Annotation sémantique et validation terminologique en texte intégral en SHS

Résumé. Nos travaux se focalisent sur la validation d'occurrences de candidats termes en contexte. Les contextes d'occurrences proviennent d'articles scientifiques des sciences du langage issus du corpus SCIENTEXT[1]. Les candidats termes sont identifiés par l'extracteur automatique de termes de la plate-forme TTC-TermSuite et sont ensuite projetés dans les textes. La problématique générale de cet article est d'étudier dans quelle mesure les contextes sont à même de fournir des critères linguistiques pertinents pour valider ou rejeter chaque occurrence de candidat terme selon qu'elle relève d'un usage terminologique en sciences du langage ou non (langue générale, transdisciplinaire, autre domaine scientifique). Pour répondre à cette question, nous comparons deux méthodes d'exploitation (l'une inspirée de la textométrie et l'autre de Lesk) avec des contextes d'occurrences du même corpus annotés manuellement et mesurons si une annotation sémantique des contextes améliore l'exactitude des choix réalisés automatiquement.

Abstract. Our work is in the field of the validation of term candidates occurrences in context. The textual data used in this article comes from the freely available corpus SCIENTEXT. The term candidates are computed by the platform TTC-TermSuite and their occurrences are projected in the texts. The main issue of this article is to examine how contexts are able to provide relevant linguistic criteria to validate or reject each occurrence of term candidates according to the distinction between a terminological and a non terminological use (general language, transdisciplinary use, use coming from another science). To answer this question, we compare two methods (a textometric one and another inspired from Lesk) with the manual annotation of the same corpus and we evaluate if a semantic annotation of contexts improves the accuracy of the choices made automatically.

Mots-clés : Annotation sémantique - extraction et désambiguïsation terminologique - textométrie - texte intégral
Keywords: Semantic Annotation - Terminological Extraction and Disambiguation - Textual Metrics (Specificity) - Full Text

1 Introduction

Nos travaux se situent dans le champ de l'extraction terminologique à partir de textes intégraux dans le domaine des SHS et plus précisément celui des sciences du langage. A la suite de (Daille 1994), (Toussaint et *al.* 1998), (Bourigault et Slozdian 1999), (Bourigault et. *al* 2001) parmi beaucoup d'autres, nous privilégions une approche allant du texte (réalisations linguistiques de termes dans les textes) aux termes (objets conceptuels). Accéder aux réalisations linguistiques des termes dans les textes suppose de les reconnaître comme telles. Parmi les travaux qui abordent cette problématique, une première partie s'appuie sur l'utilisation d'extracteurs automatiques de candidats termes qui sont ensuite validés par des experts des domaines de spécialités concernés : *Acabit* (Daille 1994 ; 2003), *Yatea* (Aubin et Hamon 2006), *TermoStat* (Drouin 2003) ou encore la plate-forme *TTC-TermSuite* (Daille et *al.*, 2011). D'autres travaux, qui peuvent être connexes, s'intéressent à la validation, l'extraction de candidats termes ou de relations entre termes dans les textes, en mettant en œuvre différentes exploitations des textes dans une approche distributionnelle. Les travaux de Daille (2003), Toussaint et *al.* (1998), Namer et Zweigenbaum (2004) ou M-C L'Homme (2004a) s'appuient sur des connaissances relevant de la morphologie dérivationnelle ou constructionnelle. Les travaux de Baneyx et *al.* (2005), Jacques et Aussenac-Gilles (2006), Aussenac-Gilles et Condamines (2009), Manser (2012), Périnet et Hamon (2013) détectent et exploitent des patrons lexico-syntaxiques pour l'identification de relations entre (réalisations linguistiques de) termes. Enfin, les travaux de Grabar et Zweigenbaum (2004), Claveau et M-C L'Homme (2005), Poibeau (2005) ou Condamines et Péry-Woodley (2007) reposent sur l'utilisation de structures sémantiques, textuelles ou discursives.

Notre objectif est d'analyser des contextes d'occurrences de candidats termes qui sont sémantiquement enrichis afin de sélectionner automatiquement les occurrences de candidats qui relèvent d'un usage terminologique et de rejeter les autres. Autrement dit, nous procédons à un type particulier de désambiguïsation sémantique que nous appellerons « désambiguïsation terminologique ». En effet, on peut constater, à la suite de M.C. L'Homme (2004b), que même si le terme en tant qu'étiquette de concept, pour une terminologie donnée et une application définie, n'est pas ambigu, ses

[1] Scientext permet d'accéder à un outil d'interrogation pour l'ensemble de la base de textes du projet. Une partie de la base est accessible est sous licence Creative Common : http://scientext.msh-alpes.fr/scientext-site/?article8 [pages consultées le 12/02/2014]

réalisations linguistiques peuvent l'être. Ceci est vrai en particulier lorsque les « termes » sont des « candidats termes » extraits automatiquement par une plate-forme d'extraction terminologique. Ce phénomène pourrait aussi se manifester lorsque les termes dont on observe les occurrences en texte intégral appartiennent à des thésaurus ou des référentiels terminologiques.

– ambiguïté avec le lexique ou la phraséologie transdisciplinaire : *argument, corpus, définition, énoncé, exemple, objet, référence*
  [+term] *En chemin, nous avons souligné la grande flexibilité des SN définis pluriel, qui en fait le lieu possible d'une négociation de la* ***référence*** *et de la désignation* (Figures et référence plurielle en corpus journalistique - Lecolle M. (2000). Cahiers de grammaire (25))
  [-term] *Les auteurs font* ***référence*** *à [...]* (Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de texte - Piérard S. et Begsten Y. (2007). TAL(47/2))

– ambiguïté avec un autre domaine de spécialité : *patient*
  [+term] *[...] ou plus rarement, à des rôles argumentaux (agent,* ***patient****, objet,...) [...]* (Les relations sémantiques : du linguistique au formel - Aussenac-Gilles N. et Séguéla P. (2000). Cahiers de grammaire (25))
  L[-term] *es* ***patients*** *cérébrolésés [...]* (Nouveaux habits de la lexicographie spécialisée : Intégration de la métaphorique dans le dictionnaire du football - Leroyer P. et Moller B. (2004). EURALEX)

– ambiguïté avec un emploi lexical ou phraséologique de langue générale : *argument, définition, énoncé, expression, objet*
  [+term] *[...] les* ***expressions*** *du type le jour suivant.* (Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de texte - Piérard S. et Begsten Y. (2007). TAL(47/2))
  [-term] ***L'expression*** *de telle ou telle relation [...]* (Variabilité des outils de TAL et genre textuel : cas des patrons lexico-syntaxiques - Jacques M.-P. et Aussenac-Gilles N. (2006). TAL (47))

Comme le montrent ces quelques exemples de réalisations linguistiques de candidats termes, c'est le contexte au sens large qui nous permet de sélectionner les occurrences relevant d'un emploi terminologique. Deux interrogations apparaissent alors : en ce qui concerne les différentes manières d'exploiter les contextes d'occurrences des candidats et la sélection des informations à exploiter dans ces contextes. Dans cet article, nous développons une approche qui s'appuie, d'une part, sur une analyse statistique des contextes d'occurrences, analyse fondée sur le calcul de spécificité (Lafon 1980), et d'autre part, sur des contextes annotés sémantiquement à l'aide de traits sémantiques ou quasi-sèmes qui sont extraits de deux ressources lexicales, le TLFi et WiktionnaireX. L'approche présentée est ensuite comparée à l'algorithme de Lesk (1986).

2 Données et méthodes

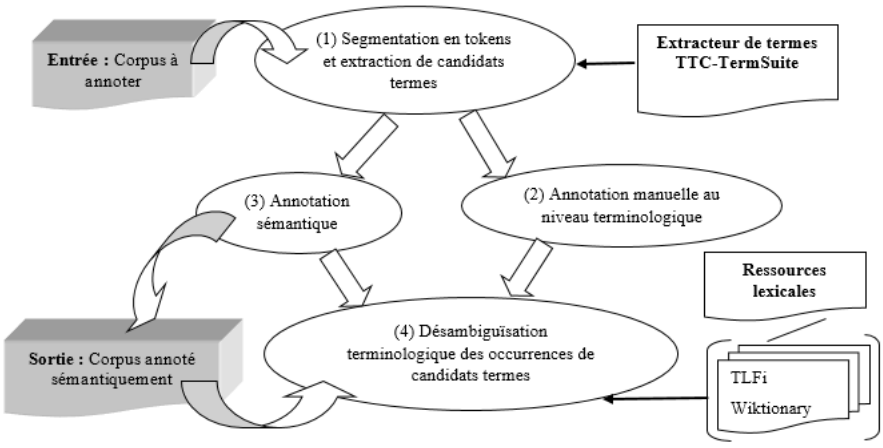

Figure 1: Méthodologie générale de l'expérience

La plate-forme TTC-TermSuite[2] segmente le corpus de travail en tokens, l'annote avec TreeTagger puis en extrait une terminologie, c'est-à-dire une liste de candidats termes qui est projetée dans les textes. Le corpus enrichi en candidats

[2] La plate-forme TTC-TermSuite est librement utilisable et open source. Elle est accessible sous licence Apache 2.0 https://code.google.com/p/ttc-project/downloads/detail?name=ttc-term-suite-1.4.jar [page consultée le 12/02/2014]

termes subit ensuite deux traitements parallèles. D'un côté, il est pris en charge dans une plate-forme d'annotation manuelle au sein de laquelle un annotateur linguiste expert désambiguïse manuellement les occurrences de candidats termes en les validant ou les rejetant d'un point de vue terminologique. A l'issue de l'annotation, les occurrences désambiguïsées de candidats termes sont considérées comme des occurrences de termes. Cet enrichissement fournit une version du corpus qui sert de corpus de référence. D'un autre côté, le corpus enrichi automatiquement en candidats termes est pris en charge par le module d'annotation sémantique. Le nouveau corpus obtenu, enrichi en candidats termes et en annotations sémantiques, est ensuite pris en charge par le module de désambiguïsation terminologique dont la tâche est de sélectionner pour chaque candidat ses occurrences terminologiques et de rejeter ses occurrences non terminologiques. Enfin, nous évaluons les performances de la désambiguïsation terminologique réalisée automatiquement en la comparant avec celle effectuée manuellement dans la plate-forme d'annotation manuelle.

2.1 Corpus de travail, enrichissement terminologique et corpus de référence

Le corpus de travail rassemble 62 articles appartenant au domaine scientifique des sciences du langage, extraits de la base Scientext. Ce corpus, au format xml-tei, comporte 397 695 occurrences. L'ensemble des textes se répartit en 47 articles de conférences (75,81% des documents et 57,06% des occurrences), et 15 articles de revues (24,19% des documents et 42,94% des occurrences). Ainsi, en nombre d'occurrences, le corpus utilisé est assez équilibré entre conférences et revues[3]. Le corpus de travail est traité par l'extracteur automatique de termes TTC-TermSuite afin d'obtenir une liste de candidats termes. Quatre paramètres définissent le filtrage de la liste des candidats termes en sortie : (1) regroupement par variantes flexionnelles et syntaxiques (la distance choisie est le *Log-likelihood ratio*) ; (2) seuil minimal de fréquence fixé à 5 ; (3) sélection des 7 500 (maximum) premiers candidats termes triés par spécificité décroissante[4] ; (4) ne sont considérés que les candidats nominaux et adjectivaux, simples et complexes. Un module de traitement interne projette ensuite les candidats termes dans les textes et les encapsule dans des balises XML tout en gérant les chevauchements tels que ceux que l'on trouve dans *graphes de dépendances sémantiques* entre *graphes de dépendances* et *dépendances sémantiques*. Le module de projection crée une nouvelle version du corpus enrichie en candidats termes.

Le corpus de référence est constitué d'une sélection de 52 documents extraits du corpus de travail enrichi en occurrences de candidats termes qui ont été évaluées manuellement du point de vue de leur caractère terminologique en contexte. Cette tâche a été réalisée au sein d'une interface d'annotation librement consultable[5] : les occurrences de candidats termes sont bornées par des crochets et leur empan est matérialisé par un surlignage dynamique. Une puce dont la couleur varie entre le vert et le rouge, cliquable dynamiquement représente les choix finaux de l'annotateur (le vert correspond à une validation, le rouge à un rejet). A l'issue de l'annotation manuelle, les évaluations sont stockées et décomptées. Parmi les 50 993 occurrences de candidats termes, correspondant à 4 431 candidats termes différents, 14 544 occurrences sont validées, soit 28,52 %. Le décompte des choix permet une classification des candidats sur deux échelles qui seront utilisées pour caractériser le jeu de test des expériences (section 3) :

- une échelle d'ambiguïté allant de 0 à 50 : nous avons opté pour cet intervalle de [0,50] suite au ratio (exprimé en pourcentage) « nombre d'occurrences validées / nombre total d'occurrences ». Plus il se rapproche de 50, plus le candidat est jugé ambigu ; pour les termes qui ont un ratio supérieur à 50, nous avons opté pour le complémentaire du ratio (100 - ratio).
- une échelle représentant la tendance terminologique du candidat (0-100): plus le ratio « nombre d'occurrences validées / nombre total d'occurrences » se rapproche de 100, plus le candidat a une tendance terminologique forte et inversement si ce ratio se rapproche de 0.

2.2 Enrichissement sémantique des données

Habituellement, lorsqu'on parle d'annotation et en particulier d'annotation sémantique, l'objectif est d'associer « une interprétation stabilisée » aux données brutes (Habert 2005). L'annotation sémantique dans cette perspective suppose la désambiguïsation. Dans l'expérience que nous menons, nous faisons le choix de procéder à un enrichissement sémantique ambiguë qui ne privilégie aucun des sens fournis par les ressources utilisées pour l'annotation. Nous

[3] Les conférences représentées sont le Cédil (Colloque international des Étudiants chercheurs en Didactique des Langues et en Linguistique), Euralex (Conférence de European Association of Lexicographie) et le colloque EID (Émotions, Interactions, Développements). Les revues sont TAL (Traitement automatique des langues), Les cahiers de grammaire et la revue LiDil (Revue de Linguistique et de Didactique des langues).

[4] Ce seuil de 7 500 a été déterminé de manière empirique par comparaison avec la distribution des types de structures de candidats complexes que l'on obtient sans filtrage.

[5] La page https://apps.atilf.fr/smarties/ [page consultée le 12/02/2014] permet un accès public en consultation. Le guide d'annotation, disponible sur le site, permet de comprendre comment est effectuée l'évaluation des candidats termes depuis leur état initial vers l'état de leur validation terminologique.

plaçant dans le cadre de la sémantique interprétative, l'une des hypothèses majeures sur laquelle nous nous appuyons est que le sens d'une unité lexicale est en grande partie co-construit par son contexte d'usage ou totalement construit pour le sens émergeant des usages[6]. C'est donc dans un deuxième temps (section 2.3 ci-dessous) que nous analysons l'information sémantique ajoutée en vue d'un type particulier de désambiguïsation sémantique, à savoir la désambiguïsation terminologique des occurrences de candidats termes apparaissant dans les contextes enrichis sémantiquement. Cette méthodologie permet de mesurer le caractère opératoire ou non du type d'annotation sémantique que nous appliquons lors du processus de désambiguïsation terminologique.

Comme nous l'avons précisé dans l'introduction (section 1), les informations sémantiques ajoutées aux unités lexicales sont des traits sémantiques ou quasi-sèmes. À la suite de Valette et *al.* (2006), nous faisons l'hypothèse que les mots pleins de toutes les définitions d'un mot vedette d'un dictionnaire représentent des traits sémantiques bruts associés à tous les sens possibles de ce mot vedette. Les traits sémantiques sont extraits sous forme lemmatisée et catégorisée en parties du discours *(adv*, adverbes ; *adj*, adjectifs ; *subst*, noms ; *v*, verbes). Ainsi, le nom *locuteur*, qui a pour définition dans le TLFi *Personne qui parle, qui produit des énoncés* sera représenté par l'ensemble de traits sémantiques {"*personne":subst, "parler":v, "produire":v, "énoncé":subst*}. Lorsque le mot vedette correspondant à une unité lexicale du texte à annoter a plusieurs définitions dans le dictionnaire, celles-ci sont ajoutées les unes aux autres. Enfin, lorsqu'un trait sémantique apparaît plusieurs fois, il est répété afin de tenir compte de toutes ses occurrences. C'est le cas du trait *note* dans l'une des définitions du verbe *annoter* dans le TLFi *"pourvoir un texte de notes" "mettre des notes en marge de ..."* : {*"pourvoir":v, "texte":subst, "note":subst, "mettre":v, "note":subst, "marge":subst*}.

La première ressource que nous utilisons est le TLFi car il s'agit d'une ressource libre sur signature d'une convention, à large couverture, comportant des définitions lemmatisées et catégorisées et disposant d'une structuration XML permettant l'extraction de ces définitions. Cependant, cette ressource n'est actuellement plus mise à jour et il est nécessaire de la compléter. La mise en ligne du WiktionnaireX par F. Sajous[7] permet d'étendre la couverture de la ressource initiale de manière significative. Nous avons sélectionné dans le WiktionnaireX (structuré en XML) les entrées de formes lemmatisées conformes au format des documents à annoter sémantiquement (les textes sont lemmatisés et annotés en parties du discours par TreeTagger via l'extracteur de candidats termes TTC-TermSuite). Pour catégoriser les traits sémantiques des définitions du WiktionnaireX, nous avons projeté les définitions catégorisées du TFLi. Nous obtenons ainsi deux ressources de traits sémantiques.

Pour choisir le mode d'annotation mis en œuvre dans les expériences (section 3), nous avons cherché à maximiser la couverture de l'annotation sémantique en utilisant les deux ressources de traits sémantiques (celle issue du TLFi prioritairement et celle issue du WiktionnaireX pour compléter). Les six types d'annotations qui ont été définis et dont le taux de couverture est représenté ci-dessous (figure 2) ont été définis en fonction de deux paramètres :

1. la ressource de traits sémantiques utilisée (TLFi exclusivement ; WiktionnaireX exclusivement ; TLFi complété par le WiktionnaireX) ;
2. les critères d'identification des entrées lexicales utilisées pour l'annotation des unités lexicales dans le textes
   - V1 = double correspondance établie à la fois sur la forme lemmatisée et la catégorie grammaticale
   - V2 = correspondance simple établie uniquement sur la forme lemmatisée

[6] On peut citer l'exemple bien connu aujourd'hui de l'usage néologique du nom *caviar* dans le discours journalistique principalement sportif. Dans ce genre textuel bien précis et pour le football en particulier, le nom *caviar* désigne une très belle passe comme l'atteste cet exemple repris de Rastier et Valette (2009 : 14) : *Confirmation ici, d'un centre précis, [David Beckham] trouvait la tête de Frank Lampard qui n'avait plus qu'à régler la mire pour ouvrir la marque et transformer ce* ***caviar*** *en but.* (Site sports.fr, 14.06.2004) .

[7] Cette ressource libre et open source est accessible à http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html [page consultée le 04/02/2014]. Elle a notamment été utilisée dans (Sajous et *al.* 2013).

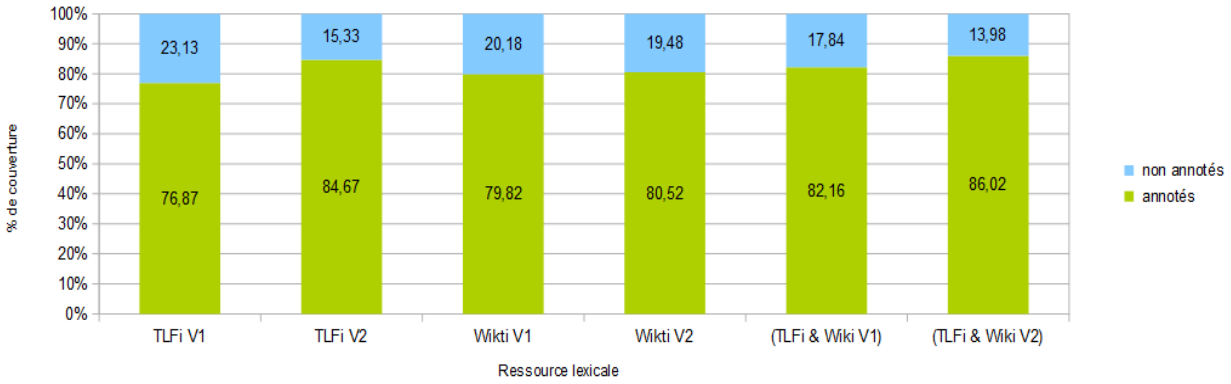

Figure 2: Couverture d'annotation de l'ensemble des mots du corpus Scientext à l'aide de différentes ressources lexicales

L'annotation (TLFi & WIKI V2) atteint une couverture de 86,02 %, elle représente le ratio entre le nombre de mots pleins de toute catégorie annotés par rapport au nombre total de mots pleins.
Pour conclure sur ce point, le tableau (1) ci-dessous détaille les taux de couverture obtenus par catégorie d'unité lexicale[8] et par type d'annotation. Là encore, le type (TLFi & Wiki V2) semble le plus satisfaisant et c'est donc ce type d'annotation qui a été choisi pour annoter le corpus de travail en traits sémantiques.

	Noms (%)		Verbes (%)		Adjectifs (%)		Adverbes (%)	
	Tokens	Types	Tokens	Types	Tokens	Types	Tokens	Types
TLFi V1	81,88	51,28	86,01	66,37	70,2	52,6	34,73	72,82
TLFi V2	82,31	52,21	95,2	67,53	75,45	55,24	85,33	91,02
Wiki V1	80,4	52,2	86,27	68,37	66,5	51,72	85,43	90,05
Wiki V2	80,74	53,27	86,43	69,08	68,25	54,3	87,64	94,17
TLFi & Wiki V1	83,27	54,37	86,35	68,73	71,77	55,48	85,43	90,05
TLFi & Wiki V2	**83,52**	**55,13**	**95,64**	**70,28**	**76,43**	**57,7**	**91,12**	**97,09**

Tableau 1: Résultat global d'annotation (couverture) du corpus Scientext à l'aide de différentes ressources lexicales

2.3 Désambiguïsation terminologique des occurrences de candidats termes

Par rapport aux champs d'application et aux difficultés relevés dans la problématique de la désambiguïsation sémantique en général (Navigli 2009), (Navigli et Lapata 2010), (Schwab et al. 2013), réaliser une désambiguïsation terminologique pourrait nous placer dans une configuration plus favorable : (1) seuls les candidats termes sont désambiguïsés et non l'ensemble des unités lexicales des textes ; (2) les « sens » entre lesquels arbitrer forment un ensemble borné et constant (usage terminologique vs. usage non terminologique).

Le type de désambiguïsation mis en œuvre appartient au champ des méthodes probabilistes supervisées. Suite à des travaux antérieurs qui ont donné des résultats encourageants (Camacho-Collados et *al.* 2014), nous utilisons une méthode basée sur la spécificité lexicale (Lafon 1980) dans laquelle l'indice de spécificité de chaque mot du corpus représente le sur-emploi ou sous-emploi des mots dans le sous-corpus par rapport au corpus de référence. Cette méthode est qualifiée de *méthode de désambiguïsation textométrique*. Le calcul de spécificité est utilisé pour établir, à partir du corpus de référence, des profils caractéristiques des deux types d'usage que l'on cherche à différencier. D'autres méthodes font usage exclusivement de ressources lexicales (sans utiliser un corpus de référence) et, plus spécifiquement de dictionnaires. On trouve parmi ces méthodes l'algorithme de Lesk (1986). Dans cette approche, un score est calculé pour la définition terminologique et un autre score pour les définitions relevant de la langue générale. Chaque score est déterminé à partir de l'intersection des mots d'une définition (terminologique ou non) avec les mots du contexte proche du candidat terme à désambiguïser. Étant donné la simplicité de l'algorithme, les résultats obtenus par Lesk sont intéressants en ce qu'ils permettent une comparaison moyennant une adaptation marginale. L'adaptation

[8] Les résultats entre types et tokens s'inversent pour les adverbes parce que les adverbes non annotés ont une fréquence d'apparition plus importante que ceux qui sont annotés. Par exemple, les séquences « a priori, ne, est-à-dire » apparaissent plusieurs fois et n'ont pas d'entrée dans le TLFi ou le le WiktionnaireX. En utilisant l'annotation (TLFi & Wiki V2), nous sommes parvenus à annoter 400 types parmi 412, alors que seulement 12899 tokens sont annotés parmi 14156.

porte sur le contexte considéré autour du mot à désambiguïser : au lieu du texte complet, nous prenons en compte le paragraphe. En effet, un texte peut contenir plusieurs occurrences d'un même candidat terme avec un comportement différent du point de vue terminologique, ce qui est beaucoup moins vrai du paragraphe.

2.3.1 Désambiguïsation textométrique

Cette première méthode de désambiguïsation, qui s'appuie sur les travaux antérieurs de Camacho Collados et al. (2014), est réalisée en deux temps : (1) établissement des profils caractérisant les usages à différencier ; (2) utilisation de ces profils pour décider, contexte par contexte, si l'occurrence analysée est terminologique ou non. Les profils terminologiques vs. non terminologiques sont établis à partir du corpus de référence en rassemblant pour chaque candidat dans deux ensembles disjoints les contextes où ce candidat relève d'un usage terminologique et les contextes où il relève d'un usage non terminologique. Chacun de ces ensembles constitue un sous-corpus par rapport au corpus de référence. Pour chaque candidat, on dispose donc d'un sous-corpus terminologique (SC_{on}) et d'un sous-corpus non terminologique (SC_{off}). Dans cette configuration, en appliquant à chaque sous-corpus (SC) un algorithme de calcul du taux de spécificité dont les résultats ont été vérifiés par comparaison avec ceux obtenus à l'aide du logiciel textométrique TXM (Heiden 2010), nous produisons une liste d'éléments supposés caractéristiques d'un usage terminologique (appelée LS_{on}) et une liste d'éléments supposés caractéristiques d'un usage non terminologique (appelée LS_{off}). Tous les éléments présents dans chacune de ces listes (LS) sont représentés avec leur taux de spécificité[9]. Pour créer les profils de traits sémantiques (LST), on répète le processus en remplaçant chaque mot plein par ses traits sémantiques. Dans le cas des contextes annotés en traits sémantiques, les profils établis sont désignés par LST_{on} et LST_{off} et ils sont construits à partir des sous-corpus disjoints annotés en traits sémantiques appelés respectivement SCT_{on} et SCT_{off}. Le tableau (2) donne une sélection des profils calculés pour le candidat *patient*.

LS_{on}		LST_{on}		LS_{off}		LST_{off}	
agent	9,51	*lar*	16,61	**douleur**	48,48	*vérifier*	123,20
planter	9,31	*eupatoriées*	10,33	**cérébrolésés**	30,06	*examen*	78,39
planteur	5,91	*dot*	9,36	**contrôle**	26,47	*pénible*	59,05
adventice	5,43	*enfoncer*	8,59	**récit**	20,68	*plaindre*	38,78
schème	4,73	*debout*	8,19	**trouble**	19,18	*poinçonner*	26,72
médecin	4,73	*primitif*	8,12	**maladie**	17,43	*platiner*	26,72
sen|sens	4,55	*actionnaire*	8,09	**adolescent**	14,48	*fisc*	26,72
primitif	4,17	*angleterre*	7,41	**frontal**	13,64	*physique*	26,71
conséquent	3,07	*agrément*	5,81	**dardier**	12,91	*examiner*	24,89
...		...		...		...	
patient (peu ambigu, peu terminologique)							

LS_{on}		LST_{on}		LS_{off}		LST_{off}	
temporel	19,09	*marquer*	35,72	**talmy**	12,98	*température*	40,43
chance	13,85	*grille*	34,54	**trajectoire**	10,67	*chaleur*	33,40
marqueur	12,39	*communiquer*	30,38	**typologie**	10,24	*place*	18,92
thème	11,99	*impact*	28,49	**cri**	10,08	*sensiblement*	18,67
bac	11,46	*choeur*	28,48	**slobin**	10,02	*typologie*	18,29
rupture	10,77	*balustrade*	24,47	**tunisien**	9,25	*sud*	13,28
thématique	9,73	*inscrire*	23,56	**interjection**	8,92	*sibérie*	12,48
émotionnel	9,11	*but*	22,55	**encoder**	8,87	*cordialité*	11,10
positif	8,81	*point*	20,34	**click**	8,59	*peur*	10,98
...		...		...		...	
expression (très ambigu)							

Tableau 2 : Profils caractéristiques des contextes terminologiques ou non terminologique extraits des contextes d'occurrences lexicaux ou enrichis en traits sémantiques[10]

[9] Le taux de spécificité de Lafon (1980) est calculé en comparant sur une distribution observée dans un corpus de référence et une distribution théorique d'un échantillon définie selon une loi hypergéométrique, : cette fréquence théorique de chaque mot est proportionnelle à la fréquence de ce mot dans le corpus. Le signe du taux de spécificité d'un élément est positif si sa fréquence observée est supérieure à sa fréquence théorique. La valeur d'un taux de spécificité se déduit de la probabilité d'obtenir la fréquence observée. (Reutenauer 2012 : chapII.2) donne une description détaillée de cet indice et le compare avec d'autres indices statistiques.

[10] Les éléments représentés dans les profils donnés ont fait l'objet d'une sélection à des fins de lisibilité : les 9 éléments indiqués sont les plus spécifiques de chaque profil et n'appartiennent pas à l'intersection des profils On x Off. Dans le calcul actuel, cette sélection n'est pas encore mise en œuvre et la différenciation des éléments communs s'appuie sur les taux de spécificité qui sont toujours très différents, par exemple *subir* avec le candidat *patient* qui a un taux de 25 en *ON* et un taux limite de 1000 en *OFF*.

Bien que le calcul des profils suive la même méthode que ceux-ci soient établis à partir des contextes lexicaux ou à partir des contextes enrichis en traits sémantiques, il existe deux différences. Les profils "lexicaux" sont constitués d'unités lexicales spécifiques des contextes et le seuil minimal du taux de spécificité est fixé à 1,5. Les profils "sémantiques" sont constitués de traits sémantiques et le seuil minimal du taux de spécificité est fixé à 2. Grâce à la multiplication des données à traiter suite à l'annotation en traits sémantiques (multiplication moyenne par 10), nous pouvons être plus sélectifs dans le choix des traits sémantiques[11].

La désambiguïsation de chaque occurrence des candidats est réalisée en comparant chaque contexte avec les profils terminologiques et non terminologiques. Les contextes lexicaux sont comparés avec les profils lexicaux LS_{on} et LS_{off} tandis que les contextes annotés sémantiquement sont comparés avec les profils sémantiques LST_{on} et LST_{off} . La comparaison est identique dans les deux cas. On détermine les éléments communs entre le contexte de l'occurrence à désambiguïser et les deux profils adéquats. Les taux de spécificité des éléments communs sont additionnés respectivement pour chaque profil, terminologique vs. non terminologique, ce qui permet d'obtenir deux scores : $score_{on}$ et $score_{off}$. L'occurrence est jugée terminologique si $score_{on}$ est supérieur à $score_{off}$ et elle est jugée non terminologique dans le cas contraire. Dans les résultats (section 3), la désambiguïsation textométrique sur les contextes lexicaux est désignée par l'abréviation *SpecLex* et celle qui est appliquée sur les contextes annotés en traits sémantiques est désignée par l'abréviation *SpecTraits*.

Le dernier point méthodologique à préciser est que les profils, lexicaux et sémantiques, sont systématiquement recalculés en éliminant le contexte de l'occurrence à désambiguïser. Ceci a pour but d'éviter de fausser les résultats de l'évaluation et d'éviter toute circularité.

2.3.2 Algorithme de Lesk

Cet algorithme (Lesk, 1986) vise à associer la bonne définition à chaque unité lexicale en contexte. Pour appliquer cet algorithme dans notre expérience, nous avons construit des définitions terminologiques pour les candidats termes en nous inspirant des usuels spécialisés en sciences du langage. Plusieurs ressources sont utilisées pour l'écriture des défintions terminologiques (Dubois J. et al. 2012 ; Ducrot O. et Schaeffer J-M. 1999 ; Neveu F. 2004). L'adaptation de l'algorithme de Lesk a consisté à choisir entre la définition terminologique d'un candidat et les définitions de ce candidat lorsqu'il est utilisé de manière non terminologique en sciences du langage. Ces défintitions sont prises dans les ressources de langue générale utilisées, le TLFi et WiktionnaireX[12]. Pour réaliser cette expérience, l'algorithme mesure l'intersection entre le contexte de chaque occurrence du candidat terme et les définitions candidates (la définition terminologique et les définitions de langue générale). Lorsque l'algorithme est appliqué aux contextes annotés en traits sémantiques (méthode appelée *LeskTraits*), l'intersection est calculée entre les traits sémantiques du contexte et les traits sémantiques des définitions candidates. Lorsque l'algorithme est appliqué aux contextes lexicaux (méthode appelée *LeskLex*), l'intersection est calculée entre les unités lexicales du contexte et les traits sémantiques des définitions candidates. L'occurrence à désambiguïser est jugée terminologique si l'intersection avec la définition terminologique contient plus d'éléments que l'intersection avec l'ensemble des définitions non terminologiques. L'occurrence est jugée non terminologique dans le cas contraire.

3 Expériences réalisées

3.1 Méthodologie de l'évaluation

Pour mesurer les performances des différentes méthodes de désambiguïsation, et l'impact de l'annotation sémantique, nous utilisons l'indice courant du taux d'exactitude (*accuracy*). Le taux d'exactitude provient d'une adaptation des métriques de *précision/rappel/F-Mesure* au champ de la désambiguïsation lexicale (Navigli 2009). Les méthodes que nous évaluons se caractérisent par une *couverture* (Nombre de réponses produites / nombre de réponses attendues) de 100 % et, par conséquent, les mesures *précision/rappel/F-Mesure* sont égales et correspondent alors au *taux d'exactitude* (Schwab et *al.* 2013). Ce taux est calculé pour chaque candidat et pour chaque méthode de désambiguïsation testée (avec ou sans annotation sémantique ; textométrie vs. Lesk).

[11] Le recours au calcul de spécificité sur les traits sémantiques pourrait de plus nous permettre de qualifier les emplois terminologiques et non terminologiques de chaque candidat terme. Ceci n'a cependant pas encore été réalisé car cela suppose une analyse plus approfondie de traits sémantiques : différencier le métalangage lexicographique (par exemple « action de *verbe* », « celui qui ») ; affiner le traitement de traits poly-lexicaux (par exemple « faire référence à ») ; classer les traits sémantiques restants en traits inhérents vs. afférents, classifiants vs. spécifiques, etc.

[12] Les définitions à contenu terminologique qui sont présentes dans le TLFi ont été supprimées afin que les définitions du dictionnaire ne relèvent que de la langue générale.

taux d'exactitude = Nombre de réponses correctes / Nombre de d'occurrences du candidat terme

3.2 Jeu de test

Nous l'avons mentionné dans la section (2.1) consacrée aux données de travail et au corpus de référence, le ratio des occurrences validées sur le nombre total d'occurrences d'un candidat terme permet de classer l'ensemble des candidats annotés selon leur taux d'ambiguïté et leur tendance terminologique. Dans le corpus de référence, certaines combinaisons ne sont pas représentées, ce qui est logique : par exemple, on ne trouvera aucun candidat qui soit à la fois très ambigu et très terminologique. Les candidats termes du jeu de test ont été choisis parmi les combinaisons existantes et en fonction des cas d'ambiguïté qu'ils représentent (cf. exemples issus du corpus, section 1). Le tableau (3) ci-dessous résume les informations quantitatives utilisées pour la sélection des candidats termes du jeu de test. L'ensemble des taux d'exactitude obtenus sur le jeu de test est résumé dans le tableau (4) dans lequel les candidats termes sont ordonnés par taux d'ambiguïté croissant.

DONNÉES DE RÉFÉRENCE			
Jeu test	Ambiguïté	Termino	Fréquence
adjectif	1,88	98,12	212
hypothèse	4,22	4,22	166
patient	5,38	5,38	93
locuteur	13,04	89,96	168
exemple	15,34	15,34	567
objet	22,51	22,51	191
corpus	24,00	76,00	550
expression	28,16	28,16	226
argument	30,00	70,00	90
référence	30,28	30,28	218
énoncé	39,31	60,69	173
définition	45,36	54,64	183
Moyenne	**21,62**	**46,28**	**236**
Écart typeP			

Tableau 3 : Candidats termes du jeu de test décrits par leur taux d'ambiguïté , leur tendance terminologique et leur fréquence

RÉSULTATS DES 4 MÉTHODES					
SpecLex	SpecTraits	LeskLex	LeskTraits	Moyenne	Écart typeP
97,44	**97,44**	30,77	94,36	80,00	*28,45*
92,99	**96,18**	82,80	76,43	87,10	*7,90*
96,70	95,60	64,84	94,51	87,91	*13,35*
93,29	**94,51**	58,54	92,68	84,76	*15,15*
68,11	**84,53**	59,25	26,04	59,48	*21,33*
65,41	62,70	**74,05**	34,59	59,19	*14,81*
91,77	91,57	57,43	66,87	76,91	*15,13*
69,52	54,29	53,81	60,95	59,64	*6,37*
82,56	69,77	32,56	41,86	56,69	*20,26*
48,69	53,93	60,21	**63,87**	56,68	*5,82*
73,33	**87,27**	53,33	58,79	68,18	*13,23*
47,73	32,39	62,50	**68,18**	52,70	*13,90*
77,30	76,68	57,51	64,93	69,10	
17,12	*20,63*	*14,06*	*21,75*	12,78	

Tableau 4 : Taux d'exactitude obtenus par méthode pour les candidats termes du jeu de test (en gras, les taux d'exactitude les plus élevés)

L'objectif de notre expérience est d'évaluer dans quelle mesure l'annotation sémantique décrite dans cet article améliore ou non le taux d'exactitude de la procédure automatique de désambiguïsation terminologique, c'est donc dans ce sens que les premiers résultats détaillés sont présentés dans la section (3.2.2). Afin de situer ces premiers résultats, nous les comparons avec une reproduction de l'algorithme de Lesk appliqué aux contextes lexicaux et aux contextes enrichis sémantiquement (section 3.2.3).

3.2.1 Mesure de l'apport de l'annotation en traits sémantiques TLFi & WiktionnaireX V2 sur le jeu de test

La première tendance générale qui se dégage était prévisible : le taux d'exactitude décroît lorsque le taux d'ambiguïté augmente (figure 4) : les moins bonnes performances apparaissent avec les candidats les plus ambigus *exemple*, *objet*, *argument*, *définition*, *référence* et *expression*. La seconde l'était aussi : le taux d'exactitude est peu sensible à la tendance terminologique des candidats termes (figure 3).

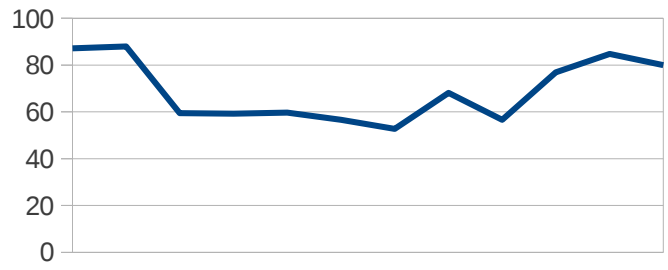

Figure 3: Moyenne des taux d'exactitude par candidats classés par tendance terminologique croissante

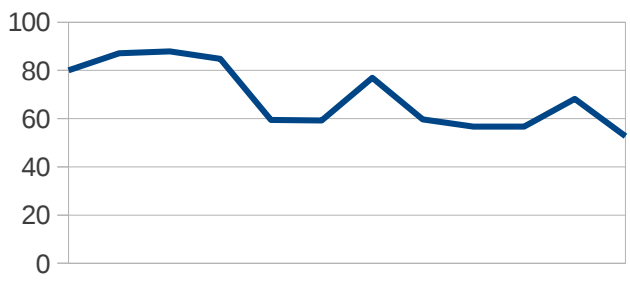

Figure 4: Moyenne des taux d'exactitude par candidats classés par taux d'ambiguïté croissant

Dans la mesure où le taux d'ambiguïté est le paramètre discriminant, c'est en fonction de celui-ci que nous présenterons la suite des résultats de l'expérience. Les courbes suivantes (figure 5) montrent l'évolution du taux d'exactitude en fonction de la mise en œuvre de la désambiguïsation sur des contextes lexicaux ou enrichis sémantiquement. Le premier constat que l'on peut faire est que l'annotation sémantique ne semble pas apporter d'amélioration sensible sur la moyenne des performances de la tâche de désambiguïsation terminologique. Cependant, dans certains cas, par exemple, *exemple* ou *énoncé* dans le jeu de test, on observe une performance supérieure et une performance très légèrement supérieure avec *référence*. Une analyse détaillée de l'ensemble des décisions divergentes sur ces trois candidats permettra dans la suite de nos travaux de mieux comprendre qualitativement l'influence de l'annotation sémantique.

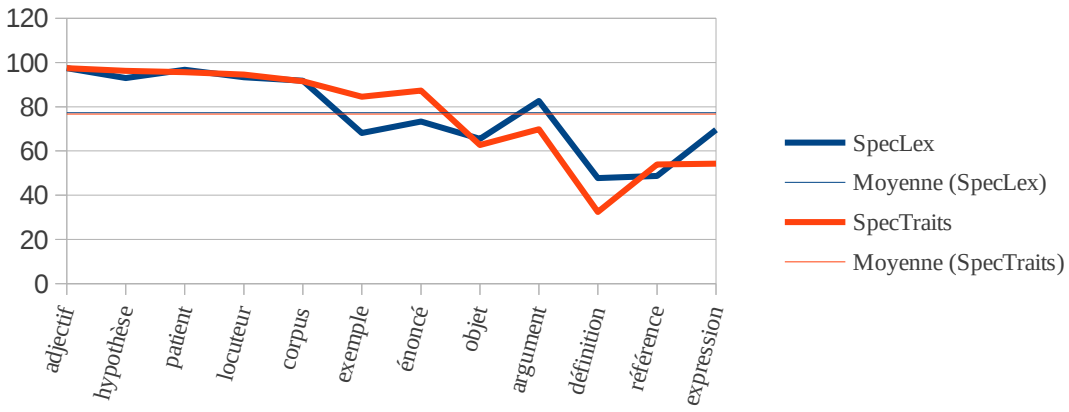

Figure 5: Taux d'exactitude par candidat pour la méthode textométrique et selon le type de contextes (lexicaux, SpecLex, ou annotés en traits sémantiques, SpecTraits)

Une analyse quantitative entre les méthodes SpecLex et SpecTraits est faite dans la section 3.3. Elle est appliquée par la suite parce qu'elle est appliquée non plus seulement au jeu de test mais à l'ensemble de candidats termes et de leurs occurrences.

3.2.2 Comparaison avec l'algorithme de Lesk sur le jeu de test

Pour situer les résultats obtenus par la méthode textométrique, nous avons implémenté la méthode de Lesk dans sa version initiale pour déterminer les taux d'exactitude obtenus à l'aide de cet algorithme fondateur dans le contexte particulier des travaux que nous développons. L'algorithme a été testé avec les contextes lexicaux (figure 6) et avec les contextes annotés en traits sémantiques (figure 7).

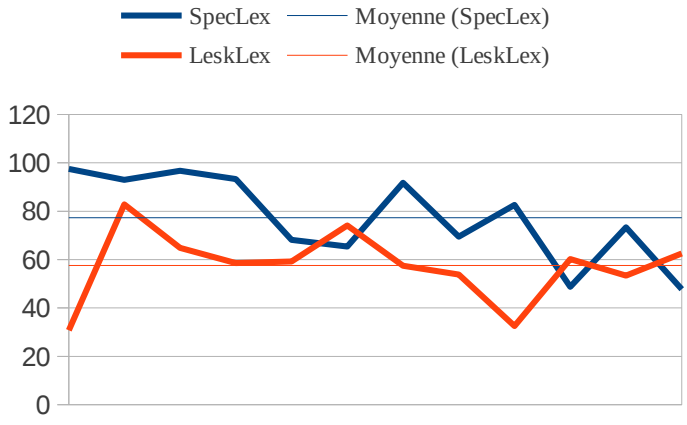

Figure 6: Comparaison des taux d'exactitude par candidat pour la méthode textométrique et pour l'algorithme de Lesk sur les contextes lexicaux

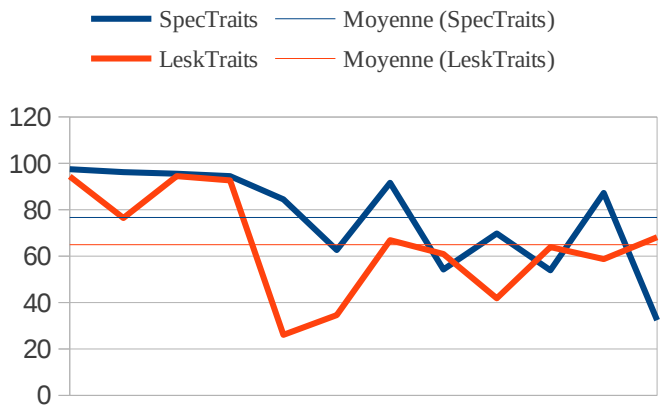

Figure 7: Comparaison des taux d'exactitude par candidat pour la méthode textométrique et pour l'algorithme de Lesk sur les contextes annotés en traits sémantiques

L'amélioration apportée par l'approche textométrique est visible : les courbes SpecLex et SpecTraits se situent au-dessus des courbes LeskLex et LeskTraits respectivement. Cependant, les faibles résultats de Lesk demandent à être explicités. Par rapport à l'état de l'art en effet (Schwab et *al.* 2013) ou (Navigli et *al.* 2007), les performances de l'algorithme de Lesk sont nettement moindres : taux moyen d'exactitude de 77 % avec une implémentation simple de l'algorithme de Lesk dans (Schwab et *al.* 2013 : 105) par exemple. La meilleure moyenne du taux d'exactitude que nous obtenons avec Lesk est de 64,93 % avec les contextes annotés en traits sémantiques.

Une explication possible repose sur la nature de l'ambiguïté que nous cherchons à résoudre. Bien que nous trouvant théoriquement dans une position beaucoup plus favorable comparativement à la difficulté que représente la désambiguïsation lexicale sémantique dans son entier, nous faisons face à des cas d'ambiguïté très fins. En témoignent les deux exemples ci-dessous du candidat *définition* : le premier montre un emploi jugé terminologique par l'annotation manuelle, le second emploi relève d'un emploi non terminologique.

> [Terminologique] : *Aujourd'hui, des relations entre termes sont de plus en plus souvent intégrées aux terminologies pour compléter les* ***définitions*** *en langage naturel.* Les relations sémantiques : du linguistique au formel - Aussenac-Gilles N. et Séguéla P. (2000). Cahiers de grammaire (25)
>
> [Non terminologique] : *La* ***définition*** *des termes est désormais le résultat d'une analyse systématique de l'usage des termes en corpus.* Les relations sémantiques : du linguistique au formel - Aussenac-Gilles N. et Séguéla P. (2000). Cahiers de grammaire (25)

Ce type d'ambiguïté terminologique entre, par ailleurs, en résonance avec une ambiguïté courante en sémantique lexicale : le lien métonymique entre une interprétation résultative et une interprétation processive que l'on rencontre souvent avec les noms déverbaux, ce qui est le cas du candidat *définition*. Le même cas de figure se présente le candidat *expression*.

> [Terminologique] : *[…] les* ***expressions*** *du type le jour suivant […]* Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de texte - Piérard S. et Begsten Y. (2007). TAL(47/2)
>
> [Non terminologique] : *L'****expression*** *de telle ou telle relation […]* Variabilité des outils de TAL et genre textuel : cas des patrons lexico-syntaxiques – Jacques M.-P. Et Assenac-Gilles N. (2006). TAL (47)

Pour ce type d'ambiguïté, la tâche de la compétition SemEval 2007 (Navigli et *al.* 2007) la plus proche est la tâche 8 : *Metonymy Resolution*. La tâche 8 de SemEval 2007 avait pour but de différencier trois niveaux de désambiguïsation (de la plus grossière à la plus fine) entre les sens métonymiques apparaissant avec les noms de lieu et les noms d'organisation. Pour chacune des catégories (lieux ou organisation), (Markert et Nissim 2007) ont défini la baseline du taux d'exactitude à 79,4 % pour la distinction entre les sens des noms de lieu et à 61,4 % pour la distinction entre les sens des noms d'organisation. Par rapport à ces taux d'exactitude, les deux variantes de la méthode Lesk telle que nous

l'avons implémentée (taux d'exactitude moyen de 57,51 % pour LeskLex et de 64,93 % pour LeskTraits) se situent autour du taux obtenu pour la distinction entre les sens des noms de lieux.

3.3 Comparaison SpecLex-SpecTraits sur un jeu de données plus étendu

La comparaison entre les résultats obtenus par la désambiguïsation textométrique (SpecLex et SpecTraits) sur le jeu de test nous a amené a conclure que l'annotation sémantique n'apportait pas d'amélioration sensible. Afin de vérifier ce constat sur un jeu de données plus important, nous avons pris deux textes du corpus Scientext (un article de revue et une communication) comme corpus d'évaluation et nous avons passé les méthodes de désambiguïsation par toutes les occurrences de candidats termes extraits par TTC-Termsuite. Dans cette nouvelle expérience, l'ambiguïté globale est plus modeste (10,34/50) et l'échantillon est plus hétérogène au niveau des fréquences. Au total, 2284 occurrences de candidats termes ont été évaluées par rapport à l'annotation manuelle.

	Types	Occurrences	Ambiguïté	SpecLex	SpecTraits
Candidats monolexicaux	339	1903	11,64	79,72	82,71
Candidats polylexicaux	92	381	3,84	80,58	80,05
Total	431	2284	10,34	**79,86**	**82,27**

Tableau 5 : Analyse des données du corpus d'évaluation

Comme le montre le tableau ci-dessus, l'ambiguïté des candidats monolexicaux (11,64/50) est bien supérieure à celle des candidats termes polylexicaux (3,84/50), c'est pourquoi les données du jeu de test (section 3.2) correspondent à des candidats monolexicaux uniquement. Sur le jeu de données plus étendu qui a été utilisé pour cette seconde expérience, on constate une amélioration des taux d'exactitude, que ce soit pour la méthode SpecLex ou la méthode SpecTraits. Cependant, contrairement à ce qu'on avait pu observer sur le jeu de test de la section précédente, on peut noter que l'utilisation des contextes enrichis en traits sémantiques (taux d'exactitude : 82,27% contre 79,86%) apporte cette fois une amélioration sensible par rapport aux résultats observés sur le jeu de test. Un étude approfondie de la complémentarité de ces deux méthodes doit donc être menée pour créer une méthode globale de désambiguïsation terminologique.

4 Conclusion et perspectives

Cet article se situe dans le champ de la désambiguïsation terminologique qui peut se voir *a priori* comme un cas de figure simplifié de désambiguïsation lexicale. La méthode que nous avons évaluée sur un jeu de candidats termes à désambiguïser s'appuie sur l'exploitation des contextes d'occurrences présentés sous deux formes : une version lexicale (contextes lexicaux) et une version enrichie en traits sémantiques (contextes sémantiques). Les résultats obtenus montrent que l'annotation sémantique n'apporte pas d'amélioration sensible pour la moyenne des performances mesurées sur le jeu de test. Cependant, nous avons vu que cette tendance s'inverse lorsque l'expérience est appliquée aux 67 candidats les plus fréquents dans le corpus de travail. Ce constat nous conduit à envisager deux perspectives immédiates : reproduire les expériences sur l'ensemble des candidats termes et analyser en détail les occurrences où les méthodes divergent. En effet, la figure (9) montre l'amplitude des variations possibles avec la méthode textométrique comme avec la méthode inspirée de Lesk. Pour pouvoir envisager une automatisation complète de la tâche de désambiguïsation, l'analyse des divergences permettrait de déterminer comment articuler les méthodes entre elles et dans quels cas privilégier l'une par rapport à l'autre.

Sur le plan des améliorations à apporter à la désambiguïsation elle-même, il serait intéressant de la faire collaborer avec des approches distributionnelles car celles-ci permettraient très probablement d'écarter d'emblée un certain nombre de cas massifs où l'occurrence d'un candidat terme fait partie d'une locution comme *par définition, faire référence à*, etc. Sur le plan des méthodes statistiques utilisées, une comparaison avec des méthodes de type SVM (*Support Vector Machines*) serait pertinente. Ces méthodes réduisent en effet la tâche de désambiguïsation à un problème de classification, en obtenant de très bons résultats par rapport à d'autres méthodes (Lee and Ng 2002 ; Navigli 2009).

Sur le plan de l'annotation sémantique enfin, nous envisageons de poursuivre nos travaux dans deux directions complémentaires. La première voie s'intéressera à deux manières possibles de limiter la dispersion de l'information sémantique due à l'utilisation des définitions d'une ressource lexicale sémantique : (1) nous appuyer sur l'hypothèse que le premier mot de même catégorie que l'unité lexicale à annoter joue le rôle d'un hyperonyme ou plutôt réfère au genre prochain dans la terminologie de la TST ; (2) utiliser les domaines d'usage qui figurent dans les définitions du TLFi. La seconde évolution possible pour l'annotation sémantique visera à utiliser une autre ressource lexicale, à savoir le

WordNet libre pour le français, la ressource Wolf telle qu'elle a été mise à disposition de la communauté par Benoît Sagot[13]. À l'aide de Wolf, nous envisageons une annotation en synsets[14] qui pourrait être réalisée de deux manières : (1) une annotation comparable à celle que nous avons développée dans cet article mais avec les synsets ou concepts possibles selon une correspondance des formes lemmatisées entre texte et ressource lexicale ; (2) une annotation désambiguïsante où le concept représentant un mot d'un article est identifié selon son contexte dans l'article. Des travaux d'indexation de sites web s'appuyant sur l'exploitation de WordNet sont déjà proposés (Desmontils et Jacquin 2001, Benyahia et *al*. 2009).

Remerciements

Nos vifs remerciements aux relecteurs/relectrices pour leurs remarques et suggestions qui ont permis d'améliorer cette proposition ; toute imprécision, erreur restante est bien sûr de notre entière responsabilité. Nos chaleureux remerciements à l'équipe du projet TermITH, au sein du laboratoire (en particulier Bertrand Gaiffe, Benjamin Husson, Etienne Petitjean, Jean-Marc Humbert et Sandrine Ollinger) ; au sein des partenaires du projet, en particulier Béatrice Daille (LINA), Agnès Tutin, Marie-Paule Jacques et Sylvain Hatier (LIDILEM), Claire François, Sabine Barreaux et Flora Badin (INIST), Yannick Toussaint et Felipe Melo-Mora (LORIA), Laurent Romary et Patrice Lopez (INRIA Saclay). Nos remerciements enfin à l'ANR pour le soutien financier accordé au projet TermITH (ANR-12-CORD-0029) ainsi qu'à nos tutelles (CNRS, INRIA, Université de Grenoble3, de Lorraine, de Nantes).

Références

Aubin S. and Hamon T. (2006) Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL)*.

Aussenac-Gilles N., Condamines A. (2009). Marqueurs de relations, genre textuel, structures syntaxiques. In Minel, J.-L. (Ed.), *Filtrage sémantique*, 115-149. Paris: Hermes/Lavoisier.

Baneyx A., Malaisé V., Charlet J., Zweigenbaum P. et Bachimont B. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. Dans Actes de la conférence *TIA-2005*, 12 pages. http://estime.spim.jussieu.fr/~jc/Files/BaneyxTIA2005.pdf [dernier accès (*d.a.*) 11/02/14]

Benyahia K., Lehireche A. et Latreche A. (2009). Annotation sémantique de pages Web. In Actes de la seconde *Conférence Internationale sur l'Informatique et ses Applications*, Saida, Algérie, 3-4 Mai, 8 pages. http://ceur-ws.org/Vol-547/54.pdf [pages consultées le 11/02/14]

Bourigault D., Jacquemin C. et l'HommeM.C. (2001). Recent Advances in Computationnal Terminology. John Benjamins :Amsterdam.

Bourigault D., Slozdian M. (1999). Pour une terminologie textuelle. Revue *Terminologies Nouvelles* 19 (Actes de la conférence *TIA 1999*), 29-32. http://www.rifal.org/cahiers/rint19/rint19.pdf. [(*d.a.*) 11/02/14]

Camacho collados J., Billami M. B., Jacquey E., Kister L. (2014). Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intégral. Dans Actes des *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Paris, France, 3-6 Juin

Claveau V. et L'Homme M.C. (2005). Apprentissage par analogie pour la structuration de terminologies - utilisation comparée de ressources endogènes et exogènes. Dans Actes de la conférence *TIA-2005*, 12 pages. http://www.irisa.fr/texmex/people/claveau/publis/Claveau-LHomme-tia05.pdf. [(*d.a.*) 11/02/14]

Condamines A., Péry-Woodley M.P. (2007). Linguistic markers of semantic and textual relations. In Alamargot, D., Terrier, P. & Cellier, J.-M. (Eds.), *Written documents in the workplace. Studies in Writing*. 3-16. Amsterdam: Elsevier.

Daille B. (1994). Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse en informatique fondamentale, Université Paris 7.

Daille B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, Bond E., Kohonen A. Carthy D.M. et Villalencio A. (eds), 9-16.

Daille B., Kageura K., Nakagawa H., Chien L.-F. (eds). (2004). Recents Trends in Computational Terminology, *Terminology*, 10(1). ISSN 0929-997.

[13] WOLF : WordNet Libre du Français, http://alpage.inria.fr/~sagot/wolf.html [page consultée le 04/02/2014].

[14] Un synset (ou un concept) correspond à un ensemble de mots qu'on peut les qualifier de synonymes entre eux représentant un sens très précis à l'aide d'une définition.

DAILLE B., JACQUIN CH., MONCEAUX L., MORIN E. et ROCHETEAU J. (2011). TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue. Démonstration au cours de la conférence *TALN 2011*. http://www.lirmm.fr/taln2011/DEMOS/DEMO_Daille_UnivNantes.pdf. [(*d.a.*) 11/02/14]

DESMONTILS E ; ET JACQUIN C. (2001). Des ontologies pour indexer un site Web. In *Journées francophones d'ingénierie des connaissances (IC)*, Jean Charlet (ed;), Presses Universitaires de Grenoble (PUG), Grenoble, 25-28 juin, 131-146,

DROUIN P. (2003). Term extraction using technical corpora as a point of leverage. In *Terminology* 9(1), 99-117.

DUBOIS J., GIACOMO M., GUESPIN L., MARCELLESI C., MARCELLESI J-B. ET MÉVEL J-P. (2012) DICTIONNAIRE DE LINGUISTIQUE. LAROUSSE.

DUCROT O. ET SCHAEFFER J-M. (1999) NOUVEAU DICTIONNAIRE ENCYCLOPÉDIQUE DES SCIENCES DU LANGAGE. ESSAIS, 832 PAGES.

GRABAR N., ZWEIGENBAUM P. (2004). Lexically-based terminology structuring. In *Terminology*, 10(1), 23-54. Résumé : http://www.ingentaconnect.com/content/jbp/term/2004/00000010/00000001/art00002. [(*d.a.*) 12/02/14]

HABERT B. (2005). Portrait de linguiste(s) à l'instrument », Texto! vol. X, n° 4, 2005.

HEIDEN S., MAGUÉ J-P., PINCEMIN B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Actes de la conférence *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome : Italie : 12 pages http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf. [pages consultées le 11/02/14]

JACQUES M-P., AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. Dans : *Traitement Automatique des Langues*, 47(1), 11-32. http://www.atala.org/Variabilite-des-performances-des [(*d.a.*) 11/02/14]

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots* (1), 127-165.

LEE, K.Y. AND NG, H. T. (2002) An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation, In Actes de la conférence *On Empirical Methods in Natural Language Processing* (EMNLP), Philadelphia, USA, 41–48.

LESK M. (1986). Automatic sense disambiguation using MRD : how to tell a pine cone fom an ice cream cone. In *Proceeding of SIGDOC' 86, ACM*, New York, USA, 24-26.

L'HOMME M.C. (2004a). Adjectifs dérivés sémantiques dans la structuration de terminologies. Dans Actes de la conférende *Terminologie, ontologie et représentation des connaissances*, Université Jean-Moulin Lyon-3, 22-23 janvier 2004. 6 pages. http://olst.ling.umontreal.ca/pdf/lhomme-lyon2003.pdf. [(*d.a.*) 11/02/14]

L'HOMME M.C. (2004b). *La terminologie : principes et techniques*, Montréal : Presses de l'Université de Montréal.

MANSER M. (2012). État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. Dans Actes de la conférence *JEP-RECITAL 2012*, 163-175. http://aclweb.org/anthology//F/F12/F12-3013.pdf. [(*d.a.*) 11/02/14]

MARKERT K., NISSIM M. (2007). SemEval-2007 Task 08 : Metonymu Resolution at SemEval-2007. In *Proceedings of the International Workshop on Semantic Evaluations (SemEval-2007)*, 36-41, Prague, Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W07/W07-2007.pdf [(*d.a.*) 12/02/2014]

NAMER F., ZWEIGENBAUM P. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *Proceedings 10th World Congress on Medical Informatics*, volume 107 of *Studies in Health Technology and Informatics*, 535-539, Amsterdam, 2004. IOS Press. http://www.ncbi.nlm.nih.gov/pubmed/15360870?dopt=Abstract [(*d.a.*) 11/02/14]

NAVIGLI, R. (2009).Word Sens Disambiguation : A Survey, *ACM Computing Surveys*, 41(2).

NAVIGLI R., LITKOWSKI K.B. ET HARGRAVES O. (2007). SemEval-2007 Task 07 : Coarse-Greined Anglish All-words Task. In *Proceedings of the International Workshop on Semantic Evaluations (SemEval-2007)*, 30-35, Prague, Association for Computational Linguistics.

NAVIGLI R. ET LAPATA, M. (2010) An Experimental Study of Graph Connectivity for Unsupervised Word Sens Disambiguation, In *IEEE Trans. Patter Anal. Mach. Intell.* , vol.32, pp. 678-692.

NEVEU F. (2004) DICTIONNAIRE DES SCIENCES DU LANGAGE. ARMAND COLIN, 316 PAGES.

PÉRINET A. , HAMON H. (2013). Hybrid acquisition of semantic relations based on context normalization in distributional analysis. Dans Actes de la conférence *TIA-2013*, 113-120. https://lipn.univ-paris13.fr/tia2013/Proceedings/actesTIA2013.pdf. [(*d.a.*) 11/02/14]

Poibeau T. (2005). Parcours interprétatifs et terminologie. Dans *Actes TIA 2005*. Rouen.

Rastier F., Valette M. (2009). De la polysémie à la néosémie. Dans *Texto!* XIV(1), 1-18. http://www.revue-texto.net/docannexe/file/2119/last_rastier_valette_polysemie.pdf

Retenauer C. 2012. Vers un traitement automatique de la néosémie : approche textuelle et statistique. Université de Lorraine., Thèse de doctorat en sciences du langage.

Sajous F., Navarro E., Gaume, B., Prévot, L., Chudy, Y. (2013). Semi-Automatic Enrichment of Crowdsourced Synonymy Networks: The WISIGOTH system applied to Wiktionary. *Language Resources & Evaluation*, 47(1), 63-96.

Schwab D., Goulian J., Tchechmedjiev A. (2013). Désambiguïsation lexicale de textes : efficacité qualitative et temporelle d'un algorithme à colonies de fourmis, In *TAL* 54(1), 99-138.

Toussaint Y., Namer F., Daille B., Jacquemin C., Royauté J., Hathout N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. Dans *Actes de la conférence TALN'98*, ATALA, Paris, France.

Valette M., Estacio-Moreno A., Petitjean E., Jacquey E. (2006). Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. Dans *Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*, P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). *Cahiers du CENTAL*, 2.1, UCL Presses Universitaires de Louvain (1), 357-366.