Utilisation de bases de données dans le cadre de traitement de texte en langage naturel
Modèle de raisonnement

Vincent Maya
LIMSI-CNRS
Bat. 508, Université Paris-sud
91403 Orsay Cedex
Email : maya@limsi.fr

Résumé *Dans le but de traiter des textes en langage naturel prenant comme références des éléments de l'environnement géographique, nous avons défini un modèle de raisonnement permettant de s'adapter simplement aux différentes organisations de bases de données géographiques, représentations de l'espace fournissant les informations permettant raisonner sur l'espace. Nous présentons ici les différents types d'informations à manipuler dans ce cadre de travail, les contraintes à respecter, les méthodes de traitement et la représentation dans le système de tous ces éléments.*

1 Cadre de travail

Notre objectif est de définir un modèle de raisonnement capable de traiter les informations sur un environnement décrit par une base de données géographiques afin d'interpréter d'une description d'itinéraire en langage naturel. Ce modèle doit pouvoir, principalement, traiter les données généralement qualitatives d'un texte, à partir de bases de données correspondant à des modèles quantitatifs de l'espace géographique ([Nun91]). Il doit également permettre de mémoriser et d'exploiter, en plus de l'environnement global de la base, le contexte correspondant à chaque étape de l'itinéraire (i.e. l'environnement, la position de l'agent du déplacement, son orientation) et d'identifier, à partir d'une description dans le texte et du contexte engendré par les recherches précédentes, les différents éléments associés à cette description parmi tous ceux présents dans la base de données. Nous signalons que ces éléments ne sont pas tous cités explicitement dans le texte.

Au rang des difficultés, nous citons en particulier les contraintes liées aux ressources. Nous devons donc définir un modèle de raisonnement permettant de prendre en compte les particularités organisationnelles de chaque base de données géographiques

Dans cet article, seules les questions liées directement au raisonnement seront traitées, le traitement syntaxique et sémantique du texte et l'analyse du contenu sémantique et conceptuel de la base de données (traitée dans [May00]) n'étant pas pris en compte. Ainsi, nous supposerons que les diverses analyses nous permettent de repérer dans le texte les repères, les différents types d'actions et les prépositions en rapport avec des notions spatiales.

Dans les parties 2 et 3, nous décrirons les différents types d'informations pouvant être contenues dans des bases de données. À partir de ces constatations, nous définirons les bases d'un modèle de raisonnement en énumérant les différentes contraintes existantes dans la partie 4 et en esquissant les méthodes de raisonnement dans la partie 5. La partie 6 sera consacrée à la description de la représentation des différentes informations utiles que nous jugeons la plus adéquate. En dernier lieu, nous parlerons des limites de notre modèle.

2 Types de repères

Notre but étant d'obtenir la représentation d'un itinéraire en terme d'éléments de la base de données, il faut pour le moins que cette base de données contienne une représentation du réseau sur lequel a lieu le déplacement. Ce réseau devra être décomposé en tronçons, tels que chaque connexion entre ses différentes branches corresponde à une extrémité de tronçons (les tronçons étant représentés par les coordonnées de leurs extrémités).

De façon facultative, les bases de données pourront contenir des informations sur la nature particulière de certains tronçons remarquables (e.g. pour le réseau routier ces tronçons peuvent être des ponts ou des tunnels) et ces tronçons pourront être regroupés, notamment suivant des notions de continuité (e.g. pour le réseau routier, les tronçons sont regroupés en rues). Nous pouvons également avoir des informations complémentaires sur la nature des connexions entre tronçons (intersections de rues, carrefours complexes ou ronds-points).

Les trois grands types de réseaux sont le réseau routier (qui peut être étendu au réseau piéton), le réseau hydrographique et le réseau ferré. À partir des différents réseaux, il est possible d'identifier des tronçons ou des intersections de tronçons, par superposition de réseaux (e.g. "Tu traverses la rivière...", “tu passes un pont.”, "...la voie qui passe sous la ligne de chemin de fer") ou par proximité entre deux réseaux ("tu longeras la ligne de chemin de fer..."). Nous opposerons le réseau sur lequel a lieu le déplacement aux autres réseaux en différenciant le réseau principal et les réseaux secondaires.

Dans les descriptions d'itinéraires, le déplacement sur le réseau principal et, en particulier, les points critiques du déplacement peuvent également être identifiés grâce des informations provenant de l'environnement immédiat. En milieu urbain, les éléments architecturaux sont fréquemment utilisés (e.g. “En sortant de la gare,...”). Ces repères peuvent correspondre non seulement à un élément de l'environnement mais aussi à un ensemble d'éléments, voire à des zones (e.g. zones industrielles ou quartiers). Nous désignerons la représentation de ces éléments dans la base de données comme repères externes (aux réseaux).

3 Informations sur les repères

L'une des principales opérations du système est de retrouver un référent à partir de sa description en langage naturel. Pour cela, nous devons "traduire" les informations du texte en ne prenant en considération que les informations disponibles dans la base de données (sur la nature des repères ou sur leur position relativement à un repère absolu ou à d'autres repères). Ces informations peuvent concerner soit la nature et la caractéristiques des repères soit leur situation spatiale.

3.1 Nature et caractéristiques des objets

Les représentations des repères dans une base de données géographiques sont, en général, regroupées suivant la nature ou les propriétés de l'élément qu'ils modélisent. Cependant, la nature des éléments décrits ne se retrouve pas nécessairement dans l'organisation des données. Ainsi, certains regroupements peuvent paraître inappropriés d'un point de vue cognitif : des objets de nature identique peuvent être éparpillés (e.g. les bâtiments étant répartis dans des classes différentes en fonction de leur superficie) ou inversement des regroupements "abusifs" (e.g. les repères étant regroupés en fonction de leur superficie, les bâtiments étant mis dans la même classe que les places, les parkings, les aérodromes, les gares, les gares routières et les zones industrielles). Pour pallier ces difficultés, ces classes d'objets possèdent un attribut pour distinguer la nature de leurs différents éléments. Ces attributs permettent également d'apporter des précisions sur des objets de même nature (permettant de distinguer les ponts ou les tunnels des tronçons "ordinaires") ou des informations complémentaires, telles que le fait qu'un tronçon de rue soit en sens unique ou qu'une rue soit en double voie, voire qu'un tronçon soit bordé ou non d'un trottoir.

Nous faisons remarquer également qu'une base de données ne peut contenir toutes les caractéristiques utilisables dans une description d'itinéraire. Il faut donc recenser celles qui sont effectivement disponibles.

3.2 Informations de nature spatiale

D'après nos spécifications (cf. la partie 2 sur les différents types de repères), les tronçons permettant de modéliser les différents réseaux et les repères externes sont tous associés à une ou à plusieurs coordonnées spatiales. Les tronçons, par exemple, sont décrits par les coordonnées de leurs extrémités. Grâce à ces coordonnées, nous disposons des informations nécessaires permettant de déduire un certain nombre de relations spatiales présentes dans des prescriptions telles que "Tourner à droite..." ou "Passer sous le pont" (prépositions projectives), "Aller vers le nord", avec éventuellement des informations numériques ("Tourne deux fois à droite", "prends la deuxième à gauche"). Si la localisation de repères externes se fait par rapport à un tronçon, le raisonnement spatial peut se ramener à la localisation d'un point (représentant la localisation du repère externe) par rapport à un segment. En effet, positionner un repère par rapport un tronçon reviendra à définir la position d'un point par rapport à un segment (droit/gauche, voire, pour le cas particulier des impasses, droite/gauche/devant). Ne sachant pas a priori le nombre de coordonnées servant à représenter les différents repères (soit un unique point, soit les contours au sol, soit le volume), ce point pourra être soit un de ceux décrivant le repère, soit calculé à partir de l'ensemble des points représentant le repère.

L'information spatiale peut également être décrite sous forme relationnelle. En effet, chaque objet de la base représentant un repère de l'environnement (ou une partie d'un repère), l'association de deux objets se rapporte à des notions géographiques. Les relations de composition, par exemple, renferment des informations spatiales : si une rue est décomposée en un ensemble de tronçons, les coordonnées de la rue correspondent aux coordonnées des tronçons qui la composent. De même, l'appartenance de tronçons à une même rue introduit des notions de connexité entre ces tronçons. Les relations entre les objets peuvent également correspondre à des notions de proximité ou d'adhérence. Dans les bases de données étudiées dans la phase préliminaire de notre travail, un type de relations associait un tronçon à un repère externe, ces repères externes liés étant toujours "remarquables" et donc ayant un intérêt pour l'itinéraire (repère critique, correspondant à un changement de direction, ou repère témoin, servant à valider les choix effectués jusque là). Les relations peuvent être même utilisées pour définir les superpositions de réseaux (e.g. entre le réseau routier et le réseau ferroviaire ou le réseau hydrographique), sans que les réseaux secondaires ne soient explicitement décrits dans la base de données.

Les attributs d'objets de la base peuvent également contenir des informations d'ordre géographique. Ainsi, à partir de l'attribut qui, à un bâtiment, associe une adresse, il est possible non seulement de savoir à quelle rue est rattaché un bâtiment, mais aussi, potentiellement, à quel tronçon de cette rue si, pour un tronçon, nous possédons les numéros des bâtiments qui le bordent (i.e. le minimum et le maximum des numéros pairs et des numéros impairs).

4 Contraintes

Pour raisonner à partir d'un texte, il faut pouvoir identifier un objet en fonction de sa nature, ce qui nécessite de connaître l'organisation de la base de données, mais aussi de transposer les descriptions de l'espace (e.g. direction et prépositions projectives), données en termes qualitatifs, en les adaptant au contenu quantitatif des bases de données. Cela revient à un passage d'une représentation qualitative de l'espace, telle que celles définies, par exemple, par Hernandez dans [Her94], à une représentation quantitative correspondant au contenu de la base de données, travail qui a été au centre d'un certain nombre d'études, telles que celle réalisées dans [Gon99].

Pour réaliser ce travail, nous devons tenir compte de nombreuses contraintes.

4.1 Contraintes de ressources

Les principaux facteurs influant sur le raisonnement à partir du contenu de bases de données sont des contraintes de ressources. Il n'est, en effet, pas possible de connaître a priori ni le contenu d'une base de données ni son organisation. De plus, en fonction de la base de données utilisée, les informations peuvent donc être obtenues sous des formes différentes (informations géométriques, relationnelles ou contenues dans les attributs d'un objet). Ainsi, la mise en relation d'un bâtiment avec le tronçon de rue "qui lui passe devant" pourrait se faire soit en considérant les coordonnées des différents éléments et en choisissant parmi les tronçons celui avec lequel la distance au bâtiment est la plus faible, soit en fonction d'une relation associant tronçon et bâtiment, soit en considérant l'adresse du bâtiment, si elle est contenue dans la base de données, qui contient le nom de la rue.

Il nous faut donc indiquer à notre système la méthode d'accès à l'information la mieux adaptée en fonction de l'organisation et du contenu de la base de données (cf. la section 6).

4.2 Contraintes liées à la nature des descriptions

Les descriptions d'itinéraires, dont les particularités ont été étudiées dans de nombreuses études (comme par exemple dans [DD97], [Den97], [Gol95]), engendrent de nombreuses contraintes.

En premier lieu, les descriptions font appel à la perception de l'agent du déplacement. Il nous faut, en quelque sorte, modéliser les limites perceptives du sujet, alors que nous travaillons sur une description globale de l'environnement (i.e. description contenue dans la base de données géographiques). Ainsi, s'il est fait mention d'un pont sans autre indication, nous supposons que ce pont fait partie de la rue à laquelle appartient le tronçon courant ou qu'il possède une intersection commune à cette rue.

De plus, les différents corpus de descriptions montrent que selon les individus, la description d'un même itinéraire peut varier dans la forme et dans le contenu, notamment en taille, en nombre de repères évoqués et en structuration du texte. Nous sommes donc contraints de définir un certain nombre d'hypothèses, à partir des études sur les descriptions d'itinéraires, afin de pouvoir limiter dans une certaine mesure la diversité des cas à traiter. Nous supposerons, par exemple, que les actions successives d'une description d'itinéraire, sont décrites globalement dans leur ordre chronologique, décision par ailleurs confortée par les différentes études.

4.3 Contraintes liées à la structure statique des bases de données

Nous devons modéliser la position mobile de l'agent du déplacement, dans le cadre d'un environnement statique. Le déplacement ayant lieu sur un unique réseau, nous baserons ce raisonnement sur le tronçon, structure minimale composant ce réseau : l'agent sera confondu avec le tronçon sur lequel il se trouve.

Le tronçon sera orienté afin de modéliser le sens de déplacement de l'agent. Cela limite le niveau de précision en fonction du découpage du réseau en tronçons. De plus, les raisonnements tournant autour du réseau principal, cela influe sur la portée du raisonnement spatial. Ainsi, la description “Le bâtiment en face de la boulangerie...” fait mention au bâtiment, qui, par rapport au tronçon routier se trouve en face de la boulangerie et donc au tronçon auquel est associée la boulangerie.

5 Raisonnement par cas

La première étape du traitement de la description consiste à reconnaître le type d'actions décrites. Les verbes utilisés (et parfois, comme il est signalé dans [Van86], la préposition qui les suivent) permettent de différencier les cas de positionnement des cas de changement de direction. Mais le véritable objectif est d'identifier les repères utilisés et la nature de ceux-ci. Cette identification débute en fait au niveau de l'analyse syntaxique, par la détermination des compléments circonstanciels de lieu, et de l'analyse sémantique, servant à associer chaque mot relatif à un repère au concept qui leur est associé.

Un changement d'itinéraire, tel que “Tourne à droite”, correspond au niveau du système à retrouver le tronçon qui possède une connexion avec la rue du tronçon courant (tronçon qui, au sein du contexte courant, indique la position de l'agent), et qui aille dans la bonne direction (gauche/ droite ou points cardinaux). Il faut cependant tenir compte de la présence éventuelle de tronçons non explicitement cités dans le texte.

Pour les positionnements, il existe plusieurs types de repères par rapport auxquels la situation géographique de l'agent mobile peut être déterminée. Le cas le plus simple consiste à atteindre un tronçon particulier à partir du tronçon courant. Ce tronçon particulier peut être identifié, par exemple par sa nature particulière (e.g. dans "Arrivé au pont...", le tronçon particulier correspond à un pont). Une fois ce tronçon déterminé, la seconde opération consiste à accéder du tronçon courant au tronçon associé au repère externe, en tenant compte des tronçons non explicitement cités. Plusieurs variantes existent. La première consiste à relier le tronçon courant non pas à un autre tronçon, mais à une extrémité de tronçon correspondant à une intersection de rues. Il peut s'agir également de positionner le mobile sur le tronçon correspondant à un repère externe. Dans ce dernier cas, la première opération consiste à associer un tronçon au repère externe concerné. Une fois ces tronçons identifiés, le problème se ramène à l'opération de déplacement sur des éléments du réseau principal, tous n'étant pas nécessairement explicitement cités dans la description.

De plus, puisque nous ne connaissons pas a priori le contenu de la base de données géographiques utilisée, notre raisonnement doit pouvoir gérer l'absence d'information : absence de représentation d'un type de repère dans la base, absence de l'attribut recherché pour un repère possédant une représentation dans la base ou absence d'objet respectant les indications fournies dans la description. Dans les deux premiers cas, il s'agit clairement d'un cas où l'information requise ne peut, du fait de l'organisation de la base de données, se trouver dans celle-ci. Face à cette situation, nous souhaitons pouvoir continuer le traitement de la description d'itinéraire : si la représentation est manquante, il faut ignorer l'information liée au repère, les redondances permettant fréquemment de poursuivre la compréhension de l'itinéraire. Dans le dernier cas, le contenu de la base de données et celui de la description d'itinéraire ne sont pas compatibles et la compréhension de l'itinéraire ne peut se poursuivre.

6 Représentation

Notre représentation devra contenir plusieurs types de connaissances : des connaissances sur la base de données utilisées, son contenu et son organisation, des connaissances sur les repères utilisables dans les descriptions d'itinéraires, et des connaissances sur les relations entre ces deux types de connaissances.

Nous utilisons un formalisme basé sur les graphes conceptuels. Nous différencions quatre types de concepts : les concepts-repères, identifiant les différents concepts relatifs aux repères utilisables dans les descriptions d'itinéraire, les concepts-objets et des concepts-attributs, utilisés pour représenter l'organisation de la base de données et les concepts-relations, correspondant aux opérations de bases prédéfinies dans notre modèle disponibles dans la base de données. Les informations relatives aux différents concepts seront représentées à l'intérieur de graphes canoniques (pour les informations sur les liens entre langage naturel et base de données) ou de graphes prototypes (pour les informations sur l'organisation de la base de données), des exemples de représentations étant donnés dans les figures 1 et 2. Le choix de cette représentation résulte notamment de l'existence de nombreuses études, tels [BCHL93], développant des méthodes permettant de lier graphes conceptuels et bases de données et d'exprimer les concepts des bases de données tels que les schémas, les données, les requêtes et les vues.

L'organisation des concepts-repères au sein du treillis des types (en considérant qu'ils représentent tous les repères utilisables dans une description d'itinéraire) reflète la hiérarchie lexicale entre ces concepts. Ainsi, les fils d'un concept sont des spécialisations de celui-ci : le concept représentant un pont ou un tunnel (qui sont des tronçons particuliers) sera un fils du concept tronçon.

À ces concepts-repères, nous associons, dans les graphes canoniques, les concepts correspondant aux objets contenus dans la base de données (concepts-objets). Pour obtenir les types d'objets qui peuvent correspondre à un repère donné, nous définissons l'opération **RelationObjetConcept**, représentée par le concept-relation de même nom, qui, pour le concept-repère qui représente un repère (les mots synonymes correspondant au même concept-repère), renvoie l'ensemble des concepts-objets et donc aux types d'objets pouvant être associés au mot (ainsi que les requêtes permettant d'obtenir ces objets dans la base de données). Ainsi, la figure 1 indique que dans l'exemple étudié les ponts sont représentés dans la base de données par des objets de la classe ObjTRONCON et spécifie la requête permettant de les différencier des autres objets de cette classe. L'absence de concept-relation à partir d'un concept-repère et la possibilité de remonter parmi ses pères permet généraliser la recherche d'objets : la recherche d'une mairie dans une base de données ne faisant pas cette distinction correspond à la recherche d'un bâtiment administratif voire, si ce type n'est pas non plus référencé, d'un bâtiment.

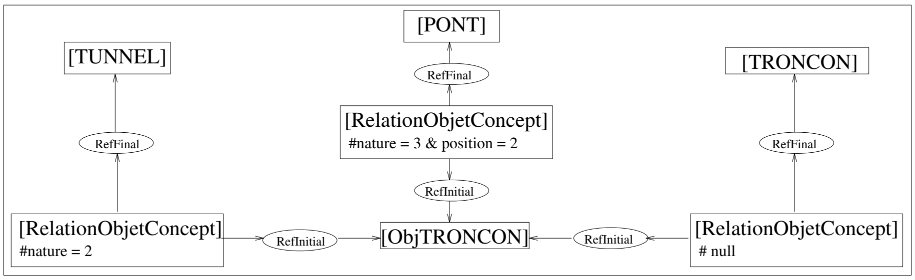

FIG. 1: Graphe conceptuel du concept-objet [Tronçon]

L'obtention des attributs accessibles pour un objet donné se fait par l'opération **Attribut** (qui présente la particularité de ne pas correspondre à un concept-relation, mais à une simple relation), qui pour un concept-objet renvoie les concepts-attributs qui lui sont associés dans le graphe prototype. Cela permet de travailler à partir de l'indication d'un attribut dans une description. Dans l'exemple présenté par la figure 2, la base de données permet de différencier les tronçons en sens unique ou à double voie.

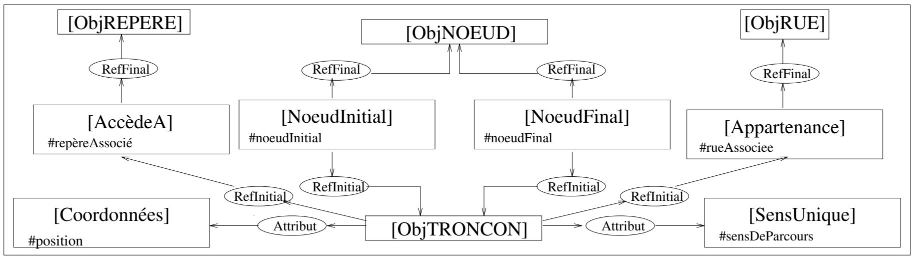

FIG. 2: Graphe prototype du concept-objet [Tronçon]

Les autres opérations prédéfinies (que nous ne détaillerons pas dans cet article) et possédant des relations équivalentes dans la base de données au niveau relationnel sont modélisées par des concepts-relations et décrites dans le graphe prototype des concepts-objets associés. Ces concepts-relations permettent d'accéder aux concepts-objets qu'elles associent, ainsi qu'à la requête correspondante. Ainsi l'opération **Compositoin**, qui renvoie pour un tronçon ses extrémités, correspond, d'après nos spécifications, obligatoirement à un concept-relation, devant être toujours définie. Elle est en fait est décomposée dans le graphe prototype de la figure 2 en deux concepts-relation, **NoeudInitial** et **NoeudFinal**. L'opération **Appartenance**, qui associe la représentation d'un tronçon à la rue à laquelle il appartient, est représentée dans la figure 2 par un concept-relation de même nom.

Ces opérations sont définies aussi bien entre des concepts appartenant au réseau principal, qu'entre un concept du réseau principal et un concept de repère externe (telle que **AccèdeA** permet d'obtenir les repères externes qui sont associés à un tronçon, l'opération inverse étant définie dans le graphe prototype des concepts-objets associés aux repères adéquats).

Cependant, certaines des opérations prédéfinies par notre modèle peuvent ne pas avoir de concept-relation associé en raison de l'absence dans la base de données utilisée de l'information associée. Pour les opérations le permettant, notre système contiendra, par défaut, une définition au niveau géométrique. Ainsi les différentes opérations **PositionRelative**, indiquant la position d'un tronçon contigu, d'un point représentant la position d'un repère externe ou d'un segment d'un réseau secondaire par rapport à un tronçon de référence, ou l'opération **Direction cardinale** renvoyant la position du tronçon relativement aux points cardinaux sont fréquemment définies au niveau géométrique.

7 Limites

Les principales limitations sont dues, de fait, au traitement par cas des phrases. Il est possible qu'une phrase, bien construite mais faisant référence, pour une même action, à de nombreux repères, ne rentre pas dans un des cas prévus, ou plus certainement, que certains de ces repères soient ignorés. Un tel cas est cependant peu probable, l'étude des différents corpus n'ayant pas, pour l'instant, révélé de telles situations, les phrases des descriptions d'itinéraires se caractérisant en général par des phrases simples ne mentionnant qu'un nombre limité de repères.

Le second type de limitation découle de la disproportion potentielle entre le contenu de la description d'itinéraire traitée et du contenu sémantique de la base de données utilisée. En effet, si certains repères-clés de la description ne sont pas représentés, le raisonnement deviendra vite impossible.

Nous sommes également restreints par nos hypothèses servant entre autres à modéliser les limites de la perception humaine et par celles caractérisant les descriptions d'itinéraires. Mais ces hypothèses étant vérifiées par la majeure partie des descriptions, cette limitation reste acceptable.

8 Conclusion

Nous avons réalisé cette étude dans le cadre d'un travail plus vaste, explorant les différentes possibilités liées au traitement de textes en langage naturel à partir du contenu des bases de données commerciales, décrivant l'environnement urbain comprenant le campus de notre université. Cela a permis, en particulier, l'implémentation de fonctionnalités à un GIS développé au sein notre laboratoire permettant la compréhension d'une large gamme de descriptions d'itinéraire.

Références

[BCHL93] C. Boksenbaum, B. Carbonneill, O. Haemmerlé, and T. Libourel. Conceptual graphs for relational databases. In Mineau and al, editors, *Conceptual Graphs for Knowledge Representations*, LNAI n 699, pages 142–161, Quebec, Canada, Aout 93. ICCS'93, Springer Verlag.

[DD97] M.-P. Daniel and M. Denis. Spatial description as navigational aids : A cognitive analysis of route directions. Technical Report 97-13, LIMSI-CNRS, 1997.

[Den97] M. Denis. The description of routes : A cognitve approach to the production of spatial discourse. In *Cahiers de Psychologie Cognitive / Current Psychology of Cognition*, pages 409–458, 1997.

[Gol95] R. G. Golledge. Path selection and route preference in human navigation : a progress report. In A. U. Franck and W. Kuhn, editors, *International Conference on Spatial Information Theory (COSIT'95) : A Theoretical Basis for GIS*, pages 207–22, 1995.

[Gon99] M.-R. Goncalves. Formalisation hybride du raisonnement spatial. application à la consultation d'une base de données géographiques. Thèse de doctorat en informatique, Université Paris XI, Septembre 1999.

[Her94] D. Hernandez. *Qualitative Representation of Spatial Knowlegde*, volume 804 of *Lecture Notes in Artificial Intelligence*. Spinger-Verlag, Berlin, 1994.

[May00] Vincent Maya. Interfacing geographical databases and natural language : A pivot level proposal. In *ECAI 2000, Workshop on Current Issues in Spatio-Temporal Reasoning*, 2000.

[Nun91] J. Nunes. Geographic space as a set of concrete geographical entities. In D. M. Mark and A. U. Frank, editors, *Cognitive and Linguistic Aspects of Geographic Space : An Introduction*, pages 9–34, Dordrecht, 1991. Kluwer Academic.

[Van86] C. Vandeloise. *L'espace en français : sémantique des prépositions spatiales*. Editions Du Seuil, Paris, 1986.