1997
BEL

Université de Nantes

Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel

THÈSE de DOCTORAT

École Doctorale : Sciences pour l'ingénieur de Nantes
Spécialité : **Informatique**

Présentée et soutenue publiquement par

Claude BELLEIL

BIBLIOTHEQUE ★ UNIVERSITE de MARNE - la - VALLEE ★

le 13 octobre 1997
à la faculté des sciences et des techniques de Nantes
devant le jury ci-dessous :

Rapporteurs :	Eric Laporte , Professeur, Université de Reims
	Dominique Laurent, Professeur, Université de Tours
Membres :	Jacques-Henri Jayez, Professeur, Université de Nantes
	Marie-Noëlle Gary-Prieur , Professeur, Université de Lille
	Denis Maurel, Maître de conférence HDR, Université de Tours
Invité :	Franz Guenthner, Professeur , Université de Munich

Directeur de thèse : Denis Maurel
Laboratoire : IRIN

1997 BEL

Remerciements:

Je tiens tout d'abord à adresser mes plus vifs remerciements à Denis Maurel, mon directeur de thèse. C'est lui qui m'a accompagné tout au long de ce voyage à l'intérieur de la recherche. Ses conseils et ses marques d'amitié m'ont été d'un précieux secours. Son niveau d'exigence et sa précision dans le travail constituaient de bonnes garanties de réussite. On garde présent à l'esprit la satisfaction que peut nous donner le sentiment d'avoir apporté une modeste pierre à l'édifice de la recherche universitaire. Bien entendu, on a déjà presque oublié qu'il a fallu tant de fois tout remettre en cause, presque tout défaire pour tout reconstruire différemment!

J'associe bien volontiers à ces marques de reconnaissance les membres du laboratoire (IRIN) et plus particulièrement ceux de l'équipe Langage Naturel. J'aurai une pensée particulière pour Jacques Henri Jayez. Très pris par ses responsabilités de président de l'université de Nantes, je sais qu'il s'est toujours tenu informé de mes travaux. J'aurai un réel plaisir à continuer à travailler au sein de son équipe.

Je n'oublie pas, qu'une partie importante de ce travail a été facilitée par la collaboration de nos partenaires. J'adresse tous mes remerciements à:

Messieurs Annarumma et Murgalé respectivement chef de projet et directeur de la rédaction de Nantes du journal Ouest-France.

Monsieur Le Piétec, qui était à l'époque où nous avons commencé nos travaux, directeur de la communication à la direction départementale de La Poste.

Monsieur Latouche, directeur du CNAM de Nantes.

Enfin, j'adresse mes plus vifs remerciements aux deux rapporteurs de ce mémoire de thèse, Dominique Laurent et Eric Laporte. Leurs remarques et suggestions m'ont aidé à compléter et à enrichir le manuscrit.

Résumé :

Ces travaux proposent le traitement automatique des toponymes français et des noms d'habitants qui leur sont associés : les gentilés.

Après un premier chapitre consacré à un état de l'art, nous introduisons dans le chapitre 2, le problème des noms propres en français. Le chapitre 3 est consacré à l'exposé de notre méthodologie. Celle-ci se décompose en deux phases principales : la reconnaissance et le typage, la résolution des coréférences.

Les chapitres suivants sont consacrés à la présentation des outils informatiques que nous avons construits pour résoudre les problèmes de reconnaissance, de détermination du type et de calcul des coréférences.

Dans le chapitre 4, nous présentons le travail de collecte en détaillant les résultats de l'enquête nationale que nous avons conduite. Ensuite, nous justifions notre choix d'une base de données relationnelles pour réaliser le dépouillement, le contrôle et la validation de ces informations. En conclusion de ce chapitre, nous notons que, pour des raisons de performance et de technique d'interrogation, la base de données relationnelle ne peut être utilisée directement dans le cadre d'un processus d'analyse automatique. Nous expliquons de quelle façon elle a été utilisée pour générer de façon automatique un dictionnaire électronique relationnel des noms de lieux et d'habitants

La construction des outils de reconnaissance, de typage et de résolution des coréférences repose sur la technique des automates et des transducteurs. Nous présentons ensuite les propriétés d'un transducteur particulier que nous avons mis au point: le transducteur d'identifiants.

Dans le chapitre 6, nous présentons le processus de reconnaissance et de typage. Enfin, dans le chapitre 7, nous exposons notre méthode de détection et de calcul des coréférences par fermeture transitive de la matrice booléenne associée. Ceci nous amène à détailler l'ensemble des informations sur les noms propres reconnus que notre système est susceptible de transmettre à un analyseur classique.

Mots clés : noms propres, toponymes, gentilés, bases de données relationelles, automates, transducteur, transducteurs d' identifiants, coréférences, fermeture transitive d'un graphe

BIBLIOTHEQUE ★ UNIVERSITE DE MARNE - LA - VALLEE ★

Recognition, typification and coreference processing
of French toponyms and their inhabitants
by relational electronic dictionary

Summary :

This study proposes the automatic processing of French toponyms and the names of the inhabitants related to them (the gentilés).

After the first chapter surveying current developments, chapter 2 introduces the problem of proper nouns in French. Chapter 3 is primarly dedicated to the description of our methodology and is actually divided into two main parts : firstly recognition and typification and on the other hand, the resolution of the coreferences.

The following chapters present the computing tools that we have made in order to overcome the problems of recognition, determination of type and assessment of the coreferences. Chapter 4 presents the work involved in data collection by describing the findings derived from our nationwide study. Further to this, we justify our choice of a relational data base to perform analysis, verification and validation of this data. This chapter ends on our stating that for purposes of performance and technique inquiry, the relational data base cannot be used directly within the context of a process of automatic analysis. We explain how this has been used to automatically generate a relational electronic dictionary of names of places and inhabitants.

The making of recognition, typification and coreference resolution tools is based on the technique of automata and transducers. We subsequently present the characteristics of a particular transducer that we have developed : the identifier transducer.

Chapter 6 presents the process of recognition and typification. Finally, in Chapter 7, we demonstrate our method of detection and assessment of the coreferences by transitive closure of the related Boolean matrix. This leads us to analyse the pool of data pertaining to the recognized proper nouns that our system is likely to transmit to the typical analyser.

Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel

En informatique linguistique, les noms propres sont des objets du langage qui n'ont été étudiés que récemment. Sur le plan linguistique, leur statut est loin d'être clair. Sont-ils soumis à des règles morphologiques et syntaxiques différentes des noms communs ? Sont-ils porteurs d'une sémantique propre et dans l'affirmative, comment celle-ci s'inscrit-elle dans l'économie générale d'un texte ? Autant de questions qui, encore aujourd'hui, font l'objet de recherches.

En traitement automatique, jusqu'à une période récente, ils n'étaient pour ainsi dire pas pris en compte. Ren et Perrault les placent dans le cadre des *non attendus* [X. Ren & F. Perrault 92] avec les fautes d'orthographe, les sigles, les néologismes et les dérivations peu communes.

Or, ils ont une importance qui est loin d'être négligeable. Une étude réalisée récemment sur un corpus constitué d'articles du journal Ouest-France a donné les résultats suivants [Piton et al. 1996]: sur 564 939 mots pour 37500 phrases, il a été compté 37684 mots différents parmi lesquels 7632 étaient des mots inconnus[1], dont 5755 noms propres.

Particulièrement présents dans les textes journalistiques, les noms propres constituent une catégorie de mots qui participe de façon décisive à la sémantique générale du texte. Dès lors, il parait logique de poser le problème de leur traitement informatique.

Les travaux que nous allons présenter ici, proposent le traitement automatique d'une catégorie particulière de noms propres : les **toponymes** du territoire français (métropole et DOM-TOM) et les noms d'habitants qui leur sont associés : les **gentilés**.

Afin d'inscrire notre démarche dans un cadre général, le premier chapitre de cet exposé sera consacré à une présentation du Traitement Automatique du Langage naturel, en détaillant certaines démarches particulièrement intéressantes pour notre propos. Nous conclurons ce tour d'horizon, qui ne peut en aucun cas viser à l'exhaustivité, par la présentation de quelques travaux représentatifs des voies actuelles de la recherche sur le traitement automatique des noms propres.

Après avoir noté les spécificités idiomatiques de cette catégorie de mots sur les plans morphologique et syntaxique, nous introduirons dans le chapitre 2, le problème des noms propres en français. Une évocation de leur statut particulier en linguistique nous conduira à mettre en évidence les phénomènes caractéristiques qui accompagnent leur reconnaissance dans le cadre d'une **lecture humaine**. Ceci nous conduira à présenter une taxinomie des noms propres du français en situant le cadre de nos travaux : **les toponymes et leurs gentilés**.

Après avoir exposé les caractéristiques de cette catégorie de noms propres, nous présenterons en détail leurs particularités morphologiques. Les problèmes soulevés par la

[1] C'est à dire n'appartenant pas au dictionnaire

présence d'ambiguïtés liées à l'homonymie et à l'existence de formes élidées nous amèneront à proposer des solutions spécifiques de traitement automatique.

Le chapitre 3 sera consacré à l'exposé de notre méthodologie. Celle-ci se décompose en deux phases principales :

- la reconnaissance et le typage
- la résolution des coréférences

Pour cela, nous illustrerons nos choix à partir d'un article du journal Ouest-France du lundi 20 mars 1995. Ce texte, qui possède la particularité de réunir la plupart des difficultés auxquelles on peut être confronté en traitement automatique de noms de lieux et d'habitants, sera utilisé comme fil conducteur tout au long de notre exposé.

*« **La Roche** - **Thouaré** : 2-0*
SANS CONVAINCRE

La Roche-sur-Yon. *- Après leur défaite à domicile contre* ***Nozay****, les* ***Ornaisiens*** *avaient besoin de se rassurer en mettant leur calendrier à jour face à* ***Thouaré****. Si l'objectif fut atteint au niveau du score, ce fut quelque peu laborieux dans la manière. Il est vrai que l'état du terrain ne facilita guère l'évolution des 22 acteurs. Face à des* ***Thouaréens*** *accrocheurs, les* ***Ornaisiens*** *parvenaient toutefois à se créer une bonne occasion après un quart d'heure de jeu (...)*

Peu avant la pause, les ***Vendéens*** *parvenaient toutefois à trouver la faille sur un ballon en profondeur (...) Ce coup du sort eut le don de réveiller les* ***Thouaréens*** *après le repos. »*

Dans le cadre d'un traitement automatique classique, tous les mots en caractère gras sont considérés comme inconnus. En étant éliminés du processus d'analyse automatique dès le niveau lexical, ils ne participent donc pas aux phases d'analyses syntaxique, sémantique et pragmatique. Ainsi, la première phrase du texte devient :

XXX *- Après leur défaite à domicile contre* ***XXX****, les* ***XXX*** *avaient besoin de se rassurer en mettant leur calendrier à jour face à* ***XXX***

On met ainsi en évidence, que si l'information dont ils sont porteurs n'est pas exploitée, l'interprétation du contenu du document se trouve sérieusement appauvrie, au risque de mettre en échec l'ensemble du processus d'analyse automatique. Comme nous l'avons déjà précisé, les travaux présentés ici s'organisent en deux phases de traitement :

La première concerne la **reconnaissance** et le **typage**. Au cours de ce processus, chaque nom propre doit être **reconnu** à partir d'une information stockée et se voir attribuer un **type** : ***La Roche*** (forme élidée de localité), ***La Roche-sur-Yon*** (localité), ***Ornaisiens*** (nom d'habitants d'un sous-ensemble de localité), etc...

La seconde se propose de **détecter et de traiter de façon automatique les relations** que les noms propres d'un texte peuvent entretenir les uns avec les autres. Ainsi, il apparaît déjà dans cette brève présentation, que des liens existent entre les noms propres :

Entité1 *habitant de* Entité2,
Entité3 *sous-ensemble de* Entité4,
Entité5 *forme élidée de* Entité6.

Un traitement automatique devrait avoir la capacité de résoudre ces coréférences directes en mettant en évidence, par exemple :

La Roche *forme élidée de* ***La Roche-sur-Yon.***

Cependant, le processus de traitement ne peut se limiter à ce premier niveau de relations. En effet, dans de nombreux cas le lien entre deux noms propres n'est pas direct : **il se construit par transitivité sur un troisième**. De plus, celui-ci peut très bien **ne pas être présent** dans le texte qui est soumis à l'analyse. Ainsi, dans l'article du journal, deux gentilés sont cités (***Vendéens, Ornaisiens***) sans que leurs référents directs ne soient présents (***Vendée, Saint-André-d'Ornay***). Or, dans la sémantique du texte, ***Vendéens*** et ***Ornaisiens*** désignent bien les habitants de ***La Roche-sur-Yon***. Il est donc indispensable que ces relations transitives puissent être détectées et traitées.

Suite à l'exposé de notre méthodologie, les chapitres suivants seront consacrés à la présentation des outils informatiques que nous avons construits pour résoudre les problèmes

- de reconnaissance et de détermination du type
- de calcul des coréférences

Ayant fait le choix de gérer l'ensemble des informations que nous nous proposons de traiter, notre première tâche a consisté à collecter toutes les données et à les organiser sur le plan logiciel. Dans le chapitre 4, nous développerons les différentes étapes de ce travail initial en détaillant les résultats de l'enquête nationale que nous avons conduite avec le soutien de trois partenaires (Ouest-France, La Poste et le CNAM de Nantes). Ensuite, nous justifierons notre choix d'une base de données relationnelles des noms de lieux et d'habitants pour réaliser le dépouillement, le contrôle et la validation des informations ainsi recueillies. En conclusion de ce chapitre, nous noterons que, pour des raisons de performance et de technique d'interrogation, la base de données relationnelle ne pouvait être utilisée directement dans le cadre d'un processus d'analyse automatique. En revanche, nous expliquerons de quelle façon elle a permis de générer de façon automatique un dictionnaire électronique relationnel des noms de lieux et d'habitants

La construction de cet outil, particulièrement bien adapté au domaine du TAL[2], repose sur la technique des automates et des transducteurs dont nous donnerons, dans le chapitre 5, les définitions ainsi que les caractéristiques de construction et de fonctionnement. Cela nous conduira à exposer les propriétés d'un transducteur particulier que nous avons mis au point et qui réalise un compromis intéressant en associant les performances d'analyse des transducteurs à la gestion des données relationnelles nécessaires à la résolution des coréférences : le transducteur d'identifiants.

Dans le chapitre 6, nous présenterons le fonctionnement informatique correspondant à la première phase de traitement : reconnaissance, typage. Nous exposerons les choix que nous avons faits pour résoudre les problèmes évoqués dans le chapitre consacré à notre méthodologie. Cela concerne essentiellement les graphies multiples, le système de typage et les différents processus de levée des ambiguïtés.

Enfin, dans le chapitre 7, nous exposerons notre technique de détection et de calcul des coréférences par fermeture transitive de la matrice booléenne associée à la « table résultat » construite au cours du processus d'analyse. Ceci nous amènera à détailler l'ensemble des informations sur les noms propres reconnus que notre système est susceptible de transmettre à un analyseur classique pour le seconder au niveau lexical, syntaxique et sémantique.

---

[2] Traitement Automatique du Langage naturel

En conclusion, après avoir exposé les tâches restant à réaliser pour améliorer l'ensemble de ces outils, nous indiquerons de quelle façon ces travaux s'inscrivent dans le cadre du projet de construction d'un dictionnaire électronique général des noms propres : le projet PROLEX [Maurel, Belleil, Eggert, Piton 1996].

1. Le Traitement automatique du langage naturel

Pour envisager un Traitement Automatique complet d'une Langue naturelle, il est nécessaire de disposer d'une description exhaustive de cette langue. Or, cette représentation, ou *grammaire* est encore à réaliser.

1.1 Une approche pluridisciplinaire

Des progrès très importants ont été réalisés dans le traitement automatique des langues, par la mise en place d'une véritable synergie entre la recherche en linguistique et en informatique. Une compréhension réelle des phénomènes de la langue implique la prise en considération de points de vue complémentaires (linguistique, logique, sociologie, psychologie, anthropologie, etc...). Dans ce cadre, les recherches menées en TAL sont assez représentatives des Sciences Cognitives, qui tentent, ces dernières années, notamment autour de la théorie des *Modèles Mentaux* de Philip N. Johnson-Laird, d'instaurer un cadre unifié dans le domaine de la cognition humaine [Ehrlich, Tardieu, Cavazza - 93].

1.1.1 La Linguistique

C'est à Ferdinand de Saussure que l'on doit la linguistique moderne. En la distinguant de la philologie, il construisit les bases d'une science véritablement structurale.

Avant les années 50, qui voient le début des travaux de Noam Chomsky, la linguistique est essentiellement classificatrice et comportementaliste. Elle se propose d'isoler les différents éléments de la langue, correspondant à divers niveaux linguistiques: phonèmes, morphèmes, mots, syntagmes, phrases... Le sens est considéré comme faisant partie plutôt du domaine de la psychologie. On oppose alors les grammaires traditionnelles, fournissant une description des langues particulières sans chercher à expliquer les régularités du langage en général, et la grammaire universelle qui vise à rendre compte de l'aspect créateur du langage, indépendamment des langues.

En 1956, Chomsky vient bouleverser cette vision en construisant une théorie où cette opposition n'est plus justifiée [Chomsky 56 :7].

> *" Cette étude traite de la structure syntaxique entendue à la fois au sens large, où elle s'oppose à la sémantique, et au sens étroit, où elle s'oppose à la phonologie et à la morphologie. Elle s'inscrit dans le cadre d'une tentative visant à construire une théorie générale formalisée de la structure linguistique et à explorer les fondements d'une telle théorie. La recherche d'une formulation rigoureuse en linguistique dépasse de loin le simple intérêt porté à des subtilités logiques ou le désir de clarifier les méthodes établies de l'analyse linguistique. Des modèles de structures linguistiques construits avec précision peuvent jouer un rôle important, à la fois négatif et positif, dans le processus de la découverte lui-même (.../..)*
> *Les résultats rapportés ci-dessous ont été obtenus avec la volonté de suivre systématiquement cette ligne de travail."*

Les deux aspects importants de l'approche de Chomsky sont:

- l'application de méthodes de recherches scientifiques à la formalisation de la syntaxe
- l'introduction de la notion de *structure profonde.*

Les méthodes linguistiques classiques étaient mal adaptées pour rendre compte de certaines ambiguïtés du langage [Sabah 89 :41] :

"... l'exemple suivant:

La critique de Chomsky est injustifiée

possède plusieurs sens bien qu'elle ne contienne pas de mot ambigu et que sa structure superficielle soit très simple. Elle peut signifier en particulier:

- Le fait que quelqu'un critique Chomsky est injustifié,
- Le fait que Chomsky critique quelqu'un est injustifié,
- Le fait de critiquer Chomsky est injustifié

Chomsky se range alors parmi ceux qui cherchent les lois cachées. Le comportement qui consiste à parler (la performance) n'est pour lui qu'un aspect des choses, le plus important étant la compétence linguistique (le savoir sous-jacent du locuteur sur sa langue) qui permet d'expliquer la créativité linguistique.[...] Il est alors amené à soutenir que la structure superficielle d'une phrase comme ***La critique de Chomsky est injustifiée*** *dissimule plusieurs structures sous-jacentes différentes, qu'il appelle* ***structures profondes.***"

Nous reviendrons dans la suite de cet exposé sur les diverses théories linguistiques qui ont fortement influencées la recherche en analyse automatique du langage naturel. Cette présentation rapide de Chomsky était nécessaire ici, dans la mesure où son oeuvre a eu une influence considérable sur l'évolution des travaux dans le domaine de l'informatique linguistique.

Par ailleurs, il faut souligner que la linguistique, qui a longtemps été une science autonome, s'est profondément transformée au contact de l'informatique. L'ordinateur a fourni des outils qui ont favorisé une progression significative des connaissances sur la langue par:

- la construction de corpus
- l'élaboration de dictionnaires électroniques
- les comptages et dénombrements
- l'ordonnancements et les classifications
- l'analyse statistique
- les simulations
- etc ...

D'une façon générale, l'utilisation de l'informatique a permis de réaliser des avancées essentielles dans le cadre de la vérification d'hypothèses linguistiques et d'améliorer ainsi la productivité de la recherche. Ces deux disciplines sont aujourd'hui associées de façon très étroite, on parle volontiers de **linguistique-informatique** ou **d'informatique-linguistique**.

1.1.2 La Logique

La linguistique a apporté aux informaticiens des descriptions de la langue ainsi que des mécanismes qui la régissent. A partir de ces modèles, on a pu envisager de construire des représentations symboliques des différents éléments du discours.

Mais pour pouvoir manipuler cet ensemble de connaissances, l'informatique a besoin de disposer de systèmes formels. Tout naturellement, on a tenté de voir dans quelle mesure la logique mathématique permettait de résoudre ce genre de problèmes.

L'émergence à la même époque des langages *Lisp* et *Prolog*, autour de la théorie du calcul des prédicats du premier ordre, a beaucoup influencé l'informatique linguistique. Concernant les rapports entre la langue et cette approche logique [Sabah 89 :173] note :

> *"Une autre question fondamentale se pose lorsque l'on cherche à appliquer ces techniques au traitement du langage naturel et en particulier lorsque l'on utilise un formalisme logique pour représenter un énoncé. Les diverses valeurs que peut prendre un même énoncé montrent qu'il n'y a pas bijection entre l'ensemble des formules logiques et celui des énoncés et qu'il n'y a donc pas équivalence entre forme linguistiques et représentation. (....) (la nature du rapport qui existe entre eux n'est d'ailleurs pas sujet à démonstration puisqu'on a d'un côté un objet mathématique et de l'autre un objet non formel...)"*

Devant les limites de cette démarche, certains chercheurs ont eu recours à des logiques non classiques sur lesquelles nous reviendrons dans la partie consacrée aux représentations conceptuelles (1.2.3.3).

1.1.3 Psychologie, Sociologie, Anthropologie

Quelles opérations mentales permettent aux hommes de parler et de comprendre ceux qui s'adressent à eux ? Comment utilisons nous les sons, les mots, les phrases, pour exprimer des idées ? Quelle est l'influence d'une situation de communication sur la chose exprimée et la façon dont on l'exprime ?

Longtemps réduite à un dialogue entre linguistes et informaticiens, le Traitement Automatique du Langage Naturel s'ouvre aujourd'hui à de nouvelles disciplines. C'est la prise de conscience de l'importance considérable du sens et du contexte dans les processus de compréhension du langage naturel qui a, peu à peu, amené les équipes à chercher des réponses dans le cadre plus large des sciences humaines.

La sociologie et l'anthropologie ont également contribué à enrichir les connaissances et à approfondir la recherche en précisant les rapports entre le langage, la société et la culture comme il est précisé dans [Carré, Degrémond, Gross, Pierrel, Sabah 91 :38].

> *"Langage et société sont, le plus souvent, considérés comme des entités séparées, étudiables l'une à travers l'autre. Dans la plus grande part des travaux, c'est la société qui est le but de la connaissance et le langage est l'intermédiaire. On parle alors de socio-linguistique. (...) Un second grand axe de travail consiste à suspendre l'opposition, la séparation entre langage et société, pour considérer le langage comme un fait social, un type de comportement. (...) cette position (...) introduit l'idée d'une réflexivité fondamentale entre la langue et ceux qui la parle, car, tout à la fois la langue n'a pas d'existence en dehors du groupe des individus qui la parlent, mais aussi la langue a une existence indépendante de chacun des individus et qui s'impose à lui."*

Dans le même ordre d'idée [Lakoff, Johnson, 85: 14] montrent que notre langage le plus quotidien est traversé par la pratique de la métaphore sans que nous en soyons toujours conscient. C'est le révélateur de l'enracinement au plus profond de l'inconscient individuel et collectif de certains aspects du langage humain :

"Pour indiquer en quoi un concept peut être métaphorique et structurer une activité quotidienne, commençons par le concept de discussion et la métaphore conceptuelle ***La discussion, c'est la guerre****. Cette métaphore est reflétée dans notre langage quotidien par une grande variété d'expressions:*

« Vos affirmations sont indéfendables.... Il a attaqué chaque point faible de mon argumentation.... Ses critiques visaient droit au but.... J'ai démoli son argumentation.... Je n'ai jamais gagné sur un point avec lui.... Tu n'es pas d'accord ? Alors défends-toi !... Si tu utilise cette stratégie, il va t'écraser.... Les arguments qu'il m'a opposés ont tous fait mouche....
(...) Voici un exemple de ce que nous voulons dire lorsque nous disons qu'un concept métaphorique, en l'occurrence, **La discussion, c'est la guerre** *structure (au moins en partie) ce que nous faisons quand nous discutons, ainsi que la façon dont nous comprenons ce que nous faisons."*

Les différents éclairages apportés par les sciences humaines constituent des contributions décisives que la recherche en informatique linguistique doit intégrer dans les difficultés qu'elle se propose de résoudre.

1.2 Les étapes de l'analyse automatique

Une des principales difficultés de l'analyse automatique du langage naturel est liée à la présence d'ambiguïtés dans les textes. Si le lecteur humain s'en accommode relativement bien, il n'en va pas de même des ordinateurs. Ainsi, dès l'étape de découpage du texte en éléments terminaux, sur laquelle nous reviendrons en détail dans la suite de ce chapitre, certains caractères rencontrés peuvent avoir plusieurs interprétations. Prenons un exemple dans les signes de ponctuation avec les différents rôles du point. Généralement, il marque la fin d'une phrase. Cependant, on peut également le rencontrer à l'intérieur d'un sigle, par exemple ***I.R.I.N*** (Institut de Recherche en Informatique de Nantes), ou bien pour indiquer la compression d'un mot: *réf.* pour références.

Comme nous l'avons déjà évoqué, la linguistique cherche à identifier les phénomènes caractéristiques du langage et à en donner des descriptions formelles. Cette démarche est généralement décomposée en étapes distinctes qui sont modélisées sur le plan informatique dans l'ordre suivant:

- La première opération consiste à découper le texte en mots et à "reconnaître" ceux-ci dans un lexique qui contient le vocabulaire "connu" de la machine. C'est l'analyse lexicale. Des connaissances en phonétique, phonologie et morphologie participent parfois à cette phase.

- A l'issue de cette étape, on doit disposer d'une liste exhaustive de toutes les interprétations possibles des mots rencontrés avec pour chacun d'entre eux, la catégorie grammaticale, les propriétés syntaxiques, la signification ...etc...(le volume et les caractéristiques des informations complémentaires contenues dans le lexique dépendent de la stratégie employée au niveau des développements).

- Cette liste est alors communiquée à un analyseur syntaxique qui va prendre en charge la description de la structure des phrases, c'est à dire l'ensemble des dépendances qui existent entre les mots et les groupes de mots. Ce niveau est important car il permet déjà, de dégager du sens.

- Ensuite intervient l'analyse sémantique, qui est chargée d'extraire la signification de l'ensemble du texte, à partir des informations des deux niveaux précédents et de formaliser cette signification dans un système de représentation symbolique sur lequel la machine pourra travailler, c'est à dire essentiellement, effectuer des inférences.

- On peut également rencontrer un niveau d'analyse pragmatique qui a pour objet de prendre en charge les relations avec le monde extérieur et les connaissances sur son

fonctionnement. C'est l'étude du sens dans un contexte particulier et par rapport à un lecteur particulier.

Voyons maintenant en détail le fonctionnement et l'articulation de ces différentes phases d'analyse.

1.2.1 L'analyse lexicale

Elle débute toujours par une segmentation du texte. A partir de chaînes de caractères, il s'agit de détecter des mots. Elle nécessite la gestion d'un lexique qui devra être le plus complet possible afin de minimiser le phénomène dit des « Non-attendus » qui correspond à l'arrivée dans l'analyseur lexical de mots non répertoriés, parmi lesquels on trouve les noms propres, mais aussi les fautes d'orthographe, les sigles, les néologismes... D'autres difficultés peuvent survenir. Il s'agit de la reconnaissance d'expressions figées comme « *prendre le taureau par les cornes* » ou « *avoir les yeux plus grand que le ventre* ».

Pour chacun des mots reconnus, le lexique doit fournir un ensemble d'informations formelles qui seront exploitables par les niveaux supérieurs. Le lexique n'est donc pas constitué d'une simple liste de mots. Il contient d'autres informations dont la forme et l'étendue varient selon les applications.

1.2.1.1 Dictionnaires du LADL

Le DELAS est le dictionnaire électronique du LADL[3] pour les mots simples du français. La définition de ce qu'est un mot simple n'est pas évidente. Il n'existe pas de norme dans ce domaine. Pour cela deux ensembles doivent préalablement être définis:

- un alphabet en relation étroite avec les codes informatiques ASCII ou EBCDIC. , et par rapport auquel certains principes doivent être arrêtés concernant, par exemple, la présence ou non des lettres majuscules.

- l'ensemble des caractères qui peuvent être considérés comme séparateurs. La liste des signes considérés comme séparateurs est le résultat de décisions parfois complexes. Par exemple l'apostrophe est généralement un séparateur, mais pas dans le mot *aujourd'hui*. La classification qui a été adoptée est la suivante [Courtois 90 :11]

> *"Par mots simples nous entendons des unités de texte (...) ne comportant aucun séparateur, en particulier pas de trait d'union, ni apostrophe, ni espace blanc."*

Cette définition a pour conséquence de ranger dans les mots simples les constituants élémentaires des locutions suivantes: *parce*, (de *parce que*), *afin*, (de *afin que*) ainsi que *tohu* et *bohu*...

1.2.1.1.1 Structure des entrées

Dans cette catégorie des mots simples, on distingue deux types de mots:

1 - ceux qui font l'objet d'une description dans le lexique-grammaire et qui correspondent aux termes du français courant.

2 - ceux qui n'ont pas de références syntaxiques, et qui correspondent généralement à des termes techniques ou scientifiques ou à des mots d'un usage très rare.

---

[3] Laboratoire d'Automatique Documentaire et Linguistique

Le dictionnaire DELAS est constitué d'environ 80 000 entrées qui sont structurées de la façon suivante:

- la graphie du mot en minuscule
- une suite d'informations grammaticales et/ou morphologiques qui comportent les éléments suivants:

1 - un symbole de partie du discours

(par exemple N pour un nom, V pour un verbe, A pour un adjectif, ADV et PREP pour respectivement un adverbe et une préposition. Ces parties du discours ont été reprises à partir des descriptions grammaticales usuelles, communes à beaucoup d'ouvrages, à savoir: ***noms, adjectifs, verbes, adverbes, déterminants, prépositions, conjonctions, pronoms, interjections*** *)*

2 - un numéro de code morphologique

(par exemple : N3 ou N21 indiquent des noms appartenant aux classes morphologiques 3 ou 21, A63 un adjectif de la classe 63 c'est à dire suivant les variations morphologiques des noms et adjectifs en ***eux*** *et codé de la façon suivante :* ***[x, se, x, ses]*** *)*

3 - d'éventuels marqueurs de restrictions de formes.

Une lettre supplémentaire est ajoutée à la suite du numéro de code morphologique pour signaler par exemple:

circuler,.V3U -> *verbe à participe passé invariable*
neiger,.V5I -> *verbe impersonnel*
frire,.V90D -> *verbe défectif, c'est à dire, n'étant pas conjugué à tous les temps, tous les modes et toutes les personnes.*

Par ailleurs, ces entrées se répartissent en deux grandes catégories selon que les mots possèdent ou non, un ou plusieurs sens.

1.2.1.1.2 les mots uniques

Ils ne donnent lieu qu'à une seule entrée dans les dictionnaires classiques. En voici quelques exemples extraits du même article déjà cité:

château,.N3
nation,.N21
fougueux,.A63
colmater,.V3
vraisemblablement,.ADV
dans,.PREP

1.2.1.1.3 les mots homographes

Une même graphie correspond à des mots différents. Pour chaque mot, tous les codes spécifiques sont donnés. En voici quelques exemples:

- déjeuner,.N1.V3 *(faisant référence au nom et au verbe)*

- avant,.N1.A80.PREP.ADV *(qui signale un nom masculin, un adjectif invariable, une préposition et un adverbe.)*

De ce fait, le nombre de mots répertoriés dans ce dictionnaire électronique est beaucoup plus important que le nombre des entrées. Les auteurs donnent les chiffres suivants:

- 50 000 noms
- 20 000 adjectifs
- 11 600 verbes
- 2800 adverbes
- 740 autres mots (*déterminants, pronoms, prépositions, conjonctions, interjections.*)

1.2.1.1.4 Les autres dictionnaires du LADL

A partir du DELAS, et de façon plus ou moins automatique, d'autres dictionnaires ont été créés:

Le Dictionnaire des formes fléchies: DELAF

Il est constitué de la totalité des formes fléchies des entrées du DELAS : pluriel et féminin des noms et adjectifs et conjugaisons des verbes. Ceci a été réalisé automatiquement, par un programme mis au point et maintenu par Blandine Courtois.

Sur le plan algorithmique, la génération est assurée par la concaténation d'un radical calculé et d'une terminaison.

Les entrées ont la forme suivante:

- forme fléchie ou conjuguée
- forme canonique et code morphologique
- identification de la forme fléchie
- noms, adjectifs : genre et nombre
- verbes: mode, temps, personne et nombre

exemple:

maison,maison,.N21:Nfs
maisons,maison,.N21:Nfp
soigneux,soigneux,.A63:Ams:Amp
soigneuse,soigneux,.A63:Afs
soigneuses,soigneux.A63:Afp
irais,aller.V16:CPr1s:CPr2s
saviez,savoir.V47:IIm2p

où les symboles terminaux ont la signification:

:Nfs → nom féminin singulier
:Nfp → nom féminin pluriel
:Ams:Amp →adjectif masculin singulier/masculin pluriel
:Afs →adjectif féminin singulier
:Afp →adjectif féminin pluriel
:CPr1s:CPr2s →cond. présent 1ère et 2ème personne sing.

:IIm2p →indicatif imparfait 2ème personne pluriel

Le nombre de formes engendrées à partir des 80 000 entrées du DELAS est de l'ordre de 600 000.

Le dictionnaire phonémique: DELAP

Les dictionnaires DELAS et DELAP décrivent les mêmes mots: les mots simples ne comportant aucun séparateur. Dans le Dictionnaire Phonémique DELAP [Laporte 88], chaque entrée est composée de trois zones:

- zone orthographique (identique au DELAS)
- zone phonémique donnant la prononciation
- zone grammaticale et flexionnelle donnant la catégorie du mot ainsi que ses variations

exemple: *discothèque,/diskotek/,.N21*

C'est un outil très important pour de nombreuses applications informatiques parmi lesquelles on notera: la production automatique de textes phonétiques, la correction orthographique, la synthèse et la reconnaissance de la parole. Nous pensons, à terme, phonémiser notre dictionnaire électronique des noms propres. Dans cette perspective, les travaux d'Eric Laporte sur les dictionnaires du LADL constituent une expérience décisive.

1.2.1.2 Le processus d'analyse lexicale : l'étiquetage

Chacun des mots détectés va être associé aux formes qui sont présentes dans le dictionnaire lexique électronique. Les informations lexicales attachées à chaque mot simple sont indépendantes de leur contexte. Dans un deuxième temps, le programme devra détecter les mots composés. Ensuite le texte sera confronté aux grammaires locales du système pour un premier processus de levée des ambiguïtés.

Voici un exemple d'étiquetage de texte tel qu'il est présenté dans « Dictionnaires électroniques et analyse automatique de textes » [Silberztein, 93]

1 - Texte à analyser :

ARTICLE L. 311-1

Les assurances sociales du régime général couvrent les risques ou charges de maladie, d'invalidité, de vieillesse, de décès, de veuvage, ainsi que la maternité, dans les conditions fixées par les articles suivants :

2 - Le texte étiqueté (extrait) :

ARTICLE	*article*	*N1 :ms*
	articler	*V3 :P1s :P3s :S1 :sS3s :Y2s*
	articler	*V3 :Kms*
L. 31-1	*L*	*CR=50*
	l	*N2 :ms :mp*
	le	*DET :ms :fs*
	le	*PRO :ms3 :fs3*
Les	*le*	*DET :mp :fp*
	le	*PRO :mp3 :fp3*
assurances	*assurances*	*N21 :fp*

sociales	*sociales*	*A76 :fp*
du	*du*	*DET :ms*
	du	*PREP*
régime	*régime*	*N1 :ms*
	régimer	*V3 :P1s :P3s :S1s :S3s :Y2s*
général	*général*	*A76 :ms*
	général	*N2S :ms*
	général	*N76 :ms*
couvrent	*couvrir*	*V33 :P3p :S3p*
les	*le*	*DET :mp :fp*
	le	*PRO :mp3 :fp3*
risques	*risque*	*N1 :mp*
	risquer	*V3 :P2s :S2s*

3 - Résultat de l'analyse lexicale (après la levée de certaines ambiguïtés)

<le, DET :fp>
(
<assurances sociales, NA :fp>+
<assurance, N21 :fp>
<sociale, A76 :fp>
)
(
<de, PREP>
<le, DET :ms>+
<du, DET :ms>
)
(
<régime, N1 :ms>
(<général, A76 :ms>+<général, N2S :ms>+<général, N76 :ms>)
<couvrir, V23 :P3p : S3p>
(
<le, DET :mp :fp>
<risque,N1 :mp>+
<le, PRO :mp3 :fp3>
<risquer,V3 :P2s :S2s>
)
../...

1.2.2 L'analyse syntaxique

Elle a pour objectif d'identifier les constituants de la phrase en indiquant comment les mots sont reliés entre eux et comment ils s'organisent les uns par rapport aux autres. L'analyse syntaxique opère à partir des éléments d'information qui lui ont été fournis par le niveau précédent celui de l'analyse lexicale.

Pour exprimer ces relations formelles entre les différents constituants de la phrase, on doit disposer de connaissances générales, qui sont décrites dans des grammaires. Mais les grammaires disponibles de type "scolaire" ne sont pas du tout adaptées aux traitements informatiques pour deux raisons qui sont soulignées par [Carré, Dégrement, Gross, Pierrel, Sabah, 1991 :59]:

> *"- elles ne sont jamais exprimées dans un formalisme (un langage artificiel de description) interprétable par un ordinateur.*

- elles ne sont pas complètes car elles contiennent des quantités considérables de sous-entendus, d'allant-de-soi, d'évidences non-dites qu'il serait fastidieux de répéter à chaque fois que cela est nécessaire."

La question du choix d'un formalisme adapté à l'outil informatique pour exprimer les phénomènes syntaxiques est donc un problème qui est au centre de l'analyse automatique du langage naturel. De plus, ces formalismes devront servir de base à tous les traitements ultérieurs, y compris ceux du niveau sémantique. Les options retenues à ce niveau auront donc un caractère stratégique dans la mesure où elles ont des conséquences importantes sur les performances de l'ensemble du processus d'analyse.

Les approches ont été très nombreuses et nous détaillerons certaines d'entre elles dans un paragraphe consacré aux formalismes de représentation.

1.2.2.1 L'analyse en constituants immédiats

Les descriptions linguistiques au niveau syntaxique sont généralement formulées en terme de décomposition des constituants. Voici un exemple de cette méthode extrait de [Chomsky 1957]:

La phrase à analyser est :

"L'homme frappe le ballon"

Le nom "*homme*" est déterminé par l'article élidé "*l'*", l'ensemble formant un Groupe Nominal (GN). "*frappe*" est le verbe de la phrase qui a pour complément "*le ballon*" qui est également un Groupe Nominal constitué d'un nom et d'un article. Le verbe et son complément forment ce que l'on appelle le Groupe Verbal (GV) qui avec le Groupe Nominal du début constitue la Phrase.

En procédant ainsi, nous avons opéré de façon intuitive des dérivations successives à partir de règles ayant la forme "X → Y" qui se lisent "réécrire Y pour X" et que nous pouvons résumer de la façon suivante:

Grammaire G:

(a) PHRASE → GROUPE NOMINAL(GN) + GROUPE VERBAL(GV)
(b) GROUPE NOMINAL(GN) → ARTICLE(Art) + NOM(N)
(c) GROUPE VERBAL(GV) → VERBE(V) + GROUPE NOMINAL(GN)
(d) ARTICLE(Art) → l', le
(e) NOM(N) → homme, ballon
(f) VERBE(V) → frappe

L'application des règles de cette grammaire G jusqu'aux éléments terminaux de la phrase donne le tableau suivant que nous appellerons (A):

Tableau de dérivation (A):

1 PHRASE
2 GN + GV
3 Art + N + GV
4 Art + N + V + GN
5 l' + N + V + GN
6 l' + homme + V + GN

7 l' + homme + frappe + Art + N
8 l' + homme + frappe + le + N
9 l' + homme + frappe + le + ballon

Chacune des lignes a été construite en appliquant les règles de la grammaire préalablement définie. Ainsi, la ligne n° 2 est construite à partir de la première en réécrivant PHRASE en GN + GV selon la règle (a) de la grammaire G.

Cette dérivation peut également être formalisée par un diagramme en forme d'arbre (B) de la façon suivante:

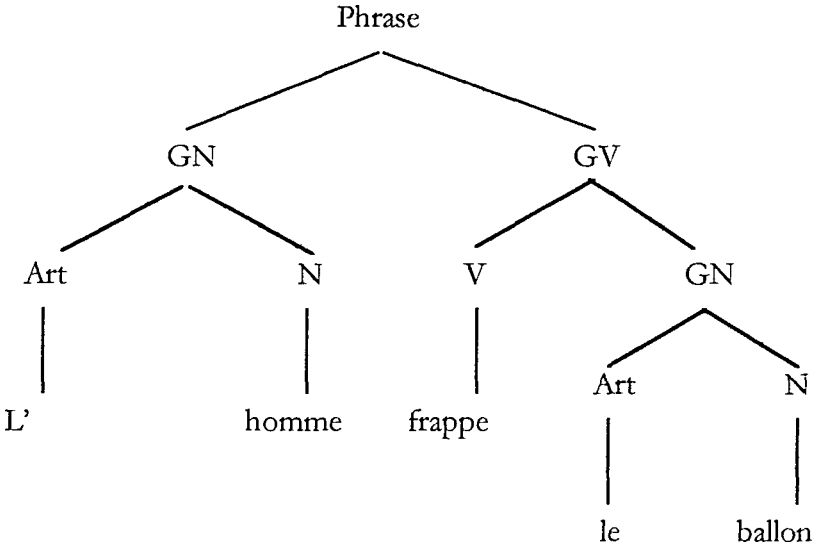


Figure n°1: Diagramme B

Cependant, une difficulté apparaît à l'occasion du passage d'un formalisme à l'autre. Noam CHOMSKY la souligne en précisant:

> *"Le diagramme (B) est moins riche en information que la dérivation (A), puisqu'il ne précise pas dans quel ordre les règles ont été appliquées dans (A). A partir de (A) on ne peut construire que (B), mais l'inverse n'est pas vrai, puisqu'il est possible de construire une dérivation représentable par (B) avec un ordre différent dans l'application des règles. (...)*
> *Sur un point une généralisation de G est manifestement nécessaire. Nous devons être en mesure de limiter l'application d'une règle à un certain contexte. Ainsi Art peut se réécrire "l'" si le nom qui suit (homme) est au singulier, mais non s'il est au pluriel. (...)"*

De plus, il arrive très souvent que les phrases soient structurellement ambiguës en conduisant les analyseur syntaxiques à formuler plusieurs hypothèses de segmentation, sans qu'il soit possible de faire un choix avec les seules informations provenant du niveau lexical.

1.2.2.2 Ambiguïtés de segmentation

Le problème est celui du rattachement des groupes nominaux sujets au verbe de la phrase:

"Mon frère le général et Lucien sont là"

Cette phrase est ambiguë. Voici les deux interprétations que l'on peut en donner:

Première interprétation: Le verbe possède deux sujets.

Cette phrase est formée d'un groupe nominal GN1 [*Mon frère le général et Lucien*] et d'un groupe verbal GV [*sont là*] formé d'un verbe et d'un adverbe. Le groupe nominal GN1 est lui même formé d'un groupe nominal GN11 [*Mon frère le général*]d'une conjonction de coordination et d'un autre groupe nominal GN12 [*Lucien*] correspondant à un nom propre. Le groupe nominal GN11 est formé d'un groupe nominal GN111 [*Mon frère*] et d'un autre groupe nominal GN112 [*le général*]. Les deux sujets du verbe sont le groupe nominal GN11 [*Mon frère le général*] et le groupe nominal GN12 [*Lucien*]. Nous sommes dans le cas de figure de l'apposition. C'est à dire que le groupe nominal [*le général*] est apposé au groupe [*Mon frère*] pour apporter une précision. Bien sûr on pourrait considérer que la présence d'une virgule entre les groupes GN111 et GN112 lèverait l'ambiguïté. En réalité il n'en est rien car même dans le cas de la coordination, cette ponctuation peut se justifier (Figure n°2).

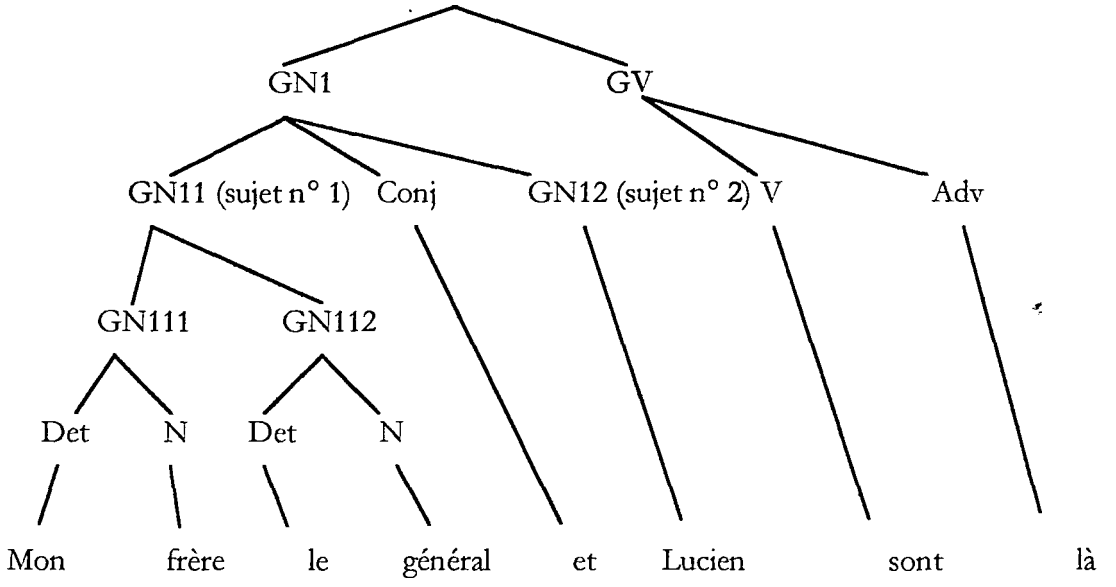


Figure n°2 : Le verbe possède deux sujets

Deuxième interprétation: Le verbe possède trois sujets GN11 [*Mon frère*], GN12 [*le général*] et GN13 [*Lucien*], la conjonction (et) opère sur trois personnes: *Mon frère, le général, Lucien.* La représentation arborescente devient la suivante (Figure n° 3):

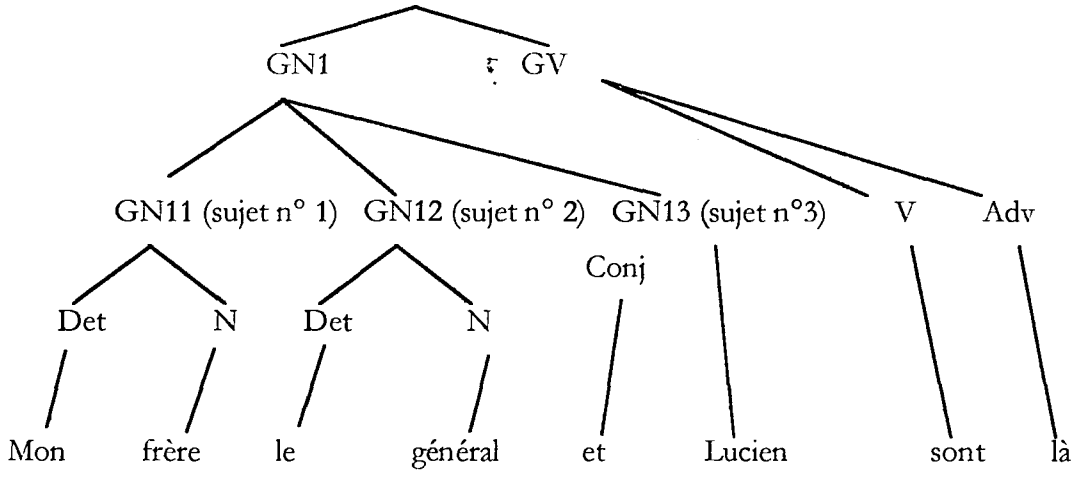


Figure n°3 : Le verbe possède trois sujets

Il n'est pas possible de lever l'ambiguïté à partir des seules informations lexicales et syntaxiques. Seules des connaissances du niveau supérieur, c'est à dire sémantiques, le permettront.

1.2.2.3 Ambiguïtés liées au lexique

Il s'agit de l'étiquetage des mots au moment de l'analyse lexicale. Certaines ambiguïtés peuvent être introduites notamment du fait de la présence de mots composés ou d'expressions figées. Voici un exemple emprunté à [R. Carré, JF Dégrement, M. Gross, JM Pierrel, G.Sabah, 1991]:

Max a sorti de terre deux pommes
Max a sorti deux pommes de terre

Si la forme n°1 n'est pas ambiguë, il n'en est pas de même pour la seconde.

Comme nous l'avions signalé en présentant les lexiques, les difficultés qui apparaissent à un niveau nécessitent des informations du niveau suivant pour être résolues. Ainsi, certains dictionnaires électroniques intègrent des grammaires dites "locales". D'une façon identique, les difficultés liées à l'analyse syntaxique trouvent des solutions soit par l'adjonction d'informations sémantiques dans les règles syntaxiques soit par la mise en oeuvre d'un fonctionnement en parallèle des deux niveaux. On parle alors d'analyseurs syntaxico-sémantiques.

Pour avoir une compréhension réelle d'un texte, il est nécessaire d'aller au delà des structures purement syntaxiques qui ne permettent pas, par exemple, de détecter des proximités de sens comme la forme active et passive d'une construction. L'analyse syntaxique, telle que nous l'avons présentée ici, a donc reçu de nombreuses extensions conduisant à des formalismes plus sophistiqués qui constituent de véritables modèles linguistiques.

1.2.3 Quelques modèles linguistiques

L'objet principal d'une grammaire réside dans sa capacité à reconnaître qu'une suite de mots constitue ou non une phrase. L'être humain réalise cela de façon "instinctive". Or, le développement de l'informatique linguistique a permis , entre autre, de mettre en évidence que de nombreuses règles de grammaire ne sont pas formalisées, car elles ne font jamais l'objet d'un apprentissage de type scolaire. Par exemple, la place de l'article en français obéit à des règles très précises, cependant c'est un aspect du langage qui ne fait jamais l'objet d'un enseignement. Ces règles sont considérées comme étant acquise par la pratique même du langage oral et aucun manuel de grammaire ne présente d'explications dans ce domaine. L'objet presque exclusif de l'enseignement de la grammaire consiste à entraîner les élèves à gérer le décalage entre *"ce que l'on entend"* et *"ce que l'on écrit"*.

A cette difficulté, liée à l'existence de règles "implicites" non formalisées, s'ajoute le fait qu'il n'existe pas aujourd'hui de grammaire couvrant la totalité des phénomènes d'une langue naturelle.

Pourtant, toute la démarche de la recherche, dans le domaine du traitement automatique du langage naturel, repose sur le postulat qu'il est possible de représenter les différents événements de l'activité langagière orale ou écrite d'une langue, sous une forme qui permette des traitements informatiques. La modélisation de l'ensemble des règles qui régissent le

fonctionnement au niveau syntaxique a donné naissance à de nombreux formalismes de représentation.

1.2.3.1 Les grammaires formelles

La notion de grammaires formelles apparaît en 1957, avec la publication de "*Sytactic Structures*" de Noam CHOMSKY. Pour lui, la linguistique a pour objectif de construire une théorie qui permette à la fois de rendre compte de deux phénomènes apparemment contradictoires:

- mettre en évidence le nombre pratiquement infini de phrases possibles qu'une langue peut engendrer,

- construire une grammaire capable de proposer une description de chacune de ces phrases, sur le plan fonctionnel et structurel.

Sa théorie s'appuie sur la constatation que la structure superficielle d'une phrase dissimule plusieurs structures sous-jacentes, qu'il nomme **structures profondes**. Cette notion permet également de rendre compte du fait que la forme active et la forme passive d'une phrase, qui sont deux structures de surface différentes, renvoient à une même structure profonde. Le principal concept caractérisant les grammaires formelles est le suivant: Le vocabulaire est constitué de deux sous ensembles finis et disjoints:

- *le vocabulaire terminal* qui correspond à l'ensemble des symboles pouvant apparaître dans les phrases d'une langue
- *le vocabulaire non terminal* comprenant les symboles nécessaires à la description de la langue et correspondant aux catégories syntaxiques.

Traditionnellement, les grammaires ne mettaient en évidences que deux niveaux:

- *les propositions* (principales, subordonnées, indépendantes etc ...)
- *les catégories lexicales* (noms, verbes, articles ..etc)

Les grammaires formelles structurent la phrase avec un niveau intermédiaire, celui des groupes:

- groupe nominal
- groupe verbal
- groupe prépositionnel
- etc ..

Cela permet de mettre en évidence certaines caractéristiques de fonctionnement d'une langue comme l'aspect récursif de l'apparition de certains groupes. Cette propriété est illustrée au travers de représentations sous forme d'arbres inversés. Sur le plan informatique, cette modélisation correspond à la notion d'automate à pile.

Les grammaires formelles ont été numérotées de 0 à 3 par Chomsky, selon que les langages engendrés allaient du plus général au plus particulier et que la forme des règles était de plus en plus contrainte. A l'époque, les modèles informatiques étant essentiellement hiérarchiques, ces théories linguistiques ont rencontré beaucoup de succès. Mais, comme elles postulent plus ou moins une autonomie de la syntaxe par rapport au lexique et à la sémantique, elles furent peu à peu délaissées au profit de modèles syntaxico-sémantiques.

1.2.3.2 Les grammaires de cas

L'émergence des techniques d'intelligence artificielle a favorisé les modèles linguistiques qui proposaient une représentation de la phrase permettant des raisonnements portant sur le sens. Dans cette optique, la représentation syntaxique apparaît comme une étape qui doit préparer et favoriser une modélisation d'une autre nature, celle qui traitera de l'aspect sémantique.

Si Chomsky est à l'origine des grammaires formelles, c'est à FILLMORE que l'on doit les travaux les plus importants sur les grammaires de cas. Voici un exemple emprunté à [Sabah 90 :92]

Jean casse la branche avec une pierre
La pierre casse la branche
La branche casse

Ces trois phrases rendent compte de la même action. Elles mettent en évidence qu'une même fonction de surface (le sujet) peut être remplie par trois participants différents:

Jean
la pierre
la branche

Bien que ces trois mots aient des rôles syntaxiques identiques, leurs rôles sémantiques sont différents. Ainsi, la structure profonde des phrases, mise en évidence par les grammaires formelles, n'est pas la plus profonde possible. Elle ne représente qu'un niveau intermédiaire entre la syntaxe et un autre aspect qu'il est indispensable de modéliser: la sémantique.

C'est ce niveau encore plus profond que FILLMORE a voulu mettre en évidence au travers de la notion de cas sémantiques. Dans les grammaires de cas, la représentation profonde d'une phrase est composée de deux éléments:

- *la modalité*, qui contient des informations sur la négation, le temps, le mode, l'aspect.

- *la proposition*, qui est une structure indépendante du temps permettant l'identification du verbe, élément central de la phrase.

Les travaux de FILLMORE reposent sur l'hypothèse que dans une langue, il existe une certain nombre de cas permettant de construire des représentations indépendantes. Voici la liste qu'il en donne:

AGENT : L'instigateur animé d'une action
INSTRUMENT : La force inanimée ou l'objet affecté
DATIF : L'animé affecté par l'action
FACTITIF : L'objet résultant de l'action
LIEU : Le lieu ou l'orientation
OBJET : Le reste

exemple: *Jean ouvre la porte avec sa clé*

AGENT : *Jean*
INSTRUMENT : *la clé*

DATIF
FACTITIF
LIEU
OBJET : *la porte*

Deux idées importantes complètent cette théorie, la première consiste à considérer qu'un cas n'a qu'une réalisation, la seconde consiste à attacher à priori à chaque verbe tous ces cas sémantiques possibles.

En proposant un système de modélisation du sens, les grammaires de cas ont été à l'origine de nombreux travaux de recherche. Le rôle important joué par la sémantique a permis le développement de liens étroits avec l'Intelligence Artificielle en permettant la réalisation d'opérations de raisonnement sur les représentations ainsi mises en oeuvre.

1.2.3.3 Les grammaires systémiques

Elles reposent sur une démarche qui s'est développée d'une façon indépendante de la linguistique américaine. Le langage y est plutôt considéré comme une activité à caractère social et profondément lié au contexte d'utilisation. D'autre part, les grammaires systémiques considèrent que l'idée de communication est à la base du langage et qu'il faut plutôt montrer les choix qui s'offrent au niveau du langage. Pour cela on va mettre en évidence les différents traits linguistiques et les différentes fonctions pour pouvoir étudier leurs interactions. Ces grammaires étudient donc davantage l'organisation fonctionnelle de la langue que sa structure grammaticale. On privilégie les liens existants entre la forme d'un texte et son contexte d'énonciation.

Terry WINOGRAD a illustré cette démarche en 1972 avec le programme SHRDLU, présenté dans son livre "Understanding Natural Language".

Celui-ci simule le comportement d'un robot dans un univers clos composé de quelques éléments (blocs posés sur une table qui vont être déplacés d'un endroit à un autre). On notera, et c'est important, que l'ensemble des concepts régissant ce monde très simple sont facilement dénombrables et donc peuvent être efficacement modélisés.

Voici comment il présente ses travaux dans une préface à un article publié dans le cadre du Laboratoire d'Intelligence Artificielle du MIT [Winograd 72 :1]:

> *"This paper describes a computer system for understanding English. The system answer questions, executes commands, and accepts information in an interactive English dialog.*
>
> *It is based on the belief that in modeling language understanding, we must deal in an integrated way with all of the aspects of language syntax, semantics, and inference. The system contains a parser, a recognition grammar of English, programs for semantic analysis, and a general problem solving system. We assume that a computer cannot deal reasonably with language unless it can understand the subject it is discussing. Therefore, the program is given a detailed model of a particular model. In addition, the system has a simple model of ist own mentality. It can remember and discuss its plans and actions as well as carrying them out. It enters into a dialog with a person, responding to English sentences with actions and English replies, asking for clarification when its heuristic programs cannot understand a sentence through the use of syntactic, semantic, contextual and physical knowledge. Knowledge in the system is represented in the form of procedures, rather than tables of rules or lists of patterns. By developing spacial procedural representations for syntax, semantics, and inference, we gain flexibility and power. Since each piece of knowledge can be a procedure, it can call directly on any other piece of knowledge in the system"*

Winograd a développé son application dans le contexte d'un monde volontairement fermé. Il a ainsi mis en évidence un des principes communément admis dans la recherche sur le traitement automatique de la langue naturelle: L'efficacité des traitements passe obligatoirement par la maîtrise des concepts régissant le contexte.

Par ailleurs, un des apports les plus précieux des grammaires systémiques a été d'intégrer, avec des formalismes différents, à la fois les aspects syntaxiques et les aspects sémantiques. Elles ont fourni à l'Intelligence Artificielle des modèles s'appuyant souvent sur la technique de gestion des arbres.

1.2.3.4 La théorie de Z. S. Harris

En traitement automatique, beaucoup d'approches ont en commun, la tentative de confronter au langage naturel, des systèmes formels extérieurs. Ainsi, chez Chomsky, la théorie du langage formel précède l'application. La démarche de Harris est inverse. Il a, vis à vis du langage, la même attitude qu'un entomologiste vis à vis du comportement des insectes. Il observe de façon empirique les phénomènes, puis élabore des théories mathématiques pour modéliser ses observations. Cette construction se réalise en quatre étapes [Harris :76]:

- Construction de **schémas** sur des phrases simples ou élémentaires correspondant à des verbes. Ces phrases sont regroupées en sous ensembles de phrases ayant le même schéma.
- Définition d'une **échelle d'acceptabilité** qui est associée à chaque phrase et qui peut être modulée entre 0 et 1. Ainsi, certaines phrases ont une acceptabilité dans un contexte donné, une autre dans un contexte différent. Sur cette échelle, est défini un préordre.
- Définition d'une **relation d'équivalence** tout en conservant le préordre. Deux phrases sont équivalentes si elles ne diffèrent que par des *ajouts*, des *omissions* ou des *permutations* d'éléments.
- Définition de **transformations** sur les schémas de phrases

Pour Harris, une langue est caractérisée par deux ensembles de générateurs :

- l'ensemble fini des schémas élémentaires
- l'ensemble fini d'opérateurs de base (transformations sur les schémas de base)

La génération s'effectue en deux temps. Si on peut décrire ces deux ensembles, on peut décrire toute la langue.

Il a également défini une *grammaire en chaîne* qui organise les données distributionnelles de façon hiérarchique en partant des schémas de phrases pour arriver aux détails des éléments de chaque schéma. Les grammaires d'arbres adjoints (TAG) sont issues des théories de Harris.

1.2.3.5 Les grammaires d'unification

Elles ont été développées dans les années 80 à partir d'une critique des grammaires transformationnelles. [A.Abeillé, 93 :9] note :

> *« Parmi les points communs des grammaires d'unification, on trouve d'une part le souci d'une articulation plus explicite du lexique, de la syntaxe et de la sémantique (.../...) d'autre part l'accent mis sur les descriptions linguistiques et le recours à un style d'analyse syntaxique plus concret, qui limite le recours à des éléments vides (.../...) et qui restreint le nombre d'étapes intermédiaires dans la production d'une phrase. »*

Les quatre modèles les plus représentatifs des grammaires d'unification sont :

- la grammaire lexicale fonctionnelle (LFG) dont nous allons présenter le formalisme

- la grammaire syntagmatique généralisée (GPSG)
- la grammaire syntagmatique guidée par les têtes (HPSG)
- la grammaire d'arbres adjoints (TAG) qui représente aujourd'hui un modèle où les recherches sont très actives et que nous détaillerons au paragraphe suivant.

Le formalisme de la grammaire lexicale fonctionnelle repose sur la notion de **structure de trait**. Il s'agit d'une notation dans laquelle plusieurs couples *attribut* = *valeur* sont placés entre crochets de façon éventuellement récursive. Par une opération d'unification, on peut ajouter des traits à une description. Voici par exemple figure 4, une opération d'unification [A.Abeillé, 93 :17] entre le nom *corps* et le déterminant *des*. On voit que par superposition, le déterminant a reçu le genre *masc* et le nom le nombre *pluriel*.

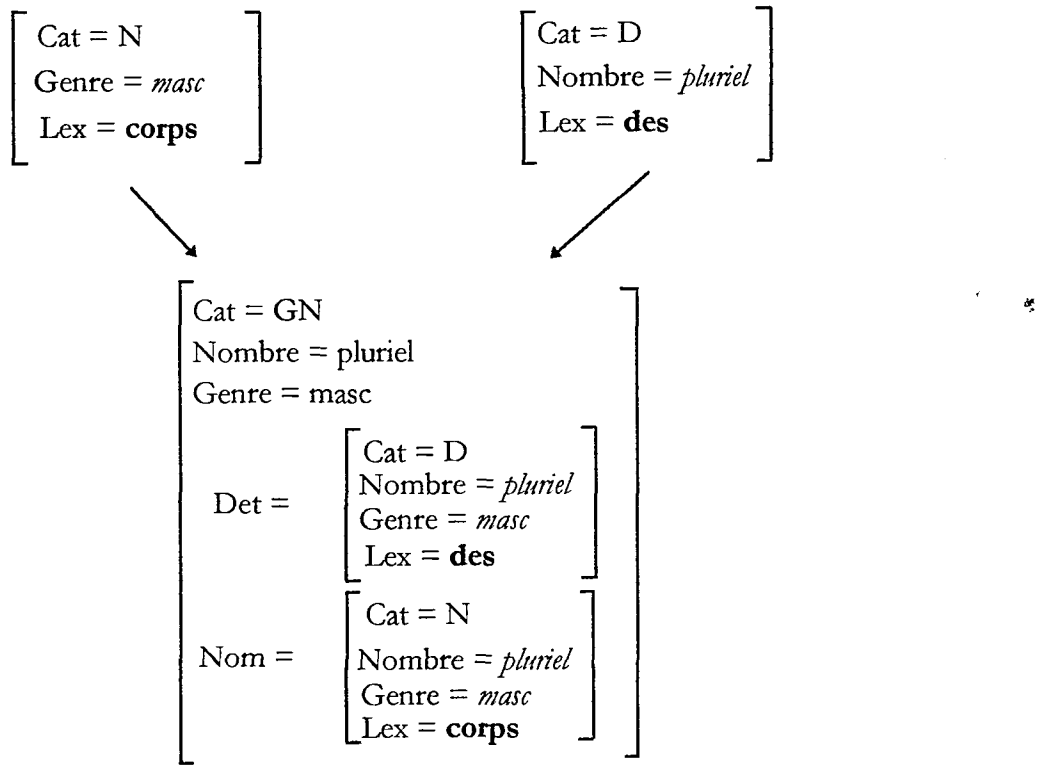


Figure n° 4 : Opération d'unification

L'avantage principal du modèle vient du fait qu'il repose sur un formalisme unique pour décrire le lexique, les règles de grammaire et les phrases de la langue, cependant [A.Abeillé, 93 :16] note :

> *« .../... on ne peut parvenir à une adéquation descriptive (pour la grammaire d'une langue donnée) qu'en multipliant les règles et les traits utilisés ; on ne voit pas se dégager de principes linguistiques généraux, de méthode, qu'on pourrait reprendre pour passer d'un phénomène à un autre ou d'une langue à une autre. »*

1.2.3.6 Les grammaires par adjonction d'arbres

Pour compléter le modèle des grammaires syntagmatiques, N. Chomsky propose une théorie transformationnelle sur les arbres syntaxiques. Il s'agit d'un système de règles qui permettent le passage d'une structure d'arbre à une autre.

Ce principe d'analyse est repris par A. Joshi en 1975 dans le cadre de la grammaire par adjonction d'arbres. Contrairement à la démarche de N. Chomsky, il s'agit d'une approche à la fois lexicale et syntaxique. L'organisation d'ensemble du modèle est la suivante [A.Abeillé,93 :202]

Cette grammaire utilise un ensemble fini d'arbres syntaxiques élémentaires qui sont soit des arbres *initiaux*, soit des arbres *auxiliaires*. A chacun des termes du lexique, un dictionnaire associe tous les arbres élémentaires qui lui correspondent. A partir d'un mot, on obtient la grammaire de ce mot, y compris les formes fléchies.

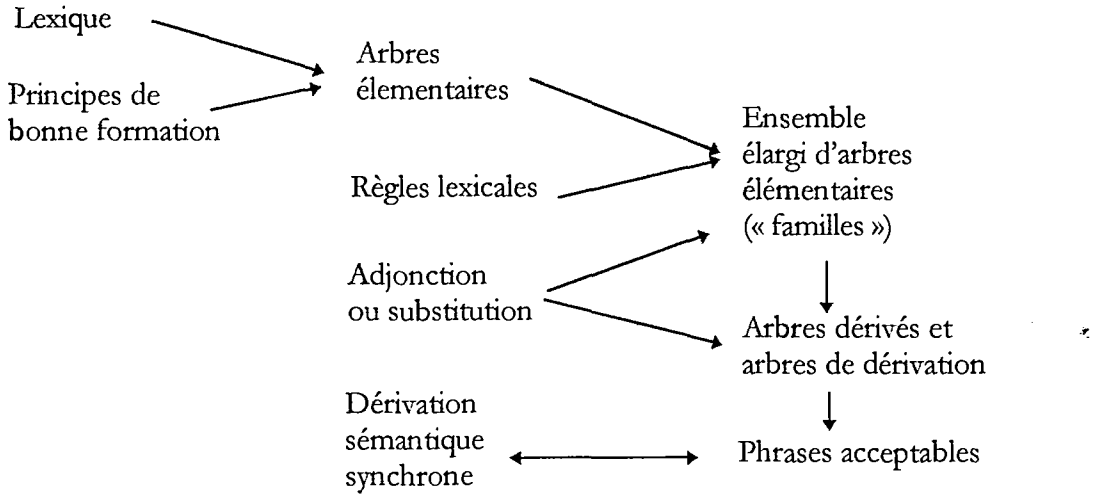


Figure n° 5: Organisation générale d'une Grammaire d'Arbres Adjoints

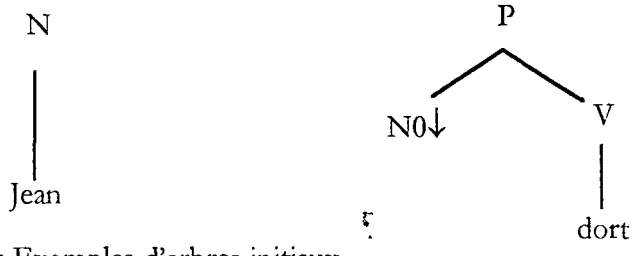


Figure n° 6 : Exemples d'arbres initiaux

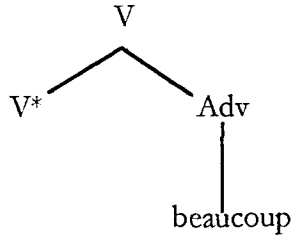


Figure n°7 : Exemple d'arbre auxiliaire

Les opérations de base sont [A. Abeillé, 93]:

• pour les arbres initiaux : la **substitution**

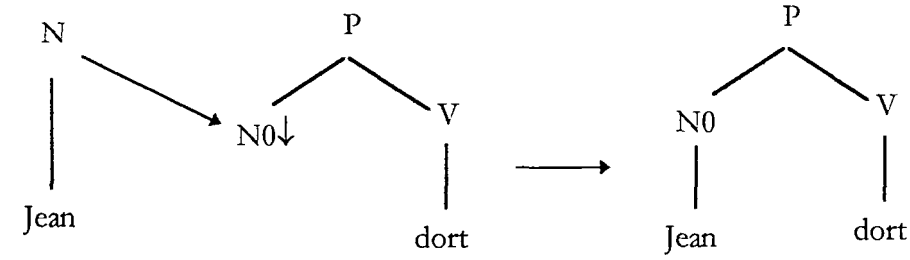


Figure n°8 : Opération de substitution

• pour les arbres auxiliaires : l'**adjonction**

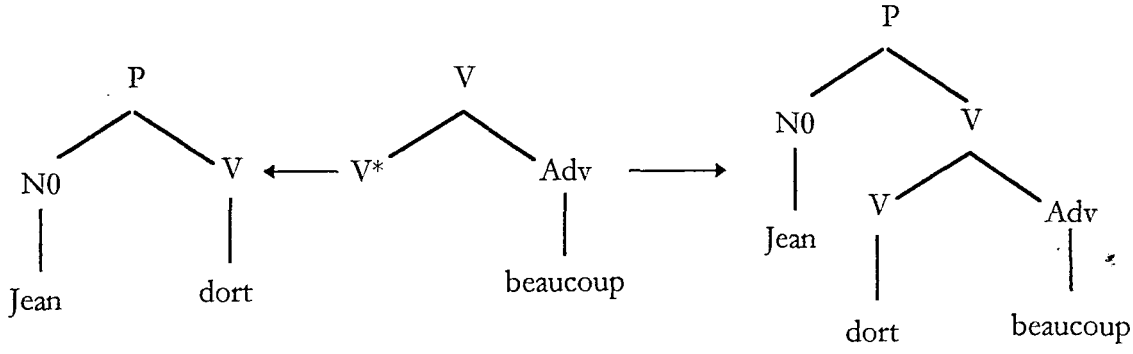


Figure n° 9 : Opération d'adjonction

Cette grammaire, plus puissante qu'une grammaire indépendante du contexte, permet de relier la syntaxe au lexique. De plus, l'opération dite « TAG synchrones », en reliant entre eux deux arbres, permet de représenter certaines transformations syntaxiques et sémantiques comme : les *formes actives-passives*, la *paraphrase* ainsi que certaines règles de *traduction.*

1.2.3.7 Les grammaires locales

Les grammaires régulières, qui constituent la classe des grammaires les plus simples dans la hiérarchie de Chomsky, sont beaucoup trop contraintes pour permettent une description du langage naturel. Cependant, si la première étape d'une analyse automatique est toujours la consultation d'un dictionnaire (analyse lexicale) on peut compléter cette étape par la consultation de grammaires régulières quand elles existent localement. Ceci permet de soulager la phase d'analyse syntaxique qui est toujours plus lourde et donc plus coûteuse sur le plan informatique.

Ces grammaires régulières utilisées par l'analyseur syntaxique du LADL [Silberztein 89] sont appelées *grammaires locales.* Elles constituent un complément à l'analyse lexicale en permettant de lever certaines ambiguïtés de façon plus rapide que d'autres techniques informatiques.

1.2.4 L'analyse sémantique

La frontière entre les niveaux syntaxique et sémantique est souvent difficile à cerner. Certains auteurs vont même jusqu'à définir la sémantique par soustraction. Elle aurait pour vocation de prendre en charge tout ce que la syntaxe n'a pas pu résoudre.

1.2.4.1 Mondes ouverts ou fermés

Le but de l'analyse sémantique est de réduire au maximum les ambiguïtés sur le sens. Généralement, elle consiste à construire un système de codification du sens, c'est à dire un ensemble de concepts formalisés auquel est adjoint un ensemble de règles permettant d'effectuer des opérations sur ces représentations formelles.

Les théories sont nombreuses et variées. Notre objectif ici, est plutôt d'en exposer les principes généraux de fonctionnement.

L'analyse sémantique se répartit en deux grandes classes d'applications, selon que l'on est plutôt dans un "monde fermé" ou plutôt dans un "monde ouvert".

Monde fermé: L'ensemble des concepts du domaine est connu et maîtrisé. Ceux-ci ont fait l'objet d'inventaires, d'analyses et ont été formalisés selon un schéma préétabli. Tous les comportements, raisonnements et dialogues des utilisateurs potentiels ont été pronostiqués. L'objectif de ce réseau sémantique est de rapprocher le sens du texte analysé du "monde fermé" dans lequel on travaille. Les ambiguïtés sont donc théoriquement rares puisqu'on a postulé un comportement "normal" de l'interlocuteur et que tout a été mis en oeuvre pour rendre ce comportement prévisible par le système.

Monde ouvert: La modélisation de l'ensemble des concepts est par définition impossible. L'objectif ne peut donc pas être l'adéquation d'un texte avec un domaine qui par nature n'est pas prévisible. Dans ce cadre, les analyseurs sémantiques ont pour but de lever les ambiguïtés "dans la mesure de leurs possibilités".

1.2.4.2 Le passage de la forme au sens

Indépendamment de la forme qu'ils peuvent prendre, ces systèmes sont tous confrontés aux mêmes problèmes que nous allons essayer de résumer ici.

Qu'il fasse référence à un *monde ouvert* ou à un *monde fermé*, un modèle sémantique constitue toujours une tentative de description d'un monde. Or, en l'état actuel des connaissances il est impossible de construire une machine ayant les capacités d'apprentissage de l'être humain, même si des progrès importants ont été réalisés dans ce domaine.

Cette description du monde qui est toujours l'oeuvre d'un informaticien est relativement figée. Sa mise au point se fait, la plupart du temps, par constatation des échecs. D'autre part, de nombreux modèles sémantiques ont fait appel à la logique mathématique pour fonder leurs règles de fonctionnement. Or, les mécanismes de la langue naturelle dans leur richesse, leur diversité et les infinies nuances dont ils sont capables se prêtent mal à des systèmes qui sont, par nature, très réducteurs. Les tentatives du côté des logiques « particulières » (floues, non monotones, etc...) que nous détaillerons plus loin n'ont pas donné de résultats décisifs.

L'association entre la forme et le sens est encore un problème qui fait, aujourd'hui, l'objet de nombreuses recherches. Une des voies possible consisterait à obtenir une certaine homogénéité des formalismes entre le niveau syntaxique et le niveau sémantique. Comme il est noté dans [Carré, Dégrement, Gross, Pierrel, Sabah, 1991 :171]:

> *"(...) le système syntaxico-sémantique (...) doit être tel que les règles de composition syntaxique d'une part et les lois de composition sémantique d'autre part soient "homomorphes" dans la correspondance entre sens et formes"*

1.2.4.3 Les représentations conceptuelles

Comme nous l'avons déjà souligné, le problème de la représentation du sens est au centre de la plupart des travaux de recherche fondamentale en informatique linguistique. Là comme ailleurs, c'est la question des modèles formels qui est posée. Dans la mesure où c'est, pour les informaticiens, un passage obligatoire avant la modélisation, de nombreux travaux ont été développé pour prendre en compte la sémantique, la pragmatique et le contexte attachés à un texte. Cependant, la limite entre la syntaxe et la sémantique reste floue. Il ne s'agit donc plus ici de descriptions de type linguistique, mais plutôt de représentations de connaissances visant à expliquer leurs conditions de vérité. L'objectif est de pouvoir effectuer des inférences au niveau sémantique et pragmatique, c'est à dire de doter les systèmes informatiques d'outils qui leur permettront de manipuler "du sens".

1.2.4.3.1 Les logiques formelles

Nous avons déjà évoqué les limites de la logique du calcul des prédicats du premier ordre dans le traitement du langage naturel. Elle possède des atouts pour prendre en compte certains aspects de la sémantique du langage naturel. Elle n'est pas ambiguë, sa sémantique est claire, et elle permet des inférences par déduction. En revanche, face à la richesse de la syntaxe et de la sémantique des langues naturelles, elle présente les handicaps suivants :

- pour la mise en oeuvre des mécanismes de raisonnement le problème est de savoir quelles règles appliquer et à quel moment.
- il existe toujours un risque d'explosion combinatoire
- il n'y a pas d'équivalence entre les formalismes linguistiques et les formalismes de représentation des connaissances.
- les connecteurs logiques ne constituent qu'un sous ensemble des connecteurs possibles d'une langue naturelle.
- les connaissances incertaines et/ou imprécises ne sont pas prises en compte
- certains paramètres comme le temps, la croyance ne peuvent se représenter que très difficilement
- il n'est pas possible de tenir compte des liens inter-phases

De nombreuses tentatives se sont développées en direction des logiques non classiques:

- *les logiques modales* : elles ont été introduites [Lewis 92] pour modéliser la notion de causalité. Dans le domaine de la représentation des connaissances, elles permettent de prendre en compte le statut distinct que peuvent prendre diverses affirmations. La notion de vérité devient relative à l'instant considéré ou à un individu particulier. Des notions nouvelles sont introduites comme: la possibilité, la nécessité, l'impossibilité, la contingence. Par ailleurs, un système d'équivalence existe entre ces diverses modalités. L'ensemble repose sur la notion de "monde possible".
- *les logiques multivaluées* qui, possèdent plus de deux valeurs de vérité peuvent traiter des informations incertaines ou imprécises
- *les logiques déontiques* traitant de connaissances juridiques
- *les logiques épistémiques* qui distinguent les croyances des connaissances
- *les logiques temporelles* susceptibles de prendre en compte des instants dont chacun correspond à un monde possible et sur lesquels on définit des modalités
- *les logiques non monotones* fonctionnant dans un monde fermé dans lequel un fait est considéré comme faux quand il n'est pas explicite.

1.2.4.3.2 Les réseaux sémantiques

Les travaux sur les réseaux sémantiques trouvent leur origine dans certaines expériences en psychologie. Il semble, en effet, que certaines informations associées à un concept soient transmissibles aux concepts hyponymes, c'est à dire aux classes plus spécifiques qui sont le résultat d'un raffinage des concepts plus généraux. La mémoire fonctionnerait donc d'une manière plus associative que hiérarchique, les éléments d'information étant stockés selon un principe d'économie.

Sur le plan de la modélisation, ils sont fondés sur la notion de *graphes*, formés de *noeuds* qui représentent les **concepts** et *d'arcs* qui expriment les **relations** entre ceux-ci. Les *noeuds* et les *arcs* sont étiquetés.

Dans le domaine des réseaux sémantiques, les travaux sont nombreux. Cependant, la théorie des graphes conceptuels de Sowa [Sowa 84] occupe une place tout à fait particulière. Il s'agit d'une base théorique générale sur la représentation des connaissances. Elle n'est pas directement liée aux problèmes du traitement du langage naturel. Elle n'est donc pas du tout "marquée" par les modèles linguistiques.

1.2.4.3.3 Les graphes conceptuels de Sowa

En développant sa théorie sur les graphes conceptuels, Sowa a construit un formalisme de représentation de réseaux sémantiques qui possède la particularité de n'avoir été conçu pour aucun domaine particulier d'applications.

J. Sowa est membre du groupe de direction de l'Institut de Recherche sur les Systèmes d'IBM. Ses premiers travaux se situaient dans le domaine des langages de programmation et de l'architecture des machines. Depuis 1972, il se consacre à l'intelligence artificielle et aux langages naturels. Ses recherches en cours sont consacrées à la conception de méthodes d'acquisition de connaissances pour les systèmes experts. Son ouvrage, "Conceptual Structures: Information, Processing in Mind and Machine" publié en 1984 est le produit de son expérience professionnelle dans les domaines du langage naturel et de la philosophie.

Dès l'introduction, il pose le problème des liens entre langage et connaissance ainsi que toutes les questions qui s'y rattachent, à la fois depuis toujours, mais aussi, et surtout, depuis l'émergence de l'informatique linguistique [Sowa 84 :1].

> "Tous les hommes, par nature désire savoir". *C'est avec ces mots qu'Aristote commence sa métaphysique. Mais qu'est-ce que la connaissance ? Qu'est-ce que les gens ont dans leur tête quand ils connaissent quelque chose ? La connaissance s'exprime-t-elle avec des mots ? Si c'est ainsi comment quelqu'un peut connaître des choses qui sont plus facile à faire qu'à dire comme taper un budget ou frapper une balle de base-ball ? Si la connaissance n'est pas exprimée en mots, comment peut-elle être transmise par le langage ? Et quelles sont les relations entre le monde extérieur, la connaissance qu'on a dans sa tête et le langage utilisé pour exprimer la connaissance sur le monde ?*
> *Ce sont les questions traditionnelles qui ont été analysées par les philosophes, les psychologues et les linguistes. Avec l'arrivée des ordinateurs, une série de questions nouvelles surgit: Peut-on programmer la connaissance dans un ordinateur ? Les ordinateurs peuvent-ils encoder et décoder cette connaissance dans le langage ordinaire ? Peuvent-ils l'utiliser pour interagir avec les gens ainsi qu'avec d'autres systèmes informatiques d'une façon plus flexible ? Ce sont les questions qui se posent dans le champ de l'Intelligence Artificielle. Elles s'ajoutent aux éternelles questions au sujet de la connaissance et des relations avec le langage et le monde que les philosophes nous adressent depuis plus de deux millénaires et demi."*

Construire une représentation sémantique, c'est décrire de façon formelle, le sens d'un énoncé en langue naturelle. En général, on fait référence aux idées (concepts) présentent dans

l'énoncé et on établit des liens (relations) entre ces idées. Bien sûr cela participe de l'hypothèse qu'il existerait une représentation formelle "au dessus" du langage, donc d'une certaine façon, indépendante de celui-ci. Cependant, cette idée d'indépendance est fortement controversée. Les tentatives de traductions de certaines expressions typiques d'une langue dans une autre sont là pour étayer fortement ce point de vue. Pour cela, entre autre, les informaticiens ont toujours utilisé des systèmes de modélisation sémantique qui étaient proches de la langue et parfois fortement influencés par la linguistique.

A l'opposé, la théorie des graphes conceptuels de Sowa a été conçue comme un modèle général de représentation des connaissances. Il a d'ailleurs été utilisé dans d'autres domaines de l'informatique qui n'ont pas de liens directs avec le traitement du langage naturel comme la modélisation objet.

La définition que Sowa [Sowa 84 : 69] donne des graphes conceptuels procède d'une approche à la fois psychologique et philosophique:

> *"La perception est le processus de construction d'un modèle de travail qui représente et interprète les perceptions sensorielles. Le modèle a deux composants: un élément sensoriel formé à partir d'une mosaïque de perceptions, chacun d'entre eux assortissant certains aspects de la perception; et un élément plus abstrait appelé un graphe conceptuel qui décrit comment les perceptions s'ajustent ensemble pour former la mosaïque. La perception est fondée sur les mécanismes suivants:*
> ** la stimulation est enregistrée en une fraction de seconde dans une forme appelée icône sensorielle (**sensory icon**).*
> ** Le comparateur associatif (**associative comparator**) recherche dans la mémoire à long terme pour percevoir des éléments de comparaison avec tout ou partie d'un icône.*
> ** L'assembleur (**assembler**) rassemble les perceptions dans un modèle de travail qui forme une approximation fermée. Un enregistrement de cet assemblage est stocké sous forme de graphe conceptuel.*
> *Les mécanismes conceptuels réalisent des concepts concrets (**concrete concepts**) qui associent perceptions et concepts abstraits."*

Quand une personne voit un chat, la lumière reflétée à partir de celui-ci est reçue comme une icône ***s***. Le comparateur associatif assortit ***s***, soit avec la perception ***p*** d'un chat seul, soit avec une collection de perceptions qui sont combinées en une image complète. Comme l'assembleur combine les perceptions, il les enregistre ainsi que leurs interconnexions dans un graphe conceptuel. Dans les diagrammes, les graphes conceptuels sont dessinés comme des enchaînements de boîtes et de cercles. Ces enchaînements représentent les associations logiques dans le cerveau (...)

Le processus de perception génère une structure **u** appelée un graphe conceptuel en réponse aux entités externes ou à une scène **e**.

1 - L'entité **e** donne naissance à un icône **s**.

2 - Le comparateur associatif trouve une ou plusieurs perceptions **p1...pn** qui s'assortissent à tout ou partie de **s**.

3 - L'assembleur combine les perceptions **p1...pn** pour former un modèle de travail qui constitue une approximation de **s**.

4 - Si un tel modèle peut être construit, l'entité **e** est dite reconnue par les perceptions **p1...pn**.

5 - Pour chaque perception **pi** dans le modèle de travail, il y a un concept **ci**, appelé l'interprétation de **pi**.

6 - Les concepts **c1...cn** sont reliés par des relations conceptuelles pour former le graphe conceptuel **u**.

Sowa précise son modèle d'analyse à l'aide d'un ensemble d'hypothèses qu'il n'est pas possible de détailler dans le cadre de ce travail. Il précise notamment l'ensemble des opérations de type relationnel qu'il est possible d'effectuer sur ces graphes. Voici l'exemple qu'il développe [Sowa 84 :92-93] pour illustrer la jointure entre deux graphes: Soient les deux graphes canoniques suivant:

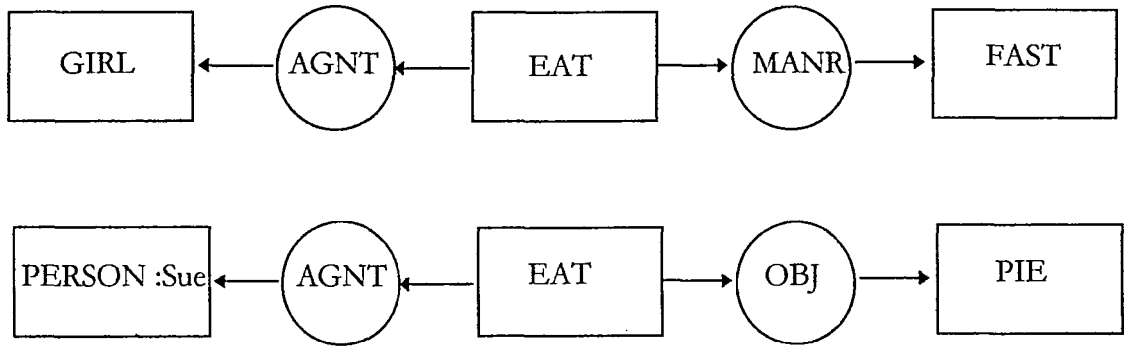


Figure n° 10 : Deux graphes canoniques : A GIRL is eating fast, A PERSON :Sue is eating pie

Si le concept de type PERSON dans le second graphe était restreint au type GIRL, alors le second pourrait être transformé de la façon suivante:

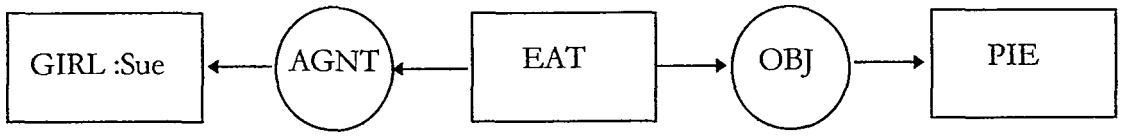


Figure n° 11 : Graphe restreint

définition: Deux relations conceptuelles de même type sont dupliquées si pour chaque *i* le *ième arc* de l'un est relié au même concept que le *ième* arc de l'autre.

Les deux paires identiques de concepts [GIRL :'Sue'] et [EAT] peuvent alors être joints. Le résultat est montré figure 12. Les deux exemplaires de (AGNT) sont dupliqués. Ensuite, l'un d'entre eux sera supprimé par simplification pour donner :

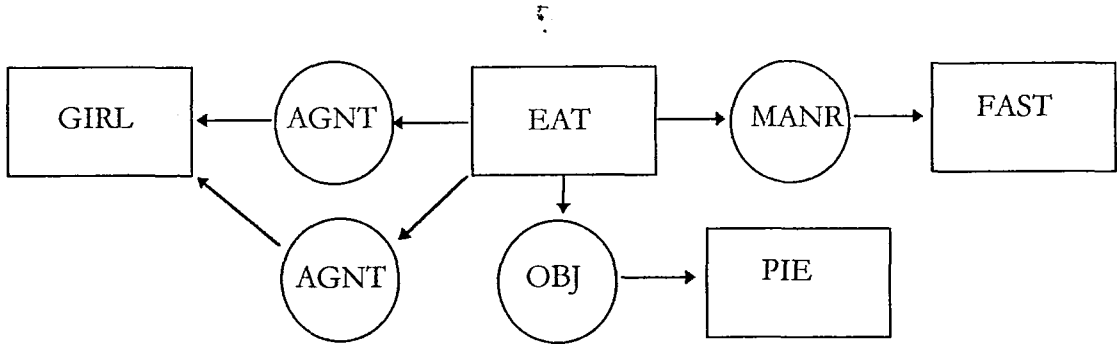


Figure n°12 : Graphe représentant *A girl, Sue, is eating pie fast*

1.2.4.3.4 Les Modèles Mentaux

Parmi les approches nouvelles, la théorie des Modèles Mentaux occupe une place un peu particulière. Pour situer notre propos, voici deux extraits qui définissent bien le cadre de cette démarche [Johnson-Laird, Ehrlich, Tardieu, Cavazza 93 : XIII].

> *"La théorie des modèles mentaux de Johnson-Laird, esquissée dans un article Mental Models in Cognitive Psychology en 1980 et développée dans un ouvrage Mental Models en 1983, est une théorie de la cognition humaine, plus particulièrement une théorie des représentations mentales mises en oeuvre dans le langage et le raisonnement. Se situant dans le cadre des sciences cognitives, la théorie s'inspire des acquis de la psychologie expérimentale, de la linguistique et de l'Intelligence Artificielle."*

Voici comment Johnson-Laird définit lui-même le modèle mental, dans le texte qu'il a écrit pour introduire cet ouvrage collectif:

> *"Un modèle mental est une représentation interne d'un état de chose (state of affairs) du monde extérieur. Il s'agit d'une forme de représentation des connaissances reconnue par de nombreux chercheurs en sciences cognitives, comme étant la façon naturelle par laquelle l'esprit humain construit la réalité, en conçoit des alternatives, et vérifie des hypothèses, lorsqu'il est engagé dans un processus de simulation mentale. La théorie des modèles mentaux doit son origine à trois éminents penseurs: le philosophe Ludwig Wittgenstein, le psychologue Kenneth Craik et le chercheur en sciences cognitives David Marr. (...)*
> *L'idée que les opérateurs humains élaborent des modèles mentaux des systèmes qu'ils contrôlent est admise de longue date, mais la nature de ces modèles - au delà de leur évidente complexité - reste inconnue. En fait, la structure des modèles mentaux est relativement plus facile à analyser dans le domaine que je considère ensuite: la compréhension du discours. Si la compréhension conduit à des modèles du monde (réel ou fictif), alors le raisonnement peut consister en la manipulation de modèles."*

La principale idée remise en cause par cette approche est celle d'une représentation du sens à partir des éléments de la langue elle-même. Comprendre un discours, c'est construire un modèle mental de la situation qui est contenu dans ce discours, mais une telle représentation semble, la plupart du temps, très éloignée de la structure syntaxique des phrases. La question que pose la théorie des Modèles Mentaux est celle de la continuité logique entre le niveau syntaxique et le niveau sémantique. Elle diffère donc des approches traditionnelles (réseaux sémantiques ou graphes conceptuels) selon lesquelles

- la détection du sens consiste à "récupérer" la forme logique de l'énoncé puis à lui appliquer des règles d'inférence.
- la représentation d'un texte encode les significations des phrases du texte.

Johson-Laird pense que dans le processus de compréhension d'un texte, l'individu construit un modèle de la situation qui est décrite et non une représentation de la signification des phrases. D'autre part, on insiste également sur la relation lecteur-texte, qui dans la plupart des théories classiques est perçue dans le sens unique du texte vers le lecteur. En réalité, un texte n'est pas seulement "lu", il est également "reçu" par un individu. C'est à dire que la compréhension qu'il en aura n'est pas réductible à un système logique d'inférences, mais dépend de l'individu lui-même, de son histoire, de sa culture ..etc..

Il étaye sa démonstration en prenant un exemple dans le domaine du raisonnement déductif [Johnson-Laird, Ehrlich, Tardieu, Cavazza 93 : 18] :

*"Si le raisonnement est fondé sur des règles formelles, il ne peut être affecté par les croyances: les règles formelles sont, par définition, imperméables au contenu des prémisses. Mais la théorie des modèles mentaux prédit de tels effets: les sujets qui parviennent à une conclusion putative qui coïncide avec leur croyances auront tendance à arrêter de rechercher des modèles alternatifs qui pourraient réfuter leur conclusion. Les gens sont « **deductive satisficers** ».*
Nous avons testé ces prédictions dans plusieurs expériences [Oakhill, Johnson-Laird, Garnham, 1989]. Quand on donne à des sujets intelligents, mais non entraînés à la logique, les prémisses suivantes:

All the Frenchmen in the room are wine-drinkers
Some of the wine-drinkers in the room are gourmets
la majorité d'entre eux tire la conclusion:
Some of the frenchmen are gourmets

Par contre, quand on leur donne les prémisses:
All the Frenchmen in the room are wine-drinkers
Some of the wine-drinkers in the room are Italians
presqu'aucun d'entre eux ne tire la conclusion:
Some of the Frenchmen in the room are Italians

et la plupart des sujets répond correctement qu'il n'y a pas de conclusion valide (...) Les sujets sont, de façon évidente, guidés par leurs connaissances lorsqu'ils construisent des modèles (...)"

L'ensemble de la démarche est donc une remise en cause de l'hypothèse selon laquelle la compréhension conduit à une représentation linguistique et que le raisonnement dépend de règles formelles d'inférences. Dans le domaine du langage, la description des processus mentaux se réfère donc à la signification de l'information plutôt qu'à sa forme, et prend en compte "l'état" du receveur.

Alan Garnham et Jane Oakhill résument ainsi les propriétés fondamentales des Modèles Mentaux [Johnson-Laird, Ehrlich, Tardieu, Cavazza 93 : 33]:

" - Les représentations mentales du contenu des textes ne correspondent à aucune de ses représentations linguistiques.
- En revanche, le traitement du langage requiert la construction et la manipulation d'analogues mentaux de parties du monde réel où d'un univers fictif. La structure d'un modèle mental correspond à celle de la situation qu'il représente.
- Les modèles mentaux sont construits de façon incrémentative: chaque nouvelle phrase (ou proposition) ajoute des informations au modèle. Pour apprécier la compatibilité des informations et du modèle, il peut être nécessaire de faire usage à la fois de connaissances générales et de connaissances spécifiques non explicitées dans le texte.
- Le modèle courant fournit une partie du contexte interprétatif de la phrase suivante.
- Les Modèles Mentaux sont calculables et finis"

En matière de conclusion, on remarquera que la plupart des idées qui fondent cette théorie ne sont pas nouvelles. Issue de la philosophie, de la psychologie et de l'intelligence artificielle elle tente de démontrer que le raisonnement humain et plus particulièrement la compréhension du langage naturel ne sont pas uniquement des processus formels ou syntaxiques. Ils relèvent de la compréhension de significations et de la manipulation de modèles fondés sur ces significations et sur des connaissances générales. Johnson-Laird précise [Johnson-Laird, Ehrlich, Tardieu, Cavazza 93 : 20]:

"Si l'on veut que les ordinateurs comprennent bien le monde, alors, comme les sujets humains, ils devront pouvoir en construire des représentation internes suffisamment riches, à partir des informations sensorielles et des formes de communication linguistique. Les données psychologiques suggèrent que les sujets humains construisent des modèles mentaux et que, le plus souvent, penser consiste à manipuler des modèles dans le but de chercher des conclusions plausibles"

1.2.5 L'analyse pragmatique

"L'addition du jambon-beurre!"

Même si le niveau sémantique réussit à détecter que "jambon-beurre" fait référence à un sandwich, intuitivement on imagine bien les difficultés qu'une machine rencontrera pour livrer le sens de cette interjection. Ce qui n'est pas le cas de la personne qui, assise à la terrasse d'un café, entend le serveur la prononcer.

C'est à ce phénomène que le niveau pragmatique fait référence. Il a pour objet de prendre en charge deux aspects qui sont essentiels pour une bonne compréhension du langage naturel:

- *le contexte d'énonciation*
- *l'implicite*

L'exemple présenté ci-dessus met bien en évidence les effets de contexte. La difficulté réside dans la façon dont on va pouvoir "donner" au système les informations nécessaires à la compréhension de la situation. En réalité, on est plus dans le domaine des connaissances mais plutôt dans celui de la culture. C'est ce que souligne encore de façon plus nette la deuxième notion: *l'implicite.*

Dans le domaine du langage naturel (texte ou dialogue), le non-dit, l'allant-de-soi constituent la plupart du temps un volume d'informations largement plus important que ce que les protagonistes ou le contenu d'un texte peuvent nous livrer.

En l'état actuel de la recherche, ce niveau d'analyse n'est pris en compte que dans des applications à "monde fermé", comme les systèmes de traductions automatiques hautement spécialisés dans des domaines techniques comme l'aviation, l'armement ou la gestion des communications.

1.2.6 Traitement automatique des noms propres : quelques voies de la recherche

D'une langue à une autre, les noms propres ont en général les mêmes propriétés sémantiques. En revanche, ils présentent sur le plan morphologique et syntaxique des particularités idiomatiques. En allemand, tous les noms (propres et communs) sont marqués par une lettre majuscule. En américain, la lettre capitale du nom propre diffuse à l'intérieur du syntagme dans lequel il est inclus :

« *The Chinese Vice-Prime Minister, Tian Jiyun* »

Le phénomène apparaît peu à peu en français. Il n'est pas rare de rencontrer dans la presse des expressions comme :

« *la Ville de Nantes* »

quand il est question de la localité géographique, alors que la lettre majuscule de « ville » dans le groupe nominal renvoie plutôt à l'entité administrative :

« La Ville de Nantes a conclu un accord de financement avec le Département de Loire-Atlantique »

« La ville de Nantes est située dans le département de Loire-Atlantique »

Autre particularité, les auteurs anglo-saxons mettent une majuscule à tous les mots qui composent un titre, à l'exception des déterminants et prépositions :

« The Analysis and Acquisition of Proper Names for the Understanding of Free Text, by Sam Coates-Stephen »

Les pratiques étant diverses, les traitements mis en œuvre sur le plan informatique le seront également. Cependant, c'est davantage au niveau des heuristiques de reconnaissance que cette diversité est marquée. Une fois cette étape réussie, la sémantique qu'il est possible d'extraire passe toujours par la détection d'une classe ou catégorie et par la mise en évidence des relations que le nom propre entretien avec le reste du texte.

Nous avons retenu trois exemples. Dans le premier, l'originalité tient à la notion de **preuve interne** et de **preuve externe** que McDonald met en œuvre pour identifier un nom propre, ainsi qu'à un système de classification en deux étapes. Dans le second, Sam Coates-Stephen présente FUNES, un système de reconnaissance fondé sur l'apprentissage. Enfin, Paik, Yu et McKenna ont mis au point un système de catégorisation et de normalisation des noms propres qui est utilisé en recherche documentaire et qui repose sur une taxinomie dont nous détaillerons les éléments.

1.2.6.1 Identification et catégorisation sémantique des noms propres de David D. McDonald

David D. McDonald propose un outil de reconnaissance et de classification des noms propres qui fonctionne de façon auxiliaire avec SPARSER, un système d'analyse du langage naturel. Le mécanisme de reconnaissance et de classification est fondé sur les notions de **preuve externe** et de **preuve interne**. La détection des catégories s'effectue, pour les preuves externes, à partir du contexte gauche droite du groupe nominal du nom propre tel qu'il a été délimité, pour les preuves internes, à partir d'une analyse de la séquence de mots et de caractères qui le compose. Par ailleurs, un modèle sémantique de chaque nom propre et de ses composants est mémorisé et utilisé pour la séquence suivante.

Il pense que l'identification et la classification sémantique des noms propres requièrent une analyse basée sur la modélisation des classes orthographiques et lexicales des différents éléments. C'est ce que l'on fait par ailleurs pour des expressions riches de contenu mais pauvres sur le plan des structures de syntaxe comme les *nombres*, les *dates*, les *citations* etc... D'autre part, il met l'accent sur le fait qu'un traitement efficace des noms propres nécessite l'utilisation d'une grammaire qui doit présenter les deux caractéristiques suivantes :

- être sensible au contexte,
- contenir des modèles sémantiques des noms propres et de leurs relations

Ce besoin de sensibilité au contexte est lié au fait que la classification d'un nom propre peut être réalisée de deux façons complémentaires :

par la **preuve interne**, qui est issue d'une analyse de la séquence de mots et de signes à l'intérieur de laquelle se trouve le nom propre. Il peut s'agir de critères considérés comme des preuves définitives tels que des abréviations comme « *Ltd.* » ou « *G.m.bH* » qui renvoient de façon explicite à la catégorie *entreprise*. La preuve interne repose sur la gestion de grands dictionnaires géographiques ou de listes de noms connus pour leurs fréquentes apparitions dans les textes.

par la **preuve externe** qui, à l'inverse, repose sur une analyse du contexte dans lequel le nom propre apparaît. Les noms propres renvoient à des types (personnes, églises, groupes de rock ...) qui ont des propriétés spécifiques et participent à des événements caractéristiques. C'est l'analyse de ces propriétés ou événements dans le contexte immédiat d'apparition qui fournira ou validera la catégorie. C'est presque toujours la façon dont une expression est utilisée qui fait qu'un nom propre appartient à une certaine catégorie. L'aboutissement d'une preuve externe peut, soit constituer une alternative à un échec de la preuve interne, soit valider cette dernière lorsqu'elle était emprunte d'ambiguïté. Par exemple, David D. McDonald cite à ce sujet l'exemple des noms de personnes dans les articles de journaux, qui peuvent tout aussi bien désigner l'individu ou l'entreprise qui porte son nom. L'apparition consécutive de deux occurrences de ce nom propre met la plupart du temps la preuve interne en échec. C'est la preuve externe, par une analyse du contexte qui lèvera l'ambiguïté.

Voici comment l'auteur présente le détail des possibilités de traitements de « PNF ». Il note que ce système, qui fait un usage important de règles de réécriture sensibles au contexte, présente des analogies avec le système FUNES [Coates-Stephen 92] que nous décrirons plus loin.

L'objectif du traitement des noms propres est de délimiter et d'interpréter les différents constituants d'expressions nominales contenant un nom propre, et de contribuer ainsi à l'analyse globale du texte. « PNF » constitue un des composants du système SPARSER. La procédure de traitement est organisée en trois étapes: **Délimitation, Classification** et **Enregistrement**.

L'opération de **délimitation** est réalisée à l'aide d'un automate à nombre fini d'états. L'algorithme reconnaît comme un groupe, toute séquence de mots commençant par une lettre majuscule. La séquence est considérée comme terminée dès la rencontre d'une virgule ou d'un mot commençant pas une minuscule.

La **classification** du nom propre est réalisée en deux temps. Tout d'abord, les algorithmes d'analyse de SPARSER sont appliqués à l'intérieur de la séquence qui vient d'être délimitée. Cela permet d'extraire des informations grammaticales sur les mots et les syntagmes comme par exemple:

- Les références internes au système sur les villes et les pays (« *Cambridge Savings Bank* »)
- L'appel à une classe de **mots clés** tels que « *Church* » ou « *Bank* », ainsi qu'à des éléments inclus comme « *Inc.* » pour les USA, « *P.T.* » pour l'Indonésie ou « *G.m.b.H.* » pour l'Allemagne.
- La classe des abréviations utilisées pour les individus comme « *Jr.* », « *Sr.* », « *Mr.* », « *Dr.* ».
- D'autres éléments comme certaines abréviations, des marques comme « & », des modifieurs ou marqueurs de générations comme « *II* » .

Une fois que les mots constituant la séquence ont été analysés et que le graphe de l'analyse grammaticale a été construit, la seconde partie du processus de classification est mis en œuvre sous forme d'une machine à états finis qui va déterminer la catégorie probable. Si aucune conclusion n'est possible, alors la catégorie générique « **nom** » sera transmise, à charge pour le mécanisme de preuve externe de compléter l'analyse. En revanche, si la catégorisation a réussie, l'arête appropriée du graphe sera étiquetée par une catégorie sémantique telle que « *personne* », « *enreprise* » ou « *journal* ».

Ensuite, le processus d'**enregistrement** prend le relais. L'auteur donne peu de précisions sur les mécanismes de formation des graphes de représentation des classes de noms propres. Il indique que les arêtes reçoivent une instanciation « **dans le modèle du discours** ». La représentation du nom propre est construite avec une étiquette, la séquence des mots telle qu'elle a été délimitée lors de l'étiquetage, ainsi que les arêtes internes au nom. Ce qui doit être produit ensuite, c'est une représentation structurée du nom sous forme d'une instance unique d'une des classes. L'objet nom peut alors être associé, dans l'ensemble du texte, à n'importe quels individus particuliers: **personne**, **entreprise**, **lieu**. Ainsi, l'ambiguïté est traitée comme un phénomène pragmatique.

La structure que le modèle sémantique produit pour les noms propres est conçue pour faciliter l'analyse de sous séquences de mots faisant référence aux individus que les noms propres désignent. Ces types de structure permettent de reconnaître des formes réduites du nom qui sont parfois utilisées. Ce critère a été adopté parce qu'il n'est pas suffisant de noter qu'un nom donné est apparu quelque part dans un article. Ainsi, le nom complet d'une entreprise est souvent cité au début d'un article, par exemple « *Sumito Electric Industries, Ltd* ». PNF doit être capable de reconnaître que c'est du même individu dont on parle plus loin quand il rencontre « *Sumito Electric* », ou « *the company* ». Il doit aussi pouvoir le retrouver dans des expressions qui font référence à d'autres entreprises partageant une partie du nom comme « *Sumito Wiring System* » ou bien de détecter une relation avec « *Sumito Electric International (Singapore)* ». De façon identique, les personnes, les entreprises et les lieux qui partagent le même élément de nom propre devraient être reconnues comme « the *Suzuki Motors Company* ....*Osamu Suzuki*, the president of the company ».

Pour faciliter la détection des références consécutives, il ne suffit pas que chaque nom propre reçoive une instanciation. Les mots qui constituent le groupe dans lequel il apparaît doivent également être instanciés et mis en relation sémantique afin de pouvoir déterminer le rôle que chacun joue. Ainsi le mot « *Suzuki* » instancie toujours le même objet sémantique. A son tour, cet objet est mis en relation avec deux autres individus : **l'entreprise de construction automobiles** par la relation « *first-word-in-name* » et son **président** par la relation « *family-name* ».

L'étiquetage

L'étiqueteur traduit les caractères en éléments représentant les mots, la ponctuation, les séquences de chiffres et d'espaces.

Un mot est connu s'il est mentionné dans une des règles de grammaire. Il a une représentation permanente et l'étiqueteur trouve et transmet cet objet quand il reconnaît la séquence désignée de caractères qui lui correspond.

Les opérations sur les mots

Le processus d'analyse de SPASER est organisé en étapes. L'étiquetage et l'instanciation des éléments terminaux du graphe constituent la première. Alors commence un ensemble d'opérations qui sont déclenchées mot par mot. L'application des règles de structures de syntagmes constitue l'étape suivante. Des heuristiques sont mises en œuvre pour désambiguiser les mots inconnus dans la séquence. L'auteur ne donne pas d'informations sur le fonctionnement des deux derniers niveaux. Seules les opérations sur les mots sont décrites.

Ainsi, les opérations déclenchées par l'identification d'un mot comprennent le traitement des initiales, des abréviations ainsi que la détection d'unités polylexicales que McDonald nomme « **polywords** » suivant [Becker 75]. Il s'agit de séquences de mots qui ne sont pas touchées par des changements morphologiques dus à la syntaxe (pluriel, conjugaison ...). C'est une façon naturelle de prédéfinir des entités comme les noms de pays, les noms d'états (US), les noms des villes importantes.

Quand PNF a terminé la reconnaissance et la classification d'un nouveau nom propre, il ajoute à la grammaire une règle « **polywords** » pour la séquence des mots constituant le « **nom-objet** » enregistré comme une instance « **polywords** ».

L'auteur indique que, lorsque le premier mot de la séquence a été reconnu, l'algorithme de contrôle des "polywords" est particulièrement rapide. Certaines fonctions permettent la détection de la ponctuation « ouvrante/fermante » comme parenthèses ou guillemets, en facilitant le regroupement des mots constituants, même s'ils ne sont pas tous connus. Cela est particulièrement efficace pour reconnaître les surnoms inclus dans les noms de personnes: "*Richard M. ("tricky Dick") Nixon*" ou les insertions d'informations "*manufactured by UNIVERSAL FLUID HEADS (Aust.) PTY. LTD.*".

Au niveau du traitement des mots, le premier traitement réalisé concerne les propriétés du mot, et plus particulièrement les propriétés de ses caractères. Chaque fois qu'une position terminale du graphe est atteinte, cela signifie que le mot suivant est en majuscule. PNF reprend alors le processus de balayage des éléments terminaux du graphe jusqu'à la rencontre d'un mot commençant pas une minuscule.

Quand le processus PNF est terminé, ses résultats sont transmis à une arête du graphe qui est construite sur la séquence des mots ayant une majuscule, avec une étiquette indiquant comment se rattacher à SPARSER, et de quelle façon la zone faisant référence à l'arête pointant le nom-objet est enregistrée.

David D. McDonald termine son article par la présentation d'un exemple que nous reprenons ici :

"An industry analyst, Robert B. Morris III in Goldman, Sachs & Co.'s San Francisco office, said ..."

La lettre majuscule du premier mot "*An*" va déclencher PNF, dont le processus de délimitation s'arrête avec la rencontre du mot suivant qui commence par une minuscule. L'analyseur continu et le groupe *an industrial analyst* est reconnu, la virgule qui suit étant marquée comme annonçant une possible apposition.

PNF est remis en oeuvre par la rencontre du mot *Robert*. Le processus de délimitation du groupe de mots reprend jusqu'à *in*. La mise en oeuvre des règles régulières de grammaire met en évidence l'abréviation et l'indicateur de génération « *III* » (le troisième). L'ensemble des prénoms n'étant pas géré sous forme d'une liste, *Robert* et *Morris* sont considérés comme des mots inconnus. En revanche, l'abréviation et l'indicateur de génération sont des éléments d'information suffisamment explicites pour permettre une classification du groupe comme **nom de personne**.

Celle-ci étant réalisée, une arête est construite sur la séquence avec l'étiquette *personne* et le processus d'enregistrement crée un **nom-objet** pour l'instance d'arête. Le modèle fourni par le classificateur est *name-initial-name-génération indicator*; ce qui est assez clair pour que le sous-type de nom *person's name with generation* soit instancié. Cet objet comprend une séquence allant du premier nom ou de la première initiale jusqu'au dernier (juste avant « *III* ». Ce nouveau nom de personne est étiqueté *Name-1*.

La partie enregistrement consiste à créer les instanciations pour les mots *Robert* et *Morris*. Les individus sont créés en tant que type *single element of a name* et les règles sont ajoutées à la grammaire. Pour assurer une bonne interprétation de la séquence Name-1, les propriétés suivantes sont attribuées aux prénoms : *Robert* est le premier nom de *Name-1* et « *Morris* » est le dernier . Le fait de laisser des mots comme *Morris* désigner des « objet name » avec des liens sémantiques vers le nom dont ils font partie permettra une reconnaissance plus rapide de sous séquences comme *Mr Morris*.

En poursuivant, PNF arrive sur le mot *Goldman*. Il sera considéré comme un mot appartenant à la séquence à cause de la virgule qui suit, cependant, il est traité en tant que mot inconnu. Il recevra l'étiquette générique « **nom** ».

PNF reprend son analyse avec *Sachs* & *Co.* , puis s'arrête à la rencontre de «'s» qui constitue une marque de fin de groupe nominal. Au cours du processus de délimitation, l'abréviation *Co.* aura été reconnue et le signe & évalué comme étant un marqueur pouvant apparaître dans les noms. Pour cette raison, la ponctuation est toujours évaluée au cours de ce processus de délimitation.

La présence de & et du mot *Company* sont des marqueurs définitifs du type **entreprise** et le processus de classification va commencer l'assemblage d'un modèle prêt pour l'enregistrement. Le nom devant la virgule n'ayant pas été classé comme nom d'entreprise, l'arête du graphe sur *Goldman* et la virgule seront considérées comme faisant partie du nom assemblé.

Le mot avec majuscules que PNF isole ensuite est *San Francisco*. Sans avoir besoin d'un dictionnaire, PNF le classe comme nom de lieu, du seul fait de la présence du mot *office* dans le contexte.

A la suite de PNF, le composant des structures de syntagmes de SPARSER prend le relais. Il réécrit les règles aussi bien celles qui sont sensibles au contexte que les autres, par exemple pour le cas ci-dessus :

name → location / _ « office »

Ainsi, une arête ayant l'étiquette « **name** » peut recevoir une nouvelle instanciation « **location** » quand le nom apparaît juste avant le mot *office*. Les règles sensibles au contexte sont mises en œuvre de la même manière que celles qui sont hors contexte, le modèle déclenché ici est le même que celui d'une règle hors contexte avec la partie *name* + *office*.

D. McDonald conclu en indiquant que SPARSER utilise un ensemble d'environ 30 règles.

1.2.6.2 FUNES, un système d'acquisition lexicale automatique de noms propres

Coates-Stephen propose un système de traitement informatique des noms propres appelé FUNES (Figuring-out Unknow Nouns from EngliSh).

Après avoir remarqué que les systèmes de traitement automatique du langage naturel doivent aujourd'hui être capable de traiter des centaines de milliers de mots différents, il note qu'il lui semble impossible que de pareils volumes soient stockés sous forme de dictionnaires. De plus, dans l'hypothèse où une telle démarche serait mise en œuvre, une catégorie particulière de mots n'est présente que rarement dans les dictionnaires : les noms propres.

Ainsi, abandonnant l'idée de pouvoir utiliser des informations sur ces noms par l'intermédiaire de dictionnaires, Coates-Stephen propose de se tourner vers d'autres sources d'information. Il note que les corpus de textes constituent des sources d'information intéressantes pour cette catégorie de mots inconnus. Il défend ainsi l'idée qu'un système de compréhension automatique des textes peut gérer et mettre à jour ses propres lexiques en utilisant la sémantique extraite du texte qu'il analyse. Il remarque que les textes eux-mêmes peuvent constituer des sources importantes d'informations descriptives et sémantiques sur les noms propres qu'ils contiennent. Un système de traitement automatique doit être capable de détecter et de traiter ce type d'information. Plutôt que d'exécuter des traitements lourds sur les corpus dans le but d'extraire des exemples de noms propres, qui seront ensuite ajoutés au lexique de façon manuel ou automatique, il pense qu'il est préférable de combiner cette acquisition lexicale au traitement lui même. Les avantages de cette démarche sont les suivants :

1 - Le lexique n'est mis à jour que lorsque les mots sont rencontrés,
2 - Le texte est considéré comme une source d'information,
3 - Le système est capable de trouver des solutions partielles pour les mots inconnus.
4 - Si cela est nécessaire, le système peut utiliser l'information extraite du texte pour compléter les entrées du lexique existant.

La façon dont les noms propres sont utilisés dans les articles de presse est représentative de la façon dont cette connaissance peut être utilisée. C'est en se fondant sur cette particularité que FUNES détecte les noms propres inconnus et utilise la sémantique extraite pour créer des définitions partielles.

Les limites des approches fondées sur la consultation des dictionnaires

Coates-Stephen remarque que beaucoup d'approches reposent sur la construction lexicale utilisant les dictionnaires électroniques. Tenter de déduire la signification des mots à partir de dictionnaire est une approche qui pose de nombreux problèmes ; cependant, certains travaux apparaissent prometteurs. Même si on peut envisager qu'un système soit capable de construire des définitions automatiques acceptables pour les mots d'un dictionnaire, cela ne résout pas le problème du traitement qu'il faut appliquer aux mots qui n'y sont pas.

Cette question fait encore l'objet de débats. Cela dépend certainement du type de textes sur lesquels on opère le traitement. L'auteur s'est intéressé aux textes de journaux et plus particulièrement aux dépêches. Il fait remarquer qu'une étude menée par [Walker , Amsler 86] et qui a consisté à comparer les entrées du *Webster's Seventh New Collegiate Dictionary* à un corpus de trois mois de dépêches du *New York Times* a donné les résultats suivants : 64% des mots contenus dans les dépêches n'étaient pas dans le dictionnaire. L'analyse en détail a mis en évidence qu'un quart était constitué de formes fléchies, un quart de noms propres, un sixième de noms composés avec un trait d'union, un douzième de fautes d'orthographes et le reste de cas non résolus parmi lesquels des mots nouveaux apparus depuis la publication du dictionnaire.

Coates-Stephen souligne qu'avec de telles comparaisons, un nombre important de mots composés viennent perturber les résultats. A moins qu'un nom composé puisse être détecté en tant que tel et traité comme un mot simple, chacun des ses éléments sera reconnu comme un mot simple. Face à cette situation, [Amsler 89] a développé un programme pour détecter de tels mots composés et comparer les noms propres provenant d'un corpus d'environ un mois du *New York Times* avec ceux qui apparaissaient dans *The World Almanac and Book of Facts*. Ce dernier ne contenait que 10% des noms propres contenus dans les dépêches. Une analyse détaillée a mis en évidence le fait que de nombreux échecs de reconnaissance venaient du fait qu'un nom propre peut avoir plusieurs formes différentes, une seule d'entre elles étant présente dans le dictionnaire[4].

Par ailleurs, l'auteur fait remarquer que dans une étude différente, [Sampson 89] a examiné les similarités entre un exemplaire du corpus *LOB (Lancaster Oslo/Bergen)* et une version électronique du *Oxford Advanced Learner's Dictionary of Current English* qui a été développée par *Roger Mitton* et qui comprend quelques 2500 noms propres avec une entrée pour chacune des formes fléchies. Le travail de Sampson allait plus loin. Il s'agissait de trouver des formes de base convergentes et à cette fin, beaucoup de problème de reconnaissance liés à la morphologie ont été évités. Toutes les lettres capitales ont été converties en minuscules, tous les mots ayant des caractères non alphabétiques ont été supprimés. De plus, si un mot n'était pas trouvé dans le dictionnaire lors d'un premier passage, plusieurs changements morphologiques étaient effectués et la recherche était répétée avec les formes nouvelles ainsi générées.

---

[4] Il existe des approches phonologiques qui permettent de reconnaître un nom propre quelque soit sa graphie.

Les résultats furent très différents de ceux obtenus par Walker et Amsler : Seulement 3,24% des mots trouvés dans le corpus *LOB* ne se retrouvent pas dans le dictionnaire de *Roger Mitton*. C'est une conclusion très satisfaisante, qui met l'accent sur le fait qu'un dictionnaire constitue, malgré tout, une très bonne base pour la couverture des mots. Cependant, c'est un fait que les mots ainsi extraits des textes n'apparaissent pas dans le format utilisé par Sampson. La plupart des problèmes rencontrés par Walker et Amsler avec les noms propres composés ne peuvent pas être résolus par des heuristiques classiques comme la suppression des traits d'union ou le passage des conventions orthographiques anglaises vers les formes américaines. De plus, étant donné la façon dont Sampson considère les noms propres, le traitement qui consisterait à les ajouter à un dictionnaire serait considéré comme héroïque et relevant d'un travail fastidieux avant que l'un d'entre eux ne puisse être reconnu en pratique.

Enfin, Seitz et Gupta [90] présentent plusieurs statistiques sur leurs travaux de constitution d'un grand dictionnaire construit à partir du *Webster's Seventh New Collegiate Dictionary* (W7), auquel ils ont ajouté les 12753 mots du *Francis and Kucera's Brown Corpus* qui ne sont pas présents dans W7 et 123 mots parmi les 5000 extraits d'articles de journaux. Ils ont ensuite traité un corpus de texte d'environ quatre mois *du Toronto Globe and Mail*, soit environ 10 millions de mots pour compter combien de mots du corpus n'étaient pas présents dans leur dictionnaire : 20 000 mots nouveaux ont été trouvés, dont la moitié commençait par une majuscule.

Coates-Stephen en conclut que les textes provenant des journaux contiennent toujours des mots qui ne figurent pas dans les dictionnaires et qu'une proportion importante d'entre eux est constituée de noms propres.

L'approche de FUNES pour l'acquisition de connaissances

Il y a eu plusieurs essais pour construire des programmes pouvant acquérir de la connaissance sur les mots. La plupart d'entre eux reposent sur la détection des contraintes, celles-ci pouvant être de différents niveaux. Les règles de syntaxe et les contraintes sémantiques telles que les restrictions de sélection induisent souvent la catégorie sémantique d'un mot. Enfin, le contexte peut fournir des hypothèses pour la signification d'un mot inconnu.

Coates-Stephen précise que FUNES est un système qui a été conçu pour se « débrouiller » avec les mots inconnus qui apparaissent dans les texte. Il est capable de contourner le problème de la non-reconnaissance lexicale qui embarrasse la plupart des analyseur classiques, en construisant des définitions partielles pour les mots inconnus qu'il rencontre. Il utilise ces définitions pour mettre à jour son propre lexique. Il traite également les mots contenant des traits d'union ainsi que d'autres formes de difficultés liées au lexique. L'information est utilisée aux différents niveaux de traitement en créant une définition partielle pour tout nom propre inconnu rencontré.

L'architecture de FUNES est la suivante :

Un traitement préalable marque les entrées et réalise un contrôle lexical des mots inconnus. Tout mot qui n'est pas présent dans le lexique est considéré comme inconnu. Pendant le traitement, la détection des mots inconnus ne s'arrête pas à cela. Un nom peut être dans le lexique en tant que nom commun et cependant avoir une signification différente quand il est utilisé en tant que nom propre (*a tornado / a Tornado fighter-bomber*), ou bien un mot peut être connu et constituer un des éléments d'un nom propre (*the Red Cross*). De tels problèmes sont détectés par la suite pendant la phase d'analyse des syntagmes nominaux. Les mots inconnus sont rangés dans une catégorie par l'examen de celle des mots du contexte immédiat. Bien qu'il ait été principalement conçu pour traiter les noms propres, FUNES est capable également de traiter les noms communs, les adjectifs inconnus, et d'une façon moins importante, les verbes.

Un composant syntaxique analyse la sortie du pré-traitement et produit un arbre syntaxique. FUNES possède une grammaire qui assure une couverture satisfaisante. Les différents composants syntaxiques sont analysés ensuite sur le plan sémantique. En particulier, les noms composés font l'objet d'une analyse détaillée, pour les raisons mentionnées plus haut, et parce que l'on trouve parmi eux, une proportion importante de noms propres.

Après le traitement de chaque phrase, une procédure regroupe toutes les informations provenant des noms inconnus de la phrase et crée une *structure de nom* qui constitue une définition partielle pour chacun d'eux.

Coates Stephen rappelle que le but de sa communication est de décrire la nature des constructions utilisées pour les noms propres et de présenter la façon dont celles-ci sont utilisées pour produire des définitions exploitables de façon automatique, avec lesquelles le lexique sera mis à jour.

Nature et analyse des définitions de noms propres

Il semble qu'il y ait un phénomène curieux autour des noms propres, et plus particulièrement de ceux qui sont présents dans les corpus de journaux: ils apparaissent souvent, accompagnés d'éléments constituant des compléments d'information. En général, ils sont composés de deux sortes d'éléments : un élément de description et un élément d'assertion. L'élément de description fournit les acteurs et les lieux, celui d'assertion indique les actions. Ainsi, les articles de presse « fonctionnent » comme le ferait un dictionnaire pour des mots ordinaires. A cet égard, ils peuvent être considérés comme des sources d'information. Les principales constructions utilisées sont présentées par l'auteur de la façon suivante :

Le syntagme nominal apposé

C'est sans doute la forme la plus aisée pour fournir une description d'une personne. L'apposition est généralement constituée de deux syntagmes nominaux séparés par une virgule . On peut relever deux cas. Dans le premier, le groupe nominal de description est situé en apposition au nom propre comme dans « *Saddam Hussein, the President of Iraq*» . Dans le second, le nom propre est en apposition au syntagme nominal de description : « *the President of Iraq, Saddam Hussein* ». La construction est analysée comme une phrase copulative contractée : *Main NP* « *is* » *Appositive NP.* Cependant, la situation peut se compliquer par l'apparition de groupes prépositionnels (*PPs*). Ainsi, un groupe de type « *of PP* » ou bien « *for PP* » peut suivre le syntagme de description. L'apposition peut se rapporter au nom propre ou au syntagme prépositionnel. Par exemple dans la phrase « *A bomb explodes outside the home of the Philippines Justice Secretary, Francis Drillon* », *Francis Drillon* ne se rapporte pas à « *the home* » mais à « *Justice Secretary* ».

Dans les articles de journaux, la plupart du temps, les appositions fournissent de l'information sur les activités. Elles peuvent également indiquer l'âge de la personne, fréquemment, elles donnent les deux comme dans « *Margaret Salmon, 43 , personnel director with the Burton Group* » . Une double apposition peut également fournir deux descriptions différentes comme dans : « *The detention of a Palestinian activist, Sari Nusseibh, a spokesman for Palestinians in the occupied territories who is accused of spying for Iraq* »

Un autre type d'apposition précise un lieu, « *The town of Larnaca, Cyprius* » . Cette construction a une syntaxe particulière : deux noms propres simples apparaissent sans déterminant. D'autres précisions peuvent également être fournies «*The town of San Nicloas, 200 miles north -west of Buenos Aires* ».

Toutes ces constructions sont traitées par FUNES qui construit une structure de cas pour représenter la phrase et une structure de nom pour recevoir la définition partielle qui sera dérivée Quand le système détecte la virgule suivie par un autre groupe nominal, le second est analysé sur le plan grammatical et les deux syntagmes sont traités sur le plan sémantique comme « *is a* ». Un

lien est ainsi construit entre la description du syntagme et le nom propre. Si c'est un mot qui indique le rôle joué comme *president* ou *journalist*, le lien prend le nom de *role*, si c'est un mot indiquant un lieu comme *town*, le lien est simplement *is a*. Une fois ce lien réalisé, l'analyse du groupe nominal de description peut être faite, en isolant du nom propre lui-même toute l'information descriptive. Ainsi, pour l'exemple de *Buenos Aires*, FUNES fournit une construction de la forme :

[location : direction (north-west), distance (measure(mile), number (200)), of ([buenos, aires])].

Les mots clés

Les mots clés sont utilisés pour identifier ce qui précède et ce qui suit un nom propre. Différents types de noms propres auront différents types de mots clés. Dans le cas des noms de personnes, ils indiquent les rôles que celles-ci jouent comme *president*, *foreign minister*, *judge*. Ils peuvent apparaître à la fin d'un syntagme nominal complexe qui fournit déjà beaucoup d'informations sur le nom propre ou bien apparaître comme un élément propre. Dans le premier cas, le groupe nominal est assimilable à une apposition comme dans «*The Senate Armed Service Committee spokesman X*».

Les objets et les lieux peuvent également être détectés par les mots clés de leur contexte. Pour les objets, le mot clé peut précéder ou suivre le nom propre. On peut ainsi trouver « *the battleship Missouri* » ou « *the Missouri battleship* ». Ici, il s'agit d'un exemple de nom propre utilisé avec un sens différent de celui de son référent habituel : le fleuve. Les noms de lieux peuvent également être précédés ou suivis de mot clés : « *the Sahara Desert* », « *the River Nile* »

FUNES détecte les mots clés pendant l'analyse du groupe des noms composés dans lequel ils apparaissent. Il crée une insertion dans la structure de nom, pour ce nom propre particulier. Pour un mot clé, la chaîne du nom composé est recherchée et les noms qui apparaissent à la suite sont définis en tant que « *role* » ou « *is a* » puis reliés au mot clé lui-même. Le type du nom propre est construit en ajoutant [name] au type du mot clé, on obtient ainsi [ship, name] pour « *cruiser Achille Lauro* » . Si le mot clé est un mot de type « *role* », le premier mot est entré en tant que « *firstname* » et le dernier est utilisé pour la suite du traitement. Ainsi une personne qui apparaît au début d'une histoire avec son nom complet, et que l'on désigne par la suite par son surnom, sera toujours identifiée comme étant la même personne. A la fin de chaque phrase, toutes ces définitions sont rassemblées pour former la structure du nom propre concerné. Par exemple, après le traitement du groupe nominal « *President François Mitterand* » FUNES a construit les déclarations *attr(mitterand,role,president)*, *attr(mitterand,firstname, françois)*, et leur a assigné la catégorie sémantique [human, name].

A la fin de l'analyse de la phrase, la procédure de traitement des structures de noms rassemble toutes les assertions sur le mot inconnu dans une structure appropriée. On obtient les éléments suivants :

[noun([françois,mitterand],[human, name], _), role,(president)], [noun(mitterand,[huaman,name],b)],[noun(françois, [human, name],_)].

syntagmes prépositionnel suivant le nom

Ils peuvent fournir de l'information descriptive de différentes façons. Une des constructions particulières est celle qui nécessite un syntagme prépositionnel du type « *of PP* » et un nom de type localisation comme dans « *The rebel Yugoslav republic of Croatia* ». Ici, le syntagme nominal du début fournit la signification du groupe prépositionnel.

Un « *of PP* » peut également indiquer l'employeur comme dans « *Jim Burnett of the National Transportation Safety Board* » ou donner des information sur l'origine comme « *President Carlos Menem of Argentina* ». Quand le nom du début du groupe prépositionnel est inconnu, il n'est pas possible de décider si une relation existe. Un mot de type « rôle » peut donner un indice, car s'il apparaît sans déterminant, la relation est plus probablement du type « origine » (on ne dit pas « *Presient X of General Motors'* »). Certains indicateurs de rôle peuvent induire d'autres relations, « *ditector* » ou « *chairman* » signalent un employeur, alors que « *kind* » et « *governor* », une origine.

Les façons avec lesquelles un syntagme prépositionnel peut qualifier un nom sont très nombreuses. Chaque type de qualification donnera un aspect différent à la définition partielle que FUNES construit pour un nom propre inconnu. Dans beaucoup de cas où le nom inconnu est au début du groupe prépositionnel, la préposition constitue parfois la seule source d'information. FUNES dispose donc de procédures d'analyse de chacun des type de prépositions. Elles utilisent l'information présente dans le contexte immédiat, pour attribuer une catégorie sémantique au mot inconnu. Elles peuvent aussi, selon le contexte, changer la nature de la relation entre le groupe prépositionnel et le nom qu'il qualifie. Ainsi, un syntagme comme « *The city of Basra* » donne la définition *attr(basra, isa, city)* alors que « *President Carlos Menem of Argentina* » donnera *attr(menem, origin, argentina)*

Autres sources d'informations descriptives

Sans doute, la façon la plus évidente de définir un terme nouveau, consiste tout simplement à déclarer qu'il « *est* » quelque chose. On ne trouve pas cela très souvent dans les articles de presse. En revanche, cette forme peut être utilisée quand on introduit un concept totalement nouveau et que l'on veut attirer l'attention du lecteur sur sa définition, par exemple : « *The counter-Scud patrol is a new type of allied air attack* »

Une autre approche consiste à utiliser un lien du type « *known as* ». C'est une méthode qui est souvent utilisée pour définir des mots étrangers qui peuvent apparaître dans le langage commun : « A *policeman, known as omawari-san* », « *A document called a shakoshome* » . Ici, il ne s'agit pas de noms propres et il sont utilisés sans lettre majuscule, FUNES est cependant capable de les analyser et d'utiliser la présence du mot de liaison pour créer une définition partielle.

La morphologie peut également donner des indices sur la catégorie du nom propre. Cependant FUNES ne réalise pas d'analyse en profondeur sur la morphologie des mots inconnus.

Une autre méthode de description des noms propres nécessite l'utilisation de références. La forme la plus commune dans les articles de journaux est l'utilisation d'un groupe nominal descriptif dans la première phrase du texte, suivi par le nom propre lui-même dans la seconde phrase, ou dans celles qui suivent. Le lecteur doit en déduire que la première description renvoie au nom propre nouvellement introduit. Par exemple :

« *A pro-Iraqi Pakistani goverment minister resigned yeseday , giving as one reason the support by the Prime Minister, Nawaz Sharif, for the US in the Gulf War. Abdul Sattar Khan Niaei released a letter he sent to Mr Sharif in which he said that he could not remain in the cabinet* »

L'utilisation des techniques traditionnelles de référence est limitée ici par le fait que le nom propre auquel on devrait se référer est inconnu. Nous n'avons donc aucune information sur le genre, le nombre ou la sémantique qui le caractérise. La technique utilisée par FUNES est applicable au cas que nous venons d'évoquer. N'importe quelle sous séquence d'un nom propre qui apparaît sans description est supposée être un référent potentiel à la description sauvegardée au préalable. Ainsi, dans l'exemple ci-dessus le syntagme « *A pro-Iraqi Pakistani goverment minister* »

sera considéré comme n'ayant pas de lien avec un nom propre dans la phrase où il apparaît. *Nawaz Sharif* est relié à *Prime Minister*, ce nom propre ne constitue donc pas n'est pas un référent potentiel. En revanche, *Abdul Sattar Khan* qui n'est pas décrit, constitue un référent candidat pour le syntagme nominal sauvegardé «*A pro-Iraqi Pakistani goverment minister*» . Cette technique nécessite des tests minutieux cependant elle peut être utilisée en tant que méthode d'acquisition des descriptions de noms propres.

Les problèmes de l'acquisition lexicale

Coates-Stephen conclut en indiquant que jusqu'à un passé récent, le problème de la présence des noms propres n'avait pas été relevé en tant que tel. Aujourd'hui, plusieurs auteurs l'ont mis en évidence. Il note par ailleurs que plus récemment [Nutter, Fox, Evens 90] ont noté, au travers d'une étude, le problème du manque d'informations lexicales dans le traitement des noms propres.

Le principal domaine d'application avec lequel le problème est apparu est celui du traitement automatique des corpus de journaux. Ainsi, le manque de vocabulaire a été cité comme la principale raison des échecs de FRUMP, un analyseur d'articles de journaux. [Kuhns 88] surmonte le problème en utilisant une base lexicale aussi complète que possible et en traitant tous les mots inconnus comme des noms propres.

Or, on n'avait pas suffisamment mis l'accent sur le fait que les corpus d'articles de journaux constituent des sources importantes d'informations lexicales, et qu'à ce jour, ils n'ont pas été suffisamment exploités. C'est cette lacune que le projet FUNES s'est proposé de combler.

1.2.6.3 Reconnaissance, catégorisation et normalisation des noms porpres de Paik, Liddy, Yu et McKenna

Paik, Liddy, Yu et McKenna ont développé un module de catégorisation et de normalisation de noms propres qui est intégré au système de recherche documentaire DR-LINK. Dans un texte, les noms propres représentent une source importante d'information qui peut être utilisée de façon profitable dans ce type d'application.

l'Identification des limites des noms propres

L'identification des limites d'un nom propre à l'intérieur d'un étiqueteur de texte est essentielle à la réussite de la catégorisation. Leur module de classification des noms propres a été conçu pour attribuer un code de catégorie à chaque entité de nom propre en utilisant 30 catégories générées à partir d'une analyse de corpus. Une normalisation des variantes d'un même nom propre s'effectue au cours du traitement.

Dans sa version la plus récente, les auteurs précisent que leur système utilise un étiqueteur probabiliste. Des heuristiques, développées à partir de l'analyse de corpus, sont appliquées pour délimiter les syntagmes de noms propres en incluant dans un même ensemble, les conjonctions et les prépositions. Par exemple, une liste de noms propres sera délimitée en tant que noms propres non adjacents, si « *of* » est une préposition incluse dans l'ensemble.

La proportion de réussite de leur module actuel d'identification de limites de noms propres est de l'ordre de 95 %.

Le plan de catégorisation des noms propres

Paik, Liddy, Yu et McKenna ont construit le plan de catégorisation des noms propres à partir de l'analyse de corpus d'articles de journaux. Celui-ci est organisé de façon hiérarchique avec neuf nœuds au premier niveau et trente nœuds terminaux. D'une façon générale, seuls les noeuds terminaux sont utilisés pour attribuer une catégorie aux noms propres d'un texte. Sur un

ensemble de 588 noms propres trouvés dans un corpus du *Wall Street Journal,* 89 % d'entre eux ont reçu une des 29 catégories, le reste s'étant vu attribué la catégorie « divers ».

Le système réalise cette classification en utilisant plusieurs méthodes. La première consiste à comparer le nom propre à une liste de tous les affixes qui ont été identifiés pour chacune des catégories. Si, à cette étape, une catégorie ne peut pas être identifiée, le nom propre est transmis à une base de données des *alias* afin de déterminer s'il a une ou des formes alternatives. Si c'est le cas, une catégorie lui est attribuée et on détermine une forme normalisée. S'il n'a pas d'équivalent dans la base des *alias*, il est transmis à une base de connaissances. Celle-ci a été construite en utilisant des ressources lexicales « *en ligne* » comme *the Gazetteer; the World Facbase* et *the Executive Desk Référence.* En cas de nouvel échec, on lui applique des heuristiques de traitement du contexte qui vont suggérer certaines catégories. Par exemple, si un nom propre est suivi par une virgule, puis par un autre nom propre qui a été identifié comme étant un Etat, le premier sera rangé dans la catégorie des noms de ville : « *Time, Illinois* ». Enfin, si le nom propre n'a toujours pas de catégorie, il est comparé à une liste de prénoms qui a été extraite d'un corpus de noms de famille. Dans tous les autres cas, il est rangé dans la catégorie « divers ».

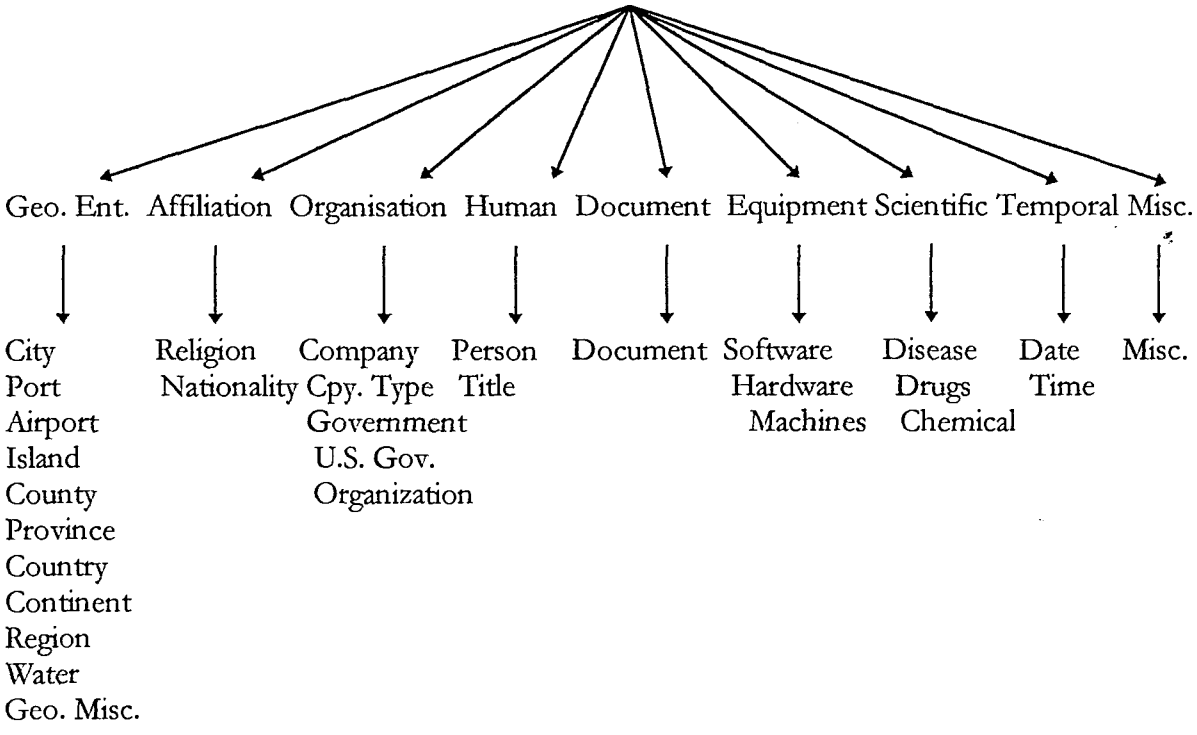


Figure n° 13 : La taxinomie des noms propres de l'article

Les entrées du lexique pour le nom propre ou le code de catégorie peuvent être utilisés pour comparer le contenu des questions au contenu des documents. Par exemple, si une question à trait à une violation de frontière, on peut limiter la liste des documents recherchés à ceux qui contiennent au moins deux noms de pays. L'utilisation de formes normalisées réduit le nombre de variantes possibles que le système a besoin de rechercher.

Pour le traitement des questions, ils utilisent actuellement une base de données de noms propres avec 168 entrées. De plus, pour améliorer les performances de la recherche documentaire, certains noms communs tels que « *socialist countries* » sont détectés de façon particulière. Il s'agit en effet de noms communs ayant des référents multiples, qui sont eux même des noms propres. Il est donc nécessaire, pour les traitements, que le système puisse leur substituer l'ensemble auquel ils renvoient. Actuellement, la base de données de ce type de noms possède 37 entrées.

Dans leur dernière version, Paik, Liddy, Yu et McKenna indiquent que les noms propres catégorisés et normalisés sont combinés à des informations issues d'un module Text Structure. Il s'agit d'un système de reconnaissance de structures de textes, qui décrit l'organisation des différents niveaux d'un document. Ainsi, les recherches peuvent s'appliquer à des sous-ensembles du document qui auront été identifiés comme susceptibles des contenir l'information souhaitée. Tous les noms propres d'une même collection de documents sont indexés dans un fichier inverse avec :

- le numéro d'accès au document,
- le composant Text Structure dans lequel le nom propre a été localisé,
- le code de catégorie.

De plus, la catégorisation des noms propres est souvent utilisée dans les autres modules du système qui sont plus particulièrement chargés d'extraire les concepts et les relations, afin de produire une représentation sémantique plus précise. Par exemple, les noms propres peuvent fournir le lieu d'implantation d'une entreprise ou la nationalité d'un individu.

2. Les noms propres en français

Les noms propres servent à désigner des individus ou des réalités individuelles. Ils véhiculent des données singulières et ont pour fonction essentielle de séparer, distinguer, rendre unique, irremplaçable.

2.1 Le statut particulier des noms propres

En faisant référence à la géographie et à l'histoire, « *Les noms propres renvoient aux trois dimensions de la deixis, la personne, l'espace et le temps* » [Molino, 1982 :19]. Quel est leur statut linguistique ? Marie-Noëlle Gary-Prieur note [Gary-Prieur 1991 :4] que « *... pour le français en tout cas, tout locuteur adulte a une intuition claire de la différence entre Nom Propre et Nom Commun* » alors que [*opus cit.*,p.7] « *La situation des noms propres dans les grammaires peut se résumer de la façon suivante : distingués d'abord des noms communs sur une base sémantique..., ils sont ensuite plus ou moins oubliés dans le chapitre consacré au nom, mais ils réapparaissent comme des cas particuliers sur le plan morphologique... On notera l'absence de toute dimension syntaxique.* »

De son côté, J. Molino remarque avec justesse que le problème des noms propres n'est pas si simple puisque [Molino 1982 :10] : « *Tout peut être un nom propre... une phrase complète... en français : N'a-qu'un-œil* ». Il ajoute ensuite que « *Le nom propre n'a qu'une possibilité minimale de productivité morphologique (morphologie dérivationnelle)* ». Cette affirmation ne tient pas compte de la plupart des noms liés à la géographie qui se dérivent en un nom et un adjectif, ce qui ne semble pas être le cas de tous les noms communs. Alain Rey fait remarquer à ce propos [Rey 1977 :30] qu' « *il faut mentionner l'important problème que pose l'absence des noms propres dans les dictionnaires de langue. En effet, les noms propres fournissent non seulement des lexicalisations (un harpagon) mais des monèmes productifs (marxiste, marxisme, martien). Certes, on trouve américain dans tous les dictionnaires, mais berrichon, puis limougeot et enfin castelpontain... ou encore giscardien, aznavourien.... ne peuvent pas tous figurer dans le lexique décrit* ».

Quels sont les éléments qui permettent de déterminer de façon claire que l'on a affaire à un nom propre ? Comme le fait remarquer Maurice Gross [Gross,1990], «*...la division entre la partie linguistique d'un dictionnaire et sa partie encyclopédique n'est pas simple à effectuer : par exemple, si le nom de pays France appartient à la partie encyclopédique, il devrait en être de même pour les noms de citoyens de pays comme un(e) Français(e), qui eux aussi comportent une majuscule, critère souvent retenu. Mais le nom de langue,*

le français et ses composés (ancien français, moyen français), l'adjectif français, le préfixe franco appartiennent plutôt à la partie langue »

En renvoyant à un concept, une idée générale, le nom commun se définit en compréhension. Le nom propre qui a en général un référent unique relève plutôt du domaine de la description. Il se définit en extension.

A la lecture d'un texte, s'il faut « **comprendre** » les mots qui sont utilisés, il est nécessaire « **connaître** » les noms propres. Ainsi, au cours d'une conversation, l'utilisation de noms propres présuppose qu'une connaissance générale de l'univers de référence soit commune aux interlocuteurs. Les noms propres employés doivent, pour chacun des participants être associés aux mêmes référents (lieu, personne, événement...). Comme le fait remarquer [M. N.Gary-Prieur 1991 :11] « *Des noms propres qui apparaissent constamment dans nos discours, et qui sont ceux de nos amis, parents ou collègues, ne requièrent pas le recours au dictionnaire, mais à un principe général de conversation qui veut qu'on n'emploie un nom propre que si on sait qu'un destinataire connaît son porteur* »

Cependant, on remarque que cette connaissance peut admettre des degrés. Si le nom d'une localité est cité, on peut ne pas être capable de la situer géographiquement. S'il s'agit d'une personne, il peut se trouver que l'on n'ait que des connaissances succinctes sur son identité, son statut ou ses activités. Cela n'empêche pas de pouvoir « suivre » la conversation ou comprendre le sens général du texte, dès lors que nous aurons perçu qu'il s'agit d'une part d'une localité et d'autre part d'une personne.

2.2 La reconnaissance

Un lecteur prenant connaissance d'un texte comportant de nombreuses références à des personnes et à des lieux, peut comprendre ce qu'il lit même sans avoir une idée précise du référent de chaque nom propre utilisé. La littérature fantastique, offre des exemples d'univers entièrement « artificiels » dans lesquels les personnages et les lieux ne renvoient pas à un « fond commun de culture générale » . Cela empêche-t-il le lecteur d'accéder à un premier niveau de compréhension du texte.

Voici deux extraits du roman fantastique de John Ronald Reuel Tolkien « *Bilbo le hobbit* ». Les personnages ne sont ni des humains ni des animaux, l'univers dans lequel ils évoluent est entièrement « reconstruit ». Il est constitué de localités, de lieux et de contrées qui sont le fruit de l'imagination de l'auteur et n'ont aucun rapport même lointain avec des localités, des lieux ou des contrées existants ou ayant existés. Un extrait comportant un nombre important de noms propres est-il intelligible, sachant qu'il cumule deux difficultés :

- c'est un extrait, il est donc par définition sorti de son contexte,
- le contexte renvoie à un univers totalement imaginé c'est à dire qu'il ne comporte aucun repère permettant au lecteur de situer les personnes, les lieux, les objets les événements...

« Au nord du **Carrock**, l'orée de **Mirkwood** se rapprochait des bords de la **Grande Rivière** et, bien qu'à cet endroit les montagnes descendissent plus près, **Beorn** leur avait conseillé de prendre ce chemin ; car, à quelques jours de chevauchée en plein nord du **Carrock**, se trouvait l'entrée d'un sentier qui traversait **Mirkwood** et menait presqu'en droite ligne vers la **Montagne Solitaire**.»

Bilbo le hobbit page 166

Malgré le double handicap signalé plus haut, nous sommes capables de placer : **Carrock, Mirkwood, Grande Rivière, Montagne Solitaire** dans la catégorie des « *noms de lieux* », et **Beorn** dans celle des « *noms de personnes* ». Pour le lecteur, cette détection des types de noms

propres a été facilitée sur le plan visuel par la présence de majuscules et au niveau sémantique par la détection des grammaires locales :

Au nord du...., l'orée de..., des bords de la..., en plein nord du..., qui traversait..., vers la...

Dans l'extrait suivant, un lecteur rangera la totalité des noms propres cités dans la catégorie des « *noms de personnes* »

« -qui va passer le premier demanda **Bilbo** ?
- Moi dit **Thorïn**, et vous viendrez avec moi, ainsi que **Fili** et **Balïn**. C'est tout ce que la barque peut contenir en une fournée. Après ce sera **Kili, Oïn, Gloïn** et **Dorin** ; ensuite, **Orin** et **Nori, Bifur** et **Bofur** ; et enfin **Dwalïn** et **Bombur.** »

Bilbo le hobbit page 179

Ce qui semble possible pour des textes dans lesquels les noms propres ont des référents plus ou moins « exotiques » le sera également lorsque l'univers de référence est plus proche du « fond commun culturel ».

La phrase suivante, extraite de « La guerre Picrocholine , Gargantua, Chapitre XXV de François Rabelais » est compréhensible même pour un lecteur n'ayant pas ou peu d'information sur l'univers Rabelaisien :

« Puis les fouaciers aidèrent à monter **Marquet,** qui était vilainement blessé, et retournèrent à **Lerné** sans poursuivre le chemin de **Parillé,** menaçant fort et ferme les bouviers, bergers et métayers de **Seuillé** et de **Sinais.** »

Si l'on substitue les types aux occurrences de noms propres, on met en évidence un premier niveau de compréhension qui se situe au delà du simple déchiffrage permettant une certaine compréhension globale de la situation décrite:

« Puis les fouaciers aidèrent à monter ***[personnage]***, qui était vilainement blessé, et retournèrent à ***[lieu ou localité]*** sans poursuivre le chemin de ***[lieu ou localité]***, menaçant fort et ferme les bouviers, bergers et métayers de ***[lieu ou localité]*** et de ***[lieu ou localité]*** . »

Pour un lecteur humain, il existe donc deux niveaux de reconnaissance pour un nom propre, ceux-ci n'étant pas exclusif l'un de l'autre. En reprenant la terminologie de Mc Donald, nous les définissons de la façon suivante :

• La **reconnaissance externe** : Le nom propre n'est pas « connu » mais le graphisme (majuscule) et la présence d'une grammaire locale induisent le type du nom propre ou précisent celui-ci en cas d'ambiguïté.

• La **reconnaissance interne** : Le nom propre est « reconnu » parcequ'il était « connu » . Pour un lecteur humain, il appartient à l'univers commun des connaissances (Louis Pasteur..., Claudie..., la Loire..., Paris..., la Charente-Maritime....). Pour un analyseur automatique, il fait partie des données stockées.

On notera à ce sujet que ces phénomènes peuvent varier de façon importante d'une langue à l'autre. Ainsi, en allemand, la présence de majuscules sur tous les noms entraîne une dilution du signe graphique de reconnaissance des noms propres. En anglais, il existe un phénomène de diffusion de la majuscule de part et d'autre du nom propre qui peut perturber la détection.

Nous verrons par la suite que la reconnaissance et le traitement informatiques des noms propres nous placent dans une problématique très différente de la reconnaissance humaine. Ainsi, en français, la reconnaissance automatique des mots commençant par une majuscule génère une tel niveau de bruit qu'elle pose plus de problèmes qu'elle n'en résout. Nous présenterons plus loin des résultats obtenus avec cette technique, sur un corpus d'articles du journal Ouest-France.

2.3 Le Typage des noms propres

La détection des types constitue une première étape dans le processus de reconnaissance et de traitement. Il est donc indispensable de mettre en place un typage des noms propres aussi précis que possible. Il existe également deux autres raisons qui militent en faveur de cette démarche :

• Comme nous essaierons de le démontrer, une partie importante du traitement que nous souhaitons mettre en œuvre au niveau sémantique consiste à détecter les coréférences entre les noms propres d'un texte en reconstruisant un réseau relationnel sur les types.

• Par ailleurs, une partie importante des ambiguïtés rencontrées en analyse automatique est liée à l'homonymie. Un même nom propre peut renvoyer à des types différents (la Loire : type entité hydrographique et type département). La levée de cette ambiguïté ne pourra s'effectuer que par un traitement des coréférences entre types ou par l'utilisation d'une grammaire locale.

2.3.1 Les différents types de noms propres

2.3.1.1 Les personnes

Il existe deux façons de nommer les individus. La première, et de loin la plus générale est constituée de l'ensemble des noms individuels. La seconde les désigne au travers de leur appartenance à un groupe.

2.3.1.1.1 Les noms individuels

Nous rangeons dans cette catégorie tous les personnages célèbres du fait de leur participation à l'une des activités humaines suivantes :

- histoire politique, économique et sociale, militaire et religieuse,
- histoire des connaissances,
- histoire de l'art au sens large, littérature, arts plastiques, musique, spectacle , architecture ...

Il faut également considérer les personnages imaginaires ou mythiques qui, du fait de la place qu'ils occupent dans l'inconscient collectif, dépassent parfois largement le cadre culturel ou artistique dans lequel ils ont été conçus (Dom Juan, Tartuffe, Madame Bovary).

2.3.1.1.2 Les noms collectifs

Nous distinguerons deux catégories principales :

Noms collectifs liés à la géographie

On y trouve principalement les gentilés (noms d'habitants de lieu) que nous avons rangés selon leurs entités géographiques de référence :

• localités (agglomération, ville, village, etc...),

- départements,
- régions administratives,
- régions historiques ou géographiques,
- pays,

Noms collectifs liés à l'histoire et à la civilisation

Dans ce cas, le nom propre désigne l'appartenance à un groupe généralement de type politique, religieux ou philosophique. L'importance, la notoriété, l'influence sur la « marche du monde » font souvent évoluer certains noms de groupes du statut de nom propre vers celui de nom commun. Ainsi protestants, musulmans ne figurent pas dans le Robert des Noms Propres alors qu'on y trouve Cathares. On notera certains cas particuliers comme Musulmans avec une majuscule qui désigne une des trois populations qui constituaient l'ex-Yougoslavie avec les Serbes et les Croates [Catherine Samary, Le Monde Diplomatique, Octobre 95]:

> *« La répression de la propagande islamiste (il y eut interdiction du voile en 1950), dont M. Alija Izetbegovic fit les frais (11), alla de pair avec la reconnaissance d'une communauté ethnico-nationale musulmane distincte de la religion : à côté des autres nations, celle des Musulmans (avec majuscule)[5] fut inscrite dans la Constitution de 1974.»*

2.3.1.2 L'espace

Nous distinguerons deux types de noms propres liés à la désignation de l'espace :

- les noms de lieux liés à la présence et à l'activité humaine,
- les noms décrivant les différentes entités de la géographie physique.

2.3.1.2.1 Les noms de lieux

On y trouve tous les noms de localités quelle que soit leur importance : villes, métropoles, agglomérations, villes moyennes, villages, bourgs auxquels il faut ajouter les sites historiques ou géographiques en relation avec l'histoire, la civilisation (*les Alignements de Carnac, la Grotte de Lascaux, le Mont-Saint-Michel* ...).

2.3.1.2.2 Les entités géographiques

Il s'agit des noms propres désignant les objets de la géographie physique : montagnes, vallées, plaines, mers, océans, baies etc... . On doit, par ailleurs, faire une place à part aux entités hydrographiques qui sont particulièrement nombreuses en France et présentent une organisation de type hiérarchique.

2.3.1.3 Les événements

Parfois formés à partir de noms communs (*la Révolution, la Deuxième Guerre Mondiale*) ces noms font référence à des notions de lieu et de date (*la Saint-Barthélémy, Mai 68*).

2.3.1.4 Les œuvres, réalisations et productions

On y trouve les noms attachés à toute la production humaine. L'ensemble est donc assez hétérogène puisque l'on peut y voir cohabiter *Hamlet, A la Recherche du Temps Perdu, le Figaro, le Joconde, le Beaujolais...*

---

[5] Dans le texte

2.3.1.5 Les entreprises

L'étude de corpus de journaux met en évidence le nombre important de noms d'entreprises. Marqués d'une majuscule et ayant un référent en général unique, ces noms appartiennent bien à la catégorie des noms propres. Cependant, les entreprises apparaissant et disparaissant au gré des conjonctures économiques, leurs noms constituent une catégories d'informations périssables dont la collecte et la gestion pose, sans doute, des problèmes de mise à jour considérables.

2.3.1.6 Les sigles

Leur nombre n'a cessé d'augmenter. Leur utilisation abusive est parfois l'objet de commentaires ironiques. Ils possèdent tous les attributs des noms propres, tout en posant des problèmes spécifiques. Ainsi, le fait pour beaucoup d'entre eux d'avoir des référents multiples nécessite de disposer d'informations pragmatiques pour lever les ambiguïtés qui accompagnent leur utilisation : dans un contexte à forte connotation économique ou commerciale le sigle **C.I.O** renvoie au **Crédit Industriel de l'Ouest**, évoqué dans un cadre scolaire, il désignera le **Centre d'Information et d'Orientation**, un sportif entendra **Comité International Olympique** ! Chacun a le souvenir de confusions au cours d'une conversation, quand des sigles sont utilisés dans un contexte qui est étranger à notre univers habituel de référence.

2.3.2 Un dictionnaire des noms propres ?

La question importante de la mise en œuvre d'un dictionnaire électronique des noms propres est donc posée. Mais quelles limites fixer à son caractère encyclopédique qui est par nature illimité ? Un dictionnaire des noms propres repose sur la notion de notoriété. Qui confère la notoriété ? Dans ce domaine, comment distinguer la mode passagère de la célébrité durable ? Il y a là un problème de choix qui est loin d'être trivial... Comment faire en sorte que les critères de sélection soient le moins subjectifs possible ? Ceux-ci varient souvent en fonction du temps et de l'espace. La notoriété n'est pas proportionnelle à l'influence réelle sur « la marche du monde ». Elle n'est pas fondée sur des critères moraux ou éthiques. En ce sens, les noms propres ont une vie, ils entrent et sortent des dictionnaires au gré du temps.

2.3.3 Le cadre de notre étude

Notre étude et les traitements informatiques qui y sont associés se limite **aux noms de lieux et à leurs gentilés avec toutes leurs flexions**. Les noms de lieux regroupent des entités très diverses liées à la géographie physique, économique et politique ainsi qu'à l'histoire . Il était donc nécessaire d'affiner le typage présenté dans le paragraphe précédent. D'autres part, nous nous sommes vite rendu compte que le système des coréférences entre noms propres est beaucoup plus dense qu'il n'y paraît . En effet, si dans un premier temps l'association entre le nom d'une localité et celui de ses habitants semble évident, on remarque vite que la localité renvoie au département, à l'hydrographie, à la région historique ou géographique qui possèdent également leurs propres gentilés, aux événements importants qui s'y sont déroulés, aux personnages célèbres qui y sont nés, y ont vécus .....

L'enquête nationale sur les noms propres que nous avons lancée en 1995 avec l'aide du journal Ouest-France, de la direction régionale de La Poste et du CNAM de Nantes et dont nous détaillerons les résultats plus loin, n'a fait que confirmer l'importance du phénomène des coréférences et de la sémantique qui y est attachée.

Aussi, tout en maintenant les limites de notre étude, il nous a paru judicieux d'anticiper d'ores et déjà la structure générale d'un dictionnaire relationnel en construisant une base de données des noms propres. Celle-ci nous permet, indépendamment des traitements à mettre en

œuvre dans le cadre d'une analyse automatique, un stockage rapide des données collectées (formes canoniques et formes fléchies), ainsi que de leurs associations.

2.4 Caractéristiques des toponymes et de leurs gentilés

Selon Jean Longnon[6] dans le **Dictionnaire National des Communes de France** « *Les lieux habités de France atteignent peut-être le nombre d'un million : villes, bourgs, écarts, fermes, moulins, maisons isolées* » . C'est dire que les 36664 noms de communes et les 28000 noms de lieux-dits recensés dans cet ouvrage sont très loin d'épuiser l'immense vocabulaire de la toponymie. Ainsi, après avoir précisé des limites au niveau des types de noms propres traités, il était nécessaire de définir des critères objectifs permettant de circonscrire les ensembles qui feraient l'objet de traitements informatiques. Nous détaillerons nos choix dans la partie consacrée à la collecte et à la validation des données.

2.4.1 Toponymes

La toponymie est une discipline auxiliaire de l'histoire qui étudie l'étymologie des noms de lieux. Il n'est pas dans notre propos de présenter l'ensemble de cette discipline. Cependant, il nous a semblé intéressant, après avoir proposé un typage des noms propres et parmi eux, des noms de lieux, d'exposer la classification par origine de Jean Longnon. Il en donne les détails dans « *Ce qu'évoquent les noms de lieux* » qui préface le dictionnaire déjà cité.

En effet, parallèlement aux recherches que nous présentons, nous nous sommes intéressés au problème de la détection de règles de formation des gentilés à partir des noms de lieux [Maurel 1995]. Ces règles pourraient en effet constituer une alternative intéressante à un échec de la reconnaissance d'un gentilé lors d'un processus de traitement automatique. Or, les travaux de [Eggert 1996] ont mis en évidence la nécessité de compléter une étude purement dérivationnelle par des informations issues d'autres disciplines comme la géographie, l'histoire et parmi lesquelles la toponymie devrait représenter une source d'informations complémentaires non négligeable.

2.4.1.1 L'environnement et la nature

Beaucoup de toponymes font référence à l'aspect du lieu où ils sont situés: ***Mont, Val, Rivière, Bois, Forêt, Roche, Etang, Lac...*** Pour introduire des éléments de distinction on ajoute :

- le nom d'un patron : *Mont-Saint-Jean, Mont-Saint-Martin, Mont-Saint-Michel*
- une région ou une situation : *Mons-en-Laonnois, Mont-sous-les-Côtes*
- une épithète : *Clermont, Montaigu, Montfort*

A l'opposé du ***Mont***, le ***Val*** engendre des phénomènes identiques : *Val-Saint-Germain, Laval, La Vallée-aux-Loups, Vaucluse* (val fermé). La ***vallée*** renvoie de façon assez naturelle à la notion de ***cours d'eau***, ce qui produit des toponymes comme: *Belle-rive, Haute-Rive, Ribaute, Rivesaltes, Moisdon-la-Rivière.*

Bois et ***Forêts*** sont à l'origine de : *Bois-le-Roi, Boiscommun, Bois-Colombes, Bois-de-Céné , Achère-la-Forêt, Apremont-la-Forêt, Forêt-la-Folie,* avec les formes plus proche du latin *boscus : Bosc-Geffroy, Bosc-Roger, Bosc-Bénard-Commin, Bosc-Edeline .*

Roche avec sa forme ***Roque*** se retrouvent dans : *Rochefort, Roquebrune, La Roche-Bernard, La Rochefoucauld.*

---

[6]Conservateur honoraire de la Bibliothèque de l'Institut de France

Des noms de localités sont formés à partir de noms d'***arbres***. Le pommier occupe dans la toponymie française une place tout à fait privilégiée : *Pommier, Pommiers, Pommeraie, Pommereau, Pommeret, Pommereux, Pommereuil, Pomerol* . D'autres noms sont formés à partir de poirier qui donne *Poiriers* mais également *Périer, Périers* ; de prunier : *Pruniers, Prunières.* D'autres arbres sont facilement identifiables dans les toponymes suivants : *Cerizy, Coignières, Chenay, Fresnaie, Charmes, Charmois, Theil, Tilleul, Tillay.*

L'eau est à l'origine de *Belleau, Longueau, Morteau* avec les formes dialectales : *Eve, Mortève, Aix, Ax, Dax*. Mais on trouve également des noms formés à partir de ***cris d'animaux*** ou de ***chants d'oiseaux***: *Chanteloup, Canteleu, Chantecoq, Chantepie.*

2.4.1.2 L'histoire

Beaucoup de toponymes font références à l'histoire. Ils constituent une mémoire sédimentaire du passage ou de l'installation des différentes peuplades venues par la mer ou par les terres. Ainsi, *Phéniciens, Grecs, Ligures, Ibères, Celtes et Romains* sont à l'origine d'un nombre important de noms de lieux.

Cependant, c'est le *christianisme* qui a le plus fortement marqué la toponymie française. Ainsi, sur les 38198 noms de lieux présents dans notre base de données, 4995 comportent le mot ***Saint*** soit 13,07 %, avec une mention particulière pour ***Saint-Martin*** qui apparaît dans 284 noms de localités.

2.4.1.3 Le développement des populations

Le développement important de la population à partir du XI^{ème} siècle provoque la création de nouvelles localités qui vont souvent prendre pour nom : *Villeneuve, Neuville...* Elles bénéficient souvent de franchises pour favoriser leur peuplement, ce qui donne naissance à toutes les *Villefranche* ou *Franqueville* dont il faudra parfois différencier les homonymes par des adjonctions telles que *Villeneuve-le-Roi, Villeneuve-l'Archevêque.*

2.4.1.4 L'industrie

L'activité humaine, industrie ou artisanat est également à l'origine de certains noms de lieux. *Fabrèges, Faverges, Fargues* font référence aux forges, *Ferrières* aux mines de fer, *Argentières* aux mines d'argent. On trouve également de nombreux *Verrières, Tuillières.* Enfin *Moulins,* qui renvoie à la minoterie ou à la fourniture d'énergie éolienne est présent dans 52 toponymes.

2.4.2 Morphologie des toponymes et de leurs gentilés

En considérant les appellations de certaines localités, on est parfois étonné de la forme que prend le nom des habitants. Certes, on tombe sur des désignations tout à fait simples et régulières, comme *Parisien,* mais on trouve également *Bellifontains* ou *Dionysiens,* où le lien entre toponyme et gentilé n'est pas évident. A l'inverse également, qui saurait dire tout de suite comment s'appellent les habitants de *Lons-le-Saulnier* ou de *Charleville-Mézières* ? Le Quid en a fait un jeu et renvoie le lecteur à sa petite liste de ces noms d'habitants.

2.4.2.1 Morphologie flexionnelle des toponymes

Les noms propres de lieux n'ont pas de flexion. Il est cependant nécessaire de leur attribuer un genre et un nombre. Mis à part les noms de villes, presque tous les noms propres liés à la géographie admettent la présence d'un déterminant. Certains noms de localité semble se construire avec un déterminant comme *Le Mans, La Roche-sur-Yon* ou *L'Aiguillon.* En fait, il s'agit de formes figées. Notons cependant que pour les noms de lieux commençant par *Le* ou par *Les,* le déterminant s'efface au profit de *au, aux, du, des,* déterminants ordinaires perdant la marque de

la majuscule, et donc signalant ainsi qu'ils n'appartiennent pas à la forme canonique. Ainsi, on aime *Le Longeron*, parce qu'on est né au *Longeron*. On passe ses vacances aux *Sables-d'Olonne* parce que l'on apprécie particulièrement *Les Sables-d'Olonne* l'été.

Ce phénomène est également souligné par la pratique administrative qui veut que l'on indique sur les documents officiels:

né à : *Le Longeron*
réside à : *Les Sables-d'Olonne*

En analyse automatique, il sera donc indispensable de pouvoir « reconnaître » les deux formes.

Pour les déterminants élidés ou pluriels, une recherche du genre est indispensable. Celle-ci peut se réaliser par la détection de modifieurs attestés dans un autre nom propre comme dans les *Hautes-Alpes*, ou dans l'examen des constituants de noms composés comme dans les *Pays de la Loire* [Maurel :95]. Certains noms de régions qui correspondent à des décompositions hétérogènes du point de vue du genre et du nombre comme *Midi-Pyrénées* ou *Provence-Alpes-Côte-d'Azur* peuvent être considérés comme ayant le même genre et le même nombre que le premier nom qui les composent.

Concernant le genre qu'il faut attribuer aux noms de localités, Grevisse note à ce sujet que *« l'usage est tout à fait flottant »* [Grevisse 1982 :549]. Des noms précédés d'un déterminant pluriel comme *Les Sables-d'Olonnes* admettent à la fois le singulier et le pluriel.

Cet *« usage flottant »* est mis en évidence par l'étude du phénomène de modalisation du nom propre [M. N. Gary-Prieur 1991 :46] . Comme l'auteur le précise « *La modalisation est généralement rapportée à l'énoncé, au verbe, à certains adverbiaux, tandis que le Nom Propre apparaît comme ce qui assure la permanence de l'individu qui le porte. [...] le Nom Propre représente dans le langage un certain paradoxe : il s'applique en permanence à un individu, alors que l'individu est ce qui change constamment.* »

L'article de Marie-Noëlle Gary-Prieur a pour objet principal de mettre en évidence qu' « *employé avec l'article indéfini et une expansion, le Nom Propre permet de dire à la fois la permanence et le changement* ». Cependant, les exemples concernant les noms de lieux qu'elle cite et qu'elle commente, illustrent bien le phénomène du changement de genre :

- *Paris est la capitale de la France.*
- *Nous avancions, main dans la main, dans les rues d'un Paris ensoleillé.*

Dans le premier exemple Paris désigne de façon permanente la ville qui est la capitale de la France ? Dans le second, il s'agit d'un « épisode de Paris ».

Certaines constructions métaphoriques et les appositions font également varier le genre. Dans celles où un nom de lieu réfère à un autre, le genre de l'un et de l'autre sont induits par le plus explicite des deux, par exemple:

Bruges, Venise verte du nord
Lyon, capitale des Gaules

induisent le féminin, alors que,

Bruges, carrefour fluvial de la Belgique
Lyon, centre nerveux de la vallée du Rhône

induisent de façon inverse un masculin.

Des phénomènes comparables existent concernant la détermination du nombre. [Maurel, Courtois 1993] remarquent que :

« Le nombre associé à un nom de lieu est d'une manière générale le singulier. La question se pose pour les noms précédés du déterminant « Les » . En principe, ils s'accordent comme un pluriel :

- *Les malouines sont revendiquées par l'Argentine*
- *Les Alpes sont partagées entre plusieurs pays »*

2.4.2.2 Morphologie flexionnelle des gentilés

Les gentilés appartiennent aux mêmes catégories flexionnelles que les noms communs. Le modèle de flexion est le suivant :

Masculin Singulier, Féminin Singulier, Masculin Pluriel, Féminin Pluriel

Pour un nom d'habitants dont le masculin singulier et le masculin pluriel ont la même forme on notera la valeur zéro en position 1 et 3. (voir l'exemple : Nantais figure 14)

En analyse automatique, la reconnaissance de certaines formes fléchies va engendrer des ambiguïtés. Ainsi, lors de la détection hors contexte du gentilé *Nantais,* le transducteur ne peut émettre que le genre (masculin), le nombre étant ambigu tant qu'il n'a pas été précisé par une analyse complémentaire du contexte gauche/droite. Il est donc nécessaire, en plus des quatre flexions habituelles (MS, FS, MP, FP) de prévoir la génération de flexions mixtes dans toutes les formes recensées qui comportent deux valeurs vides, c'est à dire : **0,e,0,es** et **0,0,s,s**. Dans le premier cas, l'ambiguïté porte sur le nombre : les formes MS et MP n'étant pas déterminées, il est nécessaire de générer une forme M(S+P). Dans le second cas (*Briviste*) c'est le genre qui pose problème, on doit donc produire (M+F)S.

N°	Catégorie	Exemples de gentilés	Localités
1	0,e,0,es	Nantais,Nantaise,Nantais,Nantaises	Nantes
2	0,ne,s,nes	Parisien,Parisienne,Parisiens,Parisiennes	Paris
3	0,e,s,es	Toulousain,Toulousaine,Toulousains,Toulousaines	Toulouse
4	0,0,s,s	Briviste, Briviste, Brivistes, Brivistes	Brive
5	0,sse,s,sses	Druyde,Druydesse,Druydes,Druydesses	Druy-Parigny
6	er,ère,ers,ères	Abondancier,Abondancière,Abondanciers,Abondancières	Abondant
7	et,ète,ets,ètes	Bussenet,Bussenète,Bussenets,Bussenètes	Bussang
8	u,lle,us,lles	Millassou,Millasolle,Millasous,Millasolles	Millas
9	au,lle,aux,lles	Manceau,Mancelle,Manceaux,Mancelles	Le Mans
10	l,le,ux,les	Martégal,Martégale,Martégaux,Martégales	Martigues
11	iaux,elle,iaux,elles	Merviandiaux,Merviandelle,Merviandiaux,Merviandelles	Mervans
12	c,que,cs,ques	Baussenc,Baussenque,Baussencs,Baussenques	Les Baux-de-Provence
13	0,te,s,tes	Angluriot, Angluriotte, Angluriots, Angluriottes	Anglure
14	0,le,s,les	Roquerol, Roquerolle,Roquerols, Roquerolles	La Roque-sur-Cèze
15	ux,lle,ux,lles	Vigeoyeux, Vigeoyelle, Vigeoyeux Vigeoyelles	Vigeois

Figure n° 14 : Tableau des catégories flexionnelles recencées

Quand il existe, l'adjectif correspondant est en général identique au gentilé, sans la lettre majuscule.

2.4.2.3 Morphologie dérivationnelle

La construction des gentilés est en général un processus dérivationnel classique, où le noms des habitants d'une ville se construit régulièrement à partir du nom de lieu. Nous présentons figure 15 la répartition des suffixes relevés sur un corpus de 2757 gentilés [Egert, Belleil, Maurel 96 : 6] :

suffixes	nombre	pourcentage	suffixes	nombre	pourcentage
-ois	995	36,1 %	-aire	9	0,3 %
-ais	691	25,1 %	-iste	8	0,29%
-ien	516	18,7 %	-at	8	0,29%
-éen	110	3,9 %	-ar	6	0,22%
-in	90	3,3 %	-asque	6	0,22%
-ain	87	3,2 %	-enc	6	0,22%
-en	37	1,3 %	-ol	6	0,22%
-on	24	0,9 %	-ant	5	0,18%
-ard	20	0,7 %	-and	4	0,15%
-ot	20	0,7 %	-ate	3	0,11%
-an	16	0,6 %	-ite	2	0,07%
-aud	12	0,4 %	-iote	2	0,07%
-ier	12	0,4 %	autre suffixe	52	1,9 %
-aux	11	0,4 %			

Figure n° 15 : Tableau de répartition statistique des suffixes de gentilés

Les noms d'habitants sont aussi souvent le résultat d'une dérivation ou d'une construction complexe à partir d'une forme dite *supplétive*. Pour A. Adouani [Adouani A. 1993:89], *"Une paire de mots est ici dite supplétive si ses deux membres sont liés entre eux par une relation dérivationnelle dont la partie sémantique est régulière mais dont la partie formelle est, soit inexistante, soit profondément altérée"*. Cet ensemble est constitué de mots éloignés du toponyme ou étymologiquement apparentés ou encore dont l'un des composants a une forme latinisée ou réduite. Ainsi les habitants de ***Saint-Etienne*** s'appellent les ***Stéphanois***. Dans la tradition chrétienne, Etienne fut le premier à recevoir la « couronne du martyr ». Il reçut alors le surnom de « couronné » *(stéphanos)*. D'une autre façon, les habitants de ***Saint-Dizier*** doivent à leur courage leur nom de ***Bragards***, François 1[er] les aurait qualifiés de *braves gars* en considérant la résistance et la détermination dont ils firent preuve pendant le siège qu'il fit à cette ville.

2.4.2.4 Les graphies multiples

Un nom de lieu peut s'écrire de plusieurs façons différentes. Comme il ne s'agit pas de variantes orthographiques, qui existent par ailleurs, nous parlons de *variantes graphiques*. Ce phénomène touche certains éléments du nom qui peuvent (ou non) être notés en abrégé. Nous avons regroupé figure 16 les formes recensées en les illustrant d'exemples .

Ces abréviations peuvent se combiner entre elles, ce qui est souvent le cas pour les noms de lieux particulièrement longs :

Saint-Etienne-de-Mont-Luc → *St-Etienne-de-Mt-Luc*

A cela s'ajoute la présence ou non du tiret de normalisation de La Poste. Prenons l'exemple d'une commune de l'agglomération nantaise dont l'orthographe normalisée est :

Pont-Saint-Martin

En combinant toutes les possibilités de graphies de *Pont* en *Pt* et de *Saint* en *St*, auxquelles on ajoutera la présence ou l'absence des tirets, on obtient 16 graphies différentes. Bien entendu, il est indispensable qu'en analyse automatique, celles-ci soient toutes « reconnues ». Comme nous le verrons dans le chapitre 6 qui traite de la mise en oeuvre, il existe des outils informatiques particulièrement bien adaptés à la résolution de ce problème.

Terme	Abréviation	exemples
Saint	St	St-Etienne
Sainte	Ste	Ste-Geneviève-des-Bois
Grand	Gd	Le Gd-Quevilly
Grands	Gds	Les Plains-et-Gds-Essarts
Grande	Gde	La Gde-Motte
Grandes	Gdes	Les Gdes-Ventes
Pont	Pt	Pt-l'Evêque
Ponts	Pts	Cubzac-les-Pts
Mont	Mt	Saint-Cyr-au-Mt-d'Or
Monts	Mts	Saint-Jean-de-Mts
sous	s	Aulnay-s-Bois
sur	/	Neuilly/Seine
-	{espace}	

Figure n° 16 : Tableau des principales variantes graphiques

2.4.3 Les ambiguïtés liées à l'homonymie

Nous envisageons de traiter trois types d'ambiguïtés liées aux noms de lieux: l'homonymie dans un même type, l'homonymie par élision , l'homonymie dans des types différents

2.4.3.1 L'homonymie dans un même type

Des localités, des personnes, des entités géographiques peuvent porter le même nom. Il est donc nécessaire de lever les ambiguïtés liées à ces situations. Par exemple, il existe en France dix sept localités portant le nom de *Saint-André*. Deux sont situées en Savoie et deux dans le Tarn et pour les autres dans les départements suivants : Alpes-Maritimes, Aude, Bouches-du-Rhône, Calvados, Charente, Haute-Garonne, Gers, Hérault, Morbihan, Nord, Pyrénées-Orientales, Haut-Rhin, Seine-Maritime.

On compte seize localités portant le nom de *Laval*, auxquelles on peut ajouter les dix-huit dont le nom est formé à partir du même nom comme *Laval-Atger* en *Lozère* ou *Laval-d'Aix* dans la *Drome*. Les noms *La Vallée* et *Les Vallées* renvoient à vingt deux localités différentes.

2.4.3.2 L'homonymie par élision

De nombreux noms de localités sont des noms composés. Ainsi, vingt trois noms composés sont formés à partir de *Neuilly*, et soixante huit à partir de *Neuville*. A part quelques exceptions (*Neuville-aux-Bois* dans les exemples cités), il est assez rare de trouver des homonymies sur ce type de noms propres composés, alors qu'il existe quatre *Neuilly*, et dix sept *Neuville*.

Or, ces formes composées font parfois l'objet d'une élision. La plupart du temps, le contexte permet au lecteur humain de reconstruire la forme ainsi effacée. Par exemple, l'évocation de *Nogent* renverra à *Nogent-sur-Loir* dans la *Sarthe* et à *Nogent-sur-Seine* dans *l'Aube.*

Ces deux types d'homonymies correspondent à un nombre de cas non négligeable. Ainsi, dans le **Dictionnaire National des Communes de France**, nous avons relevé à la lettre A, sur un total de 2 862 noms de localités, 356 cas d'homonymies simples et 243 cas d'homonymies potentielles par élision. Cela correspond à 21% du corpus analysé (voir les détails en annexe 1)

2.4.3.3 L'homonymie dans des types différents

Un même mot peut renvoyer à des noms propres appartenant à des types différents. La détection du type est essentielle pour le processus d'analyse automatique. La reconnaissance d'un type unique déclenchera des associations avec les types de noms propres en relation directe avec le type détecté. Ainsi le mot *Loire-Atlantique* appartient au type « nom de département » sans aucune ambiguïté. Dans le système relationnel du dictionnaire, il possède des liens directs avec les types suivants : **commune, région administrative, habitants département.**

Il n'en va pas de même du mot *Mayenne* qui peut appartenir à quatre types différents et donc conduire à de multiples associations « candidates » parmi lesquelles il faudra déterminer celles qui sont pertinentes.

Type détecté	Types associés	Noms propres candidats
département	région administrative	Pays de Loire
	habitants département	Mayennais
hydrographie	hydrographie	Sarthe
	commune	Mayenne, Laval, Château-Gontier
commune	habitants commune	Mayennais
	département	Mayenne
	hydrographie	Mayenne
	région historique	Vendée Militaire
	événement	Guerre de Vendée
personnage	nom collectif nom collectif alias personnage	Maison de Guise Maison de Lorraine Charles de Lorraine
	commune	Alençon (né à ...), Soissons (mort à ...)
	événement	la Sainte Ligue, Guerres de Religion

Figure n° 17 : Les quatre types différents du mot Mayenne

3. Notre méthodologie de traitement des noms de lieux et d'habitants

Certaines méthodes de reconnaissance des noms propres reposent en partie sur la détection des lettres majuscules. Nous n'avons pas fait ce choix. Nous présenterons dans le paragraphe suivant une étude justifiant notre position. Ensuite, afin d'illustrer notre méthodologie de traitement automatique des noms de lieux et d'habitants, nous reprendrons l'exemple de l'article du journal Ouest-France qui a été présenté dans l'introduction..

3.1 Les problèmes de la reconnaissance par détection des majuscules

En langue anglaise (ou américaine) certaines heuristiques de reconnaissance des noms propres reposent sur la détection d'une initiale en majuscule.

Qu'en est-il en français ? Nous avons mené une étude sur un corpus d'articles du journal Ouest-France d'environ 290 millions de caractères. Le programme d'extraction des mots commençant par une lettre majuscule a été conçu sans analyse du contexte gauche/droite. Il ne tient pas compte des mots commençant une phrase, qui par ailleurs peuvent être des noms propres. De plus, la délimitation a été réalisée du début du mot jusqu'à la rencontre d'un blanc ou d'un signe de ponctuation. Donc, lors de l'apparition d'un tiret, l'analyse se poursuit jusqu'à la rencontre d'une des conditions d'arrêt.

La liste des mots obtenue a fait l'objet d'une analyse manuelle. Elle inspire une première remarque : de nombreux noms propres sont des noms composés formés de plusieurs mots séparés par des blancs. Ainsi *Jacques Chirac, Marcel Proust* ont été délimité chacun d'entre eux comme deux mots commençant par une majuscule. Il est vrai que les travaux de Paik, Liddy, Yu et McKenna, présentés au paragraphe 1.2.5.3, ont mis en évidence que la reconnaissance des prénoms associée à des heuristiques spécifiques, permet la résolution d'un nombre de cas important. Il n'en va pas de même pour les noms de lieux et les gentilés. Ainsi, *Saint Etienne de Mont Luc* donne quatre noms propres alors qu'avec la forme normalisée de la Poste (*Saint-Etienne-de-Mont-Luc*) le programme d'analyse n'en délimite qu'un seul.

La seconde remarque tient au grand nombre de sigles apparaissant dans le texte, avec des variations très importantes d'une lettre à l'autre. Il semble évident que ce type de « mots», particulièrement présent dans les textes journalistiques devra, à terme, faire l'objet de traitements automatiques spécifiques.

Voici, présentés figure 18, l'analyse des résultats obtenus sur les dix premières lettres de l'alphabet.

Lettre	Occurrences de mots	NP ou éléments de NP	%	Sigles	%	Rem.
A	13814	4139	29,96%	213	1,54%	1
B	9721	7487	77,02%	462	4,75%	2
C	21701	5023	23,15%	2172	10,01%	3
D	12425	2385	19,20%	420	3,38%	
E	9124	849	9,31%	136	1,49%	
F	9214	4081	44,29%	596	6,47%	4
G	5781	3196	55,28%	196	3,39%	
H	2717	1627	59,88%	138	5,08%	
I	7263	356	4,90%	223	3,07%	5
J	7732	5148	66,58%	56	0,72%	6
Total	**99492**	**34291**	**34,47%**	**4612**	**4,64%**	

Figure n° 18: Tableau des résultats de l'analyse des occurrences de mots avec une majuscule

Remarques (Rem.):

* 1 - Parmi les mots délimités on compte un nombre important de : *A, Au, Autre* ..

* 2 - Pourcentage important de noms propres venant sans doute du nombre très peu élevé de prépositions commençant par cette lettre

* 3 - Beaucoup de sigles commencent par la lettre « C »

* 4 - 2553 occurrences de « F » symbole monétaire et 1646 occurrences de « France », Français » auxquelles s'ajoute un nombre important de sigles en « F... » (Fédération de ...)

* 5 - 4550 occurrences de pronoms personnels « Il », Ils »

* 6 - 753 occurrences du pronom personnel « J' »

Les résultats sur les dix premières lettres de l'alphabet font apparaître un niveau de « bruit » de l'ordre de 65 %, ce qui est déjà considérable. Il faut noter que ce chiffre ne prend pas du tout en compte le phénomène du découpage de certains noms propres tel que nous l'avons évoqué plus haut. Dès lors, un traitement automatique des noms propres ne peut pas reposer sur un tel processus. En revanche, des heuristiques de détection constituent, dans un certain nombre de cas, des outils complémentaires efficaces de reconnaissance ou de validation d'une hypothèse de détection.

3.2 Reconnaissance, typage, traitement des références : une présentation intuitive

En reprenant l'article du journal Ouest-France, que nous avons présenté lors de l'introduction, nous allons illustrer les différentes étapes des traitements automatiques que nous envisageons de mettre en œuvre pour **reconnaître, typer** et **traiter les coréférences** entre noms propres.

3.2.1 Reconnaissance et typage

*«**La Roche** - **Thouaré** : 2-0*
SANS CONVAINCRE

***La Roche-sur-Yon**. - Après leur défaite à domicile contre **Nozay**, les **Ornaisiens** avaient besoin de se rassurer en mettant leur calendrier à jour face à **Thouaré**. Si l'objectif fut atteint au niveau du score, ce fut quelque peu laborieux dans la manière. Il est vrai que l'état du terrain ne facilita guère l'évolution des 22 acteurs. Face à des **Thouaréens** accrocheurs, les **Ornaisiens** parvenaient toutefois à se créer une bonne occasion après un quart d'heure de jeu (...)*

*Peu avant la pause, les **Vendéens** parvenaient toutefois à trouver la faille sur un ballon en profondeur (...) Ce coup du sort eut le don de réveiller les **Thouaréens** après le repos. »*

Dans ce texte, le nom des habitants ***Thouaréens*** correspond à la dérivation du nom ***Thouaré*** qui est une forme élidée du nom de commune ***Thouaré-sur-Loire. Ornaisiens*** réfère à ***Saint-André-d'Ornay*** sous-ensemble (quartier) de la commune de ***La Roche-sur-Yon*** qui est présente dans le texte également sous sa forme élidée : ***La Roche. Vendéens*** renvoie, soit au nom de la région historique dans laquelle est située ***La Roche-sur-Yon***, soit au nom du département de la ***Vendée,*** ce qui conduit au même résultat. Seule ***Nozay*** n'a pas de coréférent dans l'extrait. La seule information disponible le concernant sera le type « localité ».

Dans un premier temps, un analyseur classique laissera de côté tous les noms propres rencontrés en les étiquetant ***mot inconnu.***

*« **[mot inconnu] -[mot inconnu]**: 2-0*
SANS CONVAINCRE

***[mot inconnu]**. - Après leur défaite à domicile contre **[mot inconnu]**, les **[mot inconnu]** avaient besoin de se rassurer en mettant leur calendrier à jour face à **[mot inconnu]**. Si l'objectif fut atteint au niveau du score, ce fut quelque peu laborieux dans la manière. Il est vrai que l'état du terrain ne facilita guère l'évolution des 22 acteurs. Face à des **[mot inconnu]** accrocheurs, les **[mot inconnu]** parvenaient toutefois à se créer une bonne occasion après un quart d'heure de jeu (...)*

*Peu avant la pause, les **[mot inconnu]** parvenaient toutefois à trouver la faille sur un ballon en profondeur (...) Ce coup du sort eut le don de réveiller les **[mot inconnu]** après le repos. »*

Occurrences	Type	Genre / Nombre	Référent(s)
La Roche	Inconnu	Inconnus	Inconnu(s)
Thouaré	Inconnu	Inconnus	Inconnu(s)
La Roche-sur-Yon	Inconnu	Inconnus	Inconnu(s)
Nozay	Inconnu	Inconnus	Inconnu(s)
Ornaisiens	Inconnu	Inconnus	Inconnu(s)
Thouaréens	Inconnu	Inconnus	Inconnu(s)
Vendéens	Inconnu	Inconnus	Inconnu(s)

Figure n° 19 : Tableau résultat n°1 de l'analyse des noms propres de l'article

La première tâche que nous nous sommes fixée, c'est de reconnaître les noms propres et de leur attribuer un type, un genre et un nombre. Comme nous le détaillerons plus loin, notre système réalise la reconnaissance et le typage en une seule étape. Nous nous retrouvons donc dans la situation suivante :

*« **[élision (M+F)S[localité]] -[alias (M+F)S [localité]]** : 2-0*
SANS CONVAINCRE

***[localité (M+F)S]** - Après leur défaite à domicile contre **[localité (M+F)S]** ,les **[gentilés MP [sous-ensemble (M+F)S [localité]]]**., avaient besoin de se rassurer en mettant leur calendrier à jour face à **[alias (M+F)S[localité]]**. Si l'objectif fut atteint au niveau du score, ce fut quelque peu laborieux dans la manière. Il est vrai que l'état du terrain ne facilita guère l'évolution des 22 acteurs. Face à des **[gentilés MP [alias (M+F)S[localité]]]** accrocheurs, les **[gentilés MP [sous-ensemble (M+F)S[localité]]]**., parvenaient toutefois à se créer une bonne occasion après un quart d'heure de jeu (...)*

*Peu avant la pause, les **[gentilés MP ([département] | [région historique])]**. parvenaient toutefois à trouver la faille sur un ballon en profondeur (...) Ce coup du sort eut le don de réveiller les **[gentilés MP [élision (M+F)S [localité]]]** après le repos. »*

Occurrences	Type	GENN	Référent(s)
La Roche	Elision localité	(M+F)S	Inconnu(s)
Thouaré	Alias localité	(M+F)S	Inconnu(s)
La Roche-sur-Yon	Localité	(M+F)S	Inconnu(s)
Nozay	Localité	(M+F)S	Inconnu(s)
Ornaisiens	Gentilé [Sous ensemble localité]	MP	Inconnu(s)
Thouaréens	Gentilé [Alias Localité]	MP	Inconnu(s)
Vendéens	Gentilé [Département | Région historique]	MP	Inconnu(s)

Figure n° 20 : Tableau résultat n°2 de l'analyse des noms propres de l'article

3.2.2 Détection et résolution des coréférences

Le nom propre est porteur de coréférences vis à vis d'autres noms propres. Ainsi, il est fréquent de voir dans un même texte (surtout dans les articles de journaux), des noms de lieux, d'habitants, de régions ... renvoyant les uns aux autres dans un réseau sous-jacent au texte, qui est porteur d'informations sémantiques.

Les liens que les noms de lieux et d'habitants ont les uns avec les autres étant gérés par nos outils, un traitement automatique permet de les calculer, même s'ils ne sont pas directs. Nous nous trouverons alors dans la situation suivante :

*« **[élision (M+F)S[localité 1]] -[alias (M+F)S[localité 3]]** : 2-0*
SANS CONVAINCRE

***[localité 1 (M+F)S]** - Après leur défaite à domicile contre **[localité 2 (M+F)S]** ,les **[gentilés MP [sous-ensemble (M+F)S [localité 1]]]**., avaient besoin de se rassurer en mettant leur calendrier à jour face à **[alias (M+F)S[localité 3]]**. Si l'objectif fut atteint au niveau du score, ce fut quelque peu laborieux dans la manière. Il est vrai que l'état du terrain ne facilita guère l'évolution des 22 acteurs. Face à des **[gentilés MP [alias (M+F)S [localité 3]]]** accrocheurs, les **[gentilés MP [sous-ensemble (M+F)S[localité 1]]]**., parvenaient toutefois à se créer une bonne occasion après un quart d'heure de jeu (...)*

*Peu avant la pause, les **[gentilés MP ([département[localité 1]] | [région historique[localité 1]])]** parvenaient toutefois à trouver la faille sur un ballon en profondeur (...) Ce coup du sort eût le don de réveiller les **[gentilés MP [alias (M+F)S [localité 3]]]** après le repos. »*

Occurrences	Types	G/N	Référent(s)
La Roche	Elision localité - 1	(M+F)S	La Roche-sur-Yon
Thouaré	Alias localité - 3	(M+F)S	Thouaré-sur-Loire
La Roche-sur-Yon	Localité - 1	(M+F)S	La Roche
Nozay	Localité - 2	(M+F)S	Inconnu(s)
Ornaisiens	Gentilé [Sous ensemble localité]	MP	Saint-André-d'Ornay
Thouaréens	Gentilé [Alias localité]	MP	Thouaré
Vendéens	Gentilé [Département | Région Hist.]	MP	Vendée

Figure n° 21 : Tableau résultat n°3 de l'analyse des noms propres de l'article

Comme cela est mis en évidence dans les figures 1 et 2 de l'annexe 2 , certaines coréférences sont détectées de façon directe :

La Roche → (forme élidée de ...) → La Roche-sur-Yon
Thouaré → (forme alias de ...) → Thouaré-sur-Loire

D'autres, trouvent une solution par transitivité:

Ornaisiens →La Roche-sur-Yon :

1 - Ornaisiens → (gentités de ...) → Saint-André-d'Ornay
2 - Saint-André-d'Ornay → (sous-ensemble de ...) → La Roche-sur-Yon.

Vendéens → La Roche-sur-Yon :

1 - Vendéens → (gentilés de ...) → Vendée (département de/ région hist. Géo de)
2 - Vendée → (département de / région hist. Géo de) → La Roche-sur-Yon

Thouaréens → Thouaré-sur-Loire :

1 - Thouaréens → (gentilés de ...) → Thouaré-sur-Loire
2 - Thouaré-sur-Loire → (localité dont la forme alias est...) → Thouaré

4. Une base de données relationnelle des noms de lieux et d'habitants

Nous avons fait le choix d'une reconnaissance et d'un traitement des noms de lieux et d'habitants fondés sur l'exploitation d'un dictionnaire électronique. Il était donc indispensable de procéder à une collecte des informations nécessaires à la construction de cet outil.

4.1 Collecte, validation et organisation des données

Ce travail de collecte est moins trivial qu'il n'y paraît. Si nous prenons l'exemple des noms de localités et des noms de leurs habitants (gentilés), il n'existe pas, à notre connaissance, de liste exhaustive de ce type. Certaines listes d'exemples sont parfois fournies par des dictionnaires ou des encyclopédies. La plupart du temps, elles sont incomplètes et les choix ne sont pas justifiés. Ainsi, le *Quid* pose la question « *Comment s'appellent les habitants de ...* ». Par ailleurs, certains ouvrages de grande diffusion reprennent le principe, mais en ne retenant que les localités ayant des gentilés « exotiques » (souvent des formes supplétives), sans citer la source de ces informations. Enfin, dans tous les cas, l'aspect flexionnel n'est jamais traité, seule la forme au masculin pluriel est présentée.

En novembre 1993, en réponse à notre demande, madame Sylvie Lejeune, secrétaire de la Commission de Toponymie de l'**Institut Géographique National** nous a communiqué une liste d'environ 1300 noms de lieux et d'habitants au masculin pluriel. Dans son courrier d'accompagnement, elle précisait :

> *« Je me permets toutefois de vous faire observer qu'il s'agit d'une liste établie à partir de documents d'origines diverses, parfois sujets à caution, et qu'il n'existe pas encore de véritable norme dans ce domaine. »*

Il est vrai que les gentilés peuvent être hérités de la tradition orale ou écrite, décrétés par une autorité administrative centrale comme cela a souvent été le cas au XIX^{ème} siècle ou, depuis les lois de décentralisation, fixés par le conseil municipal. Cette absence de normalisation explique sans doute le manque de documents de synthèse ; cependant l'analyse automatique a pour vocation de prendre en charge le plus grand nombre possible de phénomènes langagiers, qu'ils obéissent ou non à une norme !

Nous pensions que les préfectures disposaient de documents listant les noms d'habitants des principales localités de leur territoire d'administration. Une enquête rapide a mis en évidence qu'il n'en était rien. Certaines, souvent par l'intermédiaire des directions des Archives Départementales ont manifesté de l'intérêt pour nos recherches en souhaitant être tenu informées des résultats.

Il s'est rapidement imposé qu'il était nécessaire de mettre en place une stratégie de collecte et de validation des informations. Nous disposions des documents suivants :

- Une première liste établie en 1992 [Maurel, Leduc, Courtois 92],
- La liste de l'**IGN**

Nous pouvions compléter ces informations par celles figurant dans les **Guides Verts Michelin** et le **Robert** des noms propres. Mais le fait de devoir travailler avec des

informations d'origine uniquement documentaire nous a semblé insuffisant, étant donné la nature du sujet traité.

4.1.1 Une enquête nationale sur les noms propres

Nous avons alors envisagé de réaliser une enquête nationale adressée aux mairies. Pour mettre en œuvre une opération de cette nature, il était nécessaire de construire des partenariats avec des tiers intéressés par nos travaux et susceptibles de nous apporter une aide. Après de nombreux contacts, trois organismes ont indiqués qu'ils étaient prêts à collaborer à ce projet :

• Le journal **Ouest-France**. Monsieur Annarumma, directeur de projet était intéressé par les informations que nous étions susceptibles de collecter pour des raisons éditoriales évidentes. Il a demandé que toutes les localités de la zone de diffusion du journal soient sélectionnées pour l'enquête. En échange, le journal nous a fourni un corpus d'articles d'environ 2,5 millions de caractères sur support magnétique.

• La **Direction Départementale de la Poste**. Monsieur Le Piétec, directeur du service de communication nous a également fait part de son intérêt. Il nous a offert la possibilité d'un envoi gratuit en direction des mairies, dans des enveloppes à en-tête officielle de La Poste, pour un volume total de 6000 exemplaires. En contrepartie, il a demandé qu'une question sur la date de création des bureaux de poste locaux figure sur notre document, afin de fournir ces informations au Musée de la Poste.

• Le **Conservatoire National des Arts et Métiers** de Nantes. Depuis quelque temps, cette structure de formation souhaite développer des activités de recherche. Elle est intéressée par des collaborations avec des laboratoires universitaires. Ses dirigeants ont donc accepté de se joindre à notre projet en fournissant toute la logistique nécessaire à ce volumineux envoi.

4.1.1.1 La sélection des localités

Il était nécessaire d'établir de façon précise la liste des localités qui allaient faire l'objet de l'envoi. Nous avons retenu les critères suivants :

• Toutes les localités de plus de **10 000 habitants** (recensement 75)

• Tous les **chefs lieux** (cantons, arrondissements, départements, régions)

• Toutes le localités connues pour des raisons historiques et/ou géographique et figurant à ce titre dans les **Guides Verts Michelin**

• Toutes les communes des départements correspondant à la zone de diffusion du journal **Ouest-France**.

Soit un total de **5960**.

4.1.1.2 Le questionnaire

Comme on peut le voir en annexes 3 et 4, le questionnaire comporte deux parties, anticipant ainsi la structure de notre dictionnaire électronique (entités, associations). L'avantage d'une telle enquête était de connaître, non seulement le nom des habitants dans toutes leurs formes fléchies, mais de poser d'autres questions, dont les réponses sont pratiquement introuvables en dehors du terrain, à moins de se livrer à un gigantesque travail historico-géographique. De plus, de nombreuses localités possèdent plusieurs noms d'habitants. Il s'agit parfois de noms différents sur le plan formel mais proches sur le plan étymologique comme les **Castelroussins** et les **Châteauroussins** pour **Chateauroux** dans *l'Indre*. Dans d'autres cas, les noms ne différent que par le suffixe comme les **Tournais** ou les **Tournois** pour **Tournes** dans les *Ardennes*. Mais on trouve aussi des noms dont l'un est construit sur une

dérivation régulière, l'autre sur une dérivation supplétive comme les **Nantuatiens** et les **Catholards** pour **Nantua** dans *l'Ain.*

Mais s'il est facile de connaître l'appartenance d'une ville à un département ainsi que l'inclusion de celui-ci dans une région administrative, il est beaucoup moins aisé de déterminer si telle ou telle cité a le sentiment d'appartenir à la *Saintonge*, à *l'Argone*, au *Bugey*, à la *petite Crau*, à *l'Occitanie* ou au *Trégorrois*. De plus, certaines régions administratives portent le même nom que des régions historiques, sans faire référence à un même lieu géographique ! Ainsi, la *Vendée Militaire* s'étend largement au delà de la *Vendée*, région historique, qui elle-même ne correspond pas au département portant le même nom. Il était donc intéressant de saisir cette occasion pour poser la question suivante:

> *« En plus de la région administrative, votre localité se sent peut-être appartenir à une ou à plusieurs régions* ***traditionnelles*** *ou* ***historiques****. Par exemple, la* ***Flandre*** *(Flamands) et* ***l'Artois*** *(Artésiens) font partie de la région administrative Nord-Pas de Calais »*

Dans le paragraphe consacré au dépouillement nous verrons que ces aspects, que nous pensions accessoires au niveau des informations traitées, ont pris une importance assez inattendue. Pour chaque région historique géographique nous avons demandé les noms d'habitants dans toutes leurs formes fléchies ainsi que des informations sur la prononciation. Nous pensons, par la suite, phonémiser le dictionnaire.

La seconde partie du document d'enquête concernait les autres noms propres associés à la localité, nous donnions l'indication suivante :

> *« Une localité est parfois connue ou reconnue par un ou des personnages célèbres, un ou des événements historiques, un ou des produits typiques, ou encore un quartier, un monument ou autres ... »*

Nous sollicitions également des informations sur l'origine du nom des habitants ainsi que sur la date de création du premier bureau de poste, selon le souhait de notre partenaire La Poste.

4.1.2 Dépouillement, contrôle et validation

Nous avons reçu 1758 réponses, ce qui, aux dires des spécialistes du publipostage, constitue un taux de retour tout à fait honorable.

Une analyse rapide a permis de mettre en évidence que les informations qui « remontaient » ainsi du terrain étaient nombreuses, variées, souvent très documentées par de volumineuses annexes, mais parfois surprenantes. Ainsi, la partie consacrée aux personnages célèbres liés à une localité est revenue avec la mention *néant* pour **Colombey-les-Deux-Eglises** ! Il était donc indispensable de procéder à un contrôle et à une validation, en croisant les renseignements ainsi transmis avec ceux dont nous disposions sur le plan documentaire. Nous avons adopté la méthode suivante :

Les informations de chacune des fiches d'enquête ont été confronté à quatre sources différentes :

- Les Guides Verts Michelins
- Le dictionnaire Robert des Noms Propres
- La liste de l'IGN
- Les fichiers de l'Insee

Quand les informations divergeaient sur le nom des habitants, les différentes versions ont été collectées en signalant les sources. D'autre part, de nombreuses fiches d'enquête comportaient beaucoup d'informations supplémentaires (sites historico-géographiques, entités hydrographiques, événements historiques ...). Bien que le travail présenté ici se limite à la reconnaissance et au traitement des noms de lieux et de leur gentilés, cette abondance nous a conduit à élargir l'enregistrement des informations à toutes les entités ayant des liens avec la localité traitée, soit parce qu'elles étaient citées dans les réponses, soit parce qu'elles figuraient dans nos documents de contrôle. Ces informations appartiennent essentiellement aux catégories suivantes :

- Hydrographie (rivière, fleuve, étang, lac océan, mer...)
- Personnages célèbres
- Produits
- Sous-ensembles de localité (quartier, rue, place ...)
- Evénements
- Sites historiques ou géographiques

Si nous prenons l'exemple de la ville de Nantes, voici l'ensemble des données collectées. Nous avons mis dans la colonne « Enquête » les réponses correspondant aux questions posées, dans la suivante les information supplémentaires à l'initiative de l'enquêté, enfin en dernière colonne, les compléments que nous avons apportés lors du contrôle à partir des documents dont nous disposions :

	ENQUÊTE	UTILISATEUR	DOCUMENTATION
Localité	Nantes		
Nom habitants	Nantais		
	Nantais		
	Nantaise		
	Nantaises		
Région historique	Bretagne		
Nom habitants	Breton		
	Bretons		
	Bretonne		
	Bretonnes		
Personnages	Anne de Bretagne		Waldeck-Rousseau
	Aristide Briand		
	Cambronne		
	Jules Vernes		
	Paul Lamirault		
Hydrographie		Loire	Sèvre Nantaise
		Erdre	canal de Nantes à Brest
Sous-ensemble			passage Pommeraye
			rue Crébillon
		château des Ducs de Bretagne	
Evénement			Edit de Nantes
Produits	Muscadet		

Figure n° 22 : Synthèse du dépouillement de la fiche d'enquête de Nantes

Du fait de la stratégie choisie, le dépouillement a été beaucoup plus long que nous ne l'avions imaginé. Il a nécessité un an de travail et a permis de mettre en évidence deux phénomènes assez inattendus :

1 - Le nombre très important de petites **régions historiques et géographiques**. Nous en avons enregistré 299 en ne prenant en compte que celles dont l'existence était validée par une source documentaire. Ainsi, les liens de *Merville-Franceville* avec « *la Côte Fleurie* » ou de *Villemaure-sur-Vanne* avec la région dite « *Forêt d'Othe* » non pas pu être confirmés. En revanche, les *Gruissanais*, habitants de *Gruissan* dans le département de *l'Aude*, vivent dans le *Languedoc*, le *Narbonnais* et la *Septimanie* et sont donc à ce titre également des *Languedociens*, des *Narbonnais* et des *Septimaniens*.

2 - La très grande densité du réseau hydrographique français. De nombreuses localités ont signalé qu'elles étaient situées sur les bords d'un ru, d'une rivière, d'un étang... Dans de nombreux cas, on nous a transmis également des informations sur la localisation : traversé par, sur la rive droite... sur la rive gauche.., à proximité..., au confluent... Il n'est pas dans notre propos de construire une base de données géographique. Cependant, la collecte de ces informations et leur organisation relationnelle permettra dans le cadre de développements futurs de traiter des coréférences comme :

- riverains de la *Durance* pour les *Sistéronais*
- station balnéaire de *l'Atlantique* pour *La Baule*
- port fluvial de la *Seine* pour *Conflent-Sainte-Honorine*

On notera que l'organisation hiérarchique de ces entités hydrographiques se prête tout à fait à des traitements informatiques. A moyen terme, ces informations seront intégrées dans la cadre de notre projet de dictionnaire électronique relationnel des noms propres[7].

4.1.3 L'organisation relationnelle des données

Pour faciliter le dépouillement et préparer les traitements ultérieurs, il était important que l'on puisse mener conjointement :

- la saisie des informations sur les entités (localités, habitants, régions administratives, historiques etc...)
- la saisie des liens ou relations qui les unissent (habitants de ... , situé dans..., etc ..)

Nous avons fait le choix de construire une base de données relationnelle. Après un rappel des principaux concepts du Modèle Entités-Associations ainsi que du Modèle Relationnel, nous en présenterons une description détaillée.

4.2 Les outils informatiques

Sur le plan des développements informatiques, notre travail s'est organisé en deux phases successives :

1. **la collecte et la validation** au cours de laquelle certains prétraitements automatiques étaient nécessaires,
2. **la reconnaissance, le typage et le traitement des coréférences** dans le cadre d'une analyse automatique de textes.

---

[7] Depuis, un travail propre à l'hydrographie a été réalisé par une étudiante de la maîtrise de linguistique informatique de Paris VII : Mahnaz Ghorai Shi

4.2.1 Outils de collecte et de stockage

Il n'est pas dans notre propos d'être exhaustif sur la partie consacrée à l'analyse de l'existant et à la modélisation ainsi qu'aux caractéristiques des Systèmes de Gestion de Bases de Données Relationnelles. Il s'agit plutôt de fournir au lecteur les principaux concepts des modèles utilisés afin de suivre notre démarche. En revanche, le chapitre consacré à la reconnaissance et au traitement des noms propres présentera dans le détail les algorithmes que nous avons mis en œuvre.

D'autre part, les exemples cités pour illustrer la présentation du Modèle Conceptuel des Données sont des extraits parfois simplifiés de la base de données active dont la description détaillée sera présentée à la fin du chapitre 4 .

4.2.1.1 Analyse de l'existant et modélisation

La conception et la mise au point d'une base de données amène à distinguer plusieurs niveau de représentation de la réalité. Définir ce qu'est la **réalité** n'est pas une chose aisée. On a l'habitude de considérer que cette démarche consiste à construire une **représentation** dans un formalisme descriptif reprenant les informations qui apparaissent comme indispensables au regard des traitements que l'on souhaite mettre en œuvre dans un deuxième temps. L'informatique est un outil qui peut être utilisé dans des contextes très différents, avec des tiers à qui on ne peut pas demander d'être à la fois des experts de leur domaine d'activité et des spécialistes de l'analyse et de la programmation. En informatique le **formalisme de représentation** doit donc posséder des qualités contradictoires :

- Etre suffisamment précis et détaillé, pour représenter une base valide de conception de l'ensemble du système d'information
- Etre « lisible » par des non informaticiens, pour constituer un outil d'analyse, de discussion et de validation des choix.

Cette représentation sera constituée d'un ensemble de données élémentaires, regroupées et hiérarchisées afin d'exprimer les liens logiques existant entre les informations ainsi que les contraintes de cohérence. Il existe plusieurs techniques de modélisation. Nous allons présenter les deux approches que nous avons utilisées pour l'analyse et la conception de la base de données des noms propres :

- Le modèle Entités-Associations, support du Modèle Conceptuel des Données (MCD) de la méthode Merise, qui possède les qualités évoquées plus haut, grâce notamment à son formalisme graphique que nous détaillerons plus loin.
- Le Modèle Relationnel de E. F. Codd [Codd70] qui décrit les données dans des tables de façon relativement intuitive, mais qui repose sur des concepts rigoureux permettant l'utilisation d'outils mathématiques et algorithmiques.

4.2.1.2 Le Modèle Conceptuel des Données

Le modèle Conceptuel des Données repose sur quatre concepts principaux : **Entités, Associations, Attributs** et **Cardinalités.**

Entité

Une **Entité** est un objet de la réalité. Il peut être *concret* ou *abstrait.* Il se caractérise par le fait qu'il a une existence propre. Il est autonome. Par exemple, la localité **Nantes** est une **entité.** En fait, au cours d'un processus de modélisation, on ne s'intéresse pas directement à une entité mais plutôt à un **type d'entité.** Ainsi **Nantes** appartient au type d'entité **localité.** En analyse

Merise, pour des raison de facilité de langage, on désigne souvent le ***type d'entité*** par le terme **entité**, et dans le cas d'un individu particulier on parle alors d'**occurrence d'entité**. Ainsi, pour reprendre notre exemple, **Nantes** sera une **occurrence** de l'entité **localité**.

Dans le Modèle Entités-Associations, les entités sont représentées graphiquement sous forme d'un rectangle avec un bandeau dans lequel on indique la nom de l'entité.

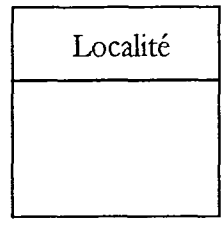


Figure n° 23 : Représentation graphique d'une entité

Association

L'**association** est un regroupement nécessaire entre deux ou plusieurs occurrences d'entités exprimant ainsi une relation formelle entre les entités considérées. Comme pour l'entité, on parle d'**association** pour désigner le ***type d'association*** et d'**occurrence d'association** pour exprimer la relation qui existe entre deux ou plusieurs **occurrences d'entités**.

L'association exprime un lien entre des objets. Elle n'a pas d'existence autonome. On la représente graphiquement sous forme d'un ovale reliant les entités considérées :

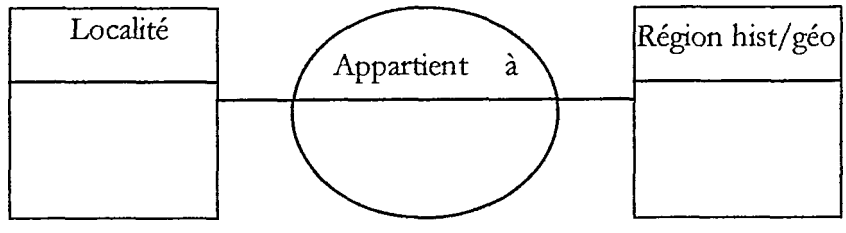


Figure n°24 : Représentation graphique d'une association reliant deux entités

La signification de cette association est simple : c'est une relation binaire, sous ensemble du produit cartésien [Localités X Régions His. Géo] :

- Du point de vue de la localité, elle exprime que des localités appartiennent à des régions historiques / géographiques.
- Du point de vue de la région (sens inverse de lecture du modèle), elle exprime que des régions historiques / géographiques comprennent des localités.

Le nombre d'occurrences participant de part et d'autre à la relation n'est pas encore exprimé dans la représentation de la figure 24.

Attributs

Un **attribut** est une caractéristique d'une entité ou d'une association que l'on juge utile, au moment de la conception du modèle, de répertorier dans la réalité perçue. Chaque attribut a un nom et un type.

Ainsi, une localité possède un nombre extrêmement important d'attributs potentiels, c'est-à-dire d'éléments qui permettent de l'identifier, de la caractériser. Ceux-ci peuvent être liés à sa situation géographique, à son histoire économique, politique, religieuse... Parmi ceux-ci, la modélisation ne retiendra que ceux qui sont jugés pertinents aux vues des traitements à mettre en œuvre.

Lorsqu'une entité a été mise en évidence, il est indispensable de pouvoir identifier de façon unique et non ambiguë chacune de ses occurrences. Pour cela, un des attributs joue le rôle particulier **d'identifiant** : on l'appelle **clé primaire** de l'**entité**. Elle se définit par ses propriétés de dépendances fonctionnelles sur l'ensemble des attributs de l'**entité**.

Il est parfois nécessaire de créer une clé de façon artificielle, aucun des attributs présents dans l'entité n'étant discriminant. Ainsi pour les localités, le code postal, qui est un des attributs, ne peut pas être utilisé comme clé car les villes importantes ont plusieurs codes postaux, alors que de petites commune se partagent le même !

Cardinalités

Un des aspects importants de la modélisation consiste à attribuer aux liens qui ont été établis entre les entité et les associations des valeurs exprimant le nombre minimum et maximum d'occurrences pouvant (ou devant) exister. C'est ce que l'on indique sous forme de **cardinalités** qui sont définies de la façon suivante [GALACSI, 84 :36] :

> *«Soit A un type d'entité et B un type d'association construit sur A. Au couple (A,B), représenté graphiquemenrt par un segment, on associe une cardinalité. Par définition, la cardinalité de (A,B) est constituée de deux nombres : le minimum et le maximum d'occurrences du type d'associations B pouvant exister pour une seule occurrence du type d'entité A.Une cardinalité est égale à 0,1 ou 1,1 ou 0,n ou 1,n. La lettre n représente un nombre entier supérieur à 1, mais qui peut se trouver fortuitement égal à 1 »*

Ainsi, l'exemple précédent complété par ces informations devient :

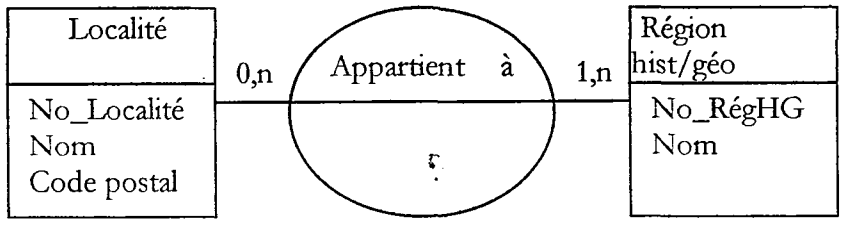


Figure n°25 : Les cardinalités d'une association

La sémantique du modèle s'exprime de la façon suivante :

- Une localité est caractérisée par :

1. un identifiant
2. un nom
3. un code postal

• Une région historique / géographique est caractérisée par :
1. un identifiant
2. un nom

• Une localité peut ne pas appartenir à une région historique/géographique (cardinalité 0).
• Une localité peut appartenir à plusieurs régions historiques/géographiques (cardinalité n).
• Une région historique/géographique comprend au moins une localité (cardinatité 1)
• Une région historique/géographique peut comprendre plusieurs localités (cardinatité n)

En revanche, les liens existant entre une localité et un département sont tout à fait différents : une localité étant toujours située dans un département et un seul, il s'agit d'une fonction totale, le modèle graphique devient :

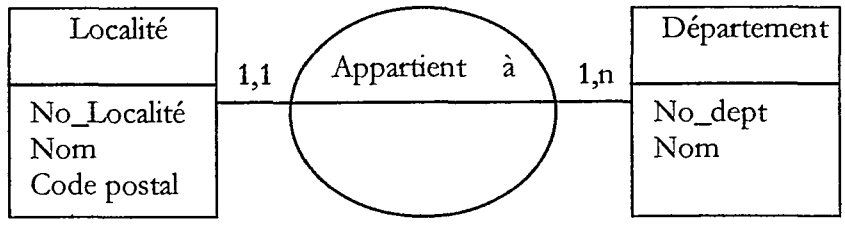


Figure n°26 : Une association avec une cardinalité **1,1**

Une des propriétés importantes du Modèle Entités-Associations est son indépendance totale par rapport aux traitements que l'on peut mettre en œuvre ensuite. Cela nous a permis de travailler au niveau de la collecte des données et de leurs relations, sans préjuger des choix que nous aurions à faire ensuite lors de la phase consacrée aux traitements informatiques des noms propres dans le cadre d'un processus d'analyse automatique de textes.

4.2.1.3 Modèle Relationnel

Une fois le modèle conceptuel mis en place, il faut faire le choix d'un modèle d'organisation des données sur le plan informatique. Comme le précisent [Delobel & Adiba 82 :10]

> *« ../.. un schéma conceptuel est le résultat d'une action de modélisation du monde réel qui respecte un modèle de données, à l'aide des termes et des expressions permises par le langage de définition des données.*
> *On différencie trois grandes classes de modèles de données qui se distinguent par la nature des associations qu'ils permettent de modéliser : les modèles hiérarchiques, réseaux et relationnels »*

Le modèle relationnel nous a semblé suffisant pour construire une base de données qui est essentiellement un outil intermédiaire de stockage et d'exploitation des informations sur les noms propres. En effet, celle-ci n'est utilisée que pour générer de façon automatique les deux listes qui permettent de construire le transducteur d'associations. Un moment, nous nous sommes interrogé sur l'opportunité d'utiliser d'un modèle objet, d'autant que les outils de construction et d'exploitation du transducteur sont écrits en C++. Cependant, la notion d'identifiant interne de la plupart des systèmes de bases de données objets nous a semblé embarrassante étant donné la nature des traitements que nous voulions réaliser sur le typage des identifiants par préfixage. Tout ceci nous a amené à faire le choix d'un SGBDR classique.

Proposé dans les années 70 [COD70], le modèle relationnel est a l'origine des Systèmes de Gestion de Bases de Données Relationnelles. Présentes sur les gros et moyens systèmes dans les années 80, les SGBDR sont aujourd'hui sur les ordinateurs individuels. Cet intérêt découle de trois raisons principales [GALACSI, 84 :45]

> *« - simplicité du modèle qui présente les données sous une forme tabulaire, donc très naturelle pour les utilisateurs non informaticiens ;*
> *- rigueur des concepts permettant l'utilisation des outils mathématiques et algorithmiques : les travaux développés autour de ce modèle ont largement contribué à la création de méthodes d'analyse moins empiriques que celles utilisées dans le passé ;*
> *- adéquation du modèle au niveau conceptuel : l'absence de notions informatiques, souvent nécessaires dans d'autres modèles, le rend particulièrement propice à la description du schéma conceptuel des données »*

4.2.1.3.1 Notion de Relation

Dans ce modèle, les informations sont vues sous forme de tables dont les titres des colonnes contiennent les noms des attributs du modèle Entités-Associations, et dont les lignes contiennent les « réalisations » de ces attributs, c'est à dire les données.

No_Localité	Nom	Code postal
1.16396	Ablis	78660
1.13384	Ablon-sur-Seine	94480
1.12703	Abondance	74360
1.10955	Abondant	28570
1.11981	Abreschviller	57560
1.13674	Abries	05460

Figure n°27 : Un extrait de la table (simplifiée) des localités

Une table possède un nom, ici, **Localité**. Elle peut se représenter sous la forme d'une relation[8] :

Localité (No_Localité, Nom, Code postal)

Le Modèle Relationnel s'exprime sous la forme d'un ensemble de relations de ce type.

4.2.1.3.2 Identifiant d'une relation

Pour l'exploitation des informations, il est indispensable de pouvoir accéder à un nuplet (une ligne) de la table. Toute relation n-aire admet au moins une clé : l'ensemble de ses attributs. Cependant, les contraintes d'exploitation nécessitent que l'on puisse disposer d'un discriminant plus simple. Pour cela, si l'un des attributs ne peut pas jouer ce rôle, une clé « artificielle » est construite. Dans l'exemple précédent, l'identifiant est le No_Localité. On le signale dans la relation en le soulignant :

Localité (<u>No_Localité</u>, Nom, Code postal)

[8] Les identifiants des entités sont préfixées d'un nombre donnant le type suivi d'un point (dans l'exemple 1.), nous expliquerons en détail ce choix dans le chapitre suivant.

4.2.1.3.3 La normalisation

Nous ne pouvons pas, ici, entrer dans le détail de l'analyse des dépendances fonctionnelles et des contraintes d'intégrités qui caractérisent le modèle relationnel. Il existe un processus de normalisation qui vise essentiellement à :

- minimiser les redondances d'informations,
- maximiser l'indépendance des relations les unes par rapport aux autres.

Pour cela, E.F. Codd a proposé une classification des relations en trois **Formes Normales.**

La définition donnée des relations implique qu'elles soient toutes en **Première Forme Normale** : les attributs d'une relation en première forme normale sont indécomposables :

par exemple la relation :

Personne (Nom, Adresse)

n'est pas en Première Forme Normale, car on peut décomposer les attributs de la façon suivante :

Nom → Nom, Prénom
Adresse → Adresse, Code postal, Commune

La relation :

Personne (Nom, Prénom, Adresse, Code postal, Commune)

est en Première Forme Normale.

Un schéma relationnel est en **Deuxième Forme Normale** si tous les constituants non clés sont pleinement dépendants des constituants clés (un attribut non clé dépend fonctionnellement de la totalité de la clé) :

par exemple la relation suivante qui exprime l'association entre des localités et des régions historiques / géographiques n'est pas en Deuxième Forme Normale:

Loc/Rhg (No_Localité, No_Rhg, Nom de localité)

No_Localité qui constitue une partie de la clé donne le nom de localité :

No_Localité → Nom de localité

Loc/Rhg (No_Localité, No_Rhg) est en Deuxième Forme Normale, par ailleurs on dispose d'une relation :

Loc (No_Localité,Nom de localité)

Un schéma relationnel est en **Troisième Forme Normale** si tous les constituants non clés sont directement et pleinement dépendants des constituants clés (un attribut non clé ne dépend pas fonctionnellement d'un autre attribut non clé) :

par exemple la relation :

Localité (No_Localité, Nom localité, Code postal, No_Dép., Nom département)

N'est pas en Troisième Forme Normale car **No_Dép.** donne **Nom département**. La relation doit donc être décomposée de la façon suivante :

Localité (No_Localité, Nom localité, Code postal, No_Dép)
Département (No_Dép. , Nom département)

Par la suite, la recherche a mis en évidence que les 3 Formes Normales ne suffisaient pas à rendre compte de tous les phénomènes de redondance et de dépendances multivaluées. D'autre formes normales ont été introduites pour compléter la normalisation [GALACSI, 84 :51] .

4.2.1.3.4 Règles de passage du modèle Entités-Associations vers le Modèle Relationnel

Le passage du Modèle Entités-Associations vers le Modèle Relationnel se fait en appliquant trois règles :

Règle 1 : Toute entité est traduite en une relation avec les mêmes attributs et comme clé l'identifiant de l'entité. (reprise de l'exemple de la figure 25)

Localité (No_localité, Nom, Code postal)

No_Localité	Nom	Code postal
1.16396	Ablis	78660
1.13384	Ablon-sur-Seine	94480
1.12703	Abondance	74360
1.10955	Abondant	28570
1.11981	Abreschviller	57560
1.13674	Abries	05460

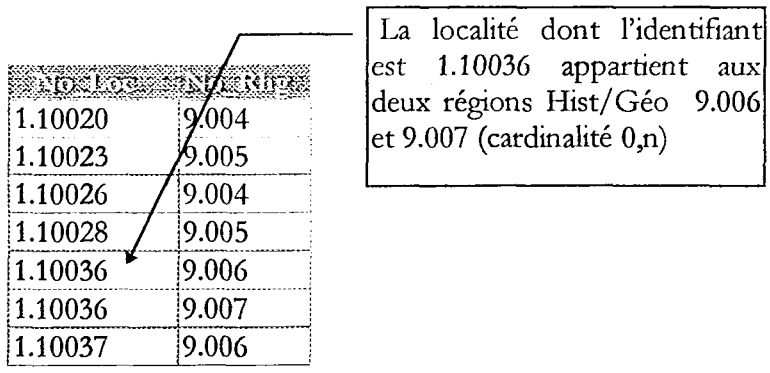
Région HG (No_RégHG, Nom)

No_Rég	Nom_Rég
9.001	Haut-Bugey
9.002	Jura
9.003	Dombes
9.004	Bugey
9.005	Bresse
9.006	Picardie
9.007	Vermandois

Figure n°28 : Relations **Localité** et **Région HG** et extrait des tables des deux entités

Règle 2 : Une association sans cardinalité 0,1 ou 1,1 est traduite par une relation dont la clé est constituée de l'ensemble des identifiants des entités associées, à laquelle on ajoute éventuellement les attributs non clé de l'association .

Loc_RHG (No_localié, No_RégHG)

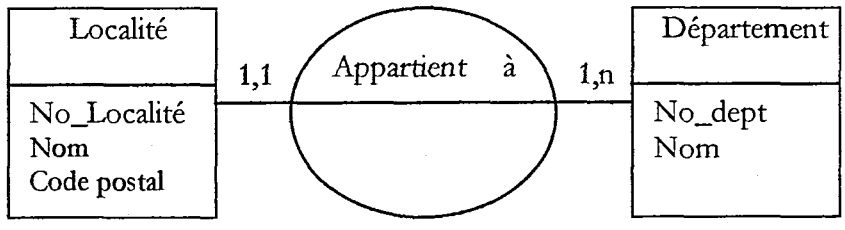


No_Loc	No_Rte
1.10020	9.004
1.10023	9.005
1.10026	9.004
1.10028	9.005
1.10036	9.006
1.10036	9.007
1.10037	9.006

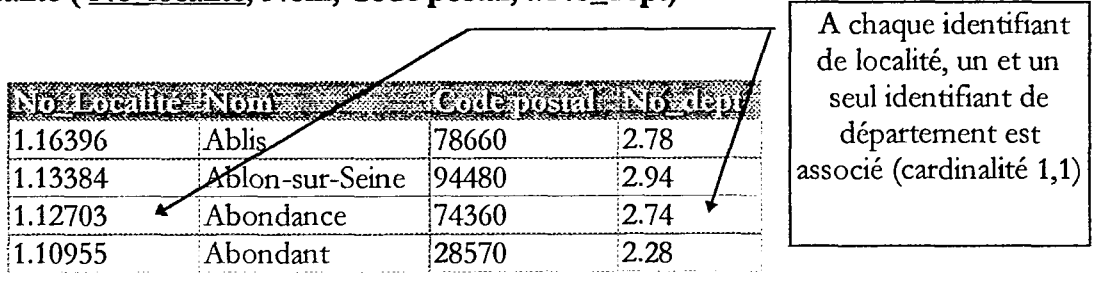
Figure n°29 : Exemple de la table Association[9] correspondant à la figure 25

Règle 3 : Toute association ayant une cardinalité 0,1 ou 1,1 avec une entité se traduit par l'ajout dans la relation équivalente de l'entité, avec le statut particulier de « clé externe » du ou des identifiants des autres entités associées (reprise du deuxième exemple de 5.2.1.2.5) :

Dans le second exemple, qui exprime le lien d'appartenance d'une localité à un et un seul département, la présence d'une cardinalité 1,1 entre Localité et Département va engendrer la création d'une clé externe[10] dans la relation localité :



Localité (<u>No_localité</u>, Nom, Code postal, #No_dept)



No_Localité	Nom	Code postal	No_dept
1.16396	Ablis	78660	2.78
1.13384	Ablon-sur-Seine	94480	2.94
1.12703	Abondance	74360	2.74
1.10955	Abondant	28570	2.28

Figure n°30 : Relation avec une clé externe et extrait de la table correspondante

4.2.1.4 L'organisation des données « Noms Propres »

Nous allons maintenant détailler l'organisation de notre base de données « noms propres ». Comme il a été signalé dans les paragraphes précédents, nous avons construit une structure globale qui va au delà du travail présenté ici. Au cours de la description, nous signalerons les entités et associations « actives » de celles qui ne sont présentes que structurellement, soit parce que nous ne disposons pas d'un ensemble de données suffisamment représentatif, soit parce que les données n'ont pas encore été collectées.

---

[9] C'est à dire que Saint-Quentin est situé à la fois en Picardie et dans le Vendermois

[10] Elles sont en général signalées par un élément graphique différent des clés primaires, ici le signe #

4.2.1.4.1 Les Formes Canoniques

Un nom propre peut être présent dans un texte sous des morphologies variées. Pour un nom de lieu, le phénomène des graphies multiples peut générer un nombre très important de formes différentes (Figure 16). Pour un nom d'habitant, on peut rencontrer une des quatre formes fléchies habituelles (MS, MP, FS, FP).

Or, ces variations n'ont pas d'influence sur la sémantique du lien qui associe deux noms propres. Ainsi, dans l'exemple suivant la relation *habitant de ...* n'est pas affectée par les variations morphologiques ou flexionnelles :

habitants de ...	
Stéphanois	Saint-Etienne
Stéphanoises	Saint Etienne
Stéphanoise	St Etienne
Stéphanoises	St-Etienne

Figure n°31 : Sémantique relationnelle indépendante des formes flexionnelles et dérivationnelles

La base de données ne contient que les formes canoniques (ligne n°1 de la figure 31):

- masculin singulier pour les gentilés
- forme normalisée de La Poste pour les noms de localité

En revanche, les outils informatiques de reconnaissance devrons prendre en charge la détection des différentes graphies ou des différentes formes fléchies d'un même nom propre **en n'émettant qu'un seul identifiant sur chacune des formes**: celui de la **forme canonique**. Ainsi pour l'exemple de la figure 31, la reconnaissance d'une des quatre formes de nom d'habitants déclenchera la production de l'identifiant de la forme canonique *Stéphanois*, et de façon symétrique, la reconnaissance d'une des quatre graphies du nom de localité, la production de l'identifiant de la forme canonique *Saint-Etienne*. L'association entre ces deux identifiants étant gérée par la base de données, toutes les formes du nom d'habitants peuvent être associées à toutes les formes du nom de localité.

De plus, en ne gérant les relations que sur les formes canonique, on minimise le volume des informations manipulées en conservant toutes la sémantique des coréférences.

4.2.1.4.2 Reconstruction des formes fléchies.

Si la seule forme canonique permet de traiter les coréférences, en revanche il est indispensable que nous puissions générer à partir de la base de données les listes des noms de lieux et d'habitants dans toutes leurs flexions. Comme nous le détaillerons plus loin, ces listes seront utilisées pour construire les outils de reconnaissance appropriés.

Pour les noms propres ayant des formes fléchies, (gentilés) le processus est le suivant :

Les tables contenant les modes de flexions et les codes de genre et nombre sont présentes[11] dans la base de données :

- à chaque nom de lieu est associé un code de genre et nombre,
- à chaque gentilé est associé un code de flexion renvoyant aux suffixes des formes fléchies et aux codes de genre et nombre.

Par exemple pour les gentilés, voici la structure de la relation **Habitants Localité** :

H_Loc (No_Hlo, Nom_Hlo, Code_Flex)

et un extrait de la table concernant les habitants d'**Abondant** dans l'**Eure et Loir**

No_Hlo	Nom_Hlo	Code_Flex
12.10003	Abondancier	7

Figure n°32 : Un extrait de la table H_Loc (habitant de Localité)

Le **Code_Flex** (7) renvoie à la relation **Flex_Gent** (Flexion des Gentilés) qui a la structure suivante :

Flex_Gent (Code_Flex, Suffixe,Offset[12], GN)

et dont voici l'extrait de la table concernant les quatre formes du code flexion N° 7 :

Code_Flex	Suffixe	Offset	GN
7	er	0	1
7	ère	2	2
7	ers	2	3
7	ères	2	4

Figure n°33 : Un extrait de la table Flex_Gent (Flexion des Gentilés)

Ces informations permettent, à partir de la **forme canonique** stockée dans la base de données de produire de façon automatique (requête SQL) une liste de la forme :

No_Hlo, Forme Fléchie, Code_GN

```
12.10003,Abondancier,1
12.10003,Abondancière,2
12.10003,Abondanciers,3
12.10003,Abondancières,4
```

Les codes GN (Genre et Nombre) 1,2,3,et 4 renvoient respectivement aux lignes de la table **Genre_Nombre**:

---

[11] Elles ne figurent pas dans le Modèle Conceptuels des Données pour ne pas alourdir la présentation, mais sont décrite dans le Modèle Relationnel.

[12] Valeur utilisée pour la reconstruction des formes fléchies à partir de la forme canonique

Code_GN	Genre_Nombre
0	Indéterminé
1	MS
2	FS
3	MP
4	FP

Figure n°34 : Un extrait de la table Genre_Nombre

4.2.1.4.3 Le problème des formes fléchies mixtes

Deux codes de flexions posent un problème particulier de reconstruction des formes fléchies: les Codes de Flexions 1 et 4 :

N°	Catégorie	Exemples de gentilés	Localités
1	0,e,0,es	Nantais,Nantaise,Nantais,Nantaises	Nantes
../..	../..	../..	../..
4	0,0,s,s	Briviste, Briviste, Brivistes, Brivistes	Brive
../..	....	../..	../..

Extrait de la Figure n° 14

Dans le premier cas le masculin singulier et le masculin pluriel ont la même formes : **Nantais**, dans le second il y a ambiguïté entre masculin singulier et féminin singulier d'une part et entre masculin pluriel et féminin pluriel d'autre part : **Briviste** et **Brivistes**.

Tant que le contexte n'a pas levé l'ambiguïté, il est nécessaire de générer toutes les formes possibles. Pour le premier cas, il faut factoriser le masculin ce qui donne une forme M(S+P), pour le second cela donnera les deux formes (M+F)S et (M+F)P. On doit donc disposer de codes de Genre et de Nombre recouvrant ces situations :

Code_GN	Genre_Nombre
5	M(S+P)
6	F (S+P)
7	(M+F)S
8	(M+F)P
9	(M+F)(S+P)

Figure n°35 : Suite de la table **Genre_Nombre**

La Table des Flexions de Gentilés doit également être complétée pour attribuer aux codes de flexions 1 et 4, la correspondance avec ces nouvelles formes.

Code_Flex	Suffixe	Offset	GN
1		0	5
1	e	0	2
1	es	0	4
4		0	7
4	s	0	8

Figure n°36 : Extrait de la table Flex_Gent pour les codes 1 e 4

On remarquera que, contrairement aux flexions normales qui comportent 4 formes donc 4 lignes dans la table, le code de flexion 1 ne compte que trois lignes (masculin singulier et masculin pluriel confondus) et le code de flexion 4, deux lignes (masculin et féminin singulier d'une part, masculin et féminin pluriel d'autre part également confondus). Ainsi, à partir des formes canoniques, et en utilisant la même technique que pour les autres formes, on reconstruit les lignes suivantes:

```
12.0015,Nantais,5
12.0015,Nantaise,2
12.0015,Nantaises,4
12.1204,Briviste,7
12.1204,Brivistes,8
```

Pour les noms propres qui n'ont pas de flexions (noms de lieux), il n'y a pas de problème de reconstruction des formes fléchies. Comme nous l'avons mis en évidence dans le paragraphe consacré à la morphologie flexionnelle des toponymes, les noms de localités ont un genre et un nombre du type (M+F)S ou (M+F)P selon le contexte d'utilisation. Pour les autres types de noms de lieux (département, régions, etc...) les situations sont très diverses. Ainsi pour les départements : **Calvados** et **Cantal** sont au masculin singulier, **Ardèche** et **Charente** au féminin singulier, **Pyrénées-Atlantiques** et **Deux-Sèvres** au féminin pluriel. En revanche, le genre pour **Aveyron** et **Oise** est indéterminé (M+F). Pour les régions historiques géographiques **Craonnais** et **Condomois** sont au masculin singulier, **Camargue** et **Balagne** au féminin singulier, mais **Charentes** au féminin pluriel (contrairement au nom du département).

Les codes correspondants sont directement gérés dans les tables des différentes entités de la base de données. Chaque table de noms de lieux possède une colonne Code_GN qui renvoie directement à la table **Genre Nombre** sans nécessité de reconstruction.

Dep	Nom_dept	Code_GN
2.01	Ain	1
2.02	Aisne	1
2.03	Allier	2
2.04	Alpes-de-Hautes-Provence	4
2.05	Hautes-Alpes	4

Figure n°37 : Un extrait de la table Département avec la colonne des Codes Genre Nombre

Ces informations ainsi regroupées dans l'ensemble des tables entités permettent de produire de façon automatique la liste générale des noms de lieux dont voici un extrait :

```
1.10002,Bourg-en-Bresse,7
1.10003,Oyonnax,7
1.10004,Hauteville-Lompnes,7
1.10005,Nantua,7
1.10006,Thoissey,7
.../...
2.19,Corrèze,2
2.21,Côte-d'Or,2
2.22,Côtes-d'Armor,4
2.23,Creuse,2
2.24,Dordogne,2
2.25,Doubs,1
```

```
.../...
4.149,Gothie,2
4.150,Catalogne,2
4.151,Salanque,0
4.152,Omois,1
4.153,Comté de Montbéliard,1
4.154,Bas-Poitou,1
```

4.2.1.4.4 Structure des identifiants d'entités

Quand un nom propre a été reconnu, la production de l'identifiant de sa forme canonique est la première action réalisée par le système. Il est donc important de générer le maximum d'informations à ce moment là. Or, dans un système relationnel classique, l'utilisateur sait sur quels objets va porter sa requête au moment où il la construit. Il interroge sa base de données relationnelles sur les CLIENTS, les PRODUITS, les COMMANDES, etc... Notre système travaille de façon tout à fait différente. Si un nom propre a été reconnu, c'est qu'il s'agit d'une des formes possibles de ce nom dont la forme canonique est présente dans une des tables de la base. Avec un identifiant classique, c'est à dire donnant la position de l'occurrence dans la table, on n'avait aucune information sur la table où cette information était stockée, et par conséquent sur le type d'entité auquel on avait affaire.

Il était donc indispensable que l'identifiant émis fournisse directement à notre système d'analyse ce type d'information. Toutes les clés des entités sont préfixées d'un numéro qui fournit le type du nom propre reconnu (nom de département, nom d'habitants de région etc...), ainsi que sa catégorie : toponyme (préfixe <10) ou gentilé (préfixe >=10). Cette double identification permet de réaliser en même temps la **reconnaissance** et le **typage** . Tous les identifiants ont donc la forme :

No Table. No Ligne

exemples :

1.10225
4.156
12.11331
1.11804
2.053

Serrières possède l'identifiant **1.10225**, le préfixe **1** indique le type ***localité*** ; **Arvénie,** l'identifiant **4.156**, le préfixe **4** indique le type ***régions historique/géographiques*** ;**Halluinois** l'identifiant **12.11371, 12** détermine le type ***habitants de localités***,

Mayenne renvoie à deux identifiants : **1.11804** et **2.053**, les préfixes renvoyant respectivement aux types *localité* et *département* (voir le détail Figure 41).

4.2.1.4.5 Le Modèle Conceptuel des données « Noms Propres »

On trouvera en annexe 5 le schéma du Modèle Conceptuel des Données[13]. Le chiffrage de l'ensemble des formes reconnues (formes fléchies, formes élidées et graphies condensées) sera précisé dans la partie consacrée aux outils de reconnaissance et de traitement (transducteurs). Nous indiquons dans l'ordre :

---

[13] Pour ne pas alourdir la représentation graphique, nous n'avons mis que les noms des entités, les associations étant identifiées par les formes abrégées des noms d'entités.

Les Entitées

Nous distinguons au niveau de nos traitements deux catégories d'entités :

1. les ENTITEES **traitées** sur lesquelles nous pensons avoir rassemblé suffisamment d'informations pour entreprendre des traitements de reconnaissance, de typage et de coréférences (en traits pleins sur le schéma) . Nous distinguerons quatre catégories :

- les ***Entités de Toponymes*** (département, localités ...)
- les ***Entités de Gentilés*** (habitants localité, habitants département ...)
- les ***Entités Complémentaires de Toponymes*** (type administratif...)
- les ***Entités Fonctionnelles*** (Flexion Gentilés, Genre_Nombre, Table des préfixes d'identifiants, Type_Elision) qui contiennent les informations morphologiques et flexionnelles de reconstruction ou d'élision ainsi que des information de type.

2. les ENTITES qui ne sont **présentes** que **structurellement**. (en pointillé sur le schéma) et qui devraient être complétées dans la suite du projet PROLEX.

Le modèle présenté obéit à la sémantique suivante[14] :

A - LES ENTITES de TOPONYME (préfixes <10)

LOCALITE :
- Identifiant de localité (format 1.xxxxx)
- Nom de localité
- Code postal

DEPARTEMENT :
- Identifiant de département (format 2.xxxxx)
- Nom de département

REGION ADMINISTRATIVE :
- Identifiant de région administrative (format 3.xxxx)
- Nom de région administrative

REGION HISTORIQUE/GEOGRAPHIQUE
- Identifiant de région historique/géographique (format 4.xxxx)
- Nom région historique/géographique

ALIAS LOCALITE
- Identifiant d'alias localité (format : 5.xxxx)
- Nom d'alias localité

Nous mettons dans cette catégorie des toponymes comme **cité phocéenne** pour *Marseille* ou **cité corsaire** pour *Saint-Malo.* Mais nous y ajoutons également :

[14] Il s'agit d'un commentaire du Modèle Conceptuel des Données

• toutes les formes élidées non ambiguës, c'est à dire celles qui ne font référence qu'à une et une seule localité, dans l'exemple de l'article : **Thouaré**, alias de *Thouaré-sur-Loire* (il n'y a qu'un seul *Thouaré-sur-Loire* en France).

• toutes les formes élidées tombées dans l'usage courant et parfois davantage utilisées que la forme complète comme **Charleville** pour *Charleville-Mézière* ou **La Baule** pour **La Baule-Escoublac.**

SOUS-ENSEMBLE DE LOCALITE
• Identifiant de sous-ensemble de localité (format :6.xxxx)
• Nom de sous-ensemble de localité

On range dans cette catégorie les quartiers des grandes villes, les noms des lieux célèbres, c'est à dire dont l'évocation entraîne immédiatement l'association avec la localité : **Montmartre** pour **Paris**, la **gare Saint-Charles**, la **Canebière** pour **Marseille**, la **place Stanislas** pour **Nancy** etc...

SITES HISTORIQUE / GEOGRAPHIQUE
• Identifiant de site historique / géographique (format :8.xxxx)
• Nom de site historique / géographique

Il s'agit de collecter tous les noms de lieux célèbres qui ne sont pas identifiables par la présence d'un nom de localité comme:

falaises d'Etretat, château de Fontainebleau, baie de Douarnenez;

ou bien d'un nom de département ou de région comme:

golfe du Morbihan, presqu'île Guérandaise, massif de l'Esterel.

On y trouve donc des noms de sites célèbres comme : le **col Bayard**, la **pointe du Raz**, **Alésia** ou **Utha-Beach**. Ils sont en général associés à la localité voisine, à la région historique géographique, au département dans lesquels ils sont situés. A ce titre, le traitement d'éventuelles coréférences de diffère pas des autres noms de lieux.

B - LES ENTITES de GENTILES (préfixes >=10)

HABITANT REGION ADMINISTRATIVE
• Identifiant habitant de région administrative (format 10.xxxx)
• Nom d'habitant de région administrative (forme canonique)

HABITANT DEPARTEMENT
• Identifiant habitant de département (format 11.xxxx)
• Nom d'habitant de département (forme canonique)

HABITANT LOCALITE :
• Identifiant habitant de localité (format 12.xxxx)
• Nom d'habitant (forme canonique)

ALIAS HABITANTS LOCALITE
• Identifiant d'alias habitants de localité (format 13.xxxx)

• Nom d'alias habitants (forme canonique)

Concernant les noms d'habitants, l'ALIAS constitue un « *Autrement dit...* » dont l'usage est courant et la forme non métaphorique. Ainsi, nous considérons **Phocéens** comme un alias de **Marseillais**, ce qui ne sera pas le cas de « **habitants de la capitale des Gaules** » pour **Lyonnais**. De la même façon, les graphies multiples des noms de lieux (Saint → St) ne constituent pas des alias, elles font l'objet d'un traitement spécifique que nous présenterons dans le chapitre consacré aux techniques de reconnaissance.

Nous disposons d'une liste de 22 alias qui nous ont tous été transmis par l'enquête. Nous n'avons pas pu croiser l'ensemble de ces informations avec d'autres sources. Elles doivent donc être considérées avec prudence. Dans la plupart des cas, il s'agit d'un nom hérité d'une ancienne dénomination de la localité, d'une particularité locale ou d'un événement du passé particulièrement marquant. Par exemple les **Contois**, habitants de **Conte** dans les **Alpes-Maritimes** ont le surnom de **Lu Tremp'oli** (trempeur d'huile) allusion à une vieille histoire de fabrication d'huile d'olive frelatée ; dans le même département, les **Lantosquois**, habitants de **Lantosque** ont le sobriquet de **Cougourdiès** à cause du produit local, la courge (en dialecte, cougourda). Les **Pontilliaciens**, de **Pontailler-sur-Saône** dans le **Douds** se surnomment les **Culs-Frits**, souvenir d'un incendie qui ravagea totalement la localité sous l'occupation romaine, etc...

HABITANT REGION HISTORIQUE/GEOGRAPHIQUE
• Identifiant d'habitants de région historique/géographique (format : 14.xxxx)
• Nom d'habitant de région historique/géographique (forme canonique)

HABITANT DE SOUS-ENSEMBLE DE LOCALITE
• Identifiant d'habitant de sous-ensemble de localité (format :15.xxxx)
• Nom d'habitant de sous-ensemble de localité

C - LES ENTITES COMPLEMENTAIRES de TOPONYMES (pas de préfixe)

TYPE ADMINISTRATIF DE LOCALITE
• Identifiant du type administratif [15]
• Libellé de type administratif

Nous avons attribué à chaque localité traitée un des cinq types administratifs suivants :

Identifiant du type	Libellé du type
0	commune
1	chef lieu de canton
2	cheflieu d'arrondissement
3	chef lieu de département
4	chef lieu de région

Figure n°38 : Les cinq types administratifs gérés

[15] Les types administratifs n'ayant pas vocation a être « reconnus » au cours du traitement, le format de l'identifiant n'est pas préfixé.

L'acquisition de ces informations au moment du traitement de coréférences est susceptible d'aider à la levée de certaines ambiguïtés.

TYPE GEOGRAPHIQUE DE LOCALITE
- Identifiant de type géographique[16]
- Nom de type géographique

Il s'agit à moyen terme, d'attribuer à chaque localité un ou plusieurs types susceptibles d'apporter des informations complémentaires au niveau du traitement des coréférences. Les types recensés sont les suivants :

No de type	Nom de type
01	lieu-dit
02	village
03	ville moyenne
04	ville
05	agglomération
06	banlieue
07	faubourg
08	port maritime
09	port fluvial
10	port d'estuaire
11	station de sports d'hiver
12	station balnéaire
13	station thermale

Figure n°39 : Les différents types de localités

D - LES ENTITES FONCTIONNELLES :

FLEXIONS DE GENTILES[17]
- Code de Flexion
- Suffixe
- Offset
- GN (Code genre et nombre)

GENRE et NOMBRE
- Code_GN
- Genre_Nombre

[16] Voir note 13

[17] La table complète est en annexe 6

Code_GN	Genre_Nombre
0	Indéterminé
1	MS
2	FS
3	MP
4	FP
5	M(S+P)
6	F (S+P)
7	(M+F)S
8	(M+F)P
9	(M+F)(S+P)

Figure n°40 : Table des Codes de Genre et Nombre

TYPE NOM PROPRE
- No_Type (préfixe de l'identifiant des entités)
- Libellé_Type

No_Type	Libellé_Type
1	Localité
2	Département
3	Région Administrative
4	Région Historique et Géographique
5	Alias Localité
6	Sous Ensemble Localité
7	Elision de localité
8	Site Historique Géographique
10	Habitant Région Administrative
11	Habitant Département
12	Habitant Localité
13	Alias Habitant Localité
14	Habitant Région Historique et Géographique
15	Habitant Sous Ensemble Localité

Figure n°41 : Préfixes des identifiants d'entités (Toponymes <10 ; Gentilés >=10)

TYPE ELISION
- No_Type_Elis
- Règle d'Elision

E - ENTITES non traitées:

ENTITES HYDROGRAPHIQUES
TYPES D'ENTITES HYDROGRAPHIQUES
EVENEMENTS
PERSONNAGES
CATEGORIE PERSONNAGES
ŒUVRES OU REALISATIONS
PRODUITS

Les Associations

Comme on le voit sur le schéma (annexe 5), l'une d'entre elles a un statut particulier :

LOC / AGGLO entre **Localité** et **Localité.** C'est une association reflexive qui fournit les indications suivantes :

- une localité appartient ou non à une agglomération (0,1)
- une localité est une agglomération comprenant plusieurs autre localités (1,n)

La notion d'agglomération est définie dans le Dictionnaire National des Communes de France. Pour chacune d'entre elles, la liste des communes constitutives est donnée.

LOC / SEL entre **Localité** et **Sous-ensemble de Localité**

- Une localité peut ne pas avoir de sous-ensemble, elle peut en avoir de 1 à plusieurs (0,n),
- Un sous-ensemble de localité appartient à une ou plusieurs localités (1,n) (il s'agit ici de traiter les homonymies dans les noms de quartier, un nom de quartier pouvant renvoyer à des communes différentes).

SEL / H_SEL entre **Sous-ensemble localité** et **Habitants sous-ensemble localité**

- un sous-ensemble de localité possède ou non un ou plusieurs noms d'habitants (0,n)
- un nom d'habitant de sous-ensemble de localité peut renvoyer à un ou plusieurs sous-ensembles (1,n)

LOC / T_E_LOC entre **Localité** et **Type Elision Localité.**

- une localité a ou non une forme élidée (0,1)
- une forme élidée de localité correspond de une à plusieurs localités (1,n)

LOC / DEP entre **Localité** et **Département.**

- une localité est située dans un et un seul département (1,1)
- un département peut compter de un à plusieurs localité (1,n)

DEP / RA entre **Département** et **Région Administrative**

- un département est situé dans une et une seule région administrative (1,1)
- une région administrative peut compter de un à plusieurs départements (1,n)

LOC / RHG entre **Localité** et **Région Historique / Géographique**

- une localité peut être située ou non dans une région historique / géographique (0,1)
- une région historique / géographique compte de une à plusieurs localités (1,n)

RA / H_RA entre **Région Administrative** et **Habitant Région Administrative**

- une région administrative possède ou non un ou des noms d'habitants (0,n)

• un nom d'habitant de région administrative correspond à une ou à plusieurs régions administratives (1,n)

DEP / H_DEP entre **Département** et **Habitant de Département**

• un département peut avoir ou non un ou plusieurs noms d'habitants (0,n)
• un nom d'habitant de département correspond à un ou plusieurs départements (1,n)

LOC / A_LOC entre **Localité** et **Alias Localité**

• une localité peut on non avoir de un à plusieurs alias (0,n)
• un alias de localité correspond de une à plusieurs localités (1,n)

LOC / T_A_LOC entre **Localité** et **Type Administratif Localité**

• une localité correspond à un et un seul type administratif principal (1,1)
• un type administratif correspond de une à plusieurs localités (1,n)

LOC / H_LOC entre **Localité** et **Habitant de Localité**

• une localité peut ne pas avoir de nom d'habitants, elle peut en avoir un ou plusieurs (0,n)
• un nom d'habitant correspond a une ou plusieurs localités (1,n)

RHG / H_RHG entre **Région Hist. Géo.** et **Habitant Région Hist. Géo.**

• une région hist. géo. peut ne pas avoir de gentilé, elle peut en avoir plusieurs (0,n)
• un gentilé de région hist. géo. correspond de une à plusieurs régions hist géo (1,n)

H_LOC / A_H_LOC entre **habitants localité** et **Alias habitants localité**

• un nom d'habitant de localité possède de zéro à plusieurs alias (0,n)
• un alias d'habitant de localité correspond à un ou à plusieurs noms d'habitants (1,n) (traitement d'éventuelles homonymies)

4.2.1.4.6 Le Modèle Relationnel « Noms Propres »

Relations construites sur les Entités Réelles :

Habitant sous-ensemble localité
H_SEL (No_HSel, nom,#Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Sous-ensemble localité
SEL (No_Sel, nom, #Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Localité
LOC (No_Loc , Localité, Code_Postal, #No_dept, #No_TLoc , #Code_GN,#No_Type_Elis)
#No_dept : clé externe renvoyant à la relation DEP

#No_Tloc : clé externe renvoyant à la relation T_A_Loc
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb
#No_Type_Elis : clé externe renvoyant à la relation fonctionnelle Type_Elision

Département
DEP(No_dept , Département, #No_Ra ,#Code_GN)
#No_Ra :clé externe renvoyant à la relation RA
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Région_administrative
RA (No_Ra, Région_Administrative,#Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Habitant_région_administrative
H_RA(No_HRa, Habitant_Ra,#Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Habitants département
H_DEP (No_HDep, Habitant_Dep, #Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Alias Localité
A_LOC (No_ALoc, Alias_Loc, #Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb
Type Administratif de Localité
T_A_LOC(No_TLoc, Type_Adm)

Alias Habitant Localité
A_H_LOC (No_AHLoc, Alias_HLoc, #Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Habitant Localité
H_LOC (No_HLoc, Habitant_Loc, #Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Habitant Région Historique Géographique
H_RHG (No_HRhg, Habitant_Rhg, #Code_Flex)
#Code_Flex : clé externe renvoyant à la relation fonctionnelle Flex_Gent

Région Historique Géographique
RHG (*No_RHG*, Région_Hg, #Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Site Historique Géographique
SHG (No_Site, Site_Hg, #Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Relations construites sur les Entités Fonctionnelles :

Flexion Gentilés

Flex_Gent (Code_Flex, Suffixe, Offset, GN)

Genre et Nombre
Gr_Nb (Code_GN, Genre_Nombre)

Type Nom Propre
Type_NP(No_Type, Libellé_Type)

Type Elision
E_LOC (No Type Elis, Règle_Elision, #Code_GN)
#Code_GN :clé externe renvoyant à la relation fonctionnelle Gr_Nb

Relations construites sur les Associations :

Habitant sous-ensemble localité / Sous-ensemble localité
H_SEL / SEL (No_HSel,No_Sel)

Sous-ensemble localité / Localité
SEL / LOC (No_sel, No_Loc)

Région_administrative / Habitant_région_administrative
RA / H_RA (No_Ra, NO_HRa)

Département / Habitants département
DEP/H_DEP (No_Dep, No_HDep)

Localité / Alias Localité
LOC / A_LOC (No_Loc, No_ALoc)

Habitant Localité / Alias Habitant Localité
H_LOC / A_H_LOC (No_Loc, No_AHLoc)

Localité / Habitant Localité
LOC / H_LOC (No_HLoc, No_Loc)

Région Historique Géographique / Habitant Région Historique Géographique
RHG / H_RHG (No_RHG, No_HRhg)

Localité / Région Historique Géographique
LOC / RHG (No_Loc, No_RHG)

Agglomération / Localité
AGGLO / LOC (No_Loc, No_Loc)

Localité / Site Historique Géographique
LOC / SHG (No_Loc, No_Site)

4.2.2 Bases de données relationnelles et TAL

Les bases de données relationnelles sont des outils informatiques qui ne sont pas adaptés aux processus de traitement automatique du langage naturel pour deux raisons principales :

• la première est d'ordre physique : la plupart du temps, le texte qui est soumis à une telle analyse est visité caractère par caractère, notamment pendant la phase dite « d'étiquetage » qui initie l'analyse lexicale. Cela suppose que les outils informatiques soient « en ligne », c'est à dire présents en mémoire centrale. Or, les bases de données sont en général sur disque et leur consultation nécessite des accès qui sont généralement coûteux.

• la seconde est d'ordre logique : on ne peut interroger une base de données relationnelle que si on a une idée précise de la nature de l'information que l'on souhaite extraire. A la reconnaissance d'une occurrence d'entité dans un texte, toutes les autres entités de la base peuvent être candidates à une association. Mais pour chacune d'entre elles, il faut construire une requête spécifique qui devra prendre en compte la nature de la sémantique que l'on souhaite extraire ainsi que les contraintes de navigation propres au modèle relationnel. Comment anticiper toutes les situations possibles ? On ne peut pas envisager de « préconstruire » des requêtes adaptées à chaque cas particulier. Pour reprendre l'exemple de l'article cité plus haut, à la détection du nom propre *La Roche-sur-Yon,* cinq entités sont candidates à une association : *Vendée* (département), *La Roche* (élision), *Yonnais* (habitant localité), *Vendée* (région historique géographique), *Saint-André-d'Ornay* (sous ensemble localité). Mais pour un autre nom de localité (*Nozay*), les associations pourront être différentes. Dès lors, il apparaît que la consultation d'une base de données relationnelle avec un langage de requête ne constitue pas une solution satisfaisante pour extraire la sémantique relationnelle qui existe entre les noms propres.

Tout ceci nous a conduit à envisager l'utilisation d'une autre catégorie d'outils, mieux adaptés aux problèmes de l'analyse automatique : les **automates** et les **transducteurs à nombre fini d'états**. Cependant, ne voulant pas perdre toute la sémantique relationnelle contenue dans la base de données, nous avons mis au point un tansducteur d'identifiants qui permet de stocker et l'exploiter « en ligne » les informations relationnelles nécessaires au traitement des coréférences.

5. Technique des automates et des transducteurs

Les automates et les transducteurs à nombre fini d'états constituent des outils particulièrement bien adaptés à la représentation et aux traitements des phénomènes linguistiques. Cela a été mis en évidence par de nombreuses applications parmi lesquelles on notera [Laporte 88] et [Kaplan & Kay 94] pour la phonologie, [Silberztein 89] pour la morphologie, [Maurel & Mohri 94] pour les grammaires locales.

5.1 Automates à nombre fini d'états

5.1.1 Arbres lexicographiques et automates

Nous nous proposons de reconnaître les noms de départements figurant dans la liste suivante : *Var, Vaucluse, Vendée, Vienne , Yonne*

Nous ne retenons pas l'idée de pouvoir simplement consulter la liste telle qu'elle est représentée ci-dessus. En analyse automatique, il n'est pas possible d'envisager, à chaque fois qu'un mot sera délimité, de déclencher la consultation de listes dont certaines peuvent atteindre des dimensions très importantes (pour notre part, la seule liste des localités traitées contient plus de 38000 noms).

Les arbres sont des structures de données très courantes en informatique, qui permettent, entre autre, de représenter une liste de la façon suivante :

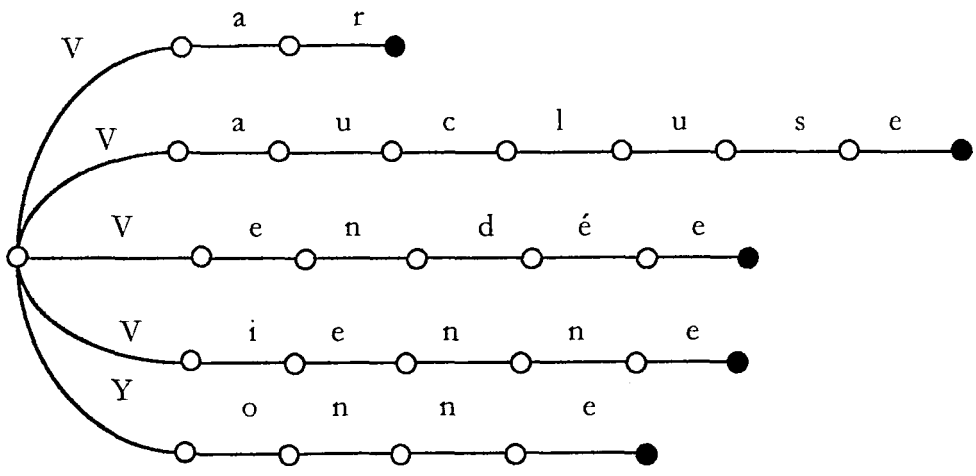


Figure n°42 : Un arbre lexicographique

Il s'agit d'un arbre qui comporte des sommets représentés par des cercles et des chemins correspondants aux traits allant d'un sommet à un autre. C'est également un automate à nombre fini d'états qui, lorsqu'on le parcourt caractère par caractère, permet de reconnaître les noms des cinq départements de la liste. Pour la structure d'automate on parle alors d'états (sommets) et de transitions (chemins). Les transitions sont étiquetées par une des lettres des différents mots à reconnaître. La visite de l'automate commence à l'état initial. Lors de la visite du texte à analyser, à la rencontre de la lettre « V », l'automate est alors parcouru en réalisant des comparaisons lettre par lettre. Tant que ces comparaisons réussissent, le parcours continue. S'il se termine à la fin du mot par la rencontre d'un état final (cercle noir), le mot a été « reconnu » par l'automate. Dans tous les autres cas, la reconnaissance a échoué.

Définition :

On appelle automate à nombre fini d'états sur un alphabet Λ le quintuplet :

$$A = \{ \Lambda, E, I, \Phi, T \} \text{ où}$$

- Λ est un ensemble non vide
  (dans l'exemple ci-dessus : {*a, c, d, e, é, g, i, l, n , o, r, s, u, V,Y* })
- E est un ensemble non vide d'états
- I est un sous-ensemble non vide de E constitué des états initiaux
- Φ est un sous-ensemble non vide de E constitué des états finaux
- T un ensemble fini non vide de triplets de E x Λ x E : les transitions

L'ensemble des chaînes reconnues par A est appelé le langage reconnu par A. On le note L(A).

5.1.2 Automates déterministes

L'automate, tel qu'il est représenté dans la figure précédente ne peut pas fonctionner de façon satisfaisante. Comme on peut le constater, à la sortie de l'état initial, il y a cinq transitions

qui sont toutes étiquetées avec le même caractère : « V ». Dès lors, si pour reconnaître le mot *Vendée*, nous nous engageons dans la première des transitions (celle qui conduit à *Var*), le mot *Vendée* ne sera pas reconnu, alors qu'il est présent dans l'automate.

Il est donc évident que pour éviter ce phénomène, à la sortie d'un état, il ne peut y avoir que des transitions étiquetées avec des caractères différents. Si de plus, l'automate n'a qu'un seul état initial et aucune transition vide, on dit alors que l'automate est **déterministe**.

Définition :

Un automate est dit déterministe si et seulement si, il vérifie les trois conditions suivantes :

- il n'a qu'un seul état initial
- il n'a pas de transitions vides
- pour tout état de l'automate, toutes les transitions qui en sont issues ont des étiquettes distinctes

Comme on peut le constater Figure 43, la règle énoncée pour la lettre « V » a également été appliquée pour la lettre « a » qui est commune à *Var* et à *Vaucluse*.

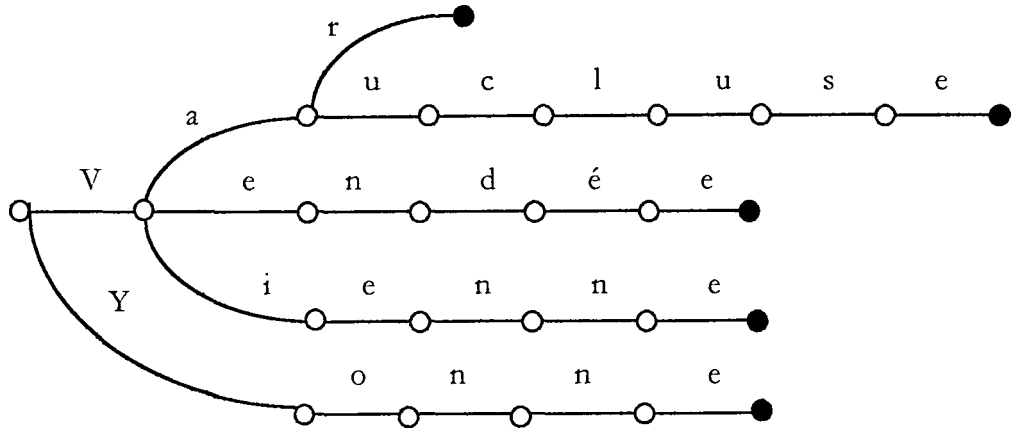


Figure ° 43 : Un automate déterministe

5.1.3 Automates déterministes minimaux

Le processus a consisté à mettre en commun les débuts de mots qui étaient identiques. On parle alors de *préfixe* commun. Or, lorsque l'on considère la fin des mots, c'est à dire leur *suffixe*, on remarque qu'il est possible de réduire le nombre d'états. Ainsi, dans notre exemple, quatre noms de département se terminent par la même lettre « **e** », une factorisation peut également être envisagée pour les « **nne** » qui finissent les mots « **Vienne** » et « **Yonne** ».

Cet automate s'obtient par construction à partir de l'algorithme dit de *pseudo-minimisation* de D. Revuz [Revuz 91].

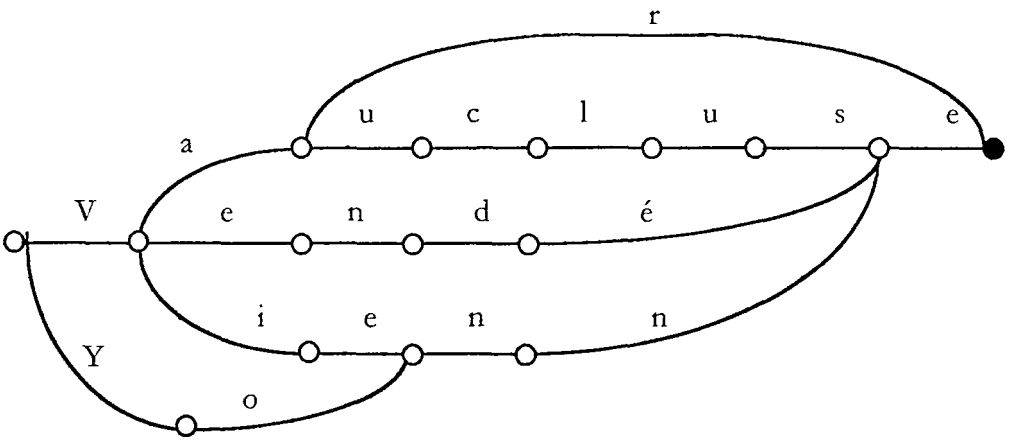


Figure n°44 : Un automate déterministe minimal

5.1.4 L'algorithme de pseudo-minimisation

Le principe de cet algorithme est le suivant. A partir d'un automate vide, les transitions et les états d'un premier mot sont mis en place. Dans notre exemple : *Var.* Pour le mot suivant, *Vaucluse,* on recherche le plus grand préfixe commun : *Va.* Ensuite, l'algorithme va rechercher le plus grand suffixe commun. Ce suffixe n'existant pas, les états et les transitions permettant de reconnaître *ucluse* sont créés. Pour le mot suivant, *Vendée,* la partie préfixe reconnue sera *V* , la partie suffixe sera *e,* ces deux éléments reconnus seront alors reliés par la construction des transitions et des états permettant de reconnaître *endé.* Et ainsi de suite jusqu'à la fin de la liste.

5.2 Transducteurs

Les automates sont des outils informatiques particulièrement bien adaptés à la reconnaissance des mots (des noms de départements dans les exemples précédents). En traitement automatique du langage naturel, il est souvent nécessaire, quand un mot a été reconnu, qu'une ou plusieurs autres informations soient alors émises, par exemple le genre, le nombre ou la catégorie lexicale ; pour nous la reconnaissance d'un nom de département pourrait entraîner l'émission du nom de ses habitants au masculin singulier.

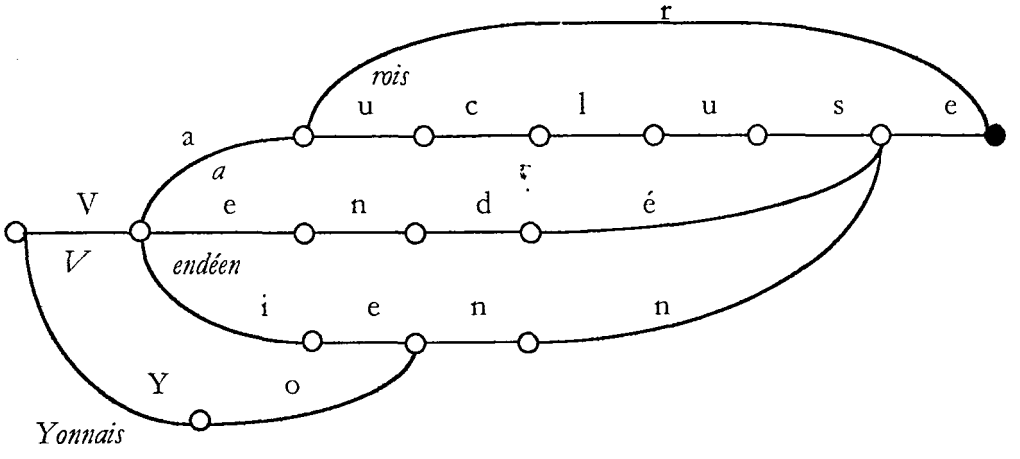


Figure n°45: Un transducteur associant des noms d'habitants (s'ils existent) à des noms de départements

Pour minimiser un transducteur [Mohri, 94], on génère la sortie dès qu'il n'y a plus d'ambiguïté. Pour Var et *Varois*, le *V* est émis de suite, le *a* ensuite et la fin de l'émission *rois* s'effectue en une seule fois sur la reconnaissance de la lettre « r ».

Définition :

Un transducteur sur Λ est un septuplet A= { Λ, Σ, E, I, Φ, T, σ} où

- Λ est un ensemble non vide de lettres

(il s'agit de l'aphabet, dans l'exemple ci-dessus : *a, c, d, e, é, g, i, l, , o, r, s, u, V*)

- Σ un ensemble non vide d'émissions (*V, a, rois, uclusien, endéen, iennois, osgien*)
- E est un ensemble non vide d'états
- I est un sous ensemble non vide de E constitué des états initiaux
- Φ est un sous ensemble non vide de E constitué des états finaux
- T un ensemble fini non vide de quadruplets de E x Λ x Σ x E, les transitions
- σ une application de Φ dans Σ qui a tout élément de Φ fait correspondre un et un seul élément dans $\Sigma \cup \{\varepsilon$ [18]$\}$

[Chauvier et Maurel, 96] donnent un algorithme de construction de transducteur basé sur les algorithmes de Revuz et Mohri.

6. Première phase de traitement : reconnaissance, typage

L'intérêt des transducteurs réside dans leur capacité à réaliser conjointement deux actions : reconnaître des mots et générer des informations. En TAL, les transducteurs sont généralement utilisés pour associer une information émise à un mot reconnu (dans l'exemple précédent : nom de département, nom d'habitants).

6.1 Le Transducteur d'identifiants

La mise au point de nos outils de traitement automatique des toponymes et des gentilés nécessite que l'on réponde à deux questions :

1. quels mots voulons nous reconnaître ?
2. quand ils ont été reconnus, quelles informations voulons nous émettre ?

La réponse à la première question est moins simple qu'il n'y paraît. Certes, nous devons reconnaître tous les noms de lieux et d'habitants qui sont présents dans notre base de données. Mais celle-ci, qui a également vocation à gérer les relations entre les différents types de noms de lieux et d'habitants, ne contient que les formes dites « canoniques ». Par exemple, pour les habitants de la *Vendée*, seule la forme du nom d'habitant au masculin singulier est présente. En revanche, le transducteur devra reconnaître toutes les formes fléchies de gentilés ainsi que toutes les abréviations de toponymes.

Pour répondre à la deuxième question, la reconnaissance d'un des noms propres gérés par notre système entraînera l'émission au moins d'un type, d'un genre et d'un nombre mais aussi, comme nous le verrons au chapitre 7, d'informations relationnelles.

---

[18] Transition vide

6.1.1 Transducteur d'identifiant et de genre et nombre

Pour réaliser la construction d'un transducteur, il est nécessaire de faire un certain nombre de choix qui prennent en compte plusieurs paramètres :

- nature des informations à reconnaître
- structures et volume des informations à émettre
- facilité de construction et mise à jour du transducteur (surtout quand le transducteur doit gérer les informations d'une base de données relationnelle)

Voici les choix que nous avons faits et leur justification sur le plan logiciel :

- les informations sur le ou les identifiants associés à une forme canonique sont générées pendant le parcours de reconnaissance, **en une seule fois** selon deux modes d'émission :

  1 - sur la première transition suivant le préfixe
  2 - sur l'état de sortie quand un nom est inclus dans un autre (***Vendéens, Vendée***)

- les informations de genre et de nombre sont sur les états terminaux.

Nous n'avons donc pas repris la technique habituelle des transducteurs qui consiste à émettre au fur et à mesure du parcours des transitions, l'information associée aux caractères reconnus. Celle-ci présente pourtant l'avantage de pouvoir factoriser les émissions quand elles ont des éléments communs.

Nous justifions notre choix:

A - L'émission de l'identifiant sur une seule transition présente l'avantage de pouvoir se faire par adresse. La structure des données correspondant à cet identifiant est donc totalement indépendante des informations reconnues par le transducteur, ce qui permet un stockage sous une forme réduite. Ainsi l'identifiant de *Frapelle* dans les *Vosges* : **1.32652** est traduit sous forme de trois caractères ASCII :

1 - préfixe de l'identifiant : 1 → transformation en caractère ASCII 001
2 - suffixe de l'identifiant : 32652 :

  2.1 - stockage du poids fort (32652 DIV 256) →ASCII 127
  2.2 - stockage du poids faible (32652 MOD 256) →ASCII 140

au moment de l'émission par le transducteur l'identifiant est reconstruit par calcul.

B - Comme nous le verrons dans la suite de cet exposé, l'information relationnelle sera rattachée aux identifiants des toponymes et des gentilés. La technique que nous utilisons permet une gestion indépendante de cette information relationnelle, c'est à dire sous forme de listes chaînées « accrochées» par adresse, à chaque identifiant d'entité.

C - Enfin, la partie **Entités** (reconnaissance,typage) étant structurellement indépendante de la partie **Associations** (sémantique relationnelle), des mises à jour peuvent être réalisées sans avoir à reconstruire entièrement le transducteur.

Exemple de construction[19] :

Liste générée à partir de la base de données (identifiant, forme fléchie, Genre Nombre) :

11.30,Vendéen,1
11.30,Vendéenne,2
11.30,Vendéens,3
11.30,Vendéennes,4
2.85,Vendée,2
11.32,Yonnais,5
11.32,Yonnaise,2
11.32,Yonnaises,4
2.89,Yonne,2

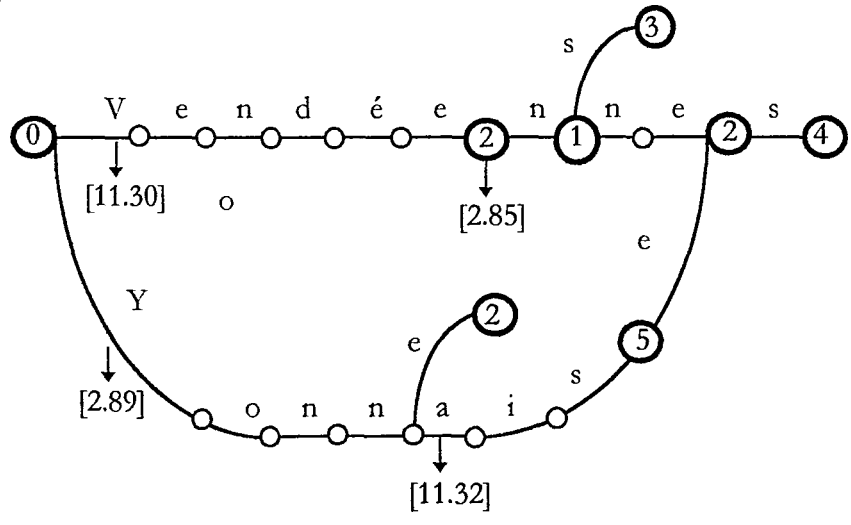


Figure n°46 : Un transducteur d'identifiants et de codes de Genre/Nombre

Lors du processus de reconnaissance, quand une transition pointe sur un identifiant, celui-ci est « ramassé » et conservé jusqu'à la rencontre d'un état final. Plusieurs identifiants peuvent être rencontrés pendant ce processus. Dans ce cas, c'est le dernier trouvé sur une transition qui est valide. Il existe cependant une exception à cette règle générale : si un identifiant est associé à l'état final, c'est ce dernier qui sera conservé.

Voyons le premier cas qui correspond à ce que nous avons appelé un mot inclus dans un autre. Comme on peut le voir sur la figure 46, au début d'une reconnaissance d'un des noms propres appartenant à la liste « *Vendée, Vendéen, Vendéens, Vendéenne, Vendéennes* », c'est l'identifiant de la première transition qui est chargé ([11.30]). Si le processus de reconnaissance s'arrête à la fin du mot *Vendée*, l'identifiant présent sur l'état final lui sera substitué. La sortie du transducteur s'effectuera avec le triplet :

2.85, Vendée, 2

[19] Les identifiants sont représentés sous leur forme numérique pour faciliter la lecture des schémas

qui contient les informations suivantes :

- type : nom de département (préfixe 2)
- mot reconnu : *Vendée*
- genre et nombre : féminin, singulier

Si le processus de reconnaissance se poursuit au delà, selon le mot reconnu, la sortie s'effectuera avec l'un des triplet suivants :

11.30, Vendéen, 1
11.30, Vendéenne, 2
11.30, Vendéens, 3
11.30, Vendéennes, 4

avec, pour le premier des triplets ci-dessus :

- type : nom d'habitant de département (préfixe 11)
- mot reconnu : *Vendéen*
- genre et nombre : masculin, singulier

Dans le second cas, lors de la reconnaissance d'un des noms propres appartenant à la liste *Yonne, Yonnais, Yonnaise, Yonnaises*, c'est l'identifiant du département qui est chargé sur la première transition et qui sera conservé jusqu'à la fin de la reconnaissance de *Yonne*. Dans ce cas, le triplet transmis sera :

2.89, Yonne, 2

qui contient les informations suivantes :

- type : nom de département (préfixe 2)
- mot reconnu : *Yonne*
- genre et nombre : féminin, singulier

Dans les autres cas, l'identifiant présent sur la transition de la lettre « a » est chargé, et la sortie s'effectue avec l'un des triplet suivants :

11.32,Yonnais,5
11.32,Yonnaise,2
11.32,Yonnaises,4

avec, pour le premier des triplet ci-dessus :

- type : nom d'habitant de département (préfixe 11)
- mot reconnu : *Yonnais*
- genre et nombre : masculin, (singulier + pluriel)

Ce transducteur d'identifiants et de codes de flexions réalise donc simultanément les opérations suivantes :

1. reconnaissance du nom propre
2. détermination du type (préfixe de l'identifiant)
3. génération du code de flexion

6.1.2 Problèmes liés à la reconnaissance et au typage

De nombreux problèmes liés à la morphologie, à l'homonymie (2.42 et 2.43) ainsi qu'aux graphies multiples restent encore à résoudre. Nous nous proposons d'illustrer les solutions que nous avons mises en œuvre en reprenant l'exemple de l'article du journal Ouest-France qui nous sert d'exemple tout au long de cet exposé.

6.1.2.1 Construction du transducteur d'identifiants et de code de flexion

Les noms propres à reconnaître sont les suivants : *La Roche, Thouaré, La Roche-sur-Yon, Nozay, Ornaisiens, Vendéens, Thouaréens.* Dans la base de données relationnelle, ils sont stockés en tant qu'entités, de la façon suivante :

Elision Localité *(extrait)* **:**

No_Elis	Elision	#Code_GN
7.10697	La Roche	7

Alias Localité *(extrait*

No_Alias	Alias_Loc	#Code_GN
5.35738	Thouaré	7

Localité *(extrait)* **:**

No_Loc	Localité	Code_Postal	#No_dept	#No_Flex	#Code_GN
1.13077	La Roche-sur-Yon		2.85		7
1.11515	Nozay		2.44		7

Habitant sous-ensemble localité *(extrait)* **:**

No_HSel	nom	#Code_Flex
15.21054	Ornaisien	2

Habitant localité *(extrait)* **:**

No_HLoc	Habitant_Loc	#Code_Flex
12.17624	Thouaréen	2

Habitant département *(extrait)* **:**

No_HDep	Habitant_Dep	#Code_Flex
11.30	Vendéen	2

Habitant région Historique Géographique *(extrait)* **:**

No_HRhg	Habitant_Rhg	#Code_Flex
14.199	Vendéen	2

Figure n°47 : Extraits des tables correspondant aux occurrences d'entités de l'article

On remarque que le nom d'habitant *Vendéen* (forme canonique) est présent dans deux tables différentes. Dans un cas, il désigne les habitants d'un département, dans l'autre ceux d'une région historique et géographique. Ils devront être présents tous les deux dans le transducteur avec leur identifiant respectif qui, selon le cas, peut être associé à des entités différentes. En effet, le département peut s'associer de façon directe avec la région administrative ce qui n'est pas le cas de la région historique géographique (voir annexe 2). Au moment de la reconnaissance des mots, rien ne permet de savoir à quel type de gentilé nous avons affaire dans le texte. Seul, un calcul des coréférences ou une analyse du contexte permettront de lever cette ambiguïté.

A partir des informations présentes dans les tables de la base de données et par association avec celles qui sont stockées dans les entités dites « fonctionnelles », le système doit construire la liste des informations nécessaires à la construction du transducteur. Les tables dans lesquelles le code de genre et de nombre est déjà présent vont pouvoir générer directement la liste. Pour notre exemple c'est le cas de **Elision Localité, Localité, Alias Localité, Région Hist. Géo, Département.** On obtient la liste suivante :

```
7.10697,La Roche,7
5.35738,Thouaré,7
1.13077,La Roche-sur-Yon,7
1.11515,Nozay,7
2.85,Vendée,2
4.134,Vendée,2
```

Pour les autres (**Habitant sous-ensemble localité, Habitant localité, Habitant département, Habitant région Historique Géographique)**, les tables contiennent un code de flexion qui renvoie à la table **Flexion Gentilés.** Par une requête SQL de jointure selon **Code_Flex**, on reconstruit les différentes formes fléchies avec pour chacune d'entre elles le code de genre et de nombre.

No_HLoc	Habitant_Loc	#Code_Flex
../..	../.	../..
12.17624	Thouaréen	2
../..	../.	../..

12.17624,Thouaréen,1
12.17624,Thouaréenne,2
12.17624,Thouaréens,3
12.17624,Thouaréennes,4

La requête de jointure ci-dessous génère la liste des quatres formes fléchies du gentilé Thouaréen

Code_Flex	Suffixe	Offset	GN
.../...	.../...	.../...	.../..
2		0	1
2	ne	0	2
2	s	0	3
2	nes	0	4

```
Select H_LOC.No_HLoc,
Left([Habitant_Loc],(Len([Habitant_Loc])-[C_FLEX].[Offset])) & [Suffixe] AS Forme_Fléchie,
C_FLEX.GN
from C_FLEX,H_LOC
where C_FLEX.[Code Flex] = H_LOC.Code_Flex;
```

Figure n°48 : Requête SQL de reconstruction des formes fléchies de gentilés

La liste complète des formes fléchies de gentilés et des noms de lieux correspondant aux noms propres de l'article est la suivante :

```
15.21054,Ornaisien,1
15.21054,Ornaisienne,2
15.21054,Ornaisiens,3
15.21054,Ornaisiennes,4
12.17624,Thouaréen,1
12.17624,Thouaréenne,2
12.17624,Thouaréens,3
12.17624,Thouaréennes,4
11.30,Vendéen,1
11.30,Vendéenne,2
11.30,Vendéens,3
11.30,Vendéennes,4
14.199,Vendéen,1
14.199,Vendéenne,2
14.199,Vendéens,3
14.199,Vendéennes,4
7.10697,La Roche,7
5.35738,Thouaré,7
1.13077,La Roche-sur-Yon,7
1.11515,Nozay,7
2.85,Vendée,2
4.134,Vendée,2
```

6.1.2.2 Le transducteur d'identifiant et de codes de genre et de nombre

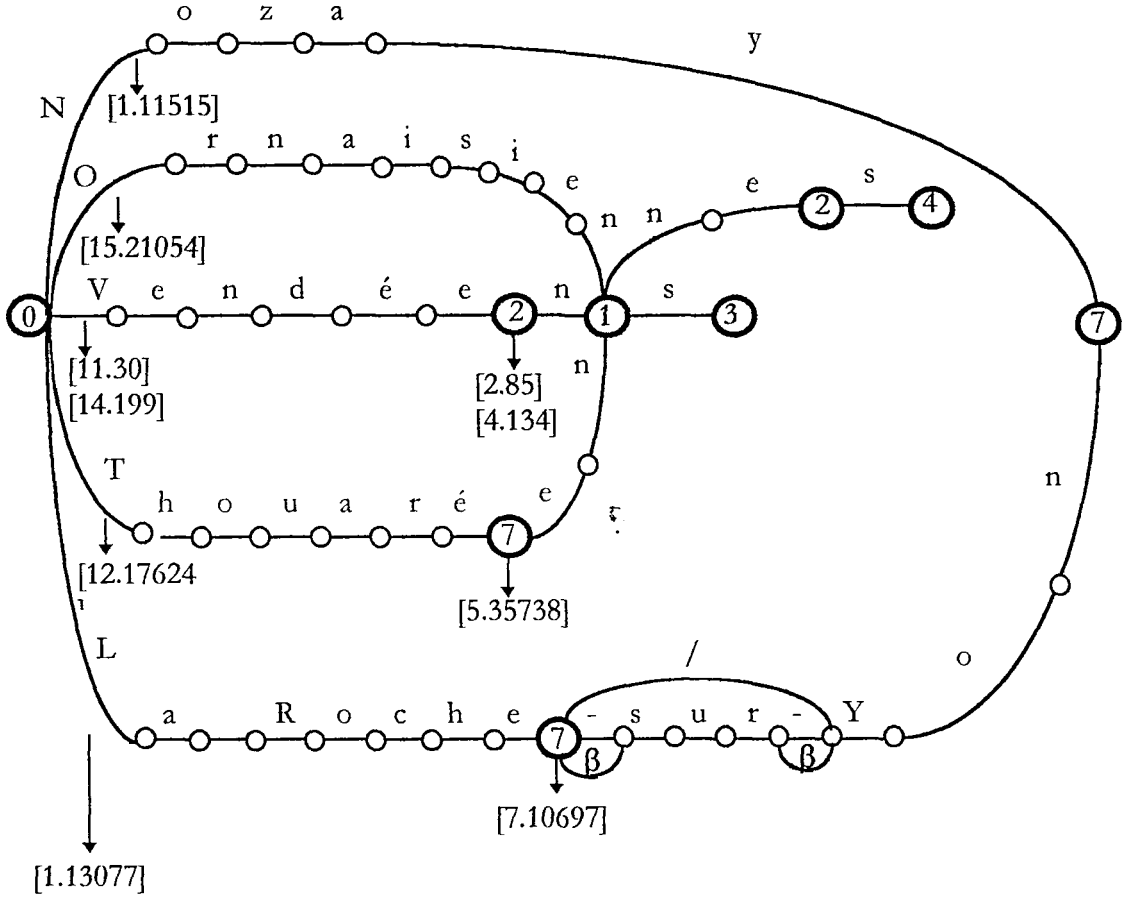


Figure n°49 : Un transducteur reconnaissant les noms propres de l'article d'Ouest-France et générant les identifiants et les codes de genre et de nombre

6.1.2.3 Traitement des particularités morphologiques et flexionnelles

Comme nous l'avions précisé dans le chapitre consacré aux caractéristiques des toponymes et de leurs gentilés certains phénomènes particuliers doivent faire l'objet de traitements automatiques. La plupart d'entre eux sont pris en charge par l'algorithme de construction du transducteur d'identifiants (voir annexe 7). D'autres, que nous préciserons par la suite sont réglés lors de traitements ultérieurs.

6.1.2.3.1 Forme canonique, formes fléchies et identifiant unique

Les relations entre les diverses entités de la base de données relationnelle se font à partir des identifiants sur les formes canoniques. Cependant, le transducteur reconnaît toutes les formes fléchies. Il est donc essentiel que pour **chacune des formes fléchies d'un même nom de gentilé ce soit le même identifiant qui soit généré.**

Notons tout d'abord que, du fait des contraintes d'unicité du modèle, il ne peut pas y avoir deux identifiants identiques pour une entité donnée. De plus, les identifiants étant tous préfixés de façon différente (typage des noms propres), il ne peut pas y avoir deux fois le même identifiant dans l'ensemble des données gérées.

Ainsi, lors de la construction du transducteur, si un identifiant est détecté sur le préfixe, il est comparé à l'identifiant courant. S'il y a égalité, on est dans le cas de figure de construction d'une forme fléchie d'un gentilé, l'une des autres formes étant déjà présente dans le transducteur. Les transitions reliant la fin du préfixe au début du suffixe seront construites sans installation de l'identifiant, celui-ci étant déjà présent.

6.1.2.3.2 Homonymie

Si lors du parcours du préfixe, on arrive sur un état de sortie et que le dernier identifiant trouvé est différent de l'identifiant courant, on est alors dans le cas d'un homonyme. Le nouvel identifiant est alors ajouté (chaînage) sur la même transition que le dernier identifiant trouvé. Ainsi, lors d'une reconnaissance, les deux identifiants seront « ramassés » et transmis à l'analyseur, à charge pour celui-ci de pouvoir lever l'ambiguïté résultant de cette situation. C'est le cas dans notre exemple des gentilés *Vendéen, Vendéenne, Vendéens* et *Vendéennes* qui ont d'abord été construits en tant que gentilés du département, avec la technique des identifiants uniques évoquée au paragraphe précédent. Lors de la tentative de construction du premier des quatre gentilés de la région historique et géographique, l'homonymie sera détectée et l'identifiant correspondant à ce nouveau type sera ajouté sur la même transition que pour les gentilés de département. Pour les trois autres formes, il n'y aura qu'un parcours aboutissant à un état final avec un identifiant déjà présent, donc aucune action de construction.

6.1.2.3.3 Mots inclus et formes élidées

Il existe des cas où un mot est inclus dans un autre. Par exemple, si les gentilés de la *Vendée* tels que nous les avons listés plus haut sont déjà présents dans le transducteur, le mot *Vendée* sera reconnu lors de sa construction comme étant une partie du préfixe d'un des quatre noms d'habitants sans que l'on arrive à un état final. Dans ce cas, un état final particulier est créé avec un identifiant[20] qui, lors de l'exploitation se substitue au dernier identifiant trouvé sur une transition si l'analyse se termine sur cet état. Dans le cas où elle se poursuit au delà, c'est toujours le dernier identifiant d'une transition qui reste actif.

---

[20] Dans le cas ordinaire, l'identifiant n'est pas situé sur un état, mais sur une transition.

Ce traitement prend également en charge le problème des formes élidées de certaines localités. Le mot *La Roche-sur-Yon* étant déjà présent dans le transducteur avec son identifiant [1.13077], *La Roche* est un mot qui correspond au parcours d'une partie du préfixe sans aboutir à un état final. Celui-ci sera donc créé dans les même condition que pour le nom de département *Vendée.*

6.1.2.3.4 Les graphies multiples

Nous avions signalé dans le paragraphe 2.4.2.3 une particularité morphologique des toponymes : les graphies multiples. Il est indispensable que le transducteur puisse « reconnaître » ces différentes formes en générant le même identifiant.

Afin d'éviter les phénomènes de divergence au moment de la construction, on procède à un traitement ultérieur. Pour cela, il est nécessaire que toutes les situations aient été recencées (voir figure 16).

Le cas le plus représentatif est celui du lemme *Saint* qui peut avoir la graphie *St.* Voici une partie d'un transducteur reconnaissant ce lemme et générant un identifiant sur ce parcours.

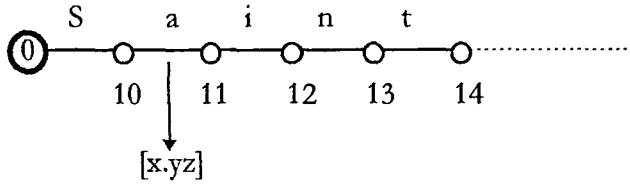


Figure n°50 : Lemme *Saint* avec un identifiant

Les actions à mettre en œuvre sont les suivantes :

- Parcourir le transducteur pour retrouver la ou les occurrences du lemme recherché
- Quand une occurrence est détectée, mémoriser le numéro de l'état suivant le « S » (dans l'exemple état n°10)
- Mémoriser le numéro de l'état suivant la lettre « t » (dans l'exemple état n°14)
- Mémoriser l'adresse d'un éventuel identifiant sur le parcours entre le « S » et le « t » (dans l'exemple l'identifiant [x.yz])
- Construire une nouvelle transition ayant comme état de départ l'état suivant le « S » (10), comme état suivant l'état suivant le « t », comme étiquette la lettre « t » et si un identifiant a été trouvé, l'adresse de celui-ci.

On obtient la configuration suivante :

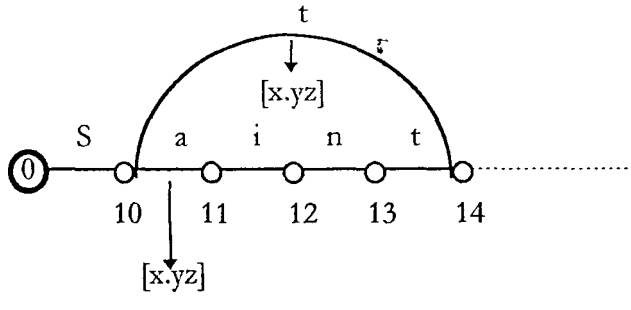


Figure n°51 : Reconstruction de la forme *St*

Dans l'exemple de l'article, le traitement concernait la mise en place des alternatives :

•*-sur-* → / - → β[21]

[21] Blanc

Ainsi, le transducteur est capable de reconnaître toutes les formes suivantes du toponyme :

La Roche-sur-Yon
La Roche sur Yon
La Roche sur-Yon
La Roche-sur Yon
La Roche/Yon

6.1.3 Informations collectées par la première phase d'analyse

Lorsque tout le texte de l'article a été parcouru par le transducteur, nous possédons pour chaque nom propre reconnu :

- le **nom propre** lui-même,
- son **identifiant** (ou ses identifiants en cas d'ambiguïté)
- son **code de Genre et Nombre.**

Ces informations sont rangées dans une liste qui a la structure suivante :

N°	NP reconnu	Identif	Code GN
1	La Roche	7.10697	7
2	La Roche-sur-Yon	1.13077	7
3	Thouaré	5.35738	7
4	Nozay	1.11515	7
5	Ornaisiens	15.21054	3
6	Thouaréens	12.17624	3
7	Vendéens	11.30	3
8	Vendéens	14.199	3

Figure n°52 : Table résultat du transducteur de la figure 49

Par jointure avec les tables d'Entités Fonctionnelles **Genre et Nombre** (selon Code GN) et **Type Nom Propre** (selon le préfixe de l'identifiant), on est en mesure de transmettre à l'analyseur principal les informations telles qu'elles sont présentées figure n°53. A ce stade, aucune information relationnelle n'a encore été prise en compte. Or, celle-ci est présente dans la base de données. Comment l'extraire et l'associer aux données du transducteur ? C'est ce que nous allons aborder dans le chapitre consacré à la deuxième phase de traitement.

N°	NP reconnu	Identif	Type NP	Code GN	GN
1	La Roche	7.10697	*Elision localité*	7	(M+F)S
2	La Roche-sur-Yon	1.13077	*Localité*	7	(M+F)S
3	Thouaré	5.35738	*Alias localité*	7	(M+F)S
4	Nozay	1.11515	*Localité*	7	(M+F)S
5	Ornaisiens	15.21054	*Hab. Sous-Ens. Localité*	3	MP
6	Thouaréens	12.17624	*Habitant Localité*	3	MP
7	Vendéens	11.30	*Département*	3	MP
8	Vendéens	14.199	*Région Hist. Géo.*	3	MP

Figure n°53 : Table résultat de la figure 52 avec les informations de Type et de Genre Nombre

7. Deuxième phase de traitement: détection et calcul des coréférences transitives

Jusqu'à présent, les traitements que nous avons mis en œuvre sont assimilables à de l'étiquetage. Nous identifions des mots correspondant aux occurrences d'entités de la base de données en prenant en compte la morphologie flexionnelle et dérivationnelle.

7.1 Un transducteur d'identifiants sur les associations directes

Ce sont les associations qui contiennent la sémantique du modèle. Il est donc indispensable de poser le problème de leur exploitation. Comme nous l'avons déjà précisé plus haut, il n'est pas question d'interroger directement la base de données. Le problème à résoudre est donc le suivant : comment gérer de l'information relationnelle avec un transducteur ?

7.1.1 Analyse de l'information relationnelle directe

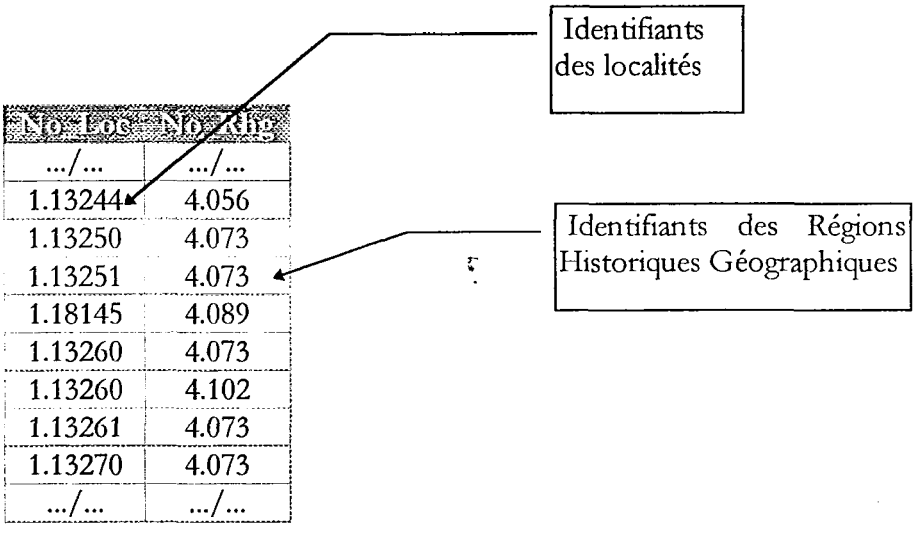
Dans le chapitre 4, nous avons mis en évidence que l'information relationnelle d'une base de données peut être stockée de deux façons :

- Dans des tables spécifiques « Associations » contenant les identifiants des entités associées avec éventuellement, une ou des informations propres à l'association elle-même. Cela correspond dans le modèle Entité-Association aux cardinalités 0,n ou 1,n.
- Dans une table « Entité», sous forme de clé externe, traduisant ainsi la présence d'une cardinalité 0,1 ou 1,1.

7.1.1.1 L'extraction de l'information relationnelle

Dans un cas comme dans l'autre, l'extraction de ces informations sous forme de listes s'effectue par une opération de projection sur les attributs correspondant aux informations relationnelles souhaitées.

Ainsi, si nous reprenons l'exemple de la figure 25 qui présentait l'association entre les **Localités** avec les **Régions Historiques Géographiques**, la table correspondant à cette relation d'association est présentée figure 54.



No_Loc	No_Rhg
.../...	.../...
1.13244	4.056
1.13250	4.073
1.13251	4.073
1.18145	4.089
1.13260	4.073
1.13260	4.102
1.13261	4.073
1.13270	4.073
.../...	.../...

Figure n°54 : Table (extrait) de la relation **LOC/RHG(No_Loc, No_Rhg)**

Pour l'association entre **Localité** et **Département**, on réalise dans la table **Localité**, une projection selon **No_Loc** et **No_Dep** :

No_Loc	Localité	No_Dep	Code Postal	Art	No_Mairie	CNI	Typ_Adm
.../...	.../...	.../...	.../...	.../...	.../...	.../...	.../...
1.15014	Aigrefeuille-sur-Maine	2.44	44140	1	2	7	1
1.11512	Ancenis	2.44	44150	1	3	7	0
1.35601	Anetz	2.44		1	1	7	0
1.35602	Arthon-en-Retz	2.44		1	1	7	1
1.11530	Assérac	2.44	44410	1	1	7	0
.../...	.../...	.../...	.../...	.../...	.../...	.../...	

No_Loc	No_Dep
.../...	.../...
1.15014	2.44
1.11512	2.44
1.35601	2.44
1.35602	2.44
1.11530	2.44
.../...	.../...

Figure n°55 : Table **Localité** (extrait) et sa projection selon **No_Loc** et **No_Dep** (extrait)

Dans la base de données, toutes les relations d'association sont binaires (y compris, par définition, l'association réflexive LOC/AGGLO). Elles expriment un lien sémantique entre deux entités. On obtient donc toujours, comme résultat, une table associant deux identifiants.

7.1.1.2 Une relation générale des associations directes

Lors de l'analyse intuitive de l'article que nous avons présentée au paragraphe 3.2 (annexe 2), nous avons mis en évidence que deux entités peuvent être associées de façon directe (chemin de longueur 1) ou indirecte (chemins de longueur 2, 3, ... n)

La première tâche que nous nous fixons consiste à donner au transducteur la capacité de gérer les **associations directes.** C'est à dire, lorsqu'une entité a été « reconnue », de fournir toutes les informations d'association avec les entités de l'entourage immédiat (chemins de longueur 1).

Ainsi, la reconnaissance du nom propre *La Roche-sur-Yon* engendre l'émission :

- de son identifiant (1.13077),
- de son code de Genre et Nombre (7)

Si l'on considère le modèle conceptuel (annexe 5), cet identifiant est potentiellement « associable» avec les entités :

- *Département* (table DEP),
- *Elision de localité* (table E_LOC),
- *Alias localité* (table A_LOC),
- *Type Administratif de localité* (Table T_A_LOC),

- *Habitants localité* (Table H_LOC),
- *Région Historique et Géographique* (RHG),
- *Sous Ensemble localité* (Table SEL) ,
- *Localité* par association réflexive LOC_AGGLO.

Cela signifie que cet identifiant peut être présent dans l'une des tables « **associations** » suivantes :

LOC/A_LOC entre Localité et Alias Localité
LOC/H_LOC entre Localité et Habitant de localité
LOC_RHG entre Localité et Région Historique Géographique
LOC/SEL entre Localité et Sous-ensemble localité
LOC/AGGLO entre Localité et Localité

ou associé directement à un des identifiants externes de *Département,Type Administratif de localité* , *Type Elision Localité* présents dans la table Localité (cardinalités **1,1**).

Nous pouvons considérer que l'ensemble de ces couples d'identifiants forme autour de l'entité **Localité** une ***relation générale d'associations***, dont le premier attribut contient les occurrences des identifiants **sources** (produits lors de la reconnaissance) et le second les occurrences des identifiants **cibles** (ceux qui conduisent par un chemin de longueur 1 vers les entités associées). Les identifiants étant préfixés, il n'y a pas d'ambiguïté sur les tables dont ils sont issus. Voici un extrait de la table correspondant à la relation que nous venons de définir. Les identifiants **cibles** sont associés à l'identifiant **source** produit lors de la reconnaissance de *La Roche-sur-Yon* (1.13077):

source	cible	
.../...	.../...	
1.13077	2.85	Vendée (Département)
1.13077	7.10697	La Roche (Elision)
1.13077	12.13162	Yonnais (Habitants Localité)
1.13077	4.199	Vendée (Région Hist. Géo.)
1.13077	6.6240	Saint-André-d'Ornay (sous-ensemble localité)
.../...	.../...	

Figure n°56 : Table contenant un extrait des identifiants de la relation générale d'association de Localité, pour l'occurrence *La Roche-sur-Yon* (1.13077):

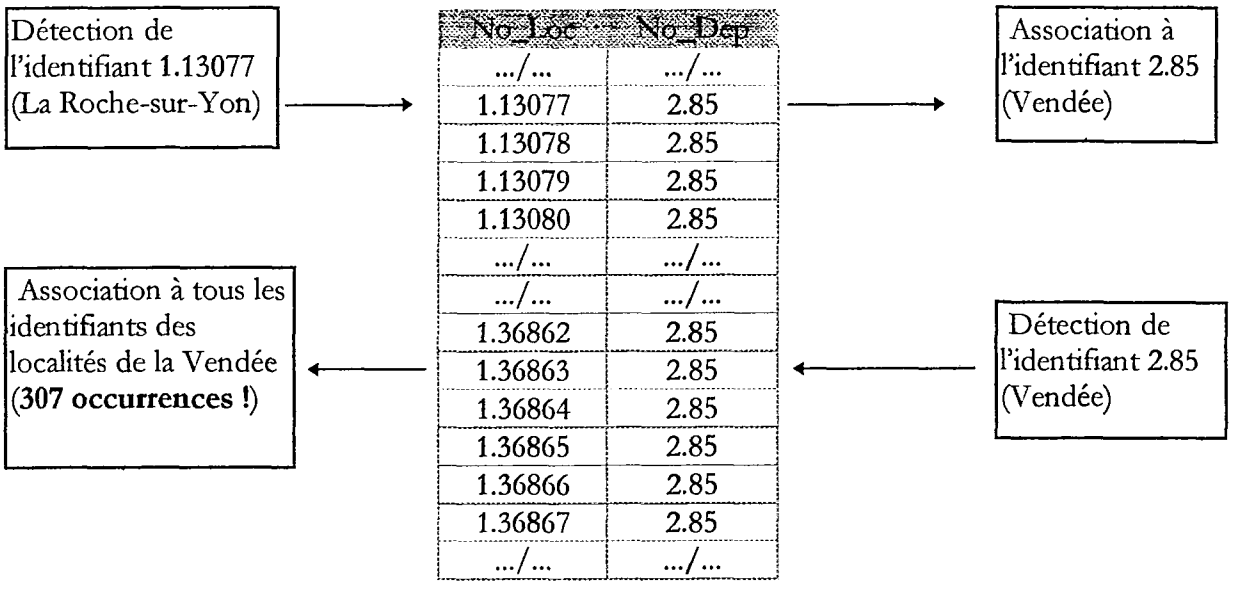
Les relations d'associations de la base de données sont toutes composées de couples d'identifiants (7.1.1.1). Chacun d'entre eux est préfixé du numéro de la table dont il est issu. La **relation générale d'association (RGA)**, telle que nous venons de la définir est donc généralisable à l'ensemble des relations d'associations de la base de données :

RGA(Source, Cible)

Cependant, cette opération de regroupement de toute la sémantique relationnelle dans une seule relation **générique** pose quelques problèmes que nous allons évoquer maintenant en proposant des solutions.

7.1.1.3 Sens de lecture d'une table association

Une table associant les identifiants de deux entités peut se lire dans deux sens. Prenons l'exemple de la table reliant les identifiants *Localité* et *Département.*



No_Loc	No_Dep
.../...	.../...
1.13077	2.85
1.13078	2.85
1.13079	2.85
1.13080	2.85
.../...	.../...
.../...	.../...
1.36862	2.85
1.36863	2.85
1.36864	2.85
1.36865	2.85
1.36866	2.85
1.36867	2.85
.../...	.../...

Figure n°57 : Les deux sens de lecture d'une table « Associations »

Conséquence de la cadinalité 1,1, un seul identifiant de *Département* peut être associé à chaque identifiant de *Localité.* En revanche, une lecture dans l'autre sens provoque la génération de 307 occurrences d'identifiants de *Localité* pour une lecture de l'identifiant de la **Vendée** (2.85). Il existe donc des cas où le nombre excessif d'informations produites peut mettre en difficulté le processus d'analyse.

Cependant, le contrôle du sens de lecture permettrait de maîtriser ce phénomène en interdisant certaines configurations. Par exemple, pour les données de la table présentée figure 57, la détection d'un identifiant de localité (source) entraînerait l'émission de l'identifiant de département (cible) . L'inverse ne sera pas autorisé.

7.1.1.4 Gestion des informations relationnelles directes : ***les choix stratégiques***

Pour que l'information relationnelle soit gérée, il est nécessaire que pour chaque table association, une ou deux listes de couples d'identifiants soient produites. Prenons l'exemple de l'association entre *Localité* et *Région Historique Géographique.* Dans la base de données relationnelle, les informations sont regroupées dans trois relations :

Entités :

LOC (No_Loc , Localité, Code_Postal, #No_dept, #No_TLoc , #Code_GN, #No_Type_Elis)
RHG (No_RHG, Région_Hg, #Code_GN)

Association :

LOC / RHG (No_Loc, No_RHG)

Selon le type de requête, à partir d'un identifiant de **Localité**, on pourra atteindre toutes les informations concernant la ou les **Régions Historiques Géographiques**. De façon symétrique, à partir d'un identifiant de **Région Historique Géographique**, on obtiendra toutes les informations sur les **Localités** qui lui sont associées.

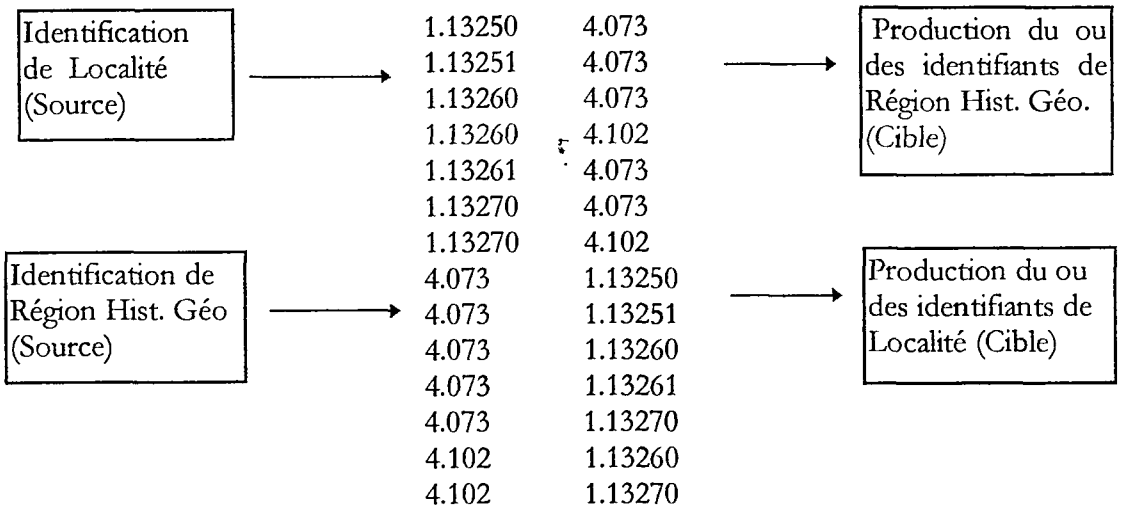
Cette technique, liée à la notion de requête, n'est pas transposable dans un transducteur. Lors du processus de reconnaissance d'un nom de localité, son identifiant est émis par le transducteur. On peut, par un procédé de construction que nous détaillerons dans les paragraphes suivants, faire en sorte qu'à ce moment là, le ou les identifiants de **Régions Historique Géographique** qui lui sont associés puissent également être générés. Mais pour obtenir un comportement analogue sur un identifiant de **Région Historique Géographique**, il faut intégrer au transducteur deux fois la liste correspondant à l'association *LOC / RHG* présentée ci-dessus.

- Une fois dans le sens No_Loc, No_RHG
- Une fois dans le sens No_RHG, No_Loc

No Loc	No Rhg
.../...	.../...
1.13250	4.073
1.13251	4.073
1.13260	4.073
1.13260	4.102
1.13261	4.073
1.13270	4.073
1.13270	4.102
.../...	.../...

Figure n°58 : Un extrait de la table correspondant à la relation *LOC/RHG(No Loc, No RHG)*

La liste produite sera la suivante :



On remarque que dans le sens **Localité → Région Hist. Géo**, pour chaque information *source* on produit une et une seule information *cible*, ce qui n'est pas le cas dans l'autre sens de lecture de la table.

Pour des raisons de performance et de dimension, toutes les informations relationnelles ne peuvent être stockées dans le transducteur. Nous avons donc été conduits à faire des choix qui correspondent aux deux situations suivantes :

L'inclusion. Le transducteur gère les relations d'associations dans le sens :

Entité1 ⊂ Entité2

Ainsi, en reprenant l'exemple de la figure n°58, le transducteur prendra en compte les informations dans le sens No_Loc → No_Rhg.

Relations	Attrib. source	Attrib. cible
LOC (No_Loc,Loc.,Code_Pos.,#No_dept,#No_TLoc ,#Code_GN,#No_T_E)	No_Loc	No_dept
DEP(No_dept , Département, #No_Ra ,#Code_GN)	No_dept	No_Ra
AGGLO / LOC (No_Loc, No_Loc)	No_Loc(Loc)	No_Loc (Aggl)
LOC / RHG (No_Loc, No_RHG)	No_Loc	No_RHG
LOC / H_LOC (No_Loc, No_HLoc)	No_Loc	No_HLoc

Figure n°59 : Relations d'association gérées dans le sens de l'inclusion

L'équivalence. C'est la situation où une forme est substituable à une autre. On range dans cette catégorie :

1. **Les élisions de localités**

Elles correspondent à des formes ambiguës de toponymes élidés. Ainsi, la forme élidée **La Roche** peut renvoyer à 22 localités différentes. Dans ce cas, la relation n'est gérée que dans le sens **Localité → Elision de Localité**. Par exemple, la reconnaissance du toponyme **La Roche-sur-Yon** (1.13077) provoque l'émission de cinq identifiants associés parmi lesquels 7.10697, identifiant de la forme élidée **La Roche**. L'inverse n'étant pas vrai.

2. **Les alias de localités**

On trouve parmi eux, les alias habituels tels que nous les avons présentés lors de l'exposé du Modèle Conceptuel des Données (4.2.1.4.5). Nous y avons également rangé toutes les formes élidées non ambiguës, c'est à dire celles pour lesquelles à un nom de toponyme correspond une et une seule élision et inversement, ainsi que les élisions tombées dans l'usage courant.

7.1.2 Construction du dictionnaire électronique relationnel

Les listes d'identifiants qui sont produites à partir de la base de données ne sont pas exploitables directement par le transducteur. La fusion de ces informations avec celles du transducteur nécessite la production d'un tableau de listes chaînées qui, au moment de la construction du transducteur d'associations permettra l'intégration de l'information relationnelle.

7.1.2.1 Construction de la liste chaînée des identifiants associés

En reprenant l'exemple cité au paragraphe précédant, la liste des identifiants de **Localité** et de **Région Historique Géographique** engendre le tableau de listes chaînées suivant :

1 [1.13250]→{4.073}
2 [1.13251]→{4.073}
3 [1.13260]→{4.073}→{4.102}
4 [1.13261]→{4.073}
5 [1.13270]→{4.073}→{4.102}
6 [1.13272]→{4.073}
7 [1.18145]→{4.089}

A la ligne 1, l'identifiant de *Courcouronnes* (1.13250) est associé à celui de *l'Ile-de-France* (4.073). A la ligne 3, l'identifiant de *Draveil* est associé aux identifiants de *l'Ile-de-France* (4.073) et de la *Brie* (4.102). On remarquera que, compte tenu des choix exposés au paragraphe précédent, aucun lien n'est géré dans le sens **Région Historique Géographique → Localité.**

7.1.2.2 Construction du transducteur d'identifiants et d'associations

Dans sa première version, le transducteur reconnaissant les noms propres de l'article du journal Ouest-France était construit à partir de la liste du paragraphe 6.1.2.1. Le transducteur d'identifiants et d'associations s'obtient en fusionnant la liste chaînée telle que nous l'avons présentée ci-dessus avec les listes d'entités (Identifiant, forme fléchie, code Genre Nombre) telles que nous les avons présentées au chapitre 6. Pour les noms propres de l'article, le tableau des listes chaînées est le suivant :

1 [15.21054]→{6.6240}
2 [12.17624]→{1.35758}
3 [11.30]→{2.85}
4 [14.199]→{4.134}
5 [7.10697]
6 [5.35738]→{1.35738}
7 [1.13077]→{2.85}→{12.13162}→{6.0624}→{4.134}→{7.10697}
8 [1.11515]→{2.44}→{12.12105}
9 [2.85]→{11.30}→{3.52}
10[4.134]→{14.199}

Conséquence des choix présentés dans le paragraphe précédent, on remarque à la ligne 5 que l'identifiant de la forme élidée *La Roche* [7.10697] n'a aucune information d'association. En revanche, ligne 6, l'identifiant de l'alias *Thouaré :* [5.35738] est associé à l'identifiant de la localité *Thouaré-sur-Loire* : {1.35738}.

Lors de la construction du transducteur d'associations, pour le traitement du triplet :

1.11515, Nozay, 7,

une recherche de l'identifiant 1.11515 est effectuée dans le tableau des listes chaînées. Celui-ci est trouvé en ligne 8, l'identifiant du triplet reçoit alors l'adresse de la liste. L'algorithme de construction du transducteur est modifié en conséquence.

A ce stade, la reconnaissance d'un nom propre entraîne l'émission :

- du nom propre lui même
- de son identifiant
- de son code de Genre Nombre
- de tous les identifiants qui lui sont associés par un lien direct

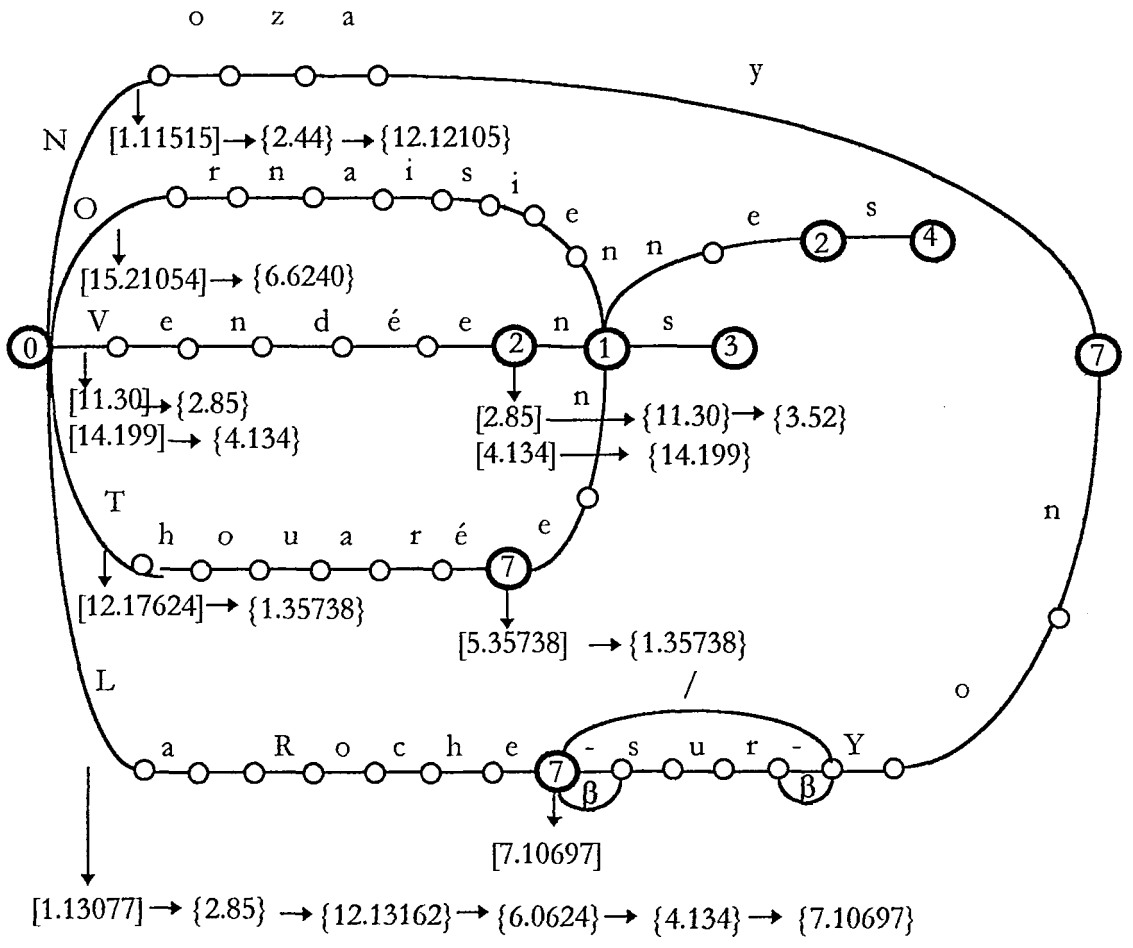


Figure n°60 : Un transducteur gérant de l'information relationnelle directe : ***On trouvera en annexe 8 une représentation graphique de la structure du transducteur d'associations***

7.2 Calcul des références transitives

Au fur et à mesure de l'exploration du texte de l'article, une table contenant l'ensemble des informations produites par le transducteur est construite. Quand l'identifiant d'un nom propre est associé à un ou plusieurs identifiants d'associations, ils sont tous rangés dans une **table résultat.** De façon simultanée, les relations directes qu'ils traduisent sont notées dans une matrice booléenne associée à cette table.

7.2.1 Analyse du texte et construction de la table des résultats et de la matrice associée

Nous allons détailler les différentes phases de construction de la **table résultat** ainsi que de la matrice associée. Les informations produites par le transducteur sont rangées selon l'algorithme suivant :

Un identifiant de nom propre peut être présent dans la **table résultat** de deux façons différentes :

- c'est l'identifiant d'un nom propre reconnu
- c'est un (ou des) identifiant associé

Ainsi, dans le tableau de la figure 61, la ligne 1 contient le nom propre reconnu : *La Roche*, son identifiant : **7.10697**, son code de Genre Nombre : 7.

Lors d'une reconnaissance d'un nom propre, l'algorithme de construction de la **table résultat** recherche tout d'abord si l'identifiant du mot reconnu n'est pas déjà présent. Si c'est le cas, l'information de la ligne correspondante sera complétée (nom propre, code de flexion , ainsi que l'éventuels identifiants associés). Pour illustrer notre propos, nous allons détailler les premières phases de ce processus.

1 - Reconnaissance du nom propre *La Roche* , chargement de son identifiant, du code de Genre et Nombre.

N°	NP reconnu	Identité	G/N
1	La Roche	7.10697	7

Figure n°61 : **Table résultat** après la reconnaissance de *La Roche*

2 - Reconnaissance de *Thouaré* et chargement

- du nom propre, de son identifiant (7.35738), du Code de G N (7) à la ligne 2
- et de l'identifiant qui lui est associé (1.35738) à la ligne 3

N°	NP reconnu	Identité	G/N
1	La Roche	7.10697	7
2	Thouaré	7.35738	7
3		1.35738	

Figure n°62 : **Table résultat** après la reconnaissance de *Thouaré*

3 - Reconnaissance de *La Roche-sur-Yon*.. Chargement du nom propre lui-même, de son code de Genre et Nombre. Dans le transducteur, 5 identifiants lui sont associés : 2.85, 12.13162, 6.0624, 4.134, 7.10697. Ce dernier étant déjà présent dans la table, les 4 autres sont chargés. La matrice associée est mise à jour.

N°	NP reconnu	Identité	G/N
1	La Roche	7.10697	7
2	Thouaré	7.35738	7
3		1.35738	
4	La Roche-sur-Yon	1.13077	7
5		2.85	
6		12.13162	
7		6.6240	
8		4.134	

Identifiant de la forme élidée de La Roche-sur-Yon déjà présent dans la table

Figure n°63 : **Table résultat** après la reconnaissance de *La Roche-sur-Yon*

4 - Reconnaissance de *Nozay*, chargement du nom propre, de son identifiant et de son code de Genre Nombre (7), ligne 9. Chargement des identifiants associés 2.44 et 12.12105 lignes 10 et 11. Construction des liens entre les lignes 9, 10 et 9, 11 dans la matrice associée.

5 - Reconnaissance d'*Ornaisiens*, chargement du nom propre, de son identifiant 15.21054, de son code de Genre Nombre (3), ligne 12. Détection de l'identifiant associé 6.6240 déjà présent dans la table ligne 7. Construction des liens entre les lignes 7 et 12 dans la matrice associée.

6 - Reconnaissance de *Thouaréens*, chargement du nom propre, de son identifiant 12.17624, de son code de Genre Nombre (3), ligne 13. Détection d'un identifiant associé 1.35738 déjà présent dans la table, ligne 3. Construction des liens entre les lignes 3 et 13 dans la matrice associée.

7 - Reconnaissance de *Vendéens*, chargement du nom propre, des deux identifiants 11.30 (habitant de département) et 14.199 (habitant de région historique géographique), de son code de Genre Nombre (3), lignes 14 et 15. Détection des identifiants associés 2.85 et 4.134 déjà présents dans la table lignes 5 et 8. Construction des liens entre les lignes 14, 15, 8 et 5. A la fin de l'analyse, nous disposons des informations suivantes :

N°	NP reconnu	Identif	G/N	
1	La Roche	7.10697	7	1
2	Thouaré	5.35738	7	2
3		1.35738		3
4	La Roche-sur-Yon	1.13077	7	4
5		2.85		5
6		12.13162		6
7		6.6240		7
8		4.134		8
9	Nozay	1.11515	7	9
10		2.44		10
11		12.12105		11
12	Ornaisien	15.21054	3	12
13	Thouaréens	6.17624	3	13
14	Vendéens	11.30	3	14
15	Vendéens	14.199	3	15

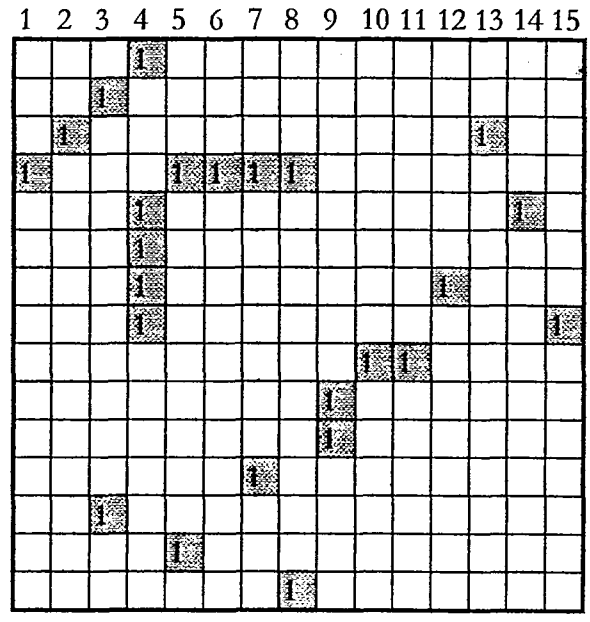


Figure n°64 : **Table résultat** et **Matrice associée** à la fin du parcours du texte de l'article

7.2.2 Fermeture transitive de la matrice booléenne associée

Dans la matrice associée à la **table résultat** de la figure 67 nous avons noté les liens directs qui existent entre les noms propres du texte (chemin de longueur 1). Par transitivité, d'autres liens peuvent être détectés (chemins de longueur 2, 3, ... n). La fermeture transitive de la matrice associée va permettre de les mettre en évidence [R. Faure,79].

« (.../...) pour la Matrice booléenne M, dans laquelle la présence d'un 1 à l'intersection de la ligne x et de la colonne y signifie : « l'arc (x,y) existe », on peut en calculer les puissances successives, en utilisant comme loi multiplicative le produit logique et comme loi additive, la somme logique ; ainsi, M^k sera encore une matrice booléenne. (.../...) la présence d'un 1 à l'intersection de la ligne x et de la colonne y de M^k signifie : il existe au moins un chemin de longueur k entre x et y »

L'algorithme de fermeture transitive se définit de la façon suivante : Soit un opérateur O_x, x étant l'indice de ligne dans la matrice. L'opérateur O_x recopie les 1 de la ligne x dans les lignes où il y a un 1 en colonne x.

Pour x = 1, on applique l'opérateur O_1 : les 1 de la ligne 1 seront recopiés dans les lignes ayant un 1 en colonne 1. Puis, sur la matrice obtenue on applique O_2 et ainsi de suite jusqu'à O_k.

N°	NP reconnu	Identif	G/N
1	La Roche	7.10697	7
2	Thouaré	5.35738	7
3		1.35738	
4	La Roche-sur-Yon	1.13077	7
5		2.85	
6		12.13162	
7		6.6240	
8		4.134	
9	Nozay	1.11515	7
10		2.44	
11		12.12105	
12	Ornaisien	15.21054	3
13	Thouaréens	6.17624	3
14	Vendéens	11.30	3
15	Vendéens	14.199	3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1			1	1	1	1	1				1		1	1
2		1	1										1		
3		1	1										1		
4	1			1	1	1	1	1				1		1	1
5	1			1	1	1	1	1				1		1	1
6	1			1	1	1	1	1				1		1	1
7	1			1	1	1	1	1				1		1	1
8	1			1	1	1	1	1				1		1	1
9									1	1	1				
10									1	1	1				
11									1	1	1				
12	1			1	1	1	1	1				1		1	1
13		1	1										1		
14	1			1	1	1	1	1				1		1	1
15	1			1	1	1	1	1				1		1	1

Figure n°65 : **Table résultat**[22] et **Matrice associée** après la fermeture transitive

La figure n°65 présente la matrice associée à la **table résultat** une fois l'algorithme de fermeture transitive appliqué.

Nous nous intéressons aux liens pouvant exister entre les noms propres reconnus dans le texte de l'article. Les lignes qui ne comportent que des identifiants ont été utilisées pendant la fermeture transitive pour calculer tous les chemins. Une fois l'opération réalisée, ces informations ne sont plus utiles. On peut donc simplifier la **table résultat** en éliminant ces lignes et en réalisant l'opération symétrique sur les lignes et les colonnes de la matrice associée.

On obtient le tableau et la matrice associée réduits tels qu'ils sont présentés figure 66.

22 Nous n'avons pas fait figurer les zéros

N°	NP reconnu	Identif	G/N
1	La Roche	7.10697	7
2	Thouaré	5.35738	7
4	La Roche-sur-Yon	1.13077	7
9	Nozay	1.11515	7
12	Ornaisien	15.21054	3
13	Thouaréens	6.17624	3
14	Vendéens	11.30	3
15	Vendéens	14.199	3

	1	2	4	9	12	13	14	15
1			1		1		1	1
2						1		
4	1				1		1	1
9								
12	1		1				1	1
13		1						
14	1		1		1			1
15	1		1		1		1	

Figure n°66 : **Table résultat**[23] et **Matrice associée** contenant la sémantique relationnelle des noms propres

7.2.3 Informations transmises à l'analyseur de TAL

Les liens de la sémantique relationnelle ayant été captés, la **table résultat** et la matrice associée peuvent être renumérotées (fig. 67), préparant ainsi l'arrivée d'autres informations relationnelles.

N°	NP reconnu	Identif	G/N
1	La Roche	7.10697	7
2	Thouaré	5.35738	7
3	La Roche-sur-Yon	1.13077	7
4	Nozay	1.11515	7
5	Ornaisien	15.21054	3
6	Thouaréens	6.17624	3
7	Vendéens	11.30	3
8	Vendéens	14.199	3

	1	2	3	4	5	6	7	8
1			1		1		1	1
2						1		
3	1				1		1	1
4								
5	1		1				1	1
6		1						
7	1		1		1			1
8	1		1		1		1	

Figure n°67 : Table résultat et matrice renumérotées

Quelles informations transmettre à l'analyseur principal ? La **table résultat** et la matrice associée de la figure 68 expriment :

Ligne n°1 : *La Roche* (**Toponyme** : élision de localité) est en relation avec (successeurs) :

- colonne 3 : *La Roche-sur-Yon* (**Toponyme** : localité)
- colonne 5 : *Ornaisiens* (**Gentilé** : habitants d'un sous-ensemble de localité)
- colonne 7 : *Vendéens* (**Gentilé** : habitants de département)
- colonne 8 : *Vendéens* (**Gentilé** : habitants de région historique géographique)

Ligne n°2 : *Thouaré* (**Toponyme** : alias de localité) est en relation avec (successeurs) :

---

[23] Nous avons soustrait la matrice unité I (1 dans la diagonale) les relations correspondant aux circuits simples n'ayant pas d'intérêt dans notre cadre de travail.

- colonne 6 : *Thouaréens* (**Gentilé** : habitants de localité)

Ligne n°3 : *La Roche -sur-Yon* (**Toponyme** : localité) est en relation avec (successeurs) :

- colonne 1 : *La Roche* (**Toponyme** : élision de localité)
- colonne 5 : *Ornaisiens* (**Gentilé** : habitants d'un sous-ensemble de localité)
- colonne 7 : *Vendéens* (**Gentilé** : habitants de département)
- colonne 8 :*Vendéens* (**Gentilé** : habitants de région historique géographique)

Ligne n°4 : *Nozay* (localité) pas de relation

Les lignes suivantes redonnant les relations déjà exprimées ci-dessus. Quand une relation existe entre des entités du même type (**Toponyme** ou **Gentilé**) les occurrences sont **substituables** entre elles :

Toponyme : *La Roche, La Roche-sur-Yon*
Gentilé : *Vendéens, Ornaisiens*

Les *Ornaisiens* désignent les habitants de *La Roche-sur-Yon* ou de *La Roche.* Ce sont aussi des *Vendéens.* On remarquera que les liens ont été mis en évidence, alors que les toponymes *Vendée* et *Saint-André-d'Ornay* ne sont pas présents dans le texte.

Toponyme : *Thouaré*
Gentilés : *Thouaréens*

Les *Thouaréens* désignent les habitants de *Thouaré.* Le toponyme *Thouaré-sur-Loire* est également absent du texte.

Toponyme : *Nozay* (Pas de référent dans le texte.)

7.3 Les volumes d'informations traitées

Nous ne faisons figurer que les formes canoniques qui sont présentes dans la base de données. Dans le transducteur, toutes les formes fléchies des gentilés sont stockées à partir des listes qui sont générées.

TOPONYMES

Départements	100
Localités	38198
Régions Administratives	26
Régions Historiques et Géographiques	299
Elisions Localités	9200
Alias Localités	3421
Sous-Ensemble Localités	138
Sites Historiques Géographiques	28
Total	**51410**

Figure n°68 : Tableau récapitulatif des toponymes présents dans la base de données

GENTILES

Habitants Département	37
Habitants Localité	3158
Habitants Région Administrative	24
Habitants Régions Historiques Géographiques	201
Alias Habitants Localité	22
Habitants Sous Ensemble Localité	29
Total	**3471**

Figure n°69 : Tableau récapitulatif des gentilés présents dans la base de données

Après la construction du transducteur selon l'algorithme cité en annexe 7 celui-ci compte 137 400 états et 174 034 transitions. Des traitements post- construction permettent d'augmenter le nombre de formes reconnues en prenant en charge les graphies multiples (2.4.2.4)

Le tableau suivant donne l'évolution du volume interne du transducteur. On remarque qu'à l'origine celui-ci reconnaît 55 508 noms propres. En fin de traitement, le nombre de formes reconnues est de 135 539 soit une augmentation de 144 %. Pendant le même temps, le nombre de transitions est passé de 174 034 à 187 760 soit une augmentation de 7,88 %. Le nombre d'identifiants d'entités et d'associations ayant augmenté dans des proportions encore inférieures.

Nombre	Etats	Trans	Ident-Ent	Ident-Asso	Nb reconnus
Transducteur d'origine	137 400	174 034	47 553	62 986	55 508
Après les Post-Traitements	137 400	187 760	49 898	67 835	135 539

Figure n°70 : Evolution du transducteur après les post-traitements des graphies multiples

Nous avons appliqué l'algorithme de fermeture transitive des identifiants de noms propres sur un corpus de 9 088 780 de caractères composé de 12 448 articles du journal Ouest-France. Les résultats bruts sont les suivants:

Occurrences de Noms propres reconnus 33935
Relations directes et transitives détectées 18224 (soit 53,70 %)

Cependant, ceux-ci doivent être analysés avec précaution. En effet dans 65,14 % des cas, l'article ne comportait qu'un nom propre, aucune association n'étant réalisable. Si on élimine ces cas, les résultats sont les suivants:

Nombre d'articles traités : 6713
Noms propres reconnus : 28200
Relations directes et transitives détectées : 18224 (soit 64,62 %)

Donc, dans un corpus de nature journalistique près des 2/3 des noms propres reconnus ont des relations sémantiques entre eux qui sont ainsi détectées et mises en évidence par notre algorithme .

Au cours du processus d'analyse, l'algorithme génère deux fichiers résultats: Le premier comptabilise par article les valeurs suivantes :

nombre de caractères traités ;nombre de noms propres reconnus ;nombre de relations

Il a la forme suivante:

extrait du fichier:
../...
689;1;0
768;8;1
1649;9;1
597;4;0
53;1;0
581;2;1
1879;1;0
5800;2;1
3044;3;0
2842;4;1
../...

Le second enregistre le détail des résultats en donnant pour chaque nom propre :

l'identifiant, le nom propre, la catégorie (Toponyme ou Gentilés), le genre nombre et le type de nom propre. Quand une relation a été calculée entre deux noms propres, une flèche signale le lien en reprenant les mêmes informations pour le coréférent.

extrait du fichier:

.../...

==========

No 70 Nb car: 1928 NP : 6 Relations : 0

1.10800; Besançon; Toponyme; (M+F)S; Localité
7.10260; Bretagne; Toponyme; (M+F)S; Elision Localité
1.24642; Bretagne; Toponyme; (M+F)S; Localité
1.35469; Bretagne; Toponyme; (M+F)S; Localité
3.53; Bretagne; Toponyme; (M+F)S; Région Administrative
4.62; Bretagne; Toponyme; (M+F)S; Région Historique et Géographique

==========

No 71 Nb car: 1074 NP : 4 Relations : 3

3.43; Franche-Comté; Toponyme; FS; Région Administrative → 2.25; Doubs; Toponyme; MS; Département
3.43; Franche-Comté; Toponyme; FS; Région Administrative → 1.33969; Doubs; Toponyme; MS; Localité
4.56; Franche-Comté; Toponyme; FS; Région Historique et Géographique
2.25; Doubs; Toponyme; MS; Département → 1.33969; Doubs; Toponyme; MS; Localité
1.33969; Doubs; Toponyme; MS; Localité

==========

No 72 Nb car: 1041 NP : 7 Relations : 2

1.16630; Benet; Toponyme; (M+F)S; Localité
7.11197; Niort; Toponyme; (M+F)S; Elision Localité
1.12893; Niort; Toponyme; (M+F)S; Localité
10.17; Aquitaine; Gentilé; FS; Habitant Région Administrative → 3.72; Aquitaine; Toponyme; FS; Région Administrative
3.72; Aquitaine; Toponyme; FS; Région Administrative
14.52; Aquitaine; Gentilé; FS; Habitant Région Historique et Géographique → 4.88; Aquitaine; Toponyme; FS; Région Historique Géographique
4.88; Aquitaine; Toponyme; FS; Région Historique et Géographique

.../...

8. Conclusion et perspectives

Notre étude s'est limitée aux principaux toponymes français et à leurs gentilés. Cependant, nous avons déjà mené une réflexion sur ce que pourrait être un dictionnaire électronique général des noms propres et son système relationnel. Une telle entreprise est-elle réalisable ?

Chaque individu est porteur d'un univers de références à l'intérieur duquel les noms propres forment un réseau dont l'étendu est fonction, des relations qu'il entretient avec les autres sur le plan familial, affectif, social, professionnel, citoyen ; de la formation qu'il a reçue, de ses centres d'intérêts culturels, techniques, etc...

Dès lors, poser le problème d'un dictionnaire électronique général des noms propres, c'est émettre l'hypothèse qu'il existe un sous ensemble de références commun à tous les individus d'un même groupe (région, pays, ethnie, civilisation..). D'ailleurs, n'est-ce pas ce système de références partagées qui fondent pour chacun son sentiment d'appartenance à la région, au pays, à l'ethnie, à la civilisation ? Cependant, avoir l'intuition que cet ensemble existe est une chose, être capable de fixer sur des critères objectifs son contenu en est une autre !

Nos travaux s'inscrivent dans le cadre de PROLEX. C'est un projet de recherche universitaire dans le domaine du traitement automatique du langage naturel qui est dirigé par Denis Maurel (MC-HDR) avec des chercheurs des universités de Tours, Nantes, Paris7, Paris1, Lille3. Il s'agit de prendre en considération l'importance de la présence des noms propres dans les corpus de texte et d'envisager des traitements spécifiques de cette classe de mots, au niveau lexicale et sémantique. Nous avons mis en évidence, que les noms propres entretiennent entre eux des relations qui sont porteuses de sémantique. Particulièrement évident pour les noms de lieux et les noms d'habitants, ce système relationnel existe pour l'ensemble des noms propres d'un texte. Un nom de lieu revoie à des noms de personnes, à des entités hydrographiques, à des événements célèbres etc... Le principe du transducteur d'associations peut donc être appliqué à l'ensemble de cette classe de mots.

Ainsi, un étudiant de Paris 7 vient de terminer un travail de recherche sur l'hydrographie à partir de données qui avaient été collectées par l'équipe PROLEX. Les informations ont été organisées de façon relationnelle : pour le territoire français, 672 entités hydrographiques (fleuves, rivières, canaux) ont été répertoriées et associées à 156 noms de localités. Cet ensemble va être intégré à notre base de données et permettra de générer un transducteur d'associations prenant en compte toute la sémantique relationnelle entre les noms de lieux, les gentilés et les noms de fleuves, rivières et canaux.

Toujours dans le cadre de PROLEX, un travail important de recherche est mené depuis deux ans sur les exonymes et la géographie internationale. Le premier but de ce travail est de constituer une base de données relationnelle des noms propres de la géographie politique internationale afin de construire un dictionnaire électronique relationnel sur le même modèle que celui que nous avons présenté, mais en prenant en compte les spécificités liées à l'exonymie[24].

Parallèlement à ce travail de collecte et de construction de dictionnaires électroniques, nous avons été amené à nous interroger sur l'existence de règles de constructions des gentilés à partir des toponymes. Comme nous l'avons indiqué au paragraphe 2.4.2.3, la construction d'un gentilés est en général un processus dérivationnel classique ou le nom des habitants se construit régulièrement à partir du nom de lieu. Mais c'est loin d'être le cas pour tous les noms d'habitants. Or, si nous possédons aujourd'hui 51 410 noms de lieux, nous n'avons collectés que 3471 gentilés. Aussi, cherchons nous à combler cette lacune par des procédures automatiques de comparaisons entre toponymes et gentilés, afin de pouvoir les relier, si les deux se trouvaient un jour associés dans un texte. En collaboration avec Elmar Eggert, chercheur de l'université de

---

[24] O. Piton , D. Maurel 1996 - Les exonymes et la géographie internationale - Etude préalable à la constitution d'une base de données dans le cadre du projet PROLEX.

Münster en Allemagne, nous travaillons sur la mise en évidence de règles de concaténation, troncation, supplétion et allomorphie dans la formation des gentilés[25]. Il s'agit à terme, de proposer une alternative à un échec de la reconnaissance d'un gentilé par notre dictionnaire électronique, en appliquant des règles de constructions à partir du toponyme.

Enfin, nous pensons que le principe du dictionnaire électronique relationnel peut s'appliquer à d'autres domaines de lexique. Ainsi, on peut envisager de construire un système relationnel entre les synonymes et les antonymes du lexique afin de fournir des équivalents ou des contraires lors d'un échec du processus d'analyse.

Dans un autre domaine, une adaptation des fichiers résultats de notre transducteur permet de réaliser un comptage des noms de lieux et des gentilés rencontrés dans un corpus. Ainsi, par rapport à la zone de référence géographique du journal Ouest-France (grand ouest), nous serons en mesure d'indiquer à notre partenaire les localités dont les noms sont cités le plus souvent, et a contrario, celles dont le journal ne parle jamais.

9. Bibliographie

Adouani A. (1993) : Traitement dérivationnel des supplétismes lexicaux, Cahiers de lexicologie, vol. 63, 87-98, Paris, Didier.

Altman E.B. (1967) On the Recognition of Personal Names in Natural Text. IBM Research

Batgelj V., Pisanski T., Kerzic D. (1992) : Automatic Clustering of Languages, Computational Linguistics, 18:3, 339-352.

Belleil C., Maurel D. (1995) : Un dictionnaire relationnel des noms propres liés à la géographie, consultés par transducteurs, 4èmes Journées scientifiques de l'AUPELF-UREF: Lexicomatique et dictionnairiques, Lyon, 28-30 septembre.

Belleil C., Maurel D. (1996) : Traitement informatique des ambiguïtés dans la reconnaissance des noms propres liés à la géographie, Bulag

Carré Degrémond Gross Pierrel Sabah (1991) Langage humain et machine - Presse du CNRS

Chomsky (1969) Structures syntaxiques - Editions du Seuil - Paris

Chomsky (1987) La nouvelle syntaxe - Editions du Seuil - Paris

Chow L. (1992) :The Morphological Structure of Proper Names California State University, Fullerton

Clémenceau D. (1993) : Structuration du lexique et reconnaisance de mots dérivés, Paris, Thèse de doctorat (Universit, Paris VII).

Coates-Stephen S.(1993) : The Analysis and Acquisition of Proper Names for the Understanding of Free Text.

Coates-Stephens S. : Automatic Lexical Acquisition Using Within-Text Descriptions of Proper Nouns

Corbin D. & Corbin P. (1991) : Vers le dictionnaire dérivationnel du français, Lexique, n°10, 147-161, Presses Universitaires de Lilles.

Courtois B (1990) Un système de dictionnaires électroniques pour les mots simples du français - Communication CNRS UA 819 - Laboratoire d'Automatique Documentaire et Linguistique.

Courtois B (1990) Un système de dictionnaires électroniques pour les mots simples du français. Dictionnaires électroniques du français Langue Française n° 87

Courtois B, Silberztein M (1990) Dictionnaires électroniques du français. Dictionnaires électroniques du français Langue Française n° 87

Danlos (1985) Génération automatique de textes en langues naturelles Masson - Paris.

Davalo N (1990) Des réseaux de neurones - Eyrolles - Paris

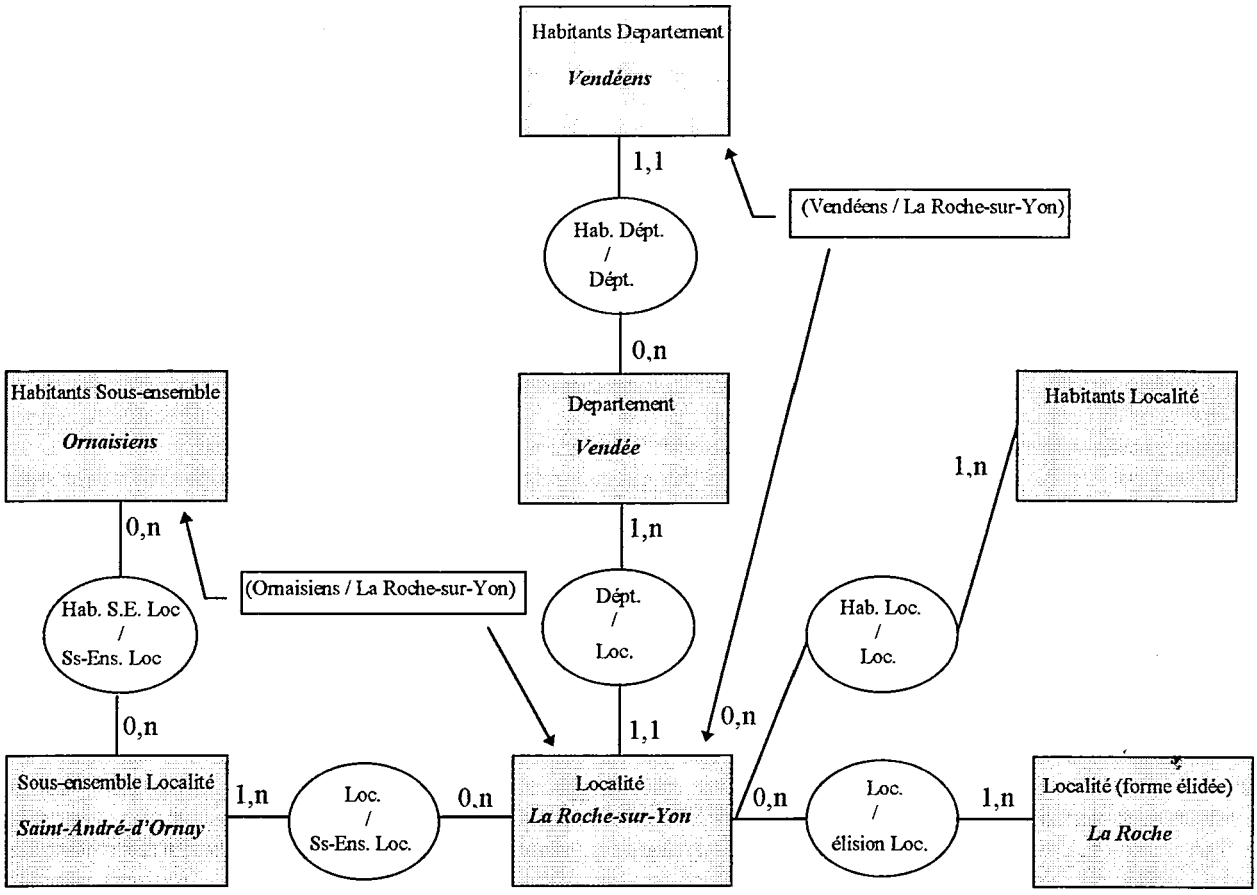
[25] E. Eggert, D. Maurel, C. Belleil 1997 Allomorphies et supplétions dans la formation des gentilés - Application au traitement informatique.

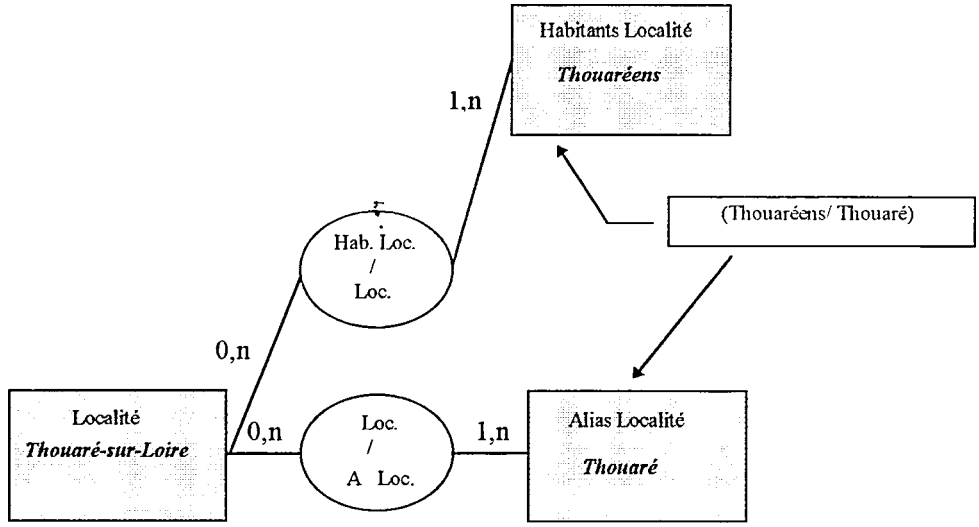
Dressler W. U. (1985) : Sur le statut de la suppléance dans la morphologie naturelle, Langages, n°78, 41-56.
Dubois J, Dubois-Charlier F (1990) Incomparabilité des dictionnaires. Dictionnaires électroniques du français Langue Française n° 87
Dubois Jouannon Lagane (1961) Grammaire Française - Larousse - Paris
Dugas A (1990) La création lexicale et les dictionnaires électroniques. Dictionnaires électroniques du français Langue Française n° 87
Eggert E. (1994) : Etude dérivationelle des dérivés de toponymes, Mémoire de maîtrise, Université Lille III.
Enguehard C. (1992) : ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique, Thèse de doctorat en Contrôle des Systèmes, Université de Technologie de Compiègne.
Faure R. (1970) : Précis de recherche opérationnelle Dunod Décision
Flaux N (1991) L'antonomase du nom propre ou la mémoire du référent, Syntaxe et sémantique des noms propres, Langue française n°92
Garrigues M. (1993) : Prépositions et noms de pays et d'îles : une grammaire locale pour l'analyse automatique des textes, Linguisticae Investigationes, volume 17, n°2, 281-305.
Gary-Prieur M N (1991) La modalisation du nom propre Syntaxe et sémantique des noms propres, Langue française n°92
Gary-Prieur M N (1991) Le nom propre constitue-t-il une catégorie linguistique ? Syntaxe et sémantique des noms propres, Langue française n°92
Grevisse M. (1982) : Le français correct - Guide Pratique, Paris Duculot.
Gross G (1990) Définition des noms composés dans un lexique-grammaire. Dictionnaires électroniques du français Langue Française n° 87
Gross M (1989) Les industries de la langue et l'étude du français. Langue française et nouvelles technologies. Langue française n°83
Gross M (1990) Le programme d'extension des lexiques électroniques. Dictionnaires électroniques du français Langue Française n° 87
Hopcroft Ullman (1979) Introduction to automata theory, languages and computation - Addison Wesley - Massachusetts
Johnson-Laird, Ehrlich, Tardieu, Cavazza (1993) Les modèles mentaux - Masson - Paris
Jonasson K (1991) Les noms propres métaphoriques : construction et interprétation. Syntaxe et sémantique des noms propres, Langue française n°92
Kleiber G (1991) Du nom propre non modifié au nom propre modifié : le cas de la détermination des noms propres par l'adjectif démonstratif. Syntaxe et sémantique des noms propres, Langue française n°92
Lakoff Johnson (1980) Les métaphores dans la vie quotidienne - Les Editions de Minuit - Paris
Laporte E (1988) Méthodes algorithmiques et lexicales de phonétisation de textes- Thèse de doctorat en informatique - Université de Paris VII
Laporte E (1990) Le dictionnaire phonémique DELAP. Dictionnaires électroniques du français Langue Française n° 87
Lomholt Jorgen (1983) : Syntaxe des noms géographiques en français contemporain, Etudes Romanes de l'Université de Copenhague. N° 25 - 1983
Mathieu-Colas (1990) Orthographe et informatique : établissement d'un dictionnaire électronique des variantes graphiques. Dictionnaires électroniques du français Langue Française n° 87
Maurel D. (1995) : "Le traitement informatique de la dérivation des noms de ville", TA information, volume 35, n°2, 111-127.
Maurel D., Belleil C. (1995) : " Un dictionnaire électronique pour les noms propres liés à la géographie ", colloque international Lexique, syntaxe et analyse automatique des textes, Université Paris X Nanterre, 3-5 mai.

Maurel D., Belleil C. (1995) : Un dictionnaire électronique pour les noms propres liés à la géographie, colloque international Lexique, syntaxe et analyse automatique des textes, Université Paris X Nanterre, 3-5 mai.
Maurel D., Belleil C. (1996) : Un dictionnaire électronique pour les noms propres liés à la géographie, Lynx.
Maurel D., Leduc B., Courtois B. (1995) : Vers la constitution d'un dictionnaire électronique des noms propres, Linguisticae Investigationes, volume 19
Mohri M (1994) Application of Local Grammars Automata : an Efficient Algorithm IGM94-16 Université de Marne-la-Vallée
Mohri M (1994) :On Some Applications of Finite-State Automata Theory to Natural Language Processing- Université de Marne-la-Vallée
Mohri M. (1992) :Syntactic Analysis by Local Grammars Automata : an Efficient Algorithm. LADL-IGM
Mohri M. (1993) :Compact Representations by Finite-state Transducers Institut Gaspard Monge-LADL
Mohri M. (1994) : Application of Local Grammars Automata an efficient Algorithm, Rapport de recherche IGM 94-16, Université de Marne-la-Vallée.
Molino J. (1982) : Le nom propre dans la langue, Langage, n° 66 Paris Larousse
Noailly M (1991) " Lénigmatique Tombouctou " : nom propre et position de l'épithète. Syntaxe et sémantique des noms propres, Langue française n°92
Nogier (1989) A natural laguage production system based on conceptual graphs - Centre scientifique IBM - Paris
Paik W, Liddy E, Yu E, McKenna M (1991) : Categorizing and Standardizing Proper Nouns for Efficient Information Retrivial.
Pfeifer U, Poersch T, Fuhr N : Searching Proper Names in Databases
Piton O. Maurel D. (1996) : Les exonymes et la géographie internationale. Etude préalable à la constitution d'une base de données dans le cadre du projet PROLEX. LI/E3i/Université François Rabelais ToursRapport interne n°180
Piton O., Taieb C., Maurel D. (1996) : L'importance du traitement informatique des noms propres ; exemple des noms de pays et de leurs dérivés, communication au 15ème Colloque Européen sur la Grammaire et le Lexique Comparés des Langues Romanes, Munich, Allemagne, 19-21 septembre.
Pitrat (1983) Réalisation d'un analyseur-générateur lexicographique général - Rapport de recherche - CNRS
Ren X., Perrault F. (1992) : The typology of Unknown Words : An Experimental Study of Two Corpora, Proceedings of COLING 92, Nantes.
Revuz D. (1991) : Dictionnaires et lexiques - Méthodes et algorithmes, Thèse de Doctorat en Informatique (Université Paris VII).
Rey A. (1977) : Le lexique : image et modèles. Du dictionnaire à le lexicologie, Paris, Armand Colin
Rey A. (1993) Dictionnaire universel des noms propres. Robert
Sabah (1988/ 1990) L'intelligence artificielle et le langage - Tomes 1 et 2 - Hermès - Paris
Silberztein M (1990) Le dictionnaire électronique des mots composés. Dictionnaires électroniques du français Langue Française n° 87
Silberztein M (1993) Dictionnaires électroniques et analyse automatique de textes - Masson - Paris
Sowa (1984) Conceptual Structures: Information Processing in Mind and Machine - IBM Systems Research Institute - Addison Wesley Publishing Company.
Wilmet M (1991) Nom propre et ambiguïté. Syntaxe et sémantique des noms propres, Langue française n°92
Winograd (1972) :Understanding natural language - Academic Press - Cambridge Massachusetts, USA.

Localité	H	E	Localité	H	E	Localité	H	E	Localité	H	E
Abancourt	2		Alleux	2		Araux	2		Aulnay	5	7
Abattoir	4		Alligny		2	Arbonne			Aulnays	3	
Abbas	2		Allonne	2		Arbre	2	3	Aulon	3	
Abbaye	8	4	Allonnes	3		Arbres	2		Aumône	3	
Abergement	2	9	Allons	2	2	Arceau	2		Aumont	4	2
Aboncourt	2	3	Allonville	2		Arces	2		Aunay		6
Achères	2		Ally	2		Arches	2		Aurel	2	
Acheux		2	Alos	2		Arcis	3	2	Auriac	4	6
Acqueville	2		Alouette	2		Arcs	2		Autels		2
Acy		2	Alpe		4	Ardon	2		Auterive	3	
Adroit	3		Alquines			Arfeuille	4		Auteuil	2	
Age	5	6	Alzon	3		Arfeuilles	3		Autheuil	2	2
Ages	5		Amagne			Argelès		2	Authie	2	
Aglan	2		Amance	3		Argelos	2		Authieux		5
Agnac	3		Ambleville	2		Argence	3		Authon	2	4
Aiglun	2		Amfreville	2	7	Argentière	4		Autigny		4
Aigrefeuille	2	3	Amigny	2		Argentières	4		Autreville	3	4
Aigremont	5		Amilly	2		Argenton	3	4	Autrey	2	4
Aigueperse	2		Ancy		4	Arguel	2		Autry		2
Aigues		5	Andelot		3	Armancourt	2		Auvers	3	4
Aiguille	4		Andillé	2		Armentière	4	3	Auvilliers	3	
Aiguillon	3	4	Andillon	2		Arnac	4	3	Auxon	3	2
Ailly	2	4	Andilly	4	3	Arnois	2	2	Auxy	3	
Ainay		2	Andreaux	2		Arnouville		2	Auzay	2	
Air	3		Andrieux	2		Aron	2		Availles		5
Aires	4		Angerville	2	4	Arques	4		Aventure	3	
Airon		2	Anglade	3		Arras		2	Avernes		2
Aisy		4	Anglard	2		Ars	5	3	Avesnes		9
Aix	4	12	Anglards		2	Arthez		3	Avoine	2	
Alainville	2	2	Angles	6		Artigues	8	2	Avrainville	4	
Albagnac	2		Annay		2	Asnières	6	11	Avricourt	3	
Albas	2		Anneville		5	Assas	2		Avrigny	2	
Albigny	3		Ansannes	2		Attigny	2		Avrillé	2	
Albon	2		Antheuil			Auberville		3	Avron	2	
Alboret		2	Antignac	2		Aubiers	3		Ayen	2	
Alex	2		Antilly	3		Aubigny	6	10	Azat	3	2
Aleyrac	2		Antony	2		Aubusson	2		Azay		8
Allas			Antras	3		Augé	3		Azé	4	
Allemant	2		Apcher	3		Augères	3		Azy	2	2
Allends	2		Apremont	6	2	Augy	2				
Total	92	66		77	40		104	55		83	82

Total homonymes simples (H)	356	
Total homonymes par élision (E)	243	
Total général	599	(21% des 2862 noms de localités dans la lettre A)


Habitants Departement
Vendéens
1,1
(Vendéens / La Roche-sur-Yon)
Hab. Dépt.
/
Dépt.
0,n
Habitants Sous-ensemble
Ornaisiens
Departement
Vendée
Habitants Localité
1,n
0,n
1,n
Hab. S.E. Loc
/
Ss-Ens. Loc
(Ornaisiens / La Roche-sur-Yon)
Dépt.
/
Loc.
Hab. Loc.
/
Loc.
0,n
1,1
0,n
Sous-ensemble Localité
Saint-André-d'Ornay
1,n
Loc.
/
Ss-Ens. Loc.
0,n
Localité
La Roche-sur-Yon
0,n
Loc.
/
élision Loc.
1,n
Localité (forme élidée)
La Roche


Habitants Localité
Thouaréens
1,n
(Thouaréens/ Thouaré)
Hab. Loc.
/
Loc.
0,n
Localité
Thouaré-sur-Loire
0,n
Loc.
/
A Loc.
1,n
Alias Localité
Thouaré

Enquête Nationale sur les Noms Propres liés à la Géographie

1 - Noms de lieux et noms d'habitants

la: 1269

Nom de la localité (avec les accents)		Code Postal
GRUISSAN	Nom Patois = Gruissan-Plage	11430

Pour une localité, il existe parfois plusieurs noms d'habitants, par exemple les habitants de *Quimperlé* sont les *Quimperlois* ou les *Quimperléens*. Si c'est votre cas, remplir les colonnes 1er nom, 2ème nom...

HY:0093 étang de Gruissan

Noms des Habitants

	Premier Nom	Deuxième Nom	Troisième Nom
masculin singulier	GRUISSANAIS	/	/
masculin pluriel	GRUISSANAIS	/	/
féminin singulier	GRUISSANAISE	/	/
féminin pluriel	GRUISSANAISES	/	/
remarques sur la prononciation	/	/	/

En plus de la région administrative, votre localité se sent peut-être appartenir à une ou plusieurs régions *traditionnelles* ou *historiques*. Par exemple, la *Flandre (Flamands)* et l'*Artois (Artésiens)* font partie de la région administrative *Nord-Pas de Calais*.

0029 0149

Noms des Habitants

	Première Région	Deuxième Région	Troisième Région
Nom de la Région Historique	LANGUEDOC	NARBONNAISE	SEPTIMANIE
masculin singulier	LANGUEDOCIEN	NARBONNAIS	SEPTIMANIEN
masculin pluriel	LANGUEDOCIENS	NARBONNAIS	SEPTIMANIENS
féminin singulier	LANGUEDOCIENNE	NARBONNAISE	SEPTIMANIENNE
féminin pluriel	LANGUEDOCIENNES	NARBONNAISES	SEPTIMANIENNES
remarques sur la prononciation	/	/	/

GOTHIE 0149

2 - Autres Noms Propres associés à votre localité

Une localité est parfois connue ou reconnue par un ou des *personnages célèbres*, un ou des *évènements historiques*, un ou des *produits typiques*, ou encore un *quartier*, un *monument* ou autre...

Satisfait, le 9 février 19

- La Tour Barberousse
- Quartier de la Vendée

MAIRIE DE GRUISSAN (AUDE)

D'autre part, connaissez-vous l'origine du nom des habitants de votre localité :

GRUISSAN : de GRUSSIUS, gentilice gallo-romain qui aurait fondé en ce lieu un établissement important : "Villa Grussiana".

Enfin, un de nos partenaires souhaite connaître la date de création du *premier bureau de poste ou point postal* dans votre localité :

?

Enquête Nationale sur les Noms Propres liés à la Géographie

1 - Noms de lieux et noms d'habitants

Nom de la localité (avec les accents)	Code Postal
Merville-Franceville-Plage	14810

Pour une localité, il existe parfois plusieurs noms d'habitants, par exemple les habitants de *Quimperlé* sont les *Quimperlois* ou les *Quimperléens*. Si c'est votre cas, remplir les colonnes 1er nom, 2ème nom...

Noms des Habitants

	Premier Nom	Deuxième Nom	Troisième Nom
masculin singulier	Mervillais Francevillais		
masculin pluriel	" "		
féminin singulier	Mervillaise - Francevillaise		
féminin pluriel	Mervillaises Francevillaises		
remarques sur la prononciation			

En plus de la région administrative, votre localité se sent peut-être appartenir à une ou plusieurs régions *traditionnelles* ou *historiques*. Par exemple, la *Flandre (Flamands)* et l'*Artois (Artésiens)* font partie de la région administrative *Nord-Pas de Calais*.

178

Noms des Habitants

	Première Région	Deuxième Région	Troisième Région
Nom de la Région Historique	Côte Fleurie "pays d'Auge"		
masculin singulier	—		
masculin pluriel	—		
féminin singulier	Côte fleurie "pays d'Auge"		
féminin pluriel	—		
remarques sur la prononciation			

2 - Autres Noms Propres associés à votre localité

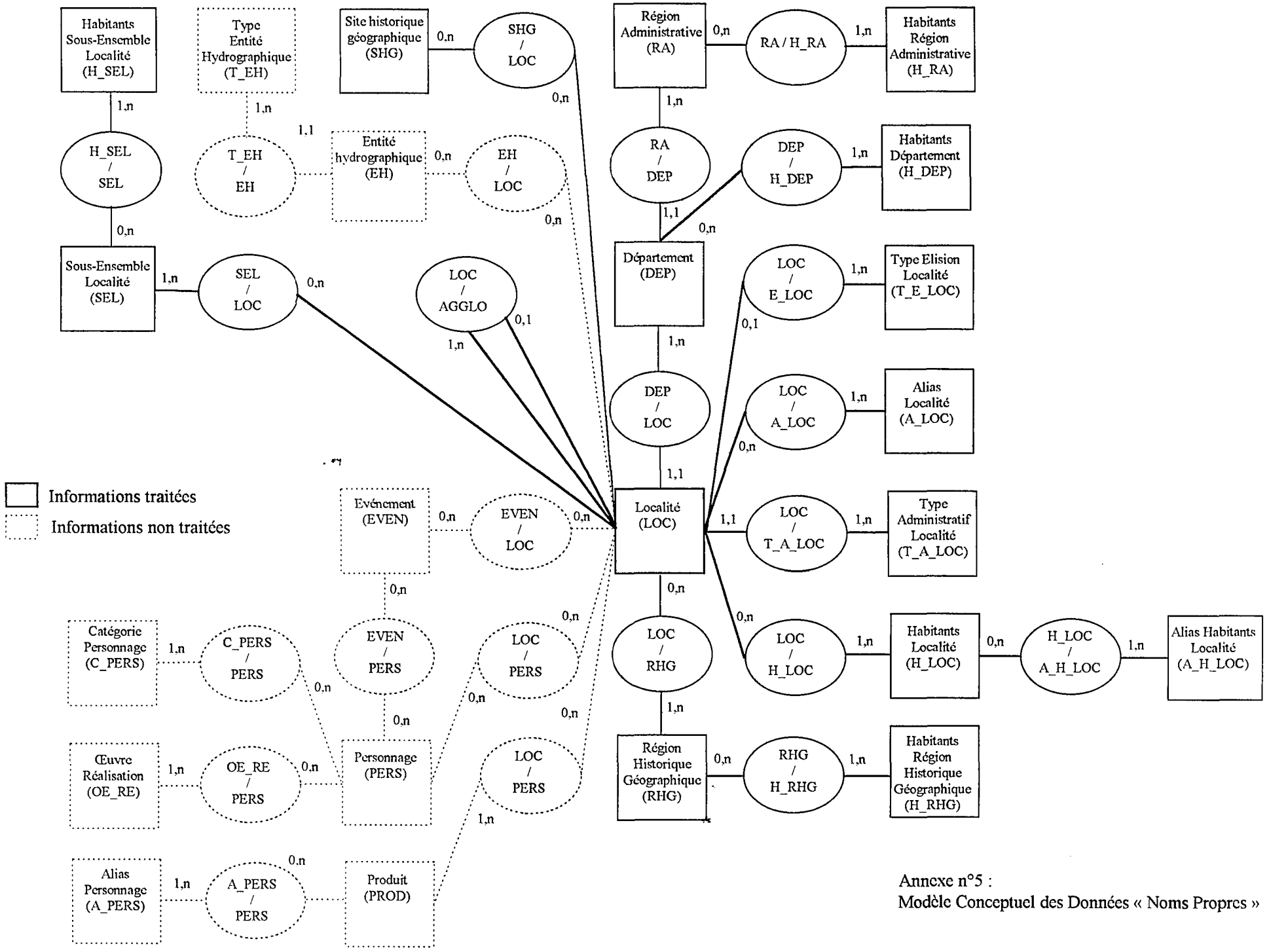
Une localité est parfois connue ou reconnue par un ou des *personnages célèbres*, un ou des *évènements historiques*, un ou des *produits typiques*, ou encore un *quartier*, un *monument* ou autre...

* La Redoute fortifications du type Vauban BAIE de l'ORNE.
* Les Batteries de Merville vestiges Allemands de la dernière guerre en cours de classement.

D'autre part, connaissez-vous l'origine du nom des habitants de votre localité : Voir ci-joint extrait du Bulletin Municipal des 1es heures 78.

Enfin, un de nos partenaires souhaite connaître la date de création du *premier bureau de poste ou point postal* dans votre localité :

Sur information le bureau existe avant la guerre de 40.



Annexe n°5 :
Modèle Conceptuel des Données « Noms Propres »

[illegible]	[illegible]	[illegible]	[illegible]
1		0	5
1	e	0	2
1	es	0	4
2		0	1
2	ne	0	2
2	s	0	3
2	nes	0	4
3		0	1
3	e	0	2
3	s	0	3
3	es	0	4
4		0	6
4	s	0	7
5		0	1
5	sse	0	2
5	s	0	3
5	sses	0	4
6	er	2	1
6	ère	2	2
6	ers	2	3
6	ères	2	4
7	et	2	1
7	ète	2	2
7	ets	2	3
7	ètes	2	4
8	u	1	1
8	lle	1	2
8	us	1	3
8	lles	1	4
9	au	2	1
9	lle	2	2
9	aux	2	3
9	lles	2	4
10	l	1	1
10	le	1	2
10	aux	1	3
10	les	1	4
11	iaux	4	1
11	elle	4	2
11	iaux	4	3
11	elles	4	4
12	c	1	1
12	que	1	2
12	cs	1	3
12	ques	1	4

Table des flexions de gentilés

algorithme de construction du transducteur d'Identifiants et de codes Genre et Nombre

```
Trans_ident_GN ( identifiant, mot, code GN )

DEBUT Trans_ident_GN
   p ← 0
   a ← caractère lu
   // traitement de la partie préfixe
   TANT QUE non fin de mot et transition (p, a ,q)
      SI un identifiant sur la transition
         ALORS
            controler si même identifiant qu'identifiant courant
            SI pas même identifiant
               ALORS
                  sauvegarde pour ajout nouvel identifiant en cas d'homonymie
               SINON
                  Indentifiant_posé ← Vrai
            FINSI
      FINSI
      SI transition pointe sur un état de sortie // qui contient le code Genre et Nombre
         ALORS
            SI le mot lu n'est pas terminé
               ALORS
                  débranche transition (p, a, etat_final)
                  création d'un nouvel état
                  branche transition (p, a, q)
                  nouvel état de sortie
            FINSI
            faire progresser p
         SINON
            SI une seule transition sur q
               ALORS
                  faire progresser p
               SINON
                  qprime ← duplique état
                  débranche (p, a, q)
                  branche qprime
                  branche q
                  q←qprime
            FINSI
      FINSI
      caractère suivant
   FIN TANTQUE
   SI tout le mot parcouru
      ALORS
         SI transition pointe sur un état de sortie
            ALORS
               SI Identifiant_posé = Faux
                  ALORS
                     // c'est un homonyme
                     ajout de l'identifiant sur la transition sauvegardée
               FINSI
            SINON
               // c'est un mot inclus
               SI Identifiant_posé = Vrai
                  ALORS // construction d'une des formes fléchies d'un gentilé
                     construire état de sortie
                  SINON // construction d'un mot inclus avec un identifiant propre
                     construire état de sortie avec identifiant
               FINSI
         FINSI
      SINON
         // traitement de la partie suffixe
            fin_préfixe ← p
            q ← etat_sortie // qui contient le code Genre et Nombre
            b ← dernier caractère
            // recherche des transitions en reculant
            TANT QUE  position (b) <> position (a) et
                      transition (p, b ,q) et
                      une seule transition à partir de b et
                      p <> fin_préfixe
               q ← p
               b ← caractère précédent
```

```
          FIN TANT QUE
          // raccordement fin du préfixe avec début du suffixe en allouant des
transitions
          // dépose de l'identifiant sur la première transition suivant la fin du
préfixe
          début_suffixe ← q ;
          p ←  fin_préfixe
          identifiant_posé ←  Faux
          TANT QUE position (b) <> position(a)
             q ← nouvel_état
             SI identifiant_posé = Faux
                ALORS
                   identifiant_posé ← Vrai
                   // création de la transition et dépose de l'identifiant
                   alloue_transition_et_identifiant (p, a, identifiant, q)
                SINON
                   // création de la transition
                   alloue_transition (p, a, q)
             FINSI
             branche_transition
             p ← q
             caractère suivant
          FIN TANT QUE
          alloue_transition (p, a, début_suffixe)
          branche_transition
   FINSI
FIN Trans_ident_GN
```

Annexe 8 : Détail de la structure interne du transducteur d'associations