Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé

Julien Lesbegueries

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

ÉCOLE DOCTORALE DES SCIENCES EXACTES ET DE LEURS APPLICATIONS ED 211

Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé

THÈSE

présentée et soutenue publiquement le 26 novembre 2007

pour l'obtention du

Doctorat de l'Université de Pau et des Pays de l'Adour

(spécialité informatique)

par

Julien Lesbegueries

Composition du jury

Rapporteurs :	Pr. Stefano Spaccapietra	EPFL Lausanne
	DR. Christian Fluhr	CEA Paris
	Pr. Patrick Gallinari	Université Pierre et Marie Curie Paris 6
Examinateurs :	Dr. Christophe Tuffery	Société Makina Corpus
	Pr Thérèse Libourel	Université de Montpellier 2
	Pr. Mauro Gaio	Université de Pau et des Pays de l'Adour
	MdC. Christian Sallaberry	Université de Pau et des Pays de l'Adour

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour — EA 3000

LIUPPA

Mis en page avec la classe thloria.

Remerciements

Je souhaite tout d'abord remercier mes encadrants de thèse, Mauro Gaio et Christian Sallaberry pour leur soutien permanent durant ces 3 ans de thèse, leurs conseils avisés et la confiance qu'ils ont bien voulu accorder. Je les remercie particulièrement pour leur ouverture d'esprit et l'émulation qu'ils ont su créer au sein de l'équipe de recherche.

Je tiens également à remercier vivement Christian Fluhr, Stefano Spaccapietra et Patrick Gallinari pour avoir accepté de rapporter ma thèse, ainsi que Thérèse Libourel et Christophe Tuffery pour avoir accepté de faire partie de mon jury.

Je remercie également les membres de l'équipe-projet DESI (feu IDEE) pour m'avoir accueilli chaleureusement, m'avoir accepté comme l'un des leurs et avec qui j'ai pu échanger des idées lors des multiples réunions. Ils ont su m'apporter une aide précieuse, de la rédaction du premier article à la relecture finale de mon mémoire.

Je n'oublie pas la « MIDR » et la Communauté d'Agglomération de Pau qui ont financé ma bourse de thèse. Je les en remercie et, de manière plus particulière, les membres la MIDR pour les échanges que nous avons eu et pour nous avoir fourni de multiples échantillons de corpus afin de tester nos réalisations.

Je souhaite aussi témoigner de ma sympathie aux caennais du GREYC et en particulier Frédérik Bilhaut, Antoine Widlöcher et Patrice Enjalbert à qui j'ai rendu quelques visites au début de ma thèse pour comprendre Linguastream ;-). Je remercie aussi vivement Victor Montes de Oca qui m'a sûrement sauvé la vie plus d'une fois à Mexico pour ma première conférence à l'étranger.

J'arrive enfin aux amis du couloir, et je commencerai par saluer Julien Lacoste, mon compagnon d'infortune, Pierre Laforcade qui avant de partir a eu le temps de bien nous faire rigoler, Fabien Romeo dont le flegme m'impressionnera toujours, puis Pierre Loustau qui a rejoint l'équipe-projet DESI et avec qui j'ai pu démarrer ma thèse. Une pensée aussi aux derniers arrivants Cyril Ballagny, Natacha Hoang, Eric Kergosien, Damien Palacio et Le Thanh Vu pour qui je souhaite un avenir plein de réussite.

Et parce qu'il n'y a pas que les doctorants, je salue aussi Laurent, Annig, Malo et Maya Lacayrelle, Eric Cariou et les autres « sportifs » (ou pas) du couloir qui m'ont fait apprécier le VTT et la bière après le sport.

Mes derniers remerciements seront pour ma famille, en particulier mes parents qui m'ont finalement toujours laissé faire ce que je voulais, Sophie et Yoann qui me pardonnent qu'on ne se soit pas vu assez ces derniers temps et mes cousins qui me soutiennent dans la voie que j'ai choisie. Je compte dans ma famille aussi toute la délégation tunisienne qui est venue se refroidir sous notre latitude trop peu tempérée : Wadi3, Fahmy d'abord, puis Ines, Najla, Asma et bien sûr Maouleti Jihene qui a su me faire aimer la cuisine épicée (et pourtant ce n'est pas évident d'après Julien).

Je dédie cette thèse à ma famille,
à ceux qui sont partis
et surtout à celle qui est revenue.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Première partie

Introduction générale

Chapitre 1

Contexte de la thèse : valorisation de fonds documentaires localisés *via* l'information géographique

Les travaux présentés dans ce manuscrit s'inscrivent au sein d'une dynamique de recherche conduite au LIUPPA dans le cadre du Projet Desi[1], centrée sur l'élaboration de nouvelles méthodes d'accès à l'information par le contenu des documents. Le contexte particulier de ma thèse, soutenue par la Communauté d'Agglomération de Pau, a entraîné un partenariat avec la Médiathèque Inter-Communale à Dimension Régionale de Pau (la MIDR). Au cours de ce partenariat, des échanges ont eu lieu sur les besoins actuels des médiathèques, notamment concernant la gestion par le contenu de documents numérisés et ont permis d'orienter nos problématiques de recherche. Pour nos expérimentations, la MIDR a mis à notre disposition leur corpus numérisé et nous avons en retour proposé des méthodes innovantes d'accès à l'information sous la forme d'un projet de recherche nommé PIV[2].

1.1 Analyse des besoins pour la valorisation d'un corpus à connotation territoriale

Les premiers travaux ont tout d'abord consisté en une étude préliminaire du corpus numérisé. Celle-ci a révélé une connotation géographique prédominante dans les documents, aussi bien dans les œuvres littéraires, traitant de récits de voyages, que dans les périodiques locaux dont les articles relatent des informations sur le territoire. Une expérimentation a montré par exemple que près de 10 000 entités nommées à connotation spatiale ont été extraites de 10 des livres du corpus (soit 600 000 mots). Une exploration de ce corpus du point de vue de l'information géographique semble donc être une démarche pertinente.

[1] http://liuppa.univ-pau.fr/DESI/
[2] pour Pyrénées Itinéraires Virtuels.

Nous nous sommes alors aussi intéressé à la caractérisation des types d'usagers susceptibles d'être intéressés par la consultation d'un tel corpus.

Caractéristiques du corpus La médiathèque, dans une optique de valorisation, a numérisé et réalisé une ROC[3] de son fonds documentaire patrimonial afin de l'indexer dans un système de recherche documentaire. De cette manière, les documents numérisés peuvent bénéficier d'une nouvelle visibilité et être parcourus par un large public. Il faut préciser que cette numérisation, compte-tenu du coût de l'opération, a été réalisée sans correction d'erreurs ni récupération de la structure des documents.

Ce corpus se compose de documents aux formats divers (œuvres littéraires, récits de voyage, journaux, cartes géographiques anciennes, lithographies, cartes postales, etc.) et qui ont pour dénominateur commun de traiter d'un territoire restreint (les Pyrénées[4]), dans une période de l'Histoire déterminée (principalement du XVIII^{e} et du XIX^{e} siècle).

Une lecture du corpus nous a permis de constater que de nombreuses informations sont associées à une localisation nommée du territoire auxquelles elles se réfèrent. Particulièrement dans les documents textuels, de nombreuses informations sont associées à des lieux géographiques, des indications spatiales, des descriptions de paysages, des indicateurs temporels et de dates impliquant un intérêt marqué de ces documents pour l'aspect géographique. La majorité des documents textuels (œuvres littéraires) sont constitués de récits de voyage (voir un extrait dans la Figure. 1.1). Les auteurs de ces œuvres utilisent la plupart du temps une structure identique. Le texte est découpé en paragraphes décrivant un passage de leur voyage. Ce passage peut consister en une description d'itinéraire, une description d'une étape, un point de vue ou une comparaison de lieux. Nous avons ainsi dégagé ces différents motifs spatiaux récurrents dans les récits de voyage.

Caractéristiques des usages potentiels Les documents qui composent le corpus sont particulièrement intéressants de par leur richesse en indications géographiques du territoire pyrénéen. Un usage touristique est envisageable, plus précisément pour un tourisme rural, permettant la découverte d'une région. L'histoire d'une région peut aussi vraisemblement intéresser les gens qui l'habitent. Un usage pédagogique peut alors être défini pour enrichir les connaissances de ses habitants et éveiller les plus jeunes à leur lieu de vie. Enfin nous avons imaginé un usage de spécialiste qui peut parcourir les ressources de la médiathèque plus efficacement en utilisant les outils avancés de Gestion Électronique de Documents (GED ou GEIDE[5]). Tous ces usages nécessitent donc l'exploitation du corpus documentaire *via* un système d'information adapté, en particulier capable d'offrir des possibilités de recherche du point de vue du territoire décrit par ce corpus. Des systèmes de recherche d'information automatiques apportent d'ailleurs de plus en plus une aide complémentaire aux professionnels des bibliothèques [FG04].

[3] Reconnaissance Optique de Caractères. Cette technique, qui interprète la forme des lettres, permet de retrouver le texte d'une œuvre au format image.

[4] http://fr.wikipedia.org/wiki/Pyrénées

[5] Gestion Electronique d'Informations et de Documents pour l'Entreprise. Par extension, système de gestion électronique de documents. Une définition complète se trouve sur http://fr.wikipedia.org/

```
1 ‘‘PARIS 6 AOÛT 1833
2 [...] Ma résolution est bientôt prise. Andiamo ! À six heures du soir je
3 monte en diligence pour Bordeaux, pour ce long voyage au milieu de
4 la poussière de l'été. C'est pourtant la route de France parcourue avec
5 plus de rapidité qu'aucune autre ; aussi étions nous bercé par les
6 mouvements ondulés de la voiture ! Nous voyons Orléans en bonnet
7 de nuit, et dans la journée, Tours devenue colonie anglaise. [...]”
```

FIG. 1.1 – Extrait de document - *Voyages inédits dans les Pyrénées, (1833-1859) Édition PyréGraph, 2002.*

Caractéristiques de l'information géographique L'élément central du corpus étant l'information géographique, nous avons cherché à définir cette notion qui intéresse de nombreux chercheurs dans divers domaines. Aussi bien des géographes et des chercheurs en géomatique[6] que des chercheurs en psychologie, des philosophes et des linguistes ont tenté de définir cette notion [Gai01].

Une définition pertinente pour une approche applicative est la base de nos travaux et est présentée dans ce mémoire. Celle-ci vient de la géomatique et considère l'entité géographique comme une molécule composée non seulement d'une composante spatiale, mais aussi d'une composante temporelle et d'une composante thématique ou « phénomène » [UTC04, Gai01]. Un exemple de texte exprimant une molécule géographique complète est « les instruments de musique dans les environs de Laruns au début du XIX^{e} siècle », les instruments jouant ici le rôle du thème.

En effet, aussi bien dans la spécification GML[7] que dans des travaux de recherche sur les bases de données [Le04] apparaît la notion de temporalité pour les entités géographiques (EG). À une EG peut être associée une ou plusieurs représentations géoréférencées (selon le modèle), qui sont valables à un moment donné de l'histoire [Gal01]. Par exemple une ville ou une forêt ont une définition spatiale variable au cours du temps, que ce soit une création, une disparition, une expansion ou une réduction. Enfin un phénomène est souvent associé, sujet d'étude de l'entité décrite (par exemple, une pollution sur un bassin géographique à une période donnée).

Objectifs de l'indexation visée Nous cherchons à améliorer la pertinence des résultats retournés lorsque une requête soumise à un système de recherche documentaire comporte des critères géographiques. Nous voulons étendre les fonctionnalités de tels systèmes afin qu'ils s'adaptent aux spécificités des corpus à forte connotation territoriale. En particulier, de nouvelles méthodes d'indexation et de requêtage seront à imaginer. Une indexation et un outil de requêtage adéquats permettront de retourner, à un utilisateur intéressé par un phénomène précis dans tel lieu et à telle époque, les documents les plus pertinents, géographiquement parlant. Par exemple, pour la requête *« les trajets*

wiki/GED

[6] http://fr.wikipedia.org/wiki/Géomatique

[7] http://www.opengis.net/gml/

Paris-Bordeaux au XIX^{e} siècle » sera retourné le paragraphe de la Figure 1.1. Des syntagmes plus complexes présents dans les documents, utilisant des notions d'adjacence, d'orientation, de distance, etc. pourront aussi être interprétés, même s'ils utilisent une combinaison de ces notions (*Ce terrain [...] couvrait tout l'espace que nous eûmes à parcourir ce jour, surtout depuis le col d'Albe jusqu'à la descente de la gorge de Malivierna...*, *Nous marquons un arrêt au Pic de Montferrat et continuons notre route par la longue arête assez aiguë qui s'en va au Sud-Est jusqu'au point côté 3147 que le comte Russell a appelé pic de Tapou...*). Ces morceaux d'information spatiale constitueront la base pour les processus d'indexation.

L'indexation devra permettre d'interpréter toute l'information géographique présente dans les documents, à tous les niveaux d'abstraction, aux moyens de méthodes spécifiques. Par niveau d'abstraction, nous entendons que l'information géographique peut être comprise à différentes granularités : tout d'abord au niveau élémentaire où chaque entité nommée est interprétée (dans le cas de l'exemple en Figure 1.1, les entités *« PARIS »*, *« pour Bordeaux »*, *« Orléans »* et *« Tours »* pour la composante spatiale et « 6 août 1833 » pour la composante temporelle). Des entités plus complexes utilisant des concepts du raisonnement qualitatif devront aussi, dans la mesure du possible, être récupérées (comme *près du sud de Pau*, *au nord de la frontière franco-espagnole*).

Toutefois, il est possible de faire une interprétation à des niveaux plus élevés dans lesquels des expressions géographiques sont associées pour être interprétées ensemble. Toujours dans l'exemple de la Figure 1.1, les quatre entités peuvent être regroupées pour former un itinéraire, de Paris à Bordeaux.

Des résumés du contenu géographique sont alors envisageables. Au niveau des phrases, des molécules sont créées pour chaque entité dégagée. Nous pouvons alors regrouper ces molécules en une nouvelle faisant la synthèse ; et ainsi de suite jusqu'à avoir une molécule pour l'œuvre entière. Par exemple, le livre contenant l'extrait de la Figure 1.1 traite de périples dans les Pyrénées. La molécule résumant l'extrait serait un itinéraire de Paris à Bordeaux se déroulant en août 1833. Puis, la molécule globale associée au livre pourrait être un itinéraire allant de Paris (lieu de départ de l'auteur) jusqu'au sud-ouest de la France et les Pyrénées. La manière de synthétiser ces entités devra elle aussi se baser sur une interprétation faite du discours. Il ne s'agit pas seulement d'agglutiner les molécules ensemble en regroupant celles qui sont spatialement, temporellement ou thématiquement proches. Nous émettons l'hypothèse que l'agencement de ces molécules dans une unité de texte peut faire sens et que leur interprétation permet de réaliser un résumé automatique. La méthode découlant de cette hypothèse doit permettre une recherche documentaire aisée et efficace, tout en préservant le sens de l'information géographique.

1.2 Travaux existants en Recherche Documentaire et Systèmes d'Information Géographique

Nous avons étudié différents outils existants dans divers domaines qui tentent de répondre aux problèmes posés. D'une part une étude du domaine de la recherche d'information et plus précisément des systèmes de gestion documentaire classiques a été réalisée

afin d'expertiser les manques potentiels pour une gestion de l'information géographique. D'un autre côté, nous avons étudié les dernières évolutions des systèmes d'information géographique concernant l'inclusion de fonctions spatiales qualitatives (ou floues) et leur intégration dans des systèmes de recherche d'information spécifiques.

La Recherche d'Information (RI) Les technologies employées dans les systèmes d'accès à l'information textuelle, dits de Recherche de l'Information ou de Recherche Documentaire ont été conçues dans les années 1970. Elles consistent en des mesures de similarité afin de retrouver des documents pertinents, à l'aide de requête en langage naturel, structuré ou bien par un ensemble de documents utilisés comme requête. À la fin des années 1990, ces techniques sont devenues insuffisantes pour l'exploitation de grandes bases de données textuelles (comme les fonds documentaires) [Sch97]. Des domaines variés de la recherche en informatique, tels que le Traitement Automatique du Langage Naturel (TALN) ou l'Aprentissage Automatique (AA), sont venus enrichir les modèles de recherche d'information, ont élargi leur domaine d'application et ont permis à un plus grand nombre d'utilisateurs de les exploiter [Zar99].

Caractéristiques des outils de GED Un système de Gestion Électronique des Documents est un système informatisé de gestion, classement, stockage, archivage, recherche de documents électroniques ou de documents à numériser. Un exemple d'utilisation courant est la numérisation de masse de documents papiers.

Le système de GED de la MIDR, comme la plupart des systèmes utilisant des moteurs de recherche, propose une indexation par notices descriptives et un regroupement des documents par thèmes en plus du système classique de recherche en texte intégral. L'indexation par notices descriptives permet de fournir une interface de requêtage précis à l'utilisateur spécialisé. Elle permet de récupérer de l'information au grain des documents et l'indexation « plein-texte » permet d'accéder à leur contenu. Cependant l'accès ne se fait alors qu'à deux niveaux, celui du document et celui beaucoup plus bas obtenu grâce à la recherche à partir de mots clés du plein-texte. Or il serait intéressant, en particulier pour les œuvres littéraires, de pouvoir proposer un accès fragmenté en plusieurs niveaux, où seulement une partie de l'œuvre, de taille variable selon le type de recherche, est retournée à l'utilisateur.

Cet outil n'offre donc qu'une réponse générique de recherche d'information et ne tient pas compte des spécificités du corpus, comme leur connotation géographique par exemple. De plus il nécessite beaucoup de manipulations manuelles pour l'indexation, notamment pour le remplissage des notices descriptives.

Caractéristiques des outils de SIG Au départ les Systèmes d'Information Géographique sont nés de besoins en gestion d'infrastructures (eau, gaz, électricité, etc.) pour l'aménagement du territoire. Ce sont des systèmes dédiés à la gestion de l'information géographique (sa composante spatiale). Ils viennent de la communauté « base de données » et ont répondu à la demande croissante d'outils exploitant des ressources géographiques (cartes, relevés de données géo-référencées, etc.). Ils ne sont cependant

Besoins \ Domaines	RI	TALN	RSQ	SIG	Classificat°
GED spécialisée	✓				
Interprétation géo.		✓	✓		
Indexation géo.			✓	✓	
Résumé spatial			✓	✓	✓

TAB. 1.2 – Tableau récapitulatif des besoins et des domaines de recherche appropriés.

pas utilisables directement dans un système de recherche documentaire spatial, car ils sont adaptés à des données structurées. Or l'indexation spatiale du contenu d'œuvres littéraires nécessite une gestion de données non-structurées et une gestion du contexte dans lequel sont exprimées les informations spatiales, difficiles à interpréter automatiquement. Pour pallier ce manque, des travaux sur l'indexation spatiale ont proposé des moyens d'utiliser les SIG en y intégrant des fonctions spécifiques au raisonnement qualitatif [Ben96], domaine approprié à la modélisation de l'expression de l'information spatiale en texte libre.

1.3 Synthèse de l'existant

Nous faisons la synthèse des besoins et des domaines existants dans le Tableau. 1.2 pour répondre aux questions de notre problématique. Notre contribution principale, exposée dans la section suivante se focalisera sur l'indexation spatiale de documents textuels tout en couvrant les domaines de Traitement Automatique du Language Naturel (TALN), de Raisonnement Spatial Qualitatif (RSQ) et des Systèmes d'Information Géographique (SIG).

En effet, notre problématique va se trouver à la jonction de plusieurs domaines de recherche qui ont pour point commun de s'intéresser à l'information géographique. Du point de vue du TALN et du RSQ, les problématiques concernent plutôt l'extraction et l'interprétation de cette information (par exemple extraire *« le sud de Pau »* ou *« Au début de XVIIIème siècle »* et leur donner un sens), tandis que du point de vue SIG, les questionnements portent surtout sur sa quantification et son indexation (affecter une représentation géo-référencée au *« sud de Pau »* et le stocker dans une base de données). Enfin le domaine de la classification, sous-domaine de la RI, utilise l'analyse faite dans les domaines précédents pour ranger automatiquement l'information selon son type (par exemple classer un paragraphe comme étant spatial ou non et dans l'affirmative, caractériser l'information spatiale qu'il contient).

Cette thèse se plaçant à l'intersection de plusieurs domaines de recherche, notre étude de l'existant pour chacun d'entre eux n'a pas la prétention d'être exhaustive. Nous avons essayé de restreindre notre discours d'état de l'art sur les éléments susceptibles de servir de socle à notre contribution.

1.4 Énoncé de la problématique

Nous nous plaçons à l'intérieur de la problématique générale qui est la construction de « molécules géographiques » pour les documents texte, tout en nous restreignant à la composante spatiale afin de pousser notre recherche pour valider certaines hypothèses, plutôt que de mener une étude en surface des trois composantes (spatiale, temporelle et thématique). De plus, la composante spatiale guide les deux autres pour définir une entité géographique.

Nous avons néanmoins réfléchi à l'intégration de notre travail et de celui nécessaire pour la composante temporelle et la composante thématique. Dans le contexte de la collaboration avec la Médiathèque, la partie thématique (ou phénomène) peut leur être déléguée (grâce notamment aux notices descriptives), tandis que la composante temporelle (restreinte à la temporalité calendaire) est un problème qui peut être traité selon une démarche similaire à celle mise en œuvre pour la composante spatiale. Des travaux ont débuté par ailleurs dans le laboratoire, se basant sur les principes de notre approche.

Notre interrogation devient alors : comment accéder à l'information spatiale contenue dans un corpus textuel et comment l'interpréter de manière à pouvoir l'utiliser dans un système de RI répondant aux besoins cités plus haut ? Quelles sont les méthodes existantes permettant d'aider à la réalisation d'un tel travail ? Comment unifier ces méthodes afin qu'elles puissent être couplées ?

1.5 Contribution

Hypothèse de départ Nous supposons que, dans le cas de la recherche d'information spatiale, il est nécessaire d'aller au delà des systèmes classiques basés-statistiques. Nous verrons en effet que l'analyse statistique peut ne pas être adaptée pour certains systèmes de recherche d'information spécifique, où la fréquence des termes dans les documents n'est pas proportionnelle à leur importance.

Un système de recherche spécialisé, interprétant le contenu spatial des documents semble plus approprié. Notre intérêt s'est donc porté sur le traitement linguistique automatique de la langue pour réaliser une analyse focalisée sur cette information. Il est à noter que nous ne voulons pas utiliser de traitements trop complexes, encore peu opérationnels dans le cas d'une mise en place dans un système de recherche d'information. Notre hypothèse, qui sera détaillée dans la partie contribution, est que des traitements relativement peu coûteux suffisent à dégager l'essentiel de l'information, c'est-à-dire les syntagmes contenant des entités nommées de lieux. Ils sont un bon point de départ pour une interprétation plus poussée ensuite, s'abstrayant d'une analyse linguistique pure et permettant d'éviter ses écueils de variabilité, de subjectivité et de dépendance du contexte. En effet, notre hypothèse de travail est que l'information spatiale peut être synthétisée sous forme de motif, décrivant un contexte particulier exprimé par l'auteur, qu'un traitement automatique se basant sur l'analyse linguistique préalable peut catégoriser. Nous nous intéressons en particulier au motif d'*itinéraire*, ainsi qu'à celui de *description de point de vue* et de *comparaison de lieux*. De cette manière, une indexa-

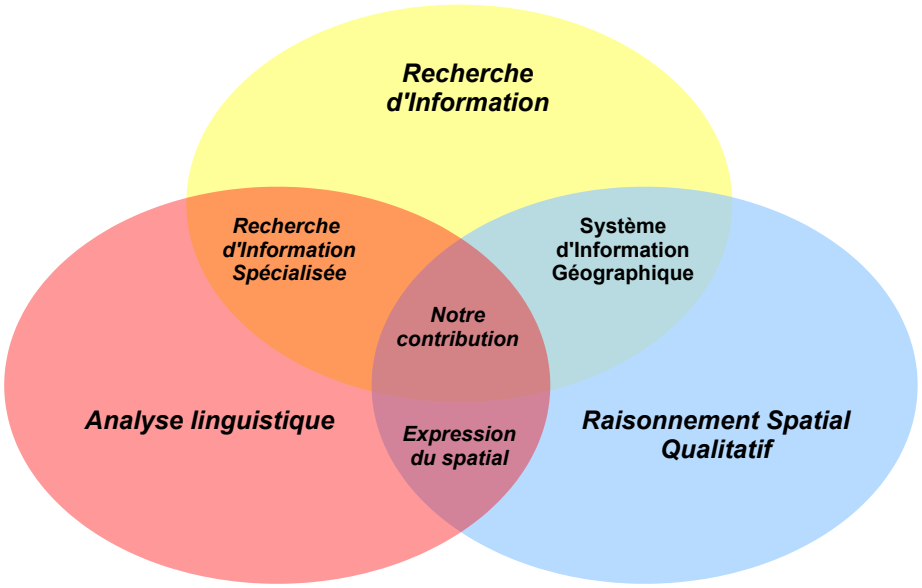


FIG. 1.2 – Intersection des principaux domaines de recherche

tion basée sur ces synthèse permet un accès aux documents à plusieurs niveaux et non plus seulement au niveau du document entier. Un paragraphe ou un chapitre peut être retourné.

Les domaines de recherche privilégiés et notre position sont illustrées en figure 1.2 : la *recherche d'information* en premier lieu qui se couple à *l'analyse linguistique* et au *raisonnement spatial qualitatif* concernant l'expression du spatial dans le texte. Des domaines liés à ces premiers sont aussi rentrés dans le cadre de notre étude comme la *recherche d'information spécialisée* et les *systèmes d'information géographique* qui sont des outils dédiés mais pouvant servir d'une part à exprimer l'information qualitative présente dans les documents et, d'autre part, à stocker des index spatiaux.

Méthodes de travail à privilégier Notre idée est d'isoler les entités spatiales présentes dans les textes, de les interpréter et de les indexer afin de mettre en place un système de RI spatial, basé sur un système d'interrogation spécifique. Ce système pourrait étendre un système existant de recherche documentaire. De plus, dans notre volonté d'interpréter de manière poussée l'information spatiale et de proposer une méthode efficace, utilisable sur un corpus conséquent d'œuvres littéraires, nous proposons de synthétiser l'information qui a été extraite du document. Notre approche privilégiera les trois aspects suivants :

– la méthode d'extraction des entités spatiales consiste à faire une analyse morpho-syntaxique et sémantique,

– l'interprétation et l'indexation se basent sur des travaux de raisonnement spatial qualitatif afin de trouver une représentation géo-référencée servant d'index pour chaque entité spatiale extraite,
– enfin la méthode employée pour la synthèse ou classification spatiale consiste à calculer des caractéristiques topologiques, métriques, etc. qui sont employées ensuite dans un système de classification.

Ces aspects font partie de deux processus globaux, d'indexation et de recherche d'information. Ces processus sont découpés sous forme de services web indépendants regroupés au sein d'une plate-forme ouverte et donc facilement évolutive. Des nouveaux modules sont facilement intégrables, l'idée étant de mutualiser plusieurs travaux se trouvant à plusieurs endroits. Une interface web permet d'appeler ces services et de visualiser les résultats de recherche.

1.6 Organisation du document

Notre mémoire de thèse se compose de deux parties principales. Dans la première partie, un premier chapitre traite de la recherche d'information et des techniques classiques pour la récupération d'information dans du texte. Nous exposons les limitations de ces techniques pour la recherche ciblée sur un domaine spécifique comme le spatial et nous abordons alors les approches linguistiques et cognitives traitant de ce problème. La présentation de ces travaux permet de déterminer les méthodes existantes utilisables dans un outil d'interprétation et d'indexation. Nous détaillons ensuite dans un deuxième chapitre les travaux existants sur l'interprétation et la représentation de motifs spatiaux définis par notre classification. Nous présentons plus particulièrement le contexte de description d'itinéraire. Le chapitre suivant porte sur la manipulation de cette information spatiale une fois qu'elle est interprétée et structurée. Des travaux sur l'indexation sont présentés, *via* l'utilisation de bases de données spécifiques à la gestion d'information géographique, les Systèmes d'Information Géographique. Nous abordons aussi des travaux de recherche relatifs à l'extraction d'information provenant de données semi-structurées.

La deuxième partie de ce mémoire est consacrée à notre contribution. Nous présentons dans le chapitre 5 nos préconisations quant à l'élaboration d'un système de recherche spécialisé dédié à l'information spatiale, basé sur une indexation multi-niveaux. Nous y définissons les principaux modèles utilisés dans ce système. Le chapitre 6 présente un premier prototype développé grâce à ces modèles afin de valider nos hypothèses et de constituer un processus de base d'indexation à un premier niveau, intraphrastique. Ce prototype (le prototype PIV) implémente donc toutes les parties d'une indexation et d'une recherche d'information spatiale. Il a évolué tout au long de la thèse et a connu de nombreuses améliorations, effectuées par des étudiants ou d'autres membres de l'équipe de recherche. Une évaluation de ce prototype est présentée. Le dernier chapitre présente les débuts d'un travail sur l'indexation multi-niveaux basée sur l'indexation au premier niveau et la classification en motifs spatiaux de documents textuels. Nous présentons les caractéristiques définies pour la classification. L'implémentation de ce travail constitue une amélioration potentielle pour le prototype PIV. Nous proposons enfin une première

expérimentation de ce système.

Un chapitre de conclusion fait la synthèse de nos résultats et liste l'ensemble des perspectives. Nous concluons ce mémoire sur les apports théoriques et concrets de notre travail dans le domaine de la RI spécialisée et des outils d'interprétation et d'indexation d'information spatiale qualitative. Nous présentons en quoi le prototype PIV, bâti sur une architecture ouverte et modulaire sous forme de services web, permet d'envisager une plate-forme de recherche pour des travaux à venir dans les domaines de la RI, du TALN, du traitement de requêtes spatiales et de la représentation / visualisation de données résultants d'une requête spatiale.

Deuxième partie

Travaux existants

Chapitre 2

Traitement de l'information dans le texte : du cas général au cas du spatial

Sommaire

2.1 Introduction

Ce chapitre explore les recherches concernant le traitement d'information contenue dans un texte et plus particulièrement l'information spatiale. Cette problématique intéresse de nombreux chercheurs dans divers domaines, notamment en recherche d'information et en traitement automatique de la langue. Plusieurs approches sont décrites, le but de ce chapitre étant de réunir et d'unifier ces connaissances afin d'identifier les

travaux pertinents par rapport au traitement de l'information spatiale contenue dans un texte. Il est à noter que nous avons restreint nos recherches à la manipulation de données non-structurées, c'est-à-dire du texte au format brut.

Les techniques de recherche d'information statistiques sont les approches les plus utilisées, leur objectif étant de faciliter l'accès aux documents les plus pertinents. C'est pourquoi nous y consacrerons la première section de ce chapitre puis verrons quelles sont les solutions proposées pour la recherche d'information spécifique (comme le cas du spatial par exemple). Nous verrons ensuite l'approche linguistique pour l'extraction d'information spatiale, basée sur des concepts cognitifs, dont les méthodes d'analyse spécifique sont intéressantes, donnant des éléments de réponse à notre problématique. Nous présenterons d'abord des travaux venant du domaine cognitif, spécialisés sur les aspects de modélisation de primitives spatiales et de leurs relations. Nous ferons enfin la synthèse de ce chapitre en soulignant les travaux les plus pertinents dans le cadre de notre problématique.

2.2 Traitement de l'information dans la Recherche Documentaire

Après un bref rappel sur les définitions des termes employés en recherche d'information, cette section décrira les méthodes employées pour l'indexation de documents, les modèles de recherche d'information, les méthodes d'évaluation de ces modèles et l'évolution de la recherche en matière de recherche documentaire, et notamment en recherche spécialisée.

2.2.1 Définitions pré-requises

La **recherche documentaire** ou **recherche d'information** [BY99] est traditionnellement définie comme l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptibles de répondre aux besoins de l'utilisateur. Il convient de définir certains de ces concepts inhérents à la recherche d'information.

La **collection de documents** (ou fond documentaire, corpus) est l'ensemble des informations accessible par le système de recherche d'information (SRI). Elle est constituée de **documents**, éléments unitaires.

Cependant la notion de **document** en elle-même est vague. Souvent définie à partir du contenant (par exemple le livre, l'objet physique qui contient le texte), elle varie souvent, dans le domaine de la recherche d'information, où la réponse attendue pour une requête peut ne pas être un livre entier mais un ou plusieurs fragments particulièrement pertinents. C'est d'ailleurs pour cela qu'est utilisé le terme « granule de document » pour définir l'unité de texte renvoyée à l'utilisateur [Baz05]. Nous utiliserons comme convention dans la suite de ce mémoire le terme « document » pour « granule de document ».

Enfin le terme **requête** correspond à l'expression du besoin en information de l'utilisateur. La requête constitue une interface entre l'utilisateur et le système de recherche

d'information. Elle constitue le paramètre d'entrée de ces systèmes et s'exprime dans un langage d'interrogation, qui est souvent basique (à partir d'un choix de mots-clés de la part de l'utilisateur). Cependant d'autres langages sont présentés dans la littérature : langage naturel, langage graphique, etc. [Baz05].

2.2.2 Méthode classique de pondération pour l'indexation

Cette section n'a pas pour vocation de faire un état de l'art exhaustif sur l'indexation de documents en RI. Elle présente seulement les méthodes les plus employées afin de déterminer dans quelle mesure nous pouvons nous en servir dans le cadre de notre problématique.

2.2.2.1 L'analyse statistique

L'approche classique d'analyse statistique de texte vient du domaine de la recherche d'information. Ce domaine spécifique a pour objectif de retrouver le meilleur appariement entre une demande d'utilisateur en quête d'information et une base documentaire susceptible de la contenir. Pour cela une indexation est effectuée, c'est-à-dire un processus visant à repérer des mots ou des expressions particulièrement significatifs (appelés *termes*) et à créer un lien entre ces termes et le texte original.

Plusieurs approches de pondération utilisant différents principes de modélisation existent : modèle booléen, modèle vectoriel, modèle probabiliste. Nous présenterons une de ces approches (l'approche vectorielle) afin d'illustrer leur fonctionnement global.

2.2.2.2 La pondération TF*IDF

L'indexation de texte et toutes les méthodes citées ici sont basées sur l'hypothèse forte que la fréquence d'apparition d'un mot dans un document permet de calculer son « importance » (tout en discriminant les mots apparaissant trop souvent dans tous les documents d'une collection).

Elle consiste ensuite en un « sac de mots » correspondant aux termes identifiés comme significatifs. Ceux-ci peuvent être ensuite lemmatisés[8]. Par exemple, pour le verbe significatif « traversa » le lemme associé « traverser » sera gardé. D'autres techniques existent, comme l'algorithme de Porter, qui propose simplement de tronquer les mots (6 ou 7 caractères).

La pondération en $tf * idf$ utilisée enfin est un calcul courant pour l'indexation d'un document. Elle permet de déterminer les termes statistiquement significatifs dans un corpus et dans les documents qui le composent. tf (*Term Frenquency*) correspond à la pondération locale, c'est à dire l'importance d'un terme dans un document (Équation. 2.1 où n_i est le nombre d'occurences du terme considéré et le dénominateur est la somme

[8]La lemmatisation est un processus de simplification morphologique permettant de regrouper les variantes d'un mot.

des occurences de tous les termes)

$$tf_i = \frac{n_i}{\sum_k n_k} \tag{2.1}$$

et *idf* (*Inverse Document Frequency*) à la pondération globale dans toute la collection, l'idée étant qu'un terme apparaissant dans tous les documents n'est pas pertinent (Équation. 2.2 où D est le nombre total de documents et d le nombre de documents contenant le terme t_i).

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2.2}$$

Ces deux formules souvent combinées forment la pondération $tf * idf$. Ce calcul a fait ses preuves dans la plupart des systèmes de recherche d'information. Il se base sur une loi empirique et une conjecture : la loi de Zipf et la conjecture de Luhn [Luh58] (figure 2.1).

La loi de Zipf est une loi énoncée en 1949 décrivant la répartition statistique des fréquences d'apparition des différents éléments d'un ensemble. À propos du texte, cette loi dicte que les mots dans les documents ne s'organisent pas de manière aléatoire mais suivant une loi inversement proportionnelle à leur rang, le rang d'un mot étant sa position dans la liste décroissante des fréquences des mots du corpus. Vient s'ajouter à cette loi, la conjecture de Luhn qui émet une hypothèse sur l'importance des termes suivant leur fréquence d'apparition (Figure. 2.1). Selon cette conjecture, les termes particulièrement significatifs (informatifs) se trouvent à l'intérieur d'une fourchette, dont le seuil maximum exclut les termes les plus fréquents correspondant à des termes employés dans n'importe quel document textuel. Le seuil minimum exclut les termes les moins fréquents, trop rares pour être réellement discriminants, l'hypothèse étant que le nombre d'occurrences d'un terme significatif est relativement important. Les seuils de cette fourchette dépendent des corpus et le calcul de leurs valeurs nécessite des tatonnements. Des travaux ont cependant été réalisés, notamment par G.Salton [Sal75] pour automatiser ce calcul.

Cette conjecture, reconnue depuis les débuts de la RI reste néanmoins une hypothèse forte sur le degré d'informativité des termes d'un corpus. Il est admis que c'est une bonne heuristique d'expression de l'information. Cependant une recherche particulière sur un sujet, non dominant en matière de fréquence de termes ne trouverait aucun résultat pertinent, l'indexation mettant de côté les termes en dehors de la fourchette.

2.2.3 Modèles de RI

À partir des hypothèses mentionnées dans la section précédente des modèles ont été créés au cours de l'évolution de la RI, se basant sur différentes théories, afin de répondre à la problématique complexe de « retourner les documents les plus pertinents » : modèle booléen (1950), modèle vectoriel (1970) [Sal71b], modèle LSI (Latent Semantic Indexing) (1990) [DDFH90], modèle probabiliste (1976) [Rob77], modèle inférentiel (1992), modèle connexionniste (1989) et modèle de langage (1998) [PC98]. Pour tous ces modèles, chaque document est représenté par un ensemble de termes d'indexation. Ces termes ont une

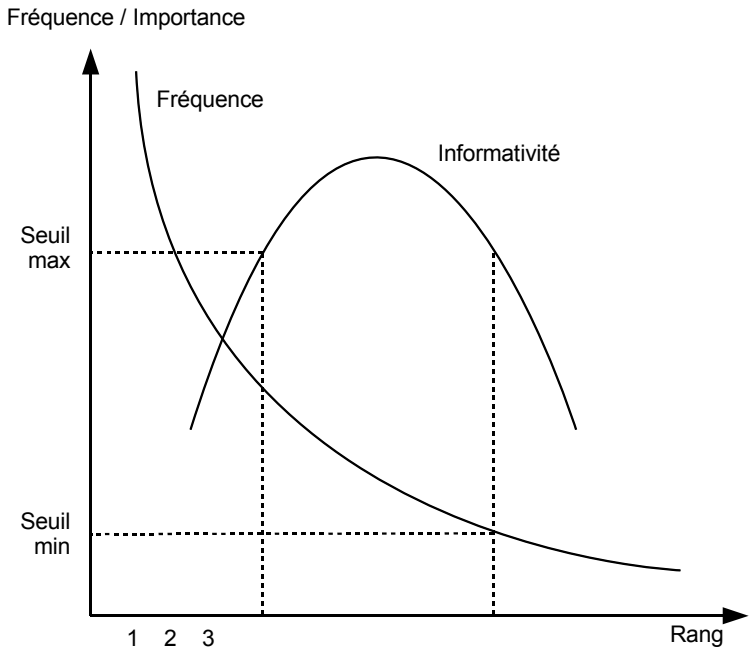
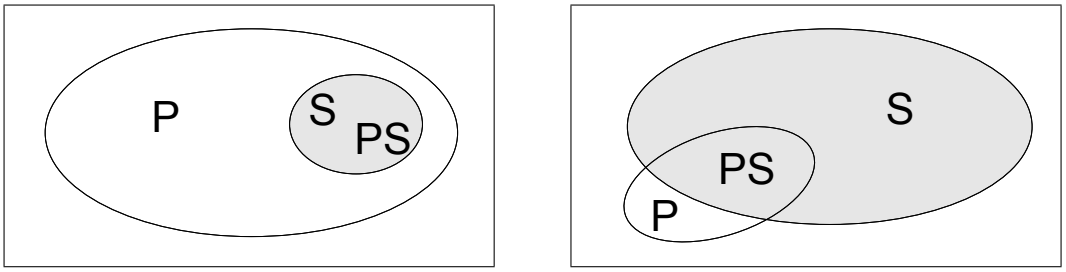


FIG. 2.1 – La conjecture de Luhn.

importance différente suivant le document, importance qui est représentée par un poids. Puis ces poids sont mis en correspondance avec ceux déterminées pour une requête et un calcul de pertinence est effectué, permettant de réaliser un appariement entre cette requête et une sous-partie des documents.

Le premier de ces modèles, le modèle booléen est basé sur la théorie des ensembles. Un document d est constitué d'une groupe de termes $t_i : d = \{t1, t4, t5\}$. Une requête q est une liste de termes avec les opérateurs booléens AND $\vee$, OR $\wedge$ et NOT $\neg$. L'appariement se base sur la présence ou l'absence des termes de la requête dans les documents : $Appariement(q, d) = 1$ *ou* 0. L'inconvénient majeur de ce modèle est que les réponses à une requête ne sont pas ordonnées. De plus la décision de pertinence binaire n'est pas une bonne approximation.

Le modèle le plus utilisé et corrigeant les limitations du booléen est le modèle vectoriel, créé par Salton dans les années 70 pour son fameux SRI nommé SMART (pour System for the Mechanical Analysis and Retrieval of Text) [Sal71a]. Dans cette méthode les documents et les requêtes sont représentés par des vecteurs dans un espace vectoriel, dont les axes sont engendrés par les termes des documents. L'appariement correspond à un calcul de similarité vectorielle entre le vecteur de la requête et ceux des documents. Plusieurs calculs existent, le plus simple étant le produit intérieur (produit scalaire). Ce modèle permet de pondérer les résultats et d'ordonner les résultats de la recherche. Son inconvénient est qu'il suppose l'indépendance entre les termes, ce qui ne reflète pas la



(a) Précision élevé, rappel très faible.

(b) Rappel élevé, précision faible.

FIG. 2.2 – Soit S les documents sélectionnés, P les documents vraiment pertinents et PS les documents pertinents sélectionnés, les figures 2.2(a) et 2.2(b) montrent la complémentarité des deux mesures.

réalité.

2.2.4 Évaluation des SRI

Très tôt une évaluation des différentes méthodes de SRI s'est avérée nécessaire. Plusieurs critères ont été défini par C.Cleverdon [Cle63] :

- La facilité d'utilisation du système,
- Le coût accès / stockage,
- La présentation des résultats,
- La capacité du système à sélectionner des documents pertinents.

À cette occasion C.Cleverdon a proposé deux mesures aujourd'hui reconnues et utilisées dans la plupart des évaluations de méthodes de recherche d'information : le rappel R et la précision P :

$$R = \frac{Nombre\ de\ documents\ pertinents\ sélectionnés}{Nombre\ de\ documents\ pertinents} = \frac{PS}{P}$$

$$P = \frac{Nombre\ de\ documents\ pertinents\ sélectionnés}{Nombre\ de\ documents\ sélectionnés} = \frac{PS}{S}$$

Le rappel évalue la capacité d'un système à sélectionner tous les documents pertinents de la collection tandis que la précision évalue la capacité d'un système à ne sélectionner que des documents pertinents. Ces 2 mesures sont nécessaires car elles sont complémentaires. La figure 2.2(a) montre qu'un résultat peut avoir une précision élevée mais un rappel faible si le système est trop restrictif et filtre trop de documents. On parle alors de silence pour les documents non récupérés. La figure 2.2(b) montre a contrario un rappel élevée et une précision faible pour un système trop large dans sa méthode de récupération de documents. Les documents non-pertinents retournés par un tel système constituent le bruit.

D'autres mesures ont ensuite été proposées[9] se basant toutes sur ces 2 premières. Par exemple la MAP (Mean Average Precision, c'est-à-dire précision moyenne) est employée

[9] http://en.wikipedia.org/wiki/Information_retrieval

pour mesurer un système de RI à partir d'un jeu de requêtes. Elle correspond à la moyenne des précisions calculées pour chacune de ces requêtes. D'autres mesures comme la précision à 5, à 10, à 20, etc. calculent une précision pour les 5, 10 ou 20 premiers résultats retournés.

L'élaboration de mesures d'évaluation pour les SRI est même devenue une action de recherche à part entière de ce domaine. Plus particulièrement, l'évaluation des systèmes de recherche d'information s'articule autour de la pertinence, notion qui s'avère complexe tant par sa modélisation (est-ce plutôt un calcul de similarité vectorielle, un degré de probabilité ?) que par son évaluation (doit-on évaluer un SRI sur le premier résultat qu'il donne ? les 5 premiers ? Quelle technique d'évaluation utiliser, comme la MAP, la F-mesure ($F = 2.(précision.rappel)/(précision+rappel)$) ? Doit-on privilégier les systèmes plutôt booléens qui retournent seulement des résultats à partir de calculs statistiques sur les documents ou plutôt souples (calculant la pertinence aussi à partir d'autres facteurs, comme par exemple la technique de PageRank de Google où les liens externes pointant une page web sont comptabilisés pour celle-ci et servent à déterminer sa pertinence), ces systèmes privilégiant la sérendipidité, c'est-à-dire le fait de faire des découvertes « heureuses[10] ».

2.2.5 Évolution des techniques de RI

Les questions majeures de Recherche d'Information restent encore ouvertes et les campagnes d'évaluation telles que celles réalisées dans TREC[11] montrent que les méthodes proposées répondent encore insuffisamment ou partiellement aux attentes (voir [HCTH99] par exemple). M.Lew [LSDJ06] fait un état de l'art des résultats en recherche d'information multimédia et traite des nouveaux challenges qui attendent ce domaine. Des méthodes récentes essayent d'améliorer les méthodes existantes citées plus haut. L'approche par modèle de langue, par exemple, est une approche prometteuse qui se base sur les modèles probabilistes. Son principe général est de définir pour un corpus, un « modèle » pour la langue utilisée [PC98, SC99], c'est à dire un modèle définissant les probabilités d'apparition des termes. En effet, un document dans un corpus est considéré comme un échantillon de langue et par entraînement, le modèle est construit. La pertinence se fait alors en calculant la probabilité que la phrase de la requête puisse être générée par le modèle de langue du corpus. Il existe aussi de plus en plus de méthodes hybrides faisant appels à diverses techniques venant d'autres domaines de recherche.

Nous allons voir dans cette section les dernières évolutions en Recherche d'Information concernant ces méthodes hybrides Nous verrons enfin une sous-partie particulière, la RI spécialisée, dont les techniques s'adaptent à un corpus spécifique.

2.2.5.1 Vers des méthodes hybrides

Les techniques décrites dans les sections précédentes sont la base de nombreux travaux plus récents, visant à corriger ses limitations par l'ajout d'autres technologies. Les

[10] http://savoirscdi.cndp.fr/culturepro/actualisation/Serres/Serres.htm
[11] http://trec.nist.gov/

principales évolutions ont pour objectif d'enrichir l'analyse essentiellement statistique par une méthode contextuelle, conceptuelle ou linguistique, l'objectif principal étant d'améliorer la précision des SRI.

Approche contextuelle La méthode contextuelle consiste à prendre en compte un contexte, que ce soit celui de l'utilisateur ou celui de la requête. Pour le contexte utilisateur, la méthode doit modéliser par exemple ses centres d'intérêt, ses préférences de recherche, tandis que pour la requête, elle doit permettre d'évaluer sa clarté, c'est-à-dire le degré d'ambiguïté par rapport aux centres d'intérêt de l'utilisateur. E.Agichtein [ABDR06] propose sur ces bases un système qui tente de prédire le comportement des utilisateurs en évaluant leurs préférences lors de recherches sur le web. J.C.Bottraud [BBB03] propose, quant à lui, une méthode de réinjection de pertinence en analysant les références documentaires rassemblées par l'utilisateur. De cette manière, un filtre est réalisé pour chaque requête qu'il pose.

Des travaux avancent l'hypothèse que la connaissance du domaine d'intérêt de l'utilisateur lors de sa recherche doit améliorer la performance du SRI si celui-ci a intégré des outils permettant d'utiliser cette information supplémentaire [Zar04]. Ces outils consistent en général en un ensemble structuré d'informations comme les ontologies ou les thesaurus.

Approche conceptuelle L'approche conceptuelle tente de récupérer une sémantique pour les documents indexés à l'aide de ressources hiérarchisées externes au documents [BBPP06, WBH+00]. Ces ressources sont nommées *thesaurus* ou *ontologies* selon le degré d'abstraction des concepts qu'elles stockent. On peut définir un thesaurus comme un dictionnaire dans lequel les relations que possèdent les termes entre eux ont plus d'importance que les définitions de ces termes. Une ontologie a plus ou moins les même propriétés à la différence que les termes sont remplacés par des concepts. De plus, elle peut être représentée à l'aide d'un graphe où les nœuds sont les concepts et les arcs des relations qui peuvent être d'ordre sémantique ou de composition et d'héritage.

Ces méthodes cherchent les termes significatifs des documents en tenant compte de leur position dans une hiérarchie de concepts. De fait, l'index contient non seulement le terme significatif mais aussi sa position, ce qui permet au moment de la recherche d'information de tenir compte des relations et des termes reliés.

Par exemple, le terme *maison* peut être rattachée dans une ontologie au terme *habitation* par une relation d'héritage (ou de généralisation). De même, le terme *ferme* peut lui être rattaché indirectement car c'est aussi une spécialisation d'*habitation* (figure 2.3). De cette manière une requête comme *la maison de la nourrice d'Henri IV de Bilhère* pourra renvoyer des résultats contenant non seulement le terme *maison* mais aussi des résultats avec le terme *habitation* voire *ferme*. Les implémentations de ce genre d'approche [Baz05, BBAG05] consistent à mesurer la possibilité d'intégrer un sous arbre de concepts d'une requête dans un arbre plus conséquent d'un document. Si les noeuds du graphe sont les mêmes et utilisés avec les mêmes relations, le document sera considéré comme pertinent.

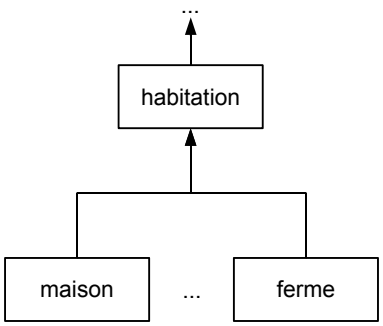


FIG. 2.3 – Exemple de relation d'héritage dans une ontologie.

Cependant, l'interprétation sémantique réalisée par de telles méthodes se limite à une représentation structurée et simplifiée du contenu des documents, n'utilisant pas le fait qu'ils sont exprimés dans une langue qui obéit à des règles. Celles-ci sont pourtant porteuses d'information sémantique qu'il serait intéressant de capturer. De plus, les ressources elles-mêmes sont forcément limitées et le plus souvent elles doivent être adaptées à un domaine particulier. Enfin il peut s'avérer relativement coûteux d'utiliser ce genre de ressources potentiellement de grande taille. Wordnet[12], un des thesaurus les plus connus et utilisés comportait par exemple en 2006, 150 000 termes regroupés en 115 000 *synsets* (synonym set, c'est-à-dire des regroupements de termes ayant le *même sens*). Une approche purement conceptuelle n'est donc envisageable que dans des cas où les domaines des corpus sont clairement identifiés et ou des ressources hiérarchiques exhaustives existent. Dans le cas d'un domaine générique, l'utilisation de ressources conséquentes telles que Wordnet sont adaptées mais alourdissent le traitement de ces systèmes.

Approche linguistique Cette approche tente de faire un pas de plus dans l'interprétation sémantique des documents. Les travaux d'analyse linguistique qui étaient au départ une fin en soi, ont trouvé au fil du temps des applications dans le monde de la Recherche d'Information. Il est de plus en plus courant de voir des techniques d'analyse linguistique associées à des techniques d'analyse statistique [DLC04]. Par exemple, la détection d'entités nommées dans un texte utilise des processus d'analyse linguistique morpho-syntaxique [MMG99]. Peu à peu les outils du Traitement Automatique de la Langue (TALN) sont venus apporter une analyse fine basée sur l'interprétation de la sémantique contenue dans les documents textuels. Des conférences telles que TREC, MUC, NAACL et SIGIR ont produit des recherches dans ce sens. Ces techniques permettent de palier le manque de structure de l'information recherchée et de la requête en les analysant de manière à leur rendre une pseudo-structure basée sur la sémantique et cela sans ressources externes. Parmi les applications émergentes, les travaux donnant le plus de résultats sont les résumés de texte et les systèmes de questions-réponses [Naz04].

[12] http://wordnet.princeton.edu/

La thèse de B.Sagot [Sag06] fait un état de l'art en matière de moyens existants pour analyser automatiquement la sémantique de la langue française. Les formalismes, les lexiques et les analyseurs y sont décrits dans le cas d'une analyse générale. Nous nous intéressons particulièrement à un schéma de traitement assez utilisé, la chaîne de traitement définie par M. Abolhassani [AFG03], qui est composée des sous-processus d'analyse et d'extraction d'information suivants :

- La lemmatisation et l'analyse morphologique découpe les termes d'un texte et retrouve leurs lemmes, c'est-à-dire des versions simplifiées et plus petites des mots. Par exemple la phrase *Je cueille des tomates à la ferme* peut être découpée en *Je (Je) , cueille (cueillir) , des (des) , tomates (tomate,) à (à), la (la), ferme (ferme).*
- L'analyse syntaxique et l'analyse des dépendances grammaticales s'intéressent à la structure des phrases et des syntagmes (groupes de termes formant des unités syntaxiques [RPR99]). Elle retrouve les liens (grammaticaux) entre les termes d'une phrase. Pour la même phrase on a *Je (pronom personnel sujet), cueille (verbe 1ère personne du singulier), des (article contracté pour de les), tomates (nom commun complément d'objet direct), à (préposition), la (article), ferme (nom commun complément circonstanciel de lieu).*
- L'analyse sémantique, souvent adaptée pour un objectif précis, interprète le texte du point de vue de cet objectif. Si l'on s'intéresse aux activités du locuteur toujours dans le même exemple, ce genre d'analyse permettra d'identifier que des tomates et le locuteur lui-même se trouvent dans la ferme.

Ces différents processus peuvent intervenir à plusieurs étapes des processus de RI. L'analyse morphologique permet de faire l'appariement sur les lemmes plutôt que sur les termes de départ, permettant d'enrichir le contenu des requêtes et l'indexation des documents et d'améliorer l'appariement. Par exemple un document contenant le terme « intègrent » pourra être retourné pour une requête avec le terme « intégration ».

L'analyse syntaxique permet de connaître le rôle des termes ou des syntagmes dans la phrase. Un exemple simple montrant l'intérêt de ce genre d'approche est le terme *ferme*, qui peut aussi correspondre au verbe conjugué *fermer*. Seule l'analyse syntaxique permet de désambiguiser cet exemple.

L'analyse sémantique quant à elle permet de s'abstraire des termes pour s'intéresser à la sémantique, c'est-à-dire au sens potentiel qu'elles véhiculent. Les ontologies peuvent alors être ré-introduites ici et permettre de fournir, avec l'aide des phases d'analyse précédentes, une interprétation du sens des mots. Par exemple, il est possible d'interpréter le terme *Vienne* comme un nom propre (grâce à l'analyse morpho-syntaxique) et plus précisément comme une ville (grâce à un thesaurus ou une ontologie) dans la phrase « Je suis allé à Vienne » et non pas comme une forme conjugué du verbe venir. Elle permet aussi par exemple de gérer la synonymie des termes : pour une requête comportant le terme *vélo* pourront être retournés des documents contenant le terme *bicyclette*.

Le TALN est donc maintenant un outil relativement utilisé dans les systèmes de RI. À ce propos, la thèse de F.Moreau [Mor06] apporte une réflexion intéressante sur l'apport des techniques du TAL dans la Recherche d'Information. Plusieurs travaux sur la RI utilisant la linguistique ont donné des résultats. Par exemple, ceux de F.Bilhaut [Bil06]

proposent une analyse de détection automatique de la structure du discours. Un autre exemple, S E.Michos [MFK99], utilise un traitement linguistique pour déterminer le « style de l'auteur », afin d'améliorer les résultats lors d'une recherche d'information contextuelle.

2.2.5.2 La RI spécialisée

La RI spécialisée a pour principe de restreindre sa problématique de recherche sur un corpus spécifique, c'est-à-dire composé de documents traitant d'un domaine particulier. Le fait de parler d'un domaine particulier implique une utilisation particulière de la langue et un vocabulaire propre, qui forme une langue dite « spécialisée ». La question posée par P.Lerat [Ler95] est de savoir dans quelle mesure on peut parler d'un *français* de la médecine, du droit, de l'audiovisuel ou de n'importe quelle spécialité possédant un vocabulaire propre. En effet, d'après lui, la spécificité des textes spécialisés tient pour une large part à leur terminologie, c'est-à-dire à l'expression des connaissances, mais aussi aux moyens linguistiques employés pour en faire le transfert. La linguistique apporte donc un plus à la compréhension des langues dites « spécialisées ».

Le TALN peut alors apporter une solution opérationnelle pour la RI spécialisée. À ce propos, A.Nazarenko [Naz04] traite dans sa thèse de l'intérêt du TAL dans ce type de systèmes et du fait que les modèles de recherche utilisés habituellement (du type modèle vectoriel) tendent à écraser la diversité des phénomènes. Il cite à ce sujet plusieurs projets alliant le traitement sémantique et statistique pour la RI spécialisée : Alvis[13] et Extra-PloDocs[14]. Alvis [BVT05] est un projet utilisant les méta-données (de type ontologiques) adaptées à un thème pour manipuler la sémantique des documents. Sa particularité est que l'architecture est basée sur le protocole peer-to-peer, devant faciliter l'intégration d'outils existants de traitements linguistiques ou de ressources web. Le projet Extra-PloDocs est spécialisée dans les corpus scientifiques (notamment en génomique[15]). Il utilise aussi une ontologie pour ce domaine. Ces projets utilisent donc le TALN et des ressources ontologiques pour répondre aux besoins spécifiques d'une problématique de recherche spécialisée.

Enfin la RI spécialisée, comme les autres types de RI, devrait tenir compte du contexte. À ce propos, A.Condamines [Con06] traite de l'interprétation de corpus spécialisé et du fait que le traitement linguistique ne doit pas être seulement introspectif. En d'autre termes, il faut considérer le locuteur et le contexte de production du texte en plus du texte lui-même. Une lecture active du document textuel, c'est-à-dire une recherche ciblée sur des éléments d'information attendus dans le texte et dépendants d'un contexte, semble donc être le meilleur moyen de reconstruire un sens du discours.

[13] http://project.alvis.info/
[14] http://www-lipn.univ-paris13.fr/~poibeau/Extra/
[15] étude du génome.

2.2.6 Cas du spatial en RI

Nous venons de décrire les méthodes les plus souvent utilisées dans le cadre d'une recherche spécialisée. Il existe plus particulièrement des travaux traitant de la problématique de la recherche d'information spatiale. Ces travaux allient les statistiques à d'autres techniques (notamment linguistiques et conceptuelles) afin de répondre à des problématiques spécifiques à ce type d'information [JPR+02, FJA05a, WP94, SVV01, LF04, BE05]. La méthodologie employée est de type recherche active. Des patrons, c'est-à-dire des outils permettant de filtrer dans un ensemble de termes des éléments répondant à des critères prédéterminés, sont définis pour caractériser l'information spatiale et sont cherchés dans les documents.

Le projet Gipsy [WP94] propose par exemple une méthode d'indexation de documents textuels basée sur l'agrégation des géo-références correspondants aux entités spatiales trouvées dans le texte. L'idée est d'utiliser cette agrégation pour retrouver la zone géographique la plus représentative, qui servira à indexer le document.

Le projet SPIRIT[16] [JPR+02] est un projet plus important d'extraction de localisations géographiques dans des pages web présentant, par exemple, des hôtels ou des restaurants. Un de leurs résultats est par exemple la modélisation de patron pour une adresse postale (de type *rue, ville, code postal, pays*). En effet, la problématique de ce projet est l'accès à l'information géographique sur le web et la constitution de système de recherche d'information spatiale. Leurs apports majeurs sont une ontologie géographique (figure 2.4), une réflexion sur le classement de pertinence géographique de documents web, une interface multi-modale spécifique et une méthode d'enrichissement des meta-données géographiques [JPR+02]. L'ontologie définit une entité géographique *(Geographical Feature)* liée à un nom unique d'entité et à une ou plusieurs variantes de dénomination et à un type géométrique (pour sa représentation) et géographique (de type *hôtel*, *restaurant*, etc.). De plus, une entité géographique peut être reliée à elle-même *via* une ou plusieurs relations spatiales. Ces relations sont utilisées au niveau de la requête et servent à définir l'entité spatiale selon qu'elle est à l'intérieur, à l'extérieur ou « près » de l'entité nommée [JAF+04].

L'ontologie sert à la fois à définir le vocabulaire et la structure spatiale des entités dans un but de recherche de l'information. Elle suit les préconisations de travaux existants [HFZ99, Har97] selon lesquelles la combinaison d'information qualitative et quantitative permet d'améliorer la recherche d'information. Cette proposition étend les fonctionnalités des *gazetteers* grâce à la modélisation de l'information qualitative. Elle permet aussi de fournir un système d'appariement spatial doté de mesures de similarité géographique. Dans ce rapport notamment [vKRAvZ04], les auteurs proposent une système classant les documents retournés à partir d'une combinaison d'un classement spatial et d'un classement thématique de la requête. Cependant le prototype développé par le projet SPIRIT n'interprète l'information qualitative qu'au niveau de la requête. Ceci est suffisant car au niveau de l'indexation, l'information spatiale correspond tou-

[16] Spatially-Aware Information Retrieval on the Internet : http://www.geo-spirit.org/index.html. Ce projet est à différencier du système SPIRIT d'exploration de données textuelles [Flu94] conçu par Christian Fluhr du Commissariat à l'Énergie Atomique (CEA) acquis par la COGEMA en 1995.

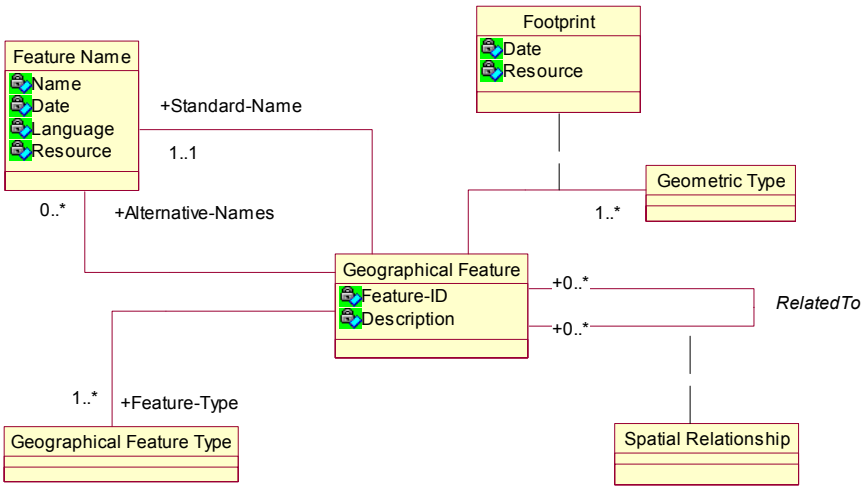


FIG. 2.4 – Schéma de l'ontologie géographique du projet SPIRIT ([SAJ04]).

jours à des adresses exactes d'hôtels ou de restaurants. De plus, l'information qualitative n'est prise en compte qu'au moyen d'une liste de choix fournie à l'utilisateur *(nord, sud, près, etc.)*.

Le projet plus récent, nommé GeoSem [BDEH07], créé dans le cadre de l'appel à projet CNRS, auquel nous avons participé *via* une collaboration avec plusieurs équipes de recherche[17], a repris cette problématique générale d'accès à l'information géographique dans des documents texte. Une méthode d'indexation générique a été proposée, basée sur une structuration de la sémantique par traits, afin de s'adapter aux critères choisis. Les index sont donc constitués d'une telle structure par critère d'indexation choisi. Des travaux [BE05] ont été réalisés en particulier autour de l'indexation de textes à connotation géographique et de la délimitation des cadres du discours. Un prototype de recherche d'information multi-critères (spatiaux, temporels, thématiques) a été développé à partir de ce modèle. La volonté de généricité de cette indexation a l'inconvénient de ne pas être optimale. En particulier pour l'information spatiale, les index sont stockés sous forme de flux XML et l'appariement se fait sans l'aide d'outils dédiés comme les Systèmes d'Information Géographique (SIG). Ils utilisent d'une part un simple calcul de distance sur des coordonnées (de type *point*) et un calcul utilisant le découpage administratif (commune, canton, département, région).

Le tableau 8.1 fait la synthèse des caractéristiques pour ces projets : comment sont définies les entités spatiales, comment sont structurés les index, le type de requête prévu et quelles sont les composantes prises en compte (spatiale, temporelle, thématique).

Le principe général de ces projets consiste donc à extraire les entités nommées du texte et à les identifier [CSJ04]. Avant l'identification, une étape préalable de désambiguisation doit souvent être opérée. De nombreux travaux dans ce sens, nommés WSD

[17] http://infodoc.unicaen.fr/geosem/

Projets	entités spatiales	index	requêtes	espace	temps	thème
GIPSY	simples avec quelques indications	1 géométrie par doc	∅	✓	∅	∅
Spirit	de type adresses	1 géométrie par page web	entités nommées avec choix d'indication qualitative	✓	∅	✓
GeoSem	complexes	un trait sémantique par entité	formulaire à 3 champs	✓	✓	✓

TAB. 2.1 – Comparaison des projets existants étudiés.

(pour *Word Sense Desambiguisation*[18]) ont été menés dans un cadre général [DEG+03]. Le moyen le plus direct d'identifier les entités nommées est d'utiliser des listes de noms générées au préalable [Pro01]. Cette méthode ne gère cependant pas les problèmes d'ambiguïté, contrairement à des travaux plus récents tentant d'intégrer le contexte dans lequel est exprimé l'information, grâce à des grammaires (bâties à la main ou de manière automatique en utilisant des processus d'apprentissage (*Machine Learning*), ou encore en utilisant des ressources ontologiques (permettant d'identifier une entité dans son contexte). Les outils LaSIE [Gai02], ANNIE (qui l'étend) proposés dans le cadre du système GATE [CMBT02] sont des systèmes proposant de telles solutions. Le projet SPIRIT, spécifique au spatial, utilise par exemple GATE et en particulier ANNIE pour l'extraction d'entités nommées.

L'indexation spatiale, qui est une indexation particulière, doit produire à partir de ces entités extraites non pas des index de termes significatifs mais des index spatiaux, c'est-à-dire contenant pour chacune d'entre elles des géo-références permettant de les identifier *sur une carte*. Nous détaillerons l'utilisation de ces géo-références dans le chapitre 4.

Suivant les usages prévus, une analyse plus ou moins fine des relations spatiales exprimées dans le texte est réalisée. Nous avons vu que les ontologies répondent à ce problème en modélisant par exemple les différentes relations topologiques possibles. Elles peuvent aussi servir à la désambiguisation et à la mise en correspondance des entités, ainsi qu'à l'expansion de requête. À titre d'exemple, F.Fu dans l'article [FJA05a] propose une ontologie (« geo-ontology ») qui joue plusieurs rôles au cours du processus d'indexation et de recherche d'information du projet SPIRIT :

1. Pour la partie d'extraction des meta-données, elle permet de faire un marquage géographique et de faire l'indexation spatiale,

[18] http://en.wikipedia.org/wiki/WSD

2. pour la partie de classement des résultats, elle fournit une mesure de distance
3. et pour la partie interface utilisateur, elle permet de désambiguiser et de faire de l'expansion de requête.

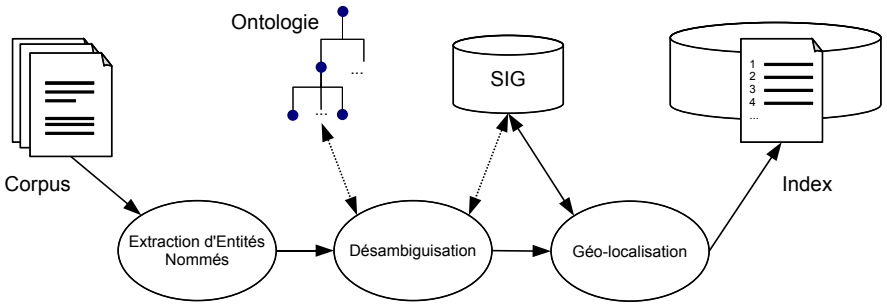


FIG. 2.5 – Chaîne de traitement type pour une indexation spatiale.

La figure 2.5 résume les phases de traitement nécessaires à la création d'un index spatial : d'abord le traitement sémantique appliqué au corpus afin d'extraire les entités nommées spatiales, puis la phase de désambiguisation qui peut utiliser soit des ressources ontologiques soit des ressources géographiques contenues soit dans des Systèmes d'Information Géographique (SIG) ou des dictionnaires géographiques de type *gazetteers* [SVV01, HGJ04, Hil04] pour déterminer de manière unique les entités citées et, enfin, la phase de géo-localisation qui elle aussi peut utiliser des SIG ou des gazetteers[19] pour produire un index spatial, contenant pour chaque entité une empreinte géo-référencée. Cette partie sera détaillée dans le 4ème chapitre présentant les travaux existants.

2.3 Analyses linguistiques et cognitives pour l'information spatiale

L'étude linguistique a pour vocation d'étudier les langues humaines. Elle recouvre des problématiques extrêmement vastes : l'étude descriptive permettant de dégager des structures, des constantes universelles (étude théorique), l'étude d'une langue à un moment donné de son histoire (synchronie) ou l'étude de son évolution dans le temps (diachronie), l'étude du langage de manière individuelle ou l'étude de son interaction avec le monde extérieur (étude contextuelle).

Nous nous intéressons dans la suite de ce document à l'approche restreinte de la linguistique théorique, nommée « sémantique », qui s'intéresse à l'interprétation du sens des mots. Ce terme, inventé à la fin du XIX^{e} siècle par le linguiste français Michel Bréal, auteur du premier traité de sémantique [20], fut plus tard employé dans le domaine infor-

[19] Index ou dictionnaire géographique, regroupant les noms, les caractéristiques et le géo-référencement des entités nommées du monde.

[20] http://fr.wikipedia.org/wiki/Michel_Br%C3%A9al

matique pour désigner l'étude de la signification des programmes vus en tant qu'objets mathématiques. Contrairement à la sémantique des langages de programmation, qui est définie de manière formelle, la sémantique de la langue naturelle repose sur des bases plus fluctuantes. La méthodologie d'analyse est inverse : dans le cas de l'étude linguistique, elle ne sert pas à définir mais à essayer de décrire un langage existant.

Nous allons voir quels sont les différents points de vue existants employés pour essayer de récupérer le sens du texte, non pas dans son ensemble mais pour une sous-partie définie au préalable. En effet, si l'on se restreint à l'information spatiale et à l'expression de celle-ci dans le langage naturel, il devient possible de mettre en place une étude descriptive et interprétative, dans la mesure où des patrons relativement simples à construire peuvent récupérer cette information. Des travaux en ce sens ont été menés pour le français [Den97, AVB97] et pour l'anglais avec la géographie naïve d'Egenhofer [EM95].

Seront ensuite présentés les principes et les méthodologies d'utilisation du raisonnement spatial qualitatif. Nous verrons enfin une théorie intéressante de l'expression de la localisation spatiale, qui sera une base pour construire notre contribution.

2.3.1 Des éléments du discours à leur interprétation dans un raisonnement spatial qualitatif

L'expression de la spatialité dans le langage humain est un domaine de recherche qui est exploré depuis longtemps par les linguistes et cognitiens. Le livre de M.Denis [Den97] fait le tour de la question spatiale dans le texte. Plusieurs points de vue y sont présentés.

Le point de vue linguistique s'intéresse à l'expression de la spatialité par le langage humain. Les facteurs syntaxiques, sémantiques et pragmatiques y sont étudiés. En effet, toutes les langues naturelles ont développé des moyens d'exprimer la référence spatiale, que ce soient des prépositions (« dans », « sur », « vers »), des verbes (« venir », « se jeter ») ou des expressions adverbiales (« ici », « à droite », « là-bas »). Les travaux résultant de ce point de vue sont des méthodes d'analyse morpho-syntaxique dont le but est d'extraire des patrons en suivant des règles syntaxiques spécifiques.

Les problématiques soulevées par ce domaine de recherche sont la polysémie des prépositions locatives et le rôle du contexte dans leur interprétation.

La psychologie cognitive s'intéresse aux mécanismes mis en œuvre dans la description et la production de descriptions spatiales. Les problématiques étudiés ici portent sur les modèles spatiaux cognitifs construits par les individus lors de la production de texte descriptif. Par exemple, il existe des travaux sur la description et la production d'itinéraires [HMW93, Fra99, FL99], qui seront détaillés dans le chapitre suivant.

Les approches formelles apportent ensuite des modèles axiomatiques ou logiques pour décrire les relations spatiales. L'étude formelle faite de la sémantique des marqueurs spatiaux permet de dégager des propriétés fondamentales pour la représentation conceptuelle de l'espace.

Nous allons présenter en détail ce point de vue car il s'insère dans notre problématique.

2.3.1.1 Approches formelles pour l'interprétation d'information spatiale qualitative

Cette axe de recherche, proche du domaine de l'intelligence artificielle en informatique, porte le nom de Raisonnement Spatial Qualitatif (RSQ). Ces travaux s'éloignent de l'étude de la langue pour s'intéresser à la cognition, c'est-à-dire à la manière de rendre explicite le savoir, du point de vue du sens commun et sans utiliser la géométrie euclidienne, sur le monde physique qui nous entoure.

Ils ont pour origine des travaux effectués par J.F.Allen [All91] sur le raisonnement temporel pour sa représentation qualitative. Des propositions pour le raisonnement spatial qualitatif ont adapté ces travaux en tenant compte des spécificités et de la plus grande complexité de l'information spatiale. Le principe général pour définir un tel formalisme est défini dans [Cha05] et procède en deux étapes :

- il faut fixer des entités spatiales basiques (par exemple des points du plan ou de l'espace, des régions, etc.). Cette définition est souvent nommée *ontologie* [CH01]. Ce concept, défini au départ par Stanislaw Lesniewski[21] et faisant parti d'un système de logique générale, correspond dans le cas du RSQ à la définition de ce qui est représenté, c'est-à-dire la primitive spatiale.
- Il faut ensuite définir des relations atomiques qualitatives qui sont susceptibles de relier ces entités spatiales entre elles. Ces relations sont principalement d'ordre directionnel ou topologique. D'autres relations peuvent apparaître, comme celles portant sur la taille des objets, la distance qui les sépare ou tout simplement leur forme. A.G.Cohn [Coh96,CH01] fait un état de l'art du RSQ et en particulier classe les relations spatiales qui peuvent être définies :
  - **Topologie :** elle décrit les propriétés de connectivité entre les primitives spatiales (voir les travaux présentés plus loin sur le RCC8 (figure 2.6) ou le modèle d'Egenhofer).
  - **Méréologie :** Comme l'ontologie, la méréologie est un concept définie au départ par Lesniewski. Elle décrit les relations de tout et de partie sur les primitives spatiales. Elle est souvent combinée avec la topologie pour construire des théories axiomatiques définissant la connexion, la partie tangentielle, la partie interne, etc.
  - **Orientation :** l'orientation dans l'espace d'un objet ; elle consiste à décrire la manière dont l'objet est disposé dans l'espace, à l'aide par exemple de deux vecteurs non parallèles portés par cet objet .
  - **Distance et Taille :** « A est à x unités de B », « A est de taille x unités ». Elle peut être représentée à une échelle relative ou absolue.
  - **Forme :** doit modéliser la forme des primitives définies. Cet aspect est difficile à définir de manière qualitative. Plusieurs solutions sont apportées :

[21] http://fr.wikipedia.org/wiki/Stanis%C5%82aw_Le%C5%9Bniewski

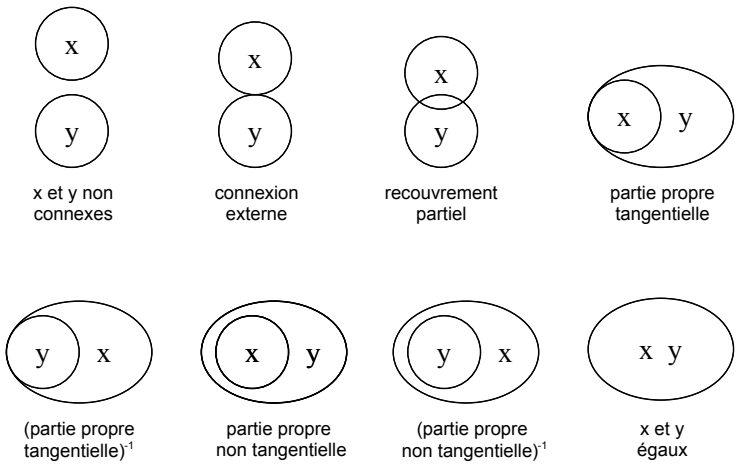


FIG. 2.6 – Les 8 relations topologiques pouvant exister entre 2 régions x et y selon le modèle RCC-8 [RCC92].

  - la description à partir de formes primitives (des cercles, des rectangles ou des sphères, des parallélépipèdes, etc.),
  - la description des contours,
  - le concept de zone convexe ou *convex hull*, correspondant à la zone de taille minimale contenant l'objet.
- **Incertitude :** elle décrit l'imprécision soit dans les régions soit dans les relations entre les régions.
- **Granularité :** elle décrit à quelle échelle sont définies les primitives spatiales. Cette relation est liée aux concepts de flou et d'incertitude. les primitives spatiales définies.

2.3.1.2 Modèles existants de raisonnement spatial qualitatif

Les modèles les plus connus, définissant en premier lieu la topologie des relations spatiales, sont ceux de Randell et Cohn [RCC92, Coh96, CBGG97, Coh97, CH01], dont la théorie nommée RCC-8 (pour *Region Connection Calculus*) est illustrée dans la figure 2.6 (les différentes relations définies) et celui d'Egenhoger [Ege91, EF91].

Le premier modèle est basé sur une axiomatisation utilisant la relation de connexion $C(x, y)$. Les 8 relations définies sont (figure 2.6) :

- $DC(x, y)$: x est déconnecté de y,
- $EC(x, y)$: x est connexe à y,
- $PO(x, y)$: x recouvre partiellement y,
- $TPP(x, y)$: x est une partie propre tangentielle de y,
- $NTPP(x, y)$: x est une partie propre non tangentielle de y,
- $TPP^{-1}(x, y)$: x possède pour partie propre tangentielle y,

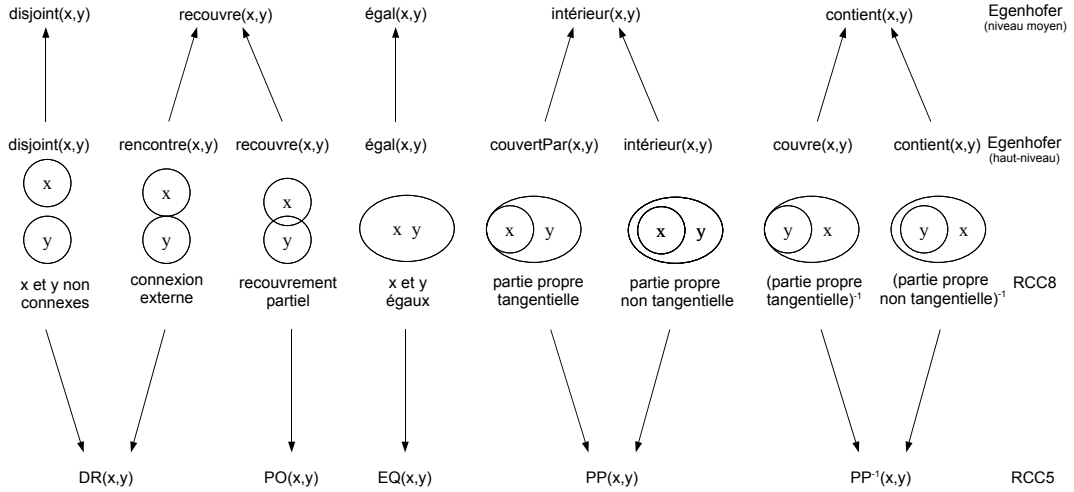


FIG. 2.7 – Lien entre les modèles modèles formels de raisonnement spatial qualitatif.

– $NTTP^{-1}(x, y)$: x possède pour partie propre non tangentielle y,
– $EQ(x, y)$: x est égal à y.

La deuxième approche définit des relations topologiques en utilisant la notion d'intérieur, de frontière et d'extérieur pour un objet spatial. Les relations entre deux objets sont alors définies par une matrice nommée « 9-intersections » permettant d'exprimer toutes les possibilités d'intersection (remplie de 0 et de 1). Du fait que sont utilisés seulement des objets réguliers fermés et non-vides et que les 3 parties des objets (intérieur, frontière, extérieur) sont connectées, seulement 8 parmi les 9 relations sont réalisables [Ege91] : *sépare*, *rencontre*, *est égal*, *est à l'intérieur*, *est couvert par*, *contient*, *couvre* et *chevauche*.

La figure 2.7 fait la comparaison de différents modèles de raisonnement spatial. Elle en montre les similitudes qui sont d'après [KRR97] assez grandes. Pour chaque approche différents niveaux de précision pour la modélisation sont proposés : d'un côté les modèles basés sur le *RCC* : RCC8, RCC5 qui ignore si les 2 régions se touchent ou non (il est à noter qu'il existe d'autres approchent qui étendent le RCC8 comme le RCC23, qui ajoute la notion de convexité dans le raisonnement). D'un autre côté les modèles proposés par Egenhofer comportent aussi plusieurs niveaux de modélisation.

Il arrive souvent que pour construire un modèle de RSQ, on ait un ensemble de contraintes à réaliser simultanément (par exemple, pour l'aménagement d'un bureau : on doit placer l'armoire à droite ou à gauche du bureau, la tablette entre les deux plans de travail, etc.). Plus généralement, on considère un nombre fini d'objets, et un ensemble de contraintes sur leurs positions relatives exprimées dans un langage symbolique. Le problème de cohérence consiste à déterminer si cet ensemble de contraintes a une solution [22]. Les travaux basés sur les théories axiomatiques [RCC92, Coh97, CH01, Ben01] formalisent les différentes contraintes possibles entre les primitives et les relations définies. La thèse de K.Challita [Cha05] présente aussi quelques formalismes spatiaux permettant notam-

[22] http://www.limsi.fr/RS98FF/CHM98FF/LC98FF/lc9.html

ment de modéliser des relations d'ordre directionnelle et topologique et les contraintes associées.

Des travaux imaginent aussi des modèles où aucune autre information spatiale que le voisinage direct, la classification en types de voisinage et la disposition pour les entités spatiales n'est disponible [LE00, EL04] (pas de relations d'orientation ou de distance). Ce modèle permet d'exprimer l'information spatiale telle qu'elle est représentée par les humains n'ayant pas une connaissance complète de l'espace dans lequel ils se trouvent. Il montre cependant que ce manque de connaissances n'entraîne pas la plupart du temps de difficultés supplémentaires pour interpréter le monde environnant.

2.3.1.3 Cas d'utilisation du raisonnement spatial qualitatif

Le RSQ peut être utile quand il est nécessaire de gérer des informations partielles. Il réduit la complexité qu'engendrerait une étude quantitative faisant plus de distinctions que nécessaire. De même il peut être plus précis en évitant les problèmes de discrétisation inhérents à l'utilisation d'information quantitative (comme la perte de l'égalité).

Il doit donc permettre à un ordinateur de faire des prédictions ou des diagnostics, d'expliquer le comportement de systèmes physiques, même dans le cas où une description quantitative précise de ces systèmes n'est pas disponible [CBGG97]. P.Muller [Mul98] parle de l'intérêt que peut avoir une réflexion sur le raisonnement qualitatif et en donne les avantages et les inconvénients. Citant A.G.Cohn il explique que les données numériques sont parfois inadaptées à la résolution d'un problème pour les raisons suivantes :

- la complexité,
- les erreurs de discrétisation,
- les problèmes survenant en cas d'information partielle,
- l'inadéquation des données numériques pour certaines descriptions vagues d'entités spatiales.

De plus, l'utilisation des données qualitatives peut présenter les avantages suivants :

- un nombre de distinctions (degré de précision et complexité des données) adaptées à la tâche visée,
- la comparaison entre valeurs potentiellement inconnues,
- et la non nécessité de fixer la granularité *a priori*.

En résumé l'approche qualitative permet de s'attaquer à une certaine classe de problèmes dans laquelle les données quantitatives peuvent être imprécises ou incomplètes. Contrairement aux Systèmes d'Information Géographique qui ont une manière totale de représenter l'espace, cette approche n'utilise pas de représentation ou de méthode de raisonnement faisant appel à un description numérique ou quantitative.

2.3.1.4 Expérimentation

M.Knauff [KRR97] propose de comparer les modèles théoriques (RCC, etc.) avec les moyens d'expression humains des relations spatiales. Cette étude montre que les différentes relations ne sont pas utilisées de manière égale. Le principe de l'expérience consiste à disposer 96 images contenant toutes 2 cercles avec des configurations différentes, dans

des classes à déterminer par les sujets. Enfin une verbalisation (description) des classes est demandée. Les résultats montrent que la relation topologique seule est la plus utilisée pour la verbalisation (62.1%). Viennent ensuite le couplage de la relation topologique avec soit la relation d'orientation, soit la relation métrique (distance, taille).

Il est donc important de connaître le contexte d'utilisation du modèle de RSQ afin de définir les relations adéquates.

2.3.2 Le concept cible / site

Un des travaux les plus intéressants parmi ceux étudiés est celui de Claude Vandeloise [Van86], linguiste et cognitien qui s'est beaucoup intéressé à l'information spatiale et notamment aux références spatiales. Lui et d'autres chercheurs [Bor98] ont émis une hypothèse concernant l'expression de l'information spatiale.

Ils montrent la manière particulière qu'a l'humain de se représenter une information spatiale lorsqu'elle est évoquée dans le langage écrit [Bor98]. Faire référence à un lieu met en jeu plusieurs éléments et ces éléments respectent une position dans la phrase. Vandeloise propose le concept de cible/site : dans le langage écrit, la cible correspond à l'objet de la description, le site à la référence. Par exemple, dans la phrase *« La voiture est près de l'arbre »*, la *voiture* est la cible et l'*arbre* est le site.

De la même manière, cette hypothèse reste valide pour l'expression des entités spatiales nommées [CMDG04]. En effet, cette hypothèse était posée au départ pour désigner des objets dans un espace directement visualisables par l'humain mais on peut l'étendre à des objets spatiaux représentés sur une carte géographique.

En effet, le concept de « cible / site » est souvent employé dans les corpus territoriaux pour aider le lecteur dans la localisation des récits de voyage. Dans ce cas, la cible est souvent définie par un ou plusieurs sites et des indicateurs d'ordre topologique, d'adjacence, d'orientation, de distance ou une imbrication de ces indicateurs. Par exemple, la phrase *« Au-delà de Pau, le paysage devient triste, [...] J'ai été à pied de Assat aux Eaux-Chaudes... »* tirée du livre *Voyages aux Pyrénées - Pimientos* contient deux lieux : (i) les environs de Pau, où dans ce cas Pau est un site (ii) et la zone entre Assat et les Eaux-Chaudes, contenant deux sites. Les indicateurs utilisés sont ici de l'ordre de l'adjacence.

2.4 Conclusion

Ce chapitre relatant les différents travaux existants sur l'analyse de contenu de documents textuels nous a permis de dégager les techniques donnant des éléments de réponse à notre problématique, qui est de proposer des moyens et des méthodes pour bâtir un système de recherche d'information spatiale.

Le domaine de la Recherche d'Information existe depuis longtemps et propose des méthodes maîtrisées et largement testées. De cette étude il ressort que nous devons nous rapprocher du sous-domaine de Recherche d'Information Spécialisée, car les méthodes classiques ne sont pas adaptées à notre problématique. En effet, nous souhaitons récupé-

rer seulement une partie de l'information contenue dans les textes, l'information spatiale et non pas travailler avec l'ensemble des termes présents dans le corpus. Nous pouvons d'ailleurs considérer que notre corpus est un corpus spécialisé dans la mesure où le vocabulaire particulier de la description d'information spatiale, est largement employé. Les méthodes de RI classiques ne sont pas adaptées parce qu'elles essayent de déterminer par le biais de la loi de Zipf et de la courbe de Luhn, un sous-ensemble de termes significatifs, compte-tenu de leur fréquence d'apparition. Cette méthode, intéressante quand le champ de recherche au moment de l'indexation est inconnu, devient inadaptée quand il est connu, c'est-à-dire dans le cas d'une recherche spécialisée. En effet, il suffit qu'un document contienne quelques termes spatialement informatifs pour devenir, dans notre cas, potentiellement pertinent (« le sud de Pau » par exemple). Ces termes qui peuvent se trouver en dehors de la fourchette déterminée par la conjecture de Luhn (car trop peu mentionnés), sont pourtant significatifs et se doivent d'être indexés dans un système de RI spatial.

Nous devons donc effectuer une recherche active pour récupérer les éléments significatifs du texte. Pour cela, nous pensons qu'il est nécessaire de réaliser une interprétation de cette information utilisée dans l'indexation. En effet, indexer le « sud de Pau » sans connaître la signification spatiale de ce syntagme n'apporte pas grand chose. Une des évolutions citées dans l'état de l'art, l'approche linguistique (utilisant les outils de TALN) semble être la méthode la plus adaptée pour effectuer cette recherche active. Elle permet d'extraire des patrons de syntagmes et d'interpréter la sémantique des relations entre les termes de ce syntagme.

Une autre évolution importante de la RI, proche de l'extraction sémantique des termes et qui pourrait donc nous intéresser, est l'approche conceptuelle et les outils ontologiques. En effet, ce qui intéresse un utilisateur posant une requête spatiale n'est pas le mot-clé qu'il utilise mais la zone géographique qu'il représente. Ce problème se pose d'ailleurs de manière assez similaire pour d'autres thématiques de recherche : un utilisateur choisit un mot-clé mais s'intéresse à une grande partie du champ sémantique auquel il appartient. Des ressources ontologiques sont alors reconstruites. Cette solution pourrait être utilisée pour le spatial, mais cela générerait une ontologie (plutôt un thesaurus vu que les concepts ici ne consisteraient qu'en des entités nommées) beaucoup trop complexe pour être utilisée dans un système de recherche d'information efficace. En effet, le nombre d'entités nommées et de relations autour de ces entités est assez grand (inclusion, distance, appartenance, adjacence, orientation, union, différence, etc.). L'utilisation d'outils spécifiques aux données spatiales géo-référencées est donc plus appropriée. Il existe par exemple les gazetteers ou les Systèmes d'Information Géographique (SIGs) qui peuvent jouer ce rôle ontologique, sans les problèmes de lourdeur puisque les relations sont calculées dynamiquement. Nous traiterons de ces outils en particulier dans le 3ème chapitre sur les travaux existants.

Cependant, ces outils ne savent gérer que les données structurées. Il convient donc de transformer le contenu spatial des documents textuels de manière à construire des structures adéquates. C'est pourquoi les travaux de Raisonnement Spatial Qualitatif vont aussi nous servir, afin d'interpréter et de structurer l'information extraite à l'aide

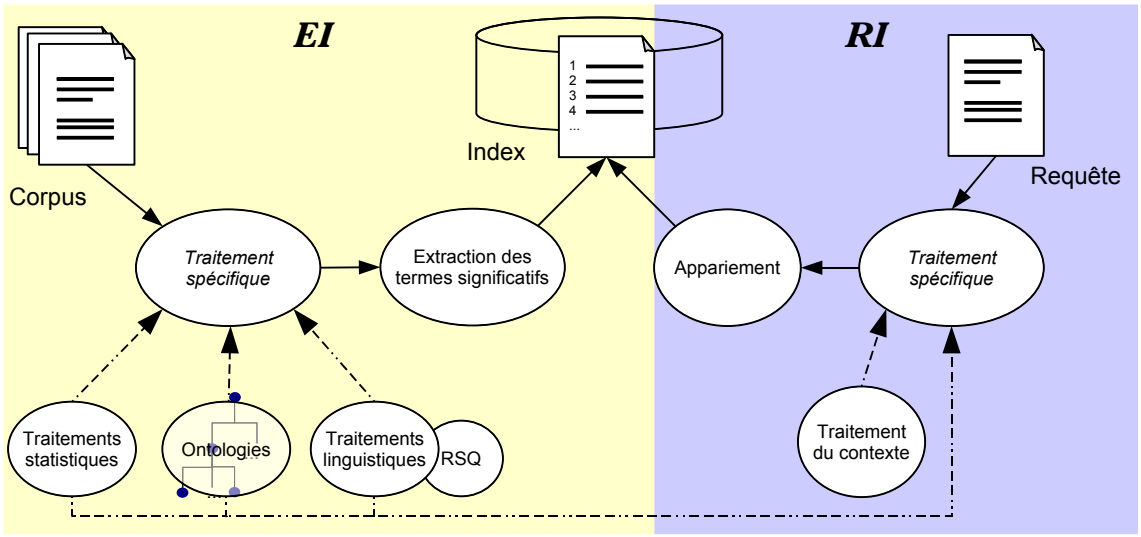


FIG. 2.8 – Outils potentiels pour les processus d'extraction et de recherche d'information spatiale (1).

des outils linguistiques. En particulier le concept de cible / site semble être un bon point de départ pour l'interprétation et la structuration d'information spatiale.

La figure 2.8 tente de synthétiser les travaux existants décrits dans ce chapitre et de voir quels sont les outils les plus pertinents dans le cadre de notre problématique. Parmi les traitements spécifiques à appliquer au texte, l'approche linguistique consistant à faire de la recherche active, c'est-à-dire à extraire des patrons, pour les syntagmes spatiaux est pour nous la plus pertinente. Le raisonnement spatial qualitatif nous permet d'envisager l'interprétation des relations sémantiques de ces syntagmes. Néanmoins l'approche ontologique, dans la mesure où l'on considère les gazetteers ou les couches de systèmes d'information géographique comme des ontologies dynamiques est intéressante dans la manière de structurer l'information recueillie. L'indexation produite par un tel processus est de niveau intra-phrastique, car les syntagmes extraits, interprétés et stockés dans les index se trouvent à l'intérieur de phrases et ont donc une portée de même niveau. Il manque à cette indexation l'approche multi-niveaux que nous voulons proposer pour l'information spatiale, afin d'accéder à différentes granularités du document selon que l'on s'intéresse à un contexte spatial de faible portée ou à un contexte spécifique de plus grande portée (un itinéraire, une description de point de vue par exemple, décrit sur quelques paragraphes, sur un chapitre ou sur une œuvre entière).

Dans la figure 2.8, la phase de recherche d'information consiste en une requête exprimée en texte libre, subissant le même traitement que les documents. Un module d'appariement spécifique utilisant les notions de raisonnement spatial qualitatif peut être alors envisagé pour retourner les documents spatialement pertinents.

Nous verrons donc dans notre premier chapitre de contribution une proposition basée sur des travaux linguistiques, cognitifs et de RI spécialisée. De plus les travaux présentés

sur l'évaluation des systèmes de RI restent valables dans le cas de systèmes spécialisés et seront donc utilisés afin d'évaluer nos contributions.

Chapitre 3

Motifs spatiaux et catégorisation de l'itinéraire

Sommaire

3.1 Introduction

Ce chapitre s'intéresse au contexte particulier dans lequel sont énoncées les informations spatiales présentes dans un texte. En effet, nous avons vu dans le chapitre précédent que les traitements linguistiques existants sont capables de détecter efficacement des syntagmes nominaux exprimant une information spatiale. Cependant cette information extraite est porteuse de sens sur une ou quelques phrases et se situe donc à une granularité très fine du texte, de l'ordre de la phrase. Or, le contexte, défini par l'auteur, exprime une information à un niveau plus élevé d'abstraction, plus significative que les syntagmes qui la composent. De manière plus générale, il est pertinent de travailler sur la sémantique d'une unité de texte plutôt que sur quelques phrases, pour interpréter une information complexe et intéressante. À titre d'exemple, les travaux de [WFB04] s'intéressent à l'extraction d'un contexte particulier, celui de discours contrastifs. Leurs résultats permettent de dégager d'un texte des passages qui s'opposent sur un thème particulier.

Dans le cadre de notre problématique, nous nous sommes penché sur des travaux de modélisation et de cognition à propos de l'expression spatiale dans un texte. À travers cette étude nous avons vu quels sont les différents contextes ou motifs spatiaux qui peuvent être extraits d'un texte. Pour cela, des travaux de psychologie cognitive ont été étudié et, en particulier, celui proposant le concept de carte cognitive. En effet, ce concept s'est révélé être au fil des expérimentations une approche d'une très bonne valeur heuristique pour modéliser ce niveau d'information [Den97]. Nous expliciterons ce concept afin de faire la transition avec la section suivante qui traitera d'un contexte spatial particulier, celui de l'itinéraire, qui émerge parmi les travaux d'analyse cognitive existants. Nous verrons comment la modélisation d'itinéraire est utilisée dans le cadre de descriptions ou de productions automatiques.

3.2 Contextes spatiaux exprimés dans un texte

De nombreux linguistes se sont intéressés à l'expression des relations spatiales dans le but d'analyser la manière dont la langue induit la cognition humaine, notamment en comparant différentes langues. En effet plusieurs hypothèses ont été soutenues [Asi04]. La première est qu'il existe des universaux linguistiques, c'est-à-dire des points de convergence entre toutes les langues sur certains aspects de l'expression (notamment l'expression du spatial) donc du point de vue linguistique. La deuxième est qu'il existe des universaux conceptuels. Quelque soit la langue parlée, la cognition humaine obéit à des points de convergence conceptuels (indépendants de la langue). La dernière hypothèse, de Sapir-Whorf [Sap21, Who69], propose que la diversité linguistique est accompagnée par une diversité conceptuelle correspondante. Cette dernière hypothèse a d'abord été réfuté puis renaît aujourd'hui avec notamment les travaux de Levinson [Lev03] pour l'expression de l'information spatiale. Par exemple, alors que les langues indo-européennes possèdent un cadre de référence relatif pour exprimer les relations spatiales (basé sur les axes corporels de l'observateur et qui peut se traduire par l'utilisation de « à gauche », « devant »), certaines langues[23] ne possèdent qu'un cadre de référence absolu semblable aux directions cardinales.

Il est donc intéressant de noter que l'expression de l'information spatiale peut être abordée de différentes manières selon le langage et que le celui-ci est susceptible de guider la représentation conceptuelle qu'on peut s'en faire. Dans notre travail nous nous restreignons au français, même si l'hypothèse peut être émise que les résultats observés pourraient être étendues à toute langue indo-européenne.

Michel Denis, dans le livre [Den97], fait l'état des réflexions portant sur le langage et la cognition spatiale. En effet le langage est à considérer à la fois comme un moyen d'acquérir et de transmettre l'information spatiale. Cependant, bien que naturel, ce mode d'expression est bien plus complexe à analyser qu'une carte ou un plan géographique. Contrairement à ceux-ci, le langage ne dispose pas *a priori* de formalisme ou de légende pour rendre une vue objective de l'information spatiale. L'homme utilise alors des outils

[23] La langue observée par Levinson était celle d'un peuple amérindien, le *Tzeltal*.

cognitifs pour rendre possible l'acquisition et la transmission de l'information spatiale. Plusieurs études ont été réalisées afin de détecter ces mécanismes cognitifs. Parmi elles l'étude sur les différentes manières de faire une description spatiale sont particulièrement pertinents pour essayer de dégager les différents motifs spatiaux pouvant exister [TT97].

- La première description étudiée est la **description de trajet**. Elle consiste à opérer une linéarisation du monde multi-dimensionel. De plus il est admis que cette linéarisation se fait selon selon l'ordre chronologique d'arrivée des événements. Le locuteur réalise en fait une *visite imaginaire* de l'endroit qu'il décrit en le parcourant virtuellement.
  Pour transmettre l'information, des points de repères sont choisis. Dans le cas d'une description de trajet, le plus souvent, c'est l'auditeur (ou lecteur) à qui s'adresse la description. La description se fait alors par rapport à sa position.
  Enfin la particularité linguistique majeure observée est l'utilisation des verbes d'action (ou verbes de mouvement).
- Une séparation est faite entre cette première description et la **description par parcours du regard**, du fait des différences existant au niveau cognitif comme au niveau linguistique. En effet, le point de repère utilisé dans ce cas est un repère extérieur à la scène décrite. De plus les objets spatiaux sont décrits les uns par rapport aux autres (à l'aide de prépositions comme « à gauche de », « en-dessous de », « derrière », etc.).
  L'observation linguistique pour cette description montrent une prédominance de verbes « statiques » (n'impliquant pas le mouvement du sujet).
- Une dernière classe étudiée est la **description en survol**, dont le repère est surélevé par rapport à la scène. Là aussi les objets spatiaux sont décrits les uns par rapport aux autres, mais à l'aide d'indications du type « au nord de », « au sud de ». Cette classe peut correspondre à la description d'une carte géographique ou depuis un point de vue élevé.

3.3 Cas particulier : les itinéraires

La description de trajet, ou itinéraire, est une problématique largement étudiée de part ses applications possibles comme par exemple la génération automatique de description pour la navigation par GPS[24]. Des études linguistiques [WR82, Kle82] et cognitives [BEJ+97, Lig92] tentent d'y répondre. Nous allons résumer ces travaux d'une part en voyant les propriétés linguistiques observables d'une description d'itinéraire. D'autre part, nous présenterons des travaux analysant la représentation cognitive d'un itinéraire afin d'aider à la production d'une description.

3.3.1 Propriétés linguistiques

Les propriétés linguistiques d'une description d'itinéraire sont étudiées dans [DD98]. Une telle description est en effet spécifique de par les caractéristiques qu'elle réunit : son

[24] Global Positioning System.

objectif final (sa fonction de transmission d'une connaissance), son contenu (du point de vue du discours), sa structure (guidée par les objets qu'elle décrit) et la perspective qu'elle impose à l'utilisateur (perspective égocentrique). Ses caractéristiques induisent un discours particulier du point de vue linguistique.

Du fait de son objectif final, qui est de susciter une action de la part de l'interlocuteur, la description d'itinéraires appartient clairement à la classe des *discours procéduraux*, c'est-à- dire des discours décrivant des actions [Fil01]. Des instructions telles que *Allez tout droit*, *Tounez à droite* sont attendues en grand nombre, sachant que les 2 actions principales édictées sont la progression, pour réduire la distance entre la destination et l'interlocuteur qui suit l'itinéraire, et la réorientation, pour garder le cap vers cette destination. En plus de ces instructions, des éléments de description, tels que des points de repère et des actions à mener comme se réorienter, longer, etc. sont attendus afin de préciser la description.

Le contenu d'une description d'itinéraire est donc lui aussi caractéristique. M.Daniel [DD98] catégorise les propositions utilisées en 5 classes selon l'utilisation des points de repères et des actions induites. Il distingue (i) les propositions demandant une action mais sans utiliser de point de repère (*Tournez à droite*), (ii) des propositions les incluant (*Traversez le parking*). Une autre classe (iii) est prévue pour les propositions sans action, utilisant des points de repères seuls (*Il y a un pont*, *Le pont passe au-dessus de la rivière*) ou faisant référence à l'interlocuteur (*La route en face de toi*). Une autre classe (iv) est prévue pour des propriétés non-spatiales mentionnées à propos des points de repères (*Le pont est en bois*). Enfin, une dernière (v) regroupe les remarques ou commentaires (*Vous ne pouvez pas le rater*). Une expérimentation montre que 80% des propositions utilisées (classes (ii), (iii) et (iv)) dans divers itinéraires comportent des points de repère. Ils constituent donc un élément essentiel de la description, largement étudiés et pris en compte dans les modèles proposés (décrits dans les sections suivantes).

3.3.2 Processus cognitif

La description d'itinéraire a donné lieu à de nombreux travaux dans le domaine de la science cognitive. D'après ces travaux, le processus de représentation mentale de la description se découpe en 2 tâches, la détermination et la description. La figure 3.1 présente schématiquement ces 2 tâches :

- La **détermination** (en haut de la figure 3.1) utilise des connaissances référentielles, c'est-à-dire la connaissance d'un environnement, ainsi que des connaissances pragmatiques, c'est-à-dire le vocabulaire nécessaire à l'interlocuteur pour projeter son déplacement et découper ses étapes. Le résultat de cette première tâche est une suite ordonnée dans le temps d'une représentation « référencée » (dans le sens où elle est constituée de points de repères, de références).
- La **description** de l'itinéraire est ensuite réalisée dans un mode d'expression (en bas de la figure 3.1). Cette tâche se compose d'abord de la structuration conceptuelle, qui consiste à utiliser un modèle d'itinéraire afin de définir l'itinéraire par exemple en une succession d'entités spatio-temporelles appelées « segments » et « relais ». Puis, la structuration textuelle détermine le contenu textuel (*quoi dire ?*)

et la forme textuelle (*comment le dire ?*) afin de générer un texte descriptif compréhensible par l'homme. Le contenu textuel se sert des étapes instanciées dans le modèle d'itinéraire précédent, tandis que la forme textuelle construit des séquences, morceaux d'itinéraires correspondant aux étapes, liées entre elles par des connexions adéquates.

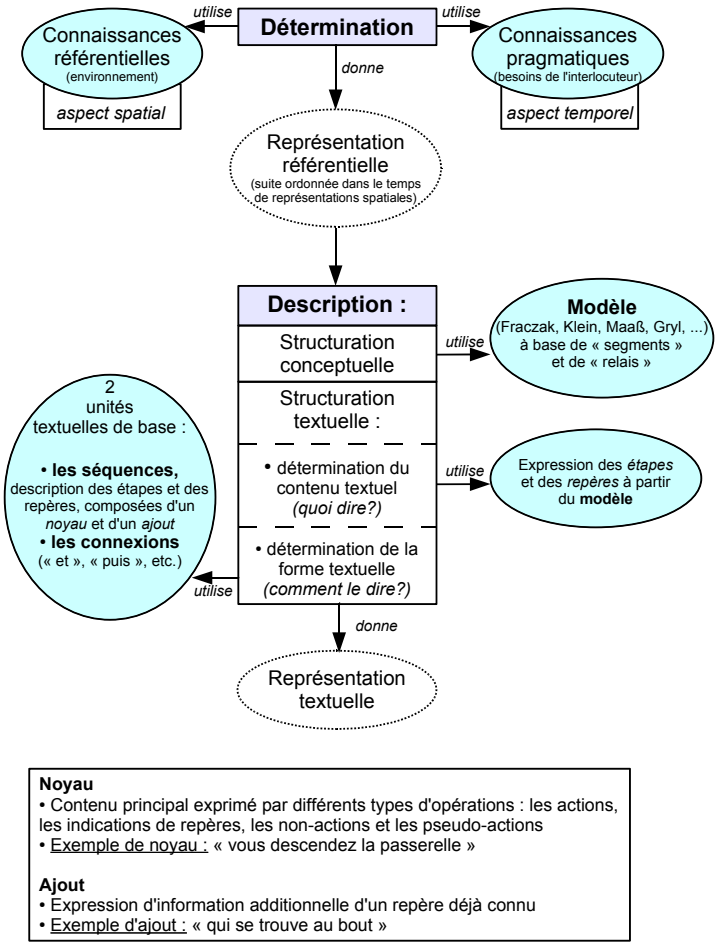


FIG. 3.1 – Production d'une description d'un itinéraire [FL99].

3.3.3 Modèles existants

Plusieurs modèles ont alors été proposés pour la représentation mentale des itinéraires [FL99, KDWH05]. Ils se placent dans le processus (figure 3.1) au moment de la 2ème tâche de *description* afin de permettre la verbalisation de la représentation mentale de

l'itinéraire. Ils sont la plupart du temps définis par trois caractéristiques : des points fixes ou repères (une église, un croisement de rues), des directions (« à gauche, au nord ») et des actions (« descends, traverse, longe »). De plus ils sont bâtis à l'aide de concepts cognitifs, le plus utilisé étant celui de *la carte cognitive*.

La carte cognitive est un processus mental composé d'une série de transformations psychologiques grâce auxquelles un individu peut acquérir, coder, stocker, se remémorer et décoder de l'information sur un phénomène présent dans son environnement. On attribue à Edward C.Tolman la création du concept (et du terme) de « carte cognitive (cognitive map) » dans son article [Tol48] tentant d'expliquer ainsi le comportement de rats dans un labyrinthe. D'après [Tar02] *« la carte cognitive traduit de manière visuelle et dynamique des représentations mentales. Elle permet d'introduire les facteurs cognitifs dans l'analyse et dans la présentation des résultats »* . Plus globalement nous pouvons la considérer comme une modélisation graphique de la cognition, permettant de structurer ses connaissances. Pour M.Denis [DB97], les cartes cognitives d'un locuteur comportent des aspects visuels reflétant les caractéristiques de l'environnement, mais aussi des aspects procéduraux, liés au souvenir des déplacements que le sujet a effectués dans cet environnement.

Même si elles peuvent être utilisées comme une métaphore pour des tâches non-spatiales [Kui77], les cartes cognitives sont donc un moyen de structurer et de représenter l'information spatiale. Elles sont à la base des modèles présentés par la suite.

L'un des premiers modèles est le modèle TOUR de B.Kuipers [Kui77, Kui00], s'appuyant sur la métaphore de la carte cognitive. B.Kuipers considère 5 catégories de connaissance : les *trajets*, séquences d'actions permettant de se déplacer d'un lieu initial à un lieu terminal, la *structure topologique* décrivant les chemins et les lieux sous forme de réseau, la *position relative* de 2 lieux par rapport à un référentiel limité, les *frontières* comme séparateurs de régions et les *régions* elles-mêmes, reliées par une relation d'inclusion. Ces connaissances sont utilisées pour construire des représentations correspondant aux connaissances spatiales, à la position courante et aux règles d'inférence qui manipulent les connaissances spatiales (pour connaître l'orientation à un moment donné de l'itinéraire ou pour gérer la recherche de chemins par exemple). La structure

```
YOU ARE HERE
PLACE: (place description)
PATH: (path description)
DIRECTION: (1-D orientation: +1 or -1)
ORIENT: (coordinate-frame description)
HEADING: (2-D orientation: 0 to 360)
```

illustre ce modèle en montrant comme est définie la représentation de la position courante : description de la position, du chemin, orientation par rapport à ce chemin, définition du système de coordonnées et orientation 2-D par rapport à cette définition. Deux autres structures sont définies de manière similaire pour les deux relations spatiales nécessaires pour la description d'itinéraire, le *Aller à (tout droit)* (*GO-TO*) et le *Tourner* (*TURN*).

D'autres modèles plus récents utilisant le concept de carte cognitive [MK99] permettent aussi de reconstruire les relations topologiques ou métriques d'un itinéraire à partir d'une séquence de « vues » (c'est-à-dire de *points de repères*) et « d'actions » expérimentées par un individu lors d'un déplacement local.

S.Wemer [WKBH00] propose le modèle **Routegraph** dont l'originalité réside dans le fait que la carte cognitive et les relations spatiales sont traduites sous forme de graphe, les arcs étant les segments et les nœuds étant les *points de repère*. La difficulté d'implémentation de ce modèle réside dans la détermination et la fusion de segments identiques afin de diminuer le nombre d'arcs.

Le modèle CORAL [DGP03] est une architecture et une implémentation de génération en langage naturel d'itinéraire pour l'aide à la navigation (dont les résultats sont comparables aux pages de résultats de ViaMichelin[25] ou de Mappy[26]). Le processus part d'une description automatique d'itinéraire, sous forme d'un tableau de type : Instruction, Nom de rue, etc. et passe par des étapes successives de raffinement, de segmentation et d'aggrégation de cette information.

Un dernier exemple (Abstract Route Directions (ARD)) [RK04] propose un modèle et une implémentation d'un système de génération de *segments* et de directions au niveau des *points de repères*. Ce travail se place donc en amont de la génération en langage naturel.

Une autre notion importante qui se dégage de tous ces modèles est celle de *point de repère*, dans la mesure où il guide la structuration conceptuelle de l'itinéraire. Nous prenons ici une section pour la définir.

3.3.4 Définitions d'un point de repère dans un itinéraire

De nombreux chercheurs ont défini les repères, ou points de repère (*landmark*), dans leurs modèles. Voici une synthèse proposée par A.Klippel [KDWH05] :

- Tout ce qui se détâche de l'arrière-plan peut être un repère.
- Dans certains cas même des routes peuvent être des repères.
- Les repères structurent la connaissance de l'environnement.
- Ils sont utilisés pour transmettre l'information sur l'itinéraire oralement ou graphiquement.
- Ils sont intégrés dans les plans de route à divers degrés, le plus souvent pour la description du départ, de l'arrivée et des points d'intérêt.
- Ils sont d'autant plus pertinents quand ils sont cités lors d'un changement de direction.
- Enfin, les repères sont plus efficaces que les panneaux de route quand on cherche son chemin.

Cette synthèse montre qu'un repère n'est *a priori* pas défini de manière stricte et claire. Dans la mesure où même une route peut faire office de repère, alors qu'on la catégoriserait plus facilement comme partie intégrante d'un itinéraire, il peut paraître

[25] http://www.viamichelin.fr/
[26] http://www.mappy.fr/

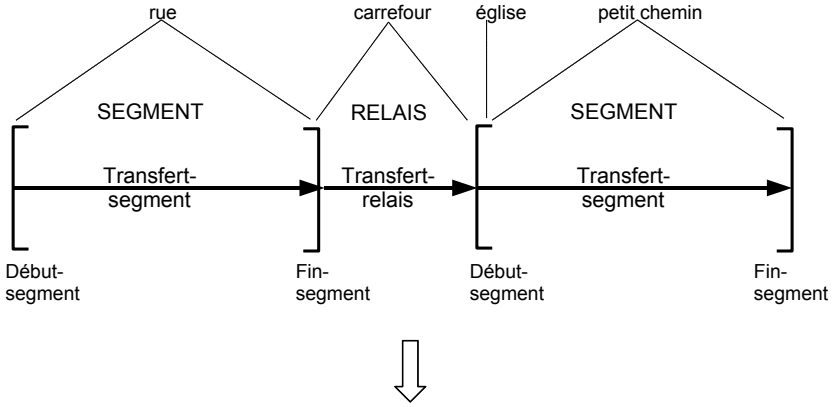


FIG. 3.2 – Structuration conceptuelle d'un itinéraire (aspect temporel). Exemple avec un résultat de production textuelle [FL99].

difficile de faire une définition générique.

La problématique inverse de détection d'itinéraire dans un texte, qui sous-tend la détection des points de repères peut alors s'avérer très complexe. La plupart des travaux présentés ici utilisent des lexiques d'entités spatiales (nommés ou non) pour les définir. Il est alors possible de faire de même pour la détection. Des travaux de constructions de tels lexiques, comme la « gazetteer » du projet Alexandria Digital Library (ADL) [Hil04,HGJ04] répondent partiellement en listant les entités nommées. Nous verrons dans le chapitre suivant les problèmes d'ambiguïté qu'entraînent ces ressources génériques.

3.3.5 Exemple de production d'une description d'itinéraire

Afin d'illustrer les concepts définies dans cette section et les exemples de modèles proposés, nous détaillons ici un exemple de reconstruction de description d'itinéraire à l'aide de la figure 3.2 et basé sur les travaux de L.Fraczak [FL99]. Pour caractériser un itinéraire, Fraczak définit une structure à un niveau « global », composée d'une succession d'entités spatio-temporelles appelées *segments* et *relais*. Les segments sont des morceaux d'itinéraires qui se caractérisent par des propriétés constantes. Les relais marquent un changement de ces propriétés. Ces concepts sont définis et utilisés de manière similaire dans d'autres modèles [Gry95]. Au sein de ces concepts apparaissent une structuration spatio-temporelle d'ordre « local ». Cette structuration est composée d'*étapes* et de *relais*. Les étapes, qui sont des entités temporelles, correspondent dans la figure 3.2 aux segments *début-segment*, *transfert-segment*, *fin-segment* et *transfert-relais*. Les repères (ou points de repères) sont comme on l'a vu des éléments spatiaux qui se dégagent de l'environnement par des caractéristiques particulières. Là aussi Fraczak les catégorisent en quatre types : les *repères-segment*, associés à un segment, les *repères-relais* associés à

un relais, les *repères-cadre* associés à une plus grande partie de l'itinéraire et les *repères auxiliaires* associés à une ou deux étapes.

En prenant l'exemple de la figure 3.2 et la figure 3.1, nous pouvons retracer la formulation de cet itinéraire. Durant la phase de **détermination**, on fait appel aux connaissances référentielles « rue », « carrefour », « église » et « petit chemin » et aux connaissances pragmatiques « Là, traverser », « juste avant la pente », etc.
A l'étape suivante de **description** une structuration conceptuelle est produite suivant le modèle de Fraczak. Le résultat ici est formé de deux segments (*« Prendre la rue... »* et *« Prendre le petit chemin... »*) reliés par un relais (*« jusqu'à un carrefour »*). Les termes « rue » et « petit chemin » sont des repères-segment ; « carrefour » est un repère-relais et « église » est un repère auxiliaire associé à une seule étape (*début-segment*).
Ensuite la phase de structuration textuelle se charge d'exprimer l'interprétation conceptuelle en construisant des séquences reliés par des connecteurs appropriés. Un résultat possible se trouve dans la partie haute de la figure 3.2.

3.4 Conclusion

À travers cette étude, nous avons vu que, du point de vue linguistique ou cognitif, les textes à connotation spatiale, ou du moins l'expression de l'information spatiale en langage naturel, sont aux centres de problématiques depuis longtemps. Plusieurs motifs spatiaux ont été étudiés, et pour chacun d'entre eux des propriétés linguistiques ont été dégagées. Ces motifs nous intéressent car dans le cadre de notre problématique d'indexation spatiale, ils peuvent constituer des éléments résumant une unité de texte et faisant office d'index spatial.

Pour retrouver ces motifs dans le texte, nous pouvons alors nous servir encore une fois de travaux linguistiques et cognitifs. Cependant notre problématique est à l'inverse de celle des travaux présentés dans ce chapitre. Les motifs spatiaux sont déjà exprimés en langage naturel et il s'agit de les identifier au milieu d'un texte. Il faut aussi faire attention aux limitations de ces travaux. M.Denis [DB97] affirme, par exemple, que les cartes cognitives sont soumises à des distorsions dues à des erreurs de précision dans la mémoire ou l'expérience des locuteurs, ou dans les mots (verbes) qu'ils utilisent. Il faut donc essayer de prendre, pour une indexation la plus précise possible, le maximum d'informations sûres, comme les entités nommées. Nous avons vu que les itinéraires sont définis pour partie grâce à des points de repères qui permettent de diriger l'interlocuteur[27]. Compte-tenu des spécificités du corpus d'étude, une restriction peut être faite sur les point de repères à détecter. Nous avons vu que les outils linguistiques peuvent extraire le entités nommées et les indicateurs spatiaux attenants. Ces entités peuvent constituer de bons repères dans la mesure où ils sont considérés par l'humain comme une information sûre de la localisation et sont géo-référençables. À un premier niveau où les problèmes d'homonymie et de variabilité orthographique sont ignorés, ils sont un sous-ensemble satisfaisant des repères mentionnés par l'auteur. Cependant, parmi ce

[27] Nous considérons que les autres motifs spatiaux sont aussi constitués de repères visuels

sous-ensemble se trouvent des repères pertinents, c'est-à-dire qu'ils sont mentionnés lors d'un changement de direction, et des repères visuels moins intéressants observables depuis l'itinéraire décrit. Il faudra donc trouver un moyen de filtrer ces derniers pour qu'ils ne bruitent pas la reconstruction des itinéraires (et des autres motifs).

Enfin, d'après les travaux des cognitiens sur les itinéraires, le phénomène est à la fois spatial et temporel. L'indexation par ce type de motifs serait donc plus complet qu'une indexation spatiale. Elle formerait un phénomène spatio-temporel comparable à une *molécule géographique complète*. Nous définirons cette molécule dans nos préconisations.

La figure 3.3 reprend le schéma de la conclusion du chapitre précédent (figure 2.8). Afin de proposer une indexation multi-niveaux, un traitement spécifique a été ajouté durant la phase d'extraction de l'information, basée sur un raisonnement sur les motifs spatiaux. Celui-ci permet d'extraire l'information spatiale à différents niveaux d'abstraction. L'indexation permet alors de retourner, durant la phase de recherche, des documents fragmentés au niveau de la phrase, du paragraphe ou du chapitre.

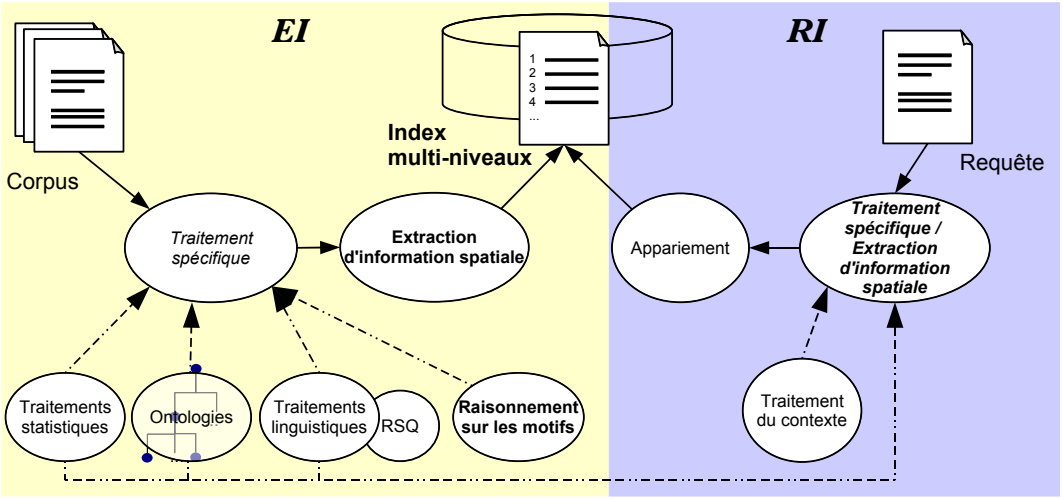


FIG. 3.3 – Processus d'extraction et de recherche d'information spatiale (2).

Chapitre 4

Manipulation de représentations géométriques relatives aux informations spatiales

Sommaire

4.1 Introduction

Ce chapitre aborde l'étude de l'information géographique traitée au départ comme une donnée structurée, en particulier dans les Systèmes d'Information Géographique (SIG). Nous allons voir que les méthodes d'analyse et de traitement fournies par ces systèmes dédiés sont souvent insuffisantes pour répondre directement aux problèmes inhérents à l'expression de l'information spatiale, notamment au niveau de l'aspect qualitatif qui peut apparaître dans le cas de son expression en langage naturel. Néanmoins, des SRI existants traitant du spatial intègrent plus ou moins les fonctionnalités propres aux SIG. Nous présenterons dans une première section des travaux utilisant la technolo-

gie des SIG, dans lesquels des données semi-structurées au départ, telles que des pages web, sont analysées puis indexées spatialement à l'aide de ces fonctionnalités spécifiques.

Nous reviendrons dans une deuxième section sur une définition des SIG et sur les fonctionnalités potentielles qu'ils pourraient apporter à des SRI spatiaux, en plus de celles déjà utilisées.

Enfin nous ferons la synthèse de ce chapitre en soulignant les avancées possibles autour de la jonction du domaine des SIG et de celui des données non-structurées dans un contexte spatial.

4.2 Extraction et recherche d'information géographique

Cette section présente comment les systèmes de recherche d'information spécialisés dans l'information spatiale peuvent utiliser les outils SIG. Ils peuvent les utiliser à quatre niveaux [JAF03] :

- Une première application possible est de valider les entités nommées trouvées dans un document (le SIG est utilisé dans ce cas comme un lexique).
- Le seconde consiste à donner une représentation géographique aux entités spatiales.
- La troisième application possible serait de gérer des index ainsi construits.
- Enfin les fonctionalités intégrées dans les SIG permettent de proposer des fonctions d'appariement et de calcul de pertinence pour un processus de recherche documentaire spatial.

Les systèmes existants de RI spatiale reçoivent des requêtes comprenant une composante spatiale qu'ils savent interpréter. La formalisation des données spatiales fournit un moyen de construire des méthodes de requêtage pour de tels systèmes. Par exemple, le projet SPIRIT (dont une des réalisations consiste à faire de la recherche d'information géographique dans les pages web [CSJ04], permet de faire des requêtes spatiales évoluées, prenant en compte des relations de distances et d'orientation [FJA05b]. En effet, à chaque page web du corpus est associée au préalable une localisation. Ce corpus étant composé de pages d'hôtel, de restaurant, etc. elles contiennent une adresse pertinente pour l'indexation. L'extraction consiste à détecter les patrons (c'est-à-dire les formes syntaxiques) que peuvent prendre une adresse. La requête interprétée permet de retourner les documents pertinents qui s'intersectent spatialement. Il est à noter que d'après [FJA05b, JAF+04] seule l'information qualitative présente dans requête est interprétée. Les fonctions SIG sont alors utilisées de manière minimale pour l'appariement : un calcul de distance est effectué pour des indications qualitatives d'adjacence et un calcul d'angle en fonction des axes cardinaux est effectué pour des indications qualitatives d'orientation. Le système d'indexation se base quant à lui sur les méthodes classiques présentes dans les SIG (grille régulière, R-tree) que nous présenterons dans ce chapitre.

D'autres projets relativement similaires utilisent les mêmes techniques comme le projet GIPSY [WP94] qui propose une indexation spatiale par agrégation. Celui-ci commence par détecter les entités nommées de lieux d'un document ainsi que d'éventuelles indications spatiales. Puis il les géo-référence ; pour gérer les « modificateurs géographiques » que sont ces indications, quelques fonctions sont implémentées. Une indication

d'orientation *(au sud du Lac Tahoe)* aura pour effet de déplacer la zone trouvée pour l'entité de départ *(le Lac Tahoe)* vers le point cardinal adéquat *(sud)*. Le choix fait par GIPSY est donc de garder la même surface pour l'entité modifiée. La dernière étape de cette indexation consiste enfin à prendre la zone géographique prédominante, c'est-à-dire la zone où les géo-réoférences s'intersectent le plus, ceci afin de limiter les erreurs.

Le projet GeoSem, mentionné dans le chapitre 2, n'utilise pas quant à lui les outils SIG. Le stockage se fait sous forme de flux XML à l'aide d'une base de données dédiée[28] et l'appariement à l'aide d'un algorithme propre basé sur des arbres hiérarchiques s'apparentant à des ressources ontologiques (telles que *Caen* se trouvant dans le *Calvados*).

Un workshop intéressant pour la recherche d'information géographique, nommé GIR [SMC+05, MS05, SVH04, RS95, NMR04, LF04] du groupe SIGIR[29], est un bon moyen de faire l'état d'avancement des travaux actuels de recherche concernant cette problématique.

4.2.1 Détection des entités nommées

Les couches de données des SIG peuvent servir de lexiques pour la détection d'entités nommées, lorsqu'ils contiennent les dénominations des lieux. De plus ces dénominations son indirectement reliées entre elles par de pseudo-relations d'hyperonymie, de méréonymie et d'hyponymie, mais dans la dimension spatiale, calculables grâce aux fonctions SIG. Si nous avons par exemple la couche des communes et la couche des départements de France, nous pouvons savoir à quel département appartient telle ville ou, à l'inverse, savoir quelles sont les communes composant un département. Ces relations d'appartenance permettent de recréer une ressource de type ontologique, les concepts étant les entités nommées (de type commune, région, forêt, cours d'eau, etc.) et les relations étant d'ordre topologique, d'appartenance, de connectivité, de distance, etc. Des problèmes d'ambiguïté peuvent cependant surgir.

La dénomination des entités n'est a priori pas le meilleur moyen de les identifier de manière unique mais c'est le seul disponible. En effet aucune règle d'unicité n'existe sur la dénomination des lieux [OR06]. Beaucoup de travaux se sont penchés sur ce problème, utilisant le contexte dans lequel sont énoncés les entités, des résultats statistiques à partir de ressources web, etc. Par exemple, J.Leveling [LH06] soulève le problème métonymique des noms de lieux. Ceux-ci peuvent en effet porter un sens différent du sens littéral dans certains contextes. Ces noms de lieux, à part le sens géographique, peuvent correpondre à des événements (« Korea turned out to be a military catastrophe for the USA. »), des gens (« Yesterday, Seoul and Peking agreed to start diplomatic relations. ») ou des produits (« Wine connoisseurs know that Chianti has to be decanted before drinking. »). J.Leveling propose de lever l'ambiguïté en classant chaque entité nommée selon qu'elle doit être considérée au sens littéral, métonymique ou mixte, grâce à une méthode d'apprentissage. S. E. Overell [OR06] propose la construction d'un modèle permettant de déterminer si une géo-localisation peut être associée à une entité, la base de tests étant

[28] Base de données eXist, http://exist.sourceforge.net/

[29] http://www.acm.org/sigs/sigir/

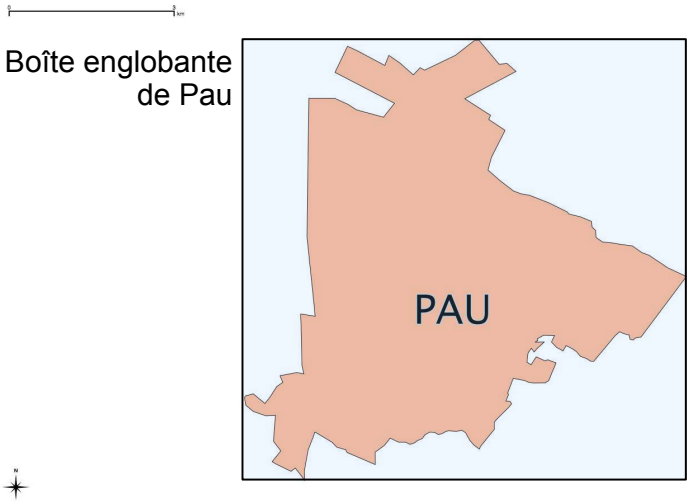


FIG. 4.1 – Représentation polygonale de la ville de Pau, et sa boîte engobante.

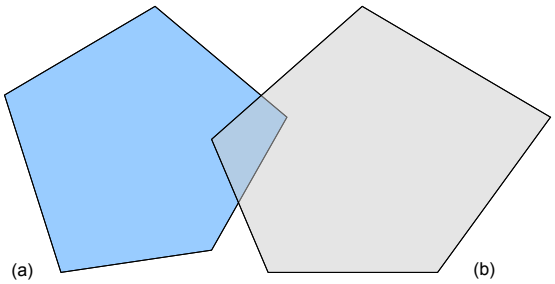
le site Wikipedia[30]. Cependant, des problèmes peuvent persister dans certains cas. Par exemple, il est courant que des lieux aux caractéristiques similaires aient le même nom (comme les quelques « Lac Bleu » des Pyrénées). Il est alors difficile d'identifier le bon géo-référencement sans l'aide du contexte.

4.2.2 Indexation et appariement avec une requête spatiale

Les SIG peuvent aussi servir pour remplacer les méthodes classiques d'appariement basée sur des similarités de vecteurs dans un espace vectoriel, ou sur des probabilités de pertinence. Le calcul de la pertinence, pour une requête d'ordre spatial, peut naturellement découler d'une proportion d'intersection entre la requête géo-référencée et un syntagme spatial extrait d'un document et géo-référencé lui aussi. Une réflexion sur l'analyse des relations topologiques, de distance, d'appartenance, etc. peut être aussi réalisée pour ce calcul d'appariement. Les projets existants utilisent la notion de *boîte englobante* afin de représenter les entités nommées. La figure 4.1 en montre un exemple pour la représentation polygonale de la ville de Pau. Nous voyons qu'une *boîte englobante* est un rectangle qui englobe de manière minimale le polygone original, ses côtés étant parallèles aux axes du repère choisi pour la représentation. Ces boîtes servent d'index afin de réaliser les calculs d'appariement.

Ensuite des fonctions topologiques d'ordre qualitatif peuvent être intégrées. À titre d'exemple le SIG *PostGIS*[31] *via* son module *Geos* implémente les fonctionnalités relatives à la matrice 9-intersection étendue (Dimensionally Extended Nine Intersection Matrix : DE 9-IM), matrice qui est un résultat de travaux de raisonnement spatial qualitatif cités dans l'état de l'art. Ces travaux ont en effet été intégrés dans la spécification de

[30] http://www.wikipedia.org/
[31] http://postgis.refractions.net/



	Intérieur	Frontière	Extérieur
Intérieur	2	1	2
Frontière	1	0	1
Extérieur	2	1	2

FIG. 4.2 – Exemple de remplissage de la matrice 9-intersections étendue pour 2 polygones qui se chevauchent (tiré de [Ope99]).

l'OGC [Ope99] (§ 2.2.13.2). Ils permettent de donner pour l'intérieur, la frontière et l'extérieur d'un objet la dimension d'intersection avec l'intérieur, la frontière et l'extérieur d'un autre objet. La figure 4.2 montre un exemple de remplissage.

4.3 Indexation dans les Systèmes d'Information Géographique

« Le Système d'Information Géographique [SIG ou GIS en anglais] est un système de gestion de base de données comprenant des types de données dédiés à l'information géographique et des opérations permettant de les manipuler »[32].

Cette citation est un bon résumé définissant les SIG. Ils consistent en effet en des outils dédiés au stockage, à la manipulation et à la représentation de données géo-référencées. L'information géographique prenant part dans de nombreux domaines, ces systèmes se voient enrichis d'une multitude de fonctionnalités spécifiques aux besoins de ces domaines. En effet, en plus des fonctionnalités habituelles d'un SGBD[33] classique, un SIG comporte des fonctionnalités propres au stockage et à la manipulation de données géographiques.

En effet, un SIG doit pouvoir répondre à ces différentes questions : « où le phénomène se trouve-t-il ? », « que trouve-t-on à cet endroit ? », « quelles sont les relations entre les

[32] http://www.gis.com/whatisgis/index.html
[33] Système de Gestion de Base de Données

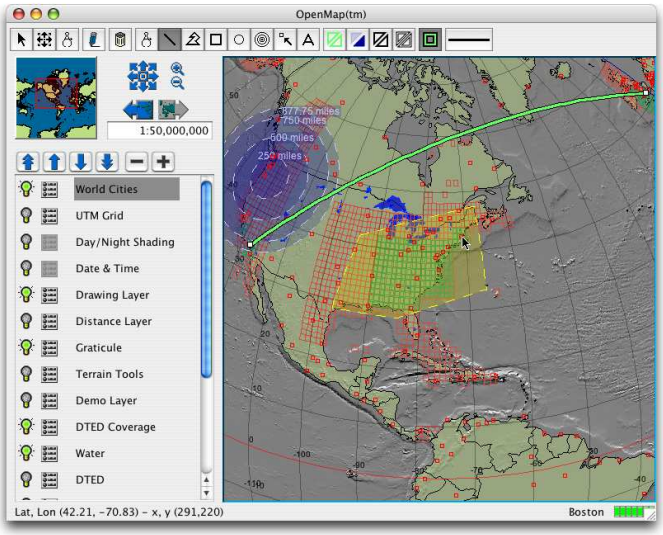


FIG. 4.3 – Visualisation de couches de données (plus des courbes faites avec des outils de dessins fournis) sous une application utilisant OpenMap.

objets représentés et le phénomène étudié ? ».

Historiquement, le premier SIG est le CGIS[34] et date du milieu des années 60. Il a été développé pour représenter les ressources du territoire canadien. Ce premier logiciel était considéré comme un calculateur de cartes, c'est-à-dire un simple générateur d'images [LGMR01]. La deuxième évolution importante date de la fin des années 60 avec le programme DIME[35] de l'organisation américaine « US Bureau of the Census ». Il a consisté à l'enregistrement et à la numérisation de l'ensemble de l'infrastructure routière des Etats-Unis. Ce projet découla sur l'important programme du laboratoire d'informatique et d'analyse spatiale de l'université de Harvard, qui déboucha à la fin des années 70 sur le SIG ODYSSEY, regroupant les fonctionnalités génériques permettant de répondre aux besoins d'applications différentes. La Guerre Froide et le développement des satellites d'espionnage permirent l'acquisition de plus en plus importante de données, et le développement massif de SIG répondant aux besoins de plus en plus variés. Enfin, l'histoire moderne des SIG débute au début des années 80 avec la baisse conséquente du prix d'ordinateurs suffisamment puissants et la démocratisation d'internet. C'est en 1993 qu'est publiée par Steve Putz du « Xerox PARC » la première carte interactive basée WEB, où les utilisateurs connectés à internet peuvent zoomer et se déplacer sur des parties de la carte, en se servant de simples clics de souris, sans avoir besoin d'installer un logiciel particulier, ou de télécharger des volumes de données importants [36].

Aujourd'hui les SIG sont des outils utilisés dans beaucoup de domaines. Pour cela, ils intègrent de nouvelles technologies adaptées en proposant de plus en plus de fonctionna-

[34] CGIS : Canada Geographic Information System
[35] DIME : Dual Independent Map Encoding
[36] le site existe encore : mapweb.parc.xerox.com/map

lités. Un travail de spécification sur ses fonctionnalités est mené par plusieurs groupes : des sociétés comme ESRI[37] et des organismes de la communauté *OpenSource* comme l'*Open Geospatial Consortium, Inc (OGC)*[38]. Ce dernier est d'ailleurs à l'origine du *Geography Markup Language (GML)*, spécification XML de l'information géographique, utilisée pour construire des modèles ou des moyens de stocker l'information tout en assurant l'interopérabilité et la portabilité.

En outre de plus en plus d'outils disponibles, en particulier sur le Web, démocratisent l'accès à l'information géographique. Ces nouvelles interfaces web et nouveaux outils permettent de visualiser des cartes tout en proposant des interactions plus ou moins poussées. Citons comme exemples Google Maps (`http://maps.google.fr`), ViaMichelin (`http://www.viamichelin.fr`), MapServer (`http://mapserver.gis.umn.edu`), MapGuide (`http://www.autodesk.fr/mapguide`) ou OpenMap (`http://openmap.bbn.com`) (figure 4.3). Ces outils proposent des APIs[39] permettant de créer soit même une interface cartographique web et interactive avec des possibilités d'interrogation de bases SIG. Le site `geotrace.net/gm` est un exemple d'utilisation de Google Map avec le SIG PostGIS[40].

De manière plus marginale, Google propose son fameux GoogleEarth (`http://earth.google.fr/`) qui est une application (non web) permettant de visualiser des données géoréférencées sur le globe terrestre 3D. Une API est prévu prochainement afin de construire son propre outil de visualisation. Du côté des travaux français, l'IGN propose une interface web 2D (`http://www.geoportail.fr/`) et une 3D (pas encore disponible) utilisant ses ressources (notamment photographiques) pour visualiser le territoire français. Malheureusement, à ce jour aucune API en ligne n'est proposée.

4.3.1 Fonctionnalités des SIG

Un SIG permet d'étudier un phénomène géographique, celui-ci étant stocké de manière à faciliter les requêtes et la sélection d'information. Ce stockage est structuré sous forme de couches d'information, chaque couche regroupant des objets géographiques de mêmes propriétés. Prenons l'exemple du stockage de l'ensemble des données géographiques contenues dans une ville : nous allons récupérer son réseau routier, son bâti, ses espaces verts, etc. Chacune de ces catégories pourra être manipulée, analysée et visualisée séparément, car elle sera rangée dans une couche qui lui est propre. Un intérêt majeur est aussi de pouvoir faire des requêtes sur plusieurs de ces couches simultanément. En effet les requêtes « basées-SIG » permettent d'interroger la base sur un ensemble de couches liées entre elles par des caractéristiques topologiques. Par exemple, les requêtes habituellement utilisées portent sur des notions d'appartenance ou de contenance d'objets géographiques dans d'autres objets géographiques. L'Open Geospatial Consortium Inc. (OGC) spécifie par exemple [Ope99] les différentes opérations à implémenter dans un SIG :

[37] `http://www.esri.com/index.html`
[38] `http://www.opengeospatial.org/`
[39] Application Programming Interface.
[40] `http://postgis.refractions.net/`. Ce SIG OpenSource suit les spécifications de l'OGC.

- pour connaître les attributs de l'objet géographique : *Dimension*() qui donne la dimension de l'objet, $IsEmpty$() qui dit si c'est un objet vide, $SpatialReference$() qui retourne le système de référence spatial utilisé,
- les opérations de base : *Clone*(), *Envelope*() qui retourne la *boîte englobante* de l'objet,
- les méthodes d'accès, réalisant les tests des relations entre objets géographiques (spécifiées à partir des modèles de raisonnement spatial qualitatif tels que celui d'Egenhofer [Ege91]) : *Contains()*, *Crosses()*, *Disjoint()*, *Equals()*, *Intersects()*, *Overlaps()*, *Touches()* et *Within()*. Une opération supplémentaire résume les opérations précédentes, nommée *Relate()*. Celle-ci calcule la matrice d'intersection d'Egenhofer (un exemple est détaillé à la fin du chapitre 4).
- Les méthodes d'analyse spatiale : *Boundary()* qui transforme l'objet en ne gardant que la fermeture, *Buffer()* qui correspond à l'objet de départ augmenté d'un facteur de distance, *ConvexHull()* qui retourne l'enveloppe convexe de l'objet, *Distance()* qui retourne la plus courte distance entre 2 objets, etc.

Des requêtes sur la distance, la surface d'intersection sont donc possibles, de même que des modifications d'objets géographiques (par union, intersection) (voir la spécification de l'Open Geospatial Consortium Inc. (OGC) pour voir l'ensemble des différentes fonctions fournies par les SIG). Des fonctionnalités plus avancées existent aussi, se basant sur les opérations spécifiées par l'OGC, notamment concernant les calculs sur les graphes routiers. À titre d'exemple, l'API PgRouting[41] implémente l'algorithme de Dijkstra, d'A^*, de $Shooting^*$, l'algorithme du « voyageur de commerce[42] ». Il pourrait être utile pour la résolution d'itinéraires à partir de points de repères se trouvant sur un graphe routier.

La méthode par couches d'informations permet aussi de faciliter l'intégration de nouvelles couches de données dans le domaine d'étude, à la condition que celles-ci soient géoréférencées dans le même système de référence, c'est-à-dire le même repère spatial (Lambert II étendu pour la France ou WGS84 (Word Geodesic System) pour le monde, par exemple).

Nous pouvons remarquer que les couches de données peuvent avoir des caractéristiques complètement différentes : certaines peuvent être définies de manière très précise. Par exemple, un réseau routier ou les bâtiments d'une ville, sont répertoriés et référencés grâce au cadastre. D'autres données comme les objets « naturels » (tels que les forêts, les cours d'eau) sont définies de manière plus vague. C'est pourquoi les SIG intègrent différentes manières de stocker l'information géographique.

4.3.2 Structures de stockage des données géographiques

Il existe principalement, deux types de structure pour stocker les données géographiques : le type « matriciel (ou raster en anglais) » et le type « vectoriel (vector) ».

Dans une représentation matricielle, l'espace géographique est divisé en un tableau

[41] http://www.postgis.fr/node/360
[42] http://fr.wikipedia.org/wiki/Problème_du_voyageur_de_commerce

de cellules habituellement carrées, ou rectangulaires. Toute variation géographique est alors exprimée par l'assignation de propriétés ou d'attributs à ces cellules. L'utilisation de ce type de données provient de la façon dont celles-ci sont récupérées : les photos satellites, les photos d'avions ou les outils de mesure qui parcourent un territoire et font des relevés à intervalles réguliers. C'est donc un type très courant. Cependant, tout détail de variation à l'intérieur des cellules est perdu, à moins d'augmenter la résolution de la grille, tout en gardant à l'esprit qu'une résolution deux fois plus grande entraînant le stockage de quatre fois plus d'informations. Elle prend donc souvent plus de place que le type vectoriel. Dans une représentation vectorielle, les objets sont définis à partir de primitives spatiales (point, ligne, polygone, etc.). La représentation est donc plus précise et permet de faire des calculs topologiques mais l'acquisition est plus complexe que pour le type matriciel. En effet, il faut trouver un modèle géométrique adapté aux objets géographiques que l'on veut représenter. Par contre il semble que sa structure soit plus efficace pour le stockage de données, même si de nombreux phénomènes géographiques ne peuvent pas être localisés avec ce système qui requiert beaucoup de précision.

Le choix entre matriciel et vectoriel doit être réalisé en adéquation avec l'information à stocker. Afin de connaître la méthode de stockage la plus adaptée, un tableau comparatif suivant le type de problématique est proposé par P.A.Longley dans [LGMR01].

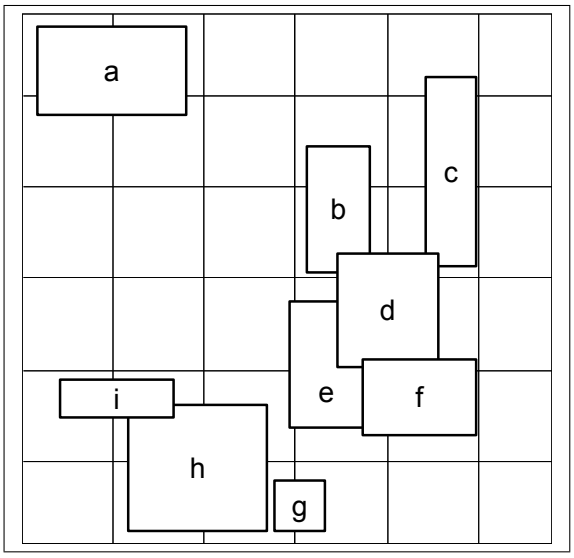
Ces structures de stockage sont admises dans l'ensemble des implémentations de SIG. Elles permettent de répondre aux besoins basiques de ces systèmes sur l'interrogation de données géographiques. Il existe cependant des propositions de modèles pour la structuration des données géographiques, d'un degré d'abstraction plus élevé, prenant en compte le caractère multi-critères de l'information géographique : son aspect spatial, son aspect temporel et les données exprimant un phénomène particulier. Le projet MADS[43] par exemple [PSZ99], définit des structures de données complexes, composées de n-uplets (nom, cardinalité, zone spatiale, zone temporelle, domaine), où la cardinalité définit si les structures sont multi-valuées ou non et le domaine définit si elles sont composées de type simples (lignes, points) ou complexes (ensembles de lignes, ensembles de points) définis récursivement.

4.3.3 Méthodes d'indexation spatiale

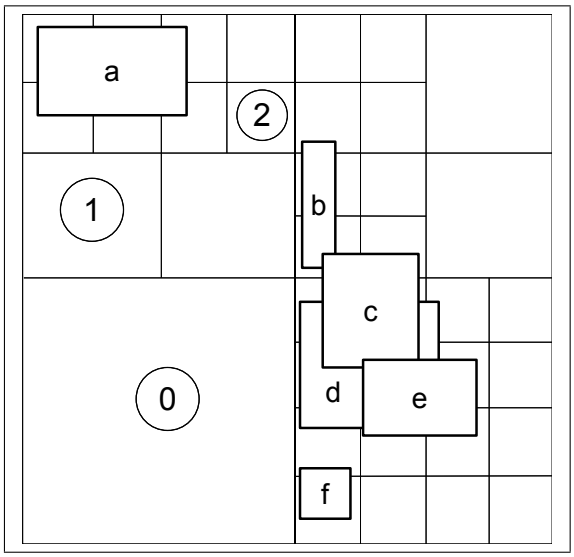
Nous décrivons ici sommairement quelques méthodes d'indexation spatiale afin de comprendre comment est manipulée l'information stockée dans les SIG. Nous vous ramenons à des travaux comme [Rob81, BKSS90, WJ96, KS97] pour voir en détails les différentes structures d'indexation proposées (comme le *K-D-B-tree*, le *R*-tree*, le *SS-tree*, etc.).

Globalement, il existe 2 idées directrices, la première étant de décomposer l'espace de référence, la deuxième consistant à recouvrir les objets spatiaux [Sag99], on dit alors qu'elle est basée-données car elle est guidée par les données stockées. L'indexation par grille (régulière, adaptative, à plusieurs niveaux) dont le découpage est illustrée en figure 4.4(a) fait partie de la première approche, de même que pour l'indexation par arbre

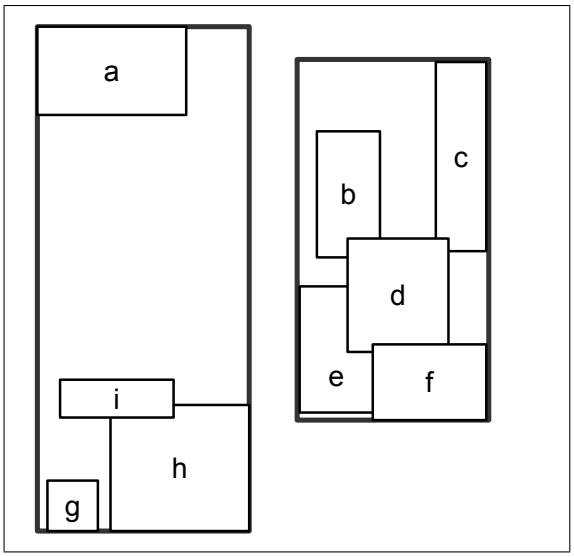
[43] *Modeling of Application Data with Spatio-temporal features.*



(a) Indexation par grille régulière.



(b) Indexation par arbre quaternaire (3 niveaux de profondeur).



(c) Indexation par arbre R^+, à l'aide des REMs.

FIG. 4.4 – Illustrations cartographiques de différentes indexations spatiales possibles.

quaternaire, qui consiste à subdiviser l'espace de référence dans les zones où la densité d'objets est grande (représentation simplifiée en figure 4.4(b)) . Au contraire, l'approche par arbre R^+ (figure 4.4(c)) est basée sur la deuxième idée directrice. Elle utilise une méthode de recouvrement par Rectangles Englobants Minimaux (REMs) afin de trier les données voisines ensemble.

Les SIG sont donc dotés d'une grande variété de modèles d'indexation, qui permettent des interrogations optimisées sur les données stockées.

4.3.4 Langage d'interrogation spatiale

La formalisation d'un langage dans les sciences de l'information géographique est un thème de recherche important des 20 dernières années. Le coeur de l'effort a été l'identification d'un répertoire de relations spatiales, et comment celles-ci se croisent. Historiquement cette recherche a pris trois voies distinctes selon Peuquet [Peu02]. La première utilise directement les principes d'algèbre et de géométrie, les meilleurs exemples étant les travaux de Tomlin [Tom83] sur l'algèbre cartographique. Une seconde voie s'est concentrée sur l'amélioration du langage de requête SQL, avec notamment les travaux d'Egenhofer sur les extensions du SQL, le PSQL, et sur le SQL3 par exemple. La troisième variante tourne autour du langage naturel dans le contexte des expressions spatiales, appuyé fortement par la recherche en cognition et en linguistique.

Une classification de ces langages est proposée par Aufaure et Trépied [APT96] et résumée dans le tableau 4.1. Les auteurs proposent deux grandes classes de langage, soit avec une approche textuelle, soit une approche graphique aussi dite par représentation. L'approche s'appuyant sur le langage naturel reste difficile car beaucoup d'ambiguïtés sont encore à résoudre au niveau de l'interprétation.

Les langages artificiels sont donc les langages les plus répandus. Ils nécessitent cependant un apprentissage et leur travers potentiel est de voir l'utilisateur passer plus de temps à identifier la commande qu'il va utiliser plutôt qu'à l'interrogation elle-même [Ege95]. Parmi ces langages, ceux correspondant à des variantes du SQL sont limités pour formuler des requêtes spatiales complexes ou symboliques. L'approche graphique permet au contraire d'exprimer ce genre de requêtes, en utilisant des composantes visuelles (des diagrammes, des icônes, des primitives graphiques) comme métaphores. Cette approche a connu une utilisation de plus en plus grande ces dernières années du fait qu'elles sont intuitives pour des non spécialistes et plus conviviales que les approches précédentes [Auf01].

4.3.5 Essais de SIG prenant en compte le qualitatif

Les données géographiques stockées dans un SIG sont des données structurées, définies de manière quantitatives, à l'aide de références numériques. Or la plupart des données géographiques existantes sont définies avec un certain degré d'incertitude. Il est par exemple difficile de définir la délimitation exacte de la zone géographique de « la vallée d'Ossau ». Nous pouvons aller plus loin en nous penchant sur les entités spatiales décrites dans des documents (textuels, images). Par exemple, l'entité « au sud de Pau »

			Exemple	Note	Problème
Langage de requête	Approche textuelle	Langage naturel		Favorable pour les utilisateurs	Beaucoup d'ambiguités à résoudre. Les requêtes peuvent être verbales et difficiles à interpréter
		Langage artificiel	TSQL, GISQL, Spatial SQL, SQL3	Extension au SQL pour permettre de traiter correctement l'espace et le temps	Manque de convivialité pour les utilisateurs
	Approche par représentation ou non-textuelle	Tabulaire	QBE	Requête par l'exemple	Exprimer les jointures
		Graphique	PQBE, SNAP	Meilleur usage du medium graphique	Concepts sous-jacents ne sont pas perçus dans un sens métaphorique
		Langage visuel	Spatial Query-by-Sketch		
Techniques hypermédias		Carte dynamique	Hypermaps Argumaps		

TAB. 4.1 – Classification des langages de requête d'après Aufaure et Trépied [APT96].

est *a priori* inquantifiable de par sa subjectivité (le sud de cette ville peut avoir plusieurs représentations dans l'esprit des gens) [Ess07] et par le caractère vague de la relation elle-même. Nous pourrions adopter un point de vue mathématique : le sud d'une ville (c'est à dire un polygone ou un point) correspond à tout ce qui se trouve en dessous de sa lattitude, comme « Los Angeles » par exemple. Mais dire que « Los Angeles » est au sud de « Pau » n'a pas de sens. Il est donc difficile de demander à un SIG de gérer ce genre de données de manière directement quantitative.

Des travaux existent néanmoins intégrant des méthodes de raisonnement qualitatif dans des modèles classiques ou proposant au contraire de nouvelles spécifications pour des SIG « qualitatifs » [Bit96, BS01, Bri98, Ben96, BCI97, Ben01, Wes03]. Ceux de T.Bittner [Bit96, BS01] par exemple proposent un modèle qualitatif composé d'une part de primitives spatiales, les *régions* représentées par des polygones et les *couvertures* qui sont des couches de *régions*. D'autre part, des relations topologiques sont définies entre les régions et les couvertures. Ce modèle spécifie la localisation d'une entité spatiale (sous forme de région) dans son environnement (sous forme de couverture). Ces structures ont pour avantage d'être aisément implémentées et stockées dans un SIG. Les travaux de B.Bennett [BCI97] montrent une implémentation de SIG dont les enregistrements consistent à des règles basées sur les relations du RCC (modèle présenté dans le premier chapitre des travaux existants). Par exemple, la règle $ec(island, sea)$ signifie que l'île a une relation de Connexion Externe avec la mer, c'est-à-dire que ces objets géographiques se touchent sans partager une surface d'intersection. Il est possible

d'imaginer un système alliant les fonctionnalités classiques et les outils manipulant les relations qualitatives utilisées par l'humain que sont la topologie, l'orientation et la métrie [KRR97]. A.Brimicombe [Bri98] propose une méthode de réduction du flou causée par l'imprécision de l'expression des entités spatiales. Pour cela il utilise, comme dans d'autres modèles existants, les ensembles flous, puis définit la notion d'*espérence floue*. Un ensemble flou A (dans l'univers U) ajoute la variable μ aux ensembles classiques, qui définit un degré d'appartenance compris entre 0 et 1. En effet les ensembles flous, en plus des éléments qui le définissent de manière sûre (comme pour les ensembles classiques) et que l'on nomme *noyau* (4.1), comportent des éléments leur appartenant plus ou moins sûrement et que l'on nomme *support* (4.2).

$$noyau(A) = \{x \in U / \mu_A(x) = 1\} \tag{4.1}$$

$$support(A) = \{x \in U / \mu_A(x) > 0\} \tag{4.2}$$

Dans ce contexte, A.Brimicombe définit l'*espérence floue* comme le pourcentage d'assurance qu'un utilisateur peut avoir en représentant une entité spatiale. Cette information est stockée en même temps que les autres caractéristiques de l'entité et sert ensuite pour l'appariement.

Enfin, M.Wessel [Wes03] propose d'intégrer aux méthodes d'indexation classiques des SIG un composant de *Description Logique (DL-Component)* qui s'occupe de stocker les relations d'ordre qualitatif entre les objets indexés.

Ces travaux traitent donc les relations topologiques entre les entités sans les interpréter de manière à leur donner une représentation quantitative potentielle. Il n'y a pas de proposition à proprement parler sur le calcul de représentation d'entité floue, complexe. La manipulation d'entités floues telles qu'elles sont exprimées dans du texte doit passer la plupart du temps par une double indexation, une pour les entités directement géo-référençables et une autre pour les relations spatiales. Des travaux seraient donc envisageables afin d'augmenter les fonctionnalités des SIG en amont de ceux présentés ici, pour proposer une indexation gérant les représentations quantitatives et qualitatives de la même façon.

Les SIG fournissent déjà des opérations testant les relations spatiales entre objets géo-référencés (comme les tests de contenance, de chevauchement, d'intersection, d'égalité, etc.) et analysant ces relations (comme le calcul de la surface d'intersection, la boîte englobante, la surface convexe, la différence, l'union, etc.), suivant la norme de l'OGC (Open Geospatial Consortium Inc.) [Ope99] (§ 2.2.1) qui elle même intègre les travaux décrits précédemment pour le raisonnement spatial qualitatif. Ces opérations peuvent alors aider à la construction d'index prenant compte d'entités décrites qualitativement.

4.4 Conclusion

Ce chapitre définit les SIG, leurs fonctionnalités, leurs cas d'utilisation, notamment dans le cas d'indexation spatiale. Nous avons vu comment sont gérées les données struc-

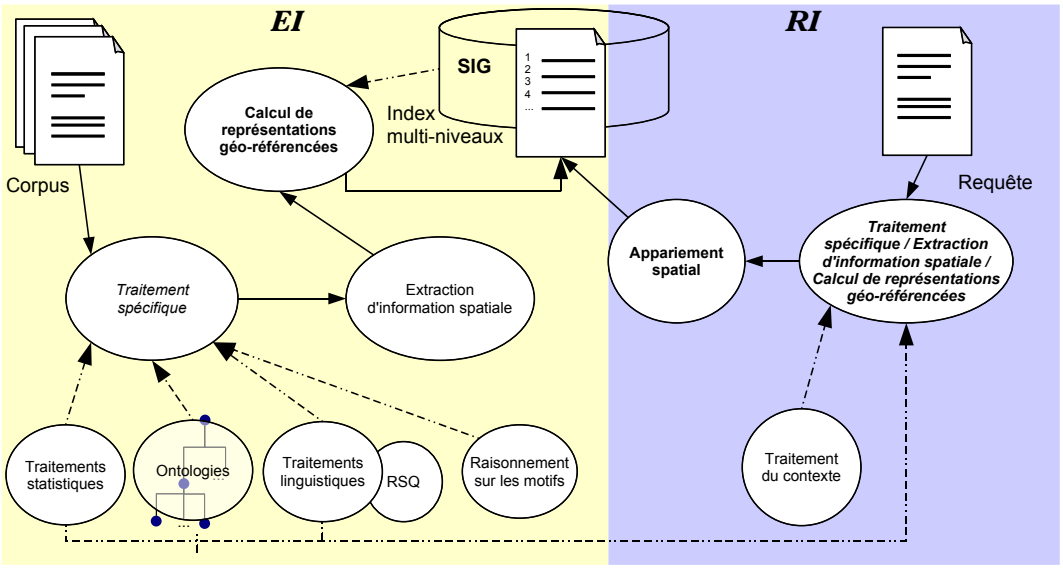


FIG. 4.5 – Processus d'extraction et de recherche d'information spatiale (3).

turées que ce type de système manipule et quelles sont les opérations qu'il doit pouvoir fournir.

Nous utiliserons dans la suite de nos travaux les données vectorielles du fait qu'il existe des fonctions de manipulation d'objets géo-référencés (comme les calculs d'intersection et de recouvrement) nécessaires à notre système de recherche d'information spatial. De plus, nous pourrons profiter de l'indexation optimisée pour ce type de données. Les données matricielles, au contraire, ne sont pas interprétées et ne sont donc pas structurées, ce qui limite leur utilisation.

Nous avons par ailleurs présentés des travaux essayant d'intégrer du raisonnement qualitatif dans les SIG. Leur méthodologie consiste à prendre des primitives spatiales et à définir des relations, en tenant compte de leur caractère *implémentable* dans ce genre d'outils. De plus des théories utilisées comme celle des ensembles flous semble pertinentes pour interpréter le raisonnement qualitatif exprimé en langage naturel.

De ces études nous retenons que les SIG seront nécessaires à l'élaboration de notre système de recherche d'information. Nous émettons l'hypothèse qu'ils pourront jouer le rôle de ressources ontologiques. Pour cela les couches de données stockées dans ces systèmes pourront représenter chaque niveau hiérarchique de l'ontologie géographique. Les relations seront alors définies implicitement *via* les opérations disponibles découlant du raisonnement qualitatif, de calculs d'intersection, d'inclusion, d'appartenance, de chevauchement, etc. Il n'est pas nécessaire dans ce cas, de stocker l'ensemble des relations existantes entre les entités stockées car le SIG pourra les calculer dynamiquement.

Des travaux existants utilisant les SIG dans un système de recherche d'information ont aussi été présenté : le projet Gipsy et le projet SPIRIT. Ceux-ci indexent leurs documents à l'aide de boîtes englobantes géo-référencées. Nous imaginons aussi utiliser

ce concept dans un premier temps, afin de valider nos hypothèses d'indexation, les boîtes englobantes étant considérées comme une bonne approximation de l'information spatiale, à un premier niveau.

Nous ajoutons à notre figure 4.5 de synthèse les modules qui font appel aux outils SIG et qui pourrons nous être utiles. Dans la phase d'extraction de l'information (à gauche), la validation des entités nommées et le calcul de représentations géo-référencées pourront avantageusement tirer partie de ces techniques, en les couplant au raisonnement qualitatif. Dans la phase de recherche (à droite), ils pourront être utilisés en particulier pour l'appariement spatial grâce aux opérateurs d'accès réalisant les tests de relations entre objets spatiaux.

Troisième partie

Contribution

Chapitre 5

Préconisations pour une recherche d'information spatiale

Sommaire

5.1 Introduction

Dans cette partie, nous décrivons les préconisations proposées, dans le cadre de notre thèse, à propos du traitement de l'information spatiale dans un système de recherche d'information. Elles serviront ensuite à introduire nos réalisations en chapitre 6 et 7.

Ces préconisations consistent principalement en un travail de modélisation de l'information spatiale et de conception d'un outil de recherche d'information spécialisé pour un corpus textuel. Parti du domaine de la recherche d'information, nous avons vu que dans le cadre d'une indexation spécialisée, les modèles classiques n'étaient pas adaptés. L'information que l'on veut rechercher dans les documents n'est pas exprimée par des termes considérés comme significatifs car par exemple ils sont parfois peu fréquents dans les documents. De plus les méthodes existantes en RI sont peu adaptées à la modélisation

des relations spatialement significatives inhérentes aux entités géographiques nommées (exprimant l'adjacence, la distance, l'inclusion, etc.) et sont pourtant nécessaires à l'élaboration d'un système de recherche spatial. D'autres domaines, notamment celui des SIG sont plus à même de manipuler ces relations, dans la mesure où ils sont structurés et décrits dans un format qu'ils reconnaissent.

Nous nous sommes alors d'abord penchés sur des travaux linguistiques et cognitifs dont la modélisation de l'expression spatiale est une fin en soi. Nous avons vu en particulier qu'une extraction partielle de la sémantique est possible en effectuant une recherche active, c'est-à-dire une recherche ciblée sur un domaine dont on a défini une structure préalable. Le raisonnement spatial qualitatif propose à ce propos des modèles exprimant les liens que peuvent avoir des objets spatiaux entre eux. Ces résultats sont alors utilisables dans un processus d'indexation par recherche active.

Nous avons enfin étudié les outils de gestion de l'information géographique, adaptés aux données géo-référencées, qui fournissent, d'une part, un moyen optimal d'indexation et, d'autre part, des fonctionnalités proches de celles qui sont nécessaires pour implémenter un modèle de RSQ.

Une indexation spatiale est alors possible pour un corpus de documents non-structurés. Cette indexation peut être construite à partir des syntagmes du texte contenant des entités nommées auxquels une représentation géo-référencée a été associée. Il en résulte une indexation très localisée au grain d'une ou de quelques phrases. Cette première extraction permet de reconstruire une indexation de fragments de texte plus importants, en liant les syntagmes entre eux à l'aide d'une modélisation basée elle aussi sur le RSQ. Cette indexation d'un grain plus élevé fait en quelque sorte la synthèse, du point de vue spatial, d'un fragment de texte. Des motifs spatiaux, correspondant à un contexte spatial exprimé par l'auteur, peuvent être définis. Dans le cas d'un corpus composé de récits de voyage, des motifs tels que « l'itinéraire », « la description d'un lieu » ou encore « la comparaison de lieux » peuvent être modélisés. Ces motifs peuvent servir ensuite à construire une indexation d'un niveau d'abstraction plus élevé. Compte-tenu que nous travaillons sur des données non-structurées, il convient de définir les « unités de texte » qui vont servir à regrouper des syntagmes spatiaux en motifs.

Le plan de ce chapitre est le suivant. Enrichis des travaux étudiés dans l'état de l'art, nous allons d'abord recadrer notre problématique. Nous proposons ensuite des modèles pour la définition d'information spatiale et expliquons comment ils permettent de concevoir une méthode d'indexation spatiale multi-niveaux.

5.2 Rappel et recentrage de la problématique

La problématique de départ était de trouver un moyen d'extraire l'information géographique d'un corpus textuel afin de construire une système de recherche d'information. Cette problématique de recherche d'information par le contenu faisant appel à de nombreux domaines possédants des techniques propres, nous devions réfléchir à un moyen de les coupler entre elles. La figure 5.1, qui reprend le schéma de l'introduction, fait la synthèse pour les principaux domaines, des méthodes les plus pertinentes par rapport à notre

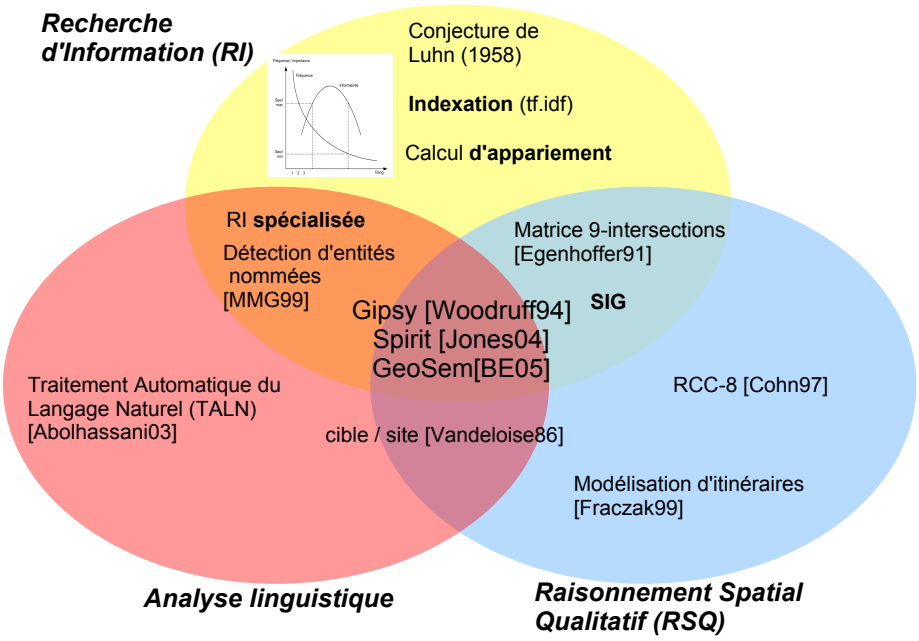
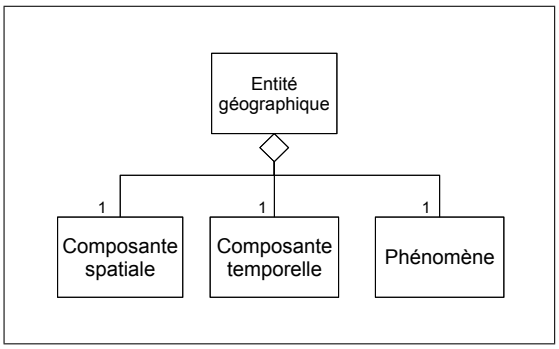


FIG. 5.1 – Intersection des trois principaux domaines d'intérêt

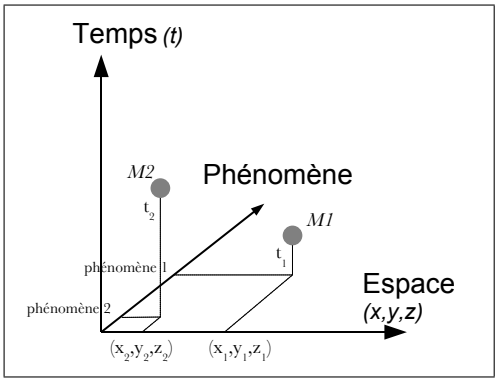
problématique. Nous nous sommes d'abord intéressés à la recherche d'information, qui ne répondait pas complètement à notre problématique. Nous avons alors étudié des travaux d'analyse linguistique à travers la recherche spécialisée. Cette analyse linguistique, dans le cas du spatial, s'est accompagnée d'une étude sur le raisonnement spatial qualitatif, pour revenir sur les systèmes de recherche d'information spatiale existants et les SIG. Nous nous trouvons donc à une intersection, tout comme le projet SPIRIT [JPR+02], le projet Gipsy [WP94] ou plus récemment le projet GeoSem [BE05].

Les questions qui se posent à nous sont donc les suivantes : Comment utiliser, par exemple, les outils linguistiques avec les outils de recherche d'information ? Comment construire un modèle linguistique qui soit opérationnel pour se coupler aux outils classiques d'indexation et de recherche d'information ? Comment apparier les documents avec les requêtes ?

Toutes ces interrogations nous ont amené à restreindre notre objectif de départ, concernant l'information que l'on veut réellement traiter dans un texte. Nous allons voir que la définition de l'information géographique que nous utilisons est une molécule composée de trois sous-molécules : l'information spatiale, l'information temporelle et le phénomène (figure 5.2). Un travail de modélisation sera fait ensuite sur la composante spatiale, sur laquelle nous concentrerons nos efforts. En effet, nous considérons cette composante comme la plus importante dans la mesure où elle guide les deux autres. Nous pensons en effet qu'un phénomène géographique peut être défini de manière minimale par sa composante spatiale, même si l'information temporelle est souvent liée à lui.



(a) Molécule définissant une Entité Géographique (EG) à partir d'une composante spatiale, d'une composante temporelle et d'un phénomène.



(b) Espace géographique avec 2 molécules $M1(x_1, y_1, z_1, t_1, ph_1)$ et $M2(x_2, y_2, z_2, t_2, ph_2)$.

FIG. 5.2 – Définition d'une entité géographique.

Les travaux existants montrent en outre la complexité d'interprétation de l'information spatiale et, contrairement à la composante thématique, il n'existe pas de propositions assez mûres pour être intégrées dans les systèmes de recherche d'information spécialisée. Enfin la problématique associée à l'interprétation temporelle est intéressante mais parait du même ordre que l'interprétation spatiale, tout en étant plus simple. De plus nous avons choisi de nous limiter à une seule composante, afin de pousser jusqu'à son terme le travail de modélisation et l'implémentation du processus d'indexation et de recherche. Enfin nos recherches sur les motifs se sont portées particulièrement sur les itinéraires, de par leur utilisation répandue dans notre corpus et de manière générale dans beaucoup de documents texte. En effet un itinéraire peut être considéré à lui seul comme une entité géographique. Cependant nous ne traiterons que de sa composante spatiale, la partie temporelle n'ayant pas été traitée au préalable. Nous justifions nos choix dans la section suivante qui détaille nos propositions de modèles.

5.3 Modélisation de l'information spatiale

Cette section décrit les différents modèles qu'il a été nécessaire de concevoir. Nous sommes partis de la modélisation de l'information qui nous intéresse, à savoir l'information géographique. Nous focalisant ensuite sur l'information spatiale (c'est-à-dire géoréférencable) nous avons défini un modèle servant de pivot dans un processus d'indexation des documents. À partir de ce dernier ont ensuite été définis des modèles d'un plus haut niveau d'abstraction permettant d'exprimer des motifs spatiaux.

5.3.1 Définition de l'Entité Géographique (EG)

L'information géographique peut être définie dans un espace à trois dimensions (figure 5.2(b)) comme une molécule formée d'une composante spatiale, d'une composante temporelle et d'une composante thématique ou phénomène (figure 5.2(a)) [Gai01, Mal03,

UTC04, PSA07]. Cette définition vient du monde des bases de données. Nous l'avons néanmoins retenu en faisant l'hypothèse que le même type d'information peut être retrouvé et extrait de données non structurées (documents textuels ou images).

Dans le cadre de nos travaux de recherche, nous proposons néanmoins une définition plus restreinte : les EG, auxquelles nous allons nous intéresser dans le corpus, possèdent forcément une composante spatiale (ES) explicite. Celle-ci consiste en une ou plusieurs entités nommées de lieux (*ville de Pau*, *les pics de la chaîne de la Maladetta*). Par contre l'ET, qui peut être une date, un intervalle de dates ou une période (*XVIII*^{e} *siècle*, *le 12 juin 1876*, *Le début des années 60*), peut être implicite, c'est-à-dire qu'elle n'est pas mentionnée directement dans le texte mais découle d'informations annexes. L'ET peut être éloignée des autres composantes formant l'EG (par exemple dans le cas d'un journal de bord, la date est marquée en début de paragraphe et le reste de l'unité de texte décrit des phénomènes se passant à cette date). L'ET peut donc aussi être associée à plusieurs EG, quand elle recouvre un paragraphe entier par exemple. Enfin le phénomène, ou composante thématique, correspond *a priori* à tout ce qui n'est pas spatial ou temporel dans le texte. C'est le sujet dont il est question à un lieu et un moment donné (botanique, thermes, etc.). Cependant, sa présence n'est pas obligatoire dans la molécule géographique dans la mesure où l'on considère que le thème prédominant peut être le phénomène géographique lui-même. En effet, dans le cadre de notre corpus, il arrive souvent que le sujet en question est l'itinéraire pris par l'auteur. La description de cet itinéraire peut alors être considérée comme un phénomène. La molécule géographique est formée dans ce cas par la représentation de l'itinéraire dans l'espace, dans le temps et dans la manière dont il a été réalisé [LGL06].

5.3.2 Modèle Pivot pour l'interprétation de l'information spatiale

D'après l'hypothèse de cible / site exposé dans l'état de l'art, nous considérons que l'information spatiale exprimée dans un texte est constituée d'au moins une entité nommée et d'un nombre variable d'indicateurs spatiaux, précisant sa localisation.

L'entité spatiale (ES) de la figure 5.2(a) est donc reprise dans notre Modèle Pivot (MP). Ce modèle, nommé pivot car il servira pour toutes les étapes d'indexation et de recherche d'information, comporte quelques similitudes avec les modèles existants, comme celui décrivant l'ontologie du projet SPIRIT (Chapitre 2, figure 2.4). En effet, il existe déjà cette notion d'entité définie par une ou plusieurs autres entités, que nous avons précisée à travers notre définition récursive présente dans le modèle. La récursivité lui confère la puissance d'expression nécessaire à la représentation d'entités complexes. Une entité spatiale définie dans notre modèle peut donc correspondre à l'une des deux options suivantes (figure 5.3) :

- une entité spatiale absolue (ESA), dans le cas où l'auteur exprime une information seulement à partir d'une entité nommée *(ex : la ville de Pau, Laruns)*. Ce sont en quelques sortes les primitives spatiales de notre modèle.
- une entité spatiale relative (ESR), dans le cas où l'auteur utilise en plus des entités nommées, des indications spatiale d'ordre topologique *(ex : près de Pau, au sud de Pau, à une heure de marche de Pau, entre Pau et Laruns)*. Nous appelons

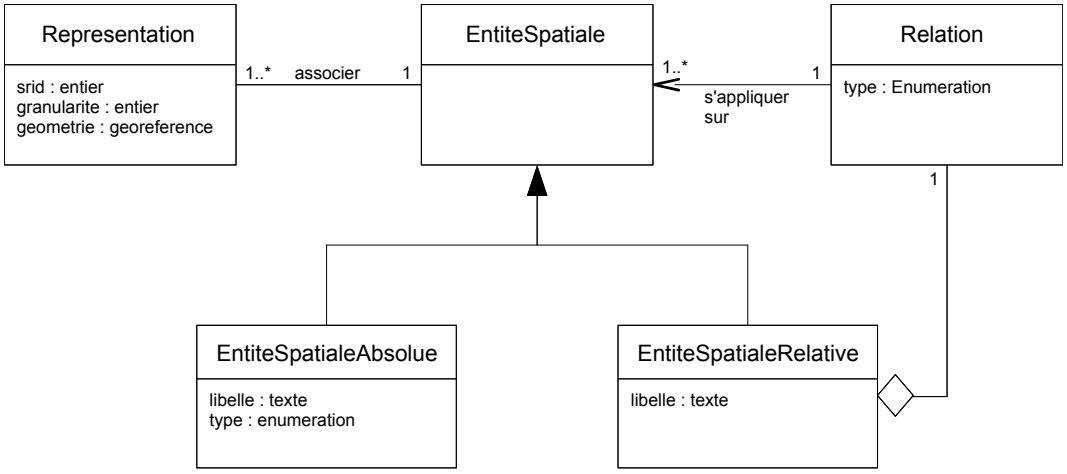


FIG. 5.3 – Définition du modèle pivot.

ces indications spatiales des *relations*. En nous basant sur les modèles de RSQ existants (voir la section 2.3.1 et [MGVOM07] pour des travaux plus récents) nous en définissons *a priori* cinq :

- l'orientation (au sud de),
- la distance (à 1 heure de marche de, à 20 km de),
- l'adjacence (près de, loin de, la périphérie de),
- l'inclusion (le quartier de, la frontière entre, le sommet de),
- l'union et l'intersection liants au moins 2 ES (entre A et B, le triangle A, B, C, à l'intersection de A et B, la frontière A-B, etc.)

Dans ce cas, une ES est donc définie par une de ces relations et au moins une autre ES (il peut y en avoir plusieurs dans le cas des relations d'union et d'intersection). La récursivité peut s'exprimer ainsi, si cette dernière ES est à son tour une ESR.

Cette notion de définition récursive est un atout de notre modèle du point de vue pragmatique mais aussi du point de vue cognitif, car l'interprétation qui en découle reste proche de l'expression de la spatialité dans du texte. À titre d'exemple, quand un auteur utilise une ES complexe, du type *au nord de la frontière franco-espagnole*, celle-ci est une ESR composée d'une relation d'orientation *(au nord de)* et d'une autre ESR *(frontière franco-espagnole)*, elle-même composée d'une relation d'inclusion *(la frontière)* et des ESA *France* et *Espagne*.

Le modèle ainsi présenté est capable d'interpréter la plupart des informations spatiales exprimées en langage naturel, utilisant toutes les relations qualitatives possibles. Nous associons ensuite à chaque ES au moins une *Représentation* (dont nous détaillerons la problématique associée dans la section 5.4). Cette représentation est géo-référencée et il peut y en avoir plusieurs, suivant les différentes échelles à laquelle est observée l'ES. En effet, une ville peut être représentée sous forme de point à l'échelle du pays, mais si le cadre d'étude est de l'ordre de la région, sa représentation correspondra plus à une forme polygonale. L'indexation est donc formée, pour chaque entité spatiale interprétée, d'une

structure correspondant à une instance du modèle pivot et d'au moins un représentation géo-référencée[44].

De la même manière, le modèle pivot est utilisé pour le processus de recherche d'information dans la mesure où les requêtes de l'utilisateur sont interprétées exactement de la même façon que les documents. La requête, qui peut être exprimée en texte libre ou en dessinant des formes géométriques sur une carte, par exemple, produit à l'aide du modèle pivot une ou plusieurs représentations géo-référencées qui sont mises en correspondance avec celles indexées pour les documents. L'appariement réalise alors un calcul sur les surfaces d'intersection afin de fournir un score de pertinence. Ces surfaces d'intersection peuvent aussi être considérées comme des instances du modèle pivot.

De plus, le fait d'utiliser le même processus d'indexation pour les documents et les requêtes permet d'imaginer une interrogation du système *via* un *document-requête*, à condition de formaliser les opérateurs logiques (*et*, *ou*, ¬) entre les entités extraites.

5.3.3 Indexation spatiale par motifs

Le modèle pivot est une première proposition pour l'indexation intraphrastique. Nous présentons maintenant une méthode d'indexation à un niveau plus élevé (de l'ordre de l'unité de texte) en regroupant les entités spatiales (ES) entre elles. L'hypothèse de cette approche est que l'analyse d'un groupe d'ES géo-référencées permet la déduction du type de contexte spatial (ou motif spatial) utilisé dans le récit, en particulier celui de la description d'un itinéraire. Cette analyse se fait en s'abstrayant de toute considération linguistique pure et considère que le *sens spatial* d'une unité de texte peut être reconstruit uniquement avec cet ensemble de zones géo-référencées égrenées dans un certain ordre le long du texte. Cette méthode, moins lourde qu'une approche purement linguistique permet d'envisager une utilisation dans un système d'indexation pour une application de recherche d'information spécialisée, tout en tenant compte des caractéristiques discursives employées par l'auteur.

Nous définissons maintenant la notion d'*unité de texte*. Puis nous présentons comment est conçue l'indexation contextuelle en faisant la *synthèse spatiale* d'unités de texte. Enfin nous parlons des *motifs spatiaux* définis dans le cadre de notre cas d'étude. Nous pouvons illustrer ces termes à travers un exemple concret d'indexation spatialement contextuelle (figure 5.4). L'auteur traite d'un voyage effectué depuis Paris pour arriver à la fin de l'extrait à Bordeaux. Les différentes entités nommées mentionnées par l'auteur sont extraites et positionnées sur une carte afin de reconstruire une représentation géo-référencée d'un itinéraire, pouvant servir par la suite dans notre système d'indexation spatial.

Unité de texte Définie à l'aide de règles typographiques, une unité de texte est un extrait d'ouvrage. Elle peut correspondre à un paragraphe, un groupe de paragraphes, une section, un chapitre et ainsi de suite jusqu'à l'ouvrage lui-même.

[44] Le modèle pivot a été implémenté sous forme de schéma XML disponible en annexe B.

Un motif faisant la synthèse spatiale est associé à chacune d'entre elles, à tous les niveaux hiérarchiques du document textuel. Ainsi le processus doit pouvoir faire une synthèse de plusieurs synthèses correspondant à des unités de texte de niveau moins élevées. Par exemple, à une unité de texte *chapitre* doit être associé un motif résumant les motifs des paragraphes compris dans celui-ci.

La difficulté à ce niveau est de découper convenablement le texte en unités qui font sens. Si un paragraphe ne contient qu'une entité spatiale et le suivant une dizaine, comment valuer l'importance du premier par rapport au deuxième ? Doit-on regrouper ces deux paragraphes pour ne former qu'une seule unité de texte ? Quelle règle d'arrêt utiliser si l'on prend des groupes de paragraphes comme unité de texte ?

De plus des règles doivent être définies pour faire la synthèse de synthèses : Comment résumer un chapitre contenant quelques paragraphes d'itinéraires et quelques paragraphes de point de vue ? etc.

Synthèse spatiale Le but est de cette synthèse est de récupérer pour une unité de texte un sens global, du point de vue spatial : décrit-il un itinéraire ? une comparaison ? ou bien fait-il une description locale ? Cette synthèse comprend donc une étape de *classification* et de calcul d'une *représentation géo-référencée*.

Pour cela nous voulons nous abstraire le plus possible d'indicateurs sémantiques dépendants du degré d'incertitude de la langue, et nous en tenir aux propriétés spatiales des entités spatiales extraites. En effet, notre hypothèse est que la disposition géographique des entités spatiales apporte une information plus facile à traiter que la sémantique du texte qui, surtout dans notre cas, peut s'avérer avoir un style très littéraire et mettre en défaut un traitement automatique trop généraliste. Cependant, des indicateurs sémantiques peuvent toutefois servir dans un deuxième temps pour compléter ou valider notre synthèse/classification.

Cette méthode prend à contre pied les méthodes classiques de synthèse de texte qui travaillent sur les mots en se servant de methodes d'apprentissage [Nen06, NVM06, ORK06]. Le fait de s'intéresser spécifiquement à l'information spatiale permet de s'abstraire du texte au moment de l'interprétation de son sens global.

Le résultat d'un tel processus n'est plus alors un texte reconstitué à partir de fragments de l'unité mais une représentation géo-référencée à laquelle on associe un motif discursif. La figure 5.4, par exemple, montre un extrait de texte et la représentation reconstruite à partir des syntagmes extraits. L'auteur part de Paris (première entité en haut à droite, correspondant à une étape du journal de voyage) pour arriver à Bordeaux (dernière entité de l'unité de texte). Des étapes intermédiaires à son voyage aident à reconstruire son itinéraire. La synthèse correspond alors au polygone géo-référencé reliant Paris à Bordeaux.

Nous allons voir maintenant quels sont les motifs que nous avons définis en nous basant sur notre corpus territorial.

Motif spatial Un motif spatial est un aspect du discours identifiable par des caractéristiques spatiales particulières. Des travaux ont montré que ces caractéristiques se

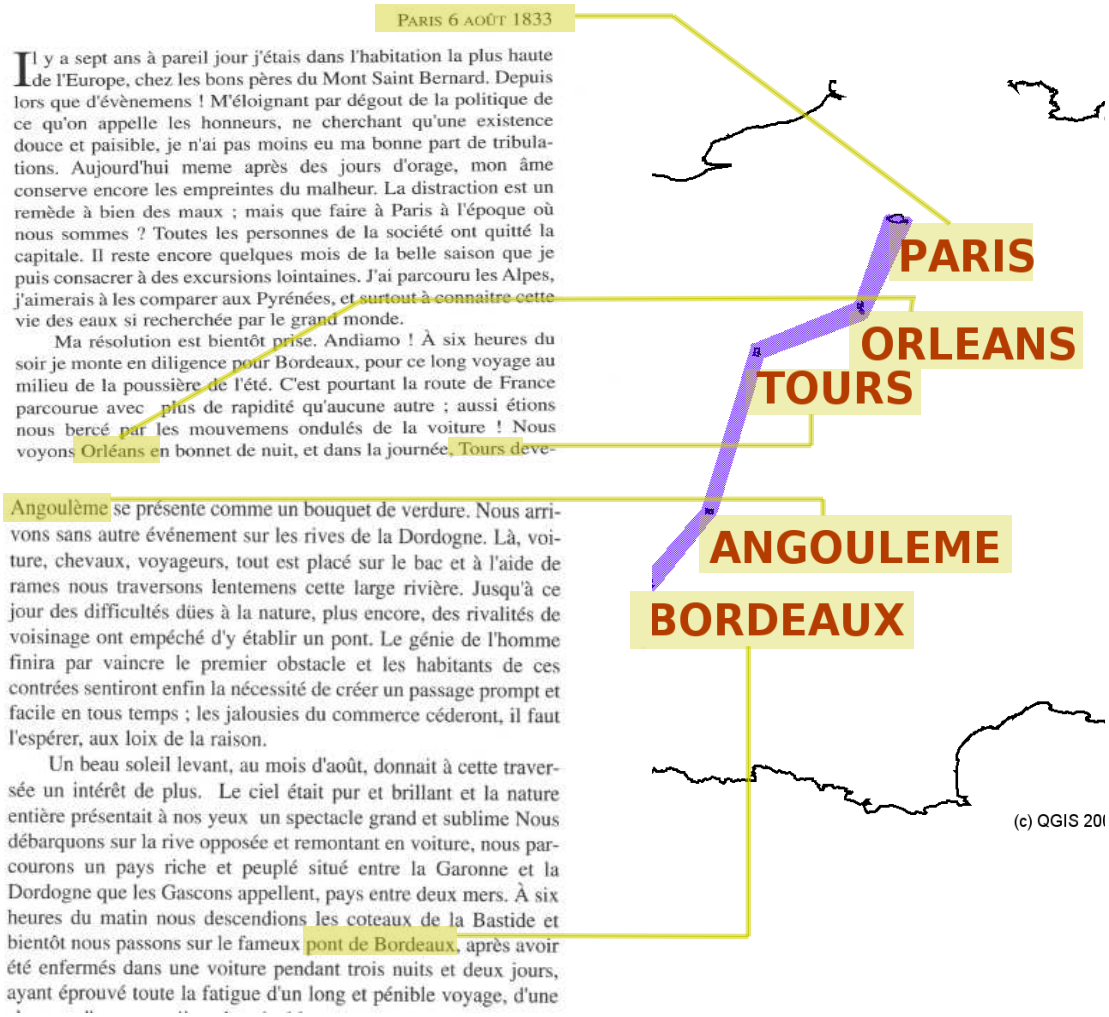


FIG. 5.4 – Synthèse d'une unité de texte en itinéraire et sa représentation géo-référencée.

traduisent par des aspects linguistiques eux aussi particuliers. B.Tversky [TT97], qui définit trois motifs de description (la description de trajet, par parcours du regard et la description en survol, mentionnés dans la section 3.2), étudie et catégorise par exemple ces aspects. Nous émettons l'hypothèse que d'autres aspects, géographiques cette fois, peuvent être dégagés pour la catégorisation.

Dans le cas de notre corpus, nous avons défini quatre motifs principaux :

– Le motif *itinéraire* (figure 5.5) correspond à la description par l'auteur d'un cheminement l'amenant d'un point A à un point B, A et B étant des lieux différents[45]. Le modèle UML montre qu'un itinéraire est défini comme une composition de

[45]La définition du Littré donne pour un itinéraire, une *« indication du chemin d'un lieu à un autre. Par extension, indication de tous les lieux par où l'on passe pour aller d'un pays à un autre. »*

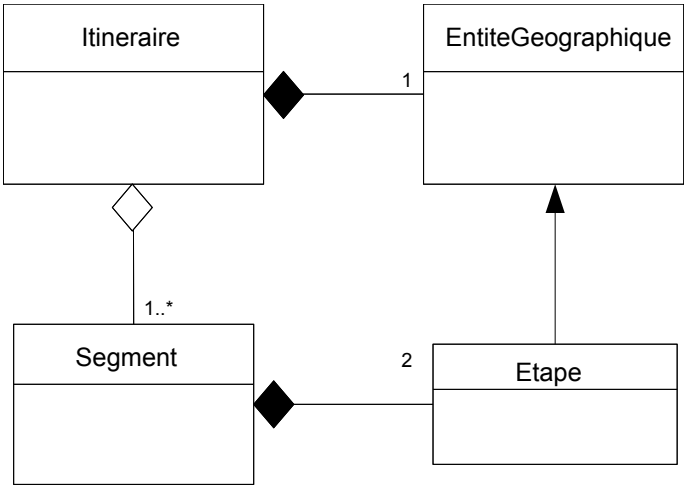


FIG. 5.5 – Modélisation d'un itinéraire à partir du modèle pivot.

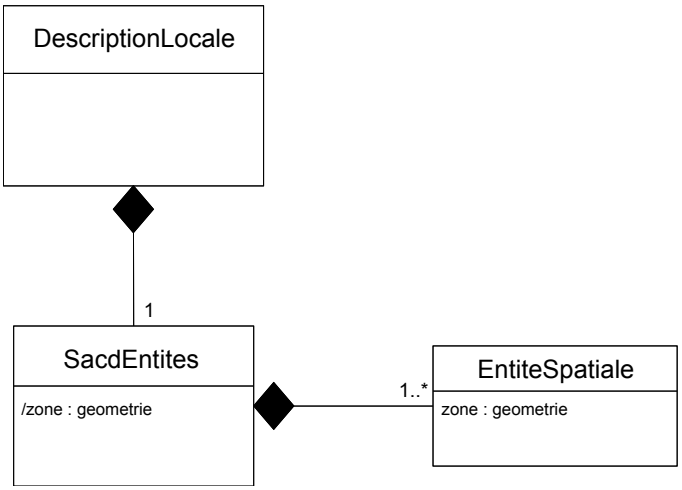


FIG. 5.6 – Modélisation d'une description locale à partir du modèle pivot.

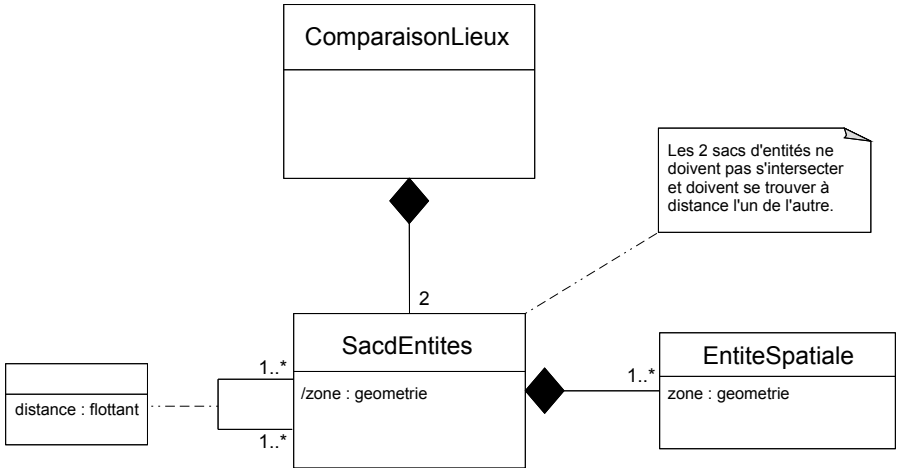


FIG. 5.7 – Modélisation d'une comparaison de lieux à partir du modèle pivot.

segments, qui sont des déplacements unitaires délimités par une *étape* à chaque extrémité. Cette étape est une spécialisation de l'entité géographique dans la mesure où l'itinéraire est considéré comme un phénomène spatio-temporel. À une étape peut en effet correspondre un lieu et une date. Un exemple de description d'itinéraire se trouve à la figure 5.4.

– Le motif *description locale* (figure 5.6) correspond à une description d'un lieu restreint, le locuteur étant dans ce lieu (*« Nous sommes à Gavarnie. [...] le cirque de Gavarnie me sembla admirable. [...] à la brêche de Roland. » Voyages aux Pyrénées, Pimientos, 2001*). Ce motif convient à une description faite sans mouvement de la part de l'auteur, contrairement au motif *itinéraire*. Le modèle le définit comme un sac d'entités réunies par des caractéristiques particulières.
– Le motif *description de point de vue* correspond à une description d'un paysage d'un point de vue surélevé (*« Le pont d'Espagne, la chute de Cerizey, le lac de Gaube, le glacier de Vignemale, quelles admirables choses », Voyages aux Pyrénées, Pimientos, 2001*). Il correspond à un *voyage virtuel* effectué par le locuteur. Le modèle 5.6 peut être utilisé pour formaliser ce motif, la différence avec la *description locale* se faisant dans les caractéristiques définies pour réunir les entités entre elles. En effet une différente disposition des entités, venant du point de vue surélevé, doit être observée.
– Enfin le motif *comparaison de lieux* (figure 5.7) correspond à évaluer un lieu en fonction d'un autre connu par l'auteur et supposé connu par le lecteur (*« Les Pyrénées sont taillés dans de plus petites proportions que les Alpes ; ses courses sont moins longues, moins pénibles qu'en Suisse, [...] » Voyages inédits dans les Pyrénées, PyréGraph, 2001*). Le modèle correspondant le définit comme une composition de deux sacs d'entités, la contrainte à vérifier étant que l'intersection des deux sacs soit vide, et que la distance les séparant soit suffisamment grande.

Cette classification en quatre motifs n'est pas arbitraire. Elle est le fruit de lectures

d'extraits de corpus, au cours desquelles nous avons essayé de définir les modèles présentés afin de clarifier la classification. L'étude menée sur des extrait de corpus a montré la présence de 36 itinéraires, 23 descriptions locales, 11 points de vue, 2 comparaisons de lieux sur les 76 paragraphes choisis (en fonction de la quantité d'entités nommées présentes). Seulement 4 paragraphes ne correspondent à aucun de ces motifs. Cet extrait de corpus servira pour l'expérimentation de l'indexation par motifs dans le chapitre 7.

Néanmoins il est possible de proposer d'autres classes, sachant que la définition se fait à une étape d'apprentissage au cours de l'indexation multi-niveaux et que c'est à cette étape que nous définissons le nombre et la nature des classes.

Pour chacun de ces motifs, nous avons analysé des caractéristiques propres aux ensembles d'ES :

- la dispersion ; les ES peuvent être contenues les unes dans les autres, être proches ou dispersées dans un espace géographique,
- l'ordonnancement ; les ES connexes dans le texte dessinent un itinéraire ou non dans un espace géographique,
- la saillance ; les ES connexes dans le texte forment des angles plats, obtus ou aigus dans un espace géographique.

De plus des propriétés intrinsèques aux ES comme leur échelle par exemple, sont utilisables pour le calcul des caractéristiques. Des caractéristiques essayant d'évaluer ces propriétés sont alors évaluées pour chaque unité de texte afin de le classifier, de lui donner une représentation géo-référencée et de l'indexer.

La figure 5.5 montre que le motif *itinéraire* est un cas particulier de l'indexation guidée par le contexte. En effet, il peut être considéré comme un phénomène spatial, mais aussi temporel. Il est composé d'étapes qui correspondent à des entités géographiques. Nous émettons alors l'hypothèse qu'il peut former une instance (un exemple) de molécule géographique complète. Cette hypothèse n'a pas encore donné lieu à une expérimentation mais semble prometteuse pour mettre en place un indexation multi-niveau de molécules géographiques.

5.4 Problématique de la représentation géo-référencée pour l'indexation

Nous avons vu que l'indexation repose sur le modèle pivot (figure 5.3) et qu'une représentation, au moins, est associée à chaque instance du modèle. C'est cette représentation qui est en fait au cœur de l'indexation. Plusieurs solutions s'offrent alors à nous pour la calculer. Elle peut consister en un positionnement dans une ressource ontologique ou en une géo-référence, produite grâce à des ressources géographiques (de types couches de SIG ou gazetteers). Ces dernières ressources fournissent des objets géo-référencés comme des *points*, des *lignes* ou des *polygones*. Dans ce cas, le calcul nécessite aussi, en ce qui concerne les ES_R, une méthode exprimant les relations définies dans le modèle pivot. Une problématique se dégage donc aussi autour du choix des primitives géométriques à associer. Enfin cette indexation particulière amène de nouveaux champs exploratoires

```
Laruns
  adjacence
    commune
      Béost
      Eaux-Bonnes
      Louvie-Soubiron
    col
      de Sieste
    gorge
      du Hourat
    pic
      de la Gentiane
      Lorry
```

FIG. 5.8 – sous-branche d'ontologie pour l'adjacence à Laruns.

autour du calcul d'appariement et du calcul de pertinence basés sur ces représentations spatiales.

5.4.1 Méthodes d'indexation disponibles

Nous classons les manières de construire un système d'indexation selon 3 méthodes intégrant plus ou moins les outils dédiés à la géomatique.

- La première consiste à utiliser des ressources de type ontologique. Par exemple pour le syntagme spatial extrait « près de Laruns », l'indexation va consister à récupérer une sous-branche d'ontologie (figure 5.8). L'index final est composé de listes de termes.
- La deuxième méthode consiste à utiliser une technique de calcul de représentation géo-référencée des entités utilisant le modèle pivot. La représentation correspond alors à une primitive géométrique (point, ligne, polygone). Pour le cas des ES_R, les relations sont interprétées par un opérateur adéquat qui transforme l'entité originale. La représentation de *Pau*, par exemple, est un polygone de taille x. Cette valeur x peut alors servir à calculer les différentes représentations résultantes des relations d'orientation, d'adjacence, d'inclusion, de distance et d'union. L'adjacence à *Pau* (*les environs de Pau, autour de Pau, près de Pau, etc.*) est donc par exemple représentée par un polygone de taille « x + *pourcentage de* x ». Cette méthode associe donc une seule géométrie précise pour chaque entité spatiale relative. Cependant, d'autres approches sont envisageables. Au lieu d'associer une seule géométrie, il est possible d'en associer plusieurs, avec des poids d'interprétation différents. Ces différentes géométries sont alors utilisées dans un calcul d'appariement spécifique, capable par exemple, de prendre en compte la géométrie la plus large en cas de trop peu de résultats retournés.
  Ces représentations géo-référencées servent ensuite à récupérer, par calcul d'intersection, les entités nommées se trouvant à l'intérieur et stockées dans une couche SIG (ou d'autres ressources). L'indexation consiste alors seulement à stocker ces listes d'entités nommées pour chaque entité extraite. L'index final est lui aussi

composé de listes de termes.
– La dernière méthode indexe la représentation géo-référencée des entités spatiales. L'indexation ne se fait donc plus sur des termes. L'index est alors composé de figures géométriques, ce qui permet l'ajout de nouvelles ressources pour les entités nommées sans faire de re-indexation totale. En effet, contrairement à la deuxième méthode qui fonctionne avec une liste statique et qui nécessiterait de mettre à jour l'index, celle-ci ne nécessite aucune action.

5.4.2 Géométries disponibles pour les représentations

La représentation associée à chaque entité spatiale est une forme géométrique dont les coordonnées se trouvent dans un référentiel terrestre telles que la projection Lambert II étendu ou la projection Lambert93 pour la France, WGS84 pour le monde, etc.

Plusieurs primitives géométriques peuvent être définies, l'idée étant d'associer les formes les plus appropriées en fonctions des entités spatiales et de l'échelle dans laquelle celles-ci sont observées. Ainsi une ligne brisée est adaptée à un cours-d'eau si l'échelle est suffisamment élevée. Une ville peut être représentée par un polygone si l'échelle est basse ou par un point si elle est vue de plus haut. D'autres formes comme les cercles peuvent être une bonne approximation de la réalité (pour exprimer la distance à une ville par exemple, comme *à 10 km de Pau*) mais les formes les plus facile à coder dans les outils SIG sont principalement les points, les lignes, les poly-lignes, les polygones, les multi-polygones[46]. Enfin, une dernière forme utilisée est la boîte englobante, dont une définition à déjà été donnée dans le chapitre 4.

5.4.3 Calcul d'appariement pour la phase de recherche

Du fait de nos choix d'indexation, basés sur des représentations géo-référencées, nous ne pouvons donc plus utiliser directement les méthodes classiques de recherche d'information (comme les modèles vectoriels). Nous proposons un appariement fait sur un pourcentage de recouvrement entre les représentations de l'index et les représentations de la requête, ces représentations pouvant être directement obtenues à partir d'un dessin sur un fonds cartographique ou interprétées à partir d'un texte libre grâce au même processus que celui utilisé pour les documents du corpus. Plus le recouvrement est grand, plus les fragments de documents correspondants seront considérés comme pertinents. Par exemple, à la requête *sud de Pau*, un fragment contenant la ville de *Gan* sera retourné avec un indice de pertinence plus élevé qu'un autre contenant la ville d' *Oloron-Sainte-Marie*, plus éloignée et décalée vers l'ouest. Le figure 5.9 montre un exemple de recouvrement avec 2 représentations possibles de la requête (en trapèze et en boîte englobante). Pour ces 2 représentations la surface commune est plus importante pour *Gan* que pour *Oloron-Sainte-Marie*. Cette figure montre aussi qu'un simple calcul d'intersection est insuffisant pour répondre à un problème d'appariement. En effet, si nous prenons l'exemple du département des *Pyrénées-Atlantiques* s'intersectant aussi avec la requête,

[46] Rien n'empêche cependant de coder un cercle à l'aide d'un polygone.

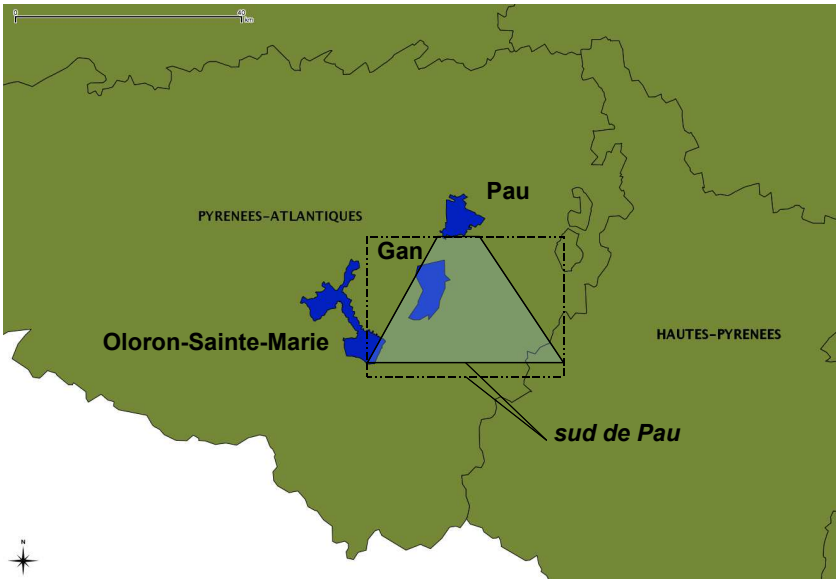


FIG. 5.9 – Deux représentations possibles du *sud de Pau* et illustration de deux entités nommées plus ou moins pertinentes *Gan* et *Oloron-Sainte-Marie*.

nous voyons qu'un fragment contenant cette entité sera aussi considéré comme pertinent, voire plus pertinent que la ville de Gan car le recouvrement sera encore plus important. Or, il ne semble pas acceptable de retourner un fragment contenant le département avant un autre contenant l'entité *Gan*, d'après la figure.

C'est pour cela que des calculs plus fins doivent être imaginés en fonction des primitives géométriques choisies. Nous détaillerons le calcul proposé pour le cas d'utilisation de boîtes englobantes dans le chapitre suivant.

5.5 Conclusion

Ce chapitre présente les préconisations pour une méthode d'indexation multi-niveaux et de recherche d'information spatiale. Il décrit les modèles et les choix pris au cours de notre recherche. Le cœur de notre travail est donc constitué de ces modèles et de leur utilisation dans la plate-forme PIV[47]. Ces modèles basés sur des travaux de raisonnement spatial qualitatif sont alors rendus opérationnels grâce aux outils de traitement automatique de la langue, dans un premier temps, puis aux outils de géomatique dans un second temps.

Le modèle définissant l'entité géographique (figure 5.2) est la base de notre réflexion sur la manière d'appréhender l'information. Déjà existant de manière similaire par ailleurs, nous proposons de l'utiliser comme modèle pour une indexation géographique des documents. Il conviendrait alors de modéliser chacune des sous-entités qui la

[47] pour *Pyrénées Itinéraires Virtuels*

composent. Cependant nous nous sommes restreint à la modélisation de la sous-entité spatiale. Le modèle pivot ainsi proposé (figure 5.3) conçoit l'interprétation de l'information spatiale de manière simple et pragmatique afin de faciliter son utilisation dans un processus opératoire. L'algorithme récursif qui en découle (algo. 1) permet en effet de fournir une méthode de construction d'index générique (à condition d'avoir les ressources géoréférencées). Les modèles proposés ensuite, définissant des motifs discursifs (figures 5.5, 5.6, 5.7), répondent à une problématique d'interprétation poussée de l'information spatiale présente dans les documents, à un niveau plus élevé dont la portée dans le texte est plus grande que celle de la phrase. Ils permettent la construction de caractéristiques nécessaires au processus d'indexation par classification. Ils permettront aussi, quand les mécanismes d'interprétation et d'indexation pour les composantes temporelle et thématique seront réalisés, de structurer les index multi-critères qu'il faudra concevoir. Par exemple, le motif *itinéraire* comportant aussi bien des connotations spatiales que temporelles et thématiques, fournira une structure interprétant la molécule géographique dans son ensemble, adaptée pour l'indexation.

Les deux chapitres suivants détaillent les implémentations de prototypes suivant ces préconisations et leurs évaluations. Le chapitre 6 décrit le prototype PIV qui implémente toutes les étapes d'indexation au niveau syntagme et les étapes de recherche d'information. Le chapitre 7 traite, quant à lui, de premiers travaux d'implémentation concernant l'indexation par motifs, à un niveau supérieur dans la hiérarchie du texte.

Chapitre 6

Système d'information spatiale pour les corpus territorialisés

Sommaire

6.1 Introduction

Ce chapitre présente notre contribution majeure concernant la création d'une plate-forme de recherche documentaire spécialisée dans l'information spatiale[48]. Cette plate-forme a pour vocation d'être une base ouverte et modulaire permettant de tester l'ensemble du projet DESI[49] de recherche documentaire : extraction d'information perti-

[48] le prototype a été baptisé "PIV" pour Pyrénées Itinéraires Virtuels.

[49] L'équipe-projet DESI, pour Document Électronique, Sémantique et Interaction, est le groupe dans lequel nous avons travaillé durant la thèse. Le site de l'équipe est disponible à l'adresse : http://liuppa.univ-pau.fr/DESI/

nente, indexation et processus de requêtage [EML05, EL05, EML05, LL06]. L'architecture et les fonctionnalités de la plate-forme seront exposées dans une première section. Nous proposons dans une deuxième section une méthode d'extraction et d'indexation d'information spatiale, basée sur notre modèle unifié et sur un traitement sémantique spécifique d'extraction spatiale [LGL06]. Nous traitons ensuite, dans une troisième section, de notre méthode de recherche d'information utilisant les fonctions SIG pour calculer des représentations géo-référencées et pour implémenter un calcul de pertinence spatiale [SBLG07a]. Une quatrième section est dévouée à l'évaluation de ce prototype aux différentes étapes du processus. Nous concluons ce chapitre en résumant les apports réalisés dans le cadre de cette plate-forme et les résultats de l'évaluation. Nous détaillons les limites qui apparaissent et les améliorations à apporter dans le futur.

6.2 Plate-forme PIV

L'architecture choisie par PIV est une architecture distribuée et modulaire, basée sur un ensemble de services web[50] effectuant chacun une des multiples tâches du processus d'extraction et de recherche d'information. Nous souhaitons en effet mettre en place une plate-forme pour la conception d'un Système spécialisé de Recherche d'Information dont les différents modules peuvent être échangés facilement et localisés en différents endroits afin de capitaliser plusieurs efforts de recherche. Nous fournissons chaque brique de PIV en tant que service web et pouvons accueillir une brique extérieure sous forme de service web. Par exemple, un module interne utilise l'analyseur morpho-syntaxique *Tree-tagger*[51]. L'architecture est faite de manière à pouvoir remplacer cet analyseur par un autre plus performant, le cas échéant. Différents outils, comme Tree-Tagger, ont été utilisés durant l'implémentation, recomposés sous forme de services web, au niveau du traitement sémantique, au niveau de l'indexation et de l'appariement spatial.

Pour le traitement sémantique, nous nous sommes essentiellement basés sur la plate-forme Linguastream[52] [Bil03, Bil06], basée sur des modules de traitements atomiques et utilisant le langage XML pour la structuration du texte. Chacun des modules nécessaires à notre traitement particulier a été transformé en service web. Un gestionnaire de base de données XML (eXist[53]) est utilisé pour stocker les fichiers nécessaires à l'indexation. Au niveau de l'indexation et de l'appariement spatial, nous avons utilisé un SIG (PostGIS[54]) et des couches de données provenant de l'IGN et de nos index. Enfin, une interface web fait appel aux différents services, jouant le rôle de chef d'orchestre de notre application.

Notons que d'autres interfaces ont été implémentées, comme des interfaces cartographiques utilisant l'API GoogleMaps (`http://maps.google.fr/maps`) et GoogleEarth (`http://earth.google.fr/`) afin d'enrichir le prototype en visualisant les résultats de manière adaptée et ainsi montrer la faisabilité d'un système complet de recherche d'in-

[50] Une liste exhaustive de ces services web se trouve en annexe.
[51] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
[52] `http://www.linguastream.org/`
[53] `http://exist.sourceforge.net/`
[54] `http://postgis.refractions.net/`

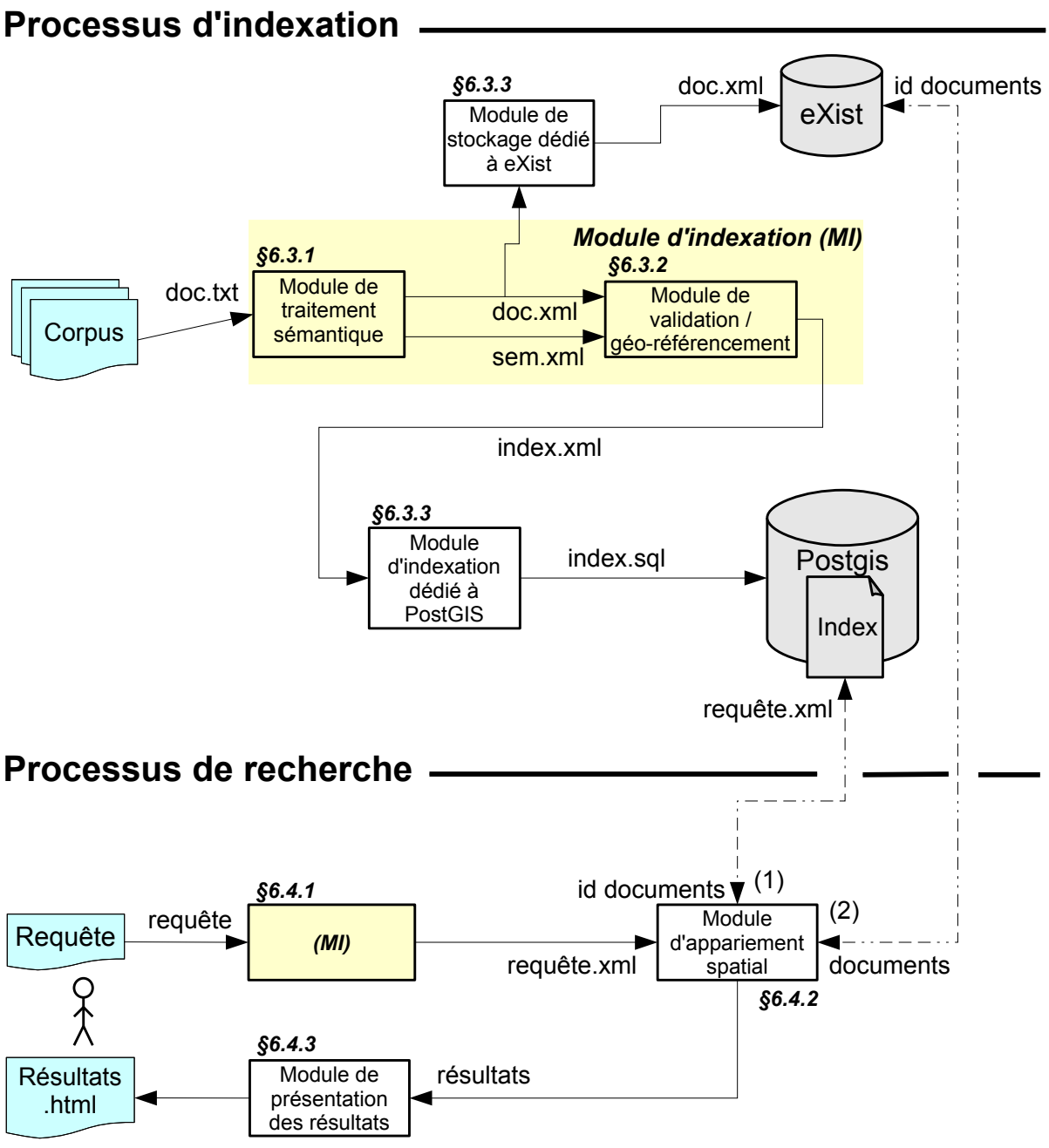


FIG. 6.1 – Schéma fonctionnel du système PIV.

formation spatiale basé sur nos index. Nous présenterons brièvement ces interfaces.

La figure 6.1 illustre les deux processus globaux d'indexation et de recherche d'information spatiale. La partie du haut correspond à l'indexation, c'est-à-dire au traitement effectué au préalable sur le corpus afin de construire le système de recherche d'information. Celui-ci est soumis d'abord à un traitement sémantique, puis à un module de validation et de géo-référencement. Ces deux traitements seront utilisés également dans la processus de recherche, pour interpréter la requête. Le résultat du processus d'indexation d'un document consiste en un stockage du texte marqué dans une base XML d'une part et, d'autre part, en un stockage de l'index à proprement parlé sous forme d'une liste d'identifiants *nom du document*, *numéro de paragraphe* auxquels sont associées des zones géo-référencées.

La partie du bas correspond au fonctionnement du système une fois mis en place. Un utilisateur pose une requête spatiale, laquelle est transformée en zone géo-référencée qui permet au module d'appariement de récupérer les fragments de documents les plus spatialement pertinents. Un dernier module s'occupe de visualiser les résultats proposés par le système.

Nous détaillons certains de ces modules correspondant à notre travail principal (les modules de traitement sémantique, de validation, de géo-référencement et d'appariement spatial) dans les sections suivantes.

6.3 Système d'extraction et d'indexation d'information

Cette section expose nos propositions dans le cadre de l'extraction et de l'indexation d'information spatiale depuis des documents textuels non-structurés. En particulier, nous utilisons le modèle pivot détaillé dans le chapitre précédent. Nous présentons donc comment il est utilisé pour le traitement sémantique, puis le géo-référencement de l'information spatiale extraite et son indexation [SGLL07, GSE+07].

6.3.1 Traitement sémantique associé au modèle

Une chaîne de traitement composée de modules de traitement sémantique a été réalisée en s'appuyant sur le modèle pivot [Lou05]. Ces modules sont relativement simples et leur capacité d'exhaustivité dans l'interprétation est en deçà de celle du modèle pivot. En effet, malgré l'apparente simplicité du modèle, un traitement complètement automatique, tel qu'il a été implémenté, ne peut extraire l'ensemble des entités spatiales définies par le modèle. Par exemple, seulement le traitement des syntagmes nominaux est pris en compte. La phrase « j'ai quitté Paris », syntagme verbal, n'est dans ce cas pas interprétée convenablement. De plus, seulement les indications spatiales placées avant l'entité nommée sont prises en compte. La phrase « la ville de Pau et ses alentours » est également mal interprétée. Ces modules permettent cependant de valider les choix faits à propos des analyses linguistiques complètement automatiques. Une expérimentation sera menée plus tard dans ce chapitre afin de déterminer le pourcentage de perte d'ES.

La figure 6.2 décrit les différents modules utilisés. Le module *TextToXML* prend en

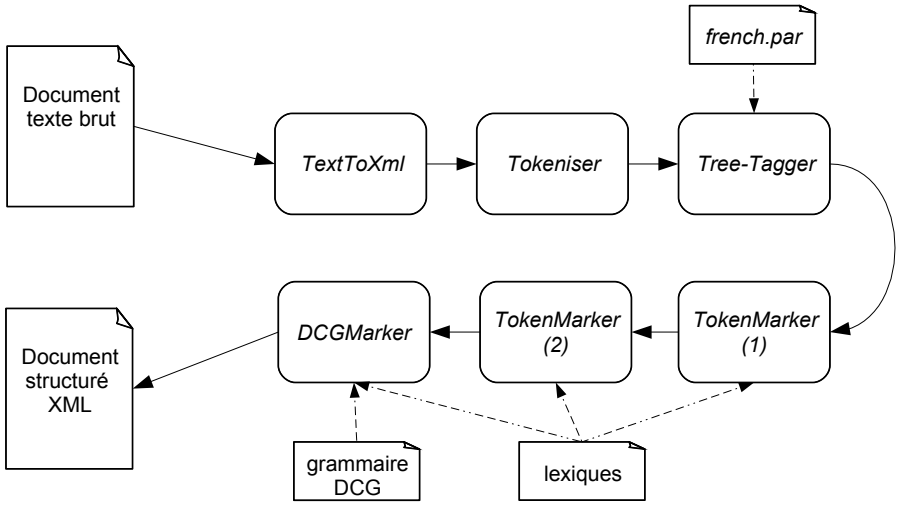


FIG. 6.2 – Modules de la chaîne de traitement extrayant les entités spatiales candidates.

entrée un fichier texte et fournit en sortie un fichier XML qui marque la structure du document. Ce module détecte par exemple les paragraphes et les numérote. Il est à noter que le texte en entrée est brut, sans récupération d'aucune structure ni de la taille de la police originale. Il est donc difficile d'avoir de meilleurs résultats que des « paragraphes marqués » grâce aux doubles retours-chariots récupérés par la reconnaissance optique de caractères.

Le module *Tokeniser* se charge de la segmentation du texte. Il prend en entrée le texte à traiter structuré sous forme d'unités de texte et produit un nouveau fichier XML dans lequel chaque mot (ou *token*) est identifié et est isolé par une balise. Par exemple, le traitement s'appliquant à la phrase « Le berger passe près de Laruns. » donne :

```
<doc>
    <ls:b type='token' id='1'>Le</ls:b>
    <ls:b type='token' id='2'>berger</ls:b>
    <ls:b type='token' id='3'>passe</ls:b>
    <ls:b type='token' id='4'>près</ls:b>
    <ls:b type='token' id='5'>de</ls:b>
    <ls:b type='token' id='6'>Laruns</ls:b>
    <ls:b type='token' id='7'>.</ls:b>
</doc>
```

Le module *Tree-Tagger* procède ensuite à l'analyse morpho-syntaxique des *tokens* précédemment isolés. Il correspond à une encapsulation du logiciel *Tree-Tagger*[55] utilisé avec le fichier de lexique et de paramétrage pour le français *french.par*. Il complète les balises par des informations sur la nature des termes, le type et le lemme associé. Pour le terme « passe » de l'exemple précédent, ce traitement donne :

```
<doc>
...
```

[55] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

```
3       <lss:sem type='token' id='3'>
4       <lss:value>
5           <tag>ver</tag>
6           <stag>pre</stag>
7           <lemma>passer</lemma>
8       </lss:value>
9   ...
10  </doc>
```

Deux modules *TokenMarker* se succèdent ensuite. Le premier a pour but de créer un squelette vide alors que le second complète ce squelette lorsqu'une entité nommée spatiale candidate est détectée. Ces deux modules fonctionnent avec des expressions régulières et sur le concept de détection de patrons, à l'aide de lexiques (voir l'annexe C). Ils prennent donc en entrée un fichier ressource contenant les expressions régulières qui formalisent les motifs à détecter et leur associent les balises à produire lors de la détection.

Ainsi le premier module prend en entrée une expression régulière de la forme $.*$ et une structure à produire de la forme *entite_nommee : non*. Cela a pour effet de marquer tous les termes avec les balises $<$ *entite_nommee* $>$ *non* $<$ */entite_nommee* $>$. Le deuxième module prend en entrée l'expression régulière $[A-Z].+$ permettant de détecter les candidats (en sélectionnant les termes commençant par une majuscule et suivis au moins d'une lettre). La structure produite est alors de la forme $<$ *entite_nommee* $>$ *oui* $<$ */entite_nommee* $>$. Le terme « Laruns » de l'exemple précédent donne :

```
<doc>
...
    <lss:sem type ="token" id="6">
        <lss:text>Laruns</lss:text>
        <lss:value>
            <tag>nom</tag>
            <stag>pro</stag>
            <lemma>Laruns</ lemma>
            <entite_nommee>oui</entite_nommee>
        </lss:value>
    </lss:sem>
...
</doc>
```

Le module *DCGMarker*[56] a pour but de détecter les indications spatiales voisines des entités nommées. Il est au coeur du traitement sémantique. Son objectif est de créer des instances du modèle pivot pour chaque syntagme contenant une information spatiale. Pour cela, il doit extraire toutes les indications spatiales se trouvant à proximité de ces entités nommées dans le texte et correspondant aux 5 relations définies dans notre modèle pivot. Pour cela le module utilise la grammaire présente en annexe D.

Cependant, nous avons tenu compte du contexte territorial de notre corpus. Nous avons en particulier omis des relations qui nécessiteraient une connaissance du contexte très précise pour être interprétées correctement (*vers quelle direction se tourne le regard de l'auteur, etc.*), sachant que l'interprétation linguistique serait très compliquée. Par exemple, les lexiques (annexe C) chargés de récupérer les indications spatiales connexes aux entités ne contiennent pas les relations d'orientation de type « à gauche de », « en

[56] Ce nom vient du module présent dans *LinguaStream*.

```
root(X) --> ega(X).
ega(es:X)-->es(X).

%----------
% Une es peut être absolue ou relative
%----------
es(es_r:X) --> es_r(X).
es(es_a:X) --> es_a(X).

%----------
% Définition d'une ES_R
%----------
es_r(name:label..relation:X..es_r:ES_R) --> relation(X), es_r(ES_R).
es_r(name:label..relation:X..es_a:ES_A) --> relation(X), es_a(ES_A).
```

FIG. 6.3 – Extrait de la grammaire : définition d'une ES comme étant une ES_A ou une ES_R , ES_R qui est ensuite définie.

dessous de », Nous avons donc restreint nos relations à des indications spatiales inhérentes à notre corpus, c'est-à-dire propres à un lieu géo-référencé et à des relations définies dans un repère plus général.

Le module effectue alors, pour chaque ES candidate extraite, une analyse basée sur des faits et des règles et remplit le modèle pivot en détectant de quel type sont les ES. Une grammaire réalise cette opération (disponible en annexe D). Elle définit une ES comme étant soit une ES_A, soit une ES_R. L'ES_R est à son tour définie comme étant composée d'une relation suivie soit d'une autre ES_R, soit d'une ES_A (figure 6.3). Ces définitions constituent les règles. Les faits sont les éléments terminaux définis dans un lexique (annexe C). Par exemple, pour la relation d'orientation, les faits définis sont les introducteurs « nord », « sud », etc. Les cinq relations sont ainsi définies. Pour chacune d'entre elle, des introducteurs sont spécifiés (figure 6.4).

Nous obtenons donc en sortie de cette première étape des ES candidates structurées au format XML (figure 6.5), qu'il convient de valider en leur associant des représentations. La figure 6.5 est un extrait contenant la structure pour l'ES_A *la ville de Lourdes* et l'ES_R *six kilomètres de Pau*.

```
%----------
% Les 5 relations définies dans le modèle pivot
%----------
relation(orientation:X) --> orientation(X).
relation(distance:X) --> distance_oriente(X),!.
relation(distance:X) --> distance(X).
relation(adjacence:X) --> adjacence(X).
relation(inclusion:X) --> inclusion(X).
relation_naire2(figure_geo:X) --> figure_geo2(X).
relation_naire3(figure_geo:X) --> figure_geo3(X).

%----------
% Relation d'orientation
%----------
intro_orientation(P)-->ls_token(_,intro:type:orientation..intro:stype:S,
    token),{P:=(pt_card:S)}.
orientation(O) --> prepOUprepart, intro_orientation(O).
orientation(O) --> prepOUprepart, intro_orientation(O), prepOUprepart.
```

FIG. 6.4 – Définition des cinq relations et détail des règles définies pour l'orientation. Les introducteurs d'orientation *intro_orientation* sont définis dans le lexique. La relation *prepOUprepart*, utilisant les résultats de l'analyse morpho-syntaxique, définit la liste des prépositions ou des articles pouvant se trouver avant et après l'introducteur d'orientation.

```
<?xml version="1.0" encoding="UTF-8"?>
<index xmlns="http://www.univ-pau.fr/~jlesbegu/specifs/piv" [...]>
  <doc doc_original="Voyages_aux_Pyrenees.txt"
    ls_doc="Voyages_aux_Pyrenees_doc.xml"
    doc_ocr="Voyages_aux_Pyrenees.txt"
    id_oeuvre="VP1"
    id_doc="1">
    <es negation="false" es_id="1" par_id="1" type_es="esa">
      <esa type_en="ville" nom="Lourdes">
        <texte>la ville de Lourdes</texte>
      </esa>
    </es>
    <es negation="false" es_id="2" par_id="1" type_es="esr">
      <esr texte="six␣kilomètres␣de␣Pau␣">
      <relation>
        <distance valeur="6000" unite="m"/>
      </relation>
      <esa type_en="inconnu" nom="Pau"/>
    </es>
     [...]
  </doc>
   [...]
</index>
```

FIG. 6.5 – Index intermédiaire avant géo-référencement.

6.3.2 Validation et géo-référencement

Le processus de validation des entités nommées s'effectue généralement grâce à des ressources ontologiques. Pour le cas du spatial, des systèmes de gestion de données géographiques peuvent jouer ce rôle, tels que des SIG ou des *gazetteers*. Nous avons choisi d'utiliser des couches de SIG (couche des communes) du fait des particularités de notre corpus territorial. Celles-ci contiennent un ratio important d'entités locales à ce territoire qu'il est difficile de retrouver sur des ressources autres que des couches de données locales. Néanmoins, il serait intéressant de coupler ces couches à des ressources externes pour valider de manière plus exhaustive les entités extraites.

Le processus consiste donc à produire une représentation géo-référencée pour chaque instance du modèle pivot (c'est-à-dire pour chaque entité candidate). Il correspond à l'implémentation de l'algorithme 1 et l'implémentation de la fonction *Ajuste_Geo()* pour les primitives géométriques choisies, c'est-à-dire les boîtes englobantes (BE)[57].

```
1  Géoréférencement Calcul_Geo(ES){
2    Si (ES est une ESA){
3      retourne Appel_SIG(ES); /* Géoréférencement */
4    }
5    Sinon /* ES est une ESR */{
6      relation <- Extraction_Relation(ES).
7
8      /* Il faut traiter le cas particulier de la relation figure
9       géométrique à part, car elle est composée de plusieurs sous-ES.
10     */
11     Si (relation = figure_geo){
12     sousESs <- Extraction_Sous_ESs(ES).
13     retourne Ajuste_Geo(relation, Calcul_Geo(sousESs)).
14     }
15     Sinon{
16       sousES <- Extraction_Sous_ES(ES).
17       retourne Ajuste_Geo(relation, Calcul_Geo(sousES)).
18     }
19   }
20 }
```

Algorithme 1 : Algorithme récursif de géo-référencement des ES.

La méthode de calcul que nous proposons, pour la représentation des ES_R, se présente sous la forme d'un algorithme récursif. Cet algorithme, *Calcul_Geo*(), suit les spécifications du modèle pivot. Il permet, de part son caractère récursif, de donner une représentation à toute instance du modèle (par exemple *à 20 km au sud Pau*, *au sud de la frontière franco-espagnole*, *nous avons quitté Bordeaux pour arriver à Pau*). Il comporte deux parties :

– la première (ligne 2-4) est le cas terminal. Si l'ES en paramètre est une ES_A, un appel est fait pour interroger les ressources *via* la fonction *Appel_SIG*().

[57] rectangles orthogonaux par rapport au système de référence qui recouvrent de manière minimale les représentations géo-référencées réelles.

– la deuxième est le cas récursif (ligne 5-18). La relation R et la sous-ES γ sont extraites puis un appel à la fonction $Ajuste\_Geo()$ modifie la zone géo-référencée en fonction de la relation R et de γ lui-même passé en paramètre de $Calcul\_Geo()$. Cette fonction, $Ajuste\_Geo()$, qui implémente les différentes relations définies par le modèle, fait partie, avec les ressources, de ce que l'on considère comme une « ontologie géographique » (dans laquelle les relations sont définies dynamiquement). En effet, les représentations géo-référencées et les relations qui les transforment constituent des ressources hiérarchisées pour les entités nommées dans lesquelles les relations (d'intersection, d'inclusion, d'adjacence, etc.) apparaissent pour chacune des représentations.

La figure 6.6 illustre les différentes interprétations prévues par l'implémentation pour les 5 relations définies dans le modèle pivot. Pour l'*adjacence*, la BE de départ, de longueur de côté c, est agrandie d'un facteur $c/2$ sur chaque côté. Pour l'*inclusion*, les côtés de la BE de départ sont réduits de $c/2$. Pour l'*orientation*, 2 côtés sont agrandis d'un facteur $c/2$ et les 2 autres d'un facteur $2c$. La BE de départ subit ensuite une translation de valeur $1.5c - c/5$ vers l'orientation choisie, afin que la BE finale garde une petite surface de recouvrement avec la BE originale. La relation de *distance* est similaire à celle de l'adjacence. Seul le facteur n'est plus proportionnel à la BE de départ mais est guidé par la distance donnée dans le texte et interprétée par un lexique (x). Enfin la relation de *figure géométrique* est simplement la BE des 2 BE originales pour l'union. Ce choix d'implémentation de la fonction *Ajuste_Geo()* est une proposition intuitive validée de manière empirique. Nous avons choisi de privilégier le rappel du système au détriment du bruit qui n'est pas considéré comme problématique dans ce cas là.

La figure 6.6 montre aussi un exemple de ce que peut donner une interprétation récursive. La phrase *entre le sud de Pau et le nord d'Oloron-Sainte-Marie* est interprétée en plusieurs temps. Une première représentation est calculée pour le sud de Pau, ainsi que pour le nord d'Oloron-Sainte-Marie. Enfin La représentation finale est produite à partir de celles-ci.

Grâce à ce processus, nous complétons donc le flux XML par des représentations géo-référencées des entités validées. La figure 6.8 montre un exemple de flux XML à cette étape. Nous pouvons voir que la représentation (passages qui sont en GML dans le flux, lignes 16-19, 35-39, 45-48) est proposée sous forme de boîte englobante. Cette solution choisie comme première approximation lors de l'implémentation du prototype, permet des calculs très rapides lors de la phase de recherche d'information.

6.3.3 Indexation au grain paragraphe

La dernière étape du processus d'indexation consiste à considérer le flux XML produit (figure 6.8) en tant qu'index, de manière à pouvoir construire un système de recherche d'information. Nous avons choisi d'utiliser la troisième proposition d'indexation décrite dans le chapitre 5, se basant sur les zones géo-référencées des entités extraites. Ce flux XML contient donc les instances du modèle pivot, les identifiants des paragraphes et des documents qui contiennent les entités spatiales correspondantes, ainsi que des représentations géo-référencées pour chaque entité. Par exemple, une instance est définie de la

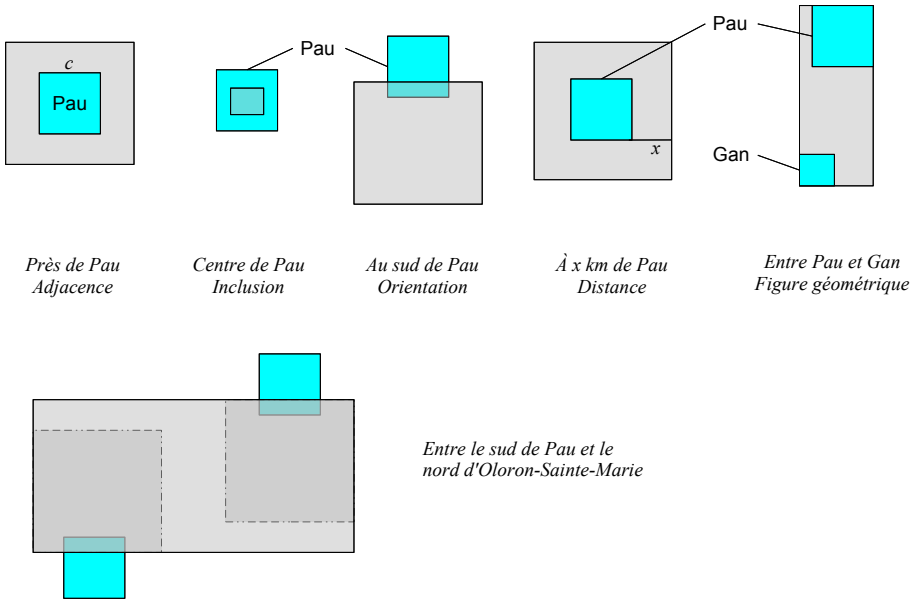


FIG. 6.6 – Interprétation des 5 relations du modèle pivot. Illustration par un exemple récursif.

gid	nom_doc	id_es	id_par	libelle	type_es	the_geom
2032	Excursions_Autour_du_Vignemale.xml	135	237	Arras	esa	POLYGON(...)
1900	Voyage_aux_Pyrenees.xml	26	69	Lourdes	esa	...
2261	Excursions_Autour_du_Vignemale.xml	647	778	Hautes-Pyrenees	esa	...
1991	Excursions_Autour_du_Vignemale.xml	43	148	dans les Pyrenees	esr	...
1911	Voyage_aux_Pyrenees.xml	46	108	Toulouse	esa	...
1912	Voyage_aux_Pyrenees.xml	47	108	Perpignan	esa	...
1808	Au_Pays_des_Isards_ori.xml	14	931	Osse	esa	...
1830	Au_Pays_des_Isards_ori.xml	93	1294	Croix	esa	...
1869	Premiere_Ascension_du_Vignemale.xml	35	31	Vignemale	esa	...
1897	Voyage_aux_Pyrenees.xml	23	69	Nerac	esa	...
1938	Voyage_aux_Pyrenees.xml	97	158	interieur de Saint-Malo	esr	...
1939	Voyage_aux_Pyrenees.xml	99	178	Pau	esa	...
1940	Voyage_aux_Pyrenees.xml	100	178	Bayonne	esa	...

FIG. 6.7 – Extrait d'index stocké dans PostGIS.

ligne 7 à la ligne 22 pour l'entité spatiale « la ville de Lourdes », présente dans le 132ème paragraphe du document *Voyages aux Pyrénées*. Une autre est définie de la ligne 23 à la ligne 49 pour l'entité relative « six kilomètres de Pau ». La représentation associée à cette entité se trouve de la ligne 43 à la ligne 46, au format GML.

Cet index peut être stocké directement dans une base de données XML (approche effectuée dans un premier temps) ou être traduite en SQL pour être stockée en tant que couche géo-référencée dans un SIG (approche retenue dans un deuxième temps par souci d'efficacité). La figure 6.7 est un extrait d'un index construit. Les champs *nom_doc* et *id_par* qui sont les identifiants du document et du paragraphe permettent de relier

```
<?xml version="1.0" encoding="UTF-8"?>
<index [...] nom_oeuvre="Voyages_aux_Pyrenees" date="2006-05-04">
  <doc doc_original="Voyages_aux_Pyrenees.txt"
  Is_doc="Voyages_aux_Pyrenees_doc.xml"
  doc_ocr="Voyages_aux_Pyrenees.txt"
  id_oeuvre="VP1" id_doc="1">
       <es negation="false" es_id="1" par_id="132" type_es="esa">
          <esa type_en="ville" nom="Lourdes">
             <texte>la ville de Lourdes</texte>
             <obj_esa>
                <type table="communes" Where="␣UPPER(nom_com)␣=␣UPPER('
                    Lourdes')" />
                <representation>
                   <gml:MultiPolygon>
                      <gml:Polygon>
                         <gml:OuterBoundaryIs>
                             398220.475063373,1796977.69432353 [...]
                         </gml:OuterBoundaryIs>
                      </gml:Polygon>
                   </gml:MultiPolygon>
                </representation>
             </obj_esa>
          </esa>
       </es>
       <es negation="false" es_id="2" par_id="149" type_es="esr">
          <esr texte="six␣kilomètres␣de␣Pau␣">
             <relation>
                <distance valeur="6000" unite="m" />
             </relation>
             <esa type_en="inconnu" nom="Pau">
                <obj_esa>
                   <type table="communes"
                         Where="UPPER(nom_com)=UPPER('Pau')" />
                   <representation>
                      <gml:MultiPolygon>
                         <gml:Polygon>
                            <gml:OuterBoundaryIs>
                                383419.161795684,1815114.29535332
                            </gml:OuterBoundaryIs>
                         </gml:Polygon>
                      </gml:MultiPolygon>
                   </representation>
                </obj_esa>
             </esa>
             <representation>
                <gml:Polygon srsName="EPSG:27582">
                   <gml:OuterBoundaryIs>377224.673367077,1812021.49164521
                       377224.673367077,1821997.33418898
                       387235.458984745,1821997.33418898
                       387235.458984745,1812021.49164521
                       377224.673367077,1812021.49164521
                   </gml:OuterBoundaryIs>
                </gml:Polygon>
             </representation>
          </esr>
       </es>
    </doc>
</index>
```

Fig. 6.8 – Index XML final après validation.

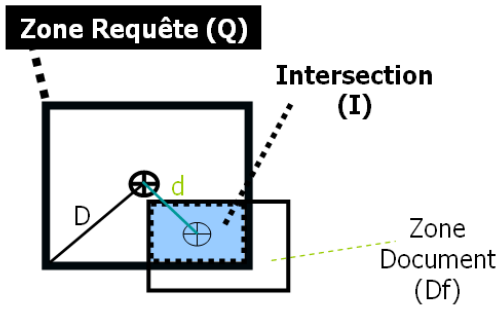


FIG. 6.9 – Variables utilisées pour le calcul d'appariement.

l'enregistrement de l'index au fragment de document correspondant.

6.4 Système de recherche d'information

Cette section décrit le fonctionnement du processus de recherche (en bas de la figure 6.1), qui exploite les index créés. Ce système doit faciliter l'accès aux documents du corpus territorial dans le cas d'interrogations ayant une connotation spatiale. Le scénario d'utilisation commence par cette interrogation exprimée au moyen d'une requête en texte libre sous forme de syntagmes nominaux qui doivent contenir des entités nommées spatiales. Le système doit alors retourner des fragments de documents contenant d'autres syntagmes nominaux indexés et spatialement pertinent avec la requête.

6.4.1 Expression et traitement de la requête

Cette phase n'est pas au cœur de notre problématique. Néanmoins une proposition est faite dans l'optique de réaliser un système complet, ainsi qu'une première évaluation.

La requête peut être exprimée en texte libre et dans ce cas, elle subit le même traitement que les documents. Elle est donc traduite en zones géo-référencées qui serviront pour l'appariement. Ce choix permettra à terme de poser un document entier en tant que requête (requête par l'exemple) afin de récupérer les documents *spatialement* similaires.

Cependant d'autres modes d'expression sont envisagés. Une implémentation réalisée au sein du même projet d'équipe a été proposée pour ce système : un mode d'expression graphique à l'aide d'une interface cartographique. Dans ce cas, la représentation dessinée par l'utilisateur (comme des primitives géométriques), n'a pas besoin d'interprétation particulière, car elle correspond directement à la zone géo-référencée qui intéresse l'utilisateur. Ce mode plus direct nécessite toutefois de connaître le territoire afin de bien positionner les primitives.

6.4.2 Calcul de la pertinence « spatiale »

L'appariement est une question primordiale en recherche d'information. De sa faculté à calculer la pertinence dépend l'efficacité du système. Les travaux existants pour les systèmes classiques utilisent principalement des méthodes vectorielles et probabilistes pour déterminer la pertinence potentielle de documents.

Nous pensons que pour l'information spatiale, le calcul de pertinence est plus simple à construire, du fait de l'interprétation réalisée sur la requête et sur les documents par le modèle pivot. Cette interprétation en représentation géo-référencée permet de modéliser l'intersection spatiale des entités comme une instance supplémentaire du modèle et peut donner la pertinence correspondante, en fonction du recouvrement requête / document.

Si nous prenons l'exemple d'une requête contenant le syntagme *au sud de Pau*, une instance est créée ainsi qu'une représentation géo-référencée. Cette représentation est comparée avec celles contenues dans les index et, par exemple, un fragment de document contenant *la ville de Gan*[58] est retourné. L'intersection calculée correspond alors à une instance du modèle pivot, qui aurait été interprétée à partir de la phrase virtuelle *l'intersection du sud de Pau et de Gan*. Or, à la manière de cette phrase qui, interprétée par un humain, peut donner un score de pertinence *qualitatif*, sa représentation géo-référencée donne un score de pertinence *quantitatif*. Nous nous basons sur cette modélisation à partir du modèle pivot pour proposer un calcul de pertinence spatial. Ce calcul se base sur des représentations à boîtes englobantes mais il pourrait être étendu à toute primitive géométrique.

Pour implémenter cette méthode, nous utilisons les caractéristiques de la figure 6.9 pour calculer un score [SBLG07a] :

$$Df_{score} = \frac{(Df_{precision} + Df_{importance})}{(2 + Df_{distance})} \tag{6.1}$$

L'équation (6.1) utilise des notions de *précision*, d'*importance* et de *distance*. Le score de précision (6.2) évalue la pertinence du document, si la surface d'intersection $I_{surface}$ (partie considérée comme pertinente) occupe un grand pourcentage de la surface du document $Df_{surface}$. Le score d'importance (6.2) évalue si cette surface d'intersection occupe un grand pourcentage de la surface de la requête $Q_{surface}$, afin de savoir si la réponse trouvée est significative. Enfin, un indice de distance (6.2) évalue la distance des centroïdes de la requête et de la surface d'intersection. Plus proches sont ces centroïdes, plus pertinent est le document D_f.

$$Df_{precision} = \frac{I_{surface}}{Df_{surface}} \quad Df_{importance} = \frac{I_{surface}}{Q_{surface}} \quad Df_{distance} = \frac{d}{D} \tag{6.2}$$

L'équation (6.1) est traduite ensuite en requête SQL utilisant des fonctionnalités SIG afin d'effectuer la recherche (figure 6.10). Le résultat retourne une liste ordonnée par ordre de pertinence des identifiants de documents résultats.

[58] Gan se situe à 10 km au sud de Pau.

area(intersection(Q_geom, Df_geom))	I_surface
area(Df_geom)	Df_surface
distance(centroid(Q_geom), centroid(Df_geom))	d
distance(centroid(Q_geom), geomfromtext('corner coordinate'))	D
SELECT pi.gid, pi.doc_name, pi.par_id, pi.SF-name, (tq.isurf/tq.dfsurf + tq.isurf/tq.qsurf)/(2 + tq.d/tq.D) AS weight FROM piv_index pi, temp_query tq WHERE pi.gid=tq.gid ORDER BY weight DESC;	

FIG. 6.10 – Calcul des surfaces, des distances et du score.

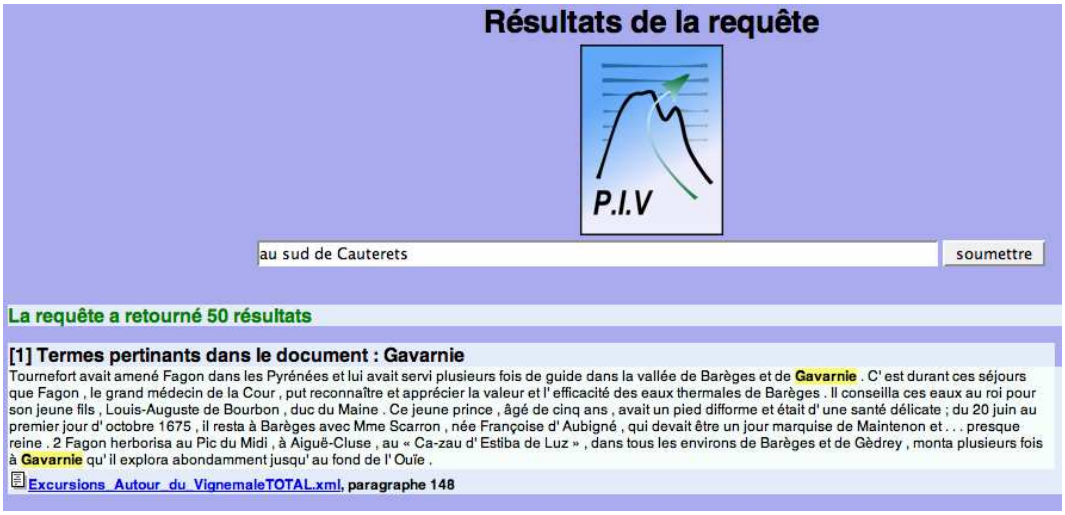


FIG. 6.11 – Visualisation des résultats dans une liste.

6.4.3 Visualisation des résultats

Nous proposons deux types de visualisation des résultats, même si cette phase, comme celle de l'expression de la requête, n'est pas au cœur de notre travail.

La liste d'identifiants retournée par la requête permet de récupérer dans une base XML où ils sont stockés, les fragments de documents pertinents. L'interface proposée au départ est une simple liste dans une page web qui pointe pour chaque élément vers le document original du corpus (figure 6.11).

Des travaux sur la visualisation cartographique des résultats ont également abouti à des résultats intéressants [EMC06, SME06, ME07] (figure 6.12). En effet, notre indexation et notre outil d'appariement permettent de fournir à l'interface les identifiants des fragments de documents pertinents ainsi que leur géo-référencement. Une réflexion est faite sur la problématique de visualisation de ces zones et notamment les problèmes d'occlusion, c'est-à-dire les problèmes de surcharge visuelle sur les fonds cartographiques.

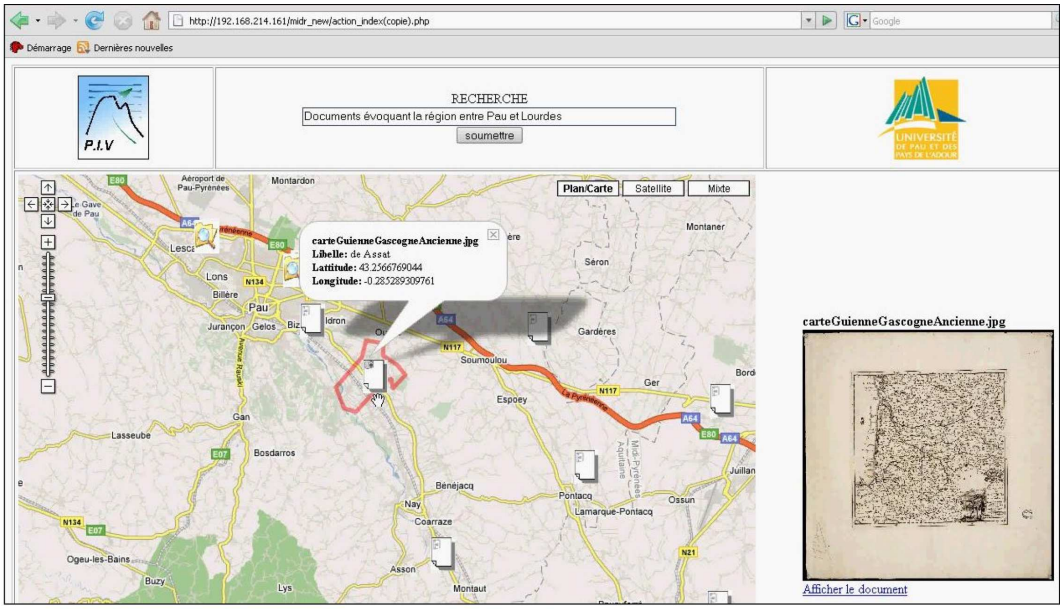


FIG. 6.12 – Visualisation des résultats sur une carte.

6.5 Évaluation intermédiaire

Après avoir implémenté l'ensemble des modules détaillés précédemment sous forme de services web (voir l'annexe E), nous avons voulu tester l'efficacité d'un tel système. Cette section expose deux évaluations effectuées sur le prototype PIV. La première, qui a pour but de valider nos hypothèses concernant la faisabilité d'un système alliant des traitements sémantiques à une indexation pour la recherche documentaire, évalue le taux de récupération de l'information spatiale extraite et identifie les problèmes liés à l'extraction. Cette évaluation a fait l'objet de la publication [SGLL07].

La deuxième évaluation compare une approche classique de RI basée sur un modèle vectoriel et notre approche, dans le cas d'un corpus territorial et de requêtes spatiales et thématico-spatiales. Nous avons publié ces résultats dans [SBLG07a, SBLG07b].

6.5.1 Évaluation de la partie EI du système PIV

Il est à noter que ces résultats évaluent la première version du prototype PIV. Nous avons lancé le processus d'extraction d'information de PIV sur un échantillon du corpus comprenant 10 livres (de type œuvre littéraire) scannés et OCR-isés. Un livre de 200 pages met une dizaine de minutes à être indexé. Le prototype a trouvé 9835 ES candidates pour ces 10 livres.

À la manière des campagnes d'évaluation de CLEF[59], une extraction et une indexation des ES ont été réalisées de manière manuelle : les participants marquent manuel-

[59] Cross-Language Evaluation Forum (CLEF) - www.clef-campaign.org

lement les ES trouvées dans l'ensemble du corpus pour avoir des résultats de référence. Nous avons ensuite comparé cette annotation manuelle à celle réalisée par PIV afin de calculer des taux de rappel et de précision à chaque étape du processus.

L'analyse des données extraites par PIV révéle que celui-ci détecte des ES candidates de types divers (ville, rue, département, commune, montagne, vallée, rivière, fleuve et route). La figure 6.13 montre la répartition de ces ES.

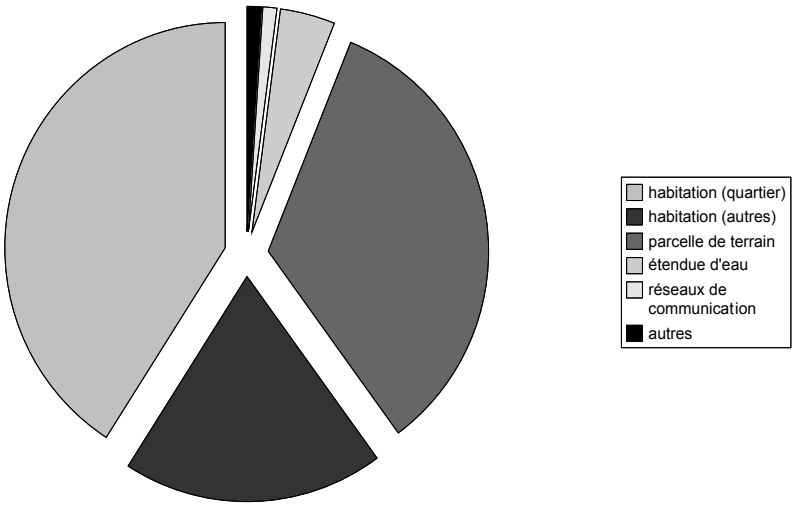


FIG. 6.13 – Répartition des ES candidates détectées.

Pour évaluer l'efficacité du système au niveau de la récupération des ES, nous avons choisi d'utiliser des mesures de rappel et de précision. Le rappel R correspond ici à

$$R = \frac{Nombre\ d'ES\ extraites\ validées\ par\ PIV\ et\ pertinentes}{Nombre\ d'ES\ pertinentes}$$

et la précision P à

$$P = \frac{Nombre\ d'ES\ extraites\ validées\ par\ PIV\ et\ pertinentes}{Nombre\ d'ES\ extraites\ validées\ par\ PIV}$$

L'évaluation révèle que le taux de rappel est de 49% et le taux de précision de 73%. Le taux de rappel un peu bas s'explique, après analyse des ES non récupérées, par le manque de ressources géographiques, puis avec une moins grande importance par le manque d'indicateurs spatiaux dans les lexiques, les variations orthographiques et les problèmes d'OCR-isation. Les erreurs de précision sont dues aux homonymies. La figure 6.14 fait la synthèse des résultats en donnant la répartition des causes d'échec de récupération des ES.

Nous détaillons maintenant nos résultats d'expérimentation pour chaque étape du processus d'indexation.

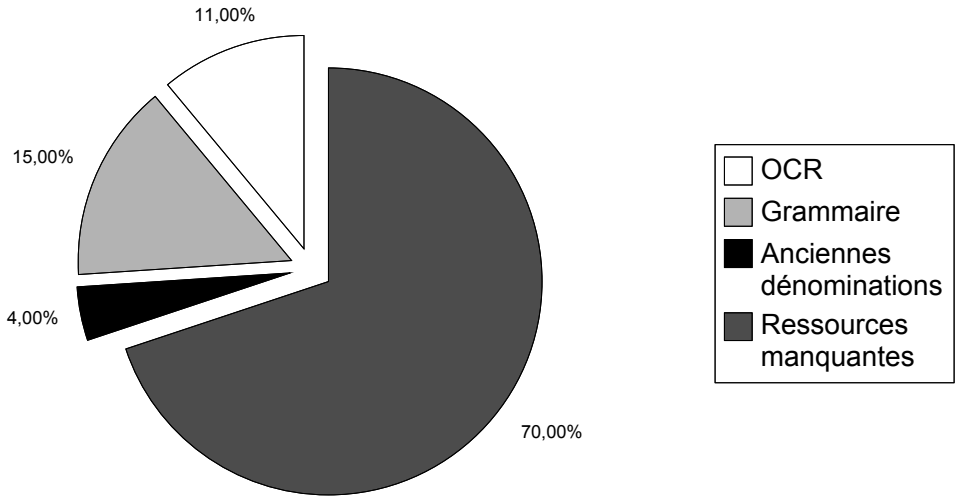


FIG. 6.14 – Répartition des causes d'échec de récupération des ES.

Étape de traitement sémantique La grande majorité des ES (80%) sont détectées et marquées. L'explication pour les ES non récupérées à cette étape venait soit du fait que les syntagmes introducteurs ne sont pas écrits correctement (français ancien), soit qu'ils sont absents (dans des carnets de voyage par exemple, les villes par lesquelles passe l'auteur sont égrenées seules comme titre de paragraphe ou encore, un extrait de l'annexe A.1 : *Nous voyons Orléans en bonnet de nuit, et dans la journée, Tours devenue colonie anglaise*), soit que les entités nommées n'ont pas de majuscule. Pour le cas des ES_R non détectées, l'explication vient du manque de règles pour récupérer certaines expressions non triviales (par exemple *Non loin de ce mont que tu appelles le Mont Maudit* qui devrait correspondre à une « adjacence au Mont Maudit » mais qui est considéré comme une simple ES_A).

Étape de validation par géo-référencement Parmi les ES extraites à l'étape 2 qui sont réellement des ES, nous n'en gardons que 30% à l'étape 3. Ceci est dû principalement au manque de ressources[60]. Les autres causes sont soit une mauvaise OCR-isation, soit une orthographe ancienne de l'entité nommée.

Conclusion de l'évaluation Ces résultats ont donc montré qu'un système de recherche d'information spatiale peut être envisagée en suivant les spécifications suivies par ce prototype.

Cette évaluation nous a amené à nous pencher sur les problèmes de variation orthographique et de mauvaise OCR. Une solution trouvée est d'utiliser un calcul de similarité

[60] Nous n'avions alors que la couche des communes dans notre base.

entre l'entité nommée que l'on ne parvient pas à valider et nos ressources. Par exemple, la *distance de Levenshtein*[61], dont une implémentation existe dans le SIG que nous utilisons (au sein de l'API *fuzzystrmatch*), donne une distance entre 2 mots en fonction du nombre d'opérations de suppression, d'insertion et de remplacement qu'il est nécessaire de réaliser pour qu'ils soient identiques.

Un processus de validation supplémentaire semi-automatique est alors envisageable au cours de la phase d'indexation. Pour une entité nommée candidate qui ne peut être validée directement, un calcul de similarité donne à un utilisateur expert une liste de propositions proches de la dénomination originale. Celui, en s'aidant du contexte peut alors retrouver la bonne correspondance et décider de valider l'entité. Une méthode d'apprentissage est aisément envisageable afin de faciliter ce processus semi-automatique.

Un exemple de cas sur lequel nous sommes tombé durant l'expérimentation, pour lequel cette méthode serait utile, est la dénomination du village de *Cauteretz* qui est souvent écrit *Cauterets* dans les documents du *XVIII*^{e} siècle. La distance de Levenshtein vaut ici 1 et *Cauteretz* est le premier résultat retourné pour *Cauterets*. Un autre exemple est la ville de *Bayonne* qui, à cause d'une erreur de reconnaissance de caractères, est quelque fois transformé en *Rayonne*. Là aussi, la distance vaut 1 et *Bayonne* arrive en premier résultat. Enfin un dernier exemple qui n'est pas résolu avec cette méthode est la troncature d'une entité nommée. La ville d'*Oloron-Sainte-Marie* est souvent dénommée *Oloron* ; plus embêtant, les villes de *Bagnères-de-Luchon* et *Bagnères-de-Bigorre* sont tronquées en *Bagnères*. Il reste donc des travaux à réaliser pour réduire la perte d'entités candidates.

Une autre amélioration envisagée à la suite de cette évaluation est l'acquisition de ressources plus hétérogènes. Nous avons fait l'acquisition, pour ce faire, de ressources géographiques de l'IGN portant sur le territoire pyrénéen et relatives aux oronymes, hydronymes et toponymes, ceci afin de réduire les échecs de récupération (figure 6.14).

6.5.2 Évaluation de la partie RI du système PIV

Cette évaluation compare la précision d'un système classique de RI basé sur un modèle de recherche vectoriel et la précision du système de recherche d'information de PIV (partie du bas de la figure 6.1).

Le corpus utilisé pour évaluer le système PIV correspond encore aux 10 livres OCRisés traitant du patrimoine culturel Pyrénéen et datant des XIXème et XXème siècles. Les livres sont découpés en paragraphes constituant 10 000 unités documentaires. Des bibliothécaires ont créé 12 requêtes dont 8 traitent de la dimension spatiale seulement et les quatre autres traitent de la dimension spatiale et thématique. Une requête spatiale peut faire référence à des entités spatiales absolues (ESA) ou relatives (ESR). Une requête spatiale et thématique telle *« instruments de musique dans les environs de Laruns au XIX*^{e} *siècle »* supporte aussi bien les entités ESA/ESR (*« environs de Laruns »*) que d'autres entités non spatiales (*« instruments de musique », « XIX*^{e} *siècle »*). Nous avons aussi utilisé l'annotation manuelle réalisée pour la première expérience.

[61] http://fr.wikipedia.org/wiki/Distance_de_Levenshtein

Toutes les requêtes	P@5	P@10	P@15	Nombre de réponses
A) Approche Spatiale (PIV)				
Avg	0.78	0.81	0.73	637
B) Approche classique				
Avg	0.50	0.43	0.40	252

TAB. 6.1 – Résultats de PIV et de l'approche classique sur les requêtes spatiales.

Évaluation de la RI spatiale Nous avons soumis huit requêtes spatiales et quatre requêtes mixtes au système PIV et à un système classique de RI (basée sur un modèle vectoriel) :

```
à l'ouest de Barèges
le triangle Accous, Arudy, Cauterets
de Pau à Bagnères-de-Luchon
autour de Barèges
les environs de Cauterets
la périphérie d'Arudy
entre Pau et Laruns
au sud de Pau
la route de Pau à Bagnères-de-Luchon
les instruments de musique dans les environs de Laruns
la périphérie de Laruns au XIXème siècle
églises et châteaux à l'ouest de Barèges
```

Nous avons comparé les premières unités documentaires restituées (top 5, 10 et 15) aux jugements manuels. Les résultats sont donnés en Table 6.1 où Avg représente la précision moyenne calculée sur toutes les requêtes utilisées et P@5, P@10 et P@15 désignent respectivement la précision pour les 5, 10 et 15 premiers résultats. La dernière colonne, Nombre de réponses, représente le nombre total de documents retournés (moyenne sur toutes les requêtes).

On peut remarquer que l'approche spatiale PIV (Table 6.1-A) donne une précision de 78% dans les cinq premiers documents (top 5) et 81% dans les dix premiers. Quand les mêmes requêtes sont soumises au système classique, les résultats se dégradent de manière significative (Table 6.1-B). La raison est que pour une requête spatiale telle *près de Laruns*, l'approche classique (basée sur une comparaison mot-mot) ne trouve jamais les documents traitant d'autres villes comme *Eaux-Bonnes* ou *Louvie-Soubiron*, qui sont localisées dans le voisinage de *Laruns*. Notre approche, en extrayant les ES à partir du texte des documents et des requêtes, propose une comparaison intégrant une interprétation de la sémantique de ces ES pour récupérer les documents pertinents. Ce résultat est aussi confirmé si nous considérons le nombre de documents restitués : en moyenne, 637 documents sont trouvés par l'approche spatiale pour toutes les requêtes ; alors que seulement 252 sont sélectionnés par l'approche classique.

Évaluation de la RI thématique et spatiale Notre approche est adaptée à une requête uniquement spatiale. Dès lors que des requêtes mixtes sont expérimentées, la P@5 est très faible (0.15) pour l'approche PIV alors que le système classique obtient une meilleure précision de 0.48. Cependant ce dernier résultat n'est pas très élevé. L'idée est alors d'évaluer un système qui combinerait les 2 approches (spatiale et classique) afin de constater les améliorations éventuelles.

Conclusion de l'évaluation Les résultats de cette expérience sont positifs pour notre prototype, qui est meilleur qu'un système classique dans le cas de requêtes posées de nature spatiale. Il répond dans ce cas de manière satisfaisante (0.78 pour la P@5). Une rapide évaluation sur des requêtes mixtes, ne contenant pas seulement de l'information spatiale, montre cependant que le système implémenté n'est qu'une sous-partie d'un véritable système de recherche d'information géographique, devant gérer les deux autres composantes (temporelle et thématique).

6.6 Bilan des réalisations et perspectives

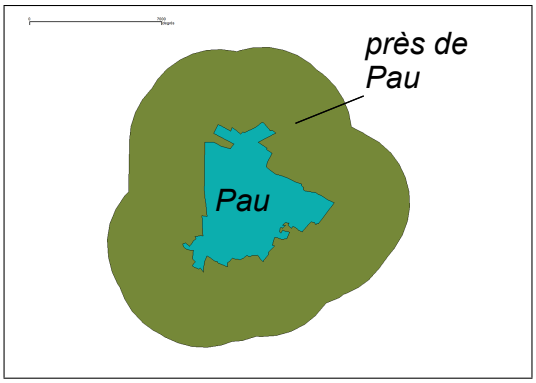
PIV est donc l'implémentation d'une architecture distribuée réunissant les outils nécessaires à l'indexation de document textuel, basés sur des techniques linguistiques. L'objectif de cette contribution est de montrer la faisabilité d'un tel SRI spécialisé et son utilité dans le cadre d'usages néophytes ou spécialistes recherchant de l'information dans des corpus territoriaux numérisés (dans les médiathèques par exemple).

Ce chapitre s'est d'ailleurs focalisé sur les documents de type texte, mais à terme il est envisagé d'indexer tous les types de documents, comprenant des lithographies, des cartes postales, des photos, des documents cartographiques, tous numérisés eux aussi. À titre d'exemple, un test a été fait sur des lithographies avec l'extraction d'information de PIV. Seule la première phase d'extraction d'ES a été réalisé à la main. La suite du processus génère automatiquement un index, comparable aux index textuels. À la suite de ce test, les documents retournés sont soit de type textuel, soit lithographiques et sont retournés à l'utilisateur de la même manière.

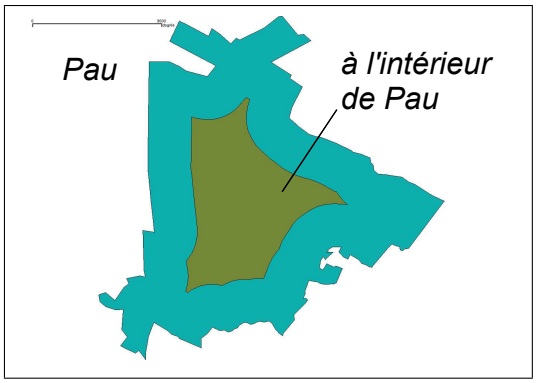
Les évaluations PIV ont montré la faisabilité de son architecture. De plus les résultats de précision du système sont encourageants. Ils montrent que PIV renvoie plus de résultats pertinents qu'un système basé sur des méthodes existantes pour des requêtes spatiales. Cependant, les résultats obtenus découlent d'expérimentations menées sur des échantillons représentatifs mais de petites tailles. Il serait intéressant, à l'avenir, de pouvoir tester notre système dans des campagnes d'évaluation, sur des corpus plus grands. La campagne GeoCLEF[62] serait en particulier pertinente, au vu de son objectif d'évaluation de systèmes de recherche d'information géographique dans plusieurs langues.

Ces évaluations nous ont aussi permis de proposer des améliorations (réalisées ou à faire) :

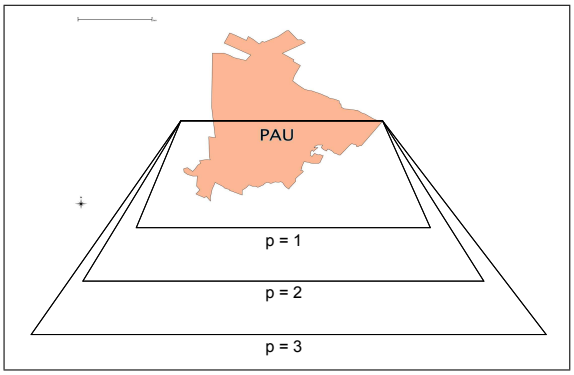
[62] http://ir.shef.ac.uk/geoclef/



(a) Adjacence à *Pau*.



(b) Inclusion de *Pau*.



(c) Représentation du *sud de Pau* intégrant la notion de poids d'interprétation pour traduire le flou inhérent aux relations qualitatives.

FIG. 6.15 – Exemples de représentations qualitatives guidées par la connaissance des ressources géographiques.

- Nous avons amélioré l'utilisation de lexiques pour les introducteurs spatiaux des entités et enrichi la grammaire de règles augmentant le nombre de patrons détectés.
- Nous avons fait l'acquisition de nouvelles ressources pour augmenter la récupération des entités spatiales candidates.
- La représentation a été améliorée grâce à l'utilisation de primitives géométriques (comme les polygones) plutôt que par des boîtes englobantes. Les figures 6.15(a) et 6.15(b) sont des illustrations de ce que peut produire un SIG pour représenter les interprétations des ES_R. De plus, d'autres travaux intégrant la notion de flou à travers la définition d'un poids d'interprétation ont été proposés. Par exemple, dans la figure 6.15(c), une interprétation d'orientation est illustrée. Pour une entité relative sont associés trois (et non plus une) représentations, ayant des poids de plus en plus faibles pour l'interprétation que la surface est grande. L'apparence trapézoïdale a été choisie pour approcher la représentation conique souvent donnée aux points cardinaux (tout en tenant compte des facilités d'implémentations). Ces choix de représentation font disparaître les biais gênants causés par l'utilisation des boîtes englobantes. Par exemple, l'interprétation peut donner une mauvaise représentation des relations d'union entre 2 ES, du fait du caractère orthogonal des boîtes englobantes (problème de bruit). Elle peut donner aussi une trop grande simplification pour la relation d'adjacence. La boîte calculée correspond en effet à un agrandissement de la boîte de départ, sans tenir compte que l'adjacence d'une région A n'inclut pas la région A dans son ensemble.
- Il est possible de récupérer des entités nommées mal orthographiées ou mal OCR-isées grâce au calcul de similarité de chaîne de caractère (distance de Levenshtein). Une interface d'indexation semi-automatique doit permettre d'augmenter le nombre d'entités extraites correctement.
- L'interprétation graphique de la requête exprimée en texte libre peut être présentée à l'utilisateur sur un fond de carte (à l'aide de GoogleMaps par exemple) en amont de la phase de recherche, afin de vérifier la justesse d'interprétation et de modifier le cas échéant la représentation. Par contre la requête en texte libre n'est pas ré-interprétée ensuite. Des propositions d'interrogation bi-modale existent comme celle proposée par F. Dumoncel [Dum06] dans sa thèse, qu'il pourrait être intéressant d'adapter. Une représentation spatiale symbolique produit des interprétations en texte libre et inversement, le texte peut modifier le schéma.
- Les opérateurs logiques interprétés dans la requête permettent de préciser la ou les zones d'intérêt ou non. Cette amélioration est aussi liée à l'amélioration du traitement sémantique qui doit interpréter la négation dans la phrase *(Je n'étais pas du tout à Pau mais à Paris)*.
- Enfin un autre travail sur la visualisation des index spatiaux dans le logiciel GoogleEarth a été réalisé [Vu06]. Celui-ci transforme les index au format KML[63] propre à ce logiciel. Cet outil permet de réaliser une visite virtuelle d'un territoire en 3D (permettant de mettre en valeur les récits de voyage[64]) et de visualiser des repré-

[63] http://earth.google.fr/userguide/v4/ug_kml.html

[64] La plupart des récits du corpus territorial se déroulent dans les Pyrénées.

sentations de documents du corpus en fonction de leur localisation.

Le travail présenté ici est donc une première étape. L'indexation se fait au niveau de la phrase (ou d'un petit groupe de phrases). Nous envisageons maintenant un système intégrant un moyen d'indexer à plusieurs niveaux un document, en regroupant les ES entre elles. Un tel système retourne des fragments de documents de différentes granularités à un utilisateur et pas seulement des paragraphes. Il y gagne ainsi en cohérence car un ensemble d'ES dans un groupe de phrases a plus de poids qu'un seul syntagme se trouvant dans un paragraphe. De nouvelles questions se posent alors sur la manière de regrouper ses ES. De la même façon, les problématiques liées aux usages et aux besoins des utilisateurs s'ouvrent à nous, dans la manière d'indexer les documents, la manière de poser la requête au système et de retourner les documents à l'utilisateur.

La suite de notre contribution s'est focalisée sur l'indexation multi-niveaux de document textuel, celle-ci étant primordiale pour construire un SRI efficace et cohérent. La recherche spécifique au domaine du spatial et plus particulièrement des récits de voyage, des itinéraires, a guidé notre manière de concevoir ces index, notamment en synthétisant l'information récoltée grâce à l'extraction d'ES et en classifiant ces synthèses suivant des motifs spatiaux, tels que mentionnés dans le chapitre 5. Le chapitre suivant se concentre donc sur la problématique de la synthèse spatiale de document textuel en se basant sur la configuration des ES citées par l'auteur dans le texte pour, à terme, proposer une indexation spatiale multi-niveaux.

Chapitre 7

Indexation spatiale par motifs

Sommaire

7.1 Introduction

L'objectif de ce chapitre est d'exposer le processus d'indexation à un niveau plus élevé que l'indexation intraphrastique détaillée dans le chapitre précédent. Nous avons vu que l'information spatiale exprimée dans un texte est utilisée dans un contexte spécifique par l'auteur, dont la portée est plus grande que la phrase. Ce contexte peut être catégorisé selon différents motifs spatiaux afin de créer une nouvelle indexation plus poussée dans le cadre d'un système de recherche d'information spécialisée.

L'idée est alors de regrouper les entités spatiales extraites dans le processus d'indexation intraphrastique et d'associer aux unités de textes les contenant un motif caractéristique, ceci afin d'indexer non plus au niveau de la phrase mais à tous les niveaux de la structure hiérarchique d'un document. Cette méthode correspond en quelque sorte à réaliser une synthèse de texte, un résumé, du point de vue spatial.

L'approche habituelle d'un processus de résumé automatique est l'agencement de phrases clés extraites du texte. Les phrases clés sont sélectionnées à partir de mots clés,

eux-mêmes déterminés grâce à des évaluations statistiques sur leur fréquence [Nen06]. A. Nenkova [NVM06] fait l'étude de trois facteurs pris en compte lors d'un processus de synthèse :

- la fréquence des mots *conteneurs de sens* (noms communs, verbes et adjectifs),
- le choix d'une fonction de composition (pour lier les phrases clés entre elles).
- l'ajustement du poids des mots selon le contexte.

Pour exemple de travail récent, J. Otterbacher [ORK06] propose une méthode de synthèse de texte hiérarchique. L'objectif est d'être compréhensible avec le minimum de mots sans forcément créer un résumé en texte libre, une application possible étant l'affichage de nouvelles sur un téléphone mobile.

Pour les mêmes raisons qui nous ont fait nous éloigner des méthodes statistiques classiques pour la recherche d'information, nous privilégions une synthèse d'information spatiale en utilisant ses spécificités. Un travail plus proche est celui du projet GIPSY [WP94] qui utilise une méthode d'agrégation spatiale pour indexer les documents, l'indexation consistant à prendre la zone géographique la plus mentionnée.

L'intérêt de synthétiser un document textuel à plusieurs niveaux de sa structure est l'amélioration de la recherche d'information et la représentation des résultats pendant le processus de recherche documentaire. Il est intéressant de pouvoir piocher dans un ouvrage un ou quelques extraits particulièrement pertinents. Il est d'autant plus intéressant de pouvoir indiquer, dans le cadre d'une recherche spatiale, dans quel contexte l'auteur parle de l'information spatiale pertinente. Nous nous intéressons donc maintenant à la manière d'interpréter à un plus haut niveau l'information spatiale présente dans les documents.

L'indexation se basant sur la catégorisation en motifs, nous exposons d'abord rapidement les méthodes existantes de catégorisation (ou classification) automatiques[65] puis, nous présentons comment sont calculées les caractéristiques nécessaires à ce processus. Enfin nous faisons une évaluation de l'indexation par motifs proposée.

7.2 Outils existants

Beaucoup d'outils ont été proposés pour répondre aux problématiques de classification et de catégorisation, présentes dans beaucoup de domaines scientifiques autres que la Recherche d'Information (analyse d'image, biologie génétique, reconnaissance de caractères, etc.). La plus ancienne proposition est celle des *« K-means »* de J.MacQueen [Mac67]. Tout comme le modèle bayesien *(« Naive Bayes Classifier »)* proposé plus tard [DP97], cette méthode utilise un modèle probabilistique permettant de connaître la distribution sur les différentes classes. D'autres méthodes, basées sur les statistiques et l'apprentissage existent aussi [Ril96], notamment avec les réseaux de neurones [Lip87,PR03].

[65] La classification est à différencier de la catégorisation du fait que la première consiste à organiser un ensemble de textes en des classes homogènes tandis que la seconde est destinée à associer à chaque texte une catégorie déjà existante [Iha04]. Cependant, nous utiliserons les deux termes pour parler de catégorisation.

Plus spécifiquement pour la catégorisation automatique supervisée, plusieurs méthodes ont été proposées [Iha04]. La plus simple est la méthode des *k plus proches voisins*. Elle consiste globalement à déterminer pour chaque nouvel élément, la liste de ses k plus proches voisins parmi les éléments déjà classés. L'élément est alors inséré dans la classe qui contient le plus des k éléments découverts. Le problème de cette méthode est qu'elle donne autant d'importance à un élément éloigné de celui à classer qu'à un élément proche. L'autre inconvénient est le choix de la valeur de k, qui est fait de manière empirique. Enfin, contrairement aux autres méthodes dont la complexité est souvent dépendante du nombre de catégories, les *k plus proches voisins* entraînent le calcul d'autant de similarités que de documents [Iha04], ce qui peut être rédhibitoire.

Une autre méthode connue est la *classification bayésienne naïve*, qui est une approche par modèle de langue. Catégoriser un texte revient à calculer la probabilité que la suite de mots qui le compose puisse être généré pour chacune des catégories. Le nouveau texte est étiqueté selon la thématique correspondant au langage de probabilité maximale.

Les *arbres de décision* permettent aussi de répondre à la problématique de catégorisation. La construction d'un tel arbre comprend :

- une méthode pour réaliser les subdivisions des nœuds,
- une règle de décision pour décider si un nœud est terminal
- et un critère permettant de choisir la classe qui correspond à chaque nœud terminal.

Le processus consiste ensuite à partir de la racine de l'arbre contenant tous les éléments du jeu d'apprentissage et de faire en sorte d'opérer des subdivisions selon des caractéristiques qui minimisent les éléments mal classés.

La dernière approche est la classification par les *SVM*, proposée par V.Vapnik [Vap98], qui parait la mieux adaptée pour la classification de texte [Joa98]. Elle réunit en effet pour K.P.Bennett [BC00] l'intuition géométrique, des mathématiques *élégantes*, des garanties théoriques et un algorithme pratique. Nous allons présenter plus en détail cette méthode que nous avons choisi d'utiliser.

7.2.1 Méthode de Support à Vastes Marges

La méthode de Support à Vastes Marges ou « SVM » (pour *Support Vector Machines*, traduit en français par *Machine à Vecteurs de Support* ou *Séparateur à Vastes Marges*) [Joa98, Joa99, Joa00, Bur98][66] est une méthode d'apprentissage basée sur les statistiques.

Le paradigme employé pour cette caractérisation est de considérer les éléments comme des vecteurs dont les différentes dimensions sont une évaluation numérique des caractéristiques. La classification s'opère alors sur la position de ces vecteurs.

Datant des années soixante et implémentée dans les années quatre-vingt-dix principalement pour des applications en génétique mais aussi pour la classification de texte, le SVM est aujourd'hui une des méthodes les plus utilisées. Elle part de deux hypothèses de base, en considérant un espace multi-dimensionel dans lesquelles se trouvent les données à classifier (les dimensions étant leurs caractéristiques) :

[66] L'article [Has06] fournit un bon tutoriel sur les SVM.

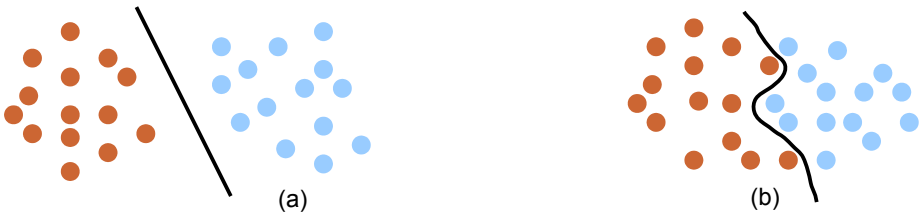


FIG. 7.1 – Exemples d'objets à classifier selon une propriété définie : ici la couleur.

- Certains calculs ne sont pas plus coûteux en transformant l'espace de représentation des données (χ) en un espace de grande dimension,
- Deux classes peuvent être linéairement séparées dans un espace de grande dimension.

L'objectif est alors de trouver le meilleur hyperplan séparateur dans l'espace de grande dimension avec des calculs peu coûteux.

La méthode consiste à associer à un élément $\vec{x}$ d'un ensemble χ, une classe notée y pouvant prendre deux valeurs (vrai ou faux). Ce couple est nommé « observation ». Une série de n observations S notée : $S = \{(\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)\}$, est utilisée comme base d'apprentissage. L'algorithme SVM construit alors des règles qui peuvent être utilisées pour classifier un nouvel objet $\vec{x} \in \chi$.

Cette technique se base sur le concept des plans de décision qui définissent une frontière séparant un ensemble d'éléments ayant des propriétés différentes. L'approche la plus simple, nommée *linear SVM* est de séparer les éléments disposés sur un plan à l'aide d'une ligne droite (Figure 7.1 (a)). Il est cependant rare d'avoir une séparation aussi nette entre les éléments et la ligne séparatrice est alors une courbe complexe difficile à modéliser (b). L'idée est alors de trouver des fonctions mathématiques (nommées *noyaux*) afin de projeter l'ensemble original (b) de manière à le séparer de nouveau par une ligne droite.

Des travaux récents sur la classification de texte utilisent les SVM. R. Angelova [AW06] propose une méthode basée-graphe pour les documents textuels afin de classifier des documents web (grâce aux hyperliens), ou des articles scientifiques (grâce aux noms d'auteurs ou d'éditeurs). V. Sindhwani [SK06] catégorise du texte avec une SVM linéaire. N.Jindal [JL06] identifie les phrases comparatives dans un texte.

Ces travaux confortent notre choix concernant l'utilisation des SVM. En particulier, nous avons utilisé l'implémentation de C.W. Hsu et C.J. Lin [HL02b, HL02a, CKSL04] qui s'appelle BSVM. Afin de choisir convenablement le jeu d'apprentissage, des mesures ont été proposées [Vap98]. En effet, celui-ci doit être le plus riche possible (afin d'améliorer la convergence d'apprentissage). Cependant il ne faut pas non plus faire du sur-apprentissage afin préserver la capacité de généralisation du classifieur. La mesure

$$R_{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} l[h(x_i), u_i]$$

représente le *risque empirique* qu'il faut minimiser, où x_i est un des m exemples d'apprentissage, u_i l'étiquette associée et l la fonction de perte mesurant la distance entre la réponse produite $h(x_i)$ et la réponse désirée u_i. Cette mesure est liée, quand le jeu d'apprentissage est convenablement choisi, au *risque réel* qui nous intéresse vraiment et qui est représenté par la formule

$$R_{réel} = \int_{X \times U} l[h(x_i), u_i] \, dF(x,u)$$

où F est la distribution inconnue des exemples sur $X \times U$, produit cartésien de l'espace des observations X et de l'espace des étiquettes U. Une trop grande richesse du jeu d'apprentissage risque alors de distendre le lien entre le risque empirique calculable et le risque réel (non calculable) et donc d'apprendre une hypothèse n'ayant rien à voir avec le vrai concept cible.

7.2.2 Construction des caractéristiques

L'hypothèse du travail présenté ici est que les motifs spatiaux, décrits dans nos préconisations par exemple, peuvent être définis par des caractéristiques utilisables dans un processus de classification tel que celui des SVM. Pour cela nous avons essayé de traduire numériquement les propriétés de dispersion, d'ordonnancement et de saillance à travers différentes caractéristiques.

Il convient en effet de placer notre cas d'étude dans un espace multi-dimensionnel. Pour cela nous définissons une unité de texte comme étant un élément $\vec{x}$ de l'ensemble du corpus χ. Les N classes y_i $(i = 1..N)$ sont définies grâce au jeu d'apprentissage avec lequel un superviseur va entraîner le système. L'élément $\vec{x}$ est défini par les m caractéristiques que nous allons définir (qui jouent le rôle de coordonnées : $\vec{x}(c_1, c_2, ..., c_m)$).

La section suivante décrit donc les différentes caractéristiques définies afin d'exprimer les propriétés de dispersion, c'est-à-dire la faculté qu'ont les entités d'être regroupées ou non, d'ordonnancement (l'ordre dans lequel elles sont citées dans le texte) et de saillance (angle qu'elles forment, de type angle plat ou angle aigu), qui permettent de distinguer les différents motifs définis. Nous décrivons la manière dont ont été implémentées les évaluations de ces caractéristiques.

7.3 Implémentation des caractéristiques

Nous avons définis six caractéristiques pour exprimer les propriétés de prédominance, d'ordonnancement, de saillance et de dispersion : $\vec{x}(P_1, P_2, O, S, D_1, D_2)$.

Les calculs de prédominance P_1, P_2 et les calculs de distance D_1 et D_2 expriment la propriété de dispersion. L'ordonnancement est exprimé par la caractéristique O, la saillance par S. Nous avons essayé de définir à travers ces caractéristiques, des moyens de distinguer les motifs définis dans nos préconisations. Le tableau 7.4 illustre pour chaque caractéristique les valeurs attendues en fonction du motif. Les valeurs attendues vont de très faible $(--)$ à très forte $(++)$.

Motif \ Caractéristiques	P1	P2	D1	D2	O	S
Itinéraire	−	−−	−+	+	++	++
Description locale	++	−−	+	++	−	−
Description de point de vue	+	−+	+	+	−+	−+
Comparaison de lieux	+	++	+	+	−	−

TAB. 7.2 – Caractéristiques et valeurs attendues pour les motifs choisis.

À titre d'exemple, une valeur faible (−) de $P1$ est attendue pour le motif *itinéraire*, étant donné que la description d'un cheminement d'un point A à un point B n'est pas censé trop s'attarder sur un lieu en particulier. Par contre l'ordonnancement O devrait avoir une valeur forte (+), tout comme la saillance S. Une valeur forte est aussi attendue pour l'évaluation de D_1, du fait que l'on s'attend, pour ce motif, à avoir une minorité d'ES très éloignées des autres.

Contrairement au motif précédent, les motifs de *description locale* et de *point de vue* impliquent une valeur forte pour la prédominance P_1. Une zone géographique doit se dégager en effet de l'ensemble des ES. D_2 doit aussi avoir une valeur forte afin d'exprimer l'adjacence des ES mentionnées pour ces types de motifs. Pour le motif de comparaison de lieux, la caractéristique P_2 est la caractéristique la plus importante. Elle évalue la prédominance de deux zones géographiques distinctes, qui tend à interpréter que l'unité de texte fait l'objet d'une comparaison.

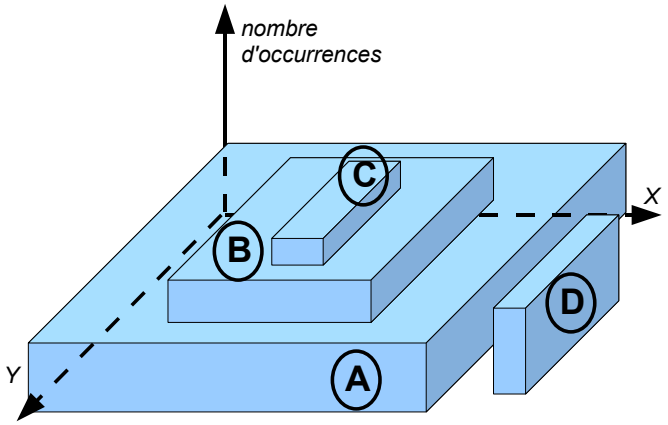


FIG. 7.2 – P_1 : (A) intersecte avec (B) et (C), mais pas avec (D). En supposant que ces quatre zones forment un ensemble d'ES d'une unité de texte, la valeur de P_1 pour celle-ci est de 3 (le nombre maximum d'occurrences en intersection) / 4 (le nombre total de zones) $= 0,75$

7.3.1 Propriété de dispersion

Prédominance d'une zone géographique Pour cette caractéristique, nous avons besoin d'évaluer un degré de prédominance d'une zone géographique par rapport à d'autres. L'algorithme 2 prend en entrée l'ensemble des ES de l'unité de texte étudiée (et qui correspond à la table de base de données géographique (T), et détermine la zone la plus intersectée (dans la figure 7.2, cette zone correspond à l'entité (C)). Une fonction SIG calcule le nombre d'intersections géométriques. La requête SQL géographique utilisée (ligne 5 de l'algorithme 2) :

```
1  SELECT sum(nb_occ) FROM T WHERE intersects(the_geom, (SELECT
2  the_geom FROM T WHERE nom_es='<Nom de l'ES>'));
```

se sert de la fonction *boolean intersects*(*geometry*, *geometry*) pour savoir si une région géographique en intersecte une autre (encodées dans le champ *geometry*). Enfin, la valeur de P_1 est un ratio fait entre les zones se trouvant dans la zone prédominante et leur nombre total.

```
1  T <- ensemble d'ES d'une unité de texte
2       (sous forme de table d'un SIG)
3  max_occ <- 0
4  Pour chaque enregistrement e de T Faire {
5       nb_occ <- Nombre d'enregistrements dont
6                 la géometrie s'intersecte avec e
7       Si (nb_occ > max_occ) Alors {
8            max_occ <- nb_occ
9       }
10 }
11 A_prev = max_occ / nombre d'enregistrements
```

Algorithme 2 : Algorithme de calcul de P_1

Le calcul de P_2 est similaire mais le but étant d'évaluer la prédominance de deux zones géographiques (afin de voir si une comparaison spatiale est faite dans l'unité de texte), le calcul d'intersection est effectué deux fois (algorithme 3).

```
T <- ensemble d'ES d'une unité de texte
     (sous forme de table d'un SIG)
max_occ <- 0
max_occ_2 <- 0
Pour chaque enregistrement e de T Faire {
     nb_occ <- Nombre d'enregistrements dont
               la géometrie s'intersecte avec e
     Si (nb_occ > max_occ) Alors {
          max_occ <- nb_occ
     }
}
nb_occ <- 0
Pour chaque enregistrement e de T Faire {
     nb_occ <- Nombre d'enregistrements dont la
          géometrie s'intersecte avec e et
          avec aucun enregistrement pris en
          compte lors du calcul de max_occ
     Si (nb_occ > max_occ_2) Alors {
          max_occ_2 <- nb_occ
     }
}
Aprev = max_occ_2 / (nombre d'ESs dans le texte - max_occ)
```

Algorithme 3 : Calcul de P_2

Calculs de distances La caractéristique D_1 consiste à utiliser la moyenne des distances entre les entités spatiales contiguës dans le texte. L'algorithme 4 (D1) montre comment est calculé le nombre de cas pour lesquels le calcul de la distance entre deux ES contiguës est au-dessous de cette moyenne.

La caractéristique D_2 est implémentée grâce à l'algorithme 4 (D2). Son but est de faire l'évaluation du rapport entre la taille des ES et la distance qui les sépare. Nous pouvons ainsi déterminer un degré de connexité ou d'adjacence. Nous choisissons de considérer que deux ES sont adjacentes quand la distance qui les sépare n'est pas inférieure à 4 fois la *taille* de la plus grande ES. La *taille* des ES est calculée en prenant la racine carrée de leurs surfaces, comme le montre la requête SQL utilisée dans la fonction *double echelle*(*int*) :

```
SELECT sqrt(area(the_geom))
FROM T WHERE pos like '%|"<position de l'ES>"|%';
```

Cette solution permet d'approcher la distance de côté moyenne de l'ES et d'obtenir une valeur en mètres comparable avec la distance. Le choix d'un facteur de 4 pour mettre en relation la taille des ES et la distance qui les sépare est un choix venant de l'expérience, à la lecture d'extraits de documents.

```
Algorithme D1 :
L <- Liste ordonnée des positions d'entités spatiales (ESs)
     de petite granularité
V : Vecteur
Pour chaque élément e de L (moins le dernier) Faire {
     // La fonction distance() donne la distance géographique
     Soit e' l'élément suivant e
     V.ajouter(distance(e,e'))
}
M <- moyenne des distances stockées dans V
Soit Nv la taille du vecteur V
num <- 0
Pour chaque élément a de V Faire {
     Si (a < M) Alors
          num <- num + 1
}
D_SFs <- num / Nv

Algorithme D2 :
V <- Vecteur ordonné des positions d'entités spatiales (ESs)
     de petite et moyenne granularités (et soit Nv la taile de V).
num <- 0;
// den prend le nombre de toutes les combinaisons
// de paires possible dans V
den <- Nv(Nv-1)/2
Pour a allant de 0 à Nv {
     echelle_a = echelle(a);
     Pour b allant de a+1 à Nv {
          echelle_b = echelle(b);
     max_echelle = max(echelle_a, echelle_b);
     Dab = distance(a,b);

     Si (Dab < 4*max_echelle) Alors {
        num <- num+1;
     }

      }
}
D_SFs <- num / den;
```

Algorithme 4 : Calculs de D_1 et D_2.

7.3.2 Propriété d'ordonnancement

L'algorithme 5 d'évaluation de l'ordonnancement (O) consiste à récupérer dans la liste L la position des petites ES dans le texte. Puis, pour chacune d'entre elles, le calcul de la distance à la première et à la dernière entité citées est sauvegardée dans un vecteur (V). Nous calculons ensuite un rapport qui donne une indication sur le rapprochement, d'ES en ES, du point d'arrivée (P_fin) et sur l'éloignement du point de départ (P_deb). Plus le pourcentage est fort, plus les ES vérifient ces conditions et sont *ordonnées*.

Pour exemple, prenons la figure 7.3 qui montre l'évaluation de deux conditions. À partir de la liste $L = \{deb, B, C, D, fin\}$ d'entités spatiales extraites d'une unité de texte, un calcul des distance à deb et à fin est effectué pour chacune d'entre elles ($d_{debB}, d_{Bfin}, d_{debC}, d_{Cfin}, d_{debD}, d_{Dfin}$). La première condition évaluée est l'ordonnancement de B à C. Nous remarquons que B est plus proche de deb et plus loin de fin que C. B se trouvant avant C dans la liste L, nous incrémentons la valeur de O. Par contre la deuxième condition évaluée montre que C est plus loin de deb et plus proche de fin que D, alors qu'il arrive avant dans la liste L. O n'est donc pas incrémenté dans ce cas.

7.3.3 Propriété de saillance

Le calcul de la saillance entre les entités spatiales contiguës dans le texte peut s'effectuer grâce à la fonction SIG $double(rad)$ $azimuth(geometry, geometry)$ qui retourne l'orientation du segment créé par les deux géométries en paramètre par rapport à l'axe vertical Sud-Nord.

La figure Figure. 7.4 illustre la valeur retournée par cette fonction et le découpage que nous avons effectué pour définir les huit orientations de base. Il pourra être intéressant d'essayer plus tard le découpage de la rose des vents proposé par Klippel [KDWH05] plus proche de la perception humaine du changement d'orientation.

Les poids choisis dans la matrice G (Algorithme 6) sont compris entre 0 et 1 et sont d'autant plus grands que le changement d'orientation est faible. Une moyenne des changements d'orientation est ainsi calculée et plus cette moyenne est grande plus la linéarité de l'ensemble d'ES est grande.

La fonction $Orientation()$ de l'algorithme 6 utilise la fonction SIG $azimuth()$.

La requête SQL suivante est utilisée pour calculer l'orientation du segment donné en paramètre :

```
1  SELECT azimuth(centroid(the_geom), (SELECT centroid(the_geom) FROM T
2  WHERE pos LIKE
3  '%|<position de la 2ème entité>|%')) FROM T WHERE pos LIKE
4   %'%|<position de la 1ère entité>|%';
```

```
L <- Liste ordonnée des positions d'entités spatiales (ES)
     de petite granularité
P_deb <- Position de la première entité de petite granularité
P_fin <- Position de la dernière entité de petite granularité

// V est un vecteur dont les éléments sont des paires
// (distance du point de départ, distance du point d'arrivée).
V : Vecteur (de paires de flottants)

Pour chaque élément e de L Faire {
     // La fonction distance() donne la distance géographique
     // entre l'ES de position P_deb et l'ES de position e.
     dist_deb <- distance(P_deb, e)
     dist_fin <- distance(e, P_fin)

     V.ajouter([dist_deb, dist_fin])
}

// V = {[d_deb1, d_fin1], [d_deb2, d_fin2], ..., [d_debn, d_finn]}
//          \        \       /      /
//           \        \     /      /
//            \        \   /      /
//             \        \/      /   ...
//              \       /\     /
//               \     /  \   /
//                \   /    >
//                  <    Condition 2
//              Condition 1

Soit Nv la taille du vecteur V
numérateur <- 0
dénominateur <- (Nv-1)+(Nv-2)+...+2+1 = Nv(Nv-1)/2

Pour a allant de 0 à Nv Faire {
     Pour b allant de a+1 à Nv Faire {
          d1 <- V[a][0]
          d2 <- V[a][1]
          d3 <- V[b][0]
          d4 <- V[b][1]
          // Condition 1 : (d1 < d3) et Condition 2 : (d2 > d4)
          Si ((d1 < d3) et (d2 > d4)) Alors {
               numérateur <- numérateur + 1
          }
     }
}
O_SFs <- numérateur / dénominateur
```

Algorithme 5 : Calcul de l'ordonnancement O.

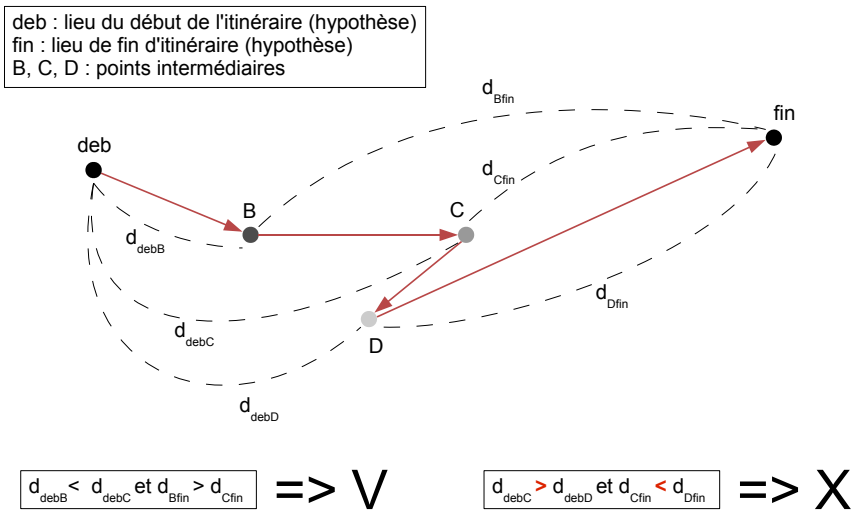


FIG. 7.3 – O : exemple de calcul des deux conditions. La première est vérifiée contrairement à la deuxième. Le résultat de O pour cet exemple est donc de 2/3 soit 67%. En effet, 2 conditions sont vérifiées (avec les points B et C, puis B et D) sur le nombre total de combinaisons possibles.

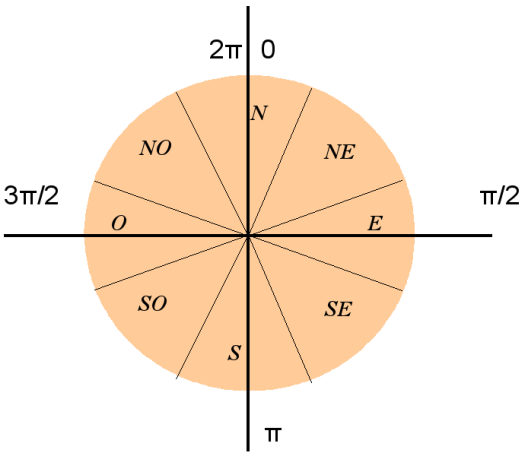


FIG. 7.4 – Rose des vents, orientation donnée en radians par la fonction SIG *azimuth*()

```
L <- Liste ordonnée des positions d'entités spatiales (ES)
     de petite granularité

// Graphe des coûts de passage d'une orientation à l'autre :
// ex. : G[N,O] = 0.5, G[E,SO] = 0.0
//              N   NE    E    SE   S    SO   O    NO
G  <-  {
             {1.0, 0.8, 0.5, 0.0, 0.0, 0.0, 0.5, 0.8}, //N
             {0.8, 1.0, 0.8, 0.5, 0.0, 0.0, 0.0, 0.5}, //NE
             {0.5, 0.8, 1.0, 0.8, 0.5, 0.0, 0.0, 0.0}, //E
             {0.0, 0.5, 0.8, 1.0, 0.8, 0.5, 0.0, 0.0}, //SE
             {0.0, 0.0, 0.5, 0.8, 1.0, 0.8, 0.5, 0.0}, //S
             {0.0, 0.0, 0.0, 0.5, 0.8, 1.0, 0.8, 0.5}, //SO
             {0.5, 0.0, 0.0, 0.0, 0.5, 0.8, 1.0, 0.8}, //O
             {0.8, 0.5, 0.0, 0.0, 0.0, 0.5, 0.8, 1.0}  //NO
       }
// En fonction de ces coûts une évaluation est faite du taux
// de «saillance» des entités.
Pour chaque élément e de L (moins le dernier) et
 e' l'élément suivant e Faire {
      V.ajouter(Orientation(e,e'))
}

Soit Nv la taille de V
numérateur <- 0.0
dénominateur <- Nv-1

Pour a allant de 0 à (Nv-1) Faire{
     numérateur <- numérateur + G[V[a]][V[a+1]]
}

S_SFs <- numérateur / dénominateur
```

Algorithme 6 : Calcul de la saillance S.

7.3.4 Calcul de la représentation associée

Un algorithme de calcul de représentation a été testé pour chaque motif à partir des résultats du calcul des caractéristiques. Il n'a pas encore donné lieu à une expérimentation poussée mais une première implémentation a donné des résultats encourageants pour quelques exemples.

En effet, pour chaque caractéristique calculée, les représentations des ES sont classées, suivant qu'elles vont dans le sens de la caractéristique, ou non. Par exemple pour la prédominance P_1, les ES (A), (B) et (C) de l'exemple figure 7.2 sont classés à part de l'ES (D). De cette manière, une fois que l'on connaît le motif grâce à la classification par SVM, on peut reconstruire une représentation géo-référencée à partir des ES sélectionnées pour les caractéristiques adéquates. Par exemple l'algorithme de représentation d'*itinéraire* est implémenté par la requête SQL :

```
SELECT GeomFromEWKB(buffer(makeline(centroid(the_geom)),
<épaisseur>)) FROM T WHERE <where_clause>;
```

où le champ *épaisseur* correspond à l'épaisseur de la ligne représentant l'itinéraire et

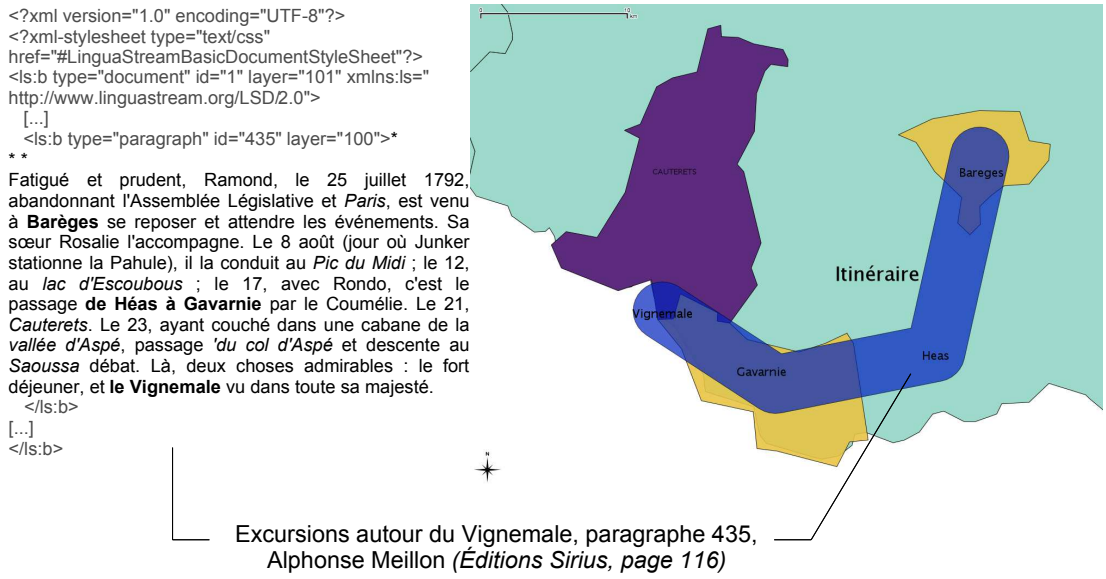


FIG. 7.5 – Exemple de représentation d'itinéraire, utilisée pour l'indexation.

le champ *where_clause* étant construit à partir du *gid* des ES validées à l'étape du calcul d'ordonnancement et de saillance (O et S). Cette requête construit alors une géométrie qui sert d'index pour l'unité de texte. La figure 7.5 montre un exemple de cette géométrie. Nous voyons que certaines entités nommées ne sont pas utilisées pour calculer la représentation finale, soit du fait de l'algorithme (pour *Paris*), soit à cause de problèmes de récupération des entités (manque de ressources ou introducteur spatial exotique comme pour *Pic du Midi, vallée d'Aspé*).

De la même manière, la requête

```
SELECT GeomFromEWKB(convexhull(boundary(collect(the_geom)))) FROM T WHERE
    <where_clause>;
```

construit une représentation pour une *description locale* ou pour une description de *point de vue*, le champ *where_clause* étant construit à partir des ES validées par D_1 et P_1.

Enfin la requête

```
SELECT GeomFromEWKB(geomunion((select convexhull(boundary(collect(the_geom
    )))
FROM T WHERE <where_clause_1>), (select convexhull(boundary(collect(
    the_geom)))
FROM T WHERE <where_clause_2>)));
```

construit une représentation pour une *comparaison de lieux*, les champs *where_clause_1* et *where_clause_2* étant construit à partir du calcul de prédominance P_2.

7.3.5 Expérimentation sur échantillon

Nous présentons ici la méthodologie de l'expérimentation mise en place pour tester l'implémentation de notre méthode de classification. Celle-ci porte seulement sur sa

faculté de classification (et non pas sur les représentations calculées pour l'indexation multi-niveaux).

Il convient tout d'abord de fixer une méthode qui délimite les unités de texte. Sachant que nous souhaitons avoir un échantillon assez conséquent, cette méthode doit être automatique et facilement implémentable. Notre choix s'est donc porté sur la structure *paragraphe* défini à l'étape de segmentation de PIV et correspondant à des morceaux de texte terminés par plus de deux retour-chariots. Nous avons finalement pris comme échantillon les paragraphes pour lesquels le système d'indexation intraphrastique de PIV détectait au moins cinq entités spatiales.

Un peu moins de 200 paragraphes ont été sélectionnés. Nous avons alors lu une partie de ces paragraphes (74 unités) et, afin d'évaluer notre processus de classification, nous les avons classés selon 5 catégories : itinéraire, description de point de vue, description locale, comparaison de lieux et une catégorie « nulle » pour les paragraphes ne faisant partie d'aucune de ces classes.

Notre base de paragraphes ainsi constituée est composée pour l'apprentissage de 12 éléments et notre base d'évaluation de 62 éléments. Ces bases sont très faibles mais s'expliquent par le fait que nous présentons ici une première évaluation et que les documents ont été supervisés par nos soins. Le résultat pour la précision de cette classification est de $43,5484\%$ (27 / 62). Nous avons lancé la classification sur le jeu d'entraînement et obtenu comme résultat 100% de réponses correctes. Le risque empirique, c'est-à-dire la somme des « erreurs », est donc nul. Une deuxième expérience a été menée afin de voir comment évolue la précision et ce risque empirique R_{emp}. Nous avons pris un jeu d'entraînement de 19 éléments (7 de plus ; il en reste 55 dans le jeu de test). La précision obtenue est de 47,27% (26/55) avec R_{emp} toujours à 0. Ces premiers résultats sont donc encourageants (une chance sur deux, pour 5 classes). Nous garderons cependant la première expérience pour l'analyse qui suit, celle-ci ayant un jeu d'entraînement plus réduit.

Nous avons analysé les erreurs de classification et avons remarqué qu'une des erreurs majeures est faite entre la description en *point de vue* et la description *locale* (27%) (figure 7.6). Cette erreur montre que des caractéristiques évaluant des propriétés linguistiques pourraient être nécessaires. En effet, nous avons vu que la description d'un point de vue est souvent basée sur une linéarisation, à la manière d'une « voyage virtuel », c'est-à-dire d'un itinéraire virtuel. Nos caractéristiques purement géographiques peuvent ne pas être assez déterminantes et l'ajout d'une évaluation du nombre de verbes de mouvements par rapport aux verbes *statiques* pourrait être intéressante. En effet il existe des travaux [TT97] montrant que l'utilisation de ce type de verbes est déterminante pour caractériser une description de point de vue ou d'itinéraire. Nous avons d'ailleurs aussi entrepris, avec un autre membre de l'équipe [LL07] une expérimentation dans ce sens.

Nous pouvons aussi remarquer que les erreurs entre l'itinéraire et la description locale arrivent en plus grand nombre quand on combine les deux sens d'erreur (description → itinéraire et itinéraire → description). Dans ce cas, ce sont les caractéristiques spatiales qui ne sont pas assez déterminantes. Des facteurs (comme celui proposé pour le calcul de D_2) nécessitent peut-être une paramétrisation plus fine.

Une autre solution est de trouver plus de caractéristiques déterminantes. Il faudrait

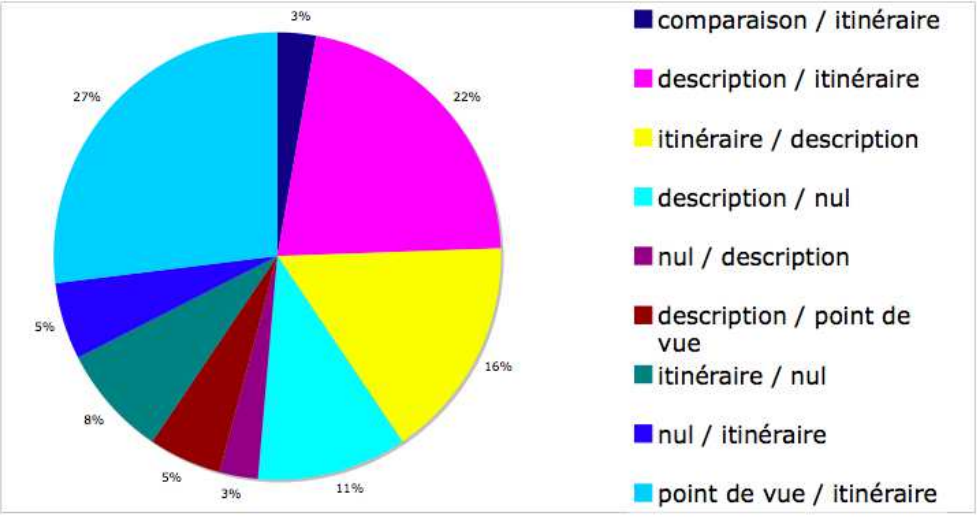


FIG. 7.6 – Répartition des erreurs de classification (classe correcte / classe trouvée).

par exemple trouver un moyen de calculer des caractéristiques à partir des propriétés intrinsèques aux ES, comme leur échelle ou le type de relations spatiales utilisées. En effet il est assez facile de récupérer ce genre de propriétés mais, par contre, il est plus difficile de les positionner sur un axe de l'espace multi-dimensionnel de classification.

La dernière expérimentation réalisée consiste à évaluer les moyennes trouvées pour les caractéristiques selon les différents motifs. Le tableau 7.4 montre les résultats trouvés. Nous pouvons remarquer que l'itinéraire a bien les caractéristiques d'ordonnancement O et de saillance S les plus fortes, même si les valeurs ne sont pas si hautes (0,293 et 0,343 alors que le maximum est 1). La valeur de P_2 (qui évalue la présence de deux groupes d'entités spatiales distantes) est la plus forte pour la comparaison de lieux. Par contre la valeur de P_1 semble trop élevée pour le motif itinéraire. La valeur de O est aussi forte pour la description de point de vue que pour l'itinéraire, mais ceci peut s'expliquer par le fait que ces deux descriptions se font de manière séquentielle. La description de point de vue correspond en effet à une sorte d'itinéraire virtuel. Néanmoins les valeurs obtenues correspondent globalement à nos attentent et expriment les différentes propriétés spatiales. L'amélioration de la catégorisation passerait plutôt par l'ajout de nouvelles caractéristiques.

7.4 Conclusion

Ce chapitre présente des travaux préliminaires concernant une méthode d'indexation multi-niveaux, basée sur une n en motifs spatiaux. La synthèse faite pour les unités de texte permet une indexation utilisant une information interprétée d'un niveau d'abstraction plus élevé que celui de l'indexation intraphrastique proposé dans le chapitre précédent. Ce travail fait partie d'un chantier visant une indexation par molécules

Motif \ Caractéristiques	P1	P2	D1	D2	O	S
Itinéraire	*0,514*	0,369	0,623	0,796	**0,293**	**0,343**
Description locale	0,524	0,491	0,476	0,861	0,163	0,177
Description de point de vue	0,384	0,344	0,448	0,461	*0,302*	*0,232*
Comparaison de lieux	0,467	**0,625**	0,425	0,9	0,283	0,225
nul	0,403	0,333	0,351	0,648	0,183	0,156

TAB. 7.4 – Caractéristiques et valeurs moyennes trouvées pour les motifs choisis.

géographiques complètes, comprenant un index temporel et un index thématique, ces molécules étant représentées sur plusieurs niveaux hiérarchiques du texte.

Nous proposons une méthode consistant à évaluer des caractéristiques afin d'associer des motifs spatiaux à des unités de texte. Cette association se fait à l'aide d'un système de catégorisation par apprentissage basé sur les SVM. Elle permet de fournir une représentation précise à ces unités de texte et ainsi de les indexer.

L'expérimentation a montré que les caractéristiques définies donnent des résultats encourageants, bien que des améliorations peuvent être apportées. En effet, il a été montré [Iha04] qu'un utilisateur ne fait confiance à un classifieur qu'à partir de 90% de réussite de classification.

D'une part, il serait intéressant d'intégrer des caractéristiques traitant les propriétés intrinsèques des entités spatiales. En effet, leurs tailles, leurs types, les relations qu'elles contiennent peuvent être de bons indicateurs.

D'autre part, il serait intéressant d'étudier les propriétés linguistiques liées aux différents motifs. Ces propriétés seraient un atout, surtout dans le cas de descriptions exprimées sans entité nommée géo-référençable. L'extrait de l'annexe A.2 par exemple, dans le deuxième paragraphe, n'en contient pas. D'autres évaluations à plus grande échelle seraient intéressantes à mener afin de valider définitivement l'ensemble de ces caractéristiques. Une expérimentation de cette indexation dans le cadre d'un système de RI reste aussi à faire, notamment afin d'évaluer les représentations proposées pour les motifs.

À plus long terme, une réflexion devra être menée sur la généralisation de l'abstraction faite avec cette indexation. En effet, la première indexation au niveau intraphrastique correspond au niveau n de l'index, et l'indexation présentée ici au niveau $n+1$. L'idée serait de proposer une méthode qui donnerait autant de niveaux d'indexation qu'il existe de niveaux d'abstraction possibles afin, à terme, d'arriver à synthétiser le corpus par œuvres entières. Pour illustrer ce point de vue, il serait par exemple intéressant de fusionner deux unités de texte contiguës, traitant pour la première d'un itinéraire d'un point A à un point B et pour la deuxième, d'un itinéraire d'un point B à un point C. Il en résulterait une indexation de niveau plus élevé représentée par la zone géographique de l'itinéraire de A à C. L'extrait de l'annexe A.1 est une bonne illustration de cet exemple. L'auteur se trouve à *Paris* au début du premier paragraphe, puis, partant pour Bordeaux,

il passe par *Tours* puis *Orléans*. Dans le deuxième paragraphe, les premières entités citées sont *pays du Poitou*, *Angoulême* et la dernière *Bordeaux*. Les deux paragraphes peuvent être interprétés comme des itinéraires séparés à un premier niveau d'abstraction alors qu'il n'en s'agit que d'un seul. Une abstraction supplémentaire pourrait alors les indexer comme un seul itinéraire de Paris à Bordeaux.

D'autres combinaisons moins évidentes seraient à envisager afin de remonter à une représentation par document. L'extrait de l'annexe A.2 montre un cas plus difficile à indexer. Le premier paragraphe est classé comme itinéraire (de *Nay* à *Bétharram*) alors que le deuxième est une description locale. Puis l'itinéraire continue de *Saint-Pé de Bigorre* jusqu'à *Lourdes*.

Ces travaux qui ont débuté avec l'objectif de proposer un système d'indexation multi-niveaux dégagent donc des problématiques complexes de catégorisation, auxquelles nous tentons d'apporter un début de réponse *via* la construction de caractéristiques spatiales et l'utilisation de techniques de classification. Cette section conclue notre partie contribution. La partie suivante fait la synthèse des améliorations possibles pour un système d'indexation multi-niveaux par catégorisation en motifs.

Quatrième partie

Conclusion

Chapitre 8

Conclusion générale

Sommaire

8.1 Synthèse

Les travaux présentés dans ce manuscrit se placent à l'intersection de trois grands domaines de recherche que sont le Traitement Automatique du Langage Naturel, la Recherche d'Information, le Raisonnement Spatial Qualitatif et les Systèmes d'Information Géographique. La problématique choisie, de l'accès à l'information spatiale contenue dans des corpus textuels, nécessite en effet une approche pluri-disciplinaire afin de mener à bien des propositions de solutions opérationnelles. Des projets existants tels que GIPSY ou SPIRIT ont jeté les bases pour un système spécialisé dans le spatial et nous ont servi comme point de départ.

L'axe de recherche choisi pour ce faire est la conception d'un système de recherche spécialisée, capable d'extraire automatiquement l'information spatiale recherchée dans le cadre d'une interrogation. Nous avons en effet considéré l'information géographique comme étant une molécule construite à partie d'une composante spatiale, d'une composante temporelle et d'une composante thématique ou phénomène (figure 8.1(a)). L'objectif est alors de localiser les documents dans un territoire précis, à une période donnée et de retrouver les thèmes majeurs abordés. Nous pouvons alors aller plus loin en regroupant les trois composantes afin de dégager le contexte choisi par l'auteur (comme par exemple celui d'une description d'itinéraire).

Cependant, nous avons restreint notre champ d'étude à la composante spatiale afin de réaliser un processus complet d'indexation et de recherche d'information. Notre travail consiste alors à récupérer la composante spatiale, d'abord au niveau du syntagme, puis à un niveau plus élevé d'abstraction, celui des unités de textes allant du paragraphe à l'œuvre complète.

Nous avons en effet imaginé un processus d'indexation multi-niveaux, composé d'une couche par niveau d'abstraction. Le premier niveau d'abstraction consiste à interpréter et indexer l'information spatiale au niveau des syntagmes nominaux. Ce premier travail permet d'avoir plusieurs interprétations d'entités, décorrélées entre elles. Le deuxième niveau consiste à regrouper ces syntagmes extraits dans des unités logiques du texte, l'idée étant que ce regroupement fait sens dans la mesure où l'auteur les utilise dans un contexte particulier. Les étapes successives d'abstraction ont pour but de retrouver ce contexte. Ainsi, le dernier niveau d'abstraction peut *in fine* synthétiser l'information jusqu'à une œuvre entière.

Pour ce faire, nous avons imaginé une méthode d'indexation par classification, en définissant des motifs spatiaux, spécifiques au corpus (itinéraire, description locale, points de vue, comparaison de lieux). Cette classification est guidée par l'évaluation de caractéristiques spatiales propres aux unités de texte indexées.

Les travaux d'expérimentation (processus d'EI et de RI de PIV) ont permis d'évaluer l'indexation intraphrastique (ainsi que la méthode de recherche d'information attenante) et une partie de l'indexation par classification. Nous avons alors montré qu'un système alliant d'une part les outils du Traitement Automatique du Langage Naturel, guidés par le Raisonnement Spatial Qualitatif et d'autre part les outils basés « données-structurées » tels que les Systèmes d'Information Géographique améliore l'accès aux documents, du point de vue spatial et, en particulier, est capable d'augmenter l'efficacité de systèmes de Recherche d'Information classiques dans le cas de requêtes mixtes (spatiales + thématiques). D'autre part, des résultats encourageants ont été présentés pour l'indexation par classification, permettant d'envisager ce genre d'indexation dans un système de recherche spécialisé. Les résultats de nos travaux ont amené la Médiathèque Intercommunale à Dimension Régionale à proposer une intégration d'une version industrialisée de notre prototype au sein de l'appel d'offre concernant son système d'information[67].

Notre apport se situe donc, en recherche d'information spécialisé, dans la manière de récupérer les syntagmes spatiaux en se basant sur notre modèle pivot. Nous proposons un système d'extraction d'information spatiale plus poussé que celui de SPIRIT par exemple, dans la mesure où nous interprétons des entités complexes exprimées en langage naturel en vue d'une indexation novatrice. Le tableau 8.1 compare les projets existants avec notre contribution et montre que celle-ci est la plus aboutie en ce qui concerne la gestion de la composante spatiale. En particulier, nos apports principaux se trouvent au niveau de l'extraction d'information spatiale (*via* notre traitement sémantique), de l'interprétation des ES_R (*via* le raisonnement spatial qualitatif), de l'indexation spatiale (et son approche multi-niveaux, permettant un accès aux documents d'un nouveau type pour ce genre de corpus) et de l'appariement pour la recherche d'information. En effet, les index créés sont manipulables par des outils dédiés à l'information géographique, ce qui nous a permis de proposer un appariement efficace et précis, se basant sur les concepts utilisés en recherche d'information (de précision, d'importance) mais en l'adaptant au spatial, permettant d'ordonner les résultats retournés selon la pertinence spatiale.

De plus, dans le domaine de la géomatique, une réflexion sur le calcul de représenta-

[67] http://www.klekoon.com/boamp/BOAMP_3_Detail.asp?ID_appel=661116

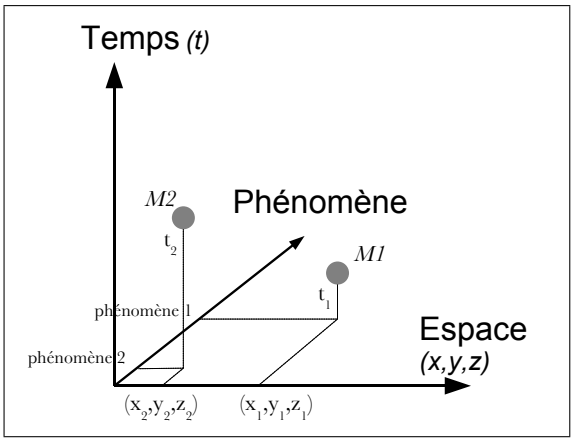
Projets	entités spatiales	index	requêtes	espace	temps	thème
GIPSY	simples avec quelques indications	1 géométrie par doc	∅	✓	∅	∅
Spirit	de type adresses	1 géométrie par page web	entités nommées avec choix d'indication qualitative	✓	∅	✓
GeoSem	complexes	un trait sémantique par entité	formulaire à 3 champs	✓	✓	✓
PIV	complexes (ES_A, ES_R)	1 géométrie par entité, par unité de texte	requête en texte libre	✓	∅	∅

TAB. 8.1 – Comparaison des projets existants étudiés et du projet PIV.

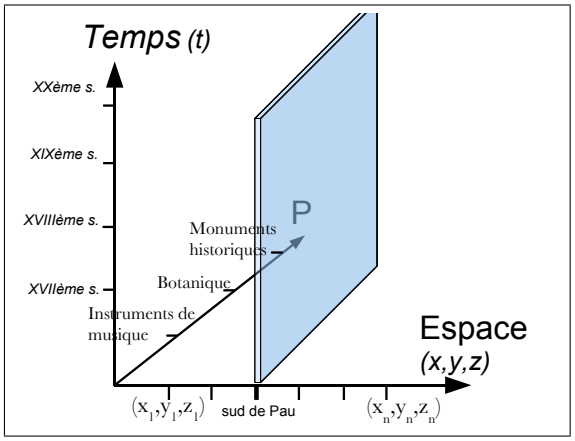
tion géo-référencée d'entités complexes a aboutie en un algorithme robuste, permettant de proposer une représentation pour n'importe quelle entité définie avec le modèle pivot. Il est envisageable d'intégrer cet algorithme dans les fonctions SIG afin de proposer des représentations pour des entités floues.

Cette thèse tente donc d'apporter des éléments nouveaux concernant la recherche d'information spécialisée dans le spatial et la représentation d'entités complexes comprenant des relations qualitatives dans les outils SIG. Nous avons tenté de montrer que la réunion de plusieurs domaines de recherche peut produire des résultats novateurs. Les possibilités offertes par la géomatique vont amener les systèmes de recherche d'information spécialisés (comme les GED par exemple) à de plus en plus utiliser et intégrer ce genre d'outils. Des projets comme SPIRIT ou les systèmes de recherche d'information spatiale en général pourraient intégrer notre méthode d'indexation ainsi que la méthode d'appariement correspondante.

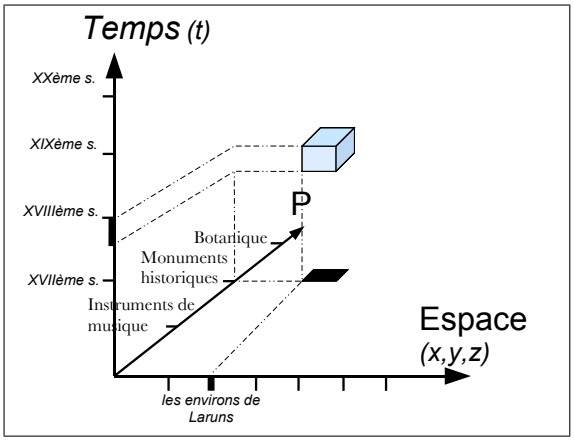
À l'inverse, les mécanismes présents dans la recherche d'information et les travaux d'analyse linguistique, adaptés à manipuler de l'information exprimée de manière qualitative, pourront éventuellement être intégrés dans les outils SIG. Ils pourraient alors améliorer leurs systèmes de requêtage et plus globalement leur gestion de l'information qualitative.



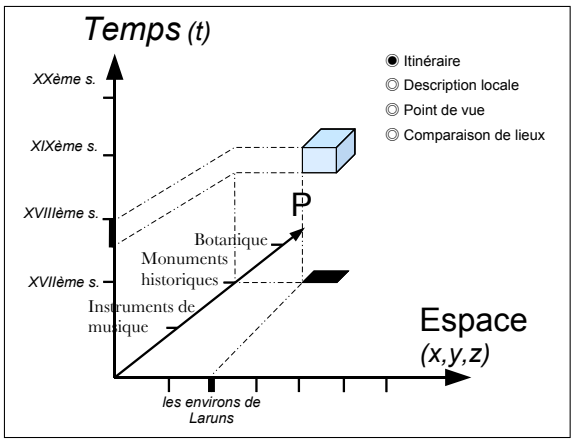
(a) Représentation originale de l'espace géographique.



(b) Tous les thèmes à toutes les époques parlant du sud de Pau.



(c) Les monuments historiques d'avant le XVIIIème siècle aux environs de Laruns.



(d) Des *itinéraires* parcourant les monuments historiques d'avant le XVIIIème siècle aux environs de Laruns.

FIG. 8.1 – Représentation imagée de l'espace géographique et trois exemples de requêtes (de la plus simple à la plus complexe).

8.2 Perspectives

Mise à part les améliorations techniques à court terme détaillées dans la synthèse du chapitre 6 (telles que la désambiguisation d'entités nommées ou l'amélioration des représentations géo-référencées pour les entités spatiales relatives), nos travaux sont le chantier d'un système de recherche d'information géographique, devant proposer une indexation sur les trois facettes que sont l'espace, le temps et le phénomène. La figure 8.1(b) montre une requête, dans l'espace géographique, pouvant être interprétée aujourd'hui par le système PIV (un utilisateur s'intéressant au *sud de Pau*). À terme, il est envisagé de pouvoir répondre à une requête du type *Les monuments historiques d'avant le XVIII*^{e} *siècle aux environs de Laruns*, représentée en figure 8.1(c). L'indexation multi-niveaux permettrait même de proposer une interrogation plus pointue à l'utilisateur. La figure 8.1(d) montre que l'exemple précédent peut être modifié par la restriction suivante : nous recherchons des ***itinéraires*** *parcourant les monuments historiques d'avant le XVIII*^{e} *siècle aux environs de Laruns*. De plus, un champ de recherche pourrait ajouter une restriction sur la granularité des résultats attendus, de l'ordre du paragraphe jusqu'au document entier.

Néanmoins, avant d'arriver à ces résultats des améliorations sont à faire pour le cas de la composante spatiale d'une part et, d'autre part, pour les deux autres composantes constituant la molécule géographique.

Perspectives pour le spatial La méthode d'indexation a été développée pour 2 niveaux comme travail de prospection. Cependant, l'idée finale est de proposer une indexation multi-niveaux complète, qui d'abstraction en abstraction, de synthèse en synthèse, indexe une œuvre entière. Pour atteindre cet objectif, il reste à concevoir des méthodes d'agrégation de motifs, une fois que ceux-ci sont définis à un premier niveau d'abstraction. En effet, une réflexion doit être menée sur la synthèse de plusieurs motifs entre eux. Par exemple, deux itinéraires seront synthétisés *a priori* de manière triviale en un itinéraire englobant. Mais comment synthétiser un itinéraire avec une description ou une comparaison de lieu ? Une méthode simple peut être de considérer que l'on synthétise les motifs 2 à 2. Par contre, si l'on veut regrouper les motifs par groupes quand ils semblent former un ensemble logique, cette problématique semble bien plus difficile.

Il reste aussi à se pencher sur le problème du découpage des unités de texte en considérant que le corpus soit toujours composé de données non-structurées. En effet ce découpage se fait selon un procédé simple sur du texte brut. Les groupes de phrases séparés par deux retours-charriots sont considérées comme des paragraphes. Cependant, d'autres cadres que ceux définis par cette logique pourraient être pertinents, comme par exemple la définition de cadres à l'aide d'analyse sémantique. Les travaux [BHDB+03] présentent à ce sujet une perspective intéressante. Ils proposent de tenir compte des syntagmes spatiaux ou temporels introductifs afin de déterminer des unités de texte parlant d'un endroit ou d'un moment précis.

D'autres ouvertures sont sujettes à réflexion à partir de ces travaux. Il serait par exemple pertinent de s'intéresser à une approche d'indexation multilingue. En effet l'indexation par classification ne nécessite que peu de dépendances à la langue française.

Celles-ci se trouvent au niveau de la phase d'indexation intraphrastique au cours de l'instanciation du modèle pivot. Il est envisageable, du moins pour les langues indo-européennes, de bâtir des lexiques d'indicateurs spatiaux pour chaque relation et de constituer une grammaire adéquate pour la détection de patrons. Ce travail n'est pas très coûteux (les 3 premiers modules du tableau E.2) et permettrait d'expérimenter notre approche par classification pour d'autres langues. De plus, les éléments de l'index servant à l'appariement n'étant constitués que de représentations spatiales, nous pourrions avoir facilement un système de recherche documentaire fonctionnant sur un corpus multilingue.

Perspectives pour la molécule géographique Des travaux pour les deux autres composantes de la molécule géographique ont été proposés. Pour la composante temporelle, une modélisation et une méthode d'indexation ont été réalisées [LPLGS07], se basant sur un modèle pivot similaire à celui présenté dans ce document mais pour le temps calendaire. Des primitives temporelles ont été définies (entités temporelles absolues de type calendaires) ainsi que des relations qualitatives (basées sur les travaux de J.F.Allen [All91]). L'approche d'indexation par classification peut ensuite être envisagée de la même façon. Enfin, nous pensons que la problématique d'extraction de phénomène peut être résolue grâce aux approches conceptuelles de la recherche d'information. Les travaux de A. Zouaq [ZFN06] notamment permettent la génération automatique d'ontologies en utilisant un traitement linguistique, afin de récupérer les thèmes principaux de documents et de les indexer. Un dernier travail consistera en l'unification de ces 3 indexations pour créer le système final. Les travaux de GeoSem proposant un système de recherche selon les 3 axes spatiaux, temporels et thématiques pourront être pertinents à envisager [BE05]. Il est à noter que l'architecture distribuée (sous forme de services web), choisie pour implémenter les différentes briques du système, permettra une évolution plus aisée.

Un premier essai a été réalisé dans [SBLG07b] à propos d'une combinaison de notre indexation spatiale (au premier niveau) avec une indexation par les statistiques pour les autres composantes de l'information. Nous avons imaginé un moyen d'intégrer deux approches au sein d'un même système d'interrogation afin d'améliorer les résultats lors de requêtes spatio-thématiques. L'idée est de subdiviser une requête en deux sous-requêtes (figure 8.2) ; la sous-requête spatiale et la sous-requête thématique. La sous-requête spatiale contient les ESs identifiées par la chaîne de traitement linguistique. La sous-requête thématique contient les termes restants de la requête (temps, événement). Comme schématisé dans la figure 8.2, « les environs de Laruns » et « instrument de musique du XIXème siècle » représentent respectivement la sous-requête spatiale et la sous-requête thématique de la requête exemple.

Les deux sous-requêtes sont ensuite soumises au système supportant l'approche appropriée. Le résultat final est ensuite construit en faisant une intersection des deux ensembles de documents sélectionnés par PIV et l'approche classique. Le classement final est basé sur celui obtenu par PIV : chaque document classé dans l'ensemble de PIV est ajouté à l'ensemble final s'il est également classé dans l'ensemble retourné par le système

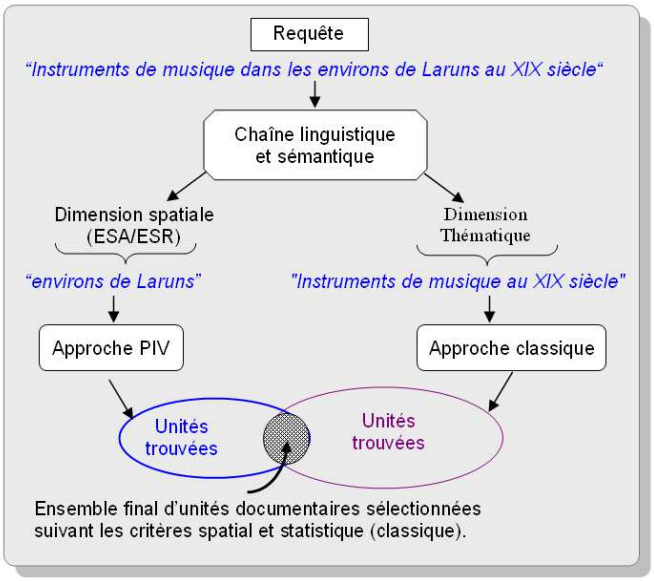


FIG. 8.2 – Combinaison des approches spatiales et thématiques.

Toutes les requêtes	P@5	P@10	P@15	Nombre de réponses
	A) Combiner les résultats de l'approche spatiale et classique			
Avg	**0.70**	**0.50**	**0.43**	**25.75**

TAB. 8.2 – Résultats de la combinaison de PIV et de l'approche classique.

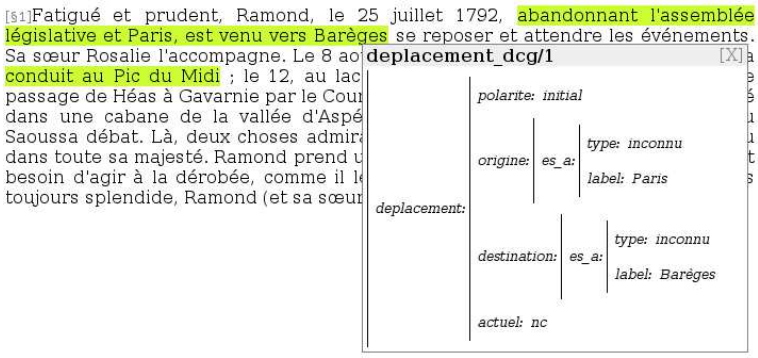


FIG. 8.3 – Analyse linguistique sur le même exemple que celui de la figure 7.5. Travail sur les verbes de mouvement [LGN07].

classique. Le détail des résultats obtenus en utilisant les requêtes précédentes (espace + thème) suivant cette stratégie sont donnés dans la Table 8.2.

Les résultats confirment l'hypothèse suivante ; combiner l'approche spatiale basée sur l'extraction et la comparaison des entités spatiales absolues et relatives avec l'approche thématique basée sur les statistiques permet d'améliorer les performances du système en classant plus de documents pertinents. Par exemple, aux cinq premiers documents, la précision atteint 70% quand les approches classique est spatiale sont combinées ; alors qu'elle n'est respectivement que de 48% et 15% quand les deux approches sont utilisées séparément. Cependant, on peut noter le nombre réduits de documents sélectionnés à cause de la méthode « prudente » de combinaison adoptée (intersection simple) : par exemple, dans le cas d'une de nos requêtes, l'approche combinée sélectionne seulement 4 unités documentaires alors que l'approche classique en retourne 233 et l'approche PIV 724. L'amélioration de la précision au niveau des premiers documents restitués s'accompagne ainsi d'une diminution au niveau du rappel.

Un domaine ouvert concerne donc le problème de fusion des ensembles de résultats pour optimiser aussi bien la précision dans les premiers documents que le rappel. Ceci nécessite l'étude d'opérateurs plus complexes (union, avec intégration pondérée de la pertinence spatiale et statistique, par exemple) que la simple intersection utilisée ici. Les travaux de [MSA05] pourraient être pertinents ici.

Une autre évolution possible de nos travaux est un travail sur la description d'itinéraires. Nous avons vu en effet qu'un des motifs définis dans nos préconisations est celui de l'itinéraire et qu'il a la particularité d'avoir des caractéristiques spatiales et temporelles. Il peut être donc considéré comme une molécule géographique à part entière, selon la définition que nous avons adoptée. Des travaux annexes menés dans notre équipe [LGN07] illustrés dans la figure 8.3 montrent comment sont utilisés les verbes de mouvement pour détecter et interpréter les itinéraires, cette analyse se faisant sur le même corpus que le notre. L'exemple de la figure 8.3 montre que la phrase *abandonnant l'assemblée législative et Paris, est venu vers Barèges* est détectée comme un morceau d'itinéraire grâce

aux verbes *abandonner* et *venir* qui sont respectivement des verbes initiaux et terminaux de mouvement. Une évolution intéressante serait alors de coupler nos efforts de caractérisation d'un itinéraire avec ces méthodes linguistiques afin d'obtenir un système de détection, d'interprétation, de représentation et *in fine* d'indexation d'itinéraires plus précis. C'est dans cette voie que nous aimerions poursuivre nos recherches, en utilisant à la fois les compétences de géomatique et de linguistique pour proposer de nouvelles méthodes de détection et d'indexation d'itinéraires, en faisant la synthèse spatiale d'unités de textes. De manière plus générale, une combinaison de la géomatique avec les domaines de la recherche d'information nous parait être une orientation porteuse et intéressante à envisager.

Annexe A

Extraits de corpus utilisés dans le cas d'étude

A.1 Extrait de l'exemple 1

PARIS 6 AOÛT 1833

II y a sept ans à pareil jour j'étais dans l'habitation la plus haute de l'Europe, chez les bons pères du Mont Saint Bernard. Depuis lors que d'évènements ! M'éloignant par dégoût de la politique de ce qu'on appelle les honneurs, ne cherchant qu'une existence douce et paisible, je n'ai pas moins eu ma bonne part de tribulations. Aujourd'hui même après des jours d'orage, mon âme conserve encore les empreintes du malheur. La distraction est un remède à bien des maux ; mais que faire à Paris à l'époque où nous sommes ? Toutes les personnes de la société ont quitté la capitale. Il reste encore quelques mois de la belle saison que je puis consacrer à des excursions lointaines. J'ai parcouru les Alpes, j'aimerais à les comparer aux Pyrénées, et surtout à connaître cette vie des eaux si recherchée par le grand monde. Ma résolution est bientôt prise. Andiamo ! À six heures du soir je monte en diligence pour Bordeaux, pour ce long voyage au milieu de la poussière de l'été. C'est pourtant la route de France parcourue avec plus de rapidité qu'aucune autre ; aussi étions nous bercé par les mouvements ondulés de la voiture ! Nous voyons Orléans en bonnet de nuit, et dans la journée, Tours devenue colonie anglaise. Peu après, sur la plus belle route du monde les chevaux s'effrayent, nous perdons le centre de gravité comme l'aurait dit Sterne, et nous voilà sens dessus dessous. Dans cette culbute il eut été beau de nous voir dénicher par la portière en face du ciel ! Nous en sommes quittes pour quelques contusions, mais une dame du coupé est grièvement blessée, et notre conducteur écrasé sous les bagages crache le sang. Le tableau que nous offrons est digne de pitié ; beaucoup de curieux arrivent, mais aucun n'offre de secours ; une anglaise seule s'intéressant au malheur, fait accepter une place dans sa voiture à la dame blessée.

Enfin à l'aide d'une voiture de rechange, on replace les bagages, et nous repartons au grand galop, peu rassurés contre les chances d'une nouvelle chute. Nous fesons quatre lieues à l'heure et traversons le plus vite possible ce triste pays du Poitou, bientôt Angoulème se présente comme un bouquet de verdure. Nous arrivons sans autre événement

sur les rives de la Dordogne. Là, voiture, chevaux, voyageurs, tout est placé sur le bac et à l'aide de rames nous traversons lentemens cette large rivière. Jusqu'à ce jour des difficultés diies à la nature, plus encore, des rivalités de voisinage ont empêché d'y établir un pont. Le génie de l'homme finira par vaincre le premier obstacle et les habitants de ces contrées sentiront enfin la nécessité de créer un passage prompt et facile en tous temps ; les jalousies du commerce céderont, il faut l'espérer, aux loix de la raison. Un beau soleil levant, au mois d'août, donnait à cette traversée un intérêt de plus. Le ciel était pur et brillant et la nature entière présentait à nos yeux un spectacle grand et sublime Nous débarquons sur la rive opposée et remontant en voiture, nous parcourons un pays riche et peuplé situé entre la Garonne et la Dordogne que les Gascons appellent, pays entre deux mers. À six heures du matin nous descendions les coteaux de la Bastide et bientôt nous passons sur le fameux pont de Bordeaux, après avoir été enfermés dans une voiture pendant trois nuits et deux jours, ayant éprouvé toute la fatigue d'un long et pénible voyage, d'une chute et d'une poussière abominable.

Voyages inédits dans les Pyrénées (par.90), Édition PyréGraph

A.2 Extrait de l'exemple 2

Ce vieux château commande la plaine de Nay, et fut toujours une des plus nobles seigneuries du Béarn. La route continue ensuite sur la rive gauche du Gave, traverse le village, et la belle plaine d'Igon pour arriver à Bétharram : M. l'abbé Menjoulet donne de ce nom plusieurs étymologies peu sûres : Beit Harram, qui signifie demeure sacrée, en langue arabe ; Beth aram, en langue hébraïque : Maison du Très-Haut ; Beth Arram, en béarnais : beau rameau, qui est une explication plus plausible. Quoi qu'il en soit, la chapelle, le calvaire et le vieux pont' tapissé de lierre, sont curieux à voir.

Sur l'emplacement de la chapelle fleurit une légende qui perpétue à sa manière la vénération ancestrale des sources et des grandes pierres. D'innocents petits bergers en conduisant leurs brebis au bord du Gave, parmi les rochers du bas de la montagne, aperçurent une petite lumière qui guida leurs pas vers une belle image de Notre-Dame. Quand les habitants du village voulurent placer cette merveilleuse figure dans un modeste oratoire dressé en un lieu plus accessible, de l'autre côté du Gave, l'image se transportait aussitôt dans les rochers d'en face. Si bien, qu'à la fin, il fallut se rendre au bon plaisir si manifeste de la madone. En récompense de cette obéissance, Notre Dame combla de grâces et de consolations ineffables la foule des pèlerins qui lui faisaient visite.

A deux kilomètres de ce sanctuaire, il ne faut pas négliger d'aller parcourir les belles grottes de Bétharram, toutes parées de concrétions aux formes bizarres. En s'infiltrant par les fissures supérieures, les eaux terribles et capricieuses ont creusé de nombreuses galeries et d'étranges cavernes. Sous l'érosion mécanique et par la corrosion chimique, voûtes et sol se sont peuplés de stalactiques et de stalagmites qui prêtent à ces grottes un fantastique aspect.

A Saint-Pé de Bigorre, le défilé s'élargit, les premiers contreforts des montagnes se soulèvent, le Gave gronde plus torrentueux, et les maisons de la cité sainte de Lourdes

se serrent comme un troupeau apeuré autour de son vieux Castellum, connu plus tard sous le nom de « Castel de Mirambal ». Ici, un souvenir.

Excursions autour du Vignemale, Alphonse Meillon, Édition Sirius

Annexe B

Schéma XML du modèle pivot

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:gml="http://www
    .opengis.net/gml">
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/gml.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/dynamicFeature.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/topology.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/coverage.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/coordinateReferenceSystems.xsd"
        />
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/observation.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/defaultStyle.xsd"/>
    <xs:import namespace="http://www.opengis.net/gml" schemaLocation="http:
        //schemas.opengis.net/gml/3.1.1/base/temporalReferenceSystems.xsd"/>
    <xs:element name="doc">
        <xs:complexType id="docType">
            <xs:sequence>
                <xs:element ref="es"/>
            </xs:sequence>
            <xs:attribute name="doc_original" type="xs:NCName" use="required
                "/>
            <xs:attribute name="ls_sem" type="xs:NCName" use="required"/>
            <xs:attribute name="doc_ocr" type="xs:NCName" use="required"/>
            <xs:attribute name="id_oeuvre" type="xs:integer" use="required"/
                >
            <xs:attribute name="id_doc" type="xs:integer" use="required"/>
        </xs:complexType>
    </xs:element>
    <xs:element name="es">
        <xs:complexType id="esType">
            <xs:sequence>
                <xs:element name="texte"/>
                <xs:choice>
                    <xs:element ref="esa"/>
                    <xs:element ref="esr"/>
                </xs:choice>
            </xs:sequence>
```

```
32              <xs:attribute name="es_id" type="xs:integer" use="required"/>
33              <xs:attribute name="par_id" type="xs:integer" use="required"/>
34              <xs:attribute name="type_es" type="xs:string" use="required"/>
35              <xs:attribute name="negation" type="xs:boolean" use="required"/>
36          </xs:complexType>
37      </xs:element>
38      <xs:element name="esa">
39          <xs:complexType id="esaType">
40              <xs:sequence>
41                  <xs:element ref="texte"/>
42                  <xs:element ref="obj_esa"/>
43              </xs:sequence>
44              <xs:attribute name="type_en" type="xs:string" use="required"/>
45              <xs:attribute name="nom" type="xs:string" use="required"/>
46          </xs:complexType>
47      </xs:element>
48      <xs:element name="esr">
49          <xs:complexType id="esrType">
50              <xs:sequence>
51                  <xs:element ref="texte"/>
52                  <xs:element ref="relation"/>
53                  <xs:choice>
54                      <xs:element ref="esa"/>
55                      <xs:element ref="esr"/>
56                  </xs:choice>
57              </xs:sequence>
58          </xs:complexType>
59      </xs:element>
60      <xs:element name="texte" type="xs:string"/>
61      <xs:element name="relation">
62          <xs:complexType id="relationType">
63              <xs:sequence>
64                  <xs:choice>
65                      <xs:element ref="figure_geo"/>
66                      <xs:element ref="adjacence"/>
67                      <xs:element ref="orientation"/>
68                      <xs:element ref="distance"/>
69                      <xs:element ref="inclusion"/>
70                      <xs:element ref="intersection"/>
71                  </xs:choice>
72              </xs:sequence>
73          </xs:complexType>
74      </xs:element>
75      <xs:element name="figure_geo">
76          <xs:complexType>
77              <xs:attribute name="extremite" type="xs:integer" use="required"/
                    >
78          </xs:complexType>
79      </xs:element>
80      <xs:element name="adjacence">
81          <xs:complexType>
82              <xs:attribute name="type_prox" type="xs:integer" use="required"/
                    >
83          </xs:complexType>
84      </xs:element>
85      <xs:element name="orientation">
86          <xs:complexType>
87              <xs:attribute name="pt_card" type="xs:integer" use="required"/>
88          </xs:complexType>
89      </xs:element>
90      <xs:element name="distance">
91          <xs:complexType>
```

```
92              <xs:attribute name="valeur" type="xs:integer" use="required"/>
93              <xs:attribute name="unité" type="xs:string" use="required"/>
94          </xs:complexType>
95      </xs:element>
96      <xs:element name="inclusion">
97          <xs:complexType>
98              <xs:attribute name="valeur" type="xs:integer" use="required"/>
99              <xs:attribute name="unité" type="xs:NCName" use="required"/>
100         </xs:complexType>
101     </xs:element>
102     <xs:element name="intersection">
103         <xs:complexType>
104             <xs:attribute name="unité" type="xs:NCName" use="required"/>
105         </xs:complexType>
106     </xs:element>
107     <xs:element name="obj_esa">
108           <xs:complexType>
109            <xs:sequence>
110                  <xs:element ref="type"/>
111                <xs:element ref="representation"/>
112            </xs:sequence>
113           </xs:complexType>
114     </xs:element>
115     <xs:element name="type">
116         <xs:complexType>
117             <xs:attribute name="table" />
118             <xs:attribute name="Where"/>
119          </xs:complexType>
120     </xs:element>
121     <xs:element name="representation">
122         <xs:complexType>
123            <xs:choice>
124                <xs:element ref="gml:Point"/>
125                <xs:element ref="gml:Polygon"/>
126                <xs:element ref="gml:MultiPolygon"/>
127            </xs:choice>
128          </xs:complexType>
129     </xs:element>
130 </xs:schema>
```

Annexe C

Lexiques utilisés dans le processus d'analyse linguistique

De la ligne 3 à la ligne 93, ce lexique permet de gérer les différentes unités de distance utilisées dans les documents. Pour chaque unité, la correspondance est donnée en mètres. Il permet alors de faire les calculs de représentations des ES_R contenant des relations de distance.

De la ligne 94 à la ligne 228, ce lexique permet de catégoriser le type des entités. Des patrons sont définis pour les différents types et ils sont classés en quelques catégories identifiables, reliées à des ressources géographiques correspondantes.

De la ligne 229 à la ligne 318, ce lexique identifie tous les indicateurs correspondant aux relations spatiales d'adjacence, d'orientation et d'inclusion. La détection de ces indicateurs permet l'instanciation du modèle pivot pour ces relations.

De la ligne 319 à la ligne 324, cette règle définit les entités spatiales candidates, au début du processus sémantique (c'est-à-dire les mots commençant par une majuscule).

```
1  <?xml version="1.0" encoding="UTF-8"?>
2 <tokenMarker>
3   <rule caseInsensitive="true">
4     <pattern>(km)|(kilom(è|e)tre(s?))</pattern>
5     <featureSet>
6       <unite>
7         <type>distance</type>
8         <original>km</original>
9         <equivalence>
10          <unite>m</unite>
11          <valeur>1000</valeur>
12        </equivalence>
13      </unite>
14    </featureSet>
15  </rule>
16  <rule caseInsensitive="true">
17    <pattern>(hm)|(hectom(è|e)tre(s?))</pattern>
18    <featureSet>
19      <unite>
20        <type>distance</type>
21        <original>hm</original>
22        <equivalence>
```

```
          <unite>m</unite>
          <valeur>100</valeur>
        </equivalence>
      </unite>
    </featureSet>
  </rule>
  <rule caseInsensitive="true">
    <pattern>(hm)|(hectom(è|e)tre(s?))</pattern>
    <featureSet>
      <unite>
        <type>distance</type>
        <original>hm</original>
        <equivalence>
          <unite>m</unite>
          <valeur>100</valeur>
        </equivalence>
      </unite>
    </featureSet>
  </rule>
  <rule caseInsensitive="true">
    <pattern>(dam)|(decam(è|e)tre(s?))</pattern>
    <featureSet>
      <unite>
        <type>distance</type>
        <original>dam</original>
        <equivalence>
          <unite>m</unite>
          <valeur>10</valeur>
        </equivalence>
      </unite>
    </featureSet>
  </rule>
  <rule caseInsensitive="true">
    <pattern>lieue(s?)</pattern>
    <featureSet>
      <unite>
        <type>distance</type>
        <original>lieue</original>
        <equivalence>
          <unite>m</unite>
          <valeur>4000</valeur>
        </equivalence>
      </unite>
    </featureSet>
  </rule>
  <rule caseInsensitive="true">
    <pattern>yard(s?)</pattern>
    <featureSet>
      <unite>
        <type>distance</type>
        <original>yard</original>
        <equivalence>
          <unite>m</unite>
          <valeur>0.91</valeur>
        </equivalence>
      </unite>
    </featureSet>
  </rule>
  <rule caseInsensitive="true">
    <pattern>pied(s?)</pattern>
    <featureSet>
      <unite>
```

```
85          <type>distance</type>
86          <original>pied</original>
87          <equivalence>
88            <unite>m</unite>
89            <valeur>0.33</valeur>
90          </equivalence>
91        </unite>
92      </featureSet>
93    </rule>
94    <rule caseInsensitive="true">
95      <pattern>agglomération(s?)</pattern>
96      <featureSet>
97        <intro>
98          <type>geo</type>
99          <stype>agglo</stype>
100       </intro>
101     </featureSet>
102   </rule>
103   <rule caseInsensitive="true">
104     <pattern>RN|nationale|N|RN</pattern>
105     <featureSet>
106       <intro>
107         <type>route</type>
108         <stype>nationale</stype>
109       </intro>
110     </featureSet>
111   </rule>
112   <rule caseInsensitive="true">
113     <pattern>D|départementale</pattern>
114     <featureSet>
115       <intro>
116         <type>route</type>
117         <stype>départementale</stype>
118       </intro>
119     </featureSet>
120   </rule>
121   <rule caseInsensitive="false">
122     <pattern>bois|forêt(s?)|bosquet(s?)</pattern>
123     <featureSet>
124       <intro>
125         <type>geo</type>
126         <stype>bois</stype>
127       </intro>
128     </featureSet>
129   </rule>
130   <rule caseInsensitive="false">
131     <pattern>montagne(s?)|pic(s?)|mont(s?)</pattern>
132     <featureSet>
133       <intro>
134         <type>geo</type>
135         <stype>mont</stype>
136       </intro>
137     </featureSet>
138   </rule>
139   <rule caseInsensitive="false">
140     <pattern>col(s?)</pattern>
141     <featureSet>
142       <intro>
143         <type>geo</type>
144         <stype>col</stype>
145       </intro>
146     </featureSet>
```

```
147   </rule>
148   <rule caseInsensitive="false">
149     <pattern>ville(s?)|village(s?)|bourg(s?)|communne(s?)</pattern>
150     <featureSet>
151       <intro>
152         <type>geo</type>
153         <stype>ville</stype>
154       </intro>
155     </featureSet>
156   </rule>
157   <rule caseInsensitive="false">
158     <pattern>rivière(s?)|(g|G)ave(s?)|fleuve(s?)|ruisseau(x?)|torrent(s?)|
          can(al|aux)</pattern>
159     <featureSet>
160       <intro>
161         <type>geo</type>
162         <stype>riviere</stype>
163       </intro>
164     </featureSet>
165   </rule>
166   <rule caseInsensitive="false">
167     <pattern>region(s?)</pattern>
168     <featureSet>
169       <intro>
170         <type>geo</type>
171         <stype>region</stype>
172       </intro>
173     </featureSet>
174   </rule>
175   <rule caseInsensitive="false">
176     <pattern>plateau(x?)|plan(s?)|terrasse|haute-plaine|belvédère</pattern>
177     <featureSet>
178       <intro>
179         <type>geo</type>
180         <stype>plateau</stype>
181       </intro>
182     </featureSet>
183   </rule>
184   <rule caseInsensitive="false">
185     <pattern>vallée(s?)|val(s?)|vallon(s?)|défilé(s?)</pattern>
186     <featureSet>
187       <intro>
188         <type>geo</type>
189         <stype>vallee</stype>
190       </intro>
191     </featureSet>
192   </rule>
193   <rule caseInsensitive="false">
194     <pattern>crête(s?)|corniche(s?)|cîme(s?)|faîte(s?)</pattern>
195     <featureSet>
196       <intro>
197         <type>geo</type>
198         <stype>crete</stype>
199       </intro>
200     </featureSet>
201   </rule>
202   <rule caseInsensitive="false">
203     <pattern>auberge|hôtel|hôte|refuge</pattern>
204     <featureSet>
205       <intro>
206         <type>geo</type>
207         <stype>hotel</stype>
```

```
208       </intro>
209     </featureSet>
210   </rule>
211   <rule caseInsensitive="false">
212     <pattern>source|fontaine</pattern>
213     <featureSet>
214       <intro>
215         <type>geo</type>
216         <stype>hydro</stype>
217       </intro>
218     </featureSet>
219   </rule>
220   <rule caseInsensitive="false">
221     <pattern>lac|marais|mare|étang</pattern>
222     <featureSet>
223       <intro>
224         <type>geo</type>
225         <stype>lac</stype>
226       </intro>
227     </featureSet>
228   </rule>
229   <rule caseInsensitive="true">
230     <pattern>proche|autour|proximité|environs|auprès|bord|fond|côté|
          périphérie|lisière|près|tout\sprès</pattern>
231     <featureSet>
232       <intro>
233         <type>adjacence</type>
234         <stype>proche</stype>
235       </intro>
236     </featureSet>
237   </rule>
238   <rule caseInsensitive="true">
239     <pattern>sud</pattern>
240     <featureSet>
241       <intro>
242         <type>orientation</type>
243         <stype>sud</stype>
244       </intro>
245     </featureSet>
246   </rule>
247   <rule caseInsensitive="true">
248     <pattern>nord</pattern>
249     <featureSet>
250       <intro>
251         <type>orientation</type>
252         <stype>nord</stype>
253       </intro>
254     </featureSet>
255   </rule>
256   <rule caseInsensitive="true">
257     <pattern>est</pattern>
258     <featureSet>
259       <intro>
260         <type>orientation</type>
261         <stype>est</stype>
262       </intro>
263     </featureSet>
264   </rule>
265   <rule caseInsensitive="true">
266     <pattern>ouest</pattern>
267     <featureSet>
268       <intro>
```

```
269         <type>orientation</type>
270         <stype>ouest</stype>
271       </intro>
272     </featureSet>
273   </rule>
274   <rule caseInsensitive="true">
275     <pattern>nord-ouest</pattern>
276     <featureSet>
277       <intro>
278         <type>orientation</type>
279         <stype>nord-ouest</stype>
280       </intro>
281     </featureSet>
282   </rule>
283   <rule caseInsensitive="true">
284     <pattern>nord-est</pattern>
285     <featureSet>
286       <intro>
287         <type>orientation</type>
288         <stype>nord-est</stype>
289       </intro>
290     </featureSet>
291   </rule>
292   <rule caseInsensitive="true">
293     <pattern>sud-est</pattern>
294     <featureSet>
295       <intro>
296         <type>orientation</type>
297         <stype>sud-est</stype>
298       </intro>
299     </featureSet>
300   </rule>
301   <rule caseInsensitive="true">
302     <pattern>sud-ouest</pattern>
303     <featureSet>
304       <intro>
305         <type>orientation</type>
306         <stype>sud-ouest</stype>
307       </intro>
308     </featureSet>
309   </rule>
310   <rule caseInsensitive="true">
311     <pattern>centre|milieu|dans|intérieur</pattern>
312     <featureSet>
313       <intro>
314         <type>inclusion</type>
315         <stype>centre</stype>
316       </intro>
317     </featureSet>
318   </rule>
319   <rule caseInsensitive="true">
320     <pattern>[A-Z]{1}.+</pattern>
321     <featureSet>
322       <egn>oui</egn>
323     </featureSet>
324   </rule>
325 </tokenMarker>
```

Annexe D

Grammaire DCG utilisée durant le processus sémantique

```
root(X) --> ega(X).

ega(es:X)-->es(X).

%----------
% Une es peut être absolue ou relative
%----------
es(es_r:X) --> es_r(X).
es(es_a:X) --> es_a(X).


%----------
% Relations binaires : définies par une relation suivie d'une ESA ou d'une
    ESR
%----------
es_r(name:label..relation:X..es_r:ES_R) --> relation(X), es_r(ES_R).
es_r(name:label..relation:X..es_a:ES_A) --> relation(X), es_a(ES_A).

%----------
% Les 5 relations définies dans le MP
%----------
relation(orientation:X) --> orientation(X).
relation(distance:X) --> distance_oriente(X),!.
relation(distance:X) --> distance(X).
relation(adjacence:X) --> adjacence(X).
relation(inclusion:X) --> inclusion(X).
relation_naire2(figure_geo:X) --> figure_geo2(X).
relation_naire3(figure_geo:X) --> figure_geo3(X).


% RELATION DE DISTANCE
distance_oriente(D..orientation:O)-->distance(D),prepart,intro_orientation(O
    ).
distance(valeur:V..unite:U..orientation:nc)-->valeur_numerique(V),
    unite_distance(U).
valeur_numerique(V) --> V@tag:num..stag:null.
valeur_numerique(V) --> V@tag:adj..stag:num.
unite_distance(D) --> ls_token(_,unite:type:distance..unite:original:O..
    unite:equivalence:E,token),{D:=(original:O..equivalence:E)}.
```

```
36
37
38 % RELATION D ORIENTATION
39 intro_orientation(P)-->ls_token(_,intro:type:orientation..intro:stype:S,
       token),{P:=(pt_card:S)}.
40 orientation(O) --> prepOUprepart, intro_orientation(O).
41 orientation(O) --> prepOUprepart, intro_orientation(O), prepOUprepart.
42
43 % RELATION D'ADJACENCE
44 adjacence(A) --> intro_adjacence(A).
45 intro_adjacence(A) --> ls_token(_,intro:type:adjacence..intro:stype:S,token)
       ,{A:=(adj_type:S)}.
46
47
48 % RELATION D'INCLUSION
49 intro_inclusion(I) --> ls_token(_,intro:type:inclusion..intro:stype:S,token)
       ,{I:=(inclusion_type:S)}.
50 inclusion(I) --> intro_inclusion(I).
51
52
53 % FIGURE GEOMETIQUE A 3 EXTREMITES
54 figure_geo3(extremite:3)-->ls_token('triangle').
55 figure_geo3(extremite:3)-->ls_token('entre').
56 es_a3s(es1:X..es2:Y..eg3:Z) --> es_a_nointro(X), separateur, es_a_nointro(Y
       ), separateur, es_a_nointro(Z).
57 es_a3s(es1:X..es2:Y..eg3:Z) --> es_a_nointro(X), es_a_nointro(Y),
       es_a_nointro(Z).
58
59
60 % FIGURE GEOGMETIQUE A 2 EXTREMITES
61 figure_geo2(extremite:2)-->ls_token('entre').
62 figure_geo2(extremite:2)-->ls_token('axe').
63 figure_geo2(extremite:2)-->ls_token('ligne').
64 figure_geo2(extremite:2)-->ls_token('dorsale').
65 figure_geo2(extremite:2)-->ls_token('route').
66 es_a2s(es1:X..es2:Y) --> es_a_nointro(X), es_a_nointro(Y).
67 es_a2s(es1:X..es2:Y) --> es_a_nointro(X), separateur, es_a_nointro(Y).
68
69
70 % Tokens de séparation
71 separateur --> ls_token('et').
72 separateur --> ls_token('-').
73 separateur --> ls_token(',').
74 separateur --> ls_token('/').
75
76
77 prepart --> prep, art.
78 prepart --> ls_token('du').
79 prepart --> ls_token('au').
80 prepart --> ls_token('Au').
81 prepOUprepart --> prep.
82 prepOUprepart --> prepart.
83 prepOUprepartOUart --> prepart.
84 prepOUprepartOUart --> prep.
85 prepOUprepartOUart --> art.
86
87 %----------
88 % Définition des introducteurs spatiaux qui donnent un type à l'entité (ex
       : la VILLE de Pau, le PIC d'Anie)
89 %----------
90 intro_geo(type:T) --> ls_token(_,intro:type:geo..intro:stype:S,token),{T:=(S
       )}.
```

```
91
92
93  % TYPES CONNUS
94  es_a(label:N..T) --> intro_geo(T), N@egn:oui.
95  es_a(label:N..T) --> prepOUprepartOUart, intro_geo(T), N@egn:oui.
96  es_a(label:N..T) --> intro_geo(T), prepOUprepart, N@egn:oui.
97  es_a(label:N..T) --> prepOUprepartOUart, intro_geo(T), prepOUprepartOUart,
        N@egn:oui.
98
99
100 % TYPE INCONNU
101 es_a(label:N..type:inconnu) --> prepOUprepartOUart, N@egn:oui.
102
103
104 % TYPE INCONNU SANS INTRODUCTEUR (pour la relation figure_geo)
105 es_a_nointro(type:inconnu..label:N) --> N@egn:oui.
```

Annexe E

Signatures des services web composant le prototype PIV

E.1 Services web de traitement sémantique

Le tableau E.2 fait la liste des services web implémentés dans le cadre de la plate-forme PIV. Nous décrivons leurs fonctions et les travaux dans lesquels ils sont intervenus. Ils s'appuient pour la partie de traitement linguistique, sur l'outil *Linguastream* [Bil03], pour le stockage de flux XML, sur l'outil de gestion de base de données XML *eXist* et pour la gestion de données géo-référencées, sur le SIG *PostGIS*.

E.1.1 Module de segmentation

Service TxtToXML

- Service : ce service prend en entrée une chaîne de caractères représentant un texte et fournit en sortie une autre chaîne qui représente la structure du document. Il détecte les paragraphes et les numérote.
- Profil : Une seule méthode est proposée par ce service : *String TxtToXml(String)*. Elle prend en paramètre une chaîne de caractère qui représente le texte à traiter et renvoie en sortie la chaîne modifiée.
- Processus : Le traitement du texte brut consiste à le transformer en flux XML *via* l'ajout de marques au début et à la fin du document, ainsi qu'à chaque espace vertical suffisamment grand pour découper le texte en paragraphes.

Service Tokenizer

- Service : ce service prend en charge la segmentation du texte. Il prend en entrée le texte sous forme de flux XML découpé en paragraphes et produit un autre flux dans lequel chaque mot est identifié et isolé dans une balise.
- Profil : La seule méthode proposée par ce service est *String Tokenizer(String)*.

Nom des services web	Commentaire	Documentation
Module de *Tokenisation* (services TxtToXML, Tokenizer)	Ce module consiste à découper le texte initial en unités lexicales minimales appelées *tokens*.	[LGL06, Ess07, Man07]
Module d'analyse morphosyntaxique (service MorphoSyntacticAnalysis)	Ce module consiste à reconnaître les mots et leur forme infléchie. Il permet aussi d'établir le genre grammatical des objets et de retrouver le lien entre ces unités.	[LGL06, Ess07, Man07]
Module d'analyse sémantique (services SetExpression, SetGrammar)	Ce module propose deux services web pour donner un sens spatial à des termes candidats.	[LGL06, Ess07, Man07]
Module d'indexation (services XSLTTransformation, Georeferencing)	Ce module réunit les données de deux flux XML fournis en entrée pour en donner un seul en sortie contenant toutes les balises. Le service de transformation utilise ensuite des algorithmes d'interprétation de la sémantique et de calcul de représentation pour les entités absolues et relatives.	[LGL06, Ess07, Man07]
Module de stockage des index (service StoreDB)	Ce module permet l'ajout des données fournies comme paramètres dans la base de données cible (XML, SIG) afin de former l'index.	[LGL06, Ess07, Man07]
	Services web de recherche d'information spatiale.	
Module de recherche d'intersection (getBooleanMatching, getRankedMatching)	Ce module permet de déterminer des intersections entre unités spatiales (requête/index) et de calculer des degrés de pertinence relatifs notamment à la surface de recouvrement requête/index. Le résultat est une liste d'unités documentaires (paragraphes) classées et pondérées.	[SBLG07a, Ess07, Man07]

Tab. E.2 – Services web d'extraction d'information spatiale.

E.1.2 Module d'analyse morpho-syntaxique

Service MorphoSyntacticAnalysis

- Service : ce service se charge de l'analyse morpho-syntaxique du texte.
- Profil : la méthode *MorphoSyntacticAnalysis()* prend en entrée le flux XML découpé en mots et le complète en y ajoutant des informations sur la nature des mots, leurs types et leurs lemmes associés.
- Processus : l'outil *Tree-Tagger* est utilisé pour réaliser cette opération. Il est appelé par la méthode *MorphoSyntacticAnalysis()*.

E.1.3 Module d'analyse sémantique

Service SetExpression

- Service : ce service fonctionne à partir d'expressions régulières et de détection de motifs. Il prend en entrée le flux XML et un fichier de ressources contenant les expressions régulières qui définissent les patrons à détecter. Le flux de sortie est augmenté de balises pour les textes correspondant à ces patrons.
- Profil : la méthode proposé par ce service est *String SetExpression(String, String)*.

Service SetGrammar

- Service : ce service a pour but de détecter les relations qui existent entre les entités spatiales candidates. Il s'appuie sur le concept de grammaires DCG (implémenté à l'aide du langage *Prolog*). Ces grammaires permettent de s'appuyer sur les mécanismes d'inférence et d'unification de Prolog à l'aide de règles simples.
- Profil : ce service offre la seule méthode *String SetGrammar(String, String)* qui prend en entré le flux XML et un ensemble de règles.

E.2 Services web d'indexation

Service XSLTTransformation

- Service : lors du traitement sémantique, chaque *token* est identifié par un numéro de paragraphe et un identifiant (dans le fichier *doc*). Ce service vient alors compléter le flux XML des identifiants des entités spatiales candidates issues de la méthode *SetGrammar()* (*sem*). Le service permet d'obtenir un flux XML valide par rapport au schéma défini (grâce au *fichier.xslt*).
- Profil : la méthode utilisée est *String XSLTTransformation(String doc, String sem, String fichier.xsl.*

Service Georeferencing

- Service : ce service se charge de trouver des représentations géo-référencées pour les entités spatiales candidates. Il prend en entrée un flux XML correspondant à une entité spatiale candidate (c'est-à-dire à une instance du modèle pivot) et retourne une représentation géo-référencée.

– Profil et Processus : Deux méthodes sont nécessaires. La première, *String Calcul_Geo(String)*, implémente l'algorithme récursif défini dans nos préconisations. La deuxième, *double[] Appel_SIG(String)*, prend en entrée le libellé d'une entité nommée et se connecte à des ressources géographiques locales afin de retourner une représentation géo-référencée.

E.3 Service web dédié au stockage dans une base de données

Service StoreDB

– Service : ce service permet le stockage des ressources nécessaires à l'indexation. Il permet d'une part de stocker les flux XML, résultats du processus du traitement sémantique, dans une base *eXist* et, d'autre part, de stocker les index spatiaux calculés grâce au service de géo-référencement précédent.
– Profil : la première méthode, *String StoreXML(String xml, String path_db, String user, String password)* se charge du stockage dans *eXist*. La deuxième méthode *String StoreGIS(String xml, String path_db, String user, String password)* se charge de stocker l'index spatial construit dans une base adéquate, adaptée à la manipulation d'objets géo-référencées. Cette dernière méthode fait appel à une autre, *String XMLtoSQL(String xml)*, qui se charge de transformer le flux XML en un fichier SQL compatible avec le SIG *PostGIS*.

E.4 Services web d'appariement

Service getBooleanMatching

– Service : ce service se charge de retourner les identifiants des documents contenant des entités spatiales dont les géo-références s'intersectent avec la zone géo-référencée de la requête. Ce calcul simple d'intersection ne différencie pas les grandes ou les petites surfaces de recouvrement entre les zones de la requête et celles des documents et tous les résultats ont la même pertinence.
– Profil : La méthode *String[] getBooleanMatching(double[])* prend en entrée une représentation géo-référencée et retourne la liste des identifiants de documents pertinents.

Service getRankedMatching

– Service : ce service opère le même processus que le précédent. Seul le calcul d'intersection change. Il implémente la proposition définie dans les préconisations sur l'appariement spatial tenant compte des surfaces de recouvrement entre la requête et les documents. Ce service permet donc d'ordonner les résultats selon une pertinence spatiale.
– Profil : La méthode *String[] getRankedMatching(double[])* prend en entrée une représentation géo-référencée et retourne la liste ordonnée des identifiants de docu-

ments pertinents.

E.5 Services web annexes

Le tableau E.3 liste les travaux annexes concernant la phase de visualisation des résultats.

Interfaces interactives de visualisation d'information spatiale.

Module de visualisation de paragraphes résultants d'une recherche (type Google)	Interface d'interrogation en texte libre et de visualisation par liste.	[LGL06]
Module de visualisation de paragraphes résultant d'une recherche (type 2D et 3D)	Interfaces 2D (ViaMichelin, GoogleMaps) et 3D (GoogleEarth).	[EMC06]
Module de visualisation 3D de ressources IGN	Interfaces 3D de navigation avec ajout de couches de ressources.	[Vu06]

TAB. E.3 – Liste de modules venant de travaux annexes.

Bibliographie

[ABDR06] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR '06 : Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10, New York, NY, USA, 2006. ACM Press.

[AFG03] M Abolhassani, N Fuhr, and N Gövert. Information extraction and automatic markup for xml documents. *Intelligence Search on XML data - LNCS Springer*, pages 159–178, 2003.

[All91] James F. Allen. Planning as temporal reasoning. In *KR, Principles of Knowledge Representation and Reasoning*, pages 3–14, 1991.

[APT96] M.A. Aufaure-Portier and C. Trepied. A survey of query languages for geographic information systems. In *IDS-3, Workshop on Interface to Database, Springer Verlag's Electronic Workshops in Computer Series, Edinburgh, UK*, pages 1–14, 1996.

[Asi04] T. Asic. *La représentation cognitive du temps et de l'espace ; analyse pragmatique de données linguistique en français et dans d'autres langues.* PhD thesis, thèse de doctorat en Sciences Cognitives, mention Sciences du langage, Université Lumière-Lyon 2, 2004.

[Auf01] M.A. Aufaure. What approach for searching spatial information ? In *Journal of Visual Languages and Computing*, volume 12, pages 351–373, 2001.

[AVB97] M Aurnague, L Vieu, and A Borillo. La représentation formelle des concepts spatiaux dans la langue. In M. Denis, editor, *Langage et cognition spatiale*, Collection 'Sciences cognitives', chapter 4, pages 69–102. Masson, 1997.

[AW06] R Angelova and G Weikum. Graph-based text classification : Learn from you neighbors. In *29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 485–492, 2006.

[Baz05] Mustapha Baziz. *Indexation conceptuelle guidée par ontologie pour la recherche d'information.* PhD thesis, Université Paul Sabatier - Institut de Recherche en Informatique de Toulouse, 2005.

[BBAG05] Mustapha Baziz, Mohand Boughanem, and Nathalie Aussenac-Gilles. Evaluating a conceptual indexing method by utilizing wordnet. In *CLEF*, pages 238–246, 2005.

[BBB03] J.C. Bottraud, G. Bisson, and M.F. Bruandet. Apprentissage de profils pour un agent de recherche d'information. In *Actes de la Conférence Apprentissage (CAp 2003)*, pages 31–46, 2003.

[BBPP06] Mustapha Baziz, Mohand Boughanem, Henri Prade, and Gabriella Pasi. A fuzzy logic approach to information retrieval using an ontology-based representation of documents. In E. Sanchez, editor, *Fuzzy Logic and the semantic web*, chapter 18, pages 363–377. Elsevier, http ://www.elsevier.com/, 2006. BougP et al. 002.

[BC00] K.P. Bennett and C. Campbell. Support vector machines : Hype or hallelujah ? *SIGKDD Explorations*, 2(2) :pages 1–13, 2000.

[BCI97] B. Bennett, A. G. Cohn, and A. Isli. A logical approach to incorporating qualitative spatial reasoning into gis (extended abstract). In *Proceedings of the International COnference on Spatial Information Theory (COSIT)*, pages 503–504, 1997.

[BDEH07] F Bilhaut, F. Dumoncel, P. Enjalbert, and N. Hernandez. Indexation sémantique et recherche d'information interactive. In *CORIA, Saint-Etienne*, pages 65–76, 2007.

[BE05] F. Bilhaut and P. Enjalbert. *Sémantique et traitement automatique du langage naturel*, chapter 10, Recherche d'information géographique, pages 371–406. Hermes, Lavoisier, 2005.

[BEJ+97] C Bessière, J Euzenat, R Jeansoulin, G Ligozat, and S Schwer. Raisonnement spatial et temporel. *Actes 6e journées nationales du PRC-GDR « intelligence artificielle », Grenoble*, pages 77–88, 1997.

[Ben96] B Bennett. The application of qualitative spatial reasoning to GIS. In R.J. Abrahart, editor, *Proc First Int. Conf. on GeoComputation*, volume I, pages 44–47, Leeds, 1996.

[Ben01] B. Bennett. A categorical axiomatisation of region-based geometry. *Fundamenta Informaticae*, 46 :pages 145–158, 2001.

[BHDB+03] F Bilhaut, M. Ho-Dac, A Borillo, T. Charnois, P. Enjalbert, A. Le Draoulec, Y. Mathet, H. Miguet, M.P. Péry-Woodley, and L. Sarda. Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *10ème Conférence Traitement Automatique du Langage Naturel (TALN)*, pages 1–6, 2003.

[Bil03] F Bilhaut. The linguastream platform. *Proceedings of the 19th Spanish Society for Natural Language Processing Conference (SEPLN)*, pages 339–340, 2003.

[Bil06] F Bilhaut. *Analyse automatique de structures thématiques discursives - Application à la recherche d'information.* PhD thesis, Université de Caen, 2006.

[Bit96] T Bittner. A qualitative model of geographic space. *International Symposium on Spatial Data Handling, Delft*, 2(10.19-10.32), 1996.

[BKSS90] N. Beckmann, H.P. Kriegel, R. Schneider, and B Seeger. The R*-tree : an efficient and robust access method for points and rectangles. In *ACM SIGMOD, USA*, pages 322–331, 1990.

[Bor98] A. Borillo. *L'espace et son expression en français*. L'essentiel. Ophrys, 1998.

[Bri98] A. Brimicombe. A fuzzy set approach to using linguistic hedges in geographical information systems. In *Cybergeo : Revue européenne de géographie*, page article 66, 1998. http://www.cybergeo.eu/index560.html.

[BS01] T Bittner and J.G Stell. Approximate qualitative spatial reasoning. *Spatial Cognition and Computation*, 2(4) :pages 435–466, 2001.

[Bur98] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2) :pages 121–167, 1998.

[BVT05] W.L. Buntime, K Valtonen, and M.P. Taylor. The alvis document model for a semantic search engine. *2nd Annual European Semantic Web Conference*, pages 1–2, 2005.

[BY99] Ricardo A. Baeza-Yates. *Modern Information Retrieval*. Addison-Wesley, ACM Press, 1999.

[CBGG97] Anthony G. Cohn, Brandon Bennett, John Gooday, and Nicholas M. Gotts. Representing and reasoning with qualitative spatial relations. In Oliviero Stock, editor, *Spatial and Temporal Reasoning*, pages 97–134. Kluwer Academic Publishers, Dordrecht, 1997.

[CH01] A G Cohn and S M Hazarika. Qualitative spatial representation and reasoning : An overview. *Fundamenta Informaticae*, 46(1-2) :pages 1–29, 2001.

[Cha05] Khalil Challita. *Problèmes de satisfaction de contraintes spatiales : de l'algèbre des régions à la géométrie affine*. PhD thesis, Université de Toulouse III, 2005.

[CKSL04] K-M Chung, W-C Kao, T Sun, and C.-J Lin. Decomposition methods for linear support vector machines. *Acoustics, Speech and Signal Processing (ICASSP)*, 16(8) :pages 1689–1704, 2004.

[Cle63] C.W Cleverdon. Comparative efficiency of indexing systems. *Cranfield, Nature, Volume 197, Issue 4863*, pages 129–130, 1963.

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE : A framework and graphical development environment for robust NLP tools and applications. In *40th Anniversary meeting of the Association for Computational Linguistics (ACL'02)*, 2002. http://gate.ac.uk/sale/acl02/acl-main.pdf.

[CMDG04] Jean Casenave, Christophe Marquesuzaà, Pantchika Dagorret, and Mauro Gaio. La revitalisation numérique du patrimoine littéraire territorialisé. In *Le numérique : impact sur le cycle de vie du document.* ENSSIB, Montréal, 2004. http://www.enssib.fr/babel/document.php?id=741.

[Coh96] Anthony G. Cohn. Calculi for qualitative spatial reasoning. In *AISMC-3 : Proceedings of the International Conference AISMC-3 on Artificial Intelligence and Symbolic Mathematical Computation*, pages 124–143, London, UK, 1996. Springer-Verlag.

[Coh97] Anthony G. Cohn. Qualitative spatial representation and reasoning techniques. In *KI '97 : Proceedings of the 21st Annual German Conference on Artificial Intelligence*, pages 1–30, London, UK, 1997. Springer-Verlag.

[Con06] Anne Condamines. Modes de construction du sens en corpus spécialisé. *Cahiers de Grammaire*, 30 :pp. 75–88, 2006.

[CSJ04] P Clough, M Sanderson, and H Joho. Extraction of semantic annotations from textual web pages. Technical report, SPIRIT is funded by EU IST Programme, 2004.

[DB97] Michel Denis and Xavier Briffault. *Les aides verbales à la navigation*, chapter 6. Langage et Cognition Spatiale, Sciences cognitives, Masson, 1997.

[DD98] M-P. Daniel and M. Denis. *Spatial Descriptions as Navigational Aids : A Cognitive Analysis of Route*, pages 45–52. Springer Berlin / Heidelberg, 1998.

[DDFH90] S. Deerwester, S.T Dumais, T.K. Furnas, and R.A. Harman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407, 1990.

[DEG+03] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, and J.Y Zien. Semtag and seeker : Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the WWW Conference*, pages 178–186, 2003.

[Den97] Michel Denis. *Langage et cognition spatiale.* Masson, 1997.

[DGP03] Robert Dale, Sabine Geldof, and Jean-Philippe Prost. Coral : using natural language generation for navigational assistance. In *CRIPTS '03 : Proceedings of the twenty-sixth Australian computer science conference on Conference in research and practice in information technology*, pages 35–44, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc.

[DLC04] C. De Loupy and E. Crestan. *Les systèmes de recherche d'informations*, chapter 6, SRI et traitement du langage naturel. Hermes, Lavoisier, 2004.

[DP97] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. In *Machine Learning*, volume 29, pages 103–137, 1997.

[Dum06] Franck Dumoncel. *Géographie et graphe - Une interaction pour exprimer des requêtes spatiales guidée par des adjacences conceptuelles*. PhD thesis, Université de Caen, 2006.

[EF91] M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2) :161–174, 1991.

[Ege91] M Egenhofer. Reasoning about binary topological relations. *Second Symposium on Large Spatial Databases, Zurich, Switzerland, Lecture Notes in Computer Science*, 525 :143–160, 1991.

[Ege95] M Egenhofer. Visual map algebra : A direct-manipulation user interface for GIS. In *IFIP, Workshop Conference on Visual Database Systems*, pages 235–253, 1995.

[EL04] G. Edwards and G Ligozat. *A formal model for structuring local perceptions of environmental space*, volume 5, pages 3–9. Springer Berlin / Heidelberg, 2004.

[EL05] C Marquesuzaà P Etcheverry and J Lesbegueries. Exploiting geospatial markers to explore and resocialize localized documents. In *Geos, Lecture Notes in Computer Science, Springer, Mexico City - Mexico*, pages 153–165, 2005.

[EM95] M Egenhofer and D Mark. Naive geography. In Springer-Verlag, editor, *COSIT - Lecture Notes in Computer Science*, volume 988, pages 1–15, 1995.

[EMC06] P. Etcheverry, C. Marquesuzaà, and S. Corbineau. Designing suited interactions for a document management system handling localized documents. In USA ISBN : 1-59593-523-1, Myrtle Beach, editor, *24th ACM International Conference on Design of Communication, SIGDOC*, pages 188–195, 2006.

[EML05] P Etcheverry, C Marquesuzaà, and J Lesbegueries. Revitalisation de documents territorialisés : Principes, outils et premiers résultats. Workshop Met-SI INFORSID - Grenoble, 2005.

[Ess07] W. Essafi. Fonctions d'interprétation d'expressions spatiales. Master's thesis, Université de Pau et des Pays de l'Adour, 2007.

[FG04] Jihad Farhat and Luc Girard. L'avenir des services de référence des bibliothèques universitaires. *Argus*, 33(2) :23–28, automne 2004.

[Fil01] L. Filliettaz. Les types de discours. In *Cercle, revue électronique* `http://www.ucm.es/info/circulo/cercle.htm`, 2001.

[FJA05a] F. FU, C.B. Jones, and A.I Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In *IASTED International Conference on Databases and Applications*, 2005.

[FJA05b] G. Fu, C.B. Jones, and A.I Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *On the Move to Meaningful Internet systems, ODBASE, Cyprus : Lectures Notes in Computer Science 3633*, pages 218–235, 2005.

[FL99] L Fraczak and G Lapalme. Utilisation de stratégies cognitives dans la génération automatique de descriptions d'itinéraires. In *Actes de la Conférence TALN-1999*, pages 145–154, 1999.

[Flu94] C Fluhr. Spirit : un système d'exploration de données textuelles. In *Le Traitement Informatique des Corpus Textuels*, 1994.

[Fra99] L Fraczak. A dynamic model of route description. *Rapport interne*, 1999.

[Gai01] Mauro Gaio. Traitements de l'information géographique : Représentations et structures. In *Mémoire d'HDR, Université de Caen*, 2001.

[Gai02] R. Gaizaukas. An information extraction perspective on text mining : Tasks, technologies and prototype applications. In *Euromap Text Mining Seminar, Sheffield*, 2002.

[Gal01] A Galton. Space, time, and the representation of geographical reality. *Topoi*, 20(2) :173–187, December 2001.

[Gry95] A. Gryl. *Analyse et modélisation des processus discursifs mis en œuvre dans la description d'itinéraires*. PhD thesis, Université de Paris XI, 1995.

[GSE+07] M. Gaio, C. Sallaberry, P. Etcheverry, C. Marquesuz'aa, and J. Lesbegueries. *Journal of Visual Languages And Computing*, chapter A Global Process to Access Documents' Contents from a Geographical Point of View. Elsevier, 2007.

[Har97] P. Harpring. *Proper words in proper places : The Thesaurus of Geographic Names*, pages 5–12. MDA Information, 1997.

[Has06] M. Hasan. Svm : Machines à vecteurs de support ou séparateurs à vastes marges. Technical report, Versailles St Quentin, France, 2006.

[HCTH99] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in Web search evaluation. *Computer Networks (Amsterdam, Netherlands : 1999)*, 31(11–16) :1321–1330, 1999.

[HFZ99] L.L. Hill, J. Frew, and Q Zheng. Geographic names. the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5, 1999. http://www.dlib.org/dlib/january99/hill/01hill.html.

[HGJ04] L.L. Hill, M.F. Goodchild, and G Janée. Research directions in georeferenced ir based on the alexandria digital library project. In *Workshop GIR - 27th annual international ACM SIGIR Special Interest Group on Information Retrieval*, 2004.

[Hil04] L.L. Hill. Georeferencing in digital libraries. *D-Lib Magazine*, 10, 2004. http://www.dlib.org/dlib/may04/hill/05hill.html.

[HL02a] C.-W. Hsu and C.-J Lin. A comparison on methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13 :415–425, 2002.

[HL02b] C.-W. Hsu and C.-J Lin. A simple decomposition method for support vector machines. *Machine Learning*, 46 :291–314, 2002.

[HMW93] G Herzog, W Maaß, and P Wazinski. Vitra guide : Utilisation du langage naturel et de représentation graphiques pour la description d'itinéraires. *Images et Langages : Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique, Paris*, pages 243–251, 1993.

[Iha04] M. Ihadjadene. *Méthodes avancées pour les systèmes de recherche d'informations*. Hermes, Lavoisier, 2004.

[JAF03] C.B. Jones, A.I Abdelmoty, and G. Fu. Maintaining ontologies for geographical information retrieval on the web. In *Proceedings of OTM Confederated International Conferences CoopIS, DOA, and OOBASE*, 2003.

[JAF+04] C.B. Jones, A.I Abdelmoty, D. Finch, G. Fu, and S Vaid. The SPIRIT spatial search engine : Architecture, ontologies and spatial indexing. In *Third International Conference on Geographic Information Science GIScience, Lecture Notes in Computer Science 3234*, pages 125–139, 2004.

[JL06] N Jindal and B Liu. Identifying comparative sentences in text documents. In *29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 244–251, 2006.

[Joa98] T. Joachims. Text categorization with support vector machines : learning with many relevant features. *Lecture Notes in Computer Scienc, Proceedings of ECML-98, 10th European Conference on Machine Learning*, series(1398) :137–142, 1998.

[Joa99] T. Joachims. Transductive inference for text classification using support vector machines. *Proceedings of ICML-99, 16th International Conference on Machine Learnings*, pages 200–209, 1999.

[Joa00] T. Joachims. Estimating the generalization performance of an svm efficiently. *Proceedings of ICML-00, 17th International Conference on Machine Learnings*, pages 431–438, 2000.

[JPR+02] C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies : An overview of the SPIRIT project. In *25th ACM SIGIR*, 2002.

[KDWH05] A. Klippel, J. Davies, S. Winter, and S. Hansen. A high-level cognitive framework for route directions. In *Proceedings of the SSC 2005 Spatial Intelligence, Innovation and Praxis : The National Bienneal Conference of the Spatial Science Institute*, 2005.

[Kle82] Wolfgang Klein. Local deixis in route directions. In *Speech, place, & action. Studies in deixis and related topics. Chichester, New York, Brisbane, Toronto, Singapore : Wiley & Sons*, 1982.

[KRR97] Markus Knauff, Reinhold Rauh, and Jochen Renz. A cognitive assessment of topological spatial relations : Results from an empirical investigation. In *Spatial Information Theory*, pages 193–206, 1997.

[KS97] Norio Katayama and Shin'ichi Satoh. The SR-tree : an index structure for high-dimensional nearest neighbor queries. In *SIGMOD*, pages 369–380, 1997.

[Kui77] Benjamin Kuipers. Modeling spatial knowledge. In *IJCAI*, pages 292–298, 1977.

[Kui00] Benjamin Kuipers. The spatial semantic hierarchy. *Artif. Intell.*, 119(1-2) :191–233, 2000.

[LE00] G Ligozat and G. Edwards. *Implicit spatial reference systems using proximity and alignement knowledge*, volume 2, pages 373–391. Spatial Cognition and Computation : An Interdisciplinary Journal, 2000.

[Le04] Yanfen Le. A feature-based temporal representation and its implementation with object-relational schema for base geographic data in object-based form. *Student poster session, UCGIS Assembly*, 2004. http://www.ucgis.org/ucgisfall2004/studentpapers/files/le.pdf.

[Ler95] Pierre Lerat. *Les langues spécialisées*. PUF, 1995.

[Lev03] S.C. Levinson. *Space in Language and Cognition. Explorations in Linguistic Diversity*. Cambridge University Press, 2003.

[LF04] Larson and Frontiera. Ranking and representation for geographic information retrieval. *Workshop on Geographic Information Retrieval - SIGIR*, 2004. http://www.geo.unizh.ch/~rsp/gir/abstracts/larson.pdf.

[LGL06] Julien Lesbegueries, M Gaio, and Pierre Loustau. Geographical information access for non-structured data. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France*, pages 83–89, 2006.

[LGMR01] Paul A. Longley, Michael F. Goodchild, David J. Maguire, and David W. Rhind. *Geographic Information Systems and Science*. Wiley, 2001.

[LGN07] P Loustau, Mauro Gaio, and T. Nodenot. Des déplacements à l'itinéraire, du syntagme au discours. In *SAGEO*, pages 1–15, 2007.

[LH06] J Leveling and S Hartrumpf. On metonymy recognition for geographic ir. In *Workshop GIR - 29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 1–5, 2006.

[Lig92] G Ligozat. Strategies for route description : An interdisciplinary approach. In *Spatial Concepts : Connecting Cognitive Theories with Formal Representations, Workshop, ECAI*, 1992.

[Lip87] R.P. Lippman. An introduction to computing with neural nets. In *IEEE ASSP Magazine*, pages 4–22, 1987.

[LL06] J Lesbegueries and P Loustau. Structuration d'information spatiale qualitative pour la recherche d'information. *Semaine de la Connaissance - RTE*, pages 1–6, 2006.

[LL07] P Loustau and J Lesbegueries. Détection d'itinéraire dans du texte : deux approches. In *RTE*, pages 1–6, 2007.

[Lou05] Pierre Loustau. Traitements sémantiques de documents dans leur composante spatiale, application au patrimoine pyrénéen. Master's thesis, Université de Pau et des Pays de l'Adour, 2005.

[LPLGS07] A. Le Parc-Lacayrelle, Mauro Gaio, and C. Sallaberry. *Extraction et recherche d'information dans des fonds documentaires patrimoniaux numérisés*. Document numérique, numéro spécial « Entreposage de documents et données semi-structurées », à paraître 2007.

[LSDJ06] M Lew, N Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval : State-of-the-art and challenges. In *ACM Transactions on Multimedia Computing, Communication, and Applications*, volume 2, pages 1–19, 2006.

[Luh58] H Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2) :159–165, 1958.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *L.M. Le Cam & J. Newman [eds.] Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

[Mal03] N. Malandain. *La relation Texte/Image, Essai de modélisation dans un corpus géographique*. PhD thesis, Université de Caen, 2003.

[Man07] J. Mansour. Développement et mise en œuvre d'une infrastructure favorisant la mutualisation des outils de traitement des informations géographiques. Technical report, Projet de fin d'études, INSAT, Tunis, Structure d'accueil : LIUPPA, Université de Pau et des Pays de l'Adour, 2007.

[ME07] C. Marqueszaà and P. Etcheverry. Implementing a visualization system suited to localized documents. In Vietnam Editions SUGER, Collection Informatique ISBN : 2-912590-4-0 Hanoi, editor, *Fifth International Conference on Research, Innovation and Vision for the Future*, pages 13–18, 2007.

[MFK99] S.E Michos, N. Fakotakis, and G. Kokkinakis. Enhancing text retrieval by using advanced stylistic techniques. *Journal of Intelligent and Robotic Systems*, 26(2) :136–156, 1999.

[MGVOM07] A.D. Miron, J. Gensel, M. Villanova-Olivier, and H. Martin. Relations spatiales qualitatives dans les ontologies géographiques avec ONTOAST. In *SAGEO*, 2007.

[MK99] B. Moulin and D. Kettani. Route generation and description using the notions of object's influence area and spatial conceptual map. *Spatial Cognition and Computation*, 1(3) :227–259, 1999.

[MMG99] A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. In *EACL, Bergen, Norway*, pages 1–8, 1999.

[Mor06] Fabienne Moreau. *Revisiter le couplage traitement automatique des langues et recherche d'information*. PhD thesis, Université de Rennes 1, 2006.

[MS05] B Martins and M.J Silva. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, pages 741–744, november 2005.

[MSA05] Bruno Martins, Màrio J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *GIR '05 : Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, New York, NY, USA, 2005. ACM Press.

[Mul98] P Muller. *Eléments d'une théorie du mouvement pour la modélisation du raisonnement spatio-temporel de sens commun*. PhD thesis, Université Paul Sabatier, Toulouse, 1998.

[Naz04] Adeline Nazarenko. *Donner accès au contenu des documents textuels - Acquisition de connaissances et analyse de corpus spécialisés*. PhD thesis, Laboratoire d'Informatique de Paris-Nord - Université Paris-Nord, 2004.

[Nen06] A Nenkova. *Understanding the process of multi-document dummarization : content selection, rewriting and evaluation*. PhD thesis, Columbia University, 2006.

[NMR04] M. Nissim, C. Matheson, and J. Reid. Recognising geographical entities in scottish historical documents. Workshop on Geographic Information Retrieval - SIGIR, 2004. http://www.ltg.ed.ac.uk/seer/papers/gir2004.pdf.

[NVM06] A Nenkova, L Vanderwende, and K McKeown. A compositional context sensitive multidocument summarizer : Exploring the factors that influence summarization. In *29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 573–580, 2006.

[Ope99] OpenGIS Project Document 99-050. Opengis simple features specification for ole/com revision 1.1. Technical report, Open GIS Consortium Inc., 1999.

[OR06] S.E Overell and S Rüger. Identifying and grounding descriptions of places. In *Workshop GIR - 29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 1–3, 2006. http://mmir.doc.ic.ac.uk/www-pub/sigir06-GIR.pdf.

[ORK06] J Otterbacher, D Radev, and O Kareem. News to go : Hierarchical text summarization for mobile devices. In *29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 589–596, 2006.

[PC98] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval, 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 275–281, 1998.

[Peu02] D.J. Peuquet. *Representations of space and time*. Guilford Press, 2002.

[PR03] L Personnaz and I. Rivals. *Réseaux de neurones formels pour la modélisation, la commande et la classification*. CNRS Editions, 2003.

[Pro01] EU Fifth Framework Programme. Spatial-aware information retrieval on the internet. Technical report, SPIRIT Contract Number : IST-2001-35047, 2001.

[PSA07] M. Perry, A. Sheth, and I.B. Arpinar. *Geospatial and Temporal Semantic Analytics*, pages 1–14. Encyclopedia of Geoinformatics, Hassan A. Karimi (Ed), Idea-Group Inc., à paraître en 2007.

[PSZ99] Christine Parent, Stefano Spaccapietra, and Esteban Zimányi. Spatio-temporal conceptual models : data structures + space + time. In *GIS '99 : Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, pages 26–33, New York, NY, USA, 1999. ACM Press.

[RCC92] David A. Randell, Zhan Cui, and Anthony Cohn. A spatial logic based on regions and connection. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *KR'92. Principles of Knowledge Representation and Reasoning : Proceedings of the Third International Conference*, pages 165–176. Morgan Kaufmann, San Mateo, California, 1992.

[Ril96] Ellen Riloff. Using learned extraction patterns for text classification. In *Connectionist, statistical, and symbolic approaches to learning for natural language processing*, pages 275–289. Springer Berlin / Heidelberg, 1996.

[RK04] Kai-Florian Richter and Alexander Klippel. A model for context-specific route directions. In *Spatial Cognition*, pages 58–78, 2004.

[Rob77] Stephen E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4) :294–304, 1977.

[Rob81] J.T. Robinson. The K-D-B-tree : a search structure for large multidimensional dynamic indexes. In *ACM SIGMOD, USA*, pages 10–18, 1981.

[RPR99] M. Riegel, J.C. Pellat, and R. Rioul. *Grammaire méthodique du français*. PUF, collection linguistique nouvelle, 1999.

[RS95] A. Roy and J. Stell. Spatial relations between indeterminate regions. In *Journal of Approximate Reasoning,. Smith, B. (1995). On Drawing Lines on a Map. Spatial Information Theory - A Theoretical Basis for GIS, COSIT*, pages 205–234, 1995.

[Sag99] J.M. Saglio. Fondements des bases de données spatiales, chapitre 7 indexation spatiale. Technical report, ENST - École nationale supérieure des télécommunications, Département informatique et réseaux, 1999.

[Sag06] Benoît Sagot. *Analyse automatique du français : lexiques, formalismes, analyseurs*. PhD thesis, Université de Paris VII Denis Diderot, 2006.

[SAJ04] P.D. Smart, A.I Abdelmoty, and C.B. Jones. An evaluation of geo-ontology representation languages for supporting web retrieval of geographic information. In *GIS Research UK*, pages 175–178, 2004.

[Sal71a] G. Salton. A comparison between manual and automatic indexing methods. In *Journal of the American Documentation*, volume 20(1), pages 61–71, 1971.

[Sal71b] G. Salton. The SMART retrieval system : Experiments in automatic document processing. *Prentice-Hall Inc., Englewood Cliffs, NJ*, pages 456–484, 1971.

[Sal75] G. Salton. *A theory of Indexing*. Society for Industrial and Applied Mathematics, 1975.

[Sap21] E. Sapir. *Language. An Introduction to the Study of Speech*. New York, Harcourt, Brace & Company, 1921.

[SBLG07a] C Sallaberry, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio. Towards a geographical ie and ir system dealing with spatial information scope in textual digital libraries – evaluation case study. *ICEIS*, 2007.

[SBLG07b] C Sallaberry, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio. Une approche d'extraction et de recherche d'information spatiale dans les documents textuels. *CORIA*, pages 1–12, 2007. http://www.irit.fr/ARIA/2007/53.pdf.

[SC99] Fei Song and W. Bruce Croft. A general language model for information retrieval (poster abstract). In *Research and Development in Information Retrieval*, pages 279–280, 1999.

[Sch97] Schatz. Information retrieval in digital libraries : Bringing search to the net. *Science*, 275 :327–334, 1997.

[SGLL07] C Sallaberry, M Gaio, J Lesbegueries, and P Loustau. *A Semantic Approach for Geospatial Information Extraction from Unstructured Documents*, chapter 9. Edited Springer Book in the Advanced Information and Knowledge Processing Series, Springer, 2007.

[SK06] V Sindhwani and S.S Keerthi. Large scale semi-supervised linear svms. In *29th annual international ACM SIGIR Special Interest Group on Information Retrieval*, pages 477–484, 2006.

[SMC+05] Mário J. Silva, B Martins, M Chaves, N Cardoso, and A.P Afonso. Adding geographic scopes to web resources. GIR - unpublished, 2005. http://www.geo.unizh.ch/~rsp/gir/abstracts/silvia.pdf.

[SME06] C. Sallaberry, C. Marquesuzaà, and P Etcheverry. Spatial information management within digital libraries. In India ISBN : 1-4244-0682-X, Bangalore, editor, *1st IEEE International Conference on Digital Information Management, ICDIM*, pages 465–475, 2006.

[SVH04] F. Schilder, Y. Versley, and C Habel. Extracting spatial information : Grounding, classifying and linking spatial expressions. In *Workshop on Geographic Information Retrieval - SIGIR*, pages 1–3, 2004.

[SVV01] Christoph Schlieder, Thomas Vögele, and Ubbo Visser. Qualitative spatial representation for information retrieval by gazetteers. In *COSIT*, pages 336–351, 2001.

[Tar02] J-C. Tarondeau. *Le management des savoirs*. Que sais-je, 2002.

[Tol48] Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, pages 189–208, 1948.

[Tom83] C.D. Tomlin. *Digital cartographic modeling techniques in environmental planning*. PhD thesis, School of Forestry and Environmental Studies, New Haven, CT, Yale University, 1983.

[TT97] B Tversky and H.A. Taylor. Langage et perspective spatiale. In *Langage et Cognition Spatiale*, chapter 2. Sciences cognitives, Masson, 1997.

[UTC04] E.L. Usery, G. Timson, and M. Coletti. Multidimensional representation of geographic features. In *International Journal of Geographic Information Science, in review*, pages 1–8, 2004. http://carto-research.er.usgs.gov/multi-dimension/pdf/usery.996.pdf.

[Van86] Claude Vandeloise. *L'espace en français*. aux Editions du Seuil, Paris, 1986.

[Vap98] V.N. Vapnik. *Statistical Learning Theory*. Wiley - Interscience, 1998.

[vKRAvZ04] M. van Kreveld, I. Reinbacher, A. Arampatzis, and R. van Zwol. Distributed ranking methods for geographic information retrieval. Technical report, Institute of information and computing sciences, Utrecht University, 2004.

[Vu06] L.T. Vu. Navigation interactive dans une carte 3d, application aux pyrénées. Technical report, Mémoire de projet Master Technologie de l'internet, Université de Pau et des Pays de l'Adour, 2006.

[WBH+00] William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green. Linguistic knowledge can improve information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 262–267, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[Wes03] Michael Wessel. Some practical issues in building a hybrid deductive geographic information system with a dl component. In *KRDB*, pages 1–12, 2003. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-79/wessel.pdf.

[WFB04] Antoine Widlöcher, Eric Faurot, and Frederik Bilhaut. Multimodal indexation of contrastive structures in geographical documents. In *RIAO*, pages 555–570, 2004.

[Who69] B.L. Whorf. *Linguistique et anthropologie*. Paris, Denoël-Gonthier, 1969.

[WJ96] D.A. White and R. Jain. Similarity indexing with the ss-tree. In *12th Conference on Data Engineering*, pages 516–523, 1996.

[WKBH00] Steffen Werner, Bernd Krieg-Brückner, and Theo Herrmann. Modelling navigational knowledge by route graphs. *Lecture Notes in Computer Science*, 1849 :295– ? ?, 2000.

[WP94] Allison Woodruff and Christian Plaunt. GIPSY : Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9) :645–655, 1994.

[WR82] D. Wunderlich and R. Reinelt. How to get there from here. In *Speech, Place and Action*, pages 183–201. Jarvella,R. J. and Klein,W., Chichester, 1982.

[Zar99] Hugo Zaragoza. *Modèles dynamiques d'apprentissage numériquepour l'accès à l'information textuelle*. PhD thesis, Université de Paris VI, 1999.

[Zar04] H Zargayouna. Contexte et sémantique pour une indexation de documents semi-structurés. *CORIA*, pages 161–177, 2004.

[ZFN06] Amal Zouaq, Claude Frasson, and Roger Nkambou. An ontology-based solution for knowledge management and elearning integration. In *Intelligent Tutoring Systems*, pages 716–718, 2006.

Résumé

Notre travail s'insère dans la problématique de l'accès à l'information spatiale présente dans des corpus textuels territoriaux. Nous proposons d'aller au-delà des systèmes de recherche d'information classiques basés sur l'analyse statistique des documents, peu adaptés pour ce cas particulier, *via* un traitement linguistique ciblé interprétant l'information spatiale. Notre hypothèse est que des traitements relativement peu coûteux suffisent à dégager l'essentiel de l'information. Ils sont un bon point de départ pour une interprétation plus poussée par la suite, utilisant les propriétés géographiques de l'information extraite afin de développer un système d'indexation à plusieurs niveaux d'abstraction.

Nous proposons en effet une méthode de recherche d'information spatiale multi-niveaux indexant un corpus textuel brut. Cette méthode qui extrait l'information d'un corpus et l'interprète, permet d'améliorer l'efficacité de systèmes de recherche d'information à chaque fois que l'interrogation comporte une connotation spatiale. L'interprétation permet en outre de retrouver le contexte dans lequel l'information spatiale a été utilisée. En particulier, elle permet d'indexer des unités de texte en leur associant des contextes de type *itinéraire*, *description locale* ou *comparaison de lieux*.

Mots-clés: Système de recherche d'information spatiale, Analyse textuelle de documents non-structurés, Raisonnement Spatial Qualitatif, Systèmes d'Information Géographique.

Abstract

The aim of our work is to provide a more easier way to access documents in territorial copora and, particularly, spatial information contents. We suggest to go further than classical based-statistic information retrieval systems that are not suitable in the case of spatial information extraction. A light linguistic process can rather be used in order to draw the information's main thing. They can be a good starting point to be used thereafter in a more precise interpretation process, using only the geographic properties extracted, in order to propose a multi-level indexing method, each level corresponding to an abstraction level of spatial information.

Thus, we propose a multi-levels spatial information retrieval system, indexing unstructured textual documents. This method, that interprets spatial information, allows to improve the efficiency of information retrieval systems each time a spatial query is performed. This interpretation can also retrieve the context in which the spatial information is used by the author. Particularly, text units can be classified in *itinerary*, *local description* or *area comparison* contexts.