Recherche de concepts et de relations sémantiques pour la géographie

Traiter automatiquement des documents de spécification de bases de données

Marion Laignelet*, Mouna Kamel, Nathalie Aussenac-Gilles****

** CLLE-ERSS - UMR 5610*
Maison de la Recherche - Université de Toulouse-Le Mirail
5, allée Antonio-Machado, F-31058 Toulouse Cedex 9

marion.laignelet@univ-tlse2.fr

*** Institut de Recherche en Informatique de Toulouse (IRIT) - CNRS*
UPS, 118, route de Narbonne, F-31062 Toulouse Cedex

mouna.kamel@irit.fr
aussenac@irit.fr

RÉSUMÉ. *Basée sur une méthodologie exploitant la notion de marqueur, nous proposons de repérer des marques linguistiques de concepts et de relations sémantiques en utilisant non seulement des informations de type syntaxique mais également la structure hiérarchique des documents. Nous travaillons sur des documents de spécification de base de données géographiques dont la structure et l'organisation, du fait notamment de leur fonction, est sémantiquement explicite. Matérialisée par un balisage XML, cette structure permet de construire un noyau d'ontologie qui est ensuite enrichi par des connaissances extraites à l'aide de patrons combinant analyse lexico-syntaxique et analyse structurelle.*

ABSTRACT.

MOTS-CLÉS : *ontologie, géographie, TAL, patrons lexico-syntaxiques, structure de document, repérage automatique de concepts et relations*

KEYWORDS: *ontology, geography, NLP, lexicale and syntactical patterns, document structure, automatic tracking of concepts and relations*

1. Introduction

Nous proposons de trouver en corpus des traces linguistiques de concepts et de relations sémantiques dans le but de construire une ontologie géographique à partir de documents de spécifications de bases de données. La méthode que nous proposons exploite non seulement des patrons lexico-syntaxiques [BOU 03, BOU 04, AUS 09, AUS 08] mais également des patrons basés sur la structure hiérarchique des documents et matérialisée dans nos textes par des balises XML. La mise en forme des documents de spécification de base de données géographiques de l'IGN, du fait notamment de leur fonction descriptive et explicative, est sémantiquement explicite et permet de construire un noyau d'ontologie. Ce noyau est ensuite enrichi par des connaissances analysées et extraites à l'aide de patrons plus locaux, combinant analyse lexicale, syntaxique et structurelle.

Ce travail s'inscrit dans le cadre d'un projet de recherche, le projet Géonto, dont l'objectif est de capitaliser des connaissances hétérogènes à partir de documents géographiques diversifiés. D'un point de vue méthodologique, l'objectif du projet est d'éprouver et d'adapter des outils et méthodes existants pour créer des ontologies à partir de données et de connaissances imprécises et hétérogènes. Du point de vue applicatif, les ontologies créées doivent permettre un accès facilité à de grandes masses de données géographiques. Parallèlement, nous espérons par cette méthode appréhender les différences de point de vue sous-jacentes au domaine.

Dans cet article, nous présentons la méthode que nous avons définie pour exploiter au mieux les documents de spécification de bases de données géographiques de manière à constituer une ontologie reflétant la structuration des connaissances définie par ces spécifications. L'originalité de cette méthode est de rechercher des connaissances en exploitant non seulement le langage naturel présent dans ces textes, mais aussi les indications explicites de structure qui encadrent les zones en langage naturel. Dans la section 2, nous présentons un panorama des différentes méthodes de construction d'ontologies à partir de texte, que ce soit pour la recherche des concepts ou celle des relations entre les concepts. La section 3 présente le projet Géonto et la problématique d'acquisition de connaissances pour la géographie et la cartographie. La section 4 décrit l'approche mise en oeuvre ainsi que les outils développés pour extraire les concepts potentiellement pertinents pour le domaine de la géographie et les relations sémantiques susceptibles de relier ces concepts. Enfin, dans la dernière section, nous évaluons quantitativement et qualitativement la méthode mise en oeuvre.

2. Construire des ontologies à partir de textes en langue naturelle

La construction d'ontologies, et plus généralement de ressources termino-ontologiques (RTO), à partir de textes nécessite la mise en place de méthodologies complexes exploitant massivement des outils du traitement automatique des langues [BOU 04, JAC 06, SCH 05, MAY 09, ROL 06]. L'utilisation de techniques de TAL permet notamment d'extraire les éléments qui vont constituer la RTO, ou qui vont servir d'indices à l'ontographe pour définir des éléments de RTO. Selon le type de RTO visé (hiérarchie de termes, thé-

saurus, hiérarchie de concepts, ontologie plus ou moins riche en relations) les besoins sont variables. Mais dans tous les cas, il est nécessaire d'extraire des termes, lesquels sont potentiellement aptes à être considérés par la suite comme des étiquettes de concepts pour le domaine en question.

2.1. *Méthodologies basées sur la notion de marqueur ou patrons*

L'utilisation de patrons pour extraire des connaissances est largement utilisée. [HEA 92] fait figure de référence en ce qui concerne les techniques de linguistique computationnelle. Meyer (2001) introduit la notion de patrons de connaissances (*knowledge rich contexts*) pour faire référence à patrons linguistiques exprimant diverses relations sémantiques liées àla définition, telles l'hyperonymie, la synonymie ou la méronymie. [HAD 02] définit la notion de marqueur comme une « *forme linguistique faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques) dont l'interprétation définit régulièrement le même rapport de sens entre les termes.* » Et, dans le contexte de la recherche de relations sémantiques en corpus, [GRA 04] définissent le patron lexico-syntaxique de la manière suivante : « *à la différence des marqueurs, les patrons identifient la relation recherchée plus précisément en définissant également des contraintes syntaxiques ou typographiques sur le contexte des termes.* » Dans tous les cas, un patron définit des contextes suffisamment fins et précis, relevant de niveaux linguistiques variables et révélateurs de connaissances ou de comportements linguistiques particuliers.

Dans le cas de la recherche de relations sémantiques, un patron permet d'indiquer l'existence d'une relation sémantique particulière entre deux entités (au moins). Appliquée à la construction d'ontologies, une telle approche fait l'hypothèse que les relations lexicales peuvent fournir des indices pour définir des relations conceptuelles et, avec elles, de nouveaux concepts et termes associés. [BOU 04] mettent en place une méthodologie fondée sur l'utilisation de patrons. Une première phase d'extraction de termes[1] construit un réseau de mots et de syntagmes (le « réseau terminologique ») dans lequel chaque syntagme est relié à sa tête et à ses expansions. À ce niveau, les éléments du réseau sont des candidats-termes potentiellement aptes à devenir des concepts. La seconde étape est une analyse distributionnelle[2] : à partir du réseau terminologique, un calcul des proximités distributionnelles entre les unités est effectué. La dernière phase extrait les relations[3]. Des outils de modélisation[4] permettent *in fine* de définir les concepts à retenir. Dans cette approche, il est important de noter que les termes sont les occurrences des mots du corpus et les relations sont extraites à l'aide de de patrons lexico-syntaxiques. D'autres travaux mettent en oeuvre cette méthodologie ([JAC 06]).

---

1. À l'aide du logiciel Syntex, analyseur syntaxique.
2. À l'aide du logiciel Upery.
3. À l'aide de Yakwa (concordancier) et Caméléon (logiciel de recherche de relations lexicales à partir de marqueurs linguistiques).
4. TermOnto, Terminae

La question des patrons lexico-syntaxiques pour la création et la population d'ontologies est également traitée dans [MAY 09]. Ces auteurs proposent la notion de patrons lexico-syntaxiques (*lexico-syntactic ontology design patterns*, ODPs) qui génèrent de l'information ontologique à partir de textes non-structurés dans le but de créer une nouvelle ontologie ou enrichir des ontologies existantes.

2.2. *Les verbes pour exprimer les relations*

Au coeur d'un patron lexico-syntaxique traduisant une relation lexicale, se trouve souvent un verbe comme indice fort de cette relation. Par exemple, *comporter*, *contenir*, *se composer de* peuvent marquer la relation d'un tout vers ses parties. [SCH 05] insiste sur le rôle des verbes comme éléments centraux pour déterminer une relation entre deux concepts d'une ontologie. Pour les auteurs, les verbes spécifient l'interaction entre les participants d'une action ou d'un événement. Leur système identifie des triplets (*i.e.* des paires de concepts reliés par une relation) qui pourront être intégrés à une ontologie existante. Cet outil extrait les termes pertinents et les verbes des textes à l'aide d'une approche combinant des techniques linguistiques et statistiques. L'approche mise en place est de type linguistique. Les structures de dépendance sont reconnues : fonction grammaticale, structure des phrases, étiquetage morpho-syntaxique et lemmatisation, normalisation des syntagmes nominaux complexes sur la base de leur tête nominale. Une reconnaissance des entités nommées et des concepts propres au domaine (le foot) est également menée (utilisation de l'ontologie du domaine) ainsi que la reconnaissance des synonymes. Enfin, un processus statistique fournit des mesures de pertinence et des mesures de cooccurrence. L'approche se veut robuste et adaptable à d'autres domaines.

2.3. *Utilisation de la structure des documents*

L'idée d'exploiter la structure des documents pour améliorer le processus d'acquisition des connaissances est proposée par [ROL 06]. Les auteurs partent du constat suivant lequel les documents ont la plupart du temps une structure logique cohérente et porteuse de sens. C'est le cas, par exemple, des manuels techniques, des dictionnaires, des codes juridiques, etc. Généralement, cette structure se matérialise en XML. Leur objectif est de construire une ontologie des plantes tropicales en s'appuyant sur la structure du document[5] à deux niveaux, d'un côté pour obtenir une hiérarchie de classes reflétant la taxinomie botanique traditionnelle (famille, genre, espèce), d'un autre pour cibler plus précisément les traitements linguistiques à effectuer pour compléter la hiérarchie de classes par des indications méronymiques. Les auteurs insistent sur le fait que la botanique est un domaine qui se prête naturellement à la représentation de taxinomies : le découpage en genres et en espèces est représenté dans la structure du document et permet ainsi d'initialiser la hiérarchie de classes de l'ontologie. Le résultat produit consiste en

5. La *Flore du Cameroun* qui représente 40 volumes publiés entre 1963 et 2001. La structure des différents volumes est assez régulière, soit une fiche par espèce.

une première ontologie intermédiaire[6] qui facilite les traitements linguistiques ultérieurs plus complexes pour créer une ontologie du domaine. D'autres travaux comme [DES 00] utilisent également la structure des documents pour construire des ontologies du domaine.

Dans tous les cas, construire des RTO à partir de textes n'est pas une tâche triviale et l'importance du domaine, de la qualité du corpus ou encore de la finalité de la ressource sont essentiels.

3. Acquisition de connaissances dans le domaine de la géographie-cartographie

Les données géographiques sont à la fois nombreuses et diverses. Par exemple, l'INSEE propose une ontologie en français[7] qui décrit les événements liés à la construction ou la modification des territoires ; le projet TOWNTOLOGY [8] développe une ontologie dédiée à l'enseignement de l'urbanisme. A notre connaissance, il n'existe pas d'ontologie dédiée aux objets cartographiques, en français du moins[9]. Dans ce contexte, le projet Géonto, dans lequel cet article s'inscrit, propose des méthodes et des outils permettant la description et l'intégration cohérente des données géographiques [MUS 09, KER 09]. Pour répondre à cet objectif, nous proposons la construction d'une ontologie du domaine qui doit permettre l'intégration de sources d'information multiples et hétérogènes. La méthodologie mise en place suit les étapes suivantes :

– création d'une première ontologie (le noyau),

– enrichissement de cette ontologie à partir des spécifications de bases de données topographiques de l'IGN.

Ce qui nous intéresse plus particulièrement concerne la conception et l'enrichissement du noyau d'ontologie.Le noyau de l'ontologie est obtenu à partir de l'exploitation de la structure des documents de spécification des bases de données. L'enrichisssement se fait à partir de l'analyse de la structure et de l'analyse linguistique de zones textuelles spécifiques, les zones de définition.

Dans les figures suivantes, un extrait du corpus est présenté, d'abord sous leur format original puis en XML.

D'une manière générale et comme cela a été dit dans la section 2, l'extraction de connaissances à partir de textes nécessite la constitution d'un corpus. Il s'agit d'une étape importante car, dans une approche automatisée, le corpus est la source essentielle d'information : la taille du corpus doit être suffisamment importante pour fournir une couverture large du domaine et il doit être suffisamment homogène pour permettre le repérage de régularités. Par ailleurs, [BOU 04] affirment qu'il est impossible de définir *a priori* des instructions méthodologiques précises car le processus de construction reste lié à l'application et aux exigences des spécialistes du domaine.

---

6. Dans notre travail, nous parlons de noyau d'ontologie.
7. `http://www.insee.fr/fr/methodes/default.asp?page=xml/xml.htm`
8. `http://liris.cnrs.fr/townto/`
9. La situation pour les ontologies en anglais est différente.

A – Voies de Communication Routière

Tronçon de Chemin

Définition : Voie de communication terrestre non ferrée destinée aux piétons, aux cycles ou aux animaux…

Regroupement : Voir les différentes valeurs de l'attribut <nature>.
Sélection : Voir les différentes valeurs de l'attribut <nature>.
Modélisation géométrique : A l'axe, au sol.

Attribut : Nature

Définition :	Permet de distinguer plusieurs types de voies de communication terrestres.
Type :	Énuméré
Valeurs :	Chemin empierré / Chemin / Sentier / Escalier / Piste cyclable

Nature = « Chemin empierré »

Définition : Route sommairement revêtue ou chemin empierré (pas de revêtement de surface ou revêtement très dégradé), mais permettant la circulation de véhicules automobiles de tourisme par tous temps.
Regroupement : Allée (carrossable) | Piste | Route empierrée
Sélection : Toutes les routes empierrées sont incluses.
…

Attribut : Franchissement

Définition :	Attribut indiquant la présence d'un obstacle physique dans le tracé d'une route et la manière dont il est franchissable.
Type :	Énuméré
Valeurs :	Bac piéton / Gué ou radier / Pont / Tunnel / Sans objet

Franchissement = « Gué ou radier »

Définition : Passage naturel ou aménagé permettant de traverser un cours d'eau sans avoir recours à un pont ou un bateau.
Regroupement : Gué | Radier
…

Attribut : Nom

Définition :	Nom du chemin.
Type :	Caractères
Valeur nulle :	Le champ contient la chaîne de caractères "Valeur non renseignée" pour tous les chemins n'appartenant pas à un grand itinéraire routier nommé (référence = BDCarto).

Figure 1. *Un extrait du corpus : sa mise en forme*

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<document> <domaine>
  <nom_domaine>A – Voies de communication routière</nom_domaine>
  <classe>
    <nom_classe>Tronçon de chemin</nom_classe>
    <definition>Voie de communication terrestre … </definition>
          …
    <regroupement> Voir les différentes valeurs de l'attribut &lt;nature&gt;
    </regroupement>
          …
    <attributs>
    <attribut>
      <nom_attribut> Nature </nom_attribut>
      <definition> Permet de distinguer plusieurs… </definition>
      <type> Énuméré</type>
      <valeurs> Chemin empierré/Chemin/Sentier/Escalier … </valeurs>
      <attribut_valeur>
        <valeur> Chemin empierré </valeur>
        <definition> Route sommairement revêtue … </definition>
        <regroupement> Allée (carrossable)|Piste|Route empierrée
        </regroupement>
      </attribut_valeur>
          …
    </attribut>
  </classe></domaine> </document>
```

Figure 2. *Un extrait de la version XML du corpus*

La construction d'une ontologie (dite de référence) dans le projet Géonto a pour but de faire collaborer des données géographiques hétérogènes. Dans ce contexte, le sens des termes et des concepts associés est susceptible de varier selon les sous-domaines de la géographie. Par exemple, le terme de grotte peut être relié à des concepts différents selon que l'on prend le point de vue d'une carte routière ou celui d'une carte de randonnée : dans un cas, "grotte" est relié au concept "curiosité touristique" par une relation d'hyperonymie alors que dans un contexte de randonnée, une relation d'hyperonymie avec le concept "cavité souterraine" serait sans doute plus appropriée, peut-être même en association avec l'idée de danger. Dans notre corpus, l'exemple du pont est plus concret : le "pont" peut être une sorte de "franchissement" lorsqu'un objet passe par dessus un autre ou une sorte d'"obstacle" lorsqu'il passe par dessous. Dans ce contexte, il est nécessaire de recourir à un corpus suffisamment large et diversifié pour être capable de mettre au jour l'ensemble des points de vue possibles pour un même objet.

Les documents sur lesquels nous travaillons sont des documents de spécification de bases de données géographiques fournies par le COGIT (IGN). Nous disposons de deux textes différents, BDCarto et BDTopo. L'ensemble de ces deux corpus constitue un corpus de 23 884 mots (17 069 pour BDCarto et 6 815 pour BDTopo). Ces textes sont en XML. Les connaissances exprimées dans ces documents est cependant relativement variée puisqu'ils décrivent les objets nécessaires à la création de cartes, qu'elles soient touristiques, routières, pédestres, etc. Des points de vue différents sur les objets manipulés dans le domaine sont ainsi mis en valeur, notamment dans les zones de définition[10]. En plus d'un travail sur la structure logique de ces textes (notamment à travers l'organisation des titres et des sous-titres) qui founit des indications de relation d'hyperonymie, nous proposons une étude approfondie des zones de définition. À ce propos, [AGU 09] constatent que dans un contexte définitoire, les relations sémantiques à considérer sont la synonymie, la méronymie, la causalité et le but.

4. Présentation de la méthode

Dans cette section, nous présentons brièvement le module de construction du noyau d'ontologie. Cette partie du travail a été décrit avec précision dans [KAM 09a] et [KAM 09b]. Nous focalisons notre attention sur la chaine de traitement destinée à l'enrichissement du noyau d'ontologie. Cette chaîne met au jour des concepts et des relations sémantiques entre ces concepts à l'aide de patrons exploitant à la fois des informations lexicales, syntaxiques et structurelles.

4.1. *Construction du noyau d'ontologie*

Partant du constat que les éléments de structure servent à organiser, subdiviser, hiérarchiser le contenu d'un document, alors cette structure est porteuse de sémantique, soit des relations hiérarchiques entre unités textuelles. Partant de là, il est possible de construire

10. Matérialisées par des balises XML spécifiques.

un noyau d'ontologie valide pour un corpus donné. L'idée de patron structurel proposé et mis en oeuvre par [KAM 09b] permet d'associer une propriété sémantique à un élément structurel. Le patron lexico-syntaxique, quant à lui, prend généralement en compte des éléments lexicaux et syntaxiques présents dans le texte non structuré. Plus précisément, les titres (emboitement ou parallélisme de titres et sous-titres) ou encore les structures énumératives peuvent permettre de mettre en lumière des relations de subordination ou de juxtaposition entre les termes contenus dans ces mêmes éléments structurels. Les auteurs insistent sur la nécessité de disposer d'un document au format qui rend explicite sa structure et son caractère hiérarchique, par exemple, le XML. Enfin, les termes doivent désigner sinon des concepts, au moins des termes valides pour le domaine visé. La démarche proposée se fait en quatre étapes : *(i)* associer une sémantique aux balises et aux relations qui les relient ; *(ii)* ré-annoter (automatiquement) le document pour mettre en évidence les propriétés des balises et pour attribuer des catégories grammaticales ou sémantiques aux unités textuelles ; *(iii)* définir les patrons structurels qui caractérisent la sémantique portée par un élément de structure et produisent un fragment d'ontologie ; et *(iv)* projeter les patrons. Cette méthode[11] décrite dans [KAM 09a] met en évidence 1183 concepts extraits à partir de la structure.

4.2. *Chaine de traitement pour l'enrichissement du noyau d'ontologie*

Cette chaîne de traitement a été produite avec le logiciel LinguaStream[12]. Son objectif est d'extraire automatiquement dans des textes en langage naturel des concepts et des relations entre ces concepts pour produire des fragments d'ontologie (de la forme : terme/concept-relation-terme/concept). La chaîne de traitement se décompose en cinq modules principaux : *(i)* la préparation du corpus en XML, *(ii)* les pré-traitements linguistiques et le repérage des syntagmes nominaux, *(iii)* la recherche de relation, *(iv)* la prise en compte de la structure (les titres exclusivement), *(v)* l'export en XML.

4.3. *La préparation du corpus en XML*

Ce module permet d'abord de sélectionner les textes du corpus à analyser ainsi que, à l'intérieur des textes, les zones textuelles délimitées par des balises particulières. Pour ce qui nous concerne, nous exploitons les zones de titres et les zones de définition (cf. 1 et 2.

4.4. *Les pré-traitements linguistiques et le repérage des syntagmes nominaux*

Cette étape est constituée de plusieurs modules. Tout d'abord, un segmenteur, fourni nativement par LinguaStream, permet de découper le texte en mots. Puis l'analyseur morpho-syntaxique TreeTagger[13] fournit un étiquetage morpho-syntaxique des mots du

11. Développée à l'aide de Gate (`http://gate.ac.uk`)
12. `http://www.linguastream.org/`
13. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

corpus. Un troisième module, basé sur des macro-expressions régulières, permet le découpage du corpus en phrase. 23 884 mots et 1 978 phrases sont annotés dans le corpus. Le repérage des syntagmes nominaux est réalisé à l'aide d'une grammaire EDCG. Il permet de repérer et d'annoter :

– des SN syntaxiquement simples (*un gué*)

– des SN avec des noms composés, [N de N] (*l'hôtel de police*)

– des SN plus complexes, constitués soit d'adjectifs (*un chemin étroit, un bâtiment industriel et commercial*), soit de compléments du nom "spéciaux" où l'un des éléments introduit une relation sémantique particulière comme *partie*, *tas*, *ensemble*, etc. (*un tronçon de route*)

En plus du repérage des syntagmes nominaux, une structure de traits sémantiques est associée automatiquement à chacune des expressions. Les informations suivantes sont associées : le type de marqueur éventuellement présent (méronyme, holonyme, hypéronyme, mot composé, mot simple), le lemme court (dans la plupart des cas le lemme du nom tête du SN), le lemme long (totalité de l'expression nominale) et des informations de type syntaxique. Ce module annote 6 244 syntagmes nominaux. Par exemple, l'occurrence *Ligne de métro* sera annotée comme cela :

```
<lss :sem type="sn" id="296">
<lss :text>Ligne de métro</lss :text>
<lss :value>
<marqueur>compose</marqueur>
<lemmasCourt>ligne</lemmasCourt>
<lemmasLong>ligne de métro</lemmasLong>
<syntaxe>null</syntaxe>
</lss :value>
</lss :sem>
```

Figure 3. *Exemple de repérage et d'annotation d'un syntagme nominal*

Sur la base de l'ensemble de ces traitements, l'étape suivante consiste à mettre au jour des relations sémantiques entre les syntagmes repérés, sans préfigurer, à ce niveau de l'étude, de leur qualité de concept ou non.

4.5. *La recherche de relations sémantiques*

Cette étape se concentre sur le repérage des relations sémantiques. Sont repérées les relations traditionnelles telles que l'hyperonymie et la méronymie. Les deux grammaires correspondantes ont été créées et adaptées à partir du travail accompli avec Caméléon [AUS 08]. Il s'agit à la fois d'une adaptation et d'un enrichissement de ces travaux relativement nos propres problématiques et nos corpus. Concernant les relations d'artefact et de fonction, elles ont été créées pour le projet Géonto, sur la base d'une étude manuelle du corpus. Dans chacune des grammaires, la relation s'établit entre un X et un Y. Dans certains cas, lorque que le X ou le Y est absent de la relation intra-phrastique, c'est le titre qui s'avère jouer, la plupart du temps, le rôle du X ou du Y.

4.5.0.1. La relation de méronymie

est transitive et anti-symétrique est repérée dans notre grammaire à l'aide de règles du type « X est une partie de Y » ou « Y est composé de X ». Le X n'est donc pas forcément le premier élément de l'expression repérée. La grammaire comporte 68 types de règles, chacune étant divisée en plusieurs configurations (présence du verbe être ou de virgules, absence explicite du X, ...). Dans le corpus, 18 cas de méronymie sont repérés[14]. Dans l'exemple suivant, la règle r12c s'applique à la phrase *Voie qui fait partie du domaine public* et extrait une relation de méronymie entre "voie" (X) et "partie du domaine public" (Y).

```
<lss :sem type="meronymie" id="1">
<lss :text>Voie qui fait partie du domaine public</lss :text>
<lss :value>
<regleMero>r12c</regleMero>
<lemmasCourtX>voie</lemmasCourtX>
<lemmasLongX>voie</lemmasLongX>
<snX>Voie</snX>
<lemmasCourtY>domaine</lemmasCourtY>
<lemmasLongY>partie domaine public</lemmasLongY>
<snY>partie du domaine public</snY>
</lss :value>
</lss :sem>
```

Figure 4. *Exemple de relation de méronymie*

4.5.0.2. La relation d'hyperonymie

est une relation transitive, anti-symétrique et hiérarchique du type « X est un Y », « X est une sorte de Y ». Dans l'exemple suivant, la règle r20a trouve une relation d'hyperonymie entre *autoroute* et *route*.

```
<lss :sem type="hyperonymieInSentence" id="1">
<lss :text>Les autoroutes sont des routes</lss :text>
<lss :value>
<regleHyperoS>r20a</regleHyperoS>
<lemmasCourtX>autoroute</lemmasCourtX>
<lemmasLongX>autoroute</lemmasLongX>
<snX>Les autoroutes</snX>
<lemmasCourtY>route</lemmasCourtY>
<lemmasLongY>route</lemmasLongY>
<snY>routes</snY>
</lss :value>
</lss :sem>
```

Figure 5. *Exemple de relation d'hyperonymie*

La grammaire dédiée à la relation d'hyperonymie comporte 20 règles. Dans notre corpus, 48 relations d'hyperonymie sont instanciées (seulement 2 dans BDTopo). Dans la

14. Il y a une différence entre le nombre de règles et le nombre d'instances : un certain nombre de règles/ patrons ne s'instancient pas sur ce corpus, les patrons ayant été développés et évalués sur des textes diversifiés et pas uniquement sur les textes de spécification de bases de données. L'intérêt de cette méthode est de se donner les moyens de traiter efficacement de nouveaux textes.

littérature, cette relation est souvent citée comme étant la plus riche [AGU 09] alors que dnas notre corpus, il semble que ce soit le contraire. Cette différence est sans doute liée au type de corpus qui, dans notre cas, est très spécialisé.

4.5.0.3. La relation d'artefact

correspond à une relation du type « X est représenté par » ou « X est employé pour ». C'est une relation symétrique. Ce type de relation se caractérise par le fait que deux concepts sont mis en relation de quasi-synonymie dans un domaine spécifique, dans notre cas, les bases de données de l'IGN. Dans le monde *réel* (*i.e.* général), ces concepts ne seraient en aucun cas considérés comme des synonymes. Cette relation est à mettre en parallèle avec la distinction entre monde réel et monde carto proposé par le Cogit[15] : un artefact est utilisé pour désigner de manière plus ou moins détournée des objets ou des ensembles d'objets du monde réel [KAS 09]. Dans l'exemple ci-dessous, le concept X n'est pas explicite dans la phrase mais doit être repris du titre, d'où la valeur « fromTitre » de « lemmasCourtX », « lemmasLongX » et « snX ».

```
<lss :sem type="artefact" id="5">
<lss :text>représentant un danger potentiel</lss :text>
<lss :value>
<regleArtef>r4aSansX</regleArtef>
<lemmasCourtX>fromTitre</lemmasCourtX>
<lemmasLongX>fromTitre</lemmasLongX>
<snX>fromTitre</snX>
<lemmasCourtY>danger</lemmasCourtY>
<lemmasLongY>danger potentiel</lemmasLongY>
<snY>un danger potentiel</snY>
</lss :value>
</lss :sem>
```

Figure 6. *Exemple de relation d'artefact*

Cette grammaire est constituée de 6 règles principales, chacune étant divisée en plusieurs configurations possibles (présence du verbe être ou de virgules, absence explicite du X dans la zone de définition, ...). Ce module repère 14 relations d'artefact dans notre corpus.

4.5.0.4. La relation de fonction

indique à quoi un objet sert. C'est une relation anti-symétrique qui introduit un Y composé soit d'un SN seul (*le transport des marchandises*) ou un syntagme verbal qui pourrait être nominalisé (*transporter les marchandises*, *le transport des marchandises*). Dans l'exemple qui suit, une relation fonctionnelle de type *faire passer des véhicules* est associée à *trajet de bateau*.

La grammaire consacrée aux artefacts est constituée de 6 règles principales, chacune étant divisée en plusieurs configurations (présence du verbe être ou de virgules, absence explicite du X, ...). Dans notre corpus, elle repère 82 relations de fonction.

15. Laboratoire de l'IGN, partenaire du projet Géonto.

```
<lss :sem type="fonction" id="4">
<lss :text>Trajet du bateau servant à passer des véhicules</lss :text>
<lss :value>
<regleFonct>r2b</regleFonct>
<lemmasCourtX>trajet</lemmasCourtX>
<lemmasLongX>trajet de bateau</lemmasLongX>
<snX>Trajet du bateau</snX>
<lemmasCourtY>véhicule</lemmasCourtY>
<lemmasLongY>passer véhicule</lemmasLongY>
<snY>passer véhicules</snY>
</lss :value>
</lss :sem>
```

Figure 7. *Exemple de relation de fonction*

4.6. *La prise en compte de la structure*

Ce dernier module repère deux types différents de relations nécessitant la prise en compte de la structure du document. Il est basé sur les observations faites dans [KAM 09b] concernant la notion de patron structurel. Trois types de relations sont mises en avant :

– le type « synonymie stricte » : un terme-concept est défini par un SN simple (un nom/terme seul). Par exemple : *cascade / définition : chute d'eau*.

– le type « hyperonymie » : un terme-concept est défini par un un nom/terme auquel sont adjoints des adjectifs/propriétés. Par exemple : *gorge / définition : vallée étroite et encaissée*

– le type « méronymie » : un terme-concept est défini par un sn introduit par un marqueur de méronymie. Par exemple : *tronçon de route / définition : portion de voie de communication*

Dans le module, le premier traitement concerne les SN constitués d'un marqueur de méronymie, d'holonymie ou d'hyperonymie. Dans ce cas, le titre est nécessaire pour la création de l'un des deux concepts de la relation (X ou Y selon les cas). Le second cas concerne les SN simples qui suivent immédiatement un titre. Deux possibilités sont envisagées : soit le SN est seul et nous concluons sur une relation de synonymie ; soit le SN ne constitue que le début de la définition et nous concluons à une relation d'hyperonymie. Ce module repère 406 relations. La figure 8 montre un exemple pour la relation de synonymie.

5. Résultats et discussion

L'évaluation des ontologies en termes de performance est un problème reconnu mais délicat [SCH 05]. Dans l'idéal, diposer d'un gold-standart permettrait de fournir une évaluation quantitative interressante. Ce type de processus nécessite un investissement humain coûteux essentiellement en termes d'annotation manuelle. Nous n'avons malheureusement pas pu mettre un tel système en place.

Nous avons mené une évaluation qualitative et manuelle des concepts et relations sé-

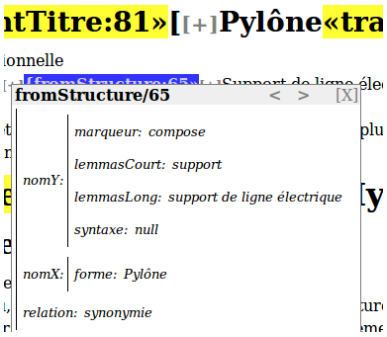

Figure 8. *Une relation de synonymie entre deux concepts extrait à l'aide de la structure et d'informations lexico-grammaticales*

mantiques associées aux concepts. Pour chaque relation annotée nous avons jugé de la pertinence de la relation elle-même ainsi que celle des concepts X et Y. Cinq valeurs sont possibles :

– la valeur est **valide** si la relation est jugée valide ainsi que les deux concepts X et Y reliés par cette relation,

– la valeur est **inverse** si la relation est valide mais est inversée (c'est le cas des relations anti-symétriques comme l'hyperonymie et la méronymie),

– la valeur est **approximative** lorsque la relation est valide mais les concepts sont approximatifs soit parce qu'ils ont été mal extraits par les programmes soit parce qu'il y a une relation de pronominalisation avec le titre et que c'est ce dernier qu'il faudrait prendre en compte,

– la valeur est **incertaine** lorsqu'une expertise particulière est requise (nécessité de recourir à des experts en cartographie),

– enfin, la valeur est **fausse** quand la relation est éronnée (dans ce cas, les concepts sont également souvent non valides).

Les résultats de cette évaluation sont indiqués dans le tableau suivant.

Relation	Artefact	Fonction	Hyperonymie	Synonymie	Méronymie	Holonymie	Total
Valide	16,7 %	14,3 %	38,8 %	25 %	19 %	58,3 %	33,1
Inverse	0 %	0 %	1 %	0 %	43 %	0 %	2,7
Approximative	16,7 %	23,8 %	12 %	25 %	19 %	8,3 %	15,4
Incertaine	66,6 %	55,5 %	19,6 %	0 %	0 %	16,7 %	22,2
Fausse	0 %	6,4 %	28,5 %	50 %	19 %	16,7 %	26,5

Ces résultats mettent en avant de grandes disparités qualitatives selon les types de relations sémantiques. Tout d'abord, les relations d'artefact et de fonction sont les relations présentant le plus haut pourcentage d'incertitude sur la pertinence de la connaissance extraite meme si d'un point de vue linguistique, elles sont tout à fait acceptables. Cette incertitude est liée à leur spécificité pour le domaine et au recours nécessaire à des experts géographes pour leur validation. La relation d'hyperonymie présente des résultats corrects

mais insuffisants. Presque 30 % des cas sont incorrects. Ces résultats vont dans le sens de ceux de [MAY 09] qui constatent, sur leurs données, une situation de surgénération de leurs patrons d'hyperonymie. Notre proposition pour le traitement de la synonymie n'est clairement pas convainquant : exploiter la structure et la syntaxe n'est pas suffisant pour distinguer l'hyperonymie de la synonymie. Enfin, les résultats concernant la méronymie et l'holonymie sont encourageants même si les patrons de méronymie doivent être revus pour ne pas générer des relations sémantiques inversant les concepts.

Une évaluation quantitative est en cours sur ces données. Nous mettons en place une procédure de comparaison des concepts existants dans le noyau ontologique avec les concepts (introduit à la section 4.4) et les relations fournis par la chaîne de traitement décrite dans cet article. Cette évaluation devrait fournir des résultats quantitatifs à même de rendre compte de l'intérêt de notre méthode d'enrichissement d'ontologie.

6. Conclusion

Cet article présente une méthodologie d'extraction de connaissances à partir de textes mise en place dans le cadre d'un projet de recherche visant l'appariement de connaissances géographiques hétérogènes. Il s'agit plus particulièrement de repérer de manière automatique des termes susceptibles d'être des concepts du domaine et des relations sémantiques particulières entre ces termes.

Dans la tradition des travaux exploitant les notions de marqueur et de patron lexico-syntaxique, notre approche utilise également la structure même des documents pour mettre en évidence des relations sémantiques entre les termes. Les résultats confirment l'idée que la structure logique permet de mettre au jour des relations d'hyperonymie entre les termes du titre et les termes contenus dans les zones de définition. Mais la situation est moins convaincante pour ce qui est de la relation de synonymie.

L'évaluation des résultats a uniquement porté sur un jugement humain de la validité des relations entre termes extraits des documents. Une seconde évaluation, en cours, va permettre de mesurer dans quelle mesure les concepts et relations extraites enrichissent le noyau d'ontologie initial. Enfin, une évaluation par les experts est nécessaire pour notamment valider le statut de concept de chacun des termes repérés.

7. Bibliographie

[AGU 09] AGUADODECEA G., ALVAREZDEMON I., MONTIEL PONSODA E., « From linguistic pattern to ontology structures », *TIA'2009*, 2009.

[AUS 08] AUSSENAC-GILLES N., JACQUES M.-P., « Designing end evaluating patterns for relation acquisition from texts with Caméléon », *Terminology*, vol. 14, n° 1, 2008, p. 45-73.

[AUS 09] AUSSENAC-GILLES N., HERNANDEZ N., « Du linguistique au conceptuel : identification de relations conceptuelles à partir de textes », *TIA*, 2009.

[BOU 03] BOURIGAULT D., AUSSENAC-GILLES N., « Construction d'ontologies à partir de textes », *Conférence TALN 2003*, 2003.

[BOU 04] BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », *Revue d'intelligence artificielle*, vol. 18, n° 1, 2004, p. 87-110.

[DES 00] DESMOULINS C., GRANDBASTIEN M., « Des ontologies pour indexer des documents texchniques pour la formation professionnelle », *IC'2000 : ingénierie des connaissances (Toulouse, 10-12 mai 2000)*, 2000.

[GRA 04] GRABAR N., HAMON T., « Les relations dans les terminologies structurées : de la théorie à la pratique », *Revue d'intelligence artificielle*, vol. 18, 2004.

[HAD 02] HADDAD M., « Extraction et impact des connaissances sur les performances des systèmes de recherche d'information », PhD thesis, Université de Grenoble, 2002.

[HEA 92] HEARST M., « Automatic Acquisition of Hyponyms From Large Text Corpora », *Proccedings of the 14th International Conference on Computational Linguistics*, 1992, p. 539-545.

[JAC 06] JACQUES M.-P., AUSSENAC-GILLES N., « Variabilité des performances des outils de TAL et genre textuel », *TAL*, vol. 47, n° 1, 2006, p. 11-32.

[KAM 09a] KAMEL M., AUSSENAC-GILLES N., « Construction automatique d'ontologies à partir de spécifications de bases de données », *Conférence IC*, 2009.

[KAM 09b] KAMEL M., AUSSENAC-GILLES N., « Utiliser la Structure du Document dans le Processus de Construction d'Ontologies (regular paper) », L'HOMME M.-C., SZULMAN S., Eds., *Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA)*, http ://www.irit.fr/, novembre 2009, IRIT, page (on line).

[KAS 09] KASSEL G., « Vers une ontologie formelle des artefacts », *20es Journées Francophones en Ingénierie des Connaissances, Hammamet, Tunisie*, 2009.

[KER 09] KERGOSIEN E., KAMEL M., SALLABERRY C., BESSAGNET M.-N., AUSSENAC-GILLES N., GAIO M., « Construction et enrichissement automatique d'ontologies à partir de ressources externes », *Conférence JFO 2009*, 2009.

[MAY 09] MAYNARD D., FUNK A., PETERS W., « Using Lexico-Syntactic Ontology Design Patterns for Ontology creation and population », EVA BLOMQVIST KURT SANDKUHL F. S. V. S., Ed., *Proceedings of the Workshop on Ontology Patterns (WOP 2009), Washington DC, USA*, vol. 516, 2009.

[MUS 09] MUSTIÈRE S., ABADIE N., AUSSENAC-GILLES N., BESSAGNET M.-N., KAMEL M., KERGOSIEN E., REYNAUD C., SAFAR B., « GéOnto : Enrichissement d'une taxonomie de concepts topographiques », *SAGEO'2009*, Marne-la-Vallée, 2009.

[ROL 06] ROLE F., ROUSSE G., « Construction incrémentale d'une ountologie par analyse du texte et de la structure du document », *Document numérique*, vol. 9, n° 1, 2006, p. 77-91.

[SCH 05] SCHUTZ A., BUITELAAR P., « RelExt : A Tool for relation Extraction from Text in Ontology Extension », *Proceedings of the 4th International Semantic Web Conference (ISWC)*, 2005.