# ” Quand rédiger c’est décrire ”: Mise en forme matérielle des textes et construction d’ontologies à partir de textes

Mouna Kamel, Mustapha Mojahid, Bernard Rothenburger

# « Quand rédiger c'est décrire » Mise en forme matérielle des textes et construction d'ontologies à partir de textes

Mouna Kamel[1], Mustapha Mojahid[1], Bernard Rothenburger[1]

[1] Institut de Recherche en Informatique de Toulouse (IRIT)-CNRS
{kamel, mojahid, rothenburger}@irit.fr

**Résumé :** La construction d'ontologie à partir de textes met classiquement en œuvre des outils issus du Traitement Automatique de la Langue et/ou des outils d'apprentissage supervisé ou non. Dans cet article nous revenons sur la possibilité d'exploiter des objets textuels à la fois facilement identifiables, souvent fertiles en connaissances ontologiques, et dont la sémantique peut clairement être explicitée par les théories du discours : les structures énumératives. Ici, nous ajoutons une nouvelle classe de relations sémantiques portée par les structures énumératives très présentes dans nos corpus : les relations lexicales telles que l'homonymie ou la synonymie. Ces relations semblent propices pour alimenter la facette terminologique d'une Ressource Termino-Ontologique. Nous montrons que ces relations peuvent être formellement caractérisées. Une évaluation de notre approche à partir d'un corpus annoté manuellement nous permet de valider notre position, ce qui constitue une première étape vers un outil d'apprentissage supervisé pour la construction d'ontologie à partir de texte.

**Mots-clés :** Construction d'ontologie à partir de texte, théories du discours, mise en forme matérielle, annotation, apprentissage supervisé.

## 1 Introduction

L'apprentissage d'ontologie à partir de texte se heurte à une difficulté bien connue : la plupart des textes introduisent des connaissances nouvelles tout en ignorant des connaissances d'arrière plan qui sont pourtant indispensables pour les interpréter (Brewster et. al, 2003). Pour affronter cette difficulté, il peut être précieux de rechercher les connaissances ontologiques dans des textes descriptifs, exhaustifs et explicites. Dit en d'autres termes, il s'agira de textes dont le but est de rendre compte de manière la plus complète et la plus claire d'une certaine réalité. Dans (Kamel & Rothenburger, 2010), (Kamel & Rothenburger, 2011), nous avons choisi de nous baser sur des textes qui possèdent les

caractéristiques énoncées ci-dessus : c'est le cas de bon nombre d'articles de l'encyclopédie en ligne Wikipédia. En particulier nous avons détaillé le parti que l'on pouvait tirer d'une propriété qui rendait ces textes particulièrement explicites : la richesse en structures textuelles énumératives. Les résultats expérimentaux sont parus encourageants montrant un taux d'élicitation (i.e. un nombre de structures ontologiques par texte) élevé.

Néanmoins, comme souvent, la diminution du silence est allée de pair avec une augmentation du bruit. Plus précisément, si les structures énumératives de ces textes sont effectivement de bons *pourvoyeurs* de structures ontologiques, elles sont aussi souvent utilisées pour décrire des phénomènes autres que la réalité ontologique d'un domaine. Nous avons identifié un certain nombre de structures énumératives comme par exemple celles ayant vocation à organiser la navigation hypertextuelle dans ces textes, celles se présentant sous forme d'une table de matière ou encore celles ayant des propriétés définitoires. En plus de ces usages il en est un qui a spécialement retenu notre attention : la mise en avant de propriétés lexicales telles que l'homonymie ou la synonymie (Hirst, 2003). Si ce nouvel usage nous éloigne de l'apprentissage de propriétés ontologiques à proprement parler, il apparaît par contre profitable de s'en servir pour la constitution ou l'enrichissement d'une Ressource Termino-Ontologique (RTO). Comment différentier la structure ontologique de la structure à visée lexicale est le sujet de cet article. Dans cette perspective nous avons essayé de mettre en évidence un certain nombre de traits discriminants pour ce nouveau type de structure énumérative.

La suite de cet article est organisée comme suit. La section suivante introduit les notions théoriques de base concernant les structures énumératives, indique des moyens de représentation de leur sémantique à l'aide des théories du discours en insistant sur les difficultés de mettre en évidence cette représentation sémantique dans le cas général. La troisième section se focalise sur les structures énumératives qui vont nous intéresser pour enrichir une RTO, soit les structures énumératives à visée ontologique ou à visée lexicale. La quatrième section caractérise ces deux types de structure énumérative ouvrant des pistes pour leur identification automatique. La dernière section évalue les premiers résultats qui ont été obtenus en utilisant ces caractérisations. Enfin, la conclusion revient sur l'intérêt de ce travail et esquisse quelques perspectives.

## 2 Analyse sémantique de la structure énumérative

### 2.1 Définition

Comme dans toute production écrite, la structure énumérative peut se présenter selon diverses formulations au sein d'un texte. Elle peut être énoncée discursivement en dehors de toute mise en forme matérielle (MFM), au sein de la même phrase ou à travers plusieurs phrases n'appartenant pas nécessairement au même paragraphe. Elle peut également être mise en évidence par l'usage de marqueurs typographiques et/ou dispositionnels. La MFM est donc un marqueur d'intentionnalité car elle manifeste la volonté de signifier conventionnellement et de manière locale dans le texte. Ces éléments typo-dispositionnels participent ainsi à la construction du sens (Virbel & Luc 2001 ; Luc *et al.* 2002). En effet, bien que les différents composants de la structure énumérative soient présentés de façon discontinue, ils constituent un tout sur le plan sémantique.

Il existe plusieurs définitions de l'énumération, celle qui nous semble le mieux prendre en compte à la fois les phénomènes architecturaux du texte et l'intention de l'auteur est celle proposée par (Virbel, 1999) : « énumérer mobilise deux actes : un acte mental d'identification des éléments d'une réalité du monde dont on vise un recensement, et où on établit une relation d'égalité d'importance par rapport au motif de recensement ; et un acte textuel qui consiste à transposer textuellement la co-énumérabilité des entités recensées, par la co-énumérabilité des segments linguistiques qui les décrivent. ». Si nous allions au delà de la relation d'égalité (égalité partielle, inégalité et asymétrie), nous pourrions couvrir la quasi-totalité des types d'énumérations en incluant les plus atypiques. Elsa Pascual (1991) a défini l'énumération de la manière suivante : « énumérer c'est attribuer un niveau égal d'importance aux entités et classer ces entités selon des critères variables ». Cette définition est certes conforme à de larges cas d'énumérations où les items sont équivalents fonctionnellement et sont réalisés à l'aide des propriétés de MFM équivalentes. Ces énumérations sont appelées parallèles (Luc *et al.* 2000). Sur le plan textuel, ce cas de structure énumérative est représenté par une structure hiérarchique. La structure énumérative est alors composée d'une amorce, d'une liste d'items et éventuellement d'une conclusion. En reprenant la terminologie de (Bush, 2003), l'amorce peut-être composée de quatre éléments dont certains sont facultatifs : l'introducteur {ci-dessous, suivant, ce qui suit…} ; l'organisateur {une liste de, les cas…} ; le classifieur, donnant le type des items {université, abbayes…} ; un segment qui « précise » la nature des items de l'énumération qu'il n'est pas possible de déduire car on ne dispose pour

cela ni d'indice sémantique ni d'indice structurel. La conclusion lorsqu'elle existe, peut avoir différents statuts comme par exemple d'illustration ou de synthèse.

En résumé, dans cette recherche, nous retenons les structures énumératives selon la définition de Pascual qui permet d'énonçer des éléments successifs d'un même champ conceptuel, et dont les éléments entretiennent un lien sémantique direct avec un concept classifieur. Ce sont les structures énumératives correspondant à cette définition que nous exploitons. Nous laissons de côté les structures énumératives dont l'énumération est non parallèle au sens de Virbel, car nous constatons que le principe de co-énumérabilité des items ne permet pas de conférer systématiquement un statut ontologique ou lexicale à la structure énumérative (exemple Figure 3).

## 2.2 Représentation sémantique des Structures Enumératives

Il s'agit alors d'identifier les liens sémantiques portés par la structure énumérative. Les approches classiques d'identification de relations sémantiques permettent uniquement de repérer les concepts et les relations lorsqu'ils sont exprimés au sein de la même phrase (Nédellec & Nazarenko, 2003), (Aussenac *et al.*, 2008), (Maedche, 2002), (Buitelaar *et al*., 2005). Mais il existe des théories du discours (ou théorie des structures rhétoriques), dont la Rhetorical Structure Theory (RST, Mann & Thompson 1998), la Segmented Discourse Representation Theory (SDRT, Asher 1993), ou la théorie de centrage (Grosz & Sidner 1986 ; Charolles 2002 ; Péry-Woodley & Scott, 2006) qui permettent d'analyser les textes « au-delà de la frontière de la phrase ». Le point de départ de ces théories réside dans le constat qu'un texte n'est pas une simple collection de phrases mais que des relations doivent exister entre les phrases d'un texte pour en assurer la cohérence (Wolf & Gibson, 2006).

L'analyse d'un discours se déroule alors en deux phases. La première étape consiste à découper le texte en segments. La seconde phase consiste à identifier la sémantique des liens ou relations du discours existant entre ces segments non forcément contigus. Les relations appartiennent à l'une des deux catégories suivantes : relations subordonnantes et relations coordonnantes. Une relation subordonnante relie un argument important (l'unité d'information la plus saillante) à un argument moins important (l'unité d'information qui supporte l'information complémentaire), tandis qu'une relation coordonnante relie des arguments de même importance. L'ensemble des relations possibles varie selon la théorie du discours : les approches réductionnistes permettent d'identifier un petit nombre de relations a priori alors que les approches multiplicatrices tendent vers un ensemble plus exhaustif et plus précis de relations. Par exemple, la relation *elaboration* donne des précisions sur un segment déjà introduit dans le discours, mais cette

relation peut être spécifiée comme une *partie d'un tout*, un *membre d'un ensemble*, une *étape d'un processus*, un *attribut d'objet*, un *détail d'une généralisation*, etc. (Lüngen *et al.*, 2008). Dans le cas général, l'identification des relations est une tâche manuelle, fastidieuse et souvent subjective car comme on l'a vu dans la définition, aucun composant d'une amorce n'est obligatoire et il arrive qu'on ne puisse disposer d'aucun indice sémantique ou structurel (Bush, 2003). En ce qui concerne l'identification automatique, la démarche adoptée est celle de l'apprentissage supervisé à partir d'un corpus annoté à la main (Carlson *et al.*, 2004) (Péry-Woodley *et al.*, 2009).

Dans le cadre de la RST, la structure rhétorique d'un discours est représentée par un arbre. La Figure 1 décrit la structure rhétorique correspondant à la structure énumérative (ayant les propriétés que nous avons retenues) selon la RST. Les items sont reliés par une relation de coordination, et chaque item est relié à l'amorce par une relation de subordination.

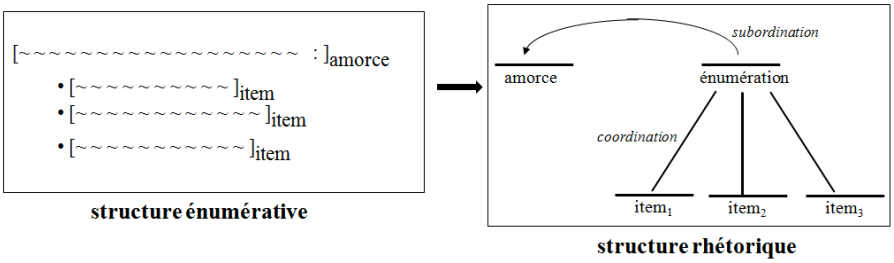


FIGURE 1 – *Structure rhétorique d'une structure énumérative.*

L'arbre issu de l'analyse rhétorique révèle souvent l'existence de connaissances ontologiques, mais parfois aussi de connaissances linguistiques, au sein des structures énumératives. Lorsqu'il s'agit de connaissances ontologiques, la correspondance entre la structure rhétorique et la structure ontologique est alors assez immédiate (Figure 2). La relation ontologique correspond à la relation du discours identifiée (généralement une des variantes de la relation elaboration décrites ci-dessus), et les concepts ou instances sont présents dans l'amorce et les items (le processus d'identification des concepts et instances fait l'objet de travaux complémentaires non discutés dans cet article).

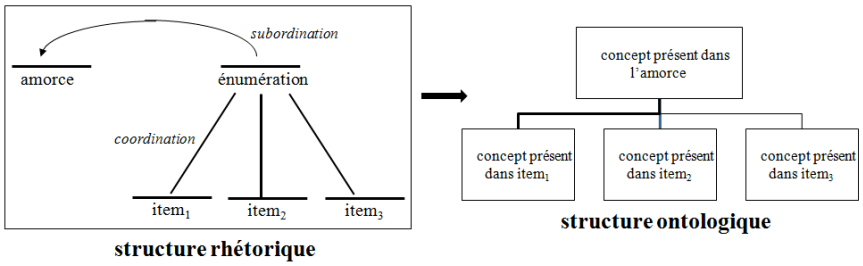


FIGURE 2 – *Structure ontologique issue d'une structure rhétorique.*

Mais il existe des cas où la correspondance structure rhétorique / structure ontologique n'est pas aussi immédiate.

## 2.3 Non équivalence entre MFM, structure rhétorique et structure énumérative

La mise en forme matérielle de la structure énumérative utilise un ensemble de caractères typo-dispositionnels et morpho-syntaxiques, tels que la ponctuation ':' en fin d'amorce, la puce, le tiret, la numérotation en début d'items, etc. Certaines structures textuelles présentent ces caractères typo-dispositionnels, sans toutefois pouvoir prétendre au statut de structure énumérative correspondant à nos critères. C'est le cas des structures textuelles présentées à la figure 3 : l'exemple de gauche comporte une dépendance syntaxique (et rhétorique) entre l'amorce (structure incomplète) et le premier item qui lui même est en relation avec le second. La deuxième énumération (de droite), malgré son caractère parallèle et sa structure complète de l'amorce, contient un classifieur (*site*) qui n'est pas repris par les items. L'analyse rhétorique de cette énumération montre que le premier item met le focus sur la « vitesse du vent » et le second sur l'aspect « propice de l'installation ».

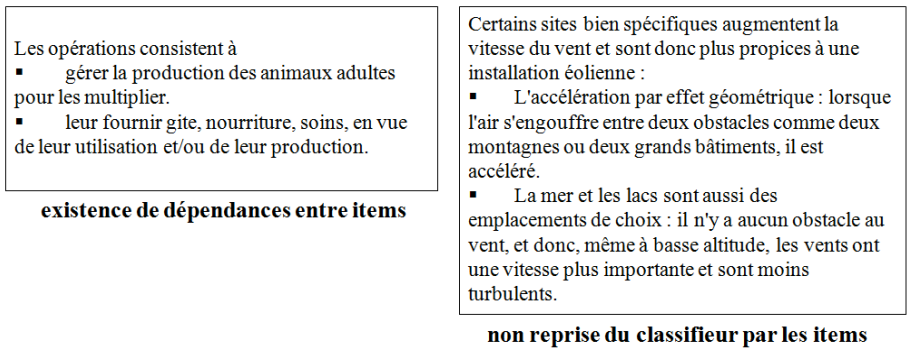


FIGURE 3 – *Exemples de structures énumératives problématiques.*

Ces observations montrent qu'il n'y a pas d'équivalence systématique entre les structures syntaxique et rhétorique et la structure matérielle d'une structure énumérative. La mise en forme matérielle fournit des indices pertinents pour repérer automatiquement en corpus des structures énumératives, mais un niveau d'analyse supplémentaire est requis pour distinguer les différentes catégories de structures énumératives.

Par ailleurs, la structure hiérarchique qui émane pour un ontologue de la structure rhétorique peut être porteuse d'une relation lexicale catégorisée comme non hiérarchique par les terminologues. Nous observons dans ce cas que la structure traduit des connaissances linguistiques et non ontologiques. Il est à noter que la notion de hiérarchie est présente à deux niveaux : au niveau rhétorique par la

relation de subordination entre l'amorce et chacun des items, et au niveau sémantique par une relation qui met en correspondance les concepts présents dans l'amorce et les items.

Lors de récents travaux (Kamel & Rothenburger, 2011), nous avons caractérisé un ensemble de traits qui nous ont permis d'identifier (1) les structures textuelles bénéficiant d'une mise en forme matérielle conforme à celle des structures énumératives, (2) celles qui sont conformes à notre définition, (3) les relations portées par ces structures hiérarchiques. Nous avons pu observer lors d'une analyse de corpus que de nouveaux indices étaient nécessaires pour pouvoir distinguer les structures ontologiques des structures dites linguistiques. C'est cet aspect qui est discuté dans les sections suivantes.

## 3 Typologie des relations sémantiques

La représentation des connaissances nécessite d'organiser les connaissances par le biais de relations sémantiques. Un concept peut alors être défini soit par les relations hiérarchiques qui le relient aux autres concepts d'une conceptualisation de domaine, soit par des liens non hiérarchiques qui expriment des rapports de sens entre les mots.

Les relations hiérarchiques sont des relations de catégorisation, qui ont pour statut de décrire et structurer les connaissances du monde réel (Rastier, 2004). La relation d'inclusion ou relation *d'hyperonymie* et la relation de *méronymie* sont des relations de catégorisation. Dans le monde des ontologies, ces relations sont respectivement appelées *is-a* et *part-of*. Ces relations participent à la structuration des terminologies et des ontologies.

Une relation lexicale non hiérarchique reflète une association entre un concept et le ou les termes qui dénotent ce concept. Ces relations sont des relations linguistiques, les plus courantes étant les relations de *synonymie*, *d'antonymie* et d'*homonymie*. Ces relations participent alors à la seule structuration des terminologies (Kister *et al*., 2011).

Les relations liant les concepts par des relations autres que lexicales, comme les relations d'*actance* (Rastier, 2004), participent également à la construction des ontologies.

Enfin, la relation *instance-of* est une relation un peu différente. Elle désigne une relation sémantique entre un concept et une instance de ce concept, une instance étant un membre de l'ensemble des individus dénoté par le concept. Elle pourrait être considérée comme une sorte de relation d'hyperonymie, mais elle est souvent différentiée dans la littérature (Hjørland, 2007). Cette relation contribue notamment à la construction de la composante factuelle des ontologies (ABox en Logique de Description).

En résumé, les relations sémantiques peuvent être catégorisées de différentes façons, selon que l'on se place du point de vue de l'ontologue ou du terminologue. La figure 4 propose une classification des relations sémantiques.

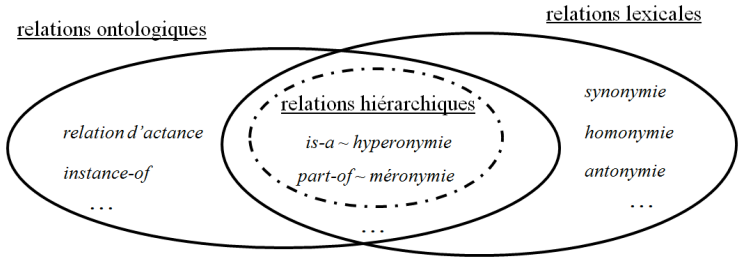


FIGURE 4 – *Classification des relations sémantiques*

La figure 5 décrit deux structures énumératives. La première matérialise des connaissances d'ordre ontologique, en établissant un lien de type *is-a* entre le concept *'force'* et les concepts *'pression hydrostatique'* et *'sous-pression'*. La seconde exprime des connaissances de nature linguistique, à travers la relation d'*homonymie*, en énonçant différents sens du terme *'arête'*.

| | |
|---|---|
| Un barrage est soumis à plusieurs forces qui sont :<br>la pression hydrostatique, exercée par l'eau sur son parement exposé à la retenue d'eau ;<br>les sous-pressions (poussée d'Archimède), exercées par l'eau percolant dans le corps du barrage ou la fondation. | Une arête est un nom commun féminin qui peut désigner :<br>l'arête, 'barbe de l'épi de graminées' (notion de botanique) ;<br>l'arête, 'partie du squelette d'un poisson' (notion d'ichtyologie) ;<br>l'arête, 'ligne d'intersection de deux plans' (notion de géométrie dans l'espace, d'architecture, etc.). |
| Structure énumérative ontologique : relation is-a | Structure énumérative homonymique |

FIGURE 5 – *Exemples de structures énumératives contenant respectivement des connaissances ontologiques et des connaissances homonymiques*.

Nous catégorisons alors les structures énumératives en fonction de la nature de la relation qu'elles portent : nous distinguons les structures énumératives ontologiques porteuses de relations hiérarchiques, des structures énumératives linguistiques porteuses de relations linguistiques.

## 4 Structure ontologique vs. structure linguistique

La typologie des structures ontologiques issues des structures énumératives que nous avons présentée dans (Kamel & Rothenburger, 2011) est basée sur un ensemble de critères typo-dispositionnels,

syntaxiques et linguistiques. Mais cet ensemble d'indices n'est plus suffisant dans la mesure où, comme le montre la figure 5, les deux structures énumératives décrites sont de nature différente, tout en étant visuellement et fonctionnellement équivalentes.

Ce travail a pour objectif de caractériser de nouveaux indices permettant de distinguer, dans un certain nombre de cas, les structures énumératives ontologiques des structures énumératives linguistiques. Pour ce qui concerne les structures linguistiques, ce travail s'intéresse plus particulièrement aux structures énumératives porteuses de la relation d'*homonymie*, qui permet d'associer différents sens à un concept. L'expression de l'homonymie est donc, par essence, un acte d'énumération pour lequel l'amorce contient le terme à définir, et chaque item contient une définition. La structure énumérative homonymique est fréquente dans les textes de type encyclopédique.

Avant de décrire de nouveaux indices permettant de caractériser les structures énumératives homonymiques, nous allons très brièvement rappeler dans la section suivante, les critères définis dans (Kamel & Rothenburger, 2011) pour identifier les relations ontologiques.

### 4.1 Identification des relations ontologiques

L'identification des relations portées par les structures énumératives paradigmatiques est basée sur les principes suivants :

- Si les items contiennent des entités nommées, alors la relation portée par la structure énumérative est la relation *instance-of* entre le concept présent dans l'amorce et les entités nommées. L'individualité est généralement exprimée à travers un nom propre (Fourour & Morin, 2003).
- Si l'amorce est syntaxiquement incomplète (le ou les constituants manquants sont fournis par les items) et se termine par un groupe verbal à la forme active (structure énumérative ontologique de la figure 3), la classe sémantique à laquelle appartient ce verbe reflète la nature de la relation.
- Si l'amorce est syntaxiquement complète et contient un marqueur linguistique qui peut être un numéral (qui en général annonce le nombre d'items présents dans la structure énumérative), un organisateur ('sorte de', 'type de', 'catégories', etc.) ou un introducteur ('suivant', 'ci-après', etc.), alors la relation est implicite et est de type *est-un*.
- Si l'amorce est syntaxiquement correcte et ne contient aucun des marqueurs linguistiques cités au paragraphe précédent, la relation est implicite est également de type *est-un*.

### 4.2 Identification de la relation d'homonymie

La relation d'homonymie est la relation entre des mots d'une langue qui ont la même forme orale et écrite mais des sens différents, selon des étymologies différentes. Elle permet donc d'associer différents sens à un même terme. Cette relation peut être identifiée par la présence des indices linguistiques suivants au sein de la structure énumérative.

**a)** Il y a reprise d'un même terme dans chaque item, voire dans l'amorce. Comme le montre la figure 6, chaque item fournit une nouvelle définition du mot 'anse.'

Science et technique
- Anse est une poignée pour porter des objets (arrosoir, cratère de Derveni, ...).
- Anse est une petite baie.
- Anse est un terme mathématique utilisé en topologie.

FIGURE 6 – *Récurrence d'un terme au niveau des items.*

**b)** Les différents sens associés à un terme sont souvent attestés dans un contexte donné. Ce contexte, identifiable par des marqueurs linguistiques, constitue un indice. La figure 7 illustre ce cas de figure. Le terme 'balise' est défini dans différents contextes, à savoir les domaines 'ferroviaire', 'maritime' et 'routier'.

Transport
- Dans le domaine ferroviaire, une balise est un module placé dans ou le long de la voie émettant des informations concernant la signalisation destinées à un équipement à bord du convoi,
- Dans le domaine maritime, une balise est un élément de la signalisation qui permet de faciliter la navigation,
- Dans le domaine routier, une balise est un dispositif implanté pour guider les usagers ou leur signaler un risque particulier.

FIGURE 7 – *Chaque sens est défini dans un contexte*.

**c)** Les verbes de référence tels que *désigner*, *référer*, *caractériser*, *dénoter*, *dénommer*, *représenter*, etc. sont souvent utilisés pour exprimer un rapport de sens entre deux unités linguistiques. La présence d'un de ces verbes en fin d'amorce constitue également un indice. Un exemple est donné par la figure 8.

Le mot bureau peut désigner :
- un meuble semblable à une table
- un lieu de travail
- une instance exécutive d'une association, d'une assemblée parlementaire ou d'autres organismes.

FIGURE 8 – *L'amorce se termine par un verbe appartenant à la classe des verbes de référence.*

### 4.3 Limites de l'approche

La présence de ces indices ne définit pas systématiquement le type de structure énumérative. Prenons le cas de la structure décrite dans la figure 9. L'amorce et chaque item contiennent bien le terme 'base', mais les items ne fournissent pas des définitions de ce terme. Par ailleurs, la caractérisation de l'inclusion lexicale entre 'base' et 'base de données', 'base' et 'base de faits', etc. ne doit cependant pas mener à identifier la relation d'hyperonymie entre ces concepts. Nous qualifions alors ce cas de « fausse homonymie ».

Base en informatique
- Une base de données est un système de stockage ordonné d'informations, généralement géré par ordinateur et exploité à l'aide du langage de requêtes SQL.
- Une base de faits est la mémoire dynamique d'un système expert, généralement organisée de manière structurée telle qu'une base de données.
- Une base de connaissances est le cœur d'un système expert contenant les connaissances d'une application experte, généralement exploitées à l'aide d'un moteur d'inférence.

FIGURE 9 – *Exemple de « fausse homonymie ».*

## 5 Evaluation du corpus

Cette section montre que l'analyse des structures énumératives dans un corpus de textes appartenant au genre encyclopédique peut contribuer à améliorer les techniques de construction d'ontologie à partir de ces textes. Notre évaluation est basée sur la validation des indices servant à caractériser les structures ontologiques et les structures linguistiques. Nous avons opté dans un premier temps pour une évaluation manuelle. Nous avons construit un corpus, défini un guide d'annotation, puis deux annotateurs ont effectué l'annotation du corpus selon ce guide d'annotation. Les annotations ont été ensuite évaluées, les indices précisés ou validés. C'est cette évaluation que nous présentons dans la suite de cette section.

### 5.1 Corpus

L'ontologie OntoTopo a été construite dans le cadre du projet GEONTO[1]. Cette ontologie est une référence pour localiser l'information relative à l'urbanisme, l'environnement et l'organisation territoriale.

[1] http://geonto.lri.fr/

L'idée est d'enrichir l'ontologie *OntoTopo* en exploitant les pages Wikipédia correspondant aux concepts de cette ontologie. Nous avons observé que les documents Wikipédia, qui sont des documents encyclopédiques, contiennent de nombreuses définitions et propriétés exprimées par des structures énumératives. En effet, les articles sont écrits selon des guides structuraux et éditoriaux. Pour ce qui concerne les énumérations bénéficiant d'une mise en forme matérielle, le « Manual of Style »[2] recommande d'utiliser la même forme grammaticale pour tous les éléments de l'énumération. Ce manuel encourage l'écriture de structures énumératives.

Nous avons construit un corpus de 2317 structures énumératives extraites de 276 pages Wikipédia référençant les concepts de l'ontologie OntoTopo : 999 ont été annotées par deux ingénieurs de la connaissance, qui ont une bonne connaissance du traitement automatique de la langue et des ontologies formelles. L'annotation a été faite selon un guide que nous avons établi : elle a eu pour but de caractériser la relation sémantique portée par toute structure textuelle assimilée à une structure énumérative (en accord avec la définition de structure énumérative que nous avons adoptée).

## 5.2 Résultats

Une première évaluation consiste à comparer les annotations en regroupant celles de type ontologique entre concepts, celles de type *instance-de* et celles de type terminologique. 'autre' fait référence d'une part aux relations dont la classe peut être identifiée mais qui n'est pas une des classes prévues, et d'autre part aux structures énumératives porteuses de plusieurs relations. Cette dualité rend le choix plus délicat et explique le taux de désaccord important pour la catégorie 'autre'. Nous constatons cependant un taux d'accord important (74%) et une valeur du Kappa de Cohen de 0,70 (Table 1).

*TABLE 1 : annotation du corpus en 4 classes par 2 annotateurs*

| | ontologique | lexicale | instance | autre | |
|---|---|---|---|---|---|
| ontologique | 323 | 3 | 15 | 54 | 395 |
| lexicale | 9 | 83 | 16 | 11 | 119 |
| instance | 7 | 7 | 108 | 6 | 128 |
| autre | 50 | 4 | 25 | 278 | 357 |
| | 389 | 97 | 164 | 349 | |

Nombre d'annotations : 999
Nombre d'accords : 792
Taux d'accord observé : 0,79
Taux d'accord aléatoire : 0,31
Kappa de Cohen : 0,7

[2] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

Une analyse plus fine des annotations (tableau ci-dessous) permet d'observer que la relation la plus présente est *is-a*, résultat prévisible au vu du genre des textes du corpus. La relation d'homonymie est relativement bien présente aussi, car une caractéristique de Wikipédia est de prévoir des pages d'homonymie. Les relations *part-of* et autres ontologiques demeurent assez fréquentes. Globalement, nous pouvons dire que ces résultats confortent notre hypothèse de départ : la présence dans certains types de textes, non seulement d'indices de fragments ontologiques, mais aussi d'indices de relations proprement linguistiques.

*TABLE 2 : annotation du corpus en 8 classes par les 2 mêmes annotateurs.*

| | est-un | partie-de | autre-ontologique | homo-nyme | syno-nyme | autre-lexicale | instance | autres | |
|---|---|---|---|---|---|---|---|---|---|
| est-un | 185 | 5 | 30 | 1 | 0 | 1 | 8 | 35 | 265 |
| partie-de | 15 | 23 | 4 | 0 | 0 | 0 | 5 | 12 | 59 |
| autre-ontologique | 12 | 1 | 48 | 1 | 0 | 0 | 2 | 7 | 71 |
| homonyme | 2 | 0 | 0 | 24 | 0 | 6 | 2 | 5 | 39 |
| synonyme | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| autre-lexicale | 6 | 0 | 1 | 10 | 0 | 42 | 14 | 6 | 79 |
| instance | 3 | 3 | 1 | 0 | 0 | 7 | 108 | 6 | 128 |
| autres | 18 | 5 | 27 | 1 | 0 | 3 | 25 | 278 | 357 |
| | 241 | 37 | 111 | 37 | 1 | 59 | 164 | 349 | |

Nombre d'annotations : 999
Nombre d'accords : 709
Taux d'accord observé : 0,71
Taux d'accord aléatoire : 0,226
Kappa de Cohen : 0,625

## 6 Conclusion

Dans cet article nous avons argumenté en faveur de l'utilisation d'objets textuels particuliers (les Structures Enumératives) pour la construction de Ressources Termino-Ontologiques. En effet, il est apparu qu'un certain nombre de textes de type descriptif tels que les textes encyclopédiques sont riches non seulement en structures énumératives à visée ontologique (décrivant une certaine réalité du monde) mais aussi en structures énumératives à visée lexicale (rendant compte de propriétés de la langue). Nous avons proposé des critères qui soient discriminants entre ces deux types de structures énumératives, mais qui permettent également l'élimination d'autres catégories de structures énumératives comme celles à visée navigationnelle ou permettant un développement

argumentatif. Notons que les critères retenus ont le statut d'indice indiquant une forte probabilité d'appartenance à l'une ou l'autre des catégories de structure énumérative.

Nous avons mis en œuvre ces critères sur un corpus de 999 structures énumératives. Le but a été d'annoter ces structures énumératives selon leur appartenance à l'une ou l'autre de six classes de relations ontologiques ou lexicales. L'évaluation des résultats obtenus a mis en évidence un bon taux d'accord entre les annotateurs.

Les perspectives de poursuite pour ce travail vont dans plusieurs directions.

D'un point de vue linguistique il conviendra d'affiner l'analyse des différents composants des structures énumératives afin de pouvoir isoler des termes dans l'amorce ou dans les items qui pourront servir d'identifiants ou de labels pour les concepts des structures ontologiques.

D'un point de vue ingénierie des connaissances il sera nécessaire de parfaire la classification des relations en analysant mieux la nature des relations non-hiérarchiques. Un autre chantier important sera l'étude des moyens d'agrégation des fragments ontologiques issus de structures énumératives avec une ontologie existante.

Enfin d'un point de vue informatique il conviendra d'assister l'ontologue avec un outil d'extraction des structures ontologiques et lexicales à partir de structures énumératives. Un premier outil basé sur une approche symbolique avait été conçu dans le cadre simplifié des seules structures énumératives à visée ontologique. Vue la nature du problème nous nous dirigeons actuellement vers une approche d'apprentissage supervisé dont le processus d'annotation manuelle décrit ci-dessus constitue une première étape. Nous attendons de cette expérimentation une évaluation en terme de précision, rappel et F-mesure.

## 7 Références

Asher, N. (1993). *Reference to abstract objects in discourse*, Dordrecht, Kluwer.

Aussenac-Gilles N., Despres S., Szulman S. (2008). The TERMINAE Method and Platform for Ontology Engineering from texts. Bridging the Gap between Text and Knowledge - Selected Contributions to *Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), IOS Press, p. 199-223.

Brewster, C., Ciravegna, F., Wilks, Y. (2003). Background and Foreground Knowledge in Dynamic Ontology Construction, in *Proceedings of the Semantic Web Workshop,* SIGIR.

Buitelaar P., Cimiano P., Magnini B. (2005). *Ontology Learning From Text: Methods, Evaluation and Applications*. IOS Press.

Bush, C. (2003). Des déclencheurs des énumérations d'entités nommées sur le Web, *Revue québécoise de linguistique*, vol. 32, n° 2, p. 47-81.

Carlson, L. Marcu, D., Okurowski. M.-E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

Charolles, M. (2002) Organisation des discours et segmentation des écrits. *Inscription Spatiale du Langage : structures et processus*, Toulouse, 29-30 janvier, Prescot, 31-39.

Fourour N., Morin E. (2003). Apport du web sémantique dans la reconnaissance des entités nommées. *Revue québécoise de linguistique*, vol. 32, n° 1, 2003, p. 41-60.

Grosz, B.J., Sidner, C.L. (1986) Attention, intentions, and the structure of discourse, *Computational Linguistics,* 12, 175-204.

Hirst, G. (2003) Ontology and the Lexicon. In Handbook on *Ontologies in Information Systems.*

Hjørland B. (2007). Semantics and knowledge organization. *ARIST* 41(1): 367-405

Kamel M., Rothenburger B. (2010). Ontology Building Using Parallel Enumerative Structure. International Conference on *Knowledge Engineering and Ontology Development (KEOD 2010)*, Valence, 25/10/2010-28/10/2010, INSTICC - Institute for Systems and Technologies of Information, Control and Communication, p. 276-281

Kamel M., Rothenburger B. (2011). Elicitation de structures hiérarchiques à partir de structures énumératives pour la construction d'ontologie. *Journées Francophones d'Ingéniérie des Connaissances (IC 2011),* Annecy du 17au 20 Mai 2011, Alain Mille (Eds.) Presse Universitaire des Antilles et de la Guyane, p. 505-522.

Kister L., Jacquey E. et Gaiffe B. (2011). Du thesaurus à l'onto-terminologie : relations sémantiques vs relations ontologiques. *CORELA* 9, 1.

Luc C. (2000). *Représentation et Composition des structures visuelles et rhétoriques du texte. Approche pour la génération de textes formatés.* Thèse de Doctorat, Université Paul Sabatier, Toulouse.

Luc, C., Mojahid, M., Péry-Woodley, M.P., Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. *CIDE'2000*, Lyon.

Luc, C., Mojahid, M., Virbel, J. (2002). Le modèle d'architecture de texte, in J. Virbel (ed.). *L'Inscription Spatiale du Langage : structures et processus*, Toulouse, IRIT-CNRS.

Luc, C., Mojahid, M., Virbel, J. (2002). L'intentionnalité communicationnelle et son impact visuel, in J. Virbel (ed*.). L'Inscription Spatiale du Langage : structures et processus, Toulouse*, IRIT-CNRS.

Lüngen, H., Bärenfänger, M., Hilbert, M., Lobin, H., Puskàs, C. (2008). Discourse relations and document structure. In Metzing, D. and Witt, A., editors, *Linguistic modeling of information and Markup Languages. Language technology*, Chapter VI, Text, Speech and Language Technology. Springer, Dordrecht

Maedche A. (2002). *Ontology learning for the Semantic Web*, vol 665. Kluwer Academic Pub.

Mann, W.C., & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8 (3). 243-281.

Nédellec C., Nazarenko A. (2003). Ontology and Information Extraction. Handbook on *Ontologies in Information Systems* (S. Staab & R. Studer eds.), Springer.

Pascual, E. (1991). *Représentation de l'architecture textuelle et génération de textes.* Thèse de l'Université Paul Sabatier. Toulouse.

Péry-Woodley, M.-P., Scott, D. (2006) Computational Approaches to Discourse and Document Processing, *T.A.L.* 47(2): 7-19.

Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L., & Widlocher (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *TALN*, Senlis, France.

Rastier F. (2004). Ontologie(s). Structuration de terminologie. J.M. Pierrel et M. Slodzian (eds.), Revue des sciences technologies de l'information, *Revue d'intelligence artificielle*, vol 18, n°1 15-40.

Virbel J., Luc C. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, XXIII, N. 1, p. 103-123.

Virbel, J. (1999). Structures textuelles, planches fascicule 1 : *Enumérations, Version 1*, Technical report, IRIT.

Wolf, F. & Gibson, E. (2006). *Coherence in Natural Language: Data Structures and Applications.* Cambridge, MA: MIT Press.