Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles

Jean-Philippe Fauconnier[1] Mouna Kamel[1] Bernard Rothenburger[1]
Nathalie Aussenac-Gilles[1]
(1) IRIT, 118 route de Narbonne 31060 Toulouse Cedex 5
{prénom}.{nom}@irit.fr

RÉSUMÉ

Ce travail s'inscrit dans le cadre de la construction et l'enrichissement d'ontologies à partir de textes de type encyclopédique ou scientifique. L'originalité de notre travail réside dans l'extraction de relations sémantiques exprimées au-delà de la linéarité du texte. Pour cela, nous nous appuyons sur la sémantique véhiculée par les caractères typo-dispositionels qui ont pour fonction de suppléer des formulations strictement linguistiques qui seraient plus difficilement exploitables. L'étude que nous proposons concerne les relations sémantiques portées par les structures énumératives parallèles qui, bien qu'affichant des discontinuités entre ses différents composants, présentent un tout sur le plan sémantique. Ce sont des structures textuelles qui sont propices aux relations hiérarchiques. Après avoir défini une typologie des relations portées par ce type de structure, nous proposons une approche par apprentissage visant à leur identification. Sur la base de traits incorporant informations lexico-syntaxiques et typo-dispositionnelles, les premiers résultats aboutissent à une exactitude de 61,1%.

ABSTRACT

A Supervised learning for the identification of semantic relations in parallel enumerative structures

This work falls within the framework of ontology engineering and learning from encyclopedic or scientific texts. Our original contribution lies within the extraction of semantic relations expressed beyond the text linearity. To this end, we relied on the semantics behind the typo-dispositional characters whose function is to supplement the strictly linguistic formulations that could be more difficult to exploit. The work reported here is dealing with the semantic relations carried by the parallel enumerative structures. Although they display discontinuities between their various components, these enumerative structures form a whole at the semantic level. They are textual structures that are prone to hierarchic relations. After defining a typology of the relationships carried by this type of structure, we are proposing a learning approach aimed at their identification. Based on features including lexico-syntactic and typo-dispositional informations, the first results led an accuracy of 61.1%.

MOTS-CLÉS : extraction de relations, structures énumératives parallèles, mise en forme matérielle, apprentissage supervisé, construction d'ontologies.

KEYWORDS: relationship extraction, parallel enumerative structures, material shaping, supervised learning, ontology learning.

© ATALA

1 Introduction

La construction d'ontologies est un processus fastidieux qui nécessite la contribution d'experts d'un domaine. Une manière de rendre ce processus moins coûteux consiste à exploiter automatiquement certains types de textes, comme les textes de nature encyclopédique ou scientifique, afin d'en extraire les connaissances. Généralement, cette exploitation de textes s'appuie sur des analyses statistiques et/ou des analyses linguistiques essentiellement focalisées sur les niveaux lexicaux et syntaxiques. Citons notamment l'approche par apprentissage automatique (Nédellec *et al.*, 2009), l'utilisation de patrons (Giuliano *et al.*, 2006) ou encore une approche hybride combinant les deux (Giovannetti *et al.*, 2008).

Cependant, ces approches souffrent de deux limites : (1) l'analyse se situe en général à un niveau intraphrastique ou, du moins, textuellement linéaire et (2) l'extraction de connaissances se base sur des indices syntaxiques sans prendre en compte les caractéristiques de mise en forme du texte. Or, il existe des relations qui se matérialisent au travers de marqueurs paralinguistiques (marqueurs typographiques et/ou dispositionnels). Ces derniers, dépassant leur rôle de mise en forme, sont des éléments structurants porteurs de sémantique. (Virbel *et al.*, 2005) théorise ces marqueurs et leur utilisation au sein de la notion de *mise en forme matérielle* (MFM). Des analyses linguistiques fines ont mis en évidence le rôle fondamental de celle-ci dans l'interprétation d'un texte et dans la caractérisation de certains objets textuels tels que les titres (Rebeyrolle *et al.*, 2009), les définitions (Pascual et Péry-Woodley, 1995) et les structures énumératives (Luc, 2001).

Pour améliorer la construction d'ontologies, nous nous intéressons aux structures énumératives (SE) parallèles avec MFM. En tant que SE, elles sont porteuses de connaissances hiérarchiques. Leur caractère parallèle implique une composition homogène d'un point de vue grammatical, typo-dispositionnel et fonctionnel. Elles disposent souvent des propriétés textuelles qui les rendent visuellement perceptibles et ces propriétés sont suffisamment stables pour que leur repérage automatique puisse être envisagé (Ho-Dac *et al.*, 2012). Enfin, leur fréquence au sein des textes scientifiques, procéduraux ou encyclopédiques reste élevée.

Les approches précédentes (Kamel et Rothenburger, 2011; Kamel *et al.*, 2012) ont montré les limites d'une approche symbolique pour l'extraction des relations sémantiques au sein des SE. Dans cet article, nous proposons deux méthodes par apprentissage supervisé. La première combine des traits linguistiques et paralinguistiques et la seconde repose sur des trigrammes. Ce travail est une première étape vers l'exploitation de SE pour la construction d'ontologies.
La section 2 introduit les SE. La section 3 présente les classes de relations, le corpus ainsi que le mode d'évaluation. La section 4 décrit le classifieur d'entropie maximale (MaxEnt) ainsi que les deux approches. La section 5 présente les résultats obtenus par validation croisée. Enfin, la conclusion revient sur l'intérêt de ce travail et esquisse quelques perspectives.

2 Les structures énumératives

L'acte d'énumération consiste à regrouper des éléments indépendants sous un même critère d'homogénéité (Pery-Woodley, 2001). La forme générale d'une structure énumérative (SE) est caractérisée par une *amorce*, une *énumération* composée d'au moins deux *items* et éventuellement une *clôture* (ou conclusion). Cette structure logique générique peut se décliner concrètement par des dispositifs linguistiques ou textuels différents. Elle peut être énoncée au fil du texte en

 © ATALA

dehors de toute *mise en forme matérielle* (MFM) et dans ce cas les items sont introduits par des marqueurs lexicaux qui sont souvent des groupes adverbiaux (par exemple « premièrement », « deuxièmement », « troisièmement » dans (1)), ou au contraire être mise en évidence par l'usage de marqueurs typographiques et dispositionnels spécifiques (comme les caractères de ponctuation, les retraits, les tirets dans (2)).

(1) *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, évitez les lampes halogènes, car une lampe halogène de 500 watts consomme l'équivalent de 23 lampes. Troisièmement, essayez de remplacer les lampes traditionnelles par des lampes basse consommation.*

(2) *Les formes de communication non parlées sont :*
- *le langage écrit,*
- *le langage des signes,*
- *le langage sifflé.*

Il existe plusieurs définitions de l'énumération. La définition 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 coénumérabilité des entités recensées, par la coénumérabilité des segments linguistiques qui les décrivent. ».

(Luc, 2001) a établi une typologie des SE permettant de distinguer les structures *homogènes* vs. *hétérogènes*, les structures *syntagmatiques* vs. *paradigmatiques*, et les structures *isolées* vs. *non isolées*. Les structures *hétérogènes* présentent des items ayant des propriétés visuelles non équivalentes et sont plus difficilement repérables automatiquement en corpus. Les structures *syntagmatiques* entretiennent des liens de dépendance entre les items, et les structures *non isolées* entretiennent des relations avec des unités textuelles localisées en dehors de la structure énumérative. Les SE *paradigmatiques*, *homogènes* et *isolées* sont dites parallèles.

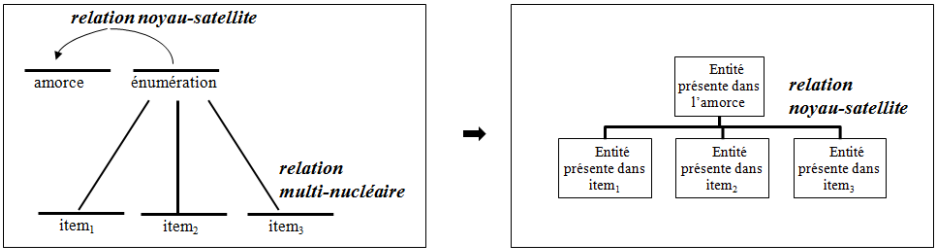

FIGURE 1 – Représentations sémantiques de la structure énumérative

Notre étude se focalise ici sur la SE parallèle car son analyse rhétorique (basée, par exemple, sur les principes de la RST (Carlson *et al.*, 2001)) permet d'établir une relation noyau-satellite qui relie l'amorce (unité d'information la plus saillante) à l'énumération (unité d'information qui supporte l'information d'arrière-plan), et une relation multi-nucléaire qui relie les items (arguments de même importance). La relation noyau-satellite sera généralement de type *elaboration* et la

© ATALA

relation multi-nucléaire de type *list*. Une relation hiérarchique entre l'amorce et chacun des items est ainsi mise en évidence. Dans ce cadre, nous envisageons de traduire cette structure rhétorique en une structure hiérarchique où les entités conceptuelles dénotées par des termes présents dans les segments textuels seraient extraites et reliées par la relation de type noyau-satellite identifiée en discours (Figure 1).

Identifier les relations portées par les SE parallèles avec MFM est l'objet de ce travail, car elles ont les propriétés (1) d'être *homogènes* et de bénéficier de traits de formatage assez réguliers pour que d'une part leur identification automatique en corpus puisse être envisagée, et que d'autre part les aspects typo-dispositionnels permettent de spatialiser le discours, et donc d'aider à la désambiguïsation, et (2) d'être *paradigmatiques* et *isolées*, ce qui assure l'unicité de la relation portée par la structure. Les processus d'identification des SE en texte ainsi que des concepts et instances au sein de ces dernières font l'objet de travaux complémentaires non discutés dans cet article. L'approche que nous développons ici est endogène dans le sens où les indices qui permettent d'identifier la relation sont recherchés au sein de la SE.

3 Classes, corpus et mode d'évaluation

3.1 Classes

L'arbre issu de l'analyse rhétorique (Figure 1) révèle souvent l'existence de connaissances ontologiques ou lexicales portées par la SE parallèle. Le but de ce travail est de détecter automatiquement la nature de la relation unique entre l'amorce et les items, lorsque cela est possible.

Classes	Description
isA	Relation hiérarchique d'hyperonymie.
partOf	Relation hiérarchique de méronymie.
instanceOf	Relation entre un concept et les instances de ce concept.
autreOntologique	Relations non-taxonomiques (e.g : *isCauseOf*, *requires*, etc.).
lexical	Relation lexicale entre termes (homonymie, synonymie, etc.).
autres	Cas ambigus et relations que l'on ne peut résoudre.

TABLE 1 – Annotation du corpus en 6 classes par 4 annotateurs

Pour transformer notre problème d'identification de relations en un problème de classification multi-classes, nous avons défini des classes correspondant aux relations recherchées. Nous distinguons 6 classes (Table 1) : *isA*, *partOf*, *instanceOf*, *autreOntologique*, *lexical* et *autres*.
À proprement parler, toutes les relations que nous désirons identifier sont lexicales, en ce sens qu'elles lient des termes au sein du texte. Cependant, dans leur dénomination, nous différencions les relations par le rôle qu'elles peuvent jouer, a posteriori, au sein d'une ressource sémantique. Par exemple, bien que les classes *isA* (exemple (2)) et *partOf* désignent les relations d'hyperonymie et de méronymie dans le domaine terminologique, nous nous référons à ces dernières sous l'appellation usitée dans le domaine des ontologies.

© ATALA

La classe *autreOntologique* comprend les relations non-taxonomiques entre concepts (exemple (3)). La classe *lexical* reprend les relations lexicales (synonymie, homonymie, etc.) et, éventuellement, les cas d'inclusion lexicale.

(3) *Tous les barrages classés (A, B, C et D) doivent disposer :*
- *d'une consigne de crue ;*
- *d'un dispositif d'auscultation adapté.*

La classe *autres* regroupe les cas ambigus, tels que ceux présentés par les SE navigationnelles et de titraille, et les relations que l'on ne peut résoudre. L'exemple (4) donne une SE de titraille qui structure un propos, un document. L'exemple (5) reprend une SE à visée navigationnelle, cas courant dans les ressources informatisées qui utilisent des liens hypertextes (indiqués ici par la mise en gras). Les liens hypertextes nous indiquent qu'il y a une plus grande probabilité que la relation portée par la SE lie des documents et non pas des termes. Enfin, l'exemple (6) présente un cas où il s'agit d'une élaboration argumentative et non ontologique.

(4) *Présentation*
1 *Fonctionnement*
2 *Terminologie*

(5) *Transports en commun*
- ***Portail des transports en commun***
- ***Portail du chemin de fer***

(6) *Le transformateur d'isolement comporte deux enroulements presque identiques au primaire et au secondaire :*
- *le nombre de spires du secondaire est souvent très légèrement supérieur au nombre de spires du primaire afin de compenser la faible chute de tension en fonctionnement ;*
- *en théorie, les sections de fil au primaire et au secondaire sont identiques, car l'intensité des courants est la même.*

3.2 Corpus

Les données utilisées dans notre travail sont issues des travaux de (Kamel et Rothenburger, 2011) visant l'enrichissement de l'ontologie OntoTopo, construite dans le cadre de l'ANR GEONTO [1]. Cette ontologie modélise les domaines de l'aménagement urbain, l'environnement et l'organisation territoriale. (Kamel et Rothenburger, 2011) ont construit leur corpus en projetant les concepts de l'ontologie OntoTopo sur les pages de Wikipédia et en extrayant, dans les pages ainsi retenues, les SE parallèles rencontrées. Par leur caractère encyclopédique, les pages Wikipédia ordonnent de nombreuses définitions et propriétés au moyen de marqueurs typo-dispositionnels. Le nombre relativement élevé de SE parallèles par page s'explique notamment par la recommandation du « Manuel of Style » de Wikipédia [2] qui préconise une forme grammaticale identique pour tous les items d'une SE. Au final, 2317 SE furent extraites de 276 pages.

À partir de ce travail, nous avons construit deux corpus respectivement nommés CORPUS_SE et CORPUS_DISTRIB. Ces derniers reprennent 1000 SE annotées par quatre annotateurs : deux ingénieurs de la connaissance, un ergonome et une étudiante. La tâche d'annotation a consisté à classer parmi les six classes définies en section 3.1 la relation sémantique portée par la SE parallèle. Un κ de Fleiss (Fleiss *et al.*, 1979), sous l'hypothèse nulle de jugements indépendants,

1. Collaboration entre le COGIT, le LRI, le LIUPPA et l'IRIT - `http://geonto.lri.fr/`
2. `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`

© ATALA

présente un accord inter-annotateurs, relativement correct pour six classes[3], de 0,559 (Table 2). Notons que la corrélation entre les deux ingénieurs de la connaissance est nettement plus grande (0,686). Seule la classe *autreOntologique* présente un accord relativement bas (0,299) qui s'explique par ses caractéristiques formelles peu stables. En cas de désaccord dans l'annotation, la classe finale d'une SE est décidée par consensus entre les différents annotateurs.

Classe	Kappa	Var	*Z*-score	p-score	IC 95%
isA	0,509	0,001033	15,859	< 1E-09	[0,446 ;0,572]
partOf	0,493	0,001275	13,808	< 1E-09	[0,423 ;0,563]
autreOntologique	**0,299**	0,000945	09,735	< 1E-09	[0,238 ;0,359]
instanceOf	0,652	0,000975	20,895	< 1E-09	[0,591 ;0,713]
lexical	0,636	0,000974	20,392	< 1E-09	[0,575 ;0,697]
autres	0,641	0,001369	17,323	< 1E-09	[0,568 ;0,713]
Corpus	**0,559**	3,112E-05	100,269	<1E-09	[0,548 ;0,570]

TABLE 2 – Annotation du corpus en 6 classes par 4 annotateurs

Les deux corpus ont été étiquetés grammaticalement et morpho-syntaxiquement par l'outil Talismane (Urieli et Fauconnier, 2012) entraîné sur le French TreeBank (Abeillé *et al.*, 2003) dans sa version en dépendances (Candito *et al.*, 2009). CORPUS_SE totalise 1000 SE et comprend 80 774 tokens. Ces 1000 SE présentent en moyenne 4,31 items par individu. Afin d'évaluer notre méthode sur sa capacité à identifier la relation entre une amorce et un item, nous avons construit le corpus CORPUS_DISTRIB en distribuant, pour chaque SE de CORPUS_SE, son amorce sur les n items qu'elle contient afin de construire n paires amorce-item. CORPUS_DISTRIB totalise 4317 paires amorce-item et comprend 119 272 tokens.

Dans CORPUS_SE, la répartition des SE dénote un déséquilibre entre la classe *autres* et le reste des classes (Figure 2). Un autre déséquilibre apparaît au niveau du nombre d'items de chaque SE (Table 3). La classe *instanceOf* contient des SE avec un nombre d'items supérieur à 20, dont l'une énumère plus de soixante types de sports collectifs. Avec une moyenne de 6,95 items par SE, la classe *instanceOf* est celle qui présente les valeurs les plus extrêmes, suivie par la classe *isA* avec une moyenne de 4,25 items par SE. Ce déséquilibre entre les items influence la répartition des paires amorce-item au sein de CORPUS_DISTRIB (Figure 2).

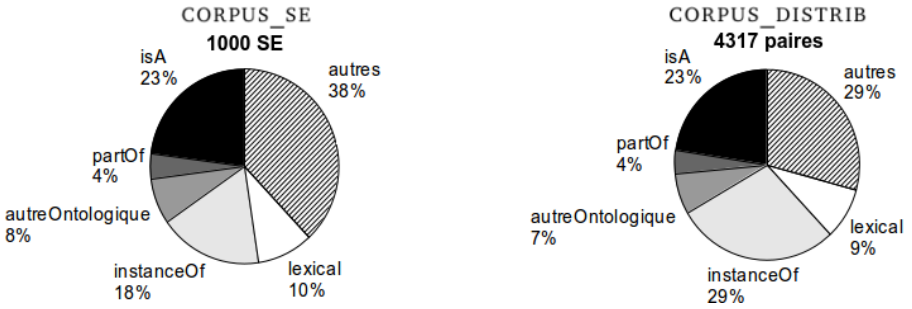

FIGURE 2 – Répartition des SE dans CORPUS_SE et des paires amorce-item dans CORPUS_DISTRIB

3. Le κ de Fleiss est sensible aux nombres de catégories.

© ATALA

Classe	SE	Nb. items	min-max	Moyenne	σ	IC 95%
isA	229	973	2 - 20	**4,25**	**2,71**	[3,90 ;4,60]
partOf	42	173	2 - 11	4,12	1,89	[3,53 ;4,71]
autreOntologique	76	299	2- 10	3,93	1,84	[3,51 ;4,36]
instanceOf	177	1231	2 - 63	**6,95**	**6,99**	[5,92 ;7,99]
lexical	96	377	2 - 14	3,93	2,03	[3,52 ;4,34]
autres	380	1264	2 - 13	3,33	1,76	[3,15 ;3,50]
Total	1000	4317	2 - 63	4,31	3,72	[4,07 ;4,54]

TABLE 3 – Nombre d'items par classes

3.3 Mode d'évaluation

Rappelons que les SE parallèles présentent la particularité que chaque item est relié à l'amorce par une même relation (Section 2). Par conséquent, il est possible d'identifier la relation entière portée par une SE si la relation entre son amorce et l'un de ses items est identifiée.
Dans cet objectif, nous proposons deux méthodes de classification par apprentissage supervisé. La première méthode utilise des traits linguistiques et paralinguistiques. La seconde repose sur des trigrammes de tokens. En outre, nous posons une *baseline* naïve qui classe tous les individus dans la classe majoritaire *autres*.

Les deux méthodes ainsi que la *baseline* ont été évaluées dans trois tâches :
– La **tâche 1** vise à la classification des SE issues de CORPUS_SE (1000 SE à classer). Dans la figure 1, il s'agit de classer la SE en identifiant la relation entre l'amorce et l'item$_1$.
– La **tâche 2** vise la classification des paires amorce-item de CORPUS_DISTRIB (4317 paires à classer). Dans la figure 1, il s'agit de classer les relations unissant amorce-item$_1$, amorce-item$_2$ et amorce-item$_3$.
– La **tâche 3** vise la classification des SE de CORPUS_SE à partir de la moyenne des prédictions de leurs paires amorce-item issues de CORPUS_DISTRIB (1000 SE à classer à partir des prédictions de 4317 paires). Dans la figure 1, il s'agit de classer la SE en calculant la moyenne des prédictions de amorce-item$_1$, amorce-item$_2$ et amorce-item$_3$.

Pour les trois tâches (et les 9 évaluations correspondantes), nous procédons à une validation croisée à 10 échantillons et mesurons l'exactitude (*micro-average*) pour la classification toutes classes confondues ainsi que rappel, précision et F-mesure pour chacune des classes :

$$\text{exactitude} = \frac{\sum_i^C \left(VP_i + VN_i\right)}{\sum_i^C \left(VP_i + VN_i + FP_i + FN_i\right)}$$

$$\text{précision}_i = \frac{VP_i}{VP_i + FP_i} \qquad \text{rappel}_i = \frac{VP_i}{VP_i + FN_i} \qquad \text{F1}_i = 2\frac{\text{précision}_i\ \text{rappel}_i}{\text{précision}_i + \text{rappel}_i}$$

où VP_i, VN_i, FP_i et FN_i sont respectivement le nombre de Vrais Positifs, Vrais Négatifs, Faux Positifs et Faux Négatifs de classes i à C où C est le nombre de classes. L'exactitude permet d'évaluer la capacité prédictive d'un modèle en donnant un poids proportionnel à chaque classe. Nous donnons aussi un écart type qui, selon (Kohavi *et al.*, 1995), peut être estimé (où N est le nombre d'individus) :

$$\sigma = \sqrt{\frac{\text{exactitude}\ (1 - \text{exactitude})}{N}}$$

© ATALA

4 Le modèle d'apprentissage

4.1 Maximum d'entropie

Pour notre tâche de classification, nous avons adopté un modèle conditionnel d'entropie maximale, dit aussi MaxEnt (Berger *et al.*, 1996). Ce modèle, qui a déjà fait ses preuves en TAL, permet de gérer de manière flexible un grand nombre de traits et repose sur le principe de maximisation d'entropie. Ce dernier vise à définir une contrainte pour chaque information observée et choisir la distribution qui maximise l'entropie tout en restant consistante vis-à-vis de l'ensemble de ces contraintes (Jaynes, 1957). Dans ce cadre d'optimisation sous contraintes, il est mathématiquement prouvé qu'une solution unique existe et un algorithme itératif garantit la convergence vers cette dernière (Ratnaparkhi, 1996). La forme classique du MaxEnt est la suivante :

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} w_i f_i(x, y) \right)$$

où $P(y|x)$ désigne la probabilité que l'individu x, ici une SE ou une paire amorce-item, appartienne à la classe y (e.g : *isA*, *partOf*, etc.). La fonction f_i est une fonction binaire appelée trait qui permet de définir les contraintes du modèle. Z(x) est une constante de normalisation qui assure que la somme des probabilités retournées pour un individu soit équivalente à 1. Chaque individu x est encodé comme un vecteur avec n traits f_i. Le paramètre w_i, dit aussi poids, de chaque trait associe à chaque individu une probabilité d'appartenance à une classe.
En pratique, le calcul de cette maximisation s'effectue au travers de différents algorithmes tels que le *Generalized Iterative Scaling* (GIS) (Darroch et Ratcliff, 1972) ou l'*Improved Iterative Scaling* (IIS) (Berger *et al.*, 1996). Dans notre travail, nous utilisons l'implémentation Apache OpenNLP MaxEnt [4] qui applique un GIS sans *correction feature* tel que recommandé par (Curran et Clark, 2003).

4.2 Approche par traits

Comme présenté dans la section 2, les SE parallèles sont des objets textuels qui conjuguent marqueurs de MFM et propriétés lexico-syntaxiques partiellement stables. La première approche proposée tente de capturer leurs régularités au moyen de deux familles de traits : (1) la première emploie les informations lexico-syntaxiques extraites de l'analyse des tokens et de leur rôle au sein de l'arbre de dépendances et (2) la deuxième famille reprend des informations paralinguistiques, dans ce cas-ci typographiques. La table 4 présente, de manière synthétique, les différents traits. La distinction entre les deux familles de traits est graduelle et certains traits reprennent des informations combinant les deux sources. Par exemple, avec un seuil de sélection de traits paramétré à 5 apparitions dans le corpus [5], le trait *LastTokenPos* renvoie dans la majorité des cas l'étiquette PONCT, car amorces et items ont tendance à être clôturés par une ponctuation.

Les traits *HasClassifier*, *HasMeronym* et *HasCircumstant* sont calculés en projetant des patrons sur les SE. Ces derniers, comme l'ensemble des traits, sont issus d'intuitions linguistiques. Une étude approfondie de leurs poids respectifs fera l'objet de travaux ultérieurs.

4. `http://opennlp.apache.org/`
5. Ce seuil de sélection des traits est appelé *cut-off* dans la littérature relative au MaxEnt.

© ATALA

Éléments	Traits	Phénomènes captés
Amorce	(First|Last)Token (Pos|Lem)	Retourne respectivement la catégorie grammaticale et le lemme du dernier/premier token de l'amorce.
	HasClassifier	Présence d'un classifieur (« sortes de », « types de », etc.)
	HasMeronym	Présence d'un marqueur de méronymie (« parties de », etc.)
	HasVerb	Présence d'un verbe conjugué ou un verbe au participe passé à la racine de l'arbre de dépendances.
	HasProperNoun	Présence d'un nom propre.
	HasPlurarlNoun	Présence d'un nom commun au pluriel.
	MultiplSentence	Présence de plusieurs phrases dans l'amorce.
Item	(First|Last)Token (Pos|Lem)	Retourne respectivement la cat. gram. et lemme du dernier/premier token de l'item.
	HasVerb	Présence d'un verbe conjugué ou un verbe au participe passé à la racine de l'arbre de dépendances.
	HasProperNoun	Présence d'un nom propre.
	HasDate	Présence d'une date en années et considérée comme NC (« 1996 », etc.)
	HasCircumstant	Présence d'un circonstant (« En Belgique », etc.)
	StartsWithINF	Présence d'un infinitif en début de phrase.
	StartsWithNUM	Présence d'un numéro en début d'item.
	StartsWithMaj	Présence d'une majuscule en début d'item.
	ContainsPonct	Présence d'une ponctuation inhabituelle au sein de l'item.
	MultiplSentence	Présence de plusieurs phrases dans l'item.

TABLE 4 – Tableau synthétique des traits utilisés

La SE en (7) de classe *isA* exemplifie l'application de traits. Les tokens entre crochets sont ceux auxquels s'appliquent les traits *(First|Last)Token(Pos|Lem)*. Les éléments en gras sont le classifieur et le verbe souligné répond au trait *HasVerb*. L'absence d'autres phénomènes (nom propre dans l'amorce, etc.) est tout autant informative pour la classification de la SE.

(7) *[Pour] un transformateur triphasé, il existe **3 types de** couplage d'enroulement [:]*
- *[le] couplage étoile, défini par la lettre Y [;]*
- *[le] couplage triangle, défini par la lettre D ou Δ[;]*
- *[le] couplage zig-zag, défini par la lettre Z[.]*

L'exemple (8) présente un cas issu de la classe *autres* où l'on voit que la présence d'un infinitif en début d'item (indiqués ici par la mise en gras) est un indice des SE procédurales.

(8) *Le déroulement*
1 *[**Mélanger**] la farine, le sucre, le sucre vanillé, les œufs, l'huile et le lait [.]*
2 *[**Verser**] la pâte dans la poêle et retourner la crêpe avec une spatule [.]*

4.3 Approche par trigrammes

La deuxième approche proposée vise à identifier les relations sémantiques au sein des SE parallèles au moyen de trigrammes. Chaque token est étiqueté soit par sa catégorie grammaticale

 © ATALA

seule, soit par sa forme lemmatisée associée à sa catégorie grammaticale. L'ajout de la catégorie grammaticale au lemme permet de diminuer les cas d'ambiguïté. Par ce choix, nous pouvons distinguer, par exemple, plat-ADJ vs. plat-NC. Pour chaque séquence de trois tokens, nous avons 2^3 trigrammes différents. Par exemple, pour la séquence « Le chat noir », les trigrammes suivants sont calculés : « le-DET chat-NC noir-ADJ », « DET chat-NC noir-ADJ », . . ., « DET NC ADJ ». Ces trigrammes sont appliqués de manière identique sur l'amorce et les items.

5 Résultats

Pour les trois tâches, nous avons procédé à une validation croisée (k=10) et présentons les résultats en termes d'exactitude (Table 5).

	Approches	Exactitude	σ	IC 95%
Tâche 1	Traits ling.	**61,10**%	0,0154	[58,08 ;64,11]
	Trigrammes	59,80%	0,0155	[56,76 ;62,83]
	Baseline autres	38,00%	0,0153	[35,00 ;40,99]
Tâche 2	Traits ling.	58,70%	0,0074	[57,25 ;60,15]
	Trigrammes	**59,50**%	0,0074	[58,04 ;60,95]
	Baseline autres	29,30%	0,0069	[27,94 ;30,65]
Tâche 3	Traits ling.	58,50%	0,0155	[55,46 ;61,53]
	Trigrammes	**59,00**%	0,0155	[55,96 ;62,03]
	Baseline autres	38,00%	0,0153	[35,00 ;40,99]

TABLE 5 – Évaluation pour les trois tâches

Au regard de cette dernière, nous constatons que, pour les trois tâches, l'approche par traits et l'approche par trigrammes aboutissent significativement à de meilleurs résultats face à la *baseline autres*. Par contre, la comparaison des deux approches est à nuancer. Contrairement aux tâches 2 et 3, l'approche par traits dépasse de peu les trigrammes dans la tâche 1, mais les intervalles de confiance nous montrent qu'il y a un possible recouvrement entre ces résultats. Cependant, l'exactitude, qui donne un poids proportionnel à chaque classe (section 3.3), ne révèle pas les difficultés éprouvées par les trigrammes pour classer les individus des classes minoritaires *partOf* et *autreOntologique* dans les trois tâches. Ces derniers sont souvent classés à tort dans les classes *autres*, *instanceOf* ou *isA* qui, dans les deux corpus, sont les classes majoritaires. La comparaison des matrices de confusion issues des deux approches pour la tâche 1 l'exemplifie (Tables 6 et 7).

	autrOnto	autres	instOf	isA	lexical	partOf	Précision	Rappel	F1
autrOnto	**13**	16	5	41	1	0	**0,54**	**0,17**	**0,26**
autres	6	**298**	29	46	0	1	0,63	0,78	0,70
instOf	0	42	**120**	9	6	0	0,66	0,68	0,67
isA	5	88	8	**120**	5	3	0,46	0,52	0,49
lexical	0	12	13	23	**47**	1	0,80	0,49	0,61
partOf	0	14	8	20	0	**0**	**0,00**	**0,00**	**0,00**

TABLE 6 – Matrice de confusion pour la Tâche 1 : Trigrammes

© ATALA

	autrOnto	autres	instOf	isA	lexical	partOf	Précision	Rappel	F1
autrOnto	**25**	11	3	29	6	2	**0,40**	**0,33**	**0,36**
autres	10	**300**	18	37	15	0	0,70	0,79	0,74
instOf	6	20	**113**	29	5	4	0,70	0,64	0,67
isA	11	66	13	**127**	12	0	0,50	0,55	0,53
lexical	6	23	12	15	**40**	0	0,50	0,42	0,45
partOf	4	10	3	17	2	**6**	**0,50**	**0,14**	**0,22**

TABLE 7 – Matrice de confusion pour la Tâche 1 : Traits linguistiques

Ainsi, de manière transversale, l'approche par traits linguistiques reste toujours préférable à l'approche par trigrammes car elle discrimine mieux les classes minoritaires *partOf* et *autreOntologique*, utiles à la construction de ressources sémantiques.

Les résultats obtenus avec l'approche par traits linguistiques dans les tâches 2 et 3 aboutissent à une identification correcte des classes utiles à la construction d'ontologies, c'est-à-dire toutes les classes sauf *autres* (Tables 8 et 9). Seules les F-mesures des classes *instanceOf* et *isA* diminuent entre la deuxième et la troisième tâche. Une difficulté à classer correctement les nombreux items de ces deux classes (Section 3.2) expliquerait cette diminution des scores.

	autrOnto	autres	instOf	isA	lexical	partOf	Précision	Rappel	F1
autrOnto	**94**	37	12	111	37	8	0,36	0,31	**0,34**
autres	35	**859**	100	195	61	14	0,61	0,68	0,65
instOf	49	140	**819**	96	104	23	0,76	0,67	**0,71**
isA	59	244	77	**543**	31	19	0,52	0,56	**0,54**
lexical	15	81	50	41	**173**	17	0,41	0,46	**0,43**
partOf	8	37	13	54	14	**47**	0,37	0,27	**0,31**

TABLE 8 – Matrice de confusion pour la Tâche 2 : Traits linguistiques

	autrOnto	autres	instOf	isA	lexical	partOf	Précision	Rappel	F1
autrOnto	**22**	10	4	31	8	1	0,36	0,29	**0,32**
autres	11	**289**	16	50	12	2	0,68	0,76	0,72
instOf	6	29	**100**	25	15	2	0,70	0,56	**0,63**
isA	16	65	9	**118**	13	8	0,47	0,52	**0,49**
lexical	5	23	11	9	**44**	4	0,46	0,46	**0,46**
partOf	1	8	2	16	3	**12**	0,41	0,29	**0,34**

TABLE 9 – Matrice de confusion pour la Tâche 3 : Traits linguistiques

En comparant la tâche 1 et la tâche 3 dans l'approche par traits linguistiques, les résultats suggèrent que l'identification de la relation entre l'amorce et le premier item est à ce jour la meilleure approche, surtout lorsque les SE présentent un grand nombre d'items (cf. *instanceOf* et *isA*). Seule la F-mesure de la classe *partOf* est significativement plus haute dans la tâche 3. Ce phénomène pourrait s'expliquer par le fait que certaines SE de cette classe présentent un premier item avec du texte rédigé et non un simple syntagme nominal (exemple (9)). Ce type de variation réduit les performances lorsque l'approche se limite au premier item (Tâche 1).

© ATALA

(9) *Les installations de la centrale électrique comprennent :*
- *un ou plusieurs postes électriques permettant la connexion au réseau électrique par l'intermédiaire d'une ou plusieurs lignes à haute tension ainsi qu'une interconnexion limitée entre tranches,*
- *les bâtiments administratifs,*
- *les bâtiments techniques,*
- *le magasin général.*

D'un point de vue qualitatif, les résultats ont aussi montré que, malgré une approche qui se veut fine et linguistique, il reste difficile de classer les individus où il y a ellipse de constituants au sein de leur amorce. Ces amorces incomplètes, tant syntaxiquement que lexicalement, sont un phénomène courant dans les documents numériques où la mise en page et les traits de formatage suppléent l'aspect lexico-syntaxique (Bush, 2003). Par exemple, la SE en (10) de classe *partOf* est, dans toutes les tâches, classée à tort dans la classe *autres*. L'absence de marqueurs de méronymie et de ponctuations ainsi que la présence de numéros en début d'item rendent difficile à distinguer cet individu d'une SE de titraille.

(10) *Système de la cordillère américaine*
1 *Montagnes rocheuses*
2 *Chaînes côtières du Pacifique*
3 *Cordillère des Andes*

Plusieurs stratégies sont envisageables pour contourner ce type de difficulté : (1) entreprendre une approche exogène du problème afin de capter des indices qui se trouvent en-dehors des SE, (2) étudier davantage l'utilisation de traits paralinguistiques (e.g : changement de police, présence de liens hypertextes dans les items, etc.) ou (3) identifier les concepts et instances déjà représentés dans l'ontologie en cours de construction et les utiliser comme de nouveaux indices pour mieux discriminer les classes de relations.

6 Conclusion

Dans cet article, nous avons défini un premier ensemble de classes des relations sémantiques portées par les SE et avons souligné leur intérêt dans la construction de ressources sémantiques. Dans ce cadre, nous avons proposé deux approches par apprentissage supervisé afin d'identifier ces relations. La première utilise des traits lexico-syntaxiques et paralinguistiques et la seconde aborde la classification au moyen de trigrammes. Les résultats montrent l'insuffisance de cette dernière pour capter des régularités pertinentes au sein des SE, notamment pour les classes minoritaires *partOf* et *autreOntologique*. La comparaison entre la tâche 1 et la tâche 3 suggère une meilleure classification des SE en se limitant à l'identification de la relation entre l'amorce et le premier item. Notons que, lors de nos expérimentations, l'augmentation du seuil de sélection des traits dans les tâches 2 et 3 a abouti à des scores plus élevés. Les causes de cette amélioration feront l'objet de travaux ultérieurs.
À terme, l'identification des relations au sein des SE s'inscrit dans un projet plus large visant, en amont, leur repérage automatique, comme cela a déjà été entrepris par (Morin, 1999), et, en aval, la construction d'une ontologie. Adjointe à un système d'extraction de relations au niveau intraphrastique, notre approche permettrait d'augmenter le rappel pour la tâche d'identification des relations en texte.

© ATALA

Par ailleurs, nos travaux ont aussi soulevé des pistes de réflexion au niveau du problème de classification en lui-même. Premièrement, il serait intéressant d'utiliser d'autres modèles d'apprentissage, tels que les CRF ou SVM, afin de mesurer l'influence, à traits égaux, du classifieur utilisé. Deuxièmement, il nous est apparu que la classe *autres* représentait avant tout une classe « par défaut » plutôt qu'un réel regroupement de relations et que peu de traits parvenaient à la discriminer correctement. Certaines méthodes de classification multi-classes préconisent l'utilisation de multiples modèles binaires et/ou la mise en place d'un seuil statique ou dynamique sur les probabilités d'appartenance. Une perspective à ce travail consisterait à établir un certain seuil en deçà duquel les individus seraient classés dans une catégorie *sansRelation*. Enfin, il serait utile de procéder à un sur-échantillonnage des classes minoritaires afin de comparer l'influence de la distribution des individus au sein de notre corpus.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *Treebanks*, pages 165–187.

Berger, A., Pietra, V. et Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Bush, C. (2003). Des déclencheurs des énumérations d'entités nommées sur le web. *Revue québécoise de linguistique*, 32(2):47–81.

Candito, M., Crabbé, B., Denis, P. et Guérin, F. (2009). Analyse syntaxique du français : des constituants aux dépendances. *In Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

Carlson, L., Marcu, D. et Okurowski, M. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. *In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1–10. Association for Computational Linguistics.

Curran, J. et Clark, S. (2003). Investigating gis and smoothing for maximum entropy taggers. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 91–98. Association for Computational Linguistics.

Darroch, J. et Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, 43(5):1470–1480.

Fleiss, J., Nee, J. et Landis, J. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5):974–977.

Giovannetti, E., Marchi, S. et Montemagni, S. (2008). Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text. *In Proc. of the 5th Workshop on Semantic Web Applications and Perspectives*. Citeseer.

Giuliano, C., Lavelli, A. et Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. *In Proceedings of the eleventh conference of the European chapter of the association for computational linguistics (EACL-2006)*, pages 5–7.

Ho-Dac, L., Fabre, C., Péry-Woodley, M., Rebeyrolle, J. et Tanguy, L. (2012). An empirical approach to the signalling of enumerative structures. *Discours. Revue de linguistique, psycholinguistique et informatique*, (10).

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

 © ATALA

KAMEL, M., MOJAHID, M. et ROTHENBURGER, B. (2012). "quand rédiger c'est décrire" mise en forme matérielle des textes et construction d'ontologies à partir de textes. *Journées Francophones d'Ingénierie des Connaissances (IC 2012)*.

KAMEL, M. et ROTHENBURGER, B. (2011). Elicitation de structures hiérarchiques à partir de structures enumératives pour la construction d'ontologie. *In Journées Francophones d'Ingénierie des Connaissances (IC 2011), Annecy*.

KOHAVI, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In International joint Conference on artificial intelligence*, volume 14, pages 1137–1145. Lawrence Erlbaum Associates Ltd.

LUC, C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *In Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 263–272.

MORIN, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes.

NÉDELLEC, C., NAZARENKO, A. et BOSSY, R. (2009). Information extraction. *Handbook on Ontologies*, pages 663–685.

PASCUAL, E. et PÉRY-WOODLEY, M. (1995). La définition dans le texte. *Textes de type consigne–Perception, action, cognition*, pages 65–88.

PERY-WOODLEY, M. (2001). Modes d'organisation et de signalisation dans des textes procéduraux. *Langages*, 35(141):28–46.

RATNAPARKHI, A. (1996). A maximum entropy model for part-of-speech tagging. *In Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA.

REBEYROLLE, J., JACQUES, M., PÉRY-WOODLEY, M. *et al.* (2009). Titres et intertitres dans l'organisation du discours 1. *Journal of French Language Studies*, 19(2):269.

URIELI, A. et FAUCONNIER, J. (2012). Talismane user manual. *CLLE-ERSS, Toulouse*.

VIRBEL, J. (1999). Structures textuelles, planches fascicule 1 : Enumérations, version 1,. Rapport technique, IRIT.

VIRBEL, J., LUC, C., SCHMID, S., CARRIO, L., DOMINGUEZ, C., PERY-WOODLEY, M., JACQUEMIN, C., MOJAHID, M., BACCINO, T. et GARCIADEBANC, 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. *Agir dans l'espace. Paris : Éditions de la Maison des sciences de l'homme*.

© ATALA