TALN 2004, Fès, 19–21 avril 2004

Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie

Véronique Malaisé (1, 2), Pierre Zweigenbaum (2) et Bruno Bachimont (1)
(1) DRE de l'Institut National de l'Audiovisuel
4, avenue de l'Europe, 94366 Bry-sur-Marne Cedex
{vmalaise, bbachimont}@ina.fr
(2) STIM/AP-HP, ERM 202 INSERM & CRIM-INaLCO
91, boulevard de l'Hôpital, 75013 Paris
pz@biomath.jussieu.fr

Résumé - Abstract

Pour construire une ontologie, un modéliseur a besoin d'objecter des informations sémantiques sur les termes principaux de son domaine d'étude. Les outils d'exploration de corpus peuvent aider à repérer ces types d'information, et l'identification de couples d'hyperonymes a fait l'objet de plusieurs travaux. Nous proposons d'exploiter des énoncés définitoires pour extraire d'un corpus des informations concernant les trois axes de l'ossature ontologique : l'axe vertical, lié à l'hyperonymie, l'axe horizontal, lié à la co-hyponymie et l'axe transversal, lié aux relations du domaine. Après un rappel des travaux existants en repérage d'énoncés définitoires en TAL, nous développons la méthode que nous avons mise en place, puis nous présentons son évaluation et les premiers résultats obtenus. Leur repérage atteint de 10% à 69% de précision suivant les patrons, celui des unités lexicales varie de 31% à 56%, suivant le référentiel adopté.

In order to build an ontology, a modeler needs to objectivate semantic information about the main terms of his domain. Some tools meant to explore corpora can help pointing out this information, and previous work has focused on the identification of hyperonyms. We propose here to rely on lay definitions to extract the information necessary to build an ontology structure: the vertical axis, related to hypernymy, the horizontal axis, related to co-hyponymy, and the transversal axis, linked to domain-related cross relations. After a survey of previous work about the extraction of definitions in NLP, we develop the method we followed, then present its evaluation criteria and the first results. The mining of lay definitions reached from 10 to 69% of precision, depending on the pattern involved, the mining of lexical items varied from 31 to 56%, following the reference considered.

Mots-clefs – Keywords

Repérage d'énoncés définitoires, relations sémantiques, patrons lexico-syntaxiques.
Mining definitions, semantic relations, lexico-syntactic pattern.

1 Introduction

La construction d'ontologie à partir de corpus textuel est un des champs explorés en Traitement Automatique des Langues (TAL). La position généralement adoptée consiste à concevoir des méthodes et outils pour aider un modéliseur d'ontologie dans sa tâche : par exemple Terminae (Szulman S., Biébow B., Aussenac-Gilles N., 2002), ou la combinaison Syntex-Upery (Bourigault D. et Lame G., 2002). L'approche retenue considère une ontologie comme une terminologie construite et structurée à partir de corpus, à formaliser. Il existe également un certain nombre d'approches concernant directement cette structuration de terminologie, soit à partir de patrons lexico-syntaxiques (de (Hearst M., 1992) à (Séguéla P. et Aussenac-Gilles N., 1999), ou dans le système KAON – `http://kaon.semanticweb.org/`), soit à partir de marqueurs associés à des règles (Le Priol F., 2001). Dans les deux cas, l'approche se focalise principalement sur l'hyperonymie. Mais d'autres types de relations sont à importantes pour la construction d'ontologie : la relation horizontale entre un terme et ses co-hyponymes, et les relations transversales qui relient un terme à un autre de manière non hiérarchique, mais actancielle (comme *personne−est_protagoniste_de →action*). La relation horizontale garantit la validité d'un «palier» ontologique et a des conséquences sur la profondeur de la hiérarchie, et les relations transversales permettent d'exprimer des connaissances à partir des termes hiérarchisés.

Certains énoncés sont particulièrement propices au repérage de ces différentes relations : il s'agit des énoncés définitoires. Rebeyrolle et Condamines (2000), par exemple, y recherchaient l'expression de la relation d'hyperonymie et de meronymie. Nous essayons de repérer à partir de ces énoncés des éléments lexicaux pouvant également aider à la structuration horizontale et transversale de l'ontologie. Notre travail s'attache principalement à la détection et à l'exploitation des énoncés définitoires, mais nous a également conduit à considérer d'autres types de contextes intéressants.

Après un rappel des travaux existants en repérage d'énoncés définitoires (section 2), nous présentons notre expérimentation sur le repérage et l'exploitation de ces énoncés, que nous avons testés sur un petit corpus, centré sur la thématique de la petite enfance (nos exemple sont extraits de ce test), puis évalués sur un corpus de diététique (section 3). Nous discutons les résultats obtenus (section 4) et concluons sur des perspectives à ce travail (section 5).

2 Travaux antérieurs

Les énoncés définitoires (ED) qui nous intéressent sont des formulations suivant la fonction métalinguistique de la langue dans sa pratique ordinaire. Auger (1997) a référencé plusieurs typologies de définitions de dictionnaires et fournit un ensemble d'indices terminologiques d'énoncés définitoires pour leur repérage en corpus. L'auteur mentionne les indices de bas niveau (autour de la ponctuation), ou de haut niveau (certaines formes syntaxiques) pour repérer ces ED, mais consacre son travail à leur repérage au moyen de marqueurs lexicaux. Condamines (1993), Rebeyrolle (2000) et Pearson (1998) présentent des éléments théoriques et des patrons lexico-syntaxiques concrets pour le repérage d'ED en corpus, intégrant dans leurs schémas des contraintes liées à la ponctuation. Muresan et Klavans (2002) proposent une alternative aux patrons lexico-syntaxiques, avec leur outil DEFINDER basé sur des règles linguistiques et des contraintes. Rebeyrolle (2000) et Muresan et Klavans (2002) présentent des évaluations de leurs outils, en prenant comme référence respectivement un corpus étiqueté manuellement et un en-

semble de textes étalon. Ils obtiennent des taux de précision de 17,95 à 79,19% pour le premier (suivant les types de patrons considérés) et de 86,95% pour le second, et de 94,75 à 100% et 75,47% pour le rappel.
Nous avons choisi de nous fonder sur des patrons lexico-syntaxiques pour le repérage d'énoncés définitoires, ce qui nous a permis de nous appuyer sur les travaux antérieurs de cette catégorie. Ceux-ci étant en majorité orientés vers les indices lexicaux, nous avons tenté de prendre en considération davantage d'indices liés à la ponctuation.

3 Repérage et exploitation de définitions en corpus

Nous détaillons ci-dessous le cadre de la constitution de nos deux corpus. Nous présentons ensuite le principe de fonctionnement du programme d'extraction, puis la mise au point de nos patrons lexico-syntaxiques, et enfin les modalités d'évaluation de ces patrons.

3.1 Constitution des corpus

Le premier corpus est centré sur le thème de la petite enfance, abordé suivant les disciplines de l'anthropologie et de la psychologie. Cette thématique était l'une de celles sélectionnées lors du projet Opales (`http://opales.ina.fr`), auquel nous avons participé. Les textes ont été collectés de deux manières : une partie du corpus a été constituée manuellement de notices documentaires (d'un volume de 5 Kmots), et une autre partie automatiquement à partir du web suivant les principes proposés par (Grabar N. et Berland S., 2001) et au moyen des programmes mis au point par les auteurs (71 Kmots, soit un total d'environ 76 Kmots). Le deuxième corpus concerne la diététique, notamment dans la perspective du maintien des personnes âgées en bonne santé. Il a été constitué automatiquement à partir des articles francophones référencés par le portail CISMeF (`http://www.chu-rouen.fr/cismef/`) sous les arborescences «diététique» et «nutrition», convertis au format texte. Il compte 480 Kmots. Les corpus sont analysés syntaxiquement par Cordial Analyseur (société Synapse – `http://www.synapse-sa.com/`), et le fichier produit associe chaque mot du corpus à son étiquette morphosyntaxique et à sa fonction grammaticale (sujet, objet, verbe recteur, etc.).

3.2 Méthode de repérage des informations sémantiques

Nous cherchons à présenter à un modéliseur d'ontologie des contextes et des éléments qui l'aident dans sa tâche. Notre outil propose pour cela une exploration ciblée du corpus textuel, dont il présente le résultat en tant que base de travail à l'utilisateur. Nous cherchons à repérer les énoncés définitoires au moyen de patrons lexico-syntaxiques, et à cibler à l'intérieur de ces énoncés l'élément défini : le definiendum. Dans un premier temps, nous avons observé que ce definiendum occupait deux places privilégiées. Nous avons donc pensé extraire les unités lexicales qui occupaient ces positions pour présenter à l'utilisateur non seulement l'énoncé définitoire, mais le definiendum concerné. Nous avons alors remarqué que les deux unités lexicales proposées dans ce cadre entretenaient des relations sémantiques intéressantes (figure 1), et nous avons tenté d'en proposer une typologie.
La première question qui se pose est celle du repérage des énoncés définitoires. Les patrons

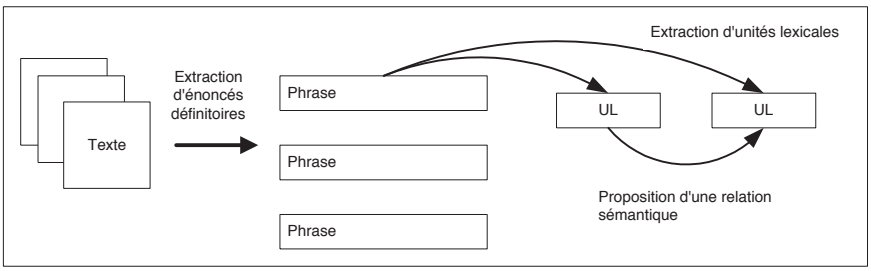

Figure 1: Suite des opérations réalisées à partir du corpus sur la base de patrons lexico-syntaxiques.

lexico-syntaxiques que nous employons pour cela s'articulent autour d'un ensemble de marqueurs, c'est-à-dire des mots ou expressions qui sont souvent révélateurs d'un énoncé définitoire : «défini comme», «c'est-à-dire», «emploie le terme de», etc. La finalité d'un patron lexico-syntaxique est de préciser les contextes lexicaux et syntaxiques dans lesquels un marqueur introduit bien l'une des relations sémantiques recherchées. Par exemple, autour du marqueur «comme», nous en avons défini trois :

- *SN* définir *MOTS** comme *SN*;
- *SN* comme *DET{1,3} SN*;
- *SN* comme (celuilceux|cellelcelles) *VERBE SN*.

Ces patrons peuvent utiliser des informations sur les formes (comme), lemmes (définir), catégories morphosyntaxiques (DET pour déterminant) et fonctions syntaxiques calculés précédemment. L'étoile signifie la présence facultative de l'élément et les accolades la possibilité de répéter une à trois fois l'élément la précédant. Nous limitons notre extraction d'énoncés définitoires au contexte de la phrase parce que c'était le plus souvent suffisant. Nous verrons dans les perspectives les traitements complémentaires que nous avons envisagés dans certains cas de figure contraires.
La deuxième question à traiter est celle du repérage des unités lexicales en relation. Nous avons défini pour cela deux types de stratégies :

- Pour les patrons possédant un marqueur verbal, on cherche à exploiter les fonctions syntaxiques associées à ce verbe : on extrait son sujet et son objet. Par exemple, dans le cas de «*Taouret dont le nom* ***signifie*** *La grande est conçue comme* [...]», les unités lexicales sujet et objet de signifier sont extraites : «*Taouret dont le nom*» et «*La grande*». Des contraintes supplémentaires s'appliquent au sujet : il doit avoir plus d'un mot (cela élimine les pronoms) et dans le cas où cette extraction fonctionnelle ne donne pas de résultat, la deuxième stratégie est également tentée;
- La seconde stratégie consiste en une exploration contextuelle : on extrait le groupe fonctionnel immédiatement à gauche et celui immédiatement à droite du marqueur, dans le cas précédemment cité ainsi que pour les marqueurs non verbaux. Par exemple dans la phrase : «[...]*les yeux qui ont une petite plaie* ***comme*** *une griffure (*[...]», «*ont une petite plaie*» et «*une griffure (*» sont extraits de manière contextuelle.

L'interface laissant la possibilité de corriger les unités lexicales proposées, nous avons essayé de proposer une séquence qui couvre au minimum l'unité lexicale intéressante, même si elle est entourée de mots parasites, à une extraction plus ciblée donnant de moins bons résultats.
La dernière question concerne la détermination des relations sémantiques. Chaque patron sert à détecter une relation sémantique entre les unités lexicales extraites, qui est proposée à l'utilisateur pour validation ou correction éventuelle. Par exemple, les patrons autour de la parenthèse permettent d'extraire des unités en relation de paradigme (des éléments d'un taxème n'étant ni synonymes ni antonymes, comme *une mère* et *un père*), d'hyperonymie (*le moïse (ou berceau)*) ou de définition «fonctionnelle» : «[...]*l'anthropologie (qui s'occupe surtout des contextes et des significations culturelles),* [...]». Cet ensemble de relations est proposé lors de l'extraction de phrases correspondant à ces patrons (voir figure 2), avec en tête de liste celle qui correspond spécifiquement à la forme syntaxique relevée autour de la parenthèse.
L'ensemble de la méthode est implémentée en XSLT, à l'aide du processeur XLST Xalan (`http://xml.apache.org/#xalan`), par souci de cohérence avec des choix antérieurs concernant le format du corpus (XML) et nos outils de gestion d'ontologie (éditeur DOE, (Troncy R. et Isaac A., 2002)) ; les patrons sont directement exprimés en XSLT. Le résultat du traitement est un formulaire HTML qui sert d'interface de validation à l'utilisateur. La figure 2 montre une copie d'écran de cette interface. Chaque rangée du tableau correspond à l'appariement d'un patron avec une phrase du corpus ; elle présente les deux unités lexicales extraites, la phrase concernée et la relation sémantique proposée entre les deux unités lexicales ; un lien permet de retourner au corpus pour y examiner la phrase dans son contexte d'origine, et la dernière colonne permet de valider l'association.

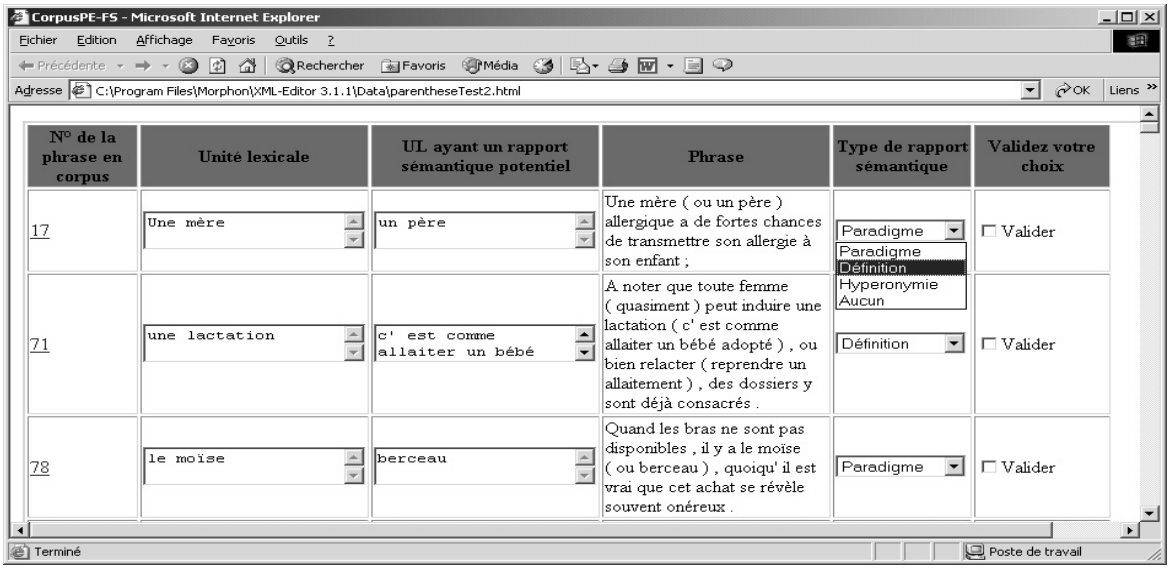

Figure 2: L'interface de visualisation.

3.3 Mise au point des patrons lexico-syntaxiques

Nous avons constitué une liste de marqueurs pour le repérage des énoncés définitoires à partir, notamment, des travaux de (Auger A., 1997), (Rebeyrolle J., 2000), et de (Fuchs C., 1994).

C'est en précisant leurs contextes d'usage que nous avons construit nos patrons lexico-syntaxiques, suivant la méthodologie de (Séguéla P. et Aussenac-Gilles N., 1999). La nécessité de leur adaptation en patrons liés au corpus tient à la polysémie des marqueurs lexicaux suivant le domaine décrit (par exemple «baptiser» peut être un marqueur de définition fiable, sauf dans le cas d'un corpus centré sur la petite enfance, où il désignera plutôt le baptême), et à une variabilité des formes syntaxiques qui sont plus ou moins complexes à décrire suivant le genre de documents pris en compte dans le corpus. Nous distinguons quatre types de marqueurs, définissant quatre groupes de patrons :

- Les marqueurs métalinguistiques à utiliser indépendamment (au nombre de 9) : appeler, baptiser, définir comme, dénommer, dénoter, désigner, nommer, signifier, vouloir dire;
- Les marqueurs métalinguistiques nominaux (11) : appellation, acception, concept, dénomination, désignation, expression, mot, nom, notion, terme, vocable, *à associer à un verbe support parmi* : appliquer, donner, employer, prendre, porter, recevoir, référer, renvoyer, réserver, utiliser;
- Les marqueurs lexicaux n'étant pas explicitement métalinguistiques, ou ceux de reformulation (21) : c'est-à-dire, en d'autres termes, soit, à savoir, en quelques sortes, une sorte de, enfin, il s'agit de, entendre par, vouloir dire, indiquer, comme, dit, par exemple, autrement dit, même chose que, équivaloir à, employer pour, marque, expliquer, préciser;
- Les ponctuations : parenthèses, guillemets et tirets d'incise sont également mentionnés dans la littérature. Nous nous sommes intéressés aux contextes définitoires autour de la parenthèse, et l'observation du premier corpus nous a permis de mettre au point quatre patrons synthétisant des contextes «intéressants» autour de cet indice de bas niveau. Nous nous sommes aperçus que, outre la définition, nous pouvions extraire des paradigmes (aide à la modélisation horizontale de l'ontologie) et des hyperonymes. Nous avons alors décidé d'inclure ces quatre schémas à l'évaluation de la méthode, même s'ils n'étaient pas exclusivement ciblés sur l'extraction de définitions.

Les relations sémantiques que l'on rencontre entre les deux unités lexicales extraites correspondent non seulement à de l'hyperonymie, mais aussi à d'autres relations : la définition ne s'exprime pas toujours suivant le canon aristotélicien de genre proche et différences spécifiques. On trouve également des définitions par synonymie («[...]*, et les représentations sociales des adultes (aussi appelées ethnothéories parentales)*[...]»), par exemplification («[...] *étudiés non pas en laboratoire ou dans des situations artificielles (comme les tests), mais dans les contextes quotidiens*.»), par composition («*Ici, les deux moments essentiels du déroulement du rituel que constituent la séparation et l'agrégation,* [...]») ou de type «fonctionnel» ([...]«*la psychologie (qui a tendance à ne s'intéresser qu'à l'individu)*»), etc. Nous avons envisagé les relations suivantes entre nos unités lexicales : hyperonymie, synonymie, antonymie, paradigme, ou «rapport fonctionnel» (correspondant à une relation transversale de l'ontologie : une relation qui n'est pas directement hiérarchisante, mais qui lie des termes de différentes branches).

3.4 Méthode d'évaluation des résultats

Les mesures classiques pour évaluer un tel outil sont la précision (la proportion d'extractions correctes parmi les résultats du système) et le rappel (la proportion d'extractions du système

parmi les résultats attendus selon un étalon de référence). Nous n'avons pas encore terminé la constitution de l'étalon de référence, et nous proposons donc à ce stade une évaluation de la précision. Celle-ci comporte deux parties : une première validation autour de l'énoncé global (énoncé définitoire ou véhiculant la relation sémantique prédite par le patron), et une autre concernant les unités lexicales extraites à partir de cet énoncé global.
La validation concernant l'énoncé global s'articule en deux points : d'une part est-il conforme à la relation sémantique annoncée, et, si ce n'est pas le cas, cet énoncé nous intéresse-t-il ? Certains patrons peuvent être sous-spécifiés par rapport au corpus, et extraire des phrases ne correspondant pas aux énoncés attendus. Mais si l'extrait s'avère être intéressant parce qu'il présente des relations structurantes et des unités lexicales propres à être intégrées à l'ontologie, nous ne le rejetons pas, mais essaierons dans un deuxième temps de créer deux patrons distincts et de gérer la nouvelle relation observée. Par exemple, le marqueur «signifier» permet d'extraire des énoncés définitoires et des phrases exprimant des implications liées au domaine, respectivement «[...] *le signe hiéroglyphique sa qui signifie la protection*.» et « *Les investissements dans le développement de la petite enfance pourraient signifier de meilleurs services* [...]». Ce deuxième type d'énoncé est «mis en réserve» pour créer deux patrons spécifiques à partir du même marqueur. Nous avons donc deux catégories dans notre évaluation : énoncé conforme à la relation sémantique annoncée (définition, hyperonymie, paradigme), ou énoncé intéressant à un autre titre.
La validation des unités lexicales (UL) extraites de la phrase s'articule également en deux points : l'UL proposée est-elle pertinente, et son extraction est-elle correcte ? C'est-à-dire qu'un premier niveau de jugement consiste à vérifier que les UL extraites sont bien celles que l'on cherche à avoir à partir de l'énoncé définitoire. Par exemple, dans la phrase : «*Le concept "éducation " est souvent défini de façon étroite, parfois même comme uniquement la scolarisation*.», il s'agit de trouver au minimum éducation et scolarisation. Un deuxième niveau concerne la délimitation de l'extrait proposé par rapport à l'UL intéressante. Trois cas de figure peuvent se produire : ou l'extrait correspond exactement à l'UL, ou celui-ci englobe l'UL (nécessitant un nettoyage manuel de la séquence), ou celui-ci ne contient pas toute l'UL (et il faut rajouter manuellement les parties manquantes). Nous considérons que l'extrait correspond exactement à l'UL même s'il compte l'article précédant le nom ou le groupe nominal constituant l'UL.

4 Application aux deux corpus

Nous avons appliqué notre méthode et les patrons au corpus sur la petite enfance (tableau 1) et sur la diététique (tableau 2). Les tableaux 1 et 2 présentent dans les premières colonnes le nombre de patrons lexico-syntaxiques correspondant à chacun des groupes définis en 3.3 : les patrons centrés sur des verbes métalinguistique (ligne Méta1), sur des éléments lexicaux métalinguistiques combinés : un terme et un verbe métalinguistique (ligne Méta2), sur des éléments lexicaux non spécifiquement métalinguistiques (ligne Ling.), et enfin sur des données de ponctuation (ligne Ponct.). Les colonnes suivantes présentent les nombres d'extraction de phrases correspondant à la relation sémantique prédite par le patron (définition, hyperonymie, paradigme), le nombre de phrases ne correspondant pas à la relation prédite mais jugées intéressantes (et servant de base pour raffiner le patron actuel suivant les différents cas). Les dernières colonnes concernent l'évaluation des unités lexicales : sont-elles correctes ou fausses, et, si correctes, sont-elles exactement extraites par nos programmes, incluses dans l'extrait proposé ou partiellement extraites ? Les données sont d'abord présentées sous la forme du nombre d'extraits validés, puis suivant leurs taux de «précision relative».

Type	Nb patrons	Nb extraits	Phrases suivant la relation sémantique	Taux de précision	Nb phrases intéressantes	Total	Précision des deux types d'énoncés
Méta1	7	32	16	50%	11	27	84%
Méta2	20	14	7	50%	4	11	78%
Ling	24	97	31+7 (2)	39%	34	72	74%
Ponct	4	79	22	27%	7+45 (1)	74	93%
Total	55	222	83	37%	101	184	82%

Type	Nb patrons	UL exacte	Taux de précision	UL incluse	Précision des deux types d'UL	UL incomplète	UL fausse	Total
Méta1	7	14	26%	14	88%	19	7	54
Méta2	20	2	9%	8	45%	3	9	22
Ling	24	6	4%	48	37%	6	84	144
Ponct	4	92	62%	29	81%	22	5	148
Total	55	114	31%	99	57%	50	105	368

Table 1: Extraction des données sur le corpus «petite enfance».

Au total, la précision des énoncés correspondant à la relation sémantique (précision «stricte») est de 37 % et celle des unités lexicales est de 31 %. Si l'on prend en compte également les phrases «intéressantes», cette précision s'élève à 63 %. De plus, 45 (correspondant au (1)) des phrases repérées par l'un des patrons de ponctuation apportaient une relation de traduction entre les deux unités lexicales ; cette relation n'était une hyperonymie, mais une reformulation dans une autre langue que l'on peut également considérer comme intéressante, amenant alors la précision à 82 %.

Les patrons des deux groupes métalinguistiques obtiennent la meilleure précision stricte (50 %), mais un nombre absolu d'énoncés plus faible (23 au total contre 32 pour le groupe Ling et 22 pour le groupe Ponct). On retrouve une opposition classique entre rendement et précision. Cette observation peut aussi être liée au fait que les textes ne relevant pas de pratiques terminographiques auraient plus souvent recours à des gloses de reformulation pour introduire des énoncés définitoires, et ces gloses sont repérées par nos patrons des groupes Ling et Ponct.

Concernant l'évaluation du système d'extraction, nous pouvons noter que certaines des unités lexicales incomplètes ne pouvaient être complétées dans le corpus par notre stratégie d'extraction parce que les données manquantes étaient implicites ou suivaient un principe de rattachement complexe. Sinon, dans le cas de 7 énoncés (2), la compréhension globale du sens de la définition nécessitait également la lecture de la phrase précédente.

Une première remarque doit précéder l'examen du tableau 2 : la conversion en format texte de certains des documents HTML du corpus a causé la segmentation excessive de certaines phrases. Des énoncés pertinents ont ainsi été coupés, empêchant les patrons de trouver les unités lexicales correctes et retournant des énoncés trop incomplets pour permettre une validation. Ces énoncés ont été considérés comme faux, faisant chuter de manière drastique le taux de précision, qui n'atteint que 16% de précision «stricte». Un choix de l'étiquetage a également induit un bruit non négligeable dans l'extraction des phrases autour de la parenthèse. Dans le cas des références de type (A), A est étiqueté comme nom commun et la forme SN (NomCommun) étant un des patrons, des extraction incorrectes ont été générées.

Type	Nb patrons	Nb extraits	Phrases suivant la relation sémantique	Taux de précision	Nb phrases intéressantes	Total	Précision des deux types d'énoncés
Méta1	7	29	20	69%	2	22	75%
Méta2	20	10	6	60%	1	7	70%
Ling	24	307	55	18%	15	70	22%
Ponct	4	365	38	10%	30+39 (3)	107	29%
Total	55	711	119	16%	87	206	29%

Type	Nb patrons	UL exacte	Taux de précision	UL incluse	Précision des deux types d'UL	UL incomplète	UL fausse	Total
Méta1	7	14	31%	4	41%	12	14	44
Méta2	20	4	28%	0	28%	6	4	14
Ling	24	16	11%	32	34%	16	76	140
Ponct	4	100	46%	56	73%	13	45	214
Total	55	134	32%	92	54%	47	139	412

Table 2: Extraction des données sur le corpus «diététique».

Nous avons à nouveau été confrontés au problème lié à la prédiction de la relation sémantique d'hyperonymie (3) : les extraits du corpus correspondaient cette fois à une expansion d'acronymes, qui peut être considéré comme une sorte de définition. Nous pouvons par ailleurs remarquer que la qualité de l'extraction des UL dépend de la complexité syntaxique du patron. Plus il est de bas niveau plus les informations contextuelles sont efficaces pour l'extraction, plus il est «linguistique», plus des informations de dépendance fonctionnelle entrent en jeu, compliquant la tâche.

5 Conclusion et perspectives

Nous avons présenté une méthode ciblant en corpus des énoncés définitoires ou intéressants dans une perspective de construction ontologique. L'évaluation de cette méthode nous a permis de soulever certains points, notamment la difficulté d'avoir des taux de précision élevés lorsque l'on s'intéresse à des marqueurs linguistiques de reformulation plutôt que des unités lexicales métalinguistiques, remarque également formulée dans (Rebeyrolle J., 2000). Nous avons également extrait des énoncés nécessitant la lecture de la phrase précédente pour en spécifier le sens. Dans certains de ces cas, la première des UL renvoyées était un pronom (personnel ou démonstratif), ce qui nous amène à envisager le traitement complémentaire suivant : extraire la phrase précédant immédiatement l'énoncé global dans la cas où la première UL renvoyée est strictement égale à un de ces pronoms.
Lorsque la méthode d'extraction sera optimisée, il faudrait idéalement pouvoir proposer à l'utilisateur une forêt de nIJuds ontologiques correspondants aux choix de validation. Nous nous penchons sur la question de la faisabilité d'une structuration en termes de nIJuds ontologiques des résultats validés : organiser les couples d'unités lexicales en fonction de la relation sémantique validée.

Références

AUGER A. (1997), *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*, Thèse de doctorat, Université de Neuchâtel.

BOURIGAULT D., LAME G. (2002), Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit, *Traitement Automatique des Langues*, Vol. 43(1), 129-150.

CONDAMINES A. (1993), Un exemple d'utilisation de connaissances de sémantique lexicale : acquisition semi-automatique d'un vocabulaire de spécialité, *Cahiers de lexicologie*, Vol. LXII, 25-65.

FUCHS C. (1994), *Paraphrase et énonciation*, Paris, Ophrys.

GRABAR N., BERLAND S. (2001), Construire un corpus web pour l'acquisition terminologique, Actes de *4e rencontres Terminologie et Intelligence Artificielle (TIA 2001)*, Nancy, 44-54.

HEARST M. (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, In Proceedings of *15th International Conference on Computational Linguistics (COLING 1992)*, Nantes, 539-545.

LE PRIOL F. (2001), Identification, interprétation et représentation de relations sémantiques entre concepts, Actes de *Xe conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, Tours, 379-384.

MURESAN S., KLAVANS J. L. (2002), A method for automatically building and evaluating dictionary resources, In Proceedings of *the Language Resources and Evaluation Conference (LREC 2002)*, Las Palmas, Spain.

PEARSON J. (1998), *Terms in context*, Amsterdam/Philadelphia : John Benjamins Publishing Company.

REBEYROLLE J. (2000), *Forme et fonction de la définition en discours*, Thèse de doctorat, Université Toulouse II - Le Mirail.

REBEYROLLE J., CONDAMINES A. (2000), , Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results, In M.-C. L'Homme, C. Jacquemin et D. Bourigault (Eds) *Recent Advances in Computational Terminology*, John Benjamins, 127-148.

SEGUELA P., AUSSENAC-GILLES N. (1999), Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, Actes de *la Xe conférence Ingénierie des Connaissances (IC 1999)*, Palaiseau, 79-88.

SZULMAN S., BIEBOW B., AUSSENAC-GILLES N. (2002), Structuration de Terminologie à l'aide d'outils de TAL avec TERMINAE, *Traitement Automatique des Langues*, Vol. 43(1), 103-128.

TRONCY R. et ISAAC A. (2002), DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies, Actes de *13e Journées Francophones d'Ingénierie des Connaissances (IC'02)*, Rouen, 63-74.