

Inclusion lexicale et proximité sémantique entre termes

Fidelia Ibekwe-SanJuan
Université de Lyon 3
4, cours Albert Thomas, 69008, Lyon
ibekwe@univ-lyon3.fr

Résumé

Nous étudions l'influence de l'inclusion lexicale sur la proximité sémantique entre termes. A partir de l'analyse des relations entre termes d'une ressource terminologique existante, lexicalement inclus dans ceux issus d'un corpus, nous formulons des hypothèses des relations engendrées. Ces hypothèses nous permettent de proposer un ordonnancement automatique des variantes des termes trouvées dans le corpus, par probabilité de proximité sémantique décroissante. Les premières expérimentations montrent que la prise en compte d'indices morpho-lexicaux comme la présence de noms composés, de noms propres et le nombre d'éléments ajoutés sont des critères à prendre en compte pour classer les variantes d'un même terme.

Mots-clés : acquisition de relations sémantiques, inclusion lexicale, proximité sémantique.

Keywords : semantic relations acquisition, lexical inclusion, semantic term proximity.

1. Introduction

La mise en relation de termes par le biais de liens lexicaux a fait l'objet de nombreuses études visant des applications différentes : enrichissement de thésaurus (Morin & Jacquemin, 2004), indexation (Jacquemin et al., 2002), veille scientifique et technologique (Ibekwe-SanJuan & SanJuan, 2004). L'inclusion lexicale désigne le fait qu'un terme t_1 est imbriqué dans t_2 : "*formation des maîtres*" et "*institut de formation des maîtres*". D'autres types d'opérations peuvent ne pas conduire à l'inclusion lexicale mais plutôt à une association lexicale comme le lien entre "*lait de vache*" et "*lait de chèvre*".

Pour identifier la relation exacte véhiculée par l'association lexicale, la majorité des études fait appel à des ressources sémantiques extérieures spécialisées comme un thésaurus du domaine (Bodenreider *et al.*, 2001 ; Grabar et Zweigenbaum, 2004). Ces ressources lorsqu'elles existent, sont rarement accessibles pour beaucoup de domaines. Elles ont de plus une couverture insuffisante qui empêche la reconnaissance des relations entre les nouveaux termes apparaissant dans un corpus spécialisé et ceux de la ressource pré-existante. Une approche alternative d'acquisition de relations sémantiques consiste en l'extraction de termes dans l'environnement immédiat des marqueurs relationnels. Ces marqueurs se réalisent sous forme de schémas lexico-syntaxiques (Hearst, 1992 ; Séguéla & Aussenac-Gilles, 1999 ; Morin & Jacquemin, 2004). Des études ont montré que ces schémas syntaxiques sont peu fréquents dans les corpus et que leur apparition dépend également du "genre de corpus" (Condamines, 2002). Les approches numériques visant l'acquisition de relations ont l'inconvénient majeur que si elles permettent de trouver des 'classes de termes' au comportement distributionnel similaire (Lin, 1998), elles ne peuvent pas caractériser les relations sémantiques entre ces termes sans recourir à une ressource sémantique extérieure (Weeds *et al.*, 2005).

Le postulat de l'existence d'une relation sémantique entre deux termes dont l'un est

inclus dans l'autre n'est pas nouveau en lexicologie et en terminologie (Kleiber & Tamba, 1990). Bodenreider *et al.* (2001) ont évalué la capacité de l'inclusion lexicale à détecter des relations d'hyponymie/hyponymie par comparaison à celles qui existaient déjà entre les concepts représentés par ces termes dans le métathésaurus UMLS¹. Cependant, ils se sont limités au cas de l'expansion adjectivale et de plus, au cas où le terme de départ et sa variante ne différaient que par un seul élément adjectival. Par exemple, le terme ternaire "*autoimmune hemolytic anemia*" est factorisé en trois termes par les auteurs: le centre (*anemia*) et ses deux expansions (*hemolytic anemia* ; *autoimmune anemia*) considérés comme ses hyponymes. Environ 50% des expansions adjectivales ont conduit à des relations hyponymiques dans l'UMLS. Dans une démarche similaire, Grabar et Zweigenbaum (2004) ont quantifié la proportion de relations de type "hyponyme/hyponyme" véhiculée par l'inclusion lexicale entre termes issus du thésaurus Mesh. La validation de ces relations s'est faite par comparaison avec les relations déjà existantes entre ces mêmes termes dans le thésaurus. Les auteurs ont montré que seulement 30% des relations hiérarchiques dans Mesh se réalisaient via l'inclusion lexicale. Notre objectif se différencie de ces deux études sur quelques points. Nous cherchons à expliciter le type de relation sémantique induite par différentes opérations d'inclusion lexicale. Nous ne nous intéressons pas à la seule expansion adjectivale ni à la seule relation d'hyponyme/hyponyme mais à toutes les relations induites par l'inclusion lexicale.

Nous cherchons également à déterminer le degré de proximité sémantique induite par chaque opération d'inclusion lexicale et quels indices morfo-lexicaux internes permettraient de prévoir ce degré. Après avoir exposé le cadre de notre travail (§2), nous précisons la notion de proximité sémantique telle que employée ici. La section §3 porte sur l'analyse des différentes opérations d'inclusion lexicale et les relations sémantiques induites. Une proposition d'ordonnement des variantes du corpus en fonction de la proximité sémantique supposée fait l'objet de la section §4. Nous terminons par une discussion sur la portée de notre étude en section (§5).

2. Cadre de travail

2.1 Travaux antérieurs

Nous avons défini dans Ibekwe-SanJuan (1998) des relations de variations basées sur l'association lexicale et la transformation structurelle. Ces relations sont utilisées pour former des classes de termes à l'aide d'un algorithme d'agrégation (Ibekwe-SanJuan & SanJuan, 2004). Les variations étudiées étaient réparties en deux catégories sommaires. Une première catégorie regroupait les variations qui affectent que les éléments modificateurs dans un terme. Il s'agit essentiellement des opérations d'expansion (Exp_g : *database design* → *real time database design*), d'insertion (Ins : *database design* → *database schema design*) ou de substitution de modificateur (Sub_m : *design of large database* ↔ *design of relational database*). Les variations qui affectent l'élément centre (tête) forment la deuxième catégorie : l'expansion droite (Exp_d : *information retrieval* → *design of information retrieval system*), l'expansion gauche_droite (Exp_gd : *information retrieval* → *chinese information retrieval system*) et la substitution de centre (Sub_c : *abundance of information retrieval* ↔ *information retrieval process*).

Jusqu'ici, les relations sémantiques induites par ces différentes variations n'étaient pas explicitées et de plus, toutes les variantes affectant les éléments modificateurs étaient intégrées dans la première phase de l'agrégation. Les variantes affectant l'élément centre étaient utilisées dans une deuxième étape pour agréger itérativement les ensembles formés à la première étape.

¹Unified Medical Language System' qui réunit plusieurs thesaurii du domaine médical.

Inclusion lexicale et proximité sémantique entre termes.

L'application de l'algorithme d'agrégation a plusieurs corpus a mis en évidence le fait qu'il était nécessaire, à l'intérieur d'une même catégorie de variations (modificateurs ou changement de centre) d'affiner les différentes opérations pour éviter la formation de classes trop hétérogènes thématiquement. Nous voudrions sonder la pertinence des indices endogènes (type d'opération lexicale ainsi que la structure interne des termes) pour ordonner les variantes d'un terme par une plus ou moins grande proximité avec celui-ci. Parmi les indices qui nous paraissent prometteurs, nous allons étudier les suivants :

- i- le nombre d'éléments ajoutés (qui peut influencer sur l'étirement du terme initial),
- ii- la catégorie morphologique de ces éléments (adverbe, adjectif, nom),
- iii- les propriétés lexicales de ces éléments (nom propre vs commun),
- iv- la fonction grammaticale affectée par l'opération d'inclusion lexicale (centre ou modifieur).

2.2 La notion de 'proximité' ou 'distance' sémantique

Par 'proximité sémantique' nous entendons l'écart entre le concept désigné par un terme t_1 et celui désigné par la variante t_2 dans laquelle t_1 est inclu. Deux variantes lexicales seront d'autant plus proches que le concept désigné par le terme englobant est apparenté à celui désigné par le terme englobé. La notion de proximité telle qu'employée ici est intuitive. Elle n'est pas l'opposée de celle de *distance* au sens mathématique du terme. Nous ne cherchons pas à quantifier, par une formule, l'écart entre un terme et une variante d'inclusion lexicale. Il est usuel de calculer la distance entre deux chaînes de caractères en utilisant la distance d'édition. Une distance d'édition entre deux chaînes correspond au coût minimal du passage de l'une à l'autre en terme d'opérations d'éditations sur des caractères (substitution, insertion, suppression). Sa possible application aux phénomènes de l'évolution terminologique par le biais de variations a été étudiée par Tartier (2004). L'utilisation de la distance d'édition pour calculer une "distance sémantique" pose un problème théorique elle implique l'attribution de poids arbitraire à chaque type d'opération étudiée, d'inclusion lexicale en fonction de sa signification linguistique. Dans le cas présent, cela impliquerait pour nous, l'attribution de poids aux différentes opérations d'inclusion lexicale que nous allons étudier, ainsi qu'une pondération de toutes les variables pouvant influencer sur la proximité sémantique des termes (nombre d'éléments ajoutés et leurs catégories et fonctions grammaticales, nature lexicale de ces éléments...). Autant de paramètres affectés de poids arbitraires, difficiles à justifier et dont le comportement lors du calcul de distance reste imprévisible. A titre d'exemple, étant donné le terme "*web site*" ($N_1_N_2$), la distance d'édition devra déterminer si une variante avec deux adjectifs insérés (*direct online web site* : $A_1_A_2_N_1_N_2$) est plus proche du premier terme que ne l'est une autre variante avec un seul nom ajouté (*intranet web site* : $N_3_N_1_N_2$). De même, il faut déterminer si une expansion droite avec un seul nom ajouté (*web site characteristics* : $N_1_N_2_N_3$) "coûte plus" que l'insertion de nombreux modificateurs (noms et adjectifs) en position initiale (*south australian public libraries network web site* : $N_3_A_1_N_4_N_5_N_6_N_1_N_2$). Face à de multiples combinaisons d'opérations élémentaires possibles pour passer d'un terme à un autre, le calcul de la distance d'édition peut conduire à des résultats erronés et contre-intuitifs. En effet, tout repose sur les valeurs initiales attribuées arbitrairement à chaque type d'opération si bien que l'ensemble du système est fragile. En revanche, il est possible de calculer la *similarité sémantique* (Pedersen *et al.*, 2004²) entre deux termes en s'appuyant sur une ressource sémantique extérieure comme WordNet. Dans ce cas, la similarité sémantique dépend du chemin parcouru dans le réseau sémantique pour atteindre un concept A à partir du concept B. Nous avons déjà évoqué le problème d'une ressource extérieure, à savoir la

² Les auteurs eux-mêmes emploient le terme de «similarité» pour présenter ces mesures car elles ne respectent pas la propriété mathématique d'inégalité triangulaire inhérente à une mesure de distance.

couverture souvent insuffisante des termes du corpus. De plus, WordNet établit des similarités entre unitermes (mots), non entre termes formés de plusieurs mots.

2.3 Corpus d'expérimentation

Afin d'élaborer nos règles d'ordonnement des variantes, nous avons constitué un corpus de 3818 mots-clés qui ont servi à indexer 3355 articles en anglais dans le domaine de la recherche d'information (désormais *corpusIR*) et 49 686 termes extraits automatiquement des résumés de ces même articles via un étiqueteur (LTChunker, développé à l'Université d'Edinburgh). Les mots-clés sont issus du lexique PASCAL maintenu par l'INIST/CNRS. Pour trouver des cas d'inclusion lexicale entre les deux vocabulaires, nous nous sommes limités aux mots-clés ayant au minimum deux éléments nominaux comme les suites "adjectifs_noms" (A_N) ou "noms_noms" (N_N). Ceci permet de réduire les cas de liens accidentels entre des unitermes et des variantes beaucoup plus longs. Par exemple, nous ne considérons pas l'inclusion lexicale entre "application" et "large distributed telecommunication application case study".

Nous limitons notre analyse aux cas des termes du corpus apparaissant sous forme de structure composée (sans attachement prépositionnel). En effet, les syntagmes nominaux (SN) anglais se réalisent majoritairement sous cette forme. Trois positions d'ajouts sont alors possibles, elles correspondent à nos définitions de l'expansion :

- i. expansion (ajout) de modifieurs à gauche du terme de départ : Exp_g
- ii. insertion de modifieurs consécutifs entre les bornes du terme : Ins
- iii. ajout d'éléments à droite du terme dont un est le nouveau centre, expansion droite : Exp_d
- iv. un quatrième type réunit les deux types élémentaires (a) et (c) : Exp_g et Exp_d.

Nous avons généré quatre listes de paires de termes : termes du corpus et les mots-clés du lexique auxquels ils sont reliés par une des opérations d'inclusion lexicale citées ci-dessus.

3. Opérations d'inclusion lexicale et relations sémantiques

Le tableau ci-dessous résume la répartition par type d'inclusion lexicale entre termes du lexique PASCAL et variantes du *corpusIR*. Ce qui est indiqué est le nombre de liens et non le nombre de termes impliqués. 1226 termes équivalents entre termes du corpus et mots-clés du lexique ont été trouvés, soit 2,5% de vocabulaire commun mais seuls 7,9% des termes du corpus ont un lien d'inclusion lexicale avec les mots-clés du lexique. Il est à noter que tous les liens d'inclusion lexicale sont uni-directionnels : ils vont tous dans le sens «mots-clés → termes du corpus», i.e., les mots-clés du lexique (essentiellement des termes binaires) sont inclus dans les termes plus longs du corpus.

<i>Type inclusion lexicale</i>	<i>Nb. liens</i>	<i>Erreurs_Rel.</i>	<i>Préc. T_Rel</i>
Expansion_gauche (Exp_g)	1961 (38,2%)	201 (10,2%)	89,7%
Insertion (Ins)	328 (6,4%)	32 (9,7%)	90,3%
Expansion_droite (Exp_d)	895 (17,4%)	12 (1,3%)	98,7%
Expansion_gauche_droite (Exp_gd)	717 (14%)	86 (12%)	78,0%
Termes équivalents (TE)	1226 (24%)		
Total (Exp_g + d + gd + Ins)	3901	331 (8,5%)	91,5%
<i>Préc. T moyenne</i>			89,1%

Tableau 1. Inclusion lexicale entre mots-clés du lexique et termes du *corpusIR*.

Inclusion lexicale et proximité sémantique entre termes.

Les lignes "*Total*" et "*Préc. T moyenne*" ne tiennent pas compte des termes équivalents (TE). La précision (*Préc. T_Rel.*) est la proportion de relations jugées 'plausibles' humainement sur l'ensemble des relations trouvées. Il est à noter que c'est un taux de précision théorique dans la mesure où les relations n'ont pas été évaluées vis-à-vis d'une tâche donnée (enrichissement du lexique, population d'ontologie, classification automatique de termes). Dans les sections suivantes, nous analyserons en détail les relations induites par chaque type d'opération d'inclusion lexicale, nous esquisserons des hypothèses sur la proximité sémantique engendrée en fonction des indices morpho-lexicaux et analyserons les sources d'erreurs. Ces sections sont organisées par type de relation induite et non par type d'inclusion lexicale. Deux types de relations se dégagent : hyponymie et association.

3.1. Relation d'hyponymie

Les opérations d'expansion gauche (Exp_g) et d'insertion (Ins) induisent généralement une relation d'hyponymie. Le nombre et la nature des éléments ajoutés ne changent pas fondamentalement le type de relation sémantique induite mais ils peuvent avoir un impact sur le degré de proximité sémantique d'avec le terme de départ. Néanmoins, ces indices ne sont pas les seuls à influencer sur ce degré. Plus influente est la combinaison d'indices "degré de compositionnalité" et "nature lexicale" des éléments ajoutés. Nous illustrons ces propos à travers ces exemples. Les termes à gauche sont les mots-clés issus du lexique PASCAL. Ils constituent nos "termes de départ".

citation analysis	→ <i>web-based</i> citation analysis	(1)
approximation operators	→ <i>rough set</i> approximation operators	(2a)
	→ <i>rough</i> approximation operators	(2b)
	→ <i>pawlak</i> approximation operators	(2c)
Archival Description	→ the <i>Encoded</i> Archival Description	(3a)
	→ the <i>General International Standard</i> Archival Description ³	(3b)
Dewey classification	→ <i>dewey decimal</i> classification	(4)
Markov model	→ <i>markov chain</i> model	(5)

• *Remarque 1. La non-compositionnalité des modifieurs favorise la proximité sémantique*

La thèse de la compositionnalité des énoncés voudrait que le sens d'un énoncé soit l'addition des sens de ses éléments constitutifs (voir Habert, 1998 pour une discussion⁴). Cependant, il est également admis que certains énoncés ou termes ont un sens non- ou faiblement compositionnel, à savoir que le sens de l'ensemble ne découle pas directement des sens de chacun des éléments pris isolément. Le degré de compositionnalité/non-compositionnalité des termes semble jouer un rôle sur la proximité sémantique avec le terme de départ. Nous observons que lorsque des modifieurs ajoutés ont un sens non-compositionnel, la proximité sémantique n'est plus une fonction directe du nombre d'éléments ajoutés. Des indices morpho-lexicaux de la non-compositionnalité peuvent être la présence des mots-composés⁵, des noms propres mais aussi de noms communs. Dans ce cas, les mots ajoutés fonctionnent comme des

³ Terme du domaine de la documentation désignant une norme d'archivage. La présence d'une abréviation après ce terme dans la phrase suivante confirme son statut de terme : «*As a data structure standard, it overlaps the General International Standard Archival Description (ISAD(G))*».

⁴ Voir aussi Nicole Wyatt. *Compositionality and context sensitivity*. Dept. of Philosophy, University of Calgary, Canada. 2004/04/23 Draft. <http://www.ucalgary.ca/~nwyatt/research/papers.html> [En ligne : consulté le 11/02/2005].

⁵ Toute combinaison de mots reliés par un trait d'union.

noms propres du domaine, des sortes d'unités pré-construites. En réalité, les variantes en (2a) et (2b) sont synonymes. Elles renvoient à deux variantes d'un même concept et sont donc deux hyponymes directs de « *approximation operators*⁶ ». Le nom « *set* » est souvent omis. Ainsi, le terme quaternaire en (2a) fonctionne comme les termes ternaires en (2b, c). La variante la plus étirée dans (3b) est un type de « *Archival Description* », donc son hyponyme direct au même titre que (3a). Les deux devraient être placés au même niveau hiérarchique dans un thésaurus du domaine. Les exemples (4) et (5) témoignent plutôt des cas d'ellipse. L'insertion de l'adjectif « *decimal* » et du nom « *chain* » n'introduit pas de nouveaux concepts plus spécifiques. La classification de Dewey est intrinsèquement de type « *décimal* ». Ce système de classification a été inventé par Dewey vers 1870 pour le classement systématique des ouvrages dans les bibliothèques. Le modèle de Markov a pour propriété intrinsèque celle d'être une chaîne. L'insertion ne conduit finalement qu'à la forme développée du terme. Les deux termes de départ seraient tout simplement des cas d'ellipse. Dans ce cas, il y aurait plutôt une relation d'équivalence, les deux variantes en (4, 5) devraient figurer au même noeud dans une ressource sémantique du domaine.

- *Remarque 2. La compositionnalité des modifieurs diminue la proximité sémantique*

Lorsque le sens des modifieurs ajoutés est compositionnel, la proximité sémantique diminue à mesure que de nouveaux modifieurs sont ajoutés. Ainsi, la variante (6b ci-dessous) est plus éloignée du terme de départ « *case study* » que celle en (6a). De même, (7b) avec trois ajouts paraît plus éloigné du terme de départ « *database management system* » où chaque modifieur ajouté modifie le précédent, « *medical image database management system* » (7a) est une sorte de « *image database management system* », lui-même sorte de « *database management system* ». Les cas d'insertion (8a,b,c) traduisent également le même phénomène de compositionnalité.

case study	→ <i>longitudinal case study</i>	(6a)
	→ <i>large distributed telecommunication application case study</i>	(6b)
database management system	→ <i>medical image database management system</i>	(7a)
	→ <i>transaction-time temporal object database management system</i>	(7b)
probabilistic model	→ <i>probabilistic retrieval model</i>	(8a)
	→ <i>probabilistic information retrieval model</i>	(8b)
	→ <i>probabilistic chinese language material model</i>	(8c)

- *Cas d'erreurs*

Il s'agit de variantes Exp_g et d'Ins qui n'ont pas induit la relation escomptée (hyponymie). 233 cas d'erreurs (10 %) sur les 2289 liens ont été recensés. Ces erreurs sont généralement dues à des modifieurs initiaux trop vagues ou à faible contenu informationnel :

Web information retrieval → **future Web information retrieval* (9)

information retrieval systems → **today information retrieval systems* (10)

information retrieval → **general information retrieval* (11)

En revanche, les modifieurs initiaux ci-dessous créent des relations acceptables :

genetic networks → *small genetic networks* (12)

game tree search → *selective game tree search* (13)

⁶ Terme informatique signifiant littéralement "opérateurs d'approximation" qui visent à caractériser un ensemble d'enregistrements à l'aide des seuls attributs connus.

Inclusion lexicale et proximité sémantique entre termes.

De ces exemples, il apparaît que la nature lexicale du modifieur initial ne permettra pas de déterminer si la variante est un bon hyponyme du terme de départ. Une solution envisageable serait de créer une liste de modifieurs initiaux (anti-dictionnaire de modifieurs initiaux) dont la présence entraînera le rejet de la relation, puis de ne vérifier que les candidats éliminés. Cela aura pour effet de réduire le nombre de vérifications manuelles. Une solution alternative serait de faire ce filtrage au niveau de l'extraction des termes.

3.2 Relation d'association

L'expansion droite et gauche-droite (Exp_d, Exp_gd) entraînent un changement de centre dans un terme existant faisant ainsi "sortir" la variante de la famille conceptuelle du terme initial. Le dernier élément ajouté doit être obligatoirement un nom (N). L'Exp_d et Exp_gd induisent une relation que nous qualifions globalement "d'association sémantique", appelée couramment « Voir aussi » dans le thésaurus. Le type de relation sémantique induite n'est pas modifié quelque soit l'écart lexical entre les deux termes. Néanmoins, l'examen de ces exemples montre que le nombre et la nature des éléments ajoutés ont un impact plus important sur la proximité sémantique. Des exemples illustrent ces propos.

african American	→ african American <i>households</i>	(14a)
	→ african American <i>low-income households</i>	(14b)
public library	→ public library <i>outlets</i>	(15a)
	→ public library <i>outlet Internet connectivity</i>	(15b)
	→ public library <i>outlet Internet connectivity data</i>	(15c)
authentic reasoning	→ authentic reasoning <i>expert systems</i>	(16)
chemical Abstracts Service	→ chemical Abstracts Service <i>Chemical Registry System</i> ⁷	(17)
clustering algorithm	→ <i>robust hierarchical clustering algorithm ROCK</i>	(18)
Computer Science	→ <i>the Networked Computer Science Technical Report Library</i>	(19)

Concernant la proximité sémantique, les mêmes remarques faites pour l'Exp_g et Ins sont valables ici. Le phénomène de compositionnalité/non compositionnalité joue le même rôle ici. En règle générale, plus de nouveaux noms sont rajoutés à droite du terme initial, plus grand est l'éloignement avec le terme de départ. (14a) est la variante la plus proche du terme initial (*african american*) puisque l'emphase est déplacée vers les « foyers » de type « afro-américains » (*african American households*). La variante plus longue (14b) rajoute au précédent concept, le nouveau concept de « foyers afro-américains à faible revenu » (*african American low-income households*). L'exemple (18) qui est un cas d'Exp_gd va dans le même sens.

En revanche, lorsqu'on considère les variantes en (15a-c), l'éloignement d'avec le terme de départ n'est plus une fonction directe du nombre d'éléments ajoutés. Alors que l'ajout d'un seul nom introduit un nouveau concept en (15a), en (15b) les deux nouveaux noms « *Internet connectivity* » fonctionnent comme un bloc dont le sens est non-compositionnel. En (15c), l'ajout d'un troisième nom (*data*) constitue un nouveau concept, aboutissant à un saut de trois concepts entre (15c) et le terme binaire de départ (*public library*). Ce phénomène de non-compositionnalité est observable dans les exemples (16-17) où « *expert systems* » et « *Chemical Registry System* » correspond chacun à un concept ou objet distinct du domaine. En (18), un cas d'Exp_gd, des nouveaux éléments rajoutés à gauche et à droite du terme

⁷ Ce terme complexe figure dans le titre d'une publication « *Chemical Abstracts Service Chemical Registry System : History, scope, and impacts* ». Il est fait en réalité de deux termes "chemical abstracts service" qui dispose d'une abréviation (CAS) et "chemical registry system".

rendent le sens de l'ensemble non-compositionnel faisant de cette variante un nom propre. La difficulté consiste à pouvoir identifier des blocs «non-compositionnels» en l'absence d'indices graphique ou lexical (noms propres, lettre majuscules). Cette difficulté s'accroît si comme dans l'exemple (19), les éléments non-compositionnels ne sont pas consécutifs.

- *Cas d'erreurs*

98 cas de bruit (6%) ont été observés parmi les relations d'Exp_d (1612 liens). Plusieurs raisons expliquent ces erreurs. La présence de chiffres à la fin d'une variante d'Exp_d ou Exp_gd peut entraîner une mauvaise relation (exemple 20 ci-dessous). Ceci est courant dans des domaines techniques où des chiffres viennent à la fin pour spécialiser un concept ou un type objet plus générique. Ces variantes fonctionnent au contraire comme des hyponymes du terme de départ. D'autres sources sont des noms qui en association avec d'autres, ont un apport sémantique faible. Ainsi, "*city* » et "*state*" rajoutés à "*New York*" dans (21a, b) n'entraînent pas un changement de référent. En fait, leur sémantique est déjà contenue dans le terme de départ (comme dans les exemples 4-5 ci-dessus). Enfin, des erreurs d'étiquetage morphologique peuvent conduire à l'extraction de mauvais termes qui vont se retrouver ensuite dans une situation d'inclusion lexicale (22).

air bus	→ *air bus <i>A320</i>	(20)
New York	→ *new York <i>City</i>	(21a)
	→ *new York <i>State</i>	(21b)
knowledge structure	→ *knowledge structure <i>due</i>	(22)

4. Ordonnement des variantes d'inclusion lexicale

Dans les sections §3.1 et §3.2, nous avons souligné le fait que la proximité sémantique entre termes ne pouvait se déterminer automatiquement à partir du nombre d'éléments ajoutés dans la variante plus longue. En revanche, les indices lexicaux que nous avons sondés nous permettent de proposer un ordonnancement de ces variantes par probabilité de proximité sémantique croissante. Ce sera l'objet de la section (§4.1). L'objectif étant de ranger en tête les variantes synonymes et hyponymes directs. Nous avons appliqué ces règles d'ordonnement aux variantes d'inclusion lexicale provenant uniquement du *corpusIR* (§4.2) car les termes disposaient des étiquettes morphologiques fournies par LTChunker. Par ailleurs, cela permettait de tester la robustesse des règles en évitant les biais éventuels liés à un vocabulaire contrôlé qui avait servi à élaborer les règles. Donc, l'automatisation de cet ordonnancement ne porte pas sur le couple "mots-clés issus du lexique PASCAL" et "leurs variantes dans le *corpusIR*" mais uniquement sur ces dernières.

4.1 Règles d'ordonnement des variantes

L'ordonnement que nous proposons exprime le fait que lorsqu'un terme t_1 (quelque que soit sa longueur) a des variantes d'inclusion lexicale, elles seront classées dans l'ordre suivant : en rang 1 (r1) les variantes Exp_g ou Ins comprenant l'un des éléments suivants : mot-composé, une abréviation, un nom propre, un syntagme adjectival (SA), un seul nom commun. En rang (r2) seront classées les autres variantes Exp_g ou Ins ordonnées par nombre d'éléments ajoutés. En rang (r3), figureront les variantes Exp_d qui contiennent un des éléments suivants : un seul N commun, un ou plusieurs noms propres ou un mot-composé. En rang (r4) nous classons tous les autres cas d'Exp_d, ordonnées par nombre d'éléments ajoutés. En rang (r5) les variantes Exp_gd seront ordonnées suivant les mêmes principes qu'en (r3). En

Inclusion lexicale et proximité sémantique entre termes.

rang (r6), tous les autres cas d'Exp_gd seront ordonnées par nombre d'éléments ajoutés. Ainsi, le nombre d'éléments ajoutés joue un rôle secondaire dans ce classement.

<i>T1</i>	<i>Rang</i>	<i>Variante</i>	<i>T1</i>	<i>Rang</i>	<i>Variante</i>
t ₁	exp_g1	: (<MC> N) t ₁ : (A N)? (U NNP) t ₁ : <SA> ^{1,2} t ₁	x ₁ x ₂	ins_1	: x ₁ (<MC> <SA> N) x ₂ : x ₁ (U NNP) x ₂
	exp_g2	autres cas de type X+ t ₁		ins_2	autres cas de type: x ₁ X+ x ₂
t ₁	exp_d3	: t ₁ (<MC'> (A? N)) : t ₁ (A N)? NNP+			
	exp_d4	autres cas de type t ₁ X+			
t ₁	exp_gd 5	: (<MC'> A N) t ₁ (<MC'> (A?N)) : (<SA> N)? NNP+ t ₁ (<SA> N)? NNP+ NNP+ N			
	exp_gd 6	: X+ t ₁ X+			

Tableau 2. Ordonnancement des variantes d'inclusion lexicale.

où :

T1 = terme de départ

t₁ = terme de la structure <SA>? N+

x₁ x₂ = sous-suites d'éléments composant un terme

< > = structure non-terminale (patrons morphologiques)

<MC> = mot-composé défini comme toute suite de mots reliés par un ou plusieurs traits d'union

<MC'> = mot composé se terminant obligatoirement par un N

{i,j} = nombre d'occurrences d'une catégorie compris entre i et j

+ = une ou plusieurs occurrences

* = opérateur de Kleene (plusieurs ou zéro occurrence)

| = symbole de l'optionnalité

? = catégorie ou symbole pouvant être vide

U = abréviation

<SA> = syntagme adjectival de la structure W? A

W = adverbe, A = adjectif, N = nom, NNP = nom propre

X = mot

4.2 Application aux variantes d'inclusion lexicale issues du corpusIR

Au total 4272 termes différents sont impliqués dans les opérations d'inclusion lexicale dans le *corpusIR*. Le tableau 3 ci-dessous donne la répartition des termes et des relations traitées par chaque règle.

Le tableau 4 ci-après illustre l'ordonnancement obtenu pour des variantes de deux termes "*information management*" et "*distributed system*".

<i>Règle</i>	<i>Nb. termes</i>	<i>Nb. relations</i>
exp_g1	5241	3408
ins_1	2845	1632
exp_g2	1508	925
ins_2	606	342
exp_d3	3138	1931
exp_d4	449	258
exp_gd5	917	574
exp_gd6	246	140
<i>Total</i>	<i>14950</i>	<i>9210</i>

Tableau 3. Application des règles d'ordonnement sur les variantes du *corpusIR*.

<i>Rang</i>	<i>Variantes</i>	<i>Rang</i>	<i>Variantes</i>
t ₁	information_NN management_NN	t ₁	distributed_JJ system_NN
exp_g1	E-healthcare_NNP information_NN management_NN	ins_1	distributed_JJ OLTP_NNP system_NN
exp_g1	global_JJ information_NN management_NN	ins_1	distributed_JJ computing_VBG system_NN
Ins_1	information_NN infrastructure_NN management_NN	ins_1	distributed_JJ database_NN system_NN
ins_1	information_NN resource_NN management_NN	ins_1	distributed_JJ object-based_NN system_NN
exp_g1	strategic_JJ information_NN management_NN	ins_1	distributed_JJ shared-nothing_JJ system_NN
ins_2	information_NN security_NN risk_NN management_NN	exp_g1	*future_JJ distributed_JJ system_NN
exp_d3	information_NN management_NN perspective_NN	exp_g1	heterogeneous_JJ distributed_JJ system_NN
exp_gd5	MI5_NNP information_NN management_NN efficiency_NN	ins2	distributed_JJ database_NN management_NN system_NN
exp_gd5	labour-intensive_JJ information_NN management_NN infrastructure_NN	exp_d3	distributed_JJ shared-nothing_JJ information-retrieval_NN system_NN
exp_gd5	personal_JJ information_NN management_NN appliance_NN	exp_gd5	distributed_JJ system_NN configuration_NN

Tableau 4. Exemples de variantes d'inclusion lexicale d'ordonnées.

On remarquera que la variante placée la plus proche du terme de départ est concerne l'ajout de d'éléments modificateurs avec une abréviation ou un nom propre (étiquette 'NNP') : *distributed_JJ OLTP_NNP system_NN* et *E-healthcare_NNP information_NN management_NN*. Les variantes d'un même rang sont rangées par ordre alphabétique. On observera également que "*distributed_JJ database_NN system_NN*" est placé avant "*distributed_JJ database_NN management_NN system_NN*" en accord avec l'hypothèse qu'en

Inclusion lexicale et proximité sémantique entre termes.

cas de compositionnalité, le nombre d'éléments ajoutés accroît la distance sémantique avec le terme de départ.

Nous avons vérifié manuellement les 2000 premières variantes ainsi ordonnées. Globalement, les règles ont bien fonctionné sur l'ensemble des variantes, nous n'avons donc pas relevé d'incohérence entre le rang des variantes et l'hypothèse de proximité sémantique proposée en tableau 2. La seule source d'erreurs est due aux règles d'extraction de termes où des séquences comme "**future_JJ distributed_JJ system_NN*" et "**several_JJ information_NN retrieval_NN application_NN*" avaient été extraites comme termes. Elles deviennent des variantes de "*distributed system*" et de "*information retrieval*" respectivement. L'ajout des modificateurs contextuels tels "*future*" et "*several*" n'engendre pas la relation escomptée. Ces cas représentent 164 termes sur les 2000 variantes examinées, soit 8,2% d'erreurs et 91,8% de précision théorique. Ce taux est quasiment identique à celui déjà relevé dans le premier corpus d'étude où nous avons comparé mots-clés issus du lexique PASCAL et leurs variantes dans le corpus (tableau 1).

5. Discussion

Nous avons sondé la pertinence des critères internes à la structure des termes pour acquérir des relations sémantiques entre termes. Cette étude a montré que les opérations lexicales très simples comme l'inclusion lexicale sont une source non-négligeable d'acquisition de relations d'hyponymie/hyponymie et d'association. Elle a également mis en évidence des difficultés du classement des variantes par proximité sémantique décroissante dues à la non compositionnalité de certains éléments ajoutés.

Notre étude a confirmé le fait qu'en général, l'expansion gauche et l'insertion induisent une relation d'hyponymie. Elle a mis en évidence également le fait que l'expansion droite et gauche_droite induisent une relation d'association. Cependant, quelques précautions doivent être prises avec ces affirmations. Nous avons observé que dans certains cas, des variantes d'expansion ou d'insertion conduisent plutôt au lien inverse (hyponymie ou méronymie) avec le terme lexicalement englobé. A titre d'exemple, « *plain semi-post algebra* » et « *alpha-rough sets* » termes de l'informatique fondamentale sont respectivement des généralisations conceptuelles de « *semi-post algebra* » (algèbres de Post) et de « *rough sets* » (ensembles flous). En français, « *algèbre de Post généralisée* » est une généralisation des « *algèbres de Post* ». Ici, l'ajout de nouveaux éléments conduit au contraire à la désignation de concepts plus génériques. Ceci serait dû à la chronologie des découvertes faites par les chercheurs, contraints de rajouter un nouveau qualificatif à un terme voisin déjà existant pour qualifier leur découverte. Nous avons également souligné le fait que la proximité sémantique entre deux termes dont l'un est lexicalement inclus dans l'autre ne se déduit pas toujours du nombre d'éléments qui les séparent. Ce constat est corroboré par Grabar & Zweigenbaum (2004) où des termes à quatre éléments pouvaient être des hyponymes directs des termes binaires dans une ressource sémantique externe : « *inhibiteurs captage agents adrenergique* » est situé directement en dessous de « *agents adrenergiques* » dans le thésaurus Mesh alors qu'un terme ternaire peut ne pas être un hyponyme direct d'un terme binaire : « *acides gras* » et « *acides gras indispensables* » ne sont pas dans un lien « père-fils » direct dans le thésaurus Mesh mais sont reliés par un chemin intermédiaire.

Une autre contribution de cette étude a été d'affiner les relations induites par l'inclusion lexicale et de proposer un ordonnancement des variantes qui en découlent par probabilité de proximité sémantique décroissante. Ce classement reste perfectible. Notamment, nous devons y intégrer les cas de substitution (modificateur et centre) ainsi que des variantes apparaissant sous

forme de structure prépositionnelle mais celles-ci ne devront pas bouleverser le classement actuel.

En outre, la précision réelle des relations acquises ici ainsi que la pertinence de l'ordonnement proposé se mesurera vis à-vis d'une tâche précise. S'agissant des tâches de construction de ressources sémantiques d'un domaine, cette précision dépendra du choix définitif d'inclure ou non une variante et sa relation dans une ressource sémantique du domaine et à quel endroit. Ce choix relève d'une série de considérations pragmatiques qui sont indépendante des phénomènes linguistiques eux-mêmes. Grabar & Zweigenbaum (2004) ont constaté que l'uniterme « *personnalité* » lexicalement inclus dans « *personnalité compulsive* » n'étaient pas dans la même hiérarchie du thésaurus Mesh, donc n'étaient pas en relation "hyponyme/hyponyme". Le premier relève des types de 'comportements' (*behaviours*) alors que sa variante lexicale Exp_g a été considérée comme un type de désordre mental (*mental disorder*) et à ce titre, rattachée à une hiérarchie différente dévolue aux maladies (*diseases*). Ce cas témoigne de l'influence des "points de vues" humains sur l'organisation des concepts dans une ressource sémantique externe. Ces points de vues peuvent évoluer en fonction des besoins. Ils rendent délicat l'évaluation des relations acquises en corpus au regard d'une ressource sémantique exogène.

Pour des tâches d'expansion de requêtes ou de classification automatique des termes à des fins de veille scientifique, il n'est pas nécessaire de comparer les relations acquises du corpus à une ressource sémantique de référence car l'objectif n'est pas structurer les termes en une hiérarchie. L'évaluation se fera à travers la pertinence des regroupements effectués par le système pour dessiner la carte du paysage scientifique du domaine ou pour affiner une requête et trouver les bonnes réponses. Dans ce cadre là, l'ordonnement que nous proposons devrait avoir pour effet de retarder le moment où des termes plus éloignés sémantiquement seront intégrés dans une même classe.

Références

- BODENREIDER O., BURGUN A., RINDFLESCHE T. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. Actes "Terminologie et Intelligence Artificielle" (TIA-2001), Nancy, 3-4 mai, 11-21.
- CONDAMINES A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, 2002, 8(1), 141-162.
- GRABAR, N., ZWEIGENBAUM, P. (2004). Lexically-based terminology structuring: Some inherent limitations. *Terminology, Special Issue on Recent Trends in Computational Terminology*, John Benjamins, 10 (1), 23-53.
- BENOÎT HABERT. Des mots complexes possibles aux mots complexes existants: l'apport des corpus. Habilitation à Diriger des Recherches, Université Lille III - Charles de Gaulle, 1998.
- HAMON, T., NAZARENKO, A. (2001). Detection of synonymy links between terms. In: Bourigault, D., Bourigault, D., JACQUEMIN, C., L'HOMME, M.-C. (eds.) (2001). *Recent Advances in Computational Terminology*. John Benjamins.
- IBEKWE-SANJUAN, F., SANJUAN, E. (2004) Mining textual data through term variant clustering: the termwatch system. Actes "Recherche d'Information assistée par ordinateur (RIAO'04). Avignon", 2004, 487-503.
- IBEKWE-SANJUAN, F. (1998). Terminological variation, a means of identifying research topics from texts. Actes "Joint International Conference on Computational Linguistics" (*COLING-ACL'98*), Montréal Québec, 10-14, August 1998, 564-570.
- HEARST, M. (1992). Automatic acquisition of hyponyms from large text corpora. Actes "Computational Linguistics" COLING'92. Nantes, 539-545.
- JACQUEMIN, C., DAILLE, B., ROYAUTE J., POLANCO X. (2002). In vitro evaluation of a program for machine-

Inclusion lexicale et proximité sémantique entre termes.

aided indexing. *Information Processing & Management*, 38(6), 765-792

KLEIBER G., TAMBA I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie, In Mortureux M-F (dir.), L'hyponymie et l'Hyperonymie, *Langages*, 98, 7-32.

LIN, D. (1998). Automatic retrieval and clustering of similar words. Proceedings of the 36th Joint International Conference on Computational Linguistics (ACL-COLING'98), Montréal, pp. 768-773.

MORIN, E., JACQUEMIN, C. (2004). Automatic acquisition and expansion of hypernym links. *Computer and the Humanities*, 36p.

NENADIC, G., SPASSIC, I., ANANIADOU, S. (2004). Mining term similarities from corpora. *Terminology, Special Issue on 'Recent Trends in Computational Terminology'*, John Benjamins, 10 (1), 55-81.

PEDERSEN T., PATWARDHAN, MICHELIZZI (2004). WordNet::Similarity – Measuring the relatedness of concepts. *Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, Mai, 3-5 2004, Boston, 4p.

SEGUÉLA, P, AUSSENAC-GILLES, N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, Actes "Ingénierie des Connaissances " (IC'99), Palaiseau, 79-88.

TARTIER, A. (2004). Analyse automatique de l'évolution terminologique : variations et distances. Thèse de Doctorat, Université de Nantes, 2004, 277p.

WEEDS J., DOWDALL J., SCHNEIDER G., KELLER B., WEIR D. (2005). Using Distributional Similarity to Organise BioMedical Terminology. *Terminology, Special issue on 'Application-driven terminology Engineering'*, John Benjamins, 11(1), 30p. (A paraître).