieur *état* désignant une partie administrative d'un pays.

4.5.2 Modélisation et algorithme

4.5.2.1 Un nouveau modèle

Notre système comprend un ensemble de n tables syntaxiques (M_1, M_2, M_3, …, M_n), un ensemble de relations entre elles et une table principale. On suppose désormais que M_1 est la table principale. Chaque table comporte une colonne primaire et peut comporter un certain nombre de colonnes secondaires[99]. Par exemple, dans notre système, nous avons deux tables syntaxiques reliées entre elles : M_1 = **NNpr** et M_2 = **PNNpr**. Leurs colonnes primaires sont les colonnes 1 des deux tables. La table principale est **NNpr**.
Une clé primaire d'une table M permet de référer directement à n'importe quelle ligne de M[100]. On adopte une numérotation absolue au sein du système pour chaque colonne clé de toutes les tables. Dans notre exemple, M_1 aura la colonne 1 (indice absolu K_1) pour colonne primaire et la colonne 2 (indice absolu K_2) pour colonne secondaire ; M_2 aura la colonne 1 (indice absolu K_3) pour colonne primaire. On a obligatoirement $K_1 \neq K_2$, $K_2 \neq K_3$ et $K_1 \neq K_3$. A chaque colonne secondaire K, on associe sa table cible. R est l'ensemble de tels couples. Dans notre exemple, les clés secondaires dans la colonne K_2 pointent sur les lignes de la table M_2 et ainsi, $R = \{(K_2,M_2)\}$.

Pour chaque ligne u de la table principale, il existe un automate (A_u) sur l'alphabet des indices (absolus) des colonnes secondaires (fig. 85). Un état correspond à une ligne d'une table et est représenté par un couple (r,m) où r est l'indice d'une ligne dans la table M_m. Soient $q = (rq,mq)$ et $p = (rp,mp)$ deux états de l'automate et K une colonne secondaire (qui est la colonne k dans la matrice M_{mq}). Une transition (q,K,p) signifie que l'élément de M_{mq} situé à l'intersection de la ligne rq et de la colonne k sélectionne la ligne rp dans M_{mp} (c'est-à-dire $(K,M_{mp}) \in R$ et $H_{mp}(M_{mq}(rq,k)) = rp$). L'automate A_u comporte un état initial qui correspond à la ligne u de la table principale M_1 (état $(u,1)$). Tous les états sont finaux. Par construction, comme les étiquettes sont des indices absolus (i.e. non relatifs à chaque table), l'automate est déterministe. Théoriquement, cet automate est susceptible de contenir des cycles. Nous donnons ci-dessous un exemple théorique avec quatre tables (M_1, M_2, M_3 et M_4) et cinq relations ($R=\{(K_1,M_4),(K_2,M_2),(K_3,M_2),(K_4,M_3),(K_5,M_3)\}$) ; v, w, x, y, z et s sont les indices de

[98] En général, dans le cas d'un classifieur de pays, *état* a une majuscule initiale (mais pas toujours).
[99] Une table peut ne pas comporter de colonnes secondaires.
[100] au moyen d'une table de hachage H_M qui associe un indice réel de ligne à chaque entrée lexicale (ou plutôt clé primaire). Par exemple, $H_{PNNpr}(2) = 3$ (équivalent de H_{PNNpr} (*mer*) = 3), car la première ligne réelle contient les intitulés des colonnes.

lignes sélectionnées par l'intermédiaire des relations à partir de la ligne *u* de la table principale. On fournit également l'automate associé à la ligne *u* de M_1. Si *q* est l'état (*v*,2) et *p* l'état (*x*,3), la transition (q,K_5,p) signifie que l'élément $M_2(v,k_5)$ sélectionne la ligne *x* de M_3 (k_5 est l'indice relatif dans M_2 correspondant à K_5).

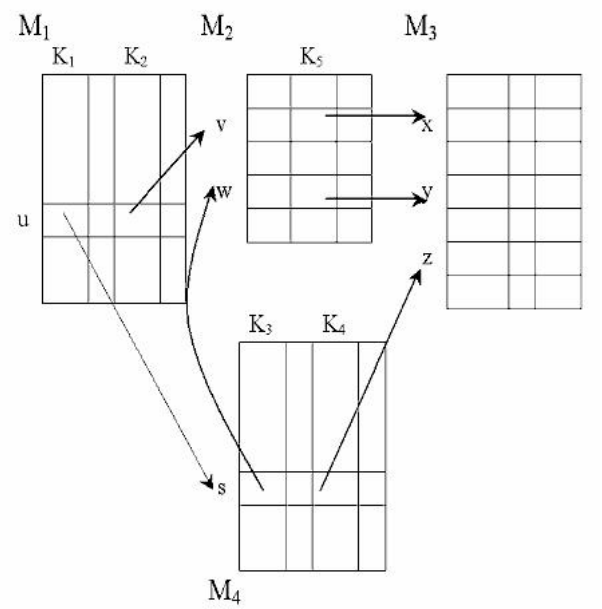

Figure 84 : système théorique

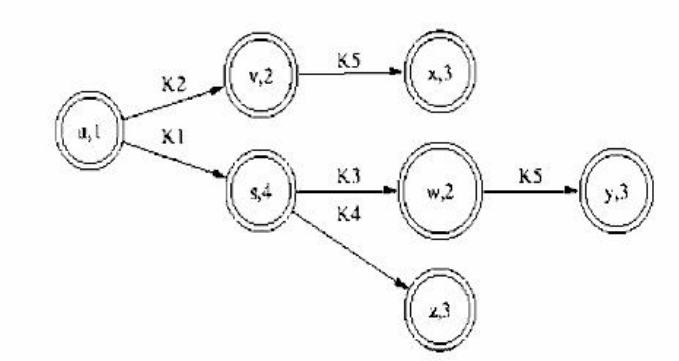

Figure 85 : automate *Au*

On dit que la ligne *v* de la table M_i est sélectionnée par la ligne *u* s'il existe un chemin dans l'automate partant de l'état initial à l'état (*v*,*i*). On dit aussi qu'une table est *m*-sélectionnée si elle contient *m* lignes sélectionnées.

4.5.2.2 Un graphe de référence avancé

Comme dans la méthode précédente de conversion, nous utilisons un graphe de référence (ou graphe paramétré). Le format des variables utilisées est différent car on traite un système avec *n* tables et non plus avec une unique table. Intuitivement, il est au moins nécessaire que les variables contiennent un indice de colonne et un indice de table. Comme les systèmes réels ne comprennent pas de relations complexes, impliquant des tables *m*-sélectionnées (avec $m > 1$), ces informations dans les variables devraient suffire dans la plupart des cas. Cependant, théoriquement, les graphes de référence requièrent plus d'informations dans les variables. C'est ce que nous allons montrer par la suite.

Comment atteindre un élément d'information ?

Pour chaque ligne u de M_1, nous souhaitons construire un graphe qui représente toutes les structures syntaxiques codées dans u et récursivement dans toutes ses sous-lignes. Chaque élément d'information se trouve dans un élément $M_i(v,j)$. Une méthode simple pour atteindre cet élément à partir de la ligne u de la table principale est d'avoir la séquence des colonnes secondaires successives utilisées pour sélectionner la ligne v. Si l'on applique l'automate déterministe A_u à cette séquence, l'état résultant est (v,i) correspondant à la ligne v de M_i. Ainsi, les variables du graphe de référence représentant des éléments d'information doivent contenir une séquence d'indices absolus de colonnes secondaires et la colonne désignant la propriété souhaitée. Comme le format de cette variable comporte une certaine complexité, pas forcément à la portée des linguistes, nous avons simplifié cette représentation.
En fait, la principale difficulté consiste surtout à déterminer une ligne parmi plusieurs lignes sélectionnées dans une table m-sélectionnée avec $m > 1$. Cependant, le cas $m > 1$ ne devrait pas être une situation très fréquente car les phénomènes linguistiques sont relativement simples. Etant donné cette remarque, nous proposons d'utiliser le format suivant pour les variables : $@i[:K_1:K_2:\ldots:K_l]:j$ où j est l'indice de la colonne contenant la propriété souhaitée ; i est l'indice de la table où l'information se trouve ; la séquence $K_1\,K_2\ldots\,K_l$ est la séquence des indices des colonnes secondaires pour atteindre la ligne sélectionnée dans la table i. Si M_i est 1-sélectionnée, $K_1\ldots K_l$ est optionnelle : en pré-calculant une table de hachage PH associant chaque table 1-sélectionnée à leur ligne sélectionnée, on accède directement à l'information. Autrement (si M_i est m-sélectionnée avec $m > 1$), la séquence de colonnes secondaires est obligatoire et le symbole i est redondant car il peut être déterminé en appliquant l'automate A_u à $K_1\ldots K_l$. Comme exemple, nous donnons le graphe de référence associé au système composé des deux tables M_1 (**NNpr**) et M_2 (**PNNpr**). Ces deux tables sont respectivement 0- et 1-sélectionnées. Les variables sont donc formées de deux paramètres (indice de table et indice de colonne). Exemple : la variable @1:6 correspond à *Npr* (colonne 6 de M_1) et la variable @2:3 correspond à l'acceptabilité de la préposition locative *dans* (colonne 3 dans M_2). Notons que nous n'utilisons plus les lettres majuscules (A,B, …,Z,AA,…) pour les indices des colonnes.

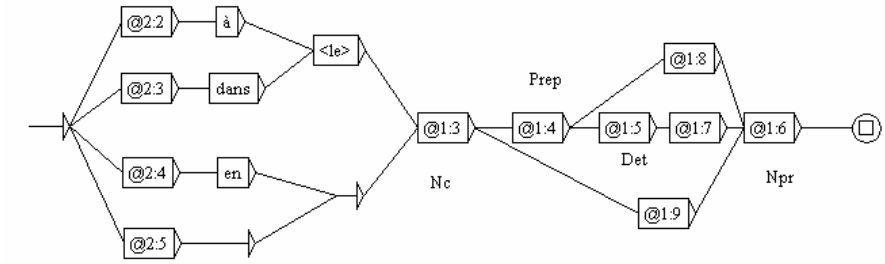

Figure 86 : graphe de référence évolué

4.5.2.3 Un nouvel algorithme

Nous proposons maintenant une extension de l'algorithme standard de conversion des tables syntaxiques en graphes. Soit $T_i = <Q_i, S_i, A_i, \Sigma\cup\Delta, \delta_i>$ une collection de graphes. T_0 est le graphe de référence. Pour $i > 0$, T_i est le graphe associé à la ligne i de M_1. L'algorithme suivant construit automatiquement T_u à partir de T_0 et le système de tables (Σ) :

```
Pour toute ligne u de M1
    Tu←T0.copie()
    Au←∑.constructionAutomate(u,M1)
    PH←∑.calculTableHachage(u,M1)
    Pour toute transition t = (q,a,q')∈ δu
        Si a ∈ Δ
            Si a = [!]@i:j //Mi est 1-sélectionnée
                v←PH(i)
            sinon //a= [!]@i:K1:… :Ki:j
                (v,i) ← Au.appliquerAutomate(K1…Kl)
            finSi
            val ← M(i,j)
            si négation logique dans a
                si M(i,j) = + alors
                    val ← -
                finSi
                si M(i,j) = - alors
                    val ← +
                finSi
            finSi
            si val = + alors
                Tu.modifierEtiquette(t,ε)
            Sinon
                Si val = - alors
                    Tu.supprimerTransition(t)
                Sinon
                    Tu.modifierEtiquette(t, val)
                FinSi
            FinSi
        finSi
    finPour
finPour
```

L'application de cet algorithme à notre système formé de deux tables fonctionne comme suit. Le programme commence à la première ligne de M_1 (entrée : *département du Nord*). Il réalise une copie du graphe de référence (T_0) et l'assigne à T_1. A_1 est automatiquement construit à partir des données générales sur le système entrées par le linguiste (tables, colonnes clés, relations). La table de hachage PH est ensuite construite. Chaque transition de T_1 est alors examinée et transformée si nécessaire (si l'étiquette contient une variable) :

- les transitions contenant @2:2, @2:4 et @2:5 sont supprimées car $M_2(1,2)$, $M_2(1,4)$ et $M_2(1,5)$ ont la valeur '-' ;
- les transitions contenant @1:8 et @1:9 sont également supprimées ;
- les étiquettes @2:3 et @1:7 sont remplacées par <E> car $M_2(1,3)$ et $M_1(1,7)$ ont la valeur '+' ;
- les étiquettes @1:3, @1:4, @1:5 et @1:6 sont respectivement remplacées par le nom *département*, la préposition *de*, le déterminant *le* et le nom propre *Nord*.

Ainsi, cette première étape du processus produit le graphe T_1 ci-dessous. Puis, le même processus continue pour les autres lignes de M_1.

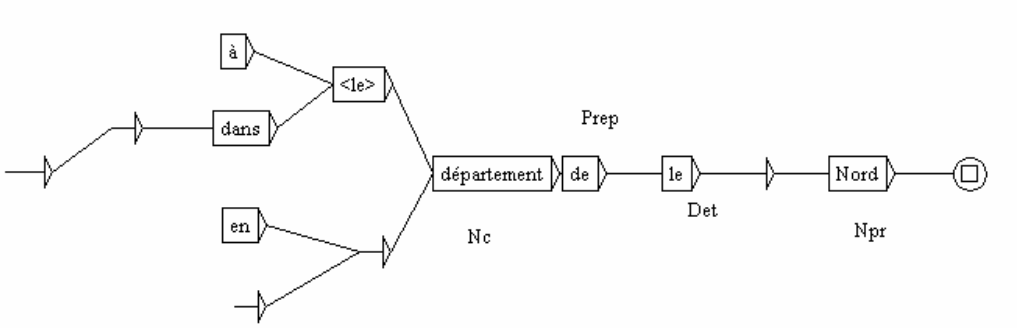

Figure 87 : entrée *département du Nord*

4.5.3 Application

L'analyse linguistique précédente (cf. 4.4) a conduit à la construction d'une table syntaxique de classifieurs représentant leur distribution prépositionnelle dans des groupes prépositionnels locatifs lorsqu'ils sont utilisés à l'intérieur de noms propres composés de lieu. Nous avons également élaboré des tables décrivant le comportement interne de différentes classes de noms propres. Nous obtenons un ensemble de systèmes relationnels. Nous proposons de convertir cet ensemble en graphes à l'aide d'une table générique (ou méta-table) et de la méthode mise en place ci-dessus. Comme précédemment, nous utilisons une table générique afin de générer le graphe paramétré pour chaque classe de noms propres. Pour construire cette table, nous reprenons la table générique élaborée pour décrire le comportement interne des noms propres. Les variables sont mises à jour : on transforme notamment les indices en entiers et l'on ajoute l'indice de la table. Par exemple, @C est remplacé en @1:3. On ajoute également des colonnes correspondant à la distribution prépositionnelle. Comme c'est indiqué dans la table ci-dessous, les noms propres dont la distribution de la forme courte dépend uniquement du classifieur, utilisent les informations codées dans la table 2 pour représenter leur comportement prépositionnel (ex : les noms de mer, de lac, etc.) ; pour les autres, ils utilisent les informations codées dans la table 1 (ex : les noms d'îles). Nous associons à cette table un méta-graphe paramétré au format standard (i.e. un seul paramètre par variable) décrivant l'ensemble des structures d'une entrée fictive qui rentre dans toutes les propriétés se trouvant dans les colonnes. Les variables correspondent aux propriétés de la table générique. Si l'on applique l'algorithme de conversion standard de la méta-table au méta-graphe paramétré, on obtient, pour chaque ligne (chaque table de noms propres), le graphe paramétré associé. Nous donnons ci-dessous le graphe de référence associé à la table **NNpr-île** généré à partir de ce processus automatique.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
Table	Dets singulier	Dets Pluriel	Nc	Modif M2	Adj	Prep	Det (forme longue)	Npr	Variante Npr (forme longue)	LE Nc Adj Det Npr	LE Nc Adj de Npr	LE Nc Adj Npr	LE Nc de Det Npr	LE Nc de Npr	LE Nc Npr	Npr-a	LE Nc Npr-a	(Dets Npro+Det Npr) être un N	LE archipel de Det Npr	Det (forme courte)	Variante Det (forme courte)	Det Npr	Variante Npr (forme courte)	dans LE Npro	à LE Npro	en Npro	Npro	dans Det Npr	à Det Npr	en Npr	à Npr
MNpr-ile	I@1:2	@1:2	@1:4	@1:5	.	@1:6	@1:7	@1:8	.	.	.	.	*	@1:9	@1:10	@1:12	@1:13	*	@1:11	@1:14	.	@1:16	.	@2:2	@2:1	@2:4	@2:5	@1:12	@1:9	@1:7	.
MNpr-departeme	+	.	@1:3	+	.	@1:4	@1:5	@1:6	.	.	.	.	*	@1:7	.	@1:8	@1:9	*	-	@1:5	.	@1:10	.	@2:2	@2:1	@2:4	@2:5	@1:12	@1:11	@1:11	.
MNpr-mer	+	.	@1:3	@1:4	.	@1:5	@1:6	@1:7	@1:8	.	.	.	*	@1:9	@1:10	.	.	@1:11	.	@1:12	.	@1:13	.	@2:2	@2:1	@2:4	@2:5	@2:2	@2:1	@2:4	.
MNpr-ocean	+	.	@1:3	.	@1:4	.	@1:5	.	.	.	.	.	.	.	*	.	.	*	.	@1:6	.	@1:7	.	@2:2	@2:1	@2:4	@2:5	@2:2	@2:1	@2:4	.
MNpr-republique	+	.	@1:3	.	@1:4	@1:5	@1:6	@1:7	.	.	@1:8	@1:3	@1:10	@1:11	.	@1:12	@1:13	*	.	@1:14	@1:15	@1:16	@1:17	@1:13	@2:1	@2:4	@2:5	@1:5	@1:20	@1:18	@1:21
MNpr-region	+	.	@1:3	*	.	@1:4	@1:5	@1:6	.	.	.	.	*	@1:7	@1:8	@1:9	@1:10	*	.	@1:11	.	@1:12	.	@1:13	@2:1	@2:4	@2:5	@1:5	@1:16	@1:14	.
MNpr-lac	+	.	@1:3	*	.	@1:4	@1:5	@1:6	.	.	.	.	*	@1:7	@1:8	.	.	+	.	@1:9	.	@1:10	.	@2:2	@2:1	@2:4	@2:5	@2:2	@2:1	@2:4	.

Figure 88 : méta-table

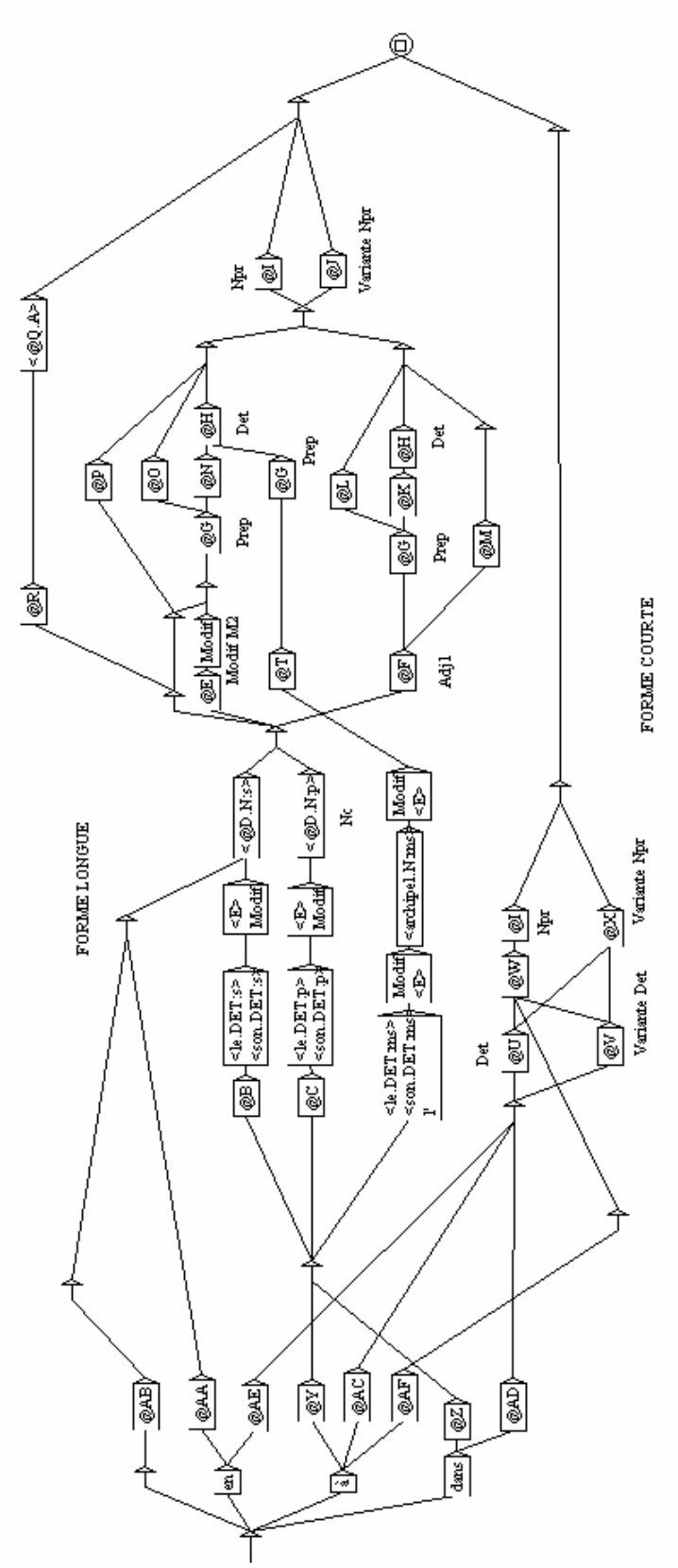

Figure 89 : méta-graphe paramétré PNNpr

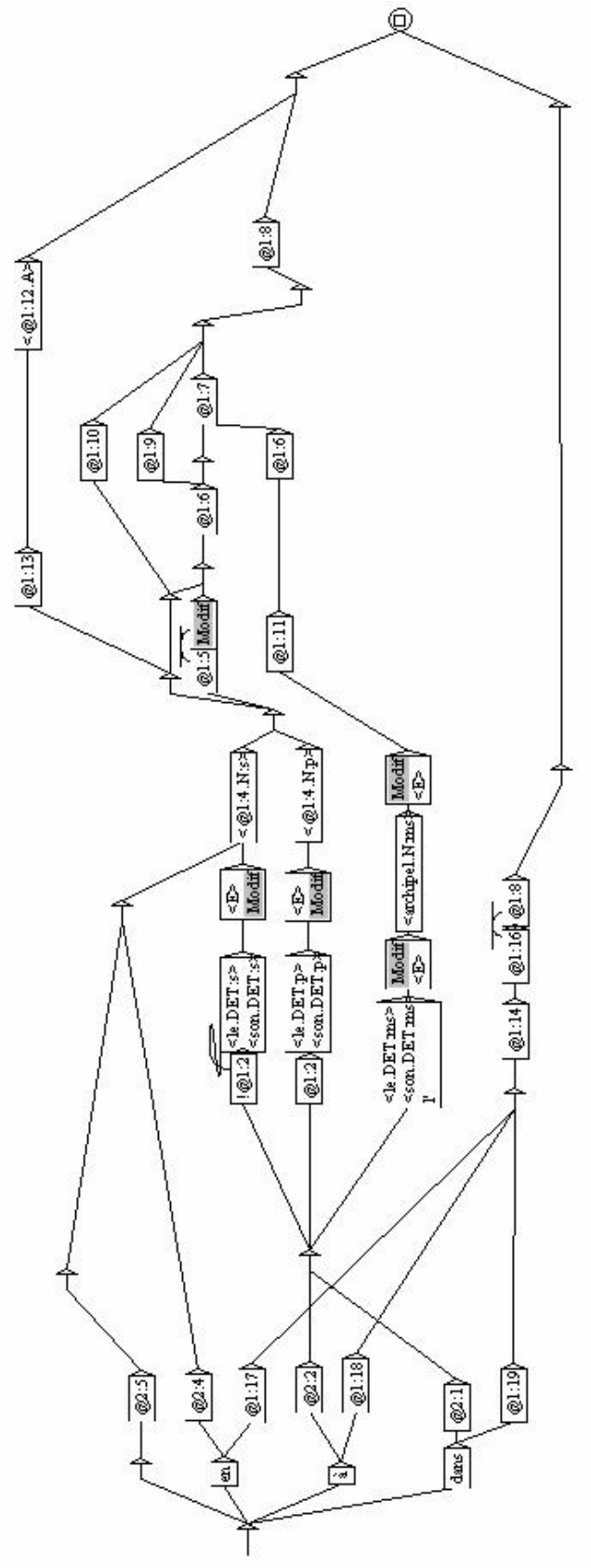

Figure 90 : graphe paramétré pour *NNpr-île*

Ensuite, nous construisons une autre table qui va nous permettre de définir chacun des systèmes relationnels. Chaque ligne de cette table correspond à un système relationnel. Chaque colonne contient une information sur le système. Les colonnes s'organisent consécutivement comme suit :

- nombre de tables du système
- pour chacune des tables : nom de la table, colonne primaire
- nombre de relations
- pour chacune des relations : table de départ, colonne secondaire de table de la table de départ, table cible
- nom du graphe patron associé au système

Ainsi, on construit manuellement la table suivante[101] :

2	NNpr-ile	1	PNNpr	7	1	1	3	2	PNNpr-ile
2	NNpr-departement	1	PNNpr	7	1	1	2	2	PNNpr-departement
2	NNpr-mer	1	PNNpr	7	1	1	2	2	PNNpr-mer
2	NNpr-republique	1	PNNpr	7	1	1	2	2	PNNpr-republique
2	NNpr-region	1	PNNpr	7	1	1	2	2	PNNpr-region

Table 24 : table de définition des systèmes relationnels

La première ligne signifie que le système comporte deux tables : M_1 = **NNpr-ile** (table principale) et M_2 = **PNNpr**. L'indice de la colonne primaire de la première est 1 et celui de la seconde est 7. Il existe une relation qui relie M_1 à M_2 par la colonne secondaire 3 de M_1. Le graphe patron associé à ce système est le graphe **PNNpr-ile** (voir ci-dessus).
Nous implémentons une procédure générale permettant d'extraire, pour chaque ligne, toutes les informations du système associé et puis appliquer le nouvel algorithme afin de générer les graphes de groupes prépositionnels locatifs pour chaque nom propre répertorié dans le système courant. Pour l'entrée lexicale *île de Bornéo* de **NNpr-ile**, nous obtenons le graphe suivant :

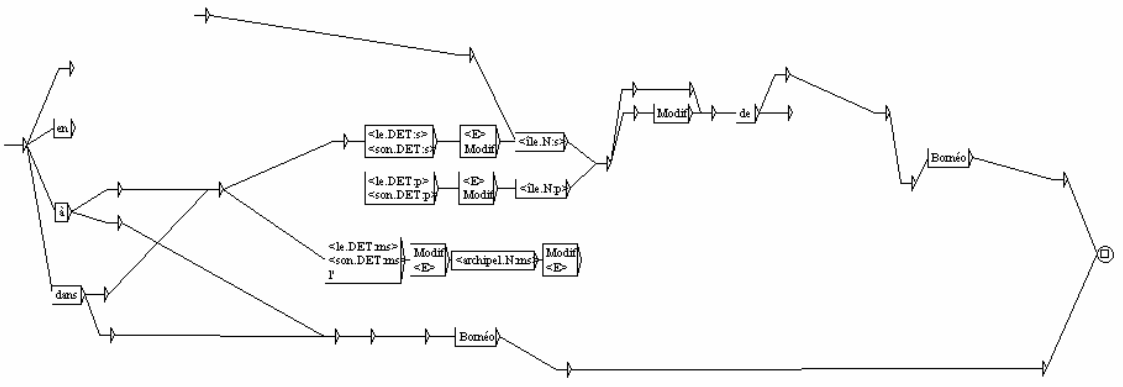

Figure 91 : *île de Bornéo* dans un adverbe de lieu

[101] Il est clair que, pour chaque ligne, le nombre de colonnes utiles varient selon le système défini (nombre de tables, nombre de relations).

4.6 Conclusion

L'analyse automatique des adverbes locatifs ne peut se résumer à un simple repérage des prépositions locatives et des groupes nominaux. L'étude des contraintes syntaxiques entre les constituants des adverbes est un premier moyen de régler le problème (M. Gross 1986, 1996). C'est ce que nous avons cherché à faire en nous attaquant à la distribution prépositionnelle des noms propres de lieu géographique au sein d'adverbes locatifs. Ce domaine n'est pas facile à traiter du fait de la spécificité des noms propres. Le travail que nous avons réalisé est une ébauche méthodologique dans le cadre du lexique-grammaire et mériterait d'être continué par des experts du domaine des lieux géographiques. Nous avons également mis au point une nouvelle méthode de conversion des tables en graphes s'appliquant à des systèmes de tables relationnelles (M. Constant, 2003). Cette représentation relationnelle ne se limite pas seulement à notre problème particulier (adverbes locatifs). En effet, les adjectifs avec noms appropriés semblent pouvoir être formalisés de cette manière (E. Laporte, 1995). Il en est de même pour certaines grammaires de dates (D. Maurel, communication personnelle).