Elicitation de Structures Hiérarchiques à partir de Structures Enumératives pour la Construction d'Ontologie

Mouna Kamel, Bernard Rothenburger

Institut de Recherche en Informatique de Toulouse (IRIT) – CNRS
UPS, 118, Route de Narbonne, 31062 Toulouse Cedex, France
{kamel, rothenburger}@irit.fr

Résumé : Un texte est une suite de phrases dont la cohérence sémantique et rhétorique doit être assurée par des relations du discours. Un texte met aussi en œuvre un ensemble de moyens typographiques, de ponctuations et d'agencements qui contribuent, eux aussi, à identifier son sens. Ces deux propriétés ont été respectivement formalisées, par différentes théories du discours, et par des modèles de structure de textes. Les correspondances entre les représentations des structures du discours et les mises en forme des textes ne sont pas généralement bijectives. Néanmoins, certaines structures discursives comme les structures énumératives ont des caractéristiques de typographie, de ponctuation ou/et de disposition qui (1) les rendent facilement repérables et (2) traduisent des relations hiérarchiques qui leur confèrent le statut d'indice de fragment d'ontologie. Dans cet article nous montrons comment les objets textuels ayant les propriétés (1) et (2) peuvent être exploités pour améliorer considérablement le processus d'enrichissement d'ontologies à partir de textes.

Mots-clés : construction d'ontologie, théorie du discours, mise en forme matérielle, structure énumérative.

1. Introduction

Une démarche classique pour la construction d'ontologies à partir de texte est synthétisée par le fameux "Ontology Learning Layer Cake" (Cimiano, 2006). Dans cette approche on procède par étapes successives : d'abord on identifie les entités atomiques que sont les termes, puis pour chacun des niveaux successifs (groupe de synonymes, concepts, relations, hiérarchie de relation, schéma d'axiomes et enfin axiomes) on utilise le niveau inférieur pour obtenir les constituants du nouveau niveau. En fait, ces méthodes permettent surtout de repérer les concepts et les relations intraphrastiques. Ces résultats peuvent être obtenus en utilisant des

méthodes d'apprentissage statistiques ou des outils issus du traitement automatique de la langue (Cimiano, 2006). Nous prétendons que dans certains types de textes, il est possible d'identifier automatiquement des structures du discours allant au-delà des limites de la phrase et qui exhibent des fragments de connaissances ontologiques. Pour cela nous allons analyser la structure des textes à deux niveaux complémentaires. Le premier niveau est celui de la structure discursive des textes (Mann & Thomson, 1988), (Asher, 1993). Il s'agira de se baser sur des relations de discours ontologiques parfois appelées 'Elaboration-Derivation' (Lüngen & al. 2010) articulant des segments répartis sur plusieurs phrases. Le second niveau est celui de la structure logique ou typo-dispositionnelle des textes qui apparait fournir des indices fiables pour identifier les structures discursives hiérarchiques. En particulier, nous utiliserons les structures énumératives verticales qui sont, d'une part, facilement repérables et analysables et d'autre part très souvent porteuses d'indices de relations ontologiques.
La section 2 énonce les principes de base sur lesquels nous nous appuyons pour ce travail : les structures discursives, la mise en forme matérielle des textes et les correspondances entre ces deux notions. La section 3 décrit les structures énumératives en général et les structures énumératives paradigmatiques et verticales en particulier. La section 4 indique les moyens que nous avons mis en œuvre pour assurer la correspondance entre les structures énumératives paradigmatiques et les structures hiérarchiques à portée ontologique. La section 5 décrit une application de ces principes, les résultats obtenus et propose des perspectives envisageables. Enfin la section 6 conclut ce travail

2. Structures discursives pour l'identification des structures ontologiques

Considérons le texte **(T1)** extrait du « Wall Street Journal ». On peut trouver actuellement ce type de texte à l'adresse suivante 'http://online.wsj.com/article/BT-CO-20110106-713738.html'. Quand un lecteur est confronté à un tel texte il déduit que 'prime rate', 'federal funds', 'discount rates', 'call money' and 'commercial paper' sont des cas particuliers de 'U.S. and foreign annual interest rates'. On obtient là une structure hiérarchique qui pourrait être représentée par le fragment de la Fig. 2. Le repérage automatique d'une telle structure ontologique semble hors de portée des outils classiques de construction d'ontologie à partir de texte. En effet, les relations taxinomiques interphrastiques ne peuvent pas être identifiées par des techniques habituelles, comme les patrons lexico-syntaxiques, l'inclusion de termes, etc.

Un autre moyen d'identifier de telles relations est de segmenter le texte et de lier les différents segments par des relations du discours. Le texte **(T2)** décrit une segmentation du texte **(T1)**, conformément aux prescriptions de

Carlson et Marcu dans le cadre de la Rhetorical Structure Theory (RST) (Carlson & Marcu, 2001).

Texte T1 : extrait du Wall Street Journal

```
The key U.S. and foreign annual interest rates below are a guide to
general levels but don't always represent actual transactions.
       PRIME RATE: 10 1/2%. The base rate on corporate loans at large
U.S. money center commercial banks.
       FEDERAL FUNDS:
8 3/4% high, 8 11/16% low, 8 5/8% near closing bid, 8 11/16% offered.
Reserves traded among commercial banks for overnight use in amounts of $1
million or more.
Source: Fulton Prebon (U.S.A.) Inc.
       DISCOUNT RATE: 7%. The charge on loans to depository institutions
by the New York Federal Reserve Bank.
       CALL MONEY: 9 3/4% to 10%. The charge on loans to brokers on stock
exchange collateral.
       COMMERCIAL PAPER : placed directly by General Motors Acceptance
Corp.:
8.50% 30 to 44 days;
8.25% 45 to 65 days;
8.375% 66 to 89 days;
8% 90 to 119 days;
7.875% 120 to 149 days;
7.75% 150 to 179 days;
7.50% 180 to 270 days.
```

Texte T2 : Segmentation du texte T1 selon la RST

```
(190) [The key U.S. and foreign annual interest rates below are a guide
to general levels but don't always represent actual transactions.A]
[PRIME RATE: 10 1/2%. The base rate on corporate loans at large U.S.
money center commercial banks.B] [FEDERAL FUNDS: 8 3/4% high, 8 11/16%
low, 8 5/8% near closing bid, 8 11/16% offered. Reserves traded among
commercial banks for overnight use in amounts of $1 million or more.
Source: Fulton Prebon (U.S.A.) Inc.C] [DISCOUNT RATE: 7%. The charge on
loans to depository institutions by the New York Federal Reserve
Bank.D] [CALL MONEY: 9 3/4% to 10%. The charge on loans to brokers on
stock exchange collateral. E] [COMMERCIAL PAPER placed directly by
General Motors Acceptance Corp.: 8.50% 30 to 44 days; 8.25% 45 to 65
days; 8.375% 66 to 89 days; 8% 90 to 119 days; 7.875% 120 to 149 days;
7.75% 150 to 179 days; 7.50% 180 to 270 days.F]wsj_0602
```

Chacun des segments est annoté en l'encadrant par des crochets et en les indexant par une lettre. Le diagramme de la Fig. 1. décrit la structure rhétorique correspondant à **(T1)** dans la représentation de la RST. *Elaboration-Set-Member* et *List* sont deux relations de la RST, les lettres majuscules correspondent aux fragments.

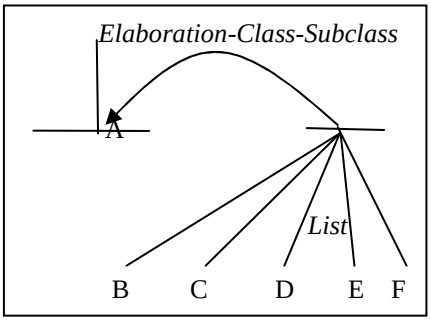

Fig. 1 – structure rhétorique du texteT1

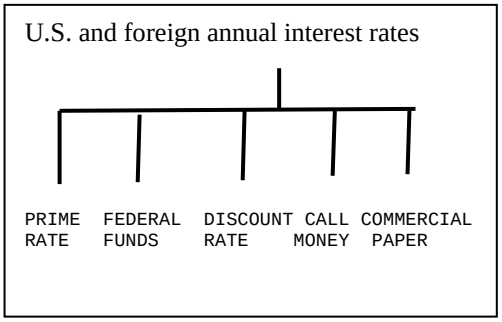

Fig. 2 – structure hiérarchique apprise à partir du texte T1

Le passage de la structure rhétorique de la Fig. 1. à la structure ontologique de la Fig. 2. est immédiate. Par contre les étapes de segmentation et de représentation en RST sont en général faites manuellement. L'idée que nous défendons dans cet article est qu'il existe, dans des cas spécifiques, une procédure qui permet de passer automatiquement du texte à la structure ontologique. Dans les deux sections suivantes nous présentons brièvement les théories du discours et montrons comment les propriétés typo-dispositionnelles de textes écrits peuvent être utilisées dans le cadre de ces théories.

2.1. Théorie des structures du discours

Plusieurs théories du discours (ou théorie des structures rhétoriques) ont été proposées pour analyser la sémantique des textes « au-delà de la frontière de la phrase ». Le point de départ 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).

D'une manière générale les théories pour l'analyse du discours visent à étudier la cohérence des discours d'un point de vue rhétorique (l'intention du rédacteur) ou sémantique (la description du monde). Pour le lecteur, ces deux types de relations sont repérés par des indices multiples : marqueurs lexicaux, syntaxiques, typographiques, ponctuationnels ou dispositionnels.

L'analyse d'un discours se déroule alors en deux phases. La première étape consiste a découper le texte en segments appelés *text span* ou parfois *Elementary Discourse Unit* (*EDU*). Ces constituants peuvent être contigües ou non. La seconde phase consiste à identifier la nature des liens ou relations du discours existant entre ces segments. Ces relations appartiennent à l'une des deux catégories suivantes : relations subordonnantes ou relations coordonnantes qui organisent le discours de manière hiérarchique (Asher & Vieu, 2005). Dans le cadre de la RST, une relation subordonnante relie un noyau à un satellite et une relation

coordonnante relie plusieurs noyaux (structure multinucléaire). Par exemple, dans la Fig. 1. *Elaboration-Set-Member* est une relation subordonnante reliant le noyau A à une structure multinucléaire satellite au sein de laquelle la relation coordonnante *List* relie B, C, D, E et F. Le choix de l'ensemble des relations possibles dans une théorie du discours est sujet à débats ; elles vont des approches réductionnistes (où peu de relations sont identifiées a priori) à des approches multiplicatrices (où l'on tend vers un ensemble exhaustif de relations) (Hovy & Maier, 1991). Dans le cas général, l'identification des relations est une tâche délicate et en partie subjective. Même lors d'une identification par des experts, l'unanimité n'est pas acquise. 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.

2.2. Mise en forme et identification de structures du discours

Les aspects visuels associés à un texte écrit contribuent aussi à la construction de son sens (Virbel & al., 2005) , (Luc, 2000). Ces aspects de mise en forme matérielle incluent des choix typographiques pour permettre de repérer des composants particuliers du texte (titres, légendes, notes, etc.) et des choix de mise en forme (indentation, saut de ligne, etc.). Ils organisent de manière immédiatement perceptible la structure logique d'un texte.

La mise en forme du texte peut aussi être utilisée pour produire la structure discursive. Pour segmenter et organiser la structure du texte (**T3**) ci-dessous, on pourra utiliser des indices linguistiques (par exemple : 'est en grande partie responsable' est un indice de relation de causalité) mais aussi des indices de mise en forme (les deux paragraphes à puce sont considérés comme des justifications de l'assertion initiale)

La structure du discours de ce texte que nous proposons est la suivante :

justify(non-volitionnal-cause(A,B),
sequence (motivation(contrast(D,E),C),
non-volitionnal-cause(F,G))

Cette structure indique que le premier paragraphe à puce donne un premier argument motivé par le contraste entre la situation en France et au Danemark ou en Allemagne et un second argument pour justifier l'assertion initiale.

La structure logique où typo-dispositionnelle du même texte est la suivante : *SE(Amorce([A,B]),enum(item([C,E,D]),item([F,G]))*
SE est une structure énumérative telle que celles décrites dans la section suivante. Elle est composée d'une amorce suivie d'un ensemble de deux items.

Dans ce cas il existe une correspondance partielle entre la structure sémantique et la structure typo-dispositionnelle. Cette correspondance n'est pas toujours assurée (voir par exemple (Power & al., 2003)). On peut aussi noter que les relations associées à cette structure sont de type argumentatif

ou rhétorique. Dans les sections suivantes nous allons décrire des structures typo-dispositionnelles particulières : les structures énumératives verticales. Contrairement à l'exemple ci-dessus, ces structures sous-tendent le plus souvent des relations du discours de type ontologique.

Texte T3 : Un exemple de texte mis en forme

Or cette surconsommation est en grande partie responsable des résistances croissantes des bactéries aux antibiotiques :

- ce sont les pays les plus grands consommateurs d'antibiotiques qui constatent aussi les plus fortes résistances des bactéries : les staphylocoques dorés sont résistants à la méthicilline dans 57% des cas en France, alors que la fréquence observée au Danemark n'est que de 1% et en Allemagne de 9%
- et chaque fois qu'une baisse sensible et durable de consommation d'antibiotiques est constatée, ces phénomènes de résistance diminuent.

Texte T4: Segmentation du texte **T3**

```
[Or cette surconsommation A][est en grande partie responsable des
résistances croissantes des bactéries aux antibiotiques B]:[ce sont les
pays les plus grands consommateurs d'antibiotiques qui constatent aussi
les plus fortes résistances des bactéries C][les staphylocoques dorés sont
résistants à la méthicilline dans 57% des cas en France D][alors que la
fréquence observée au Danemark n'est que de 1% et en Allemagne de 9% E ][
et chaque fois qu'une baisse sensible et durable de consommation
d'antibiotiques est constatée, F][ces phénomènes de résistance diminuent.
G]
```

3. Structures énumératives

L'acte d'énumération consiste à énoncer les éléments successifs d'un même champ conceptuel, ces éléments entretenant un lien hiérarchique direct ou indirect avec un concept classifieur. Sur le plan textuel, cet acte est retranscrit par une structure hiérarchique dite structure énumérative. La structure énumérative est composée d'une amorce, d'une liste d'items (constituant l'énumération) et éventuellement d'une conclusion. L'amorce contient le concept classifieur et le lien sémantique qui le relie à au moins un des items ; elle introduit aussi la liste d'items. Un item est une entité co-énumérée ayant un lien sémantique avec le concept classifieur ou un autre item. La conclusion lorsqu'elle existe synthétise les différentes propositions exprimées à travers les items.

Par ailleurs, la structure énumérative trouve divers modes de rédaction au sein d'un texte. Elle peut être énoncée au fil du texte en dehors de toute mise en forme matérielle (MFM), ou au contraire être mise en évidence par l'usage de marqueurs typographiques et/ou dispositionnels spécifiques.

Dans ce qui suit, nous proposons une typologie des structures énumératives et caractérisons celles que nous exploitons pour la construction d'ontologie.

3.1. Structure énumérative sans MFM vs Structure énumérative avec MFM

3.1.1. Structure énumérative sans MFM

Les structures énumératives sans MFM sont exprimées de façon linéaire, et sont interphrastiques. Les items sont alors introduits par des marqueurs de relation du discours, qui sont souvent des groupes adverbiaux (*premièrement*, *deuxièmement*, *troisièmement* dans la Fig. 3).

Comment faire pour économiser 68% d'électricité par rapport à une dépense habituelle? Premièrement, en éteignant la lumière dès votre sortie d'une pièce. Cela peut paraître banal, mais ça ne l'est absolument pas. Deuxièmement, éviter les lampes halogènes car une lampe allogène de 500 watts consomme l'équivalent de 23 lampes. Troisièmement, essayez de remplacer les lampes traditionnelles par des lampes basse consommation. Elles coûtent plus cher, mais permettent d'économiser 53KWH par an et durent environ 4 fois plus longtemps.

Fig. 3. : Enumération sans MFM (http://www2c.ac-lille.fr/pneruda-wattrelos/idd0405bis/planetesoli/solutions/energies.htm

Comme le montre cet exemple, les items peuvent être des éléments discursifs relativement complexes, et un marqueur de discours peut porter sur plusieurs phrases, ce qui suppose de résoudre des phénomènes linguistiques complexes tels que les anaphores, les ellipses, etc. L'identification de telles structures énumératives a fait l'objet de nombreux travaux (Virbel et al., 2005), (Ho-Dac et al., 2010), notamment à travers l'étude de la relation d'Elaboration (Bras et al., 2008).

3.1.2. Structure énumérative avec MFM

Lorsque l'auteur souhaite mieux mettre en évidence une structure énumérative dans un document, il use de moyens de mise en forme (ponctuation, caractères typo-dispositionnels) pour organiser, subdiviser et hiérarchiser les différents composants. Ces structures énumératives ont alors la capacité d'être perçues avant d'être interprétées (Ho-Dac et al., 2010).

Une structure énumérative mise en forme uniquement à l'aide de caractères ponctuationnels est dite horizontale. Un exemple est donné à la Fig. 4.

Selon les définitions de l'IAU, il y a huit planètes dans le système solaire (Mercure, Venus, Terre, Mars, Jupiter, Saturne, Uranus et Neptune).

Fig. 4. : Un exemple de Structure Enumérative horizontale avec MFM

Les différents composants sont marqués par des caractères spécifiques : la liste des items est délimitée par des parenthèses, les items ("*Mercure*", "*Venus*", "*Mars*", etc.) sont séparés par une virgule ou par le marqueur lexical "*et*", l'amorce ("*huit planètes dans le système solaire*") est l'unité

textuelle qui précède immédiatement la liste d'items. La relation, dans ce cas implicite, est de nature taxinomique.

Les structures énumératives sont encore plus saillantes lorsque leurs différents composants sont présentés de façon discontinue, selon des critères typo-dispositionnels. Leurs éléments constituent cependant un tout sur le plan sémantique. Ces structures énumératives sont dites verticales. Un exemple est donné à la Fig. 5.

Les formes de communication non parlées sont :
- *le langage écrit*
- *le langage des signes*
- *le langage sifflé*

Fig. 5. : Un exemple de structure énumérative verticale

3.2. Structure énumérative syntagmatique vs Structure énumérative paradigmatique

Des études ont montré que les auteurs sont enclins à placer l'information la plus importante au début des unités textuelles (Ho-Dac, 2007). Nous avons pu attester ce phénomène sur les items de structures énumératives appartenant à des corpus différents. Partant de ce constat, nous considérons dans ce travail que les composants cibles de l'énumération sont situés en début d'item.

3.2.1. Structure énumérative syntagmatique

Les items peuvent entretenir des relations de dépendance (syntaxiques ou rhétoriques) entre eux, exprimées généralement par des marqueurs de relation du discours. Nous qualifions alors la structure énumérative de syntagmatique.

Le Lindy Hop est :
- une danse swing
- dont la naissance remonte aux années 20

Fig. 6. : Un exemple de structure énumérative syntagmatique (Luc, 2000)

L'exemple décrit ci-dessus montre une structure énumérative dans laquelle le second item est subordonné au premier (les deux items ne peuvent être inversés). Par ailleurs, cette structure exprime deux relations sémantiques : la relation *est-un* entre l'amorce et le premier item, et une relation de type *date* entre les deux items.

3.2.2. Structure énumérative paradigmatique

Lorsqu'il n'existe pas de relation de dépendance entre les items, et que les têtes d'items sont syntaxiquement équivalentes, les items sont considérés comme fonctionnellement équivalents. Le même lien sémantique relie alors l'amorce à chaque item. Dans ce cas, nous qualifions la structure énumérative de paradigmatique. Les structures énumératives décrites dans les figures 4, 5 sont paradigmatiques.

3.3. Construction d'ontologie vs Peuplement d'ontologie

Les structures énumératives paradigmatiques expriment un lien hiérarchique entre l'amorce et chacun des items. L'élément père contenu dans l'amorce correspond généralement à un concept, les éléments fils contenus dans les items peuvent correspondre à des concepts ou à des instances.

La démarche consistant à identifier la nature exacte des éléments (concept ou instance) et leur intégration dans l'ontologie sort du cadre de cette étude. Notre objectif est de proposer une méthode pour formaliser, sous forme de structure hiérarchique, les connaissances ontologiques contenues dans une structure énumérative paradigmatique. Cette structure hiérarchique contribuera à la construction d'ontologie ou au peuplement d'ontologie selon que les éléments fils auront été expertisés comme concepts ou comme instances.

3.4. Quelles structures énumératives pour la construction / le peuplement d'ontologie ?

Les structures énumératives que nous ciblons pour la construction ou le peuplement d'ontologie sont celles qui sont paradigmatiques avec MFM, structures que nous désignerons dans la suite par SEP. Ces structures ont l'avantage (1) d'être facilement identifiables grâce aux marqueurs ponctuationnelles ou typo-dispositionnelles, (2) de pouvoir être représentées sous forme de structures hiérarchiques, (3) de permettre une analyse automatique du discours, analyse basée à la fois sur les propriétés sémantiques de ces marqueurs et sur les propriétés syntaxiques de l'amorce et des items, et (4) d'être fréquentes dans les textes scientifiques, procéduraux ou encyclopédiques. Nous décrivons dans la section suivante le processus de traduction d'une SEP en Structure Hiérarchique.

4. Elicitation d'une SEP en Structure Hiérarchique

Lorsque la structure énumérative a été identifiée comme paradigmatique, les liens sémantiques existant entre l'élément père présent dans l'amorce et les éléments fils présents dans les items sont représentés par une structure hiérarchique. La Fig.7 donne un exemple de cette correspondance.

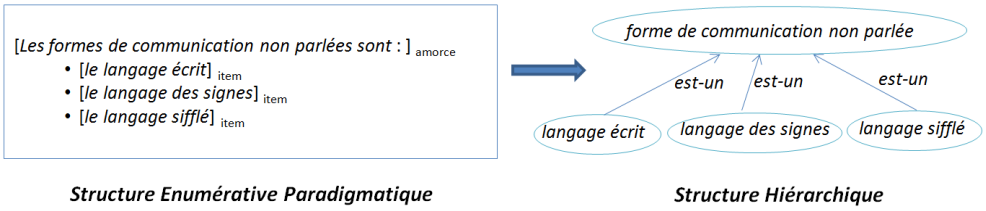

Fig.7. : Elicitation d'une SEP en Structure Hiérarchique

Le processus décrit se fait en trois étapes : (1) identification de la structure énumérative avec MFM, (2) identification de la propriété paradigmatique, (3) identification des éléments père et fils présents dans la SEP, et de la relation sémantique qui les lie.

4.1. Identification des structures énumératives avec MFM

Nous avons défini un ensemble de patrons lexico-syntaxiques permettant d'identifier les composants des structures énumératives avec MFM grâce aux annotations morpho-syntaxiques produites par un Tokenizer et un Segmenteur en phrases. Nous donnons deux exemples de patrons.

```
(P1) : ({Token.string== "-"} (({Token})* ):a {SL}) -->
                                        :a.Item={ kind = "Item", rule = "RItem"}
(P2) : ({Sentence,Token} ({Token})* {Token.string==":"} {Item} ):a -->
                              :a.AmorcePotentielle={ kind = "Amorce", rule = "RAmorce"}
```

Ces patrons sont écrits en JAPE[1]. La partie gauche de la règle correspond à un schéma d'annotation, la partie droite à des instructions de manipulation de ces annotations. Le patron (P1) annote *Item* toute la chaine lexicale comprise entre un tiret et un saut de ligne (préalablement annoté SL). Le patron (P2) annote *Amorce* toute chaîne lexicale correspondant à une phrase contenant le caractère deux-points suivi d'un item.

4.2. Identification des SEP

Il s'agit de déterminer dans cette étape si la structure énumérative avec MFM (annotée à l'étape précédente) est paradigmatique, c'est-à-dire si les items sont fonctionnellement équivalents. Les têtes d'items sont analysées (1) pour déceler la présence éventuelle de liens de subordination (exprimés à l'aide de marqueurs de relation du discours) et (2) pour déterminer si les têtes d'items ont même structure syntaxique. Un lexique répertoriant les marqueurs de relation du discours (Hovy & Maier, 1991) et un analyseur syntaxique sont utilisés.

[1] JAPE : Java Annotation Patterns Engine

4.3. Identification des connaissances ontologiques

L'identification des connaissances ontologiques est essentiellement basée sur l'analyse syntaxique de l'amorce et des items, et sur l'utilisation de lexiques.

4.3.1. Identification de l'élément père et de la relation sémantique : analyse de l'amorce

La structure syntaxique de l'amorce permet d'identifier dans certains cas l'élément père et dans tous les cas la relation, qui est soit dénotée explicitement par un verbe, soit correspond à la relation *est-un.* Trois types d'amorce ont été caractérisées :

Type 1 : l'amorce est une proposition syntaxiquement non correcte. Il y a alors deux possibilités :

- l'amorce est composée d'un syntagme nominal (Fig. 8.a). Ce syntagme correspond au terme associé à l'élément père, la relation sémantique implicite est la relation *est-un*.
- l'amorce se termine par un groupe verbal à la forme active (Fig. 5). Dans ce cas, le ou les constituants manquants sont fournis par les items. La classe sémantique à laquelle appartient ce verbe reflète la nature de la relation. Le terme associé à l'élément père est le terme contenant l'unité lexicale désignée comme sujet de ce verbe.

Type 2 : L'amorce est complète (Fig. 8.b). Elle est syntaxiquement correcte et contient des indices linguistiques tels qu'un numéral qui annonce le nombre d'items, ou un des termes répertoriés dans un lexique ("*suivant*", "*sorte*", "*type*", etc.) L'élément père est le terme qui co-occure avec ce marqueur, et la relation à nouveau implicite est de type *est-un*.

Anomalies chromosomiques : - délétion - insertion - duplication	En France , on distingue deux catégories de fouilles archéologiques : - la fouille de sauvetage ou fouille préventive , - la fouille programmée
8.a	**8.b**
Un aérodrome comprend éventuellement des bâtiments, des installations et des matériels : - les balises - l' aire de trafic (de stationnement) - l' aire à signaux - …	
8.c	

Fig.8. : Exemples de SEP issues de Wikipédia

Type

3 : L'amorce est syntaxiquement correcte et non complète (Fig. 8.c). L'élément père peut alors être désigné par le sujet ou l'objet de la proposition principale. La décision est généralement basée sur des connaissances d'arrière plan. La relation est aussi implicite, de type *est-un*.
Une analyse détaillée des propriétés syntaxiques des amorces est présentée dans (Kamel & Rothenburger, 2010). Nous ne traitons actuellement que les SEP ayant des amorces de type 1 et 2.

4.3.2. Identification des éléments fils : analyse des items

Le constat selon lequel les auteurs placent l'information la plus importante en début d'unités textuelles nous amène à considérer que l'élément fils à relier à l'élément père est présent en tête de syntagme. Lorsque l'item commence par un syntagme nominal, c'est ce syntagme nominal qui est associé au concept fils ou à l'instance.

Ces traitements ont été réalisés à l'aide de la plate-forme GATE[2] dont le principe est d'appliquer successivement sous forme de pipeline des ressources linguistiques et/ou des ressources de traitement sur un corpus. Le résultat est un corpus annoté, et ces annotations peuvent faire l'objet de divers traitements à l'aide de règles écrites en JAPE ou en Java. Comme GATE dispose d'une Ontology API, des fragments d'ontologie peuvent être construits à partir de l'exploitation de ces annotations.

5. Application, résultats et perspectives

Nous avons appliqué les outils que nous venons de décrire dans le cadre d'une application d'enrichissement d'ontologie.

5.1. Application

L'ontologie OntoTopo concernant la localisation de l'information relative aux problématiques d'aménagement, d'environnement ou d'urbanisme a été construite dans le cadre du projet GEONTO[3]. Par ailleurs, nous avons observé que les documents Wikipédia (qui sont des documents encyclopédiques) contiennent beaucoup de définitions et de propriétés exprimées sous forme de SEP.

Différents travaux exploitent les documents Wikipedia pour l'enrichissement d'ontologie, en utilisant soit le champ catégorie pour

---

[2] General Architecture for Text Engineering : plate-forme d'ingénierie linguistique développée à l'Université de Sheffield (http://gate.ac.uk)

[3] Projet ANR-07-MDCO-005, http://www.lri.fr/geonto : collaboration entre le COGIT, le LRI (Université paris Sud), le LIUPPA (Université de Pau) et l'IRIT (Université de Toulouse)

étendre les taxonomies (Chernov et al., 2006), soit les infoboxes pour peupler la base DBpedia (Auer et al., 2007). Notre approche est différente dans la mesure où nous tirons profit de la structure textuelle du document. De plus, les articles Wikipédia sont rédigés selon le guide "the Manual of Style" (http://en.wikipedia.org/wiki/Wikipedia:Manual of Style) qui préconise l'utilisation de SEP et recommande pour ces structures d'utiliser la même forme grammaticale pour tous les items.

5.2. Résultats

Les 728 concepts de l'ontologie OntoTopo ont permis d'obtenir 406 pages désambigüisées (correspondant donc à des concepts), dont 182 qui contiennent au moins une SEP. Ces 182 pages ont constitué notre corpus. La chaîne de traitement que nous avons mis en œuvre a permis d'observer les résultats expérimentaux suivants.

Table 1. Nombre de structures énumératives dans le corpus

corpus (nombre de pages wikipedia)	nombre de structures énumératives avec MFM	nombre de SEP
182	434	276

Sur les 276 SEP, 149 des SEP présentent une amorce syntaxiquement correcte et non complète (amorces de type 3). Le tableau 2 présente les fréquences d'apparition des SEP en fonction du type de l'amorce.

Table 2. Fréquences des SEP en fonction du type de l'amorce

Amorce syntaxiquement non correcte	Amorce syntaxiquement correcte et complète	Amorce syntaxiquement correcte et non complète
0.27% (75/276)	**0.18%** (52/276)	**0.54%** (149/276)

Les 127 PES restantes ont pu être traduites en structures hiérarchiques. Leur élicitation de 127 SEP ont fourni 349 nouveaux concepts et 201 instances validés par les experts impliqués dans le projet.

5.3. Discussions et perspectives

L'évaluation de la qualité d'une ontologie est un problème encore ouvert. Elle peut être basée sur des mesures quantitatives (proximité avec une autre ontologie, couverture par rapport à un corpus, qualité des résultats recherchés, etc.) ou sur des aspects qualitatifs (consistance logique, validité conceptuelle, validation d'un expert, etc.). Ne disposant pas d'ontologie de référence évaluée par des experts qui puisse être comparée à notre ontologie produite automatiquement, nous avons choisi ici d'évaluer a posteriori le

nombre de concepts et de relations pertinents obtenus lors de notre phase d'enrichissement. Nous avons constaté que le nombre de concepts a été amélioré de 50%. Nous avons identifié plus de 80% de relations taxinomiques et 15% de relations méréonymiques.

Les résultats obtenus ne sont fonction que des seules SEP dont l'amorce est syntaxiquement non correcte ou syntaxiquement correcte et complète (situation 1 et 2, section 4.3.1).

Une première perspective est d'exploiter le troisième type de SEP décrit ci-dessus à l'aide de techniques d'apprentissage pour repérer le concept père présent dans l'amorce.

Par ailleurs, se limiter au premier syntagme nominal rencontré dans l'item est, dans certains cas, insuffisant pour caractériser précisément l'élément fils. L'exemple de la Fig. 9 montre que lorsque deux items ont même tête syntagmatique, un seul concept est alors généré. Pour être conforme à la SEP (i.e. trois concepts fils) nous envisageons d'élargir l'analyse de chaque item à l'intégralité de ses constituants.

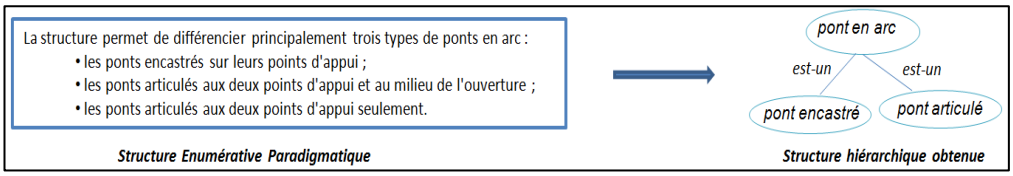

Fig. 9. : Un exemple de SEP où deux items ont même tête syntagmatique

Il est aussi des cas, comme dans l'exemple de la Fig. 10, où les items ne contiennent pas de syntagmes nominaux en leur tête. Cela correspond généralement à un phénomène d'ellipse. Individuellement, les items ne font sens ni en tant que label de concept, ni en tant que valeur d'instance. Ces items adjectivaux qualifient pourtant (en le spécialisant) l'élément père

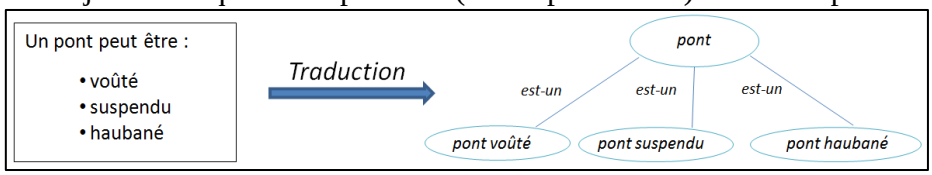

Fig. 10. : Un exemple de SEP où les items pris individuellement ne font pas sens

Des règles de correction ont déjà été proposées pour les items adjectivaux, les items énumératifs et l'inclusion lexicale (Kamel et al., 2010). Nous envisageons d'une part d'améliorer ces règles en ce qui concerne les items itératifs, et d'autre part d'élargir cet ensemble de règles au traitement des items quantitatifs ou numériques et aux items phrastiques.

Enfin, afin d'exploiter au mieux le contenu textuel des pages Wikipédia, il serait intéressant de coupler notre approche à celle de (Herbelot & Copestake, 2006) qui exploite les définitions des documents Wikipédia.

6. Conclusion

Le travail que nous avons présenté vise à améliorer la construction d'ontologie à partir de textes. Pour cela nous sommes partis de la facilité de repérage de structures énumératives possédant des caractéristiques de mise en forme bien établies. Puis, nous avons mis en évidence qu'il existait souvent une correspondance entre les structures énumératives de ce type et des structures du discours. C'est par ce biais que nous avons pu identifier le caractère ontologique de ces structures énumératives. A partir de là nous avons pu établir une chaîne de traitement qui partant d'une ontologie déjà établie permet de l'enrichir à partir de textes riches en structure énumératives. Les résultats obtenus sont suffisamment encourageants pour nous inciter à poursuivre ce travail dans plusieurs directions. La première direction est d'affiner l'exploitation des structures énumératives, nous avons indiqué des pistes dans ce sens dans la section 5.3. Une autre direction est de généraliser la démarche à d'autres types de mise en forme d'une part, et à d'autres structures telle que la définition. Enfin, nous envisageons d'utiliser des outils d'apprentissage supervisé déjà mis en œuvre pour l'identification de structures du discours.

Références

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

ASHER, N., VIEU, L. (2005) Subordinating and Coordinating Discourse Relations. Dans : Lingua, Elsevier, Vol. 115 N. 4, p. 591-610, 2005.

AUER, S., BIZER, C., LEHMANN, J., KOBILAROV, G., CYGANIAC, R., IVES, Z. (2007). DBpedia : a nucleus for a web of open data. In: Proceedings of the Sixth International Semantic Web Conference and Second Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, vol. 4825, pp 715-728

BRAS, M., PREVOT, L., VERGEZ-COURET, M. (2008). Quelle(s) relation(s) de discours pour les structures énumératives ? Actes du Colloque Mondial de Linguistique Française CMLF'08, Durand, J., Habert, B., Laks, B. (éds.), pp. 1945-1964, Paris

CARLSON, L and MARCU D. (2001). Discourse Tagging Manual. Unpublished manuscript, http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

CHERNOV, S., IOFCIU, T., NEJDL, W., ZHOU, X. (2006). Extracting semantic relationships between Wikipedia categories. In: Proceedings of the First International Workshop : SemWiki'06 - From Wiki to Semantics. Co-located with the Third Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro

CIMIANO P. (2006) Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer, November,

HERBELOT, A., COPESTAKE, A. (2006). Acquiring ontological relationships from Wikipedia using RMRS. In: Proceedings of the International Semantic

Web Conference 2006. Workshop on Web Content Mining with Human Language Technologies, Athens, GA

HO-DAC, L-M. (2007). Exploration en corpus de la position initiale dans l'organisation du discours. Thèse de doctorat en sciences du langage. Université de Toulouse 2.

HO-DAC L.-M., PÉRY-WOODLEY M.-P., TANGUY L. (2010). Anatomie des Structures Énumératives. Dans Actes de la conférence TALN 2010 - Traitement Automatique des Langues Naturelles, Canada

HOVY E.H., MAIER E. (1991). Parsimonious or Profligate: How many and which Discourse Structure Relations? Manuscrit.

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., AUSSENAC-GILLES N., LAIGNELET M. (2010). Correction d'ontologies construites à partir de la structure de documents. *Journées Francophones d'Ingénierie des Connaissances (IC 2010)*, Nîmes (France), 08/06/2010-11/06/2010, Sylvie Despres (Eds.), Ecole des Mines d'Alès, p. 29-40

LÜNGEN H., BÄRENFÄNGER M., HILBERT M., LOBIN H., AND 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.

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

NGUYEN, D.P.T., MATSUO, Y., ISHIZUKA, M. (2007). Relation extraction from Wikipedia using subtree mining. In: Proceedings of the AAAI'07 Conference, Vancouver, Canada, July 2007, pp. 1414-1420

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

POWER, R., SCOTT, D., BOUAYAD-AGHA, N. (2003). Document structure. Computational Linguistics, 29 (2), 211-260.

VIRBEL, J., LUC, C., SCHMID, S., CARRIO, L., DOMINGUEZ, C., PERY-WOODLEY, M-P., JACQUEMIN, C., MOJAHID, M., BACCINO, T. & GARCIA-DEBANC, C. (2005). Approche cognitive de la spatialisation du langage. De la modélisation de structures spatio-linguistiques des textes à l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. In Bullier, J. & Thinus-Blanc, C. (coord.) Agir dans l'espace. Paris : Editions de la Maison des sciences de l'homme.

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