Méthode semi-compositionnelle pour l'extraction de synonymes des termes complexes

Amir Hazem — Béatrice Daille

Université de Nantes, LINA UMR CNRS 6241
2 rue de la Houssinière, BP 92208
F-44322 Nantes cedex 3
{amir.hazem,beatrice.daille}@univ-nantes.fr

RÉSUMÉ. *L'extraction de synonymes et des mots sémantiquement liés est une tâche utile en recherche d'information et en traitement automatique des langues. L'analyse distributionnelle a fourni un cadre théorique et opérationnel pour la détection de synonymes en corpus qui a principalement été exploité pour la découverte des synonymes de mots simples relevant de la langue générale. Dans cet article, nous nous intéressons à la découverte de synonymes de phrasèmes nominaux relevant de domaines de spécialités. Nous proposons une méthode semi-compositionnelle non supervisée qui mêle analyse compositionnelle et analyse distributionnelle. Nous montrons que cette méthode permet d'identifier nombre de termes complexes synonymes non découverts par la méthode état de l'art fondée sur une analyse compositionnelle seule, tout en étant beaucoup plus précise que la méthode exploitant la seule analyse distributionnelle.*

ABSTRACT. *Automatic synonyms and semantically related word extraction is a challenging task, useful in many NLP applications such as question answering, search query expansion, text summarization, etc. While different studies addressed the task of word synonym extraction, only a few investigations tackled the problem of acquiring synonyms of multi-word terms (MWT) from specialized corpora. To extract pairs of synonyms of multi-word terms, we propose in this paper an unsupervised semi-compositional method that makes use of distributional semantics and exploit the compositional property shared by most MWT. We show that our method outperforms significantly the state-of-the-art.*

MOTS-CLÉS : *synonymes, termes complexes, compositionnalité, sémantique distributionnelle, méthode non supervisée.*

KEYWORDS: *synonyms, multi-word terms, compositionality, distributional semantics, unsupervised method.*

1. Introduction

L'extraction de synonymes et des mots sémantiquement liés est une tâche utile en recherche d'information et en traitement automatique des langues. L'analyse distributionnelle a fourni un cadre théorique et opérationnel pour la détection de synonymes en corpus qui a principalement été exploité pour la découverte des synonymes de mots simples relevant de la langue générale. La synonymie des termes est un phénomène couramment rencontré dans les textes. En revanche, les synonymes des termes ne figurent généralement pas dans une ressource dictionnairique (Kremer *et al.*, 2014) à l'exception de quelques sigles et acronymes. Dans cet article, nous nous intéressons à la découverte de synonymes de phrasèmes nominaux relevant de domaines de spécialités, et en particulier aux synonymes de termes complexes où l'un des lexèmes du terme est substitué par un de ses synonymes comme *énergie éolienne* ↔ *courant éolien* dans le domaine de spécialité des énergies renouvelables. Le principe de compositionnalité sémantique est utilisé pour générer les synonymes d'un terme complexe : deux termes complexes sont synonymes si l'un des composants est substitué par l'un de ses synonymes ; l'existence du terme complexe généré est vérifiée en corpus. Cette méthode est identique à celle de Hamon et Nazarenko (2001) mais diffère sur la fourniture des synonymes des composants des termes complexes. Hamon et Nazarenko (2001) s'appuient sur un dictionnaire de langue générale. Nous proposons d'exploiter plutôt les mots sémantiquement liés identifiés par une analyse distributionnelle un peu bridée. Nous démontrons que cette méthode appelée *semi-compositionnelle* permet d'extraire des synonymes de termes complexes avec une bonne précision là où l'approche à base de dictionnaire échoue. Des expérimentations sont faites sur trois langues et deux domaines de spécialités.

La suite de cet article est organisée de la manière suivante. La section 2 fait un tour d'horizon des principales méthodes employées pour la détection automatique de synonymes dans les textes. La section 3 décrit le principe de compositionnalité et de synonymie. La section 4 présente l'approche semi-compositionnelle utilisée pour la génération de synonymes de termes complexes et les filtrages en corpus opérés. La section 5 décrit les différentes ressources linguistiques utilisées dans nos expériences, et en particulier les listes de référence exploitées. La section 6 évalue la contribution de divers paramètres de l'analyse distributionnelle vis-à-vis de la qualité des termes complexes synonymes extraits. Les sections 7 et 8 détaillent les acquis et les points restant à améliorer.

2. Identification automatique des synonymes

L'identification automatique de synonymes, et plus généralement des mots sémantiquement liés, est possible grâce à trois grandes familles d'approches, à savoir : (i) les approches distributionnelles qui effectuent une analyse distributionnelle en corpus dans un cadre purement monolingue, (ii) les approches qui exploitent les énoncés définitoires en corpus ou présents dans des ressources lexicales, et (iii) les approches d'alignements lexicaux multilingues qui s'appuient sur des textes traduits considérés

comme des réservoirs de synonymes. Une exploitation par combinaison des approches de ces trois familles permet, elle aussi, d'améliorer la qualité des synonymes extraits (Wu et Zhou, 2003).

L'analyse distributionnelle est la méthode la plus populaire pour la découverte de synonymes en corpus. Elle est fondée sur la conception contextualiste du sens d'un mot : « On reconnaît un mot à ses fréquentations »[1]. Le sens d'un mot est défini par l'ensemble des contextes dans lequel il apparaît. Deux mots, ou plus largement deux unités lexicales, partageant des contextes similaires, vont être sémantiquement liés (Harris, 1954). Plus les contextes seront similaires plus les mots le seront également. Plusieurs études ont exploité l'analyse distributionnelle automatique pour la détection de synonymes en corpus (Hindle, 1990 ; Grefenstette, 1994 ; Lin, 1998 ; Hagiwara, 2008 ; Ferret, 2010 ; Ferret, 2013). Elles diffèrent selon la définition du contexte adoptée, les méthodes pour comparer ces contextes et l'ordonnancement de ces mots identifiés comme sémantiquement liés. Lin (1998), par exemple, a introduit l'idée que les mots partageant le plus de relations syntaxiques étaient les plus favorables à être en relation de synonymie. Cette idée a été reprise et étendue au chemin syntaxique (Hagiwara, 2008) dans le but de pallier le manque de relations de dépendance syntaxique directes. Lorsqu'un nombre important de mots sémantiquement liés sont détectés, Ferret (2013) améliore leur ordonnancement en détectant et en déclassant les mots les plus ambigus.

Les énoncés définitoires sont des réservoirs à synonymes. Ceux-ci sont détectés dans les textes à l'aide de patrons morphosyntaxiques ou extraits de ressources lexicographiques telles que les dictionnaires de langue générale ou les dictionnaires terminologiques. Les patrons lexico-syntaxiques ont montré leur efficacité pour la détection de relations sémantiques telles que l'hyperonymie ou la méronymie. Dans le cas de la synonymie, il existe peu de constructions syntaxiques ou partiellement lexicalisées la dénotant. Blondel et Senellart (2002), par exemple, ont adopté l'hypothèse selon laquelle les synonymes partagent plusieurs mots en commun dans leurs définitions. Ainsi, des graphes d'énoncés définitoires sont construits à partir d'un dictionnaire. Chaque mot du dictionnaire représente un sommet du graphe, et deux mots u et v seront reliés entre eux par un arc, si et seulement si le mot u apparaît dans la définition du mot v ou inversement. Les synonymes sont extraits en fonction de la similarité entre les sommets des graphes.

Se positionner dans un contexte multilingue peut aussi favoriser l'extraction de synonymes. Plusieurs méthodes exploitent des corpus multilingues sous l'hypothèse que les mots qui ont des contextes de traduction similaires auront tendance à être en relation sémantique (Wu et Zhou, 2003 ; Van der Plas et Tiedemann, 2006). Partant de l'observation que les approches fondées sur la similarité distributionnelle étaient incapables de distinguer les synonymes des autres relations sémantiques, telles que l'hyperonymie ou l'antonymie, et qu'il y avait plus de chances de trouver des synonymes d'un mot en le caractérisant par ses traductions que par son contexte distri-

1. « *You shall know a word by the company it keeps.* »(Firth, 1957).

butionnel, Van der Plas et Tiedemann (2006) ont utilisé l'alignement multilingue des mots à partir de corpus parallèles dans plusieurs langues afin d'extraire les synonymes. L'hypothèse est qu'un mot n'est jamais traduit par son antonyme, son hyperonyme ou son co-hyponyme. Le mot anglais *apple*, par exemple, ne serait jamais traduit en français par *fruit* ou par *poire*. Ainsi, chaque mot est caractérisé par un vecteur de ses traductions dans dix langues cibles. Cette méthode qui est une extension de celle proposée par Wu et Zhou (2003) concernant les corpus parallèles, a produit de meilleurs résultats que l'approche distributionnelle monolingue exploitant les relations de dépendance syntaxique avec une augmentation moyenne de 7 % de F-mesure.

Certains travaux ont montré l'utilité de combiner plusieurs ressources pour améliorer la qualité des synonymes extraits. Wu et Zhou (2003), par exemple, combinent plusieurs ressources dont un dictionnaire monolingue, un corpus parallèle bilingue et un large corpus monolingue. Dans un premier temps, chaque ressource est utilisée individuellement pour extraire les synonymes. L'approche utilisant un dictionnaire monolingue est fondée sur la méthode de Blondel et Senellart (2002) citée plus haut. L'approche utilisant un corpus parallèle, quant à elle, est fondée sur la traduction d'un mot pour exprimer son sens. Ainsi, chaque mot est représenté par le vecteur de ses traductions accompagnées de leur probabilité de traduction. L'extraction des synonymes se fait en mesurant la similarité entre les vecteurs traduits selon la mesure du cosinus. Enfin, l'approche utilisant un large corpus monolingue est fondée sur l'hypothèse distributionnelle selon laquelle les synonymes auront tendance à apparaître dans les mêmes contextes lexicaux. Chaque mot est associé aux mots avec lesquels il est en relations de dépendance syntaxique. Un contexte est représenté par un triplet <mot1, type de relation, mot2>, où mot2 est appelé *attribut*. Ainsi, un mot sera caractérisé par un vecteur contenant l'ensemble de ses attributs. Deux mots seront synonymes s'ils partagent des attributs similaires. Par la suite, les trois approches sont combinées de telle sorte que la similarité entre deux mots sera la somme pondérée de la similarité renvoyée par chaque approche individuelle.

Quelle que soit la méthode utilisée, celle-ci produit une liste de mots potentiellement synonymes pour une unité lexicale donnée. Parmi ces candidats, il n'est pas évident d'établir une distinction claire et précise entre la catégorie des synonymes et les autres catégories de relations sémantiques (Lin *et al.*, 2003 ; Van der Plas et Tiedemann, 2006). Concernant l'approche distributionnelle, Resnik (1993 :18)[2] déclare que « l'information capturée en utilisant les méthodes distributionnelles n'apparaît comme ni vraiment syntaxique, ni purement sémantique ». La sémantique reliant les éléments lexicaux ne relève pas des seules synonymie ou quasi-synonymie mais aussi des autres relations sémantiques classiques telles que l'antonymie, l'hyperonymie, la co-hyponymie, la méronymie, et les relations sémantiques non classiques telles que les relations action-agent (Morris et Hirst 2004). Morlane-Hondère (2013) a réalisé une étude exhaustive et approfondie des relations sémantiques générées à l'aide d'une ana-

2. « *It would seem that the information captured using distributional methods is not precisely syntactic, nor purely semantic - in some sense the only word that appears is* distributional. »

lyse distributionnelle automatique pour le français dans le domaine général et confirme le vaste ensemble de relations sémantiques qu'elle met à jour.

Toutes ces approches ont été appliquées à l'extraction de synonymes de termes simples. L'extraction de synonymes de termes complexes composés de plusieurs unités lexicales, quant à elle, a été très peu étudiée alors que les synonymes de termes complexes sont de loin les plus fréquents quand il s'agit de langues de spécialités, particulièrement les langues romanes. Seuls Hamon et Nazarenko (2001) se sont intéressés à cette tâche et la méthode qu'ils ont proposée fait office de référence.

3. Compositionnalité et synonymie

Une définition générale de la compositionnalité admise par tous est celle proposée par Parnee *et al.* (1990) : une expression est compositionnelle si son sens est exprimé en fonction du sens de ses parties en respectant les règles syntaxiques de combinaison[3].

Hamon et Nazarenko (2001) font l'hypothèse que les synonymes des termes complexes sont compositionnels si leurs parties sont des synonymes. Ils ont défini trois règles pour détecter les relations de synonymie. Étant donné les candidats termes complexes $CCT_1 = (T_1, E_1)$ et $CCT_2 = (T_2, E_2)$ où T_1 (respectivement T_2) correspond à la tête du terme complexe et E_1 (respectivement E_2) correspond à son expansion et $syn(CT_1, CT_2)$ une relation de synonymie entre les candidats termes CT_1 et CT_2, les règles d'inférences suivantes sont utilisées :

– $R_1 : T_1 = T_2 \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$
– $R_2 : E_1 = E_2 \wedge syn(T_1, T_2) \supset syn(CCT_1, CCT_2)$
– $R_3 : syn(T_1, T_2) \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$

La règle R_1 signifie que les têtes syntagmatiques sont identiques et les expansions sont des synonymes (*collecteur général*/*collecteur commun*). Les synonymes des constituants lexicaux d'un terme complexe comme *général* dans cet exemple sont fournis par un dictionnaire de synonymes de la langue générale.

Kraft (2007) note que les expressions et leurs parties sont habituellement ambiguës et qu'il est difficile de leur assigner un seul sens. Nous illustrons cette remarque par l'analyse des synonymes des termes complexes qui sont présents dans les banques de données terminologiques. En examinant les parties des synonymes des termes du domaine des énergies renouvelables partageant au moins une partie commune, nous rencontrons des relations diverses :

3. « *A compound expression is composional if its meaning is a function of the meaning of the parts and of the syntactic rule by which they are combined.* »

– synonyme : *energy output*/*energy production* donnés par Termium[4] où *output*/*production* sont synonymes ;

– hyperonyme : *turbine noise*/*turbine sound* donnés par le GDT (*Le grand dictionnaire terminologique*)[5] où *sound* est un hyperonyme de *noise* ou encore *implantación de las máquinas*/*implantación de aerogeneradores* donnés par le Lexique panlatin où *máquina* 'machine' est un hyperonyme de *aerogenerador* 'aérogénérateur' ;

– indéfini : *nuclear plant*/*nuclear station* donnés par Termium où *plant* et *station* ne sont pas reliés sémantiquement ou encore *arbre lent*/*arbre primaire* donnés par Terminalf sans relation sémantique entre *lent* et *primaire*.

Notre hypothèse est que la sémantique distributionnelle qui permet d'identifier les mots en relation sémantique dans un domaine de spécialité doit aider à découvrir d'autres synonymes de termes complexes. Pour intégrer l'analyse distributionnelle au sein de la méthode compositionnelle, nous avons besoin d'adapter la méthode proposée dans (Hamon et Nazarenko, 2001).

4. Méthode semi-compositionnelle

Notre méthode s'inspire des travaux de Morin et Daille (2012) qui s'étaient intéressés à l'amélioration de l'alignement de termes complexes à partir de corpus bilingues comparables en combinant une approche compositionnelle avec une approche distributionnelle. Nous utilisons la même approche pour la tâche d'extraction de synonymes de termes complexes (TC). Nous partons de l'hypothèse qu'un terme complexe et ses synonymes adoptent le principe de compositionnalité. Par exemple, le synonyme d'*énergie renouvelable* peut être obtenu par : (i) la décomposition en parties du TC, puis (ii) la détection des mots en relation sémantique avec *énergie* et/ou *renouvelable* en utilisant la méthode distributionnelle, et enfin (iii) la recomposition des parties des termes candidats et le filtrage des termes recomposés à l'aide de listes construites à partir du corpus monolingue spécialisé. Notre méthode semi-compositionnelle diffère en deux points de la méthode proposée par Hamon et Nazarenko (2001), à savoir : (i) la manière d'extraire les synonymes des termes simples, et (ii) la longueur des termes complexes traités.

4.1. *Approche distributionnelle*

Contrairement à Hamon et Nazarenko (2001) qui utilisent un dictionnaire de langue générale pour déterminer les synonymes de chaque partie du terme complexe, notre approche exploite la palette des relations sémantiques des mots fournie par une analyse distributionnelle. L'approche distributionnelle propose des synonymes et des relations sémantiques en contexte, ce qui n'est pas le cas des dictionnaires de langue

4. www.btb.termiumplus.gc.ca
5. www.oqlf.gouv.qc.ca/

générale et surtout des lexiques recensant les vocabulaires spécialisés. De plus, et comme cité précédemment, les parties d'un terme complexe et de son synonyme ne sont pas obligatoirement en relation de synonymie et peuvent être reliées par d'autres relations comme la quasi-synonymie, l'hyperonymie ou l'hyponymie. Ainsi, l'utilisation d'un dictionnaire de synonymes de la langue générale limite l'identification de synonymes de termes complexes au seul cas où un lien de synonymie est attesté entre deux constituants de deux termes complexes.

Nous adoptons l'hypothèse classique de l'analyse distributionnelle qui dit que deux mots sont en relation sémantique s'ils partagent les mêmes contextes lexicaux. Ainsi, pour identifier ces liens sémantiques, nous modélisons le contexte des mots à l'aide de vecteurs, appelés *vecteurs de contexte*. Étant donné un corpus, nous calculons le contexte de chaque mot modélisé dans un vecteur, puis nous mesurons la similarité entre tous les vecteurs de contexte construits. Un score élevé de similarité entre deux mots induit une similarité sémantique forte.

Le vecteur de contexte $v_{w_i^s}$ d'un mot source donné w_i^s[6] par exemple, contiendra tous les mots qui apparaissent avec w_i^s dans une fenêtre contextuelle de n mots autour de lui. Soit $occ(w_i^s, w_j^s)$ la valeur de cooccurrence de w_i^s avec w_j^s qui est un mot appartenant à son contexte. Une mesure d'association, comme l'information mutuelle (notée IM) (Fano, 1961), le taux de vraisemblance (noté TV) (Dunning, 1993) ou le *discounted odds-ratio* (noté DOR) (Evert, 2008), est utilisée pour mesurer la corrélation entre w_i^s et w_j^s ainsi que tous les autres mots de son contexte. Enfin, une mesure de similarité est utilisée également pour chaque mot candidat w_i^t en fonction de son vecteur de contexte, $v_{w_i^s}^t$. Plusieurs mesures de similarité peuvent être utilisées. La plus populaire est la mesure du cosinus (Salton et Lesk, 1968) (notée COS) ou la mesure du Jaccard pondéré (notée JAC) (Grefenstette, 1994). Les candidats en relation sémantique avec w_i^s seront les mots candidats, ordonnés en fonction de leur score de similarité.

Ci-dessous la table de contingence et les mesures d'association et de similarité :

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

Tableau 1. *Table de contingence*

$$\begin{aligned} TV(i,j) = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(N)\log(N) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) \\ & -(c+d)\log(c+d) \end{aligned} \qquad [1]$$

6. L'exposant s de w_i^s fait référence au corpus source, et l'indice i au i-ième mot de ce corpus.

avec $N = a + b + c + d$.

$$\text{IM}(i,j) = log\frac{a}{(a+b)(a+c)} \quad [2]$$

$$DOR(i,j) = \log\frac{(a+\frac{1}{2}) \times (d+\frac{1}{2})}{(b+\frac{1}{2}) \times (c+\frac{1}{2})} \quad [3]$$

$$Cosinus_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t assoc_t^{l\,2}}\sqrt{\sum_t assoc_t^{k\,2}}} \quad [4]$$

$$Jaccard_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t max(assoc_t^l, assoc_t^k)} \quad [5]$$

avec $assoc_t^l$ par exemple qui fait référence à une mesure d'association donnée (TV, IM ou DOR) entre les deux mots t et l.

4.2. *Longueur des termes complexes*

À la différence de la méthode de Hamon et Nazarenko (2001), notre approche semi-compositionnelle ne se limite pas à des termes complexes de longueur 2. Elle peut être appliquée à des termes complexes de n'importe quelle longueur à partir du moment où les termes complexes suivent les règles R_1 et R_2. Nous proposons de généraliser ces règles à la fois en acceptant une plus large gamme de relations sémantiques et en autorisant les substitutions sémantiques aux termes complexes de n'importe quelle longueur. Nous étendons les règles R_1 et R_2 en remplaçant les relations synonymiques $syn(CCT_1, CCT_2)$ par les relations sémantiques $sem(CCT_1, CCT_2)$. La règle R_1^G correspond à la généralisation de la règle R_1 (respectivement la règle R_2^G correspond à la généralisation de la règle R_2) et T_1, T_2, E_1, E_2 sont des termes complexes. De plus, nous supprimons la règle R_3 en nous appuyant sur les résultats obtenus par Hamon et Nazarenko (2001) où ils ont montré que cette règle était peu productive et extrêmement bruitée. Nous obtenons donc les règles suivantes :

– $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
– $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$

Le tableau 2 illustre quelques exemples de termes complexes et de leurs synonymes en français, en anglais et en espagnol dans le domaine des énergies renouvelables. De même, le tableau 3 illustre quelques exemples de termes complexes

et de leurs synonymes en français et en anglais dans le domaine du cancer du sein. Ces termes et leurs variantes synonymiques ont été extraits de ressources terminologiques. La plupart des exemples de synonymes correspondent bien à nos règles où l'un des composants du terme complexe reste inchangé. Quelques exemples :

En : dans le domaine de l'énergie éolienne *wind turbine*/*wind machine*, *power supply*/*energy supply*, dans le domaine médical *invasive carcinoma*/*infiltrating carcinoma*, *adjuvant therapy*/*adjuvant treatment*.

Fr : dans le domaine de l'énergie éolienne *énergie renouvelable*/*énergie durable*, dans le domaine médical *curage du ganglion*/*ablation du ganglion*.

Pour chaque terme complexe (TC), nous fixons alternativement sa partie gauche pour extraire les mots en relation sémantique avec sa partie droite et sa partie droite pour extraire les mots en relation sémantique avec sa partie gauche. Ceci correspond aux règles R_1^G et R_2^G.

L'inconvénient de cette méthode est l'impossibilité de traiter les synonymes qui ne suivent pas les règles citées plus haut, par exemple : le terme complexe *moulin à vent* et son synonyme : *éolienne*.

Nous filtrons ensuite les termes complexes candidats obtenus en les comparant avec deux listes extraites du corpus : une liste de n-grammes et une liste de termes complexes candidats.

4.3. *Filtrage*

Le filtrage consiste à comparer la liste des termes synonymes candidats obtenue à l'aide de la méthode semi-compositionnelle à des listes extraites du corpus. Le but est de s'assurer de la correction syntagmatique des candidats, l'analyse distributionnelle par définition s'exonère de cette contrainte.

4.3.1. *Filtrage par n-grammes*

Une liste de n-grammes est construite automatiquement à partir du corpus préalablement lemmatisé. Partant du principe de filtrage des suites de mots improbables dans le corpus, une méthode simple consiste à collecter tous les n-grammes apparaissant dans le corpus. Tout n-gramme est considéré comme un terme candidat potentiel. Les termes synonymes candidats qui ne sont pas des n-grammes sont éliminés.

4.3.2. *Filtrage par extracteur terminologique*

Les termes synonymes candidats sont comparés à une liste de termes candidats extraits par un extracteur de termes. Un terme synonyme candidat non proposé par l'extracteur est éliminé. Nous avons utilisé TermSuite[7] (Rocheteau et Daille, 2011).

7. `logiciels.lina.univ-nantes.fr/redmine/projects/termsuite`

TermSuite détecte les occurrences de termes simples et complexes en exploitant leurs structures morphosyntaxiques caractéristiques. Il normalise et regroupe les différentes variantes des candidats termes. Ceux-ci sont ensuite ordonnés en fonction de leur spécificité. Les hapax et les termes avec une spécificité équivalente à celle d'un hapax sont éliminés. Voici un extrait de la liste produite par TermSuite où le premier chiffre indique le rang, T s'il s'agit d'un terme, V s'il s'agit d'une variante, suivi de la forme fléchie du terme ou de la variante la plus fréquente.

```
20 T tower
21 T wind farm
21 V wind energy farm
21 V wind power farms
21 V wind turbine farms
```

Le terme *wind farm* fait partie de la liste des termes candidats extraits du corpus des énergies éoliennes (cf. section 5). Il est le 21^e candidat et est accompagné de trois variantes terminologiques.

5. Ressources expérimentales

Dans cette section nous détaillons les corpus et les listes de référence utilisées dans nos expériences ainsi que les paramètres des différentes approches.

5.1. *Corpus*

Les expériences ont été menées sur deux corpus comparables spécialisés.

Le premier corpus comparable relève du domaine des énergies renouvelables et est disponible en trois langues : le français, l'anglais et l'espagnol. Les corpus ont été collectés à partir des pages Web en utilisant le *crawler* BABOUK (De Groc, 2011). BABOUK prend en entrée une liste de termes spécifiques à un domaine, appelée *liste d'amorces de termes*, et trouve des textes sur le Web qui traitent du domaine spécialisé dénoté par ces termes. Lors de la première itération (de collection des pages Web), cette liste d'amorces de termes est étendue en s'appuyant sur les nouveaux termes se trouvant dans les pages Web collectées. Pour élargir la recherche, les amorces de termes sont combinées de manière aléatoire pour former une requête. Cette dernière est ensuite soumise à un moteur de recherche qui va retourner les n meilleures pages qui correspondent à cette recherche. Ensuite, des filtres définis choisissent les pages qui sont riches en terminologies spécifiques au domaine et ignorent les pages qui ne sont pas propres, après conversion au format texte, celles qui sont de très petite taille ou de très grande taille, etc. Certains corpus monolingues ont été étendus avec des documents recueillis manuellement sur le Web afin d'atteindre la taille minimale fixée

Synonymes des termes anglais

wind turbine	wind machine
power supply	energy supply
power plant	electricity plant
savonius model	savonius type
energy output	energy production
sea wind farm	offshore wind farm
wind farm	wind power plant
wind turbine	aeroturbine

Synonymes des termes français

énergie renouvelable	énergie durable
centrale électrique	centrale éolienne
unité de stockage	dispositif de stockage
arbre primaire	arbre lent
force du vent	vitesse du vent
éolienne	moulin à vent

Synonymes des termes espagnols

ángulo de paso	ángulo de calaje
extremo de la pala	punta de la pala
mapa de vientos	mapa eólico
coeficiente de potencia	coeficiente de rendimiento
implantación de las máquinas	implantación de aerogeneradores
aerogenerador	torre eólica

Tableau 2. *Exemples de synonymes anglais, français et espagnols de termes complexes extraits de ressources terminologiques du domaine des énergies renouvelables*

à 300 000 mots. La collecte a été effectuée en 2010. Le tableau 4 résume la taille des corpus français, anglais et espagnol en nombre de mots et d'articles[8].

Le deuxième corpus comparable relève du domaine du cancer du sein. Il est disponible en français et en anglais. La collection des textes a été construite à partir de publications scientifiques collectées sur le portail *Elsevier*[9] et sur *Google Scholar*[10]. Les documents ont été collectés de manière à ce qu'ils répondent aux critères de comparabilité suivants :

8. Le corpus des énergies renouvelables est téléchargeable à l'URL : `www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`
9. `www.elsevier.com`
10. `scholar.google.com`

Synonymes des termes anglais

anti-oestrogens therapy	anti-oestrogens treatment
invasive carcinoma	infiltrating carcinoma
tumour tissue	cancerous tissue
adjuvant therapy	adjuvant treatment
complete breast prosthesis	full breast prosthesis
invasive ductal carcinoma	invasive ductal cancer

Synonymes des termes français

radiographie mammaire	radiographie du sein
cellule de la tumeur	cellule cancéreuse
curage du ganglion	ablation du ganglion
reconstruction du sein	reconstruction mammaire
reconstruction mammaire différée	reconstruction mammaire secondaire
reconstruction mammaire différée	reconstruction du sein différée

Tableau 3. *Exemples de synonymes anglais et français de termes complexes dans le domaine du cancer du sein extraits de Termium*

	Fr	**En**	**Es**
Nb. de mots	313 943	314 549	453 953
Nb. d'articles	11	28	46

Tableau 4. *Caractéristiques des corpus du domaine des énergies renouvelables*

(a) contenir le terme clé *cancer du sein* pour le français et son équivalent *breast cancer* pour l'anglais ;

(b) être publiés dans la période 2001-2008.

Le tableau 5 précise les caractéristiques du corpus du cancer du sein.

	Fr	**En**
Nb. de mots	267 180	198 244
Nb. de documents	78	70

Tableau 5. *Caractéristiques des corpus du domaine du cancer du sein*

Tous les corpus sont prétraités linguistiquement en utilisant la tokénisation, l'analyse morphosyntaxique et la lemmatisation à l'aide de TreeTagger.

5.2. *Listes de référence*

Les listes de référence ont été construites à partir de diverses ressources terminologiques. Seules ont été retenues les ressources qui recensaient dans leurs fiches des termes synonymes. Dans de telles bases, les synonymes ne sont pas systématiquement présents, et pour les fiches en comprenant, la nature de la synonymie est diverse. Nombre de synonymes proposés dans ces ressources sont des mots liés par d'autres types de relations sémantiques, comme la quasi-synonymie ou l'hyperonymie. De nombreuses variantes de termes complexes comme les abréviations, les constructions syntaxiques concurrentes comme N P N A/N A P N ou N P N/N Ar avec Ar adjectif relationnel pour le français, les réductions lexicales y sont listées. Toutes ces variantes de termes complexes ont été écartées de nos listes car elles peuvent être détectées à l'aide d'une analyse syntagmatique.

Concernant le domaine des énergies renouvelables, les termes complexes français ont été sélectionnés à partir de Terminalf[11]et ses 84 fiches terminologiques. Les synonymes sont indiqués sous le champ `forme concurrente`. Il existe 56 termes qui acceptent entre 1 et 3 formes concurrentes soit 76 synonymes. À ces 76 synonymes, nous avons ajouté 14 autres synonymes ou quasi-synonymes déduits à la lecture des fiches terminologiques, en particulier dans les champs `définition` ou `genre`. Le contenu du champ `genre` fait souvent référence à un hyperonyme qui peut être utilisé comme une variante synonymique en contexte. N'ont pas été considérées comme des synonymes, les formes concurrentes qui sont des variantes de réduction portant :

– sur les éléments fonctionnels comme *rotor Darrieus* et *rotor de Darrieus* ;

– sur les composants lexicaux comme la variante *traînée* pour le terme *traînée aérodynamique* ou encore la variante *coefficient de puissance* pour le terme *coefficient de puissance du rotor*.

Pour l'espagnol, nous avons collecté 64 termes complexes et leurs synonymes à partir du Lexique panlatin de l'énergie éolienne[12]. Ce lexique contient 300 entrées en 8 langues. Certains termes comprennent des synonymes disponibles dans deux zones géographiques différentes : l'Espagne et le Mexique. Par exemple, le terme espagnol *aerogenerador de dos palas* (*éolienne bipale*) aura l'équivalent suivant en espagnol mexicain *aerogenerador de doble aspa*. Nous avons considéré ces variantes terminologiques géographiques comme des synonymes. En comparaison de Terminalf qui ne contient que des termes simples ou des termes complexes composés de deux unités lexicales, le Lexique panlatin contient aussi de nombreux termes de longueur supérieure à deux unités lexicales pleines comme *aerogenerador de eje vertical con geometría variable* (*éolienne à axe vertical et géométrie variable*). Les synonymes recensés contiennent de nombreuses variantes que nous n'avons pas incluses dans nos listes :

11. `terminalf.scicog.fr`
12. `www.realiter.net/wp-content/uploads/2013/06/pan-energie-power.pdf`

– les variantes morphologiques portant sur les composants du terme complexe comme les équivalences groupe prépositionnel et adjectif *efecto de aceleración*/*efecto acelerador* avec *acelerador* dérivé du nom *aceleración* ;

– les variantes de réduction ou d'augmentation de terme complexe que la réduction ou l'augmentation porte sur l'un des composants autonomes du terme comme *efecto de pérdida aerodinámica*/*efecto de entrada en pérdida aerodinámica* ou sur l'un des composants morphologiques comme *acoplamiento dinámico*/*acoplamiento aerodinámico* ;

– les variantes de substitution des prépositions comme *enfriamiento del aire*/*enfriamiento por aire* ;

– les constructions syntaxiques concurrentes où les composants en position d'expansion du terme complexe ont été inversés : *aerogenerador para vientos de baja velocidad*/*aerogenerador para bajas velocidades de viento*.

Les termes anglais ont été sélectionnés à partir du glossaire en ligne (Gipe, 2004) et de la banque de données terminologiques Termium[13]. Nous avons collecté 84 termes complexes et leurs synonymes en écartant les variantes qui sont moins diversifiées que pour le français et l'espagnol. Les variantes prédominantes sont les variantes de réduction lexicale comme *wind power plant*/*wind plant*.

Après projection dans les corpus du domaine de l'éolien respectifs de chaque langue, nous avons obtenu 34 termes français, 20 termes anglais et 26 termes complexes espagnol associés à leurs synonymes.

Nous avons procédé de manière semblable pour constituer les synonymes de termes complexes du domaine du cancer du sein. Nous avons principalement extrait les synonymes de Termium, là encore en supprimant les mêmes types de variantes que pour le domaine de l'éolien. Après projection dans le corpus du cancer du sein dans chaque langue, nous avons obtenu 20 termes français et 16 termes anglais associés à leurs synonymes.

La taille réduite des listes de référence que ce soit pour le domaine de l'éolien ou du cancer du sein peut s'expliquer par le fait que les termes suivent le principe de monosémie et mononymie comme le rappellent Bowker et Hawkins (2006), p. 83 : « un terme correspond à un concept, et un concept n'est désigné que par un seul terme »[14]. Une autre raison est que les corpus de spécialité étant souvent de petite taille, ceci induit un nombre limité de termes spécialisés et de variantes synonymiques. Enfin, nombre de variantes synonymiques sont contextuelles, et il est difficile pour un terminologue de prédire et de détecter toutes les variantes synonymiques qui peuvent être produites.

13. `www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng`

14. « *A term should be applied to a single concept, and a concept should be designed by only one term. So, synonyms of terms are rare phenomena.* »

5.3. *Méthode à base de dictionnaire*

Nous avons utilisé la méthode proposée par Hamon et Nazarenko (2001). Pour extraire les synonymes des termes simples en français nous avons utilisé le dictionnaire en ligne DES[15]. Il contient 49 168 entrées et 201 511 relations de synonymie relevant de la langue générale. Pour l'anglais, le dictionnaire de synonymes a été construit en utilisant la base de données lexicale WordNet[16] qui contient environ 117 000 entrées (synsets). La principale relation présente dans WordNet est la synonymie.

5.4. *Paramètres de l'approche distributionnelle*

Nous avons besoin de fixer trois paramètres concernant l'approche distributionnelle, à savoir :

1) la taille de la fenêtre contextuelle qui sert à construire le vecteur de contexte (Morin *et al.*, 2007 ; Gamallo Otero, 2008) ;

2) la mesure d'association (le taux de vraisemblance (TV) (Dunning, 1993), l'information mutuelle (IM) (Fano, 1961), le *discounted odds-ratio* (DOR) (Evert, 2008)) qui sert à mesurer la force de la relation entre les mots ;

3) la mesure de similarité (l'indice du Jaccard pondéré (JAC) (Grefenstette, 1994), le cosinus (COS) (Salton et Lesk, 1968)) qui sert à mesurer la similarité entre les vecteurs de contexte des mots.

Pour construire le vecteur de contexte, nous avons choisi une taille de fenêtre égale à 7, c'est-à-dire 3 mots précédant et trois mots suivant le mot à caractériser. Comme mesures d'association nous avons utilisé IM, TV et DOR et comme mesures de similarité nous avons utilisé COS et JAC.

Pour évaluer les différentes combinaisons de paramètres de l'analyse distributionnelle, nous avons adopté la mesure de la MAP (*Mean Average Precision*) (Manning *et al.*, 2008) ainsi que la précision aux TOP1, TOP5 et TOP10 qui examine le premier candidat, puis les ensembles des 5 et 10 premiers candidats.

$$MAP = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rang_i} \qquad [6]$$

où $|W|$ correspond à la taille de la liste d'évaluation et $Rang_i$ correspond au rang du candidat synonyme correct i.

Si les TOP1, TOP5 et TOP10 renvoient la précision des systèmes aux premiers rangs, la MAP, quant à elle, renvoie la précision moyenne globale des systèmes. Ainsi,

15. www.crisco.unicaen.fr/des/synonyms
16. wordnetweb.princeton.edu/perl/webwn/

un système A ayant une meilleure précision au TOP1 qu'un système B peut avoir une moins bonne MAP, si dans la suite du classement, ses synonymes sont moins bien classés que ceux du système B.

6. Résultats

Dans cette section, nous présentons les résultats des expériences menées sur le corpus des énergies renouvelables pour le français, l'anglais et l'espagnol, et sur le corpus du cancer du sein pour le français et l'anglais. La méthode de Hamon et Nazarenko (2001) est notée *référence* et notre approche est notée *Semi-Comp*. Les candidats renvoyés par Semi-Comp sont ordonnés de trois manières différentes :

– le rang par score d'association où les termes complexes candidats sont ordonnés par rapport à la valeur décroissante du score d'association du composant substitué (Semi-Comp (assoc)) ;

– le rang par effectif où les termes complexes candidats sont ordonnés par rapport à leur fréquence d'occurrence dans le corpus (Semi-Comp (effec)) ;

– le rang par rapport de fréquence où les termes complexes candidats sont ordonnés en fonction de leur rapport de fréquence avec le terme complexe dont les synonymes sont recherchés (Semi-Comp (rdf)).

Nous étudions deux manières de filtrer les candidats termes synonymes, à savoir un filtrage à l'aide d'une liste de n-grammes, noté *sans patrons syntaxiques* (illustré dans la deuxième colonne de chaque tableau de résultats) et un filtrage par une liste de termes complexes candidats, noté *avec patrons syntaxiques* (illustré dans la troisième colonne de chaque tableau de résultats). Enfin, nous expérimentons trois configurations correspondant à trois combinaisons impliquant chacune une mesure d'association et une mesure de similarité différentes. Ces trois combinaisons ne représentent qu'un sous-ensemble de la combinatoire entre mesures d'association et de similarité mais sont classiquement utilisées en analyse distributionnelle et ont montré leur efficacité comparativement à d'autres combinaisons. Les configurations sont indiquées sur l'axe des ordonnées des tableaux. Par exemple, la légende IM-COS signifie que l'approche distributionnelle utilisée dans Semi-Comp exploite l'information mutuelle comme mesure d'association et le cosinus comme mesure de similarité.

Pour chaque configuration, les meilleurs résultats au TOP1, TOP5, TOP10 et MAP sont indiqués en gras. Le tableau 6 donne les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus français des énergies renouvelables. La première remarque concerne l'approche de référence où les résultats obtenus sont très faibles en comparaison avec l'approche semi-compositionnelle quelle que soit la configuration choisie. Les meilleurs résultats sont obtenus avec l'approche Semi-Comp (assoc) avec la configuration IM-COS associée à un filtrage par patrons syntaxiques et donnent une précision au TOP1 de 26,4 %, au TOP5 de 55,8 % et une MAP de 38,5 %. La meilleure précision au TOP5 de 55,8 % est partagée par toutes les configurations avec Semi-Comp (effec) quel que soit le filtrage. La meilleure

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	0,25	-	-	-	0,25
TV-JAC	Semi-Comp (assoc)	17,6	52,9	52,9	31,7	23,5	52,9	58,8	38,2
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2
IM-COS	Semi-Comp (assoc)	17,6	47,0	52,9	29,4	**26,4**	**55,8**	61,7	**38,5**
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2
DOR-COS	Semi-Comp (assoc)	14,7	44,1	47,0	26,9	23,5	52,9	**70,5**	37,9
	Semi-Comp (effec)	14,7	**55,8**	64,7	31,2	17,6	**55,8**	64,7	34,8
	Semi-Comp (rdf)	5,88	23,5	35,2	14,9	5,88	47,0	52,9	23,2

Tableau 6. *Résultats des expériences sur le corpus français des énergies renouvelables*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	3,63	-	-	-	3,63
TV-JAC	Semi-Comp (assoc)	20,0	55,0	65,0	36,1	15,0	60,0	70,0	38,2
	Semi-Comp (effec)	45,0	80,0	90,0	59,2	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,2	15,0	40,0	45,0	26,7
IM-COS	Semi-Comp (assoc)	15,0	55,0	65,0	32,8	15,0	60,0	70,0	35,8
	Semi-Comp (effec)	**45,0**	**80,0**	**90,0**	59,2	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,2	15,0	40,0	45,0	26,7
DOR-COS	Semi-Comp (assoc)	15,0	40,0	55,0	27,2	10,0	40,0	60,0	26,8
	Semi-Comp (effec)	**45,0**	**80,0**	**90,0**	60,0	**45,0**	**80,0**	**90,0**	**61,1**
	Semi-Comp (rdf)	20,0	40,0	40,0	28,3	15,0	40,0	45,0	26,7

Tableau 7. *Résultats des expériences sur le corpus anglais des énergies renouvelables*

précision au TOP10 de 70,5 % est fournie par Semi-Comp (assoc) avec la configuration DOR-COS associée à un filtrage par patrons syntaxiques.

Le tableau 7 donne les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus anglais des énergies renouvelables. Nous constatons que les résultats de l'approche de référence sont de nouveau très faibles

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	8,09	-	-	-	8,09
TV-JAC	Semi-Comp (assoc)	7,69	38,4	42,3	19,7	11,5	50,0	61,5	29,8
	Semi-Comp (effec)	15,3	34,6	42,3	24,8	15,3	38,4	46,1	25,7
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	23,1
IM-COS	Semi-Comp (assoc)	15,3	38,4	42,3	25,3	**23,0**	**57,6**	**65,3**	**37,0**
	Semi-Comp (effec)	15,3	34,6	42,3	24,8	15,3	38,4	46,1	25,7
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	22,5
DOR-COS	Semi-Comp (assoc)	0	30,7	42,3	14,4	19,2	53,8	**65,3**	32,6
	Semi-Comp (effec)	15,3	34,6	42,3	24,9	15,3	38,4	50,0	25,8
	Semi-Comp (rdf)	3,84	11,5	15,3	7,90	11,5	30,7	42,3	22,5

Tableau 8. *Résultats des expériences sur le corpus espagnol des énergies renouvelables*

en comparaison de ceux de l'approche semi-compositionnelle. Les meilleurs résultats sont obtenus avec Semi-Comp (effec) associé à un filtrage par patrons syntaxiques et donnent une précision au TOP1 de 45 %, au TOP5 de 80 %, au TOP10 de 90 % et une MAP de 61,1 %. La configuration n'influence pas les résultats et le filtrage ne l'influence qu'à la marge.

Les résultats de la dernière expérience dans le domaine des énergies renouvelables et qui concernent le corpus espagnol sont représentés dans le tableau 8. Pour le corpus espagnol comme pour le corpus français, c'est le classement par score d'association Semi-Comp (assoc) exploitant la configuration IM-COS associée à un filtrage par patrons syntaxiques qui donne les meilleurs résultats avec une précision au TOP1 de 23 %, au TOP5 de 57,6 %, au TOP10 de 65,3 % et une MAP de 37 %. La meilleure précision au TOP10 de 65,3 % est obtenue par Semi-Comp (assoc) avec la configuration DOR-COS associée à un filtrage par patrons syntaxiques.

D'une manière générale, l'approche Semi-Comp donne toujours de meilleurs résultats que l'approche de référence quelle que soit la configuration choisie. Concernant la manière de classer les candidats, les classements par score d'association et par effectif sont à privilégier et à associer à un filtrage par patrons syntaxiques. Les trois configurations testées donnent des résultats très proches qui varient selon la langue et la méthode de classement choisies. Le classement par effectif montre une grande stabilité vis-à-vis de la configuration choisie et ceux pour les trois langues.

Les tableaux 9 et 10 fournissent les résultats de l'approche de référence ainsi que ceux de l'approche semi-compositionnelle sur le corpus du cancer du sein, respectivement pour le français et l'anglais.

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	4,92	-	-	-	4,92
TV-JAC	Semi-Comp (assoc)	5,26	15,7	47,3	13,9	26,3	47,3	57,8	34,7
	Semi-Comp (effec)	**31,5**	**42,1**	**73,6**	40,3	31,5	52,6	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3
IM-COS	Semi-Comp (assoc)	10,5	36,8	52,6	19,9	21,0	**52,6**	57,8	35,7
	Semi-Comp (effec)	**31,5**	42,1	**73,6**	40,3	**31,5**	**52,6**	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3
DOR-COS	Semi-Comp (assoc)	15,7	42,1	52,6	27,1	26,3	57,8	57,8	38,5
	Semi-Comp (effec)	**31.5**	42,1	**73,6**	**40,4**	**31,5**	**52,6**	63,1	39,9
	Semi-Comp (rdf)	0	15,7	31,5	8,6	15,7	47,3	47,3	30,3

Tableau 9. *Résultats des expériences sur le corpus français du cancer du sein*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
	Référence	-	-	-	7,03	-	-	-	7,03
TV-JAC	Semi-Comp (assoc)	6,66	20,0	26,6	13,3	0	26,6	40,0	16,7
	Semi-Comp (effec)	20,0	46,6	46,6	29,0	**26,6**	**53,3**	**53,3**	**38,1**
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0
IM-COS	Semi-Comp(assoc)	6,66	20,0	26,6	12,6	6,66	26,6	53,3	18,1
	Semi-Comp (effec)	20,0	46,6	46,6	29,0	**26,6**	**53,3**	**53,3**	**38,1**
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0
DOR-COS	Semi-Comp (assoc)	6,66	13,3	20,0	11,0	6,66	20,0	33,3	15,7
	Semi-Comp (effec)	20,0	40,0	40,0	25,7	**26,6**	46,6	46,6	34,8
	Semi-Comp (rdf)	0	6,66	6,66	3,0	0	20,0	33,3	9,0

Tableau 10. *Résultats des expériences sur le corpus anglais du cancer du sein*

À nouveau, l'approche semi-compositionnelle donne de meilleurs résultats que l'approche de référence. Pour le français, nous constatons que Semi-Comp (effec) en utilisant la configuration DOR-COS associée à un filtrage sans patrons syntaxiques donne les meilleurs résultats avec une précision pour le TOP1 de 31,5 %, pour le TOP10 de 73,6 % et une MAP de 40,4 %. Néanmoins, la précision au TOP1 de 31,5 % est obtenue par tous les classements quels que soient la configuration et le filtrage, et la précision pour le TOP5 de 52,6 % est proposée par Semi-Comp (effec) pour le filtrage avec patrons syntaxiques et à 42,1 % quelle que soit la configuration sans patrons syn-

taxiques. L'amélioration de la MAP pour le filtrage sans patrons syntaxiques apparaît entre les rangs 5 à 10 uniquement, ce qui en atténue l'intérêt.

Pour l'anglais, c'est encore Semi-Comp (effec) en utilisant les configurations TV-JAC et IM-COS associées à un filtrage avec patrons syntaxiques qui donne les meilleurs résultats avec une précision au TOP1 de 26,6 %, au TOP5 de 53,3 %, au TOP10 de 53,3 % et une MAP de 38,1 %.

Les résultats de Semi-Comp qui utilisent le classement par rapport de fréquence sont très faibles et confirment l'inefficacité de ce type de classement. La méthode semi-compositionnelle fournissant les meilleurs résultats au TOP1 en moyenne, un classement par effectif et un filtrage des candidats par patrons syntaxiques, ne propose qu'une fois sur quatre un synonyme. En revanche, la méthode semi-compositionnelle adoptant un classement par effectif, la configuration IM-COS et un filtrage des candidats par patrons syntaxiques donne toujours les meilleurs résultats au TOP5 quels que soit le domaine ou la langue. Le score le plus faible de précision au TOP5 est de 52,6 % pour le français et le plus élevé est de 80 % pour l'anglais dans le domaine des énergies renouvelables. Cela signifie que l'examen du TOP5 fournit en moyenne entre 2 à 3 synonymes. Le dépouillement des résultats fournis par la méthode semi-compositionnelle peut donc à moindre effort aider à recenser les variantes synonymiques contextuelles.

Certains résultats quantitatifs nous ont donné envie de les examiner plus finement. Le filtrage à l'aide de patrons syntaxiques fournit généralement de meilleurs résultats que le filtrage par n-grammes. Ce résultat est vérifié pour l'anglais dans le domaine du cancer du sein. Par exemple le filtrage par patrons syntaxiques propose en rang 2 le couple *nipple prosthesis*/*nipple reconstruction* et en rang 12 le couple *os biopsy*/*bone biopsy* qui n'apparaissent pas avec le filtrage par n-grammes. Le filtrage par patrons syntaxiques donne aussi généralement de meilleurs rangs aux couples synonymes, comme *breast tissue*/*mammary tissue*, rang 7 avec filtrage par patrons syntaxiques, et au rang 29 avec filtrage par les n-grammes. En revanche, dans le domaine des énergies renouvelables, le filtrage par patrons syntaxiques est équivalent à celui proposé par les n-grammes. Les mêmes couples sont proposés au rang 1 comme *wind generator*/*wind turbine*, *lattice construction*/*lattice tower*, *drive train*/*power train*.

Bien entendu, certains couples de synonymes sont filtrés par les patrons syntaxiques alors qu'ils ne le sont pas avec les n-grammes. Il s'agit pour l'anglais de termes comportant un participe présent comme *infiltrating carcinoma* dans le domaine du cancer du sein. Le patron $V_{\text{participe présent}}$ N produit trop de candidats termes incorrects et ne fait pas partie des patrons de termes complexes anglais. Des remarques similaires peuvent être faites dans d'autres langues. Pour le français, les couples *protocole radiothérapie*/*protocole cmf* et *protocole chimiothérapie*/*protocole cmf* sont proposés en rang 1 et 2 avec le filtrage par n-grammes et sont absents du filtrage par patrons syntaxiques, sans doute à cause d'une erreur de lemmatisation du corpus.

Pour l'espagnol et le domaine des énergies renouvelables, le meilleur classement est obtenu très nettement par le classement par association contrairement aux résultats

pour le français et l'anglais dans les deux domaines étudiés qui privilégie un classement par effectif. La liste de référence de l'espagnol comportait des spécificités par rapport aux autres listes de référence : termes complexes de longueur plus importante, variantes géographiques d'espagnol. Plusieurs candidats dans les premiers rangs sont partagés par les deux classements comme *torre eólica*[ESP]*turbina eólica*[ESP] ou *flujo de viento*[ESP]/*corriente de viento*[ESP]. Le fait d'avoir inclus des variantes géographiques Espagne [ESP] et Mexique [MEX] ne pose pas de problèmes particuliers. Leurs occurrences dans le corpus apparaissent en haut des deux classements comme *energía eólica*[MEX]/*potencia eólica*[ESP] et *mapa de los vientos*[ESP]/*mapa eólico*[MEX]. Tous les synonymes de rang 1 fournis par le classement par effectif et par le classement par association sont des termes synonymes proposés par le Lexique panlatin. Les différences portent sur les positions de certains couples du Lexique panlatin qui apparaissent plus loin dans le classement par effectif que dans le classement par association, comme par exemple *implantación máquina*[ESP]/*implantación aerogeneradores*[ESP], *implantación aerogeneradores*[ESP]/*instalación aerogeneradores*[ESP] en rang 8 proposé pour le classement par l'effectif et en rang 1 pour le classement par association. Un couple discutable de notre liste de référence puisqu'il n'apparaît ni dans le Lexique panlatin, ni dans Termium : *abrigo de torre*/*sombra de torre* apparaît en rang 1 pour le classement par association et en rang 22 pour le classement par effectif. La taille des listes étant réduite, ces quelques faits expliquent la différence des résultats.

En conclusion, les résultats obtenus sur les deux corpus confirment l'efficacité et l'opérationabilité pour la mise à jour de ressources terminologiques de l'approche semi-compositionnelle pour l'extraction de termes complexes reliés sémantiquement. Les variantes synonymiques extraites relèvent de causes différentes : dénominations concurrentes géographiques comme pour l'espagnol et variantes contextuelles. L'approche semi-compositionnelle n'est pas sensible à ces différentes variantes.

7. Discussion

Peu de travaux se sont intéressés à l'identification de synonymes de termes complexes. À notre connaissance, seuls Hamon et Nazarenko (2001) se sont attelés à cette tâche en exploitant le principe de compositionnalité des termes complexes. Les faibles résultats obtenus par l'approche de référence peuvent s'expliquer de deux façons : la première est que dans le domaine de spécialité les dictionnaires de synonymes sont rares ou non disponibles. La seconde est que le fait de s'appuyer sur un dictionnaire de langue générale peut conduire à l'extraction de termes complexes inadaptés et/ou hors domaine comme cela a été montré dans nos expériences.

La principale contribution de notre approche est l'utilisation de l'analyse distributionnelle à la place des dictionnaires pour identifier les synonymes et les mots sémantiquement liés des composants d'un terme complexe. L'analyse distributionnelle permet d'identifier des mots en relation sémantique qui sont utilisés pour construire des synonymes de termes complexes en exploitant, là encore, le principe de compositionnalité.

Nous avons aussi identifié le fait que les synonymes des termes complexes ne sont pas toujours composés de synonymes de leurs parties, et que l'utilisation de mots sémantiquement liés était plus souhaitable pour cette tâche. Notre approche peut être appliquée à n'importe quel terme complexe qui suit les règles R_1^G et R_2^G et pour n'importe quelle catégorie de variantes synonymiques, géographiques ou contextuelles. Le classement des candidats synonymes par leur effectif associé à un filtrage par patrons syntaxiques fournit les meilleurs résultats.

Si nos différentes expériences ont montré que le classement des cinq premiers candidats synonymes comprenait la moitié des synonymes corrects, il reste une marge de progression. Outre l'identification des synonymes de termes complexes, et sachant que l'analyse distributionnelle peut renvoyer des termes simples antonymes (*chaud/froid*) ou contrastifs (*blanc/noir*), il est légitime de s'intéresser à l'extraction de termes complexes antonymes ou contrastifs *via* l'approche semi-compositionnelle. Pour ce faire, nous avons construit une liste d'antonymes de termes complexes incluant des termes constrastifs pour le corpus du cancer du sein et nous avons mené une expérience supplémentaire pour identifier les antonymes des termes complexes qui suivent les règles R_1^G et R_2^G. Nous considérons que deux termes complexes sont antonymes si leurs parties le sont. Les listes de référence pour l'anglais et le français dans le domaine médical contiennent respectivement 12 et 9 paires de termes complexes antonymes. Ces listes ont été encore plus difficiles à construire que pour les termes synonymes. Les antonymes étaient absents de toutes les ressources terminologiques que nous avions utilisées pour la construction les listes de synonymes. Nous avons consulté le *Trésor de la langue française*, et pour les antonymes listés comme *artificiel/naturel* cherché des termes complexes ayant au moins un élément en commun et apparaissant dans le corpus comme *ménopause artificielle/ménopause naturelle*. Nous nous sommes limités au corpus du domaine médical. Pour l'anglais, nous avons traduit les couples d'antonymes français et vérifié leur appartenance au corpus et inclus quelques antonymes marqués lexicalement comme *malignant tissue/non-malignant tissue*.

Comme le montrent les résultats des tableaux 11 et 12, l'approche semi-compositionnelle est aussi capable d'identifier les antonymes des termes complexes. Dans la majorité des cas, c'est le classement par score d'association qui donne les meilleurs résultats avec une MAP de 40,6 % pour le français (Semi-Comp(assoc) avec DOR-COS avec filtrage par patrons syntaxiques) et 27,6 % pour l'anglais (Semi-Comp (assoc) avec TV-JAC avec filtrage sans patrons syntaxiques). Le classement par effectif comme celui par rapport de fréquence donnent de très faibles résultats.

Une première interprétation de ces résultats est de dire qu'en utilisant l'analyse distributionnelle, les termes simples antonymes sont mieux classés que les termes simples synonymes si l'on s'appuie sur un classement par score d'association. Ainsi, un classement par effectif favoriserait plutôt l'identification des synonymes de termes complexes et un classement par score d'association l'extraction d'antonymes de termes complexes. Ce résultat conforte l'observation faite par les linguistes Morlane-Hondère (2013) que les contextes partagés par un terme et son antonyme sont peu nombreux et diffèrent des contextes partagés par un terme et ses synonymes. Le score d'associa-

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
TV-JAC	Semi-Comp (assoc)	**37,5**	37,5	37,5	39,7	25,0	**50,0**	**62,5**	37,2
	Semi-Comp (effec)	0	0	25,0	3,4	0	12,5	37,5	7,4
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4
IM-COS	Semi-Comp(assoc)	25,0	**50,0**	50,0	34,5	37,5	**37,5**	50,0	40,0
	Semi-Comp (effec)	0	0	25,0	3,4	0	12,5	37,5	7,4
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4
DOR-COS	Semi-Comp (assoc)	12,5	37,5	50,0	27,8	**37,5**	37,5	**62,5**	**40,6**
	Semi-Comp (effec)	0	0	25,0	3,8	0	12,5	37,5	7,6
	Semi-Comp (rdf)	0	0	0	2,0	0	25,0	37,5	8,4

Tableau 11. *Résultats des expériences pour l'identification des antonymes sur le corpus français du cancer du sein*

	Méthode	Sans patrons syntaxiques				Avec patrons syntaxiques			
		P1	P5	P10	MAP	P1	P5	P10	MAP
TV-JAC	Semi-Comp (assoc)	**25,0**	**25,0**	33,3	**27,6**	16,6	**25,0**	**41,6**	23,6
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8
IM-COS	Semi-Comp(assoc)	16,6	16,6	16,6	17,8	16,6	16,6	25,0	18,1
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8
DOR-COS	Semi-Comp (assoc)	16,6	16,6	25,0	18,8	16,6	16,6	25,0	18,9
	Semi-Comp (effec)	8,33	8,33	16,6	10,8	8,33	16,6	25,0	12,8
	Semi-Comp (rdf)	0	8,33	16,6	6,1	8,33	16,6	16,6	12,8

Tableau 12. *Résultats des expériences pour l'identification des antonymes sur le corpus anglais du cancer du sein*

tion mettrait en lumière le caractère remarquable au sens collocationnel des contextes partagés par un terme et son antonyme et le classement par effectif privilégierait l'importance du nombre de contextes partagés par un terme et son synonyme.

8. Conclusion

Nous avons présenté dans cet article l'approche semi-compositionnelle pour l'extraction de synonymes de termes complexes. Fondée sur le principe de distributionnalité et de compositionnalité, notre approche a montré des gains significatifs en comparaison avec l'approche état de l'art. Si plus d'expériences sont sûrement nécessaires, les résultats encourageants de l'approche semi-compositionnelle confirment l'intérêt de combiner les analyses distributionnelle et compositionnelle pour l'identification de termes complexes synonymes. Dans des travaux futurs, nous nous attellerons à l'extraction de synonymes de termes complexes qui ne sont pas compositionnels et inversement, nous explorerons l'extraction de synonymes de termes simples. Perinet et Hamon (2013) proposent une méthode hybride pour l'acquisition de relations sémantiques fondées sur une normalisation contextuelle qui pourrait donner de bons résultats pour l'extraction de termes synonymes tout comme une extraction des contextes par une analyse syntaxique.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet TERMITH (`www.atilf.fr/ressources/termith`) a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-2-CORD-0029.

9. Bibliographie

Blondel V. D., Senellart P., « Automatic Extraction of Synonyms in a Dictionary », 2002.

Bowker L., Hawkins S., « Variation in the organization of medical terms - Exploring some motivations of term choice », *Terminology*, vol. 12, n° 1, p. 79-110, 2006.

De Groc C., « Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction », *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT'11, IEEE Computer Society, Washington, DC, USA, p. 497-498, 2011.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Evert S., *Corpus Linguistics. An International Handbook*, vol. 2, De Gruyter Mouton, chapter Corpora and collocations, p. 1212-1248, 2008.

Fano R. M., *Transmission of Information : A Statistical Theory of Communications*, MIT Press, Cambridge, MA, USA, 1961.

Ferret O., « Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus. », *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds), *LREC*, 2010.

Ferret O., « Identifying bad semantic neighbors for improving distributional thesauri », *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, p. 561-571, 2013.

Firth J. R., « A synopsis of linguistic theory 1930-1955 », *in* J. R. Firth, W. Haas, M. A. K. Halliday (eds), *Studies in Linguistic Analysis*, Special volume of the Philological Society, Blackwell, Oxford, p. 1-32, 1957.

Gamallo Otero P., « Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora », *Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, Marrakech, Morroco, p. 19-26, 2008.

Gipe P., *Wind power : renewable energy for home, farm, and business*, Chelsea Green Pub. Co., 2004.

Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publisher, Boston, MA, USA, 1994.

Hagiwara M., « A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features », *Proceedings of the ACL-08 : HLT Student Research Workshop*, Association for Computational Linguistics, Columbus, Ohio, p. 1-6, June, 2008.

Hamon T., Nazarenko A., « Detection of synonymy links between terms : experiment and results », *Recent Advances in Computational Terminology*, John Benjamins, p. 185-208, 2001.

Harris Z. S., « Distributional structure. », *Word*, 1954.

Hindle D., « Noun Classification from Predicate-Argument Structures. », *ACL*, p. 268-275, 1990.

Kraft M., « Compositionality : The very Idea », *Research on Language and Computation*, vol. 5, n° 3, p. 287-308, 2007.

Kremer G., Erk K., Padó S., Thater S., « What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus », *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, p. 540-549, April, 2014.

Lin D., « Automatic retrieval and clustering of similar words », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 768-774, 1998.

Lin D., Zhao S., Qin L., Zhou M., « Identifying Synonyms among Distributionally Similar Words », *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

Manning D. C., Raghavan P., Schütze H., *Introduction to information retrieval*, Cambridge University Press, 2008.

Morin E., Daille B., « Revising the Compositional Method for Terminology Acquisition from Comparable Corpora », *24th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, Coling'12*, Mumbai, India, p. 1797-1810, 2012.

Morin E., Daille B., Takeuchi K., Kageura K., « Bilingual Terminology Mining – Using Brain, not brawn comparable corpora », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, p. 664-671, 2007.

Morlane-Hondère F., Une approche linguistique de l'évaluation des ressources extraites par l'analyse distributionnelle automatique, PhD thesis, Université Toulouse II Le Mirail, 2013.

Morris J., Hirst G., « Non-classical lexical semantic relations », *HLT-NAACL Worlshop on Computational Lexical semantics (CLS'04)*, ACL, p. 46-51, 2004.

Parnee B. H., Ter Meulen A., Wall R. E., *Mathematical Methods in Linguistics*, vol. 30 of *Studies in Linguistics and Philosophy*, Kluwer Academic Publishers, Dordrecht, 1990.

Périnet A., Hamon T., « Hybrid acquisition of semantic relations based on context normalization in distributional analysis », *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA2013)*, Paris Nord, France, p. 113-122, October, 2013.

Resnik P., Selection and Information : A class-based approach to lexical relationships, PhD thesis, University of Pennsylvania, 1993.

Rocheteau J., Daille B., « TTC TermSuite - A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora », *Proceedings of the IJCNLP 2011 System Demonstrations*, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, p. 9-12, November, 2011.

Salton G., Lesk M. E., « Computer evaluation of indexing and text processing », *Journal of the Association for Computational Machinery*, vol. 15, n° 1, p. 8-36, 1968.

Van der Plas L., Tiedemann J., « Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity », *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia, 2006.

Wu H., Zhou M., « Optimizing synonym extraction using monolingual and bilingual resources », *In Proceedings of the second international workshop on Paraphrasing*, p. 72, 2003.