Article

« Un dictionnaire relationnel des noms propres liés à la géographie consulté par transducteurs »

Claude Belleil et Denis Maurel

Meta : journal des traducteurs / Meta: Translators' Journal, vol. 42, n° 2, 1997, p. 273-282.

UN DICTIONNAIRE RELATIONNEL DES NOMS PROPRES LIÉS À LA GÉOGRAPHIE CONSULTÉ PAR TRANSDUCTEURS*

CLAUDE BELLEIL ET DENIS MAUREL[1]
IRIN, Université de Nantes, et LI/E3i Université François-Rabelais, Tours, France

Résumé

Peu traités informatiquement jusqu'à présent, les noms propres et les noms qui leur sont associés demandent des informations supplémentaires à leur véritable interprétation. Le projet développé à Nantes portant sur les noms propres prévoit la création d'une base de données relationnelle entre les noms propres liés à la géographie. Il sera ainsi possible de reconnaître et d'associer, par exemple, des noms propres directement à la lecture même s'il s'agit d'un lien métaphorique entre le nom d'une ville et celui de ses habitants.

Abstract

Until recently proper names and those nouns generally associated with them have not been systematically linked electronically. Additional information is often required for them to be properly understood. The aim of the project being carried out at Nantes is to create a database linking proper names and place names. This will make it possible to automatically recognize and associate place names even if the association between the name of a city and its inhabitants is metaphorical.

INTRODUCTION

Lorsqu'on utilise, dans le traitement automatique de la langue naturelle, un dictionnaire électronique[2], l'une des premières difficultés est constituée par la présence de mots inconnus, difficulté qui entre dans le cadre du *non-attendu*. Or, un échec lors de la phase d'analyse lexicale peut compromettre la totalité du traitement, en laissant subsister des obscurités qui ne pourront pas être résolues dans les étapes ultérieures.

La notion de *non-attendu* (Ren et Perrault 1992) renvoie à des situations très diverses parmi lesquelles on notera :

- les mots mal orthographiés ;
- les noms propres ;
- les néologismes ou les dérivations peu communes (Clemenceau 1993) ;
- les structures syntaxiques non prévues ou incorrectes ;
- les métaphores.

La rencontre d'un nom propre peut varier en fréquence selon la nature des textes analysés. Dans certains cas, par exemple les textes journalistiques ou ceux faisant référence à des notions de géographie ou d'histoire, leur présence est très importante et nécessite véritablement un traitement particulier.

Comment intégrer le traitement automatique des noms propres dans l'analyse automatique ? Quels sont les processus à mettre en œuvre pour y arriver ? Ce sont les travaux de recherche que nous avons entrepris à Nantes dans le cadre de l'IRIN (Institut de Recherche

en Informatique de Nantes) en construisant un dictionnaire électronique relationnel des principaux noms propres liés à la géographie, dictionnaire qui sera consulté par transducteur directement lors de la lecture du texte. Ces travaux sont réalisés en association avec trois partenaires : la direction départementale de la Poste, le CNAM de Nantes et le journal *Ouest France*.

1. LES CARACTÉRISTIQUES PARTICULIÈRES DES NOMS PROPRES

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. 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). Les dictionnaires électroniques utilisés actuellement ne comportent pas ou peu de noms propres. Parfois, ceux-ci font l'objet d'un traitement *ad hoc*.

Quels sont les éléments qui permettent de déterminer de façon claire que l'on a affaire à un nom propre ? Comme le fait remarquer Gross (1990 : 123), «...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 des 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».

La question importante de la mise en œuvre d'un dictionnaire électronique des noms propres est donc posée. Quelle limite fixer à son caractère encyclopédique qui est par nature illimité ?

Un dictionnaire des noms propres repose sur la notion de notoriété. Mais 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 choix 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.

Nous avons fait le choix de limiter notre étude, dans un premier temps, aux noms propres liés à la géographie. Les noms propres se répondent les uns aux autres au travers de réseaux relationnels qui sont porteurs de sens. Ces associations et ces relations de dépendance sont particulièrement mises en évidence dans des corpus spécialisés liés à l'histoire et à la géographie, comme les guides touristiques, mais également dans les articles de journaux. Cependant, la mise en place de réseaux de type relationnel nous a conduit à étendre le cadre de nos travaux, tout d'abord au niveau de la collecte des données (régions historiques, personnages célèbres, événements marquants...), mais également au niveau de la définition des structures qui permettront par la suite de traiter l'ensemble de ces informations.

Ces recherches visent donc à doter les systèmes d'analyse automatique du langage naturel de moyens de traiter les noms propres sur le plan lexical et pragmatique. Ils s'articulent autour des axes suivants :

- reconnaissance et typage par transducteur ;
- interrogation d'un dictionnaire électronique relationnel permettant de mettre en évidence les associations possibles.

2. RECONNAISSANCE PAR TRANSDUCTEUR, GÉNÉRATION D'UNE CLÉ ET TYPAGE

2.1. Détermination d'une forme canonique

2.1.1. Les graphies multiples

Il existe tout d'abord un réel problème de graphies multiples des noms propres liés à la géographie. Il ne s'agit pas seulement des variantes orthographiques qu'il nous faudra, bien sûr, répertorier. Ce phénomène concerne aussi certains éléments du nom propre qui peuvent être notés en abrégé, par exemple :

Grand → Gd
Pont → Pt
Saint(e) → St(e)
Mont → Mt
sur → /
sous → s

À cela s'ajoute la présence ou non du tiret de normalisation de la Poste pour les noms de lieux formés de plusieurs mots et celle, en principe obligatoire, d'une majuscule au début de chaque composant, à l'exception des prépositions. Ainsi, 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 écritures différentes. Ce nombre passe à 64^3 si l'on veut gérer également la présence des majuscules au début de chacun des deux composants internes. Le stockage de toutes ces combinaisons dans un dictionnaire n'est pas envisageable. En revanche, la reconnaissance de toutes ces formes peut être réalisée simplement en utilisant un automate à nombre fini d'états (figure 1).

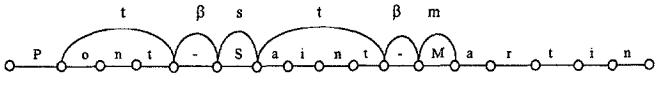


Figure 1 :
Un automate[4] reconnaissant les 64 graphies de *Pont-Saint-Martin*

Il en est de même pour les noms des personnes, avec la présence ou non du prénom, celui-ci pouvant être abrégé, en principe, par son initiale en majuscule suivie (ou non) d'un point : *Albert Camus*, *A. Camus* ou encore *Camus*.

2.1.2. Morphologie flexionnelle

Les noms des habitants des communes, des régions et des pays ont une morphologie flexionnelle. Il est donc nécessaire de reconnaître les quatre formes fléchies correspondant

au genre et au nombre pour leur associer une **forme canonique** qui sera seule présente dans le dictionnaire relationnel (Maurel et Belleil 1995).

2.1.3. Morphologie dérivationnelle

Si de nombreux noms géographiques engendrent par dérivation les noms d'habitants correspondants, il existe aussi, dans ce domaine, une très forte proportion de dérivations à partir de formes supplétives. Ainsi, les habitants de *Verzenay* en Champagne se nomment les *Bouquins* ! Il est donc exclu de fonder uniquement la reconnaissance des noms d'habitants sur un calcul à base de règles (Maurel 1994), sauf dans le cadre du *non-attendu* pour des noms n'appartenant pas au dictionnaire. Une réflexion est en cours à ce sujet à partir de Eggert (1994).

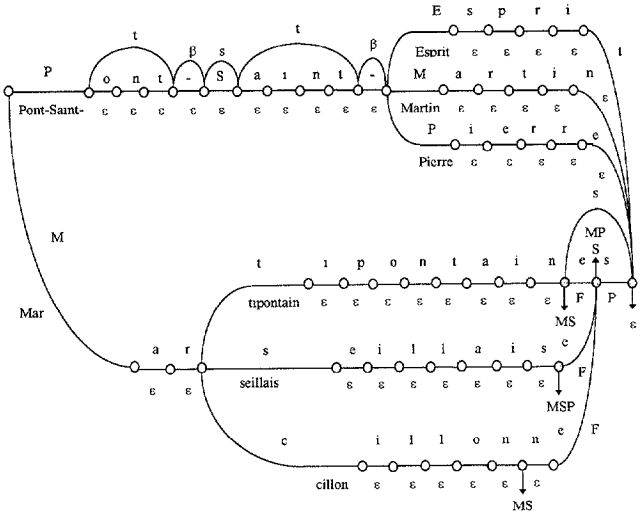


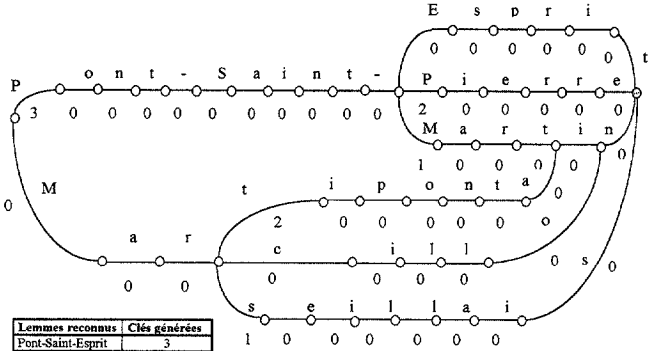
Figure 2 :
Un transducteur[5] générant des formes canoniques de noms propres

Pour traiter la morphologie flexionnelle sur le plan informatique, il est possible de conserver le même modèle que celui qui a été présenté à la figure 1, chaque état final pouvant alors porter de l'information. Il s'agit d'automates à multiterminaux (Revuz 1991 : 62). Cependant, des études récentes ont mis en évidence l'intérêt qu'il y avait à générer l'information dès que possible, en autorisant aussi les transitions à porter de l'information (Mohri 1994). On parle alors de transducteurs à nombre fini d'états. Reprenons l'exemple de *Pont-Saint-Martin* en ajoutant la reconnaissance du nom des habitants de la commune, les *Martipontains*. Le transducteur présenté en figure 2 illustre la génération de *Pont-Saint-Martin* sitôt effectuée la reconnaissance de la lettre *M* entre son prédécesseur

Pont-Saint-Esprit[6] et son successeur *Pont-Saint-Pierre*[7] ainsi que celle de *Martipontain* sitôt faite la rencontre de la lettre *t* entre ses prédécesseurs *Marcillon*, *Marcillons*, *Marcillonne*, *Marcillonnes*, habitants de *Marcilly-le-Hayer*[8] et ses successeurs *Marseillais*, *Marseillaise*, *Marseillaises* : habitants de *Marseille*.

2.2. Un transducteur de hachage

Une fois le nom propre reconnu et la forme canonique générée, un transducteur de hachage (Revuz 1991 : 67) permet de fournir une clé unique pour chaque lemme. Celles-ci se situent dans un intervalle [0, n-1], n correspondant au nombre de mots reconnus. La méthode consiste à compter pendant le parcours le nombre de mots lexicographiquement inférieurs au mot recherché. Ce nombre correspond à la position qu'occuperait le mot s'il était rangé dans un tableau. On obtient cette clé en additionnant les émissions successives, par exemple, pour *Pont-Saint-Martin* : 3+0+0+0+0+0+0+0+0+0+0+1+0+0+0+0+0=4.



Lemmes reconnus	Clés générées
Pont-Saint-Esprit	3
Pont-Saint-Pierre	5
Pont-Saint-Martin	4
Martipontain	2
Marcillon	0
Marseillais	1

Figure 3 :
Un transducteur de hachage délivrant des clés uniques pour chaque lemme

2.3. Le typage

La forme normalisée, puis la clé, étant générées, il faut ensuite provoquer une entrée dans le dictionnaire relationnel afin de rechercher les associations possibles. Se pose alors le problème du type. Un même mot peut correspondre à des noms propres appartenant à des types différents. Cette détermination est essentielle pour le processus d'analyse automatique. La détection d'un type unique déclenchera des essais d'associations avec les types de noms propres en relation directe avec le type détecté. Ainsi le mot *Loire-Atlantique* appartient sans aucune ambiguïté au type «nom de département». 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 qui va donc conduire à de multiples associations parmi lesquelles il faudra déterminer celles qui sont pertinentes (figure 4).

TYPE DÉTECTÉ	TYPE ASSOCIÉS	ASSOCIATIONS
département	région administrative	Pays de Loire
	habitants département	Mayennais
	commune	Mayenne, Laval, Château-Gontier
commune	habitants commune	Mayennais
	département	Mayenne
	hydrographie	Mayenne
	région historique	Vendée Militaire
	événement	Guerre de Vendée
personnage	alias personnage	Charles de Lorraine
	commune	Alençon (né à...). Soissons (mort à ...)
	type personnage	homme politique
	événement	la Sainte Ligue, Guerres de Religion

Figure 4 :
Le lemme *Mayenne* peut appartenir à quatre types différents

Souvent, une grammaire locale : *ville de*, *préfecture de*, *duc de*, permet de lever l'ambiguïté en déterminant le typage de façon plus rapide parmi tous les types candidats. Nous allons donc poursuivre notre travail de reconnaissance par une analyse linguistique du contexte des noms propres et par la construction de grammaires locales (Garrigues 1993). Pour cela, nous disposerons d'un corpus de deux années d'articles du journal *Ouest France*.

Ensuite, c'est par un index associant la clé générée par le transducteur de hachage aux clés d'accès dans le dictionnaire électronique relationnel que les associations seront déclenchées. On notera à ce sujet l'indépendance totale entre les clés d'accès aux tables relationnelles et celles du transducteur de hachage, ce qui permettra de gérer de façon indépendante ces deux outils informatiques.

2.4. Les différents types

La structure générale de ce dictionnaire repose donc sur un recensement des différents types de noms propres. Nous avons commencé ce travail en partant des noms de lieux et d'habitants mais, comme on peut le constater sur le modèle conceptuel des données de la base, un type en appelle souvent un autre, qui lui-même fera référence à un troisième...

Voici la liste des types que nous avons envisagé de traiter :

COMMUNE (agglomération, ville, banlieue, ville moyenne, village, bourg, hameau)
ALIAS COMMUNE
TYPE DE COMMUNE (port, station balnéaire, station thermale)
QUARTIER, LIEU (sous-ensemble de la commune)
HABITANT COMMUNE
ALIAS HABITANT COMMUNE
DÉPARTEMENT
HABITANT DÉPARTEMENT
RÉGION ADMINISTRATIVE
HABITANT RÉGION ADMINISTRATIVE
RÉGION HISTORIQUE ou GÉOGRAPHIQUE
ALIAS RÉGION HISTORIQUE ou GÉOGRAPHIQUE
HABITANT RÉGION HISTORIQUE ou GÉOGRAPHIQUE
ALIAS HABITANT RÉGION HISTORIQUE ou GÉOGRAPHIQUE
LANGUE RÉGIONALE
SITE ou LIEU HISTORIQUE ou GÉOGRAPHIQUE
HYDROGRAPHIE
ÉVÉNEMENT
PERSONNAGE
ALIAS PERSONNAGE
TYPE de PERSONNAGE
ŒUVRE(s) ou RÉALISATION(s)
PRODUIT

Ces choix ont des conséquences sur la collecte et la saisie des informations. Pour ce qui concerne les noms de villes (communes) et d'habitants, contrairement à ce que nous pensions, nous avons découvert que de telles listes n'existaient pas. Nous avons pu rassembler quelques éléments partiels en interrogeant l'IGN, en consultant certains ouvrages et dictionnaires, ainsi que les guides verts Michelin. Cependant, dans la plupart des cas, on ne trouve que le nom des habitants au masculin pluriel. Il est vrai qu'il n'existe pas de normalisation dans ce domaine. Les noms d'habitants de lieux peuvent être hérités de la tradition orale ou écrite, décrétés par une autorité administrative centrale, comme cela a parfois été le cas au XIX^{e} siècle ou, surtout depuis les lois de décentralisation, être fixés par une décision de l'autorité territoriale (conseil municipal).

Le caractère hétéroclite et partiel des matériaux dont nous disposions nous a donc conduit à envisager le lancement d'une enquête nationale en recherchant le partenariat d'organismes susceptibles d'être intéressés par le résultat de nos travaux, en l'occurrence le journal *Ouest France* et la Poste.

L'avantage d'une telle enquête était de connaître non seulement les noms des habitants des différentes communes concernées, mais aussi de poser d'autres questions dont les réponses sont pratiquement introuvables en dehors du «terrain», par exemple, le lien que certaines communes entretiennent avec les régions que nous avons hérité de l'histoire ou de la géographie. En effet, 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 commune a le sentiment d'appartenir à la *Saintonge*, à l'*Argonne*, au *Bugey*, à la *Petite Crau*, à l'*Occitanie* ou au *Trégorrois*. Ajoutons à cela que certaines régions administratives portent le même nom que des régions historiques sans faire référence à un même espace géographique ! Ainsi, la *Vendée Militaire* s'étend largement au-delà du département de la *Vendée*.

L'objectif que nous nous sommes fixé, dans un premier temps, était de rassembler les données concernant :

- toutes les villes de plus de 10 000 habitants (recensement INSEE 1975) ;
- toutes les villes connues pour leur intérêt historique, géographique, touristique... et citées de ce fait dans les guides verts Michelin ;
- enfin, à la demande d'un de nos partenaires, le journal *Ouest France*, la grande majorité des communes de l'Ouest.

Ce qui correspond à un ensemble de 5 980 communes.

3. RECONSTRUCTION DES ASSOCIATIONS PAR CONSULTATION DU DICTIONNAIRE ÉLECTRONIQUE

Comme nous l'avons déjà précisé, les noms propres n'appartiennent pas à l'ordre de la définition mais à celui de la description. Ainsi, quand on étudie un article consacré à un nom propre dans un dictionnaire imprimé, on constate qu'il s'agit d'une description de type encyclopédique visant, entre autres, à recenser l'ensemble des relations que le nom propre entretient avec d'autres noms propres.

Un nom propre est porteur de coréférences avec d'autres noms propres. Ainsi, il est fréquent de voir dans un même texte (surtout dans les articles journalistiques) des noms de lieux, d'habitants et de régions renvoyant les uns aux autres sans que ces liens puissent être mis en évidence au travers de règles morphologiques. Voici, par exemple, un extrait d'un article de commentaire sportif du journal *Ouest France* du lundi 20 mars 1995 :

> *«ESO **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 cet exemple, le nom des habitants *Thouaréens* correspond à la dérivation du nom de la commune *Thouaré*. *Ornaisiens* fait référence au nom des habitants de *Saint-André-d'Ornay*, quartier (sous-ensemble) appartenant à la commune de *La Roche-sur-Yon*. Enfin, les *Vendéens* renvoient soit au nom de la région historique dans laquelle est située la commune de *La Roche-sur-Yon*, soit au nom du département de la *Vendée*, ce qui conduit au même résultat.

Il est évident que ces relations, qu'elles soient duales ou multiples, sont porteuses de sens et que, dans un processus d'analyse automatique, la possibilité de les reconstruire contribuera à la compréhension globale du texte (figure 5).

Comme nous l'indiquions au début de cet article, la construction d'un réseau relationnel de noms propres liés à la géographie nous a conduit à un recensement des associations possibles. Cette construction a été réalisée à partir de l'étude de certains corpus spécialisés comme les articles de journaux et les extraits de guides touristiques (figure 6).

L'enquête réalisée a également fait ressortir des éléments qu'un travail documentaire n'aurait pas permis avec une aussi grande précision dans le détail. En premier lieu, le très grand nombre de petites régions historiques et/ou géographiques auxquelles les habitants de certaines communes ont le sentiment d'appartenir. Chacune d'entre elles renvoie également à des noms d'habitants qui correspondent soit à des formes flexionnelles (les *Dombistes* pour la *Dombes*, les *Thiérachiens* pour la *Thiérache*...), soit à des

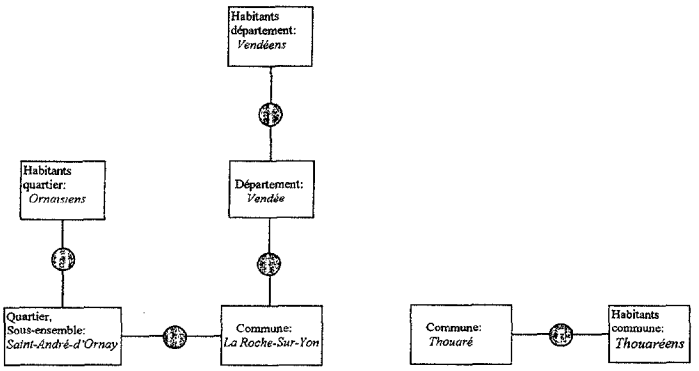


Figure 5 :
Exemple d'associations de noms propres

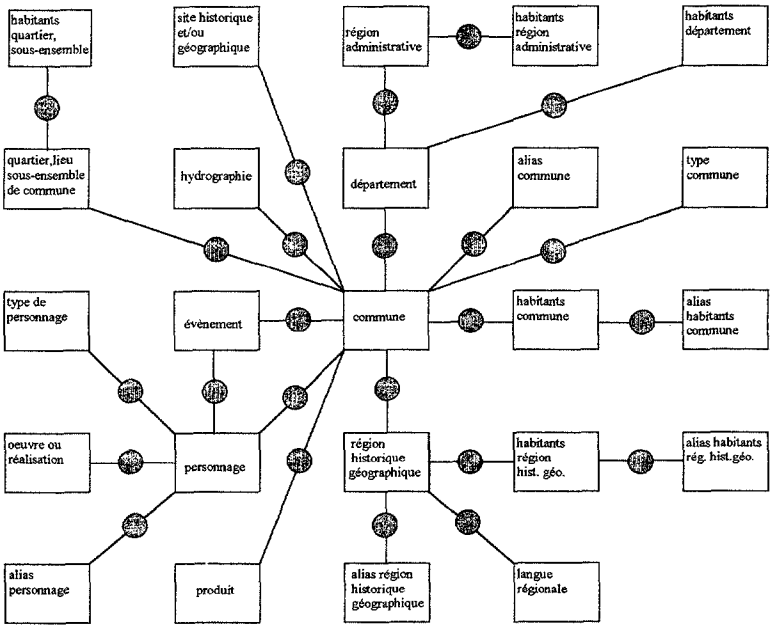


Figure 6 :
Le réseau des noms propres du dictionnaire relationnel

formes supplétives (les *Gavots* pour le *Comté de Nice*). Il faudra probablement prendre également en compte la très grande densité du réseau hydrographique du pays : rivières, canaux, étangs, lacs, qui apportent eux aussi autant de noms propres en lien direct avec le nom des communes où ils sont situés.

CONCLUSION

Ces travaux ont donc mis en évidence, s'il en était besoin, la difficulté de proposer un traitement morphologique et pragmatique des noms propres. Le problème du volume des données à stocker, et donc à consulter, est sans aucun doute l'un des plus importants. Le risque d'un encyclopédisme ingérable sur le plan informatique nous a conduit à limiter dans un premier temps notre travail aux noms propres liés à la géographie.

Cependant, l'utilisation de représentations informatiques adaptées et d'algorithmes performants nous permettra d'apporter une réponse à ce problème et d'intégrer à court terme le traitement des noms propres à celui du langage naturel.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV^es^ Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. Denis Maurel est aussi membre associé du laboratoire d'informatique de l'École d'Ingénieurs en Informatique pour l'Industrie de l'Université de Tours.
2. Ce qui n'est pas toujours le cas, voir par exemple Enguehard (1992).
3. Ou 128 si l'on souhaite aussi reconnaître le mot avec une minuscule initiale.
4. Le symbole β désigne ici le blanc (caractère espace).
5. M = Masculin, F = Féminin, S = Singulier, P = Pluriel, ε = pas d'émission.
6. Localité du Gard.
7. Localité de l'Eure.
8. Localité de l'Aube.

RÉFÉRENCES

CLÉMENCEAU, D. (1993) : *Structuration du lexique et reconnaissance de mots dérivés*, Paris, Thèse de doctorat, Université Paris VII.

EGGERT, E. (1994) : *Étude dérivationnelle des dérivés de toponymes*, Mémoire de maîtrise, Université de Lille III.

ENGUEHARD, C. (1992) : *ANA, Apprentissage Naturel Automatique d'un réseau sémantique*, Thèse de doctorat en contrôle des systèmes, Université Technologique de Compiègne.

GARRIGUES, M. (1993) : «Prépositions et noms de pays et d'îles : une grammaire locale pour l'analyse automatique des textes», *Linguisticae Investigationes*, 17 (2), pp. 281-305.

GROSS, M. (1990) : «Le programme d'extension des lexiques électroniques», *Langue française*, n° 87, Paris, Larousse.

MAUREL, D. (1994) : «Le traitement informatique de la dérivation des noms de ville», *TA information*, 35 (2), Paris, pp. 111-127.

MAUREL, D. et C. BELLEIL (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.

MOHRI, M. (1994) : *Application of Local Grammars Automata an Efficient Algorithm*, Rapport de recherche IGM 94-16, Université Marne-la-Vallée.

REN, X. et F. PERRAULT (1992) : «The Typology of Unknown Words: An Experimantal Study of Two Corpora», *COLING 92*, Nantes.

REVUZ, D. (1991) : *Dictionnaires et lexiques — Méthodes et algorithmes*, Thèse de doctorat en informatique, Université Paris VII.