Analyse distributionnelle appliquée aux textes de spécialité

Réduction de la dispersion des données par abstraction des contextes

Amandine Périnet* — **Thierry Hamon*,****

* *Université Paris 13, Sorbonne Paris Cité, France*
amandine.perinet@edu.univ-paris13.fr
** *LIMSI-CNRS, BP133, Orsay, France*
hamon@limsi.fr

RÉSUMÉ. *Les modèles vectoriels utilisés pour l'analyse distributionnelle souffrent de la dispersion des données dans la matrice des contextes et du nombre important de dimensions de cette matrice. Ces limitations rendent difficile leur application aux corpus de spécialité, et les termes ne sont habituellement pas pris en compte alors qu'ils sont essentiels. Dans cet article, nous proposons une adaptation de l'analyse distributionnelle afin de pouvoir l'utiliser efficacement sur des textes de spécialité. L'approche proposée réalise une abstraction des contextes distributionnels pour réduire la dispersion des données et ainsi améliorer la qualité des regroupements tout en y incluant les termes. Nous avons évalué notre approche sur deux corpus médicaux. L'analyse des résultats montre que tout en permettant la prise en compte des termes dans l'analyse distributionnelle, l'abstraction des contextes, notamment grâce à l'inclusion lexicale, permet d'obtenir des regroupements sémantiques de meilleure qualité et plus homogènes.*

ABSTRACT. *The vector space models used for the distributional analysis suffer from data sparseness in the context matrix and from the great number of dimensions of the matrix. These limitations make its use on domain-specific corpora difficult and terms are usually not considered while they are essential. In this paper, we propose an adaptation of the distributional analysis in order to apply it efficiently on domain-specific corpora. The proposed approach performs an abstraction of the distributional contexts in order to reduce data sparseness and thus improve the quality of the distributional classes, and also to include terms in these classes. We evaluated our approach on two medical corpora. The results analysis shows that context abstraction, especially thanks to the lexical inclusion, leads to more semantically homogeneous classes with a better quality. The approach also achieves to take terms into account in the classes.*

MOTS-CLÉS : *analyse distributionnelle, corpus de spécialité, dispersion des données.*

KEYWORDS: *distributional analysis, domain-specific corpora, data sparseness.*

TAL. Volume 56 – n° 2/2015, pages 77 à 102

1. Introduction

Les relations entre termes jouent un rôle prépondérant dans les terminologies pour définir le sens des termes, mais aussi lors de l'utilisation des ressources terminologiques dans des applications en langue de spécialité pour augmenter leur couverture (Bodenreider *et al.*, 2002 ; McCray *et al.*, 2002) ou permettre leur adaptation (Cohen et Demner-Fushman, 2013). De nombreuses méthodes ont été proposées pour acquérir des relations sémantiques entre termes (Nastase *et al.*, 2013). Parmi celles-ci, l'analyse distributionnelle conduit au regroupement de mots sémantiquement proches en tenant compte des contextes qu'ils partagent : plus le nombre de contextes communs est élevé, plus les deux mots cibles sont sémantiquement proches (Harris, 1954 ; Firth, 1957). L'acquisition de relations sémantiques est vue comme un problème de regroupement de mots sémantiquement proches fondé sur des informations statistiques liées aux contextes et des mesures de similarité sémantique. À partir de ces regroupements, il est ainsi possible d'obtenir un nombre important de relations sémantiques qui ne sont, cependant, pas typées : il peut s'agir aussi bien de relations classiques telles que la synonymie ou l'hyperonymie, que de relations propres au domaine (Morlane-Hondère, 2013).

La mise en œuvre de l'analyse distributionnelle dans un processus automatique s'appuie généralement sur une représentation vectorielle des mots du corpus (Curran, 2004 ; Sahlgren, 2006). Un mot cible est alors un point dans un espace à n-dimensions où chaque dimension correspond à des contextes possibles, et où la valeur associée aux dimensions est le nombre d'occurrences du contexte correspondant. Le vecteur de chaque mot cible représente donc les informations contextuelles mais aussi des données statistiques distributionnelles comme le nombre de contextes et le nombre d'occurrences des contextes partagés (Turney et Pantel, 2010 ; Lund et Burgess, 1996). La similarité sémantique entre deux mots cibles est ainsi définie comme une proximité dans cet espace, calculée à l'aide du cosinus de l'angle par exemple. Les modèles vectoriels ont l'avantage de permettre une quantification facile de la proximité sémantique entre deux mots. Cependant, ils souffrent d'un problème de dispersion des données, car ils s'appuient sur des espaces aux très grandes dimensions et sur la redondance des contextes partagés alors qu'il s'agit d'évènements souvent rares (Chatterjee et Mohan, 2008). Ainsi, en considérant les espaces vectoriels comme des matrices de contexte, où les lignes sont les mots cibles du texte et les colonnes sont les contextes, on dispose généralement de matrices creuses ou éparses où beaucoup d'éléments sont à zéro car peu de contextes sont associés à un mot cible (Turney et Pantel, 2010).

Lorsqu'il s'agit de corpus de spécialité, ce problème de dispersion des données est accentué par des tailles de corpus beaucoup plus petites, un faible nombre d'occurrences du vocabulaire et un nombre de contextes partagés plus faible. Or, quand l'analyse distributionnelle est utilisée sur des corpus de spécialité, il est essentiel de prendre en compte les termes simples et complexes, à la fois dans les mots cibles, c'est-à-dire les mots regroupés, et dans les contextes des mots cibles (pour le calcul distributionnel). Ceci est généralement difficile à réaliser. Dans le cadre de traitement

de corpus monolingues, très peu de travaux existent. En revanche, pour l'extraction de lexiques bilingues à partir de corpus comparables, plusieurs travaux ont recours à l'analyse distributionnelle et portent un intérêt particulier aux termes complexes ((Daille et Morin, 2005), (Déjean *et al.*, 2002) et (Zweigenbaum et Habert, 2006) pour un aperçu général). Pour faire face au problème du faible nombre d'occurrences, Morin et Hazem (2014) utilisent un modèle de régression en amont de l'analyse distributionnelle. Ce modèle, entraîné sur des corpus de petite et de grande taille, leur permet de prédire le nombre d'occurrences de chaque contexte de manière à rendre ces valeurs plus fiables sur de nouveaux textes. Notre problématique est proche de ces travaux, qui s'appuient également sur les travaux fondateurs de Grefenstette (1994).

Ainsi, en raison de leur très faible nombre d'occurrences, les termes complexes se retrouvent généralement écartés du calcul de similarité, et les mots du corpus sont généralement considérés indépendamment du fait qu'il s'agisse de termes ou non.

Dans cet article, nous nous intéressons à l'adaptation d'une méthode d'analyse distributionnelle en prenant en compte les termes simples et complexes. Cette adaptation nous amène ainsi à aborder le problème de la dispersion des données et la réduction de la dimension de l'espace vectoriel dans le contexte des textes de spécialité. Nous émettons l'hypothèse que la suppression de la variation terminologique dans les contextes[1] ne dégrade pas trop leur sémantique, et permette d'améliorer la qualité des regroupements tout en prenant en compte les termes complexes. Pour cela, nous réalisons une abstraction des contextes à l'aide de relations sémantiques acquises automatiquement à partir de nos corpus de travail. Ainsi, les relations calculées par trois méthodes d'acquisition de relations d'hyperonymie (l'inclusion lexicale, les patrons lexico-syntaxiques et la variation terminologique) et d'une méthode d'acquisition de relations de synonymie sont utilisées pour généraliser ou normaliser les contextes distributionnels. Cette étape d'abstraction des contextes permet de réduire leur diversité ; le nombre de contextes différents est ainsi réduit.

Dans la suite de cet article, nous revenons en détail, à la section 2, sur le problème de la dispersion des données et nous exposons les solutions proposées dans les travaux précédents. La section 3 est consacrée à la description générale de la méthode distributionnelle mise en œuvre. Le processus d'abstraction des contextes est exposé à la section 4. Après avoir présenté le matériel utilisé (section 5), nous montrons les expériences réalisées et les résultats obtenus à la section 6.

2. Réduction de la dispersion des données

Les modèles vectoriels sont limités par la dispersion des données dans la matrice des contextes : beaucoup d'éléments de la matrice sont à zéro car généralement peu de contextes sont associés à un mot cible. On dispose alors d'une matrice de très faible densité, considérée comme creuse (Turney et Pantel, 2010). Sahlgren (2006) constate

1. Nous entendons par *contexte* une unité lexicale qui apparaît dans le voisinage du mot cible.

à travers ses expériences que plus de 99 % des entrées d'une matrice sont égales à zéro. Cet inconvénient est dû notamment à la distribution des mots dans le corpus (Baroni, 2009) : quelle que soit la taille du corpus, la plupart des mots ont un faible nombre d'occurrences, et un nombre de contextes très limité au regard du nombre de mots dans le corpus. La dispersion des données touche à la fois les corpus de langue générale, habituellement très volumineux (Weeds et Weir, 2005 ; van der Plas, 2008), et les textes de spécialité, souvent de plus petite taille et caractérisés par un vocabulaire avec un plus petit nombre d'occurrences. Ainsi, même dans un gros corpus tel que le BNC (100 millions de mots), moins de 14 % des mots ont un nombre d'occurrences de 20 ou plus (Baroni, 2009). Comme conséquence, les méthodes fondées sur l'analyse distributionnelle obtiennent de meilleures performances lorsque beaucoup d'informations sont disponibles, et notamment sur ces corpus volumineux, caractérisés par des nombres d'occurrences des mots du vocabulaire plus élevés. La réduction de la dispersion des données devient donc un enjeu majeur dans le cas des corpus de spécialité.

Il existe une forte corrélation entre la densité de la matrice et la performance du modèle vectoriel. Ainsi, même s'il est difficile de saisir la structure sémantique sous-jacente des matrices creuses, plus une matrice est creuse, moins le modèle vectoriel est performant sur la tâche donnée. Ce rapport est équivalent à celui liant le nombre d'occurrences des mots et la qualité des vecteurs (Bullinaria et Levy, 2007) : plus le nombre d'occurrences des mots est élevé, plus le modèle vectoriel est performant (Ferret, 2013 ; Weeds et Weir, 2005 ; van der Plas, 2008). *A contrario*, la similarité entre les mots cibles ayant un faible nombre d'occurrences est calculée à partir de très peu d'information dans les contextes. Ces mots cibles ont donc une plus grande tendance à être mal regroupés (Caraballo, 1999). Cependant, les mots avec un faible nombre d'occurrences ont un rôle essentiel sur la qualité des relations extraites, qu'ils soient en position de mot cible ou dans le contexte (Gorman et Curran, 2006), car ces mots rares peuvent correspondre à des contextes caractéristiques.

Ce premier problème a des conséquences sur le coût des traitements. Pour construire l'espace vectoriel, la méthode distributionnelle est fondée sur des éléments statistiques. Ainsi, si les données ne sont pas assez importantes, il n'est pas possible de disposer d'informations statistiques suffisamment fiables et significatives pour la construction du modèle distributionnel. De plus, la matrice de cooccurrence peut devenir extrêmement large quelle que soit la taille du corpus, et l'efficacité de l'algorithme en est alors affectée (Sahlgren, 2006). Le dilemme est donc le suivant : la plus grande quantité de données est nécessaire afin de construire un modèle suffisamment fiable, mais pour que les algorithmes puissent traiter les données à un coût raisonnable la plus petite quantité de données possible est préférable.

Pour répondre à ces problèmes et notamment pallier la dispersion des données, les solutions proposées sont de deux types : les premières visent à influer sur la définition des contextes, et les secondes interviennent au niveau de la construction ou de la réduction de la matrice des vecteurs de contexte. L'objectif est toujours de réduire l'espace, c'est-à-dire la mémoire occupée, et le temps de traitement.

Parmi les méthodes visant à influer sur les contextes, certaines s'intéressent plus particulièrement à la sélection des contextes utiles ou à l'intégration des informations sémantiques de manière à modifier la distribution des contextes. Ainsi, Broda *et al.* (2009) proposent de pondérer les contextes non pas en utilisant le nombre d'occurrences des contextes à l'état brut comme il est d'usage, mais en ordonnant les contextes en fonction de leur nombre d'occurrences. Le rang est ensuite utilisé pour pondérer puis sélectionner les contextes. D'autres approches s'appuient sur des modèles de langue pour déterminer les mots plausiblement intersubstituables, c'est-à-dire les substituts les plus probables pour représenter les contextes (Baskaya *et al.*, 2013). Ces modèles assignent des probabilités à des séquences arbitraires de mots en se fondant sur le nombre de cooccurrences dans un corpus d'entraînement (Yuret, 2012). Les mots substituts et leurs probabilités sont ensuite utilisés pour créer des paires de mots de manière à alimenter une matrice de cooccurrence, avant d'utiliser un algorithme de classification. Ces méthodes sont limitées car leur performance est proportionnelle à la taille du vocabulaire et elles nécessitent de disposer de données d'entraînement importantes. L'intégration d'informations sémantiques supplémentaires peut également être un moyen d'exercer une influence sur les contextes. En effet, Tsatsaronis et Panagiotopoulou (2009) ont démontré que la modification d'une méthode distributionnelle à l'aide de relations sémantiques calculées automatiquement ou provenant d'une ressource existante permet d'améliorer sa performance. Ainsi, avec un amorçage, Zhitomirsky-Geffet et Dagan (2009) modifient les poids des éléments au sein des contextes en s'appuyant sur les voisins sémantiques trouvés à l'aide d'une mesure de similarité distributionnelle. En s'appuyant sur ces travaux, Ferret (2013) s'intéresse au problème des mots ayant peu d'occurrences en corpus. Afin de mieux prendre en compte ces informations sémantiques, il propose d'utiliser un jeu d'exemples positifs et négatifs sélectionnés de manière non supervisée à partir d'un thésaurus distributionnel, et ainsi entraîner un classifieur supervisé. Ce classifieur est ensuite appliqué pour réordonnancer les voisins sémantiques. La méthode permet ainsi d'améliorer la qualité de la relation de similarité entre des noms ayant un nombre d'occurrences moyen ou faible.

Pour faire face aux problèmes liés à la très grande dimension des vecteurs, à la dispersion des données et au bruit statistique, une autre solution consiste à limiter le nombre de composants vectoriels avec un lissage de la matrice (Turney et Pantel, 2010). En effet, le calcul de la similarité entre toutes les paires de vecteurs est une tâche coûteuse alors que seuls les vecteurs qui partagent une dimension différente de zéro doivent être comparés[2]. La plupart des modèles connus utilisent des méthodes de réduction de dimension, généralement mises au point de manière à conserver les mots dont le nombre d'occurrences est faible. Une solution consiste à projeter les données aux dimensions élevées dans un espace ayant un nombre de dimensions plus réduit, tout en préservant approximativement les distances relatives entre les points, c'est-à-dire entre les mots cibles. La décomposition aux valeurs singulières

2. On considère que deux vecteurs ne partageant pas de dimension ne peuvent pas être similaires.

(SVD) (Deerwester *et al.*, 1990) est une méthode d'algèbre linéaire permettant la factorisation de matrice. Elle peut également être utilisée pour décomposer une matrice, afin d'obtenir une matrice finale ayant beaucoup moins de colonnes (généralement quelques centaines) mais plus dense (Turney et Pantel, 2010). Les méthodes fondées sur la SVD permettent ainsi de produire des vecteurs de contexte moins creux et moins affectés par le bruit statistique. À partir des données initiales, c'est-à-dire la matrice des contextes, cette technique divise la matrice en composants linéaires indépendants. La SVD est une méthode de factorisation de matrice coûteuse utilisée dans l'analyse sémantique latente (LSA) (Landauer et Dumais, 1997) pour réduire la matrice des contextes. La LSA est une méthode permettant de calculer des vecteurs sémantiques, ou vecteurs de contexte, à grande dimension, à partir des statistiques de cooccurrence des mots cibles.

Les méthodes décrites ci-dessus correspondent toutes à un processus d'optimisation généralement non supervisé. Face à ces méthodes, un nouvel ensemble de modèles vectoriels, fondé sur un apprentissage supervisé, a vu le jour ces dernières années et fait l'objet de nombreux travaux. Ainsi, à partir des travaux fondateurs de Bengio *et al.* (2003), des modèles prédictifs s'appuyant sur des réseaux de neurones ont été proposés (Mikolov *et al.*, 2013). Contrairement aux méthodes non supervisées qui commencent par construire les vecteurs de contexte et ensuite pondèrent ces vecteurs, les réseaux de neurones fixent directement les poids des vecteurs de manière à prédire les contextes dans lesquels les mots cibles correspondants ont tendance à apparaître. Le système apprend ainsi à assigner des vecteurs similaires à des mots cibles similaires (Baroni *et al.*, 2014).

Pour répondre au problème de la dispersion des données dans l'espace vectoriel, l'approche que nous proposons consiste à ajouter des informations sémantiques dans les contextes distributionnels, à l'instar de Tsatsaronis et Panagiotopoulou (2009) et Ferret (2013). Cependant, notre objectif diffère des travaux précédents : nous intégrons des relations sémantiques acquises automatiquement pour regrouper les contextes en faisant abstraction de la variation terminologique. Le nombre de contextes est ainsi réduit et, en revanche, la valeur associée à la cooccurrence mot cible/mot en contexte augmente. De plus, si les méthodes fondées sur la SVD réduisent la dispersion des données, leur fonctionnement n'est pas très explicite et elles s'appuient sur des méthodes mathématiques. Nous proposons au contraire de généraliser les contextes distributionnels en utilisant des connaissances linguistiques, des relations sémantiques acquises sur l'ensemble du corpus de travail. Le fonctionnement de notre méthode est explicitement décrit dans la section suivante.

3. Architecture globale de la méthode distributionnelle

Comme souligné précédemment, l'analyse distributionnelle appliquée à des corpus de spécialité ou des corpus de petite taille est limitée par une dispersion des données dans la matrice des contextes : cette matrice, représentant la distribution des mots ou des termes, contient beaucoup d'éléments ayant une valeur nulle. Pour tenter de ré-

soudre ce problème, nous proposons une approche consistant à densifier la matrice des contextes. Cette approche consiste à réaliser une abstraction des variations superficielles ou des contextes soit peu significatifs statistiquement soit liés au bruit de la méthode d'identification de ces distributions.

La méthode d'analyse distributionnelle que nous avons mise en œuvre suit le schéma présenté dans la figure 1. L'abstraction des contextes se trouve au cœur de la méthode. Elle exploite les contextes des mots cibles définis à l'étape 1 et précède le calcul de similarité sémantique (étape 3). L'abstraction des contextes, qui correspond, pour nous, à leur généralisation et à leur normalisation, est réalisée à l'aide de relations sémantiques acquises automatiquement. C'est une fois que la variation morphologique et sémantique est réduite dans les contextes que nous calculons la similarité entre les mots cibles.

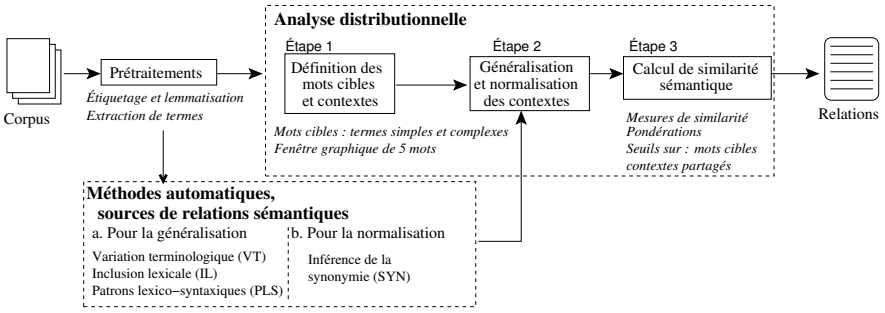

Figure 1. *Processus d'analyse distributionnelle*

3.1. *Définition des mots cibles et des contextes*

Dans le cadre d'applications en langue de spécialité, l'identification de relations sémantiques entre termes (simples et complexes) est primordiale. Les termes font référence aux notions du domaine et les relations sémantiques permettent plus d'appréhender le sens des termes. Ainsi, nous nous restreignons à l'analyse distributionnelle entre termes simples et complexes, qui constituent pour nous les mots cibles, et nous nous intéressons aux relations entre termes simples et complexes pris comme un seul ensemble (par exemple : *artère* et *souffle systolique*). Comme contextes distributionnels des mots cibles, nous avons choisi d'utiliser des fenêtres graphiques de $\pm$ 5 mots autour du mot cible. Il s'agit d'une taille reconnue comme adaptée aux textes de spécialité (Rapp, 2003 ; Généreux et Hamon, 2013). Elle permet aussi d'obtenir des contextes plus pertinents pour un mot cible, engendrant des résultats de meilleure qualité. Cependant, en limitant le nombre de contextes sélectionnés, cette taille de fenêtre accentue le problème de dispersion des données (Rapp, 2003). Les contextes sont composés de mots pleins qui cooccurent avec le mot cible au sein de la fenêtre

graphique. Nous considérons comme contexte les adjectifs, les noms, les verbes et les termes simples et complexes en écartant les mots vides (déterminants, conjonctions, adverbes, etc.).

3.2. *Calcul de similarité sémantique*

Lorsque les contextes ont été collectés, nous calculons la similarité entre deux mots cibles, en fonction de leurs contextes partagés. De nombreuses mesures de similarité et de pondération existent (Weeds *et al.*, 2004). Lors de précédentes expériences, nous avons constaté que l'indice de Jaccard obtient de meilleurs résultats avec les corpus de spécialité (Périnet et Hamon, 2014) :

$$S(w_m, w_n) = \frac{|\{\forall k, ctxt^k_{w_m}\} \cap \{\forall k, ctxt^k_{w_n}\}|}{|\{\forall k, ctxt^k_{w_m}\} \cup \{\forall k, ctxt^k_{w_n}\}|}$$

où $ctxt^k_{w_m}$ représente le contexte k du mot cible w_m, $\{\forall k, ctxt^k_{w_m}\}$ représente l'ensemble des contextes du mot cible w_m, et $|ctxt^k_{w_m}|$ correspond au nombre d'occurrences du contexte k pour le mot cible w_m.

L'indice de Jaccard compare le nombre de contextes communs à deux mots cibles à l'ensemble des contextes de ces mots (Tanimoto, 1958). Nous utilisons la généralisation pondérée de l'indice de Jaccard, telle que définie par Grefenstette (1994). Cette version généralise la similarité de Jaccard à la sémantique des valeurs non binaires, de manière à représenter chaque contexte par une valeur réelle entre 0 et 1. L'intersection est remplacée par le poids minimal et l'union par le poids maximal. Afin de valoriser les contextes les plus significatifs, nous avons pondéré les contextes par le nombre d'occurrences relatif du contexte c du mot cible w_m :

$$NbOccRel(w_m, c) = \frac{|ctxt^c_{w_m}|}{|\{\forall k, ctxt^k_{w_m}\}|}$$

Cette mesure de pondération des contextes est généralement associée à l'indice de Jaccard. Elle permet de prendre en compte l'importance d'un contexte d'un mot cible, par rapport au nombre total de contextes du mot cible.

Lors du calcul de similarité entre les mots cibles un très grand nombre de relations est généré. Garder toutes ces relations n'a pas de sens : un trop grand ensemble de relations est difficile à exploiter et à analyser *a posteriori*. Nous filtrons les relations avec la combinaison de trois paramètres : deux d'entre eux sont appliqués aux contextes (le nombre des contextes partagés et leur nombre d'occurrences) et le troisième est appliqué aux mots cibles (le nombre d'occurrences des mots cibles). Pour chaque paramètre, un seuil est calculé automatiquement en fonction du corpus. Les seuils que nous utilisons pour chaque corpus (voir section 5.1) sont répertoriés dans le tableau 1.

	Textes cliniques	Menelas
Nombre de contextes partagés	1	1
Nombre d'occurrences des contextes partagés	1	2
Nombre d'occurrences des mots cibles	3	3

Tableau 1. *Paramètres : valeurs des seuils sur les contextes et mots cibles*

4. Règles d'abstraction des contextes distributionnels

Une fois que les mots cibles et les contextes ont été définis, nous réalisons une abstraction des contextes. Cette abstraction est réalisée avec des ensembles de relations sémantiques acquises par des méthodes automatiques à partir du corpus de travail (voir section 5.2). Elle comprend une généralisation et une normalisation des contextes.

Nous partons du constat que les éléments superficiels différenciant les formes d'une même unité lexicale sont parfois effacés lors du processus de lemmatisation à l'aide d'une abstraction morphologique. Il est ainsi possible de regrouper sous une même unité lexicale ces différentes variations. Par exemple, la lemmatisation des verbes conjugués *opéré*, *opérons* et *opèrent*, permet d'effacer les marques de temps, de mode et de personne, et de regrouper ces trois formes sous le lemme *opérer*.

Une telle abstraction peut être également envisagée au niveau sémantique, où les traits « effacés » ne sont plus morphologiques mais sémantiques. Cette abstraction sémantique se traduit par exemple par le passage à un niveau supérieur dans une hiérarchie de concepts. Par exemple, les termes *chaise*, *fauteuil* et *tabouret* peuvent voir certains de leurs traits sémantiques effacés, de manière à être regroupés dans la classe sémantique des *sièges*. Le terme *tabouret* perd alors ses traits sémantiques *sans dossier* et *trois pieds*, le terme *fauteuil* perd ses traits *accoudoirs* et *confort*, et enfin la *chaise* perd ses traits *dossier* et *quatre pieds*.

Ainsi, nous émettons l'hypothèse que les contextes distributionnels peuvent être regroupés dans une même classe sémantique (par exemple, la classe des *sièges*). Cette classe serait représentée par un élément de cette classe, comme par exemple son hyperonyme (*siège*), de manière similaire au lemme *opérer* par rapport à l'ensemble des formes qu'il couvre. Le représentant *siège* serait alors utilisé comme substitut pour remplacer l'ensemble des mots appartenant à cette classe dans les contextes distributionnels. Après abstraction sémantique des contextes, la diversité des contextes est alors réduite : les contextes ne comptent alors plus qu'un seul lemme, *siège*, là où il y en avait quatre avant l'abstraction (*fauteuil*, *tabouret*, *chaise* et *siège*). Nous supposons que si ce substitut est utilisé pour remplacer les contextes, il devrait permettre de faire abstraction d'éléments superficiels, tout en gardant la même base sémantique et le même sens. L'objectif est d'une part, de diminuer la diversité des contextes distributionnels (on trouve alors dans les contextes uniquement *siège*, et non plus *chaise*, *tabouret* et *fauteuil*), et, d'autre part, d'augmenter le nombre d'occurrences

des contextes. Le contexte *siège*, s'il remplace ces trois termes, a alors une occurrence de 3.

4.1. *Généralisation des contextes*

La généralisation utilise des relations d'hyperonymie acquises par les méthodes définies à la section 5.2 : les patrons lexico-syntaxiques (PLS), l'inclusion lexicale (IL) et la variation terminologique (VT). Nous disposons alors, pour chaque mot w_i dans le contexte du mot w, de plusieurs ensembles de relations d'hyperonymie, $\mathbb{H}_s(w_i) = \{H_1, \ldots, H_n\} : \mathbb{H}_{PLS}$, $\mathbb{H}_{IL}$ et $\mathbb{H}_{VT}$, l'ensemble des hyperonymes pouvant être vide. Nous avons défini deux règles de substitution permettant de généraliser les contextes.

Ainsi, pour chaque mot w_i dans le contexte d'un mot w, nous appliquons l'une des règles suivantes :

1) si $|\mathbb{H}_S(w_i)| = 1$, alors $w_i := H_1$

Si un seul hyperonyme (H_1) acquis par une ou plusieurs méthodes S correspond au mot en contexte, le mot est remplacé par cet hyperonyme. Par exemple, si l'inclusion lexicale fournit la relation *restriction/restriction du débit coronaire*, alors *restriction du débit coronaire* est remplacée par *restriction*.

2) si $|\mathbb{H}_S(w_i)| > 1$, $w_i = \text{argmax}_{|H_i|}(\mathbb{H}_S(w_i))$

Si plusieurs hyperonymes acquis par une ou plusieurs méthodes S correspondent au mot en contexte, nous prenons en compte le nombre d'occurrences des hyperonymes $|H_1|, \ldots, |H_n|$ dans le corpus, et nous choisissons l'hyperonyme dont le nombre d'occurrences est le plus élevé dans le corpus. Par exemple, si pour le terme *artère coronaire droite* dans le contexte, les patrons lexico-syntaxiques fournissent les hyperonymes suivants : *artère coronaire*, *artère*, *vaisseau*, celui qui a le plus grand nombre d'occurrences est sélectionné et utilisé pour remplacer *artère coronaire droite* dans le contexte des mots cibles.

Quand plusieurs ensembles de relations d'hyperonymie sont disponibles, la phase de généralisation des contextes est réalisée en utilisant chaque méthode individuellement (par exemple, en généralisant avec les patrons lexico-syntaxiques) ou en combinant les méthodes. Les contextes sont alors généralisés en utilisant les ensembles de relations les uns à la suite des autres (par exemple, en généralisant avec les patrons puis avec l'inclusion lexicale) ou toutes ensemble (l'union des trois méthodes). Nous n'utilisons pas la propriété de transitivité de la relation d'hyperonymie.

4.2. *Normalisation des contextes*

Quant à la normalisation des contextes, elle utilise des relations de synonymie acquises à l'aide de la méthode définie en section 5.2. Nous avons défini une règle de normalisation qui vise à réduire les variations sémantiques. Les relations de synonymie sont tout d'abord regroupées sous la forme de groupes de synonymes et le

synonyme ayant le plus grand nombre d'occurrences est choisi comme représentant de ce groupe. Ainsi, à chaque mot w_i dans le contexte du mot cible w, correspond un groupe de synonymes $\mathbb{S}(R) = \{S_1, \ldots, S_n, R\}$ avec son représentant R.

Nous définissons une règle de normalisation des contextes, appliquée à chaque mot w_i dans le contexte d'un mot w pour substituer le mot du contexte par le représentant du groupe de synonymes auquel il appartient : $si\ \exists R | w_i \in \mathbb{S}(R), alors\ w_i\ :=\ R$ (l'ensemble de synonymes peut être vide). Si un mot dans le contexte appartient à un groupe de synonymes, il est remplacé par le représentant du groupe. Par exemple, si le terme *altération métabolique* dans le contexte d'un mot cible appartient au groupe de synonymes fourni par la méthode d'acquisition de synonymes (*anomalie métabolique*, *maladie métabolique*, *troubles métaboliques* et *altération métabolique*) celui qui a le plus grand nombre d'occurrences est sélectionné comme représentant et utilisé pour remplacer *altération métabolique* dans le contexte.

Pour la généralisation et la normalisation, si deux termes ont le nombre d'occurrences le plus élevé, le choix du terme représentatif n'ayant pas d'impact sur la méthode, nous sélectionnons le premier dans l'ordre alphabétique.

5. Matériel

Nous présentons dans cette section, les corpus de travail sur lesquels nous avons mené nos expériences, ainsi que les méthodes d'acquisition de relations sémantiques utilisées lors de la généralisation des contextes.

5.1. *Corpus*

Nous avons mené nos expériences sur deux corpus médicaux de taille et de langue différentes. Ces corpus contiennent des échanges entre spécialistes et se caractérisent par un degré de spécialisation élevé. Nous avons utilisé le corpus Menelas rédigé en français (Zweigenbaum, 1994). Le second corpus est rédigé en anglais et fourni par la compétition I2B2/2012 (Sun *et al.*, 2013). Dans les deux cas, les textes sont anonymisés.

Le corpus français Menelas comporte 84 839 mots. Il est constitué de deux grandes parties : un extrait d'un manuel de référence sur la coronarographie et les maladies coronariennes (environ 15 000 mots), et un ensemble de comptes rendus d'hospitalisation et de lettres de médecins hospitaliers aux médecins traitants concernant des malades atteints d'une maladie coronarienne (environ 70 000 mots). Les phrases de ce corpus sont longues avec 17,5 mots par phrase en moyenne. Pour ce corpus, le problème de dispersion des données est lié à la petite taille du vocabulaire. Si le manuel est bien rédigé, avec des phrases ayant une syntaxe *sujet - verbe - objet*, ce n'est pas toujours le cas des lettres des médecins, parfois produites à la hâte. Ainsi, le corpus comporte des abréviations, mais également un certain nombre d'erreurs. Les phrases

ne sont pas toujours bien construites, et peuvent, par exemple, ne pas contenir de verbe ou correspondre à une prise de notes.

Le corpus en anglais est composé de 311 documents cliniques provenant d'hôpitaux américains, fournis par Partners HealthCare et the Beth Israel Deaconess Medical Center. Il comporte 178 070 mots. Ce corpus comprend des phrases plus courtes que le corpus Menelas, avec 11 mots par phrase en moyenne, mais son vocabulaire est deux fois plus important, engendrant une plus grande diversité dans les contextes et accentuant le problème de dispersion des données.

Les corpus sont analysés à travers la plate-forme de TAL Ogmios[3] (Hamon et Nazarenko, 2008). La plate-forme a été configurée pour un étiquetage morphosyntaxique et une lemmatisation du corpus, à l'aide de TreeTagger (Schmid, 1994), et une extraction de termes analysés syntaxiquement a été réalisée à l'aide de YaTeA[4] (Aubin et Hamon, 2006). Les mots cibles et les contextes distributionnels sont définis à partir de ces prétraitements. Nous identifions ainsi les termes simples et complexes (dans les mots cibles et les contextes), et les mots dans les contextes (cf. section 3.1). Les termes extraits sont également utilisés pour l'acquisition des relations sémantiques.

5.2. *Acquisition de relations sémantiques*

La généralisation des contextes distributionnels s'appuie sur des relations sémantiques existantes, acquises sur l'ensemble du corpus de travail. Pour obtenir ces relations à partir de corpus, nous avons choisi d'utiliser plusieurs approches classiques d'acquisition de relations sémantiques entre termes : des patrons lexico-syntaxiques (PLS) dédiés à l'acquisition de relations d'hyperonymie, une méthode utilisant l'inclusion lexicale (IL), et des règles de variation terminologique (VT).

5.2.1. *Patrons lexico-syntaxiques*

Nous avons recours à des patrons définis pour l'acquisition de relations d'hyperonymie (*artère coronaire droite*/*artère* pour le français, et *disease*/*diabetes* en anglais). Pour le français, nous utilisons les patrons définis par Morin et Jacquemin (2004), comme par exemple *{quelques | plusieurs etc.} SN : LISTE*, où SN est un syntagme nominal et LISTE une liste de syntagmes. Pour l'anglais nous reprenons les patrons définis par Hearst (1992), pour acquérir des relations entre termes simples et complexes, par exemple, *SN {, SN}*{,} or other SN*, où SN est un syntagme nominal.

5.2.2. *Inclusion lexicale*

Cette approche s'appuie sur l'hypothèse, selon laquelle, si un terme en position tête (par exemple, *infarctus*) est inclus lexicalement dans un autre terme (par exemple, *infarctus du myocarde*), il existe généralement une relation d'hyperonymie entre ces

3. `http://search.cpan.org/~thhamon/Lingua-Ogmios/`
4. `http://search.cpan.org/~thhamon/Lingua-YaTeA/`

deux termes (Grabar et Zweigenbaum, 2003). Nous utilisons ici l'analyse syntaxique des termes fournie par YaTeA.

5.2.3. *Variation terminologique*

Nous utilisons la méthode d'acquisition de variantes terminologiques proposée par Jacquemin (2001) et implémentée dans Faster. Cette méthode exploite des règles de transformation morphosyntaxique décrivant la variation terminologique. Les variantes peuvent résulter de plusieurs opérations syntaxiques, morphologiques ou lexicales : principalement la permutation (*antibiotic course/courses of antibiotics*), la dérivation (*sténose de l'aorte/sténose aortique*) et l'insertion (*abdominal pain/abdominal muscle pain*). Par ailleurs, bien que Faster offre la possibilité d'obtenir des variantes sémantiques, nous avons choisi de ne pas les acquérir de cette manière. En effet, SynoTerm propose également des relations sémantiques dont une partie pourrait être acquise par Faster. De plus, les relations acquises par SynoTerm sont typées sémantiquement alors que Faster n'offre pas cette possibilité.

Dans le corpus en français (Menelas), les règles utilisées pour identifier des relations sémantiques entre termes sont essentiellement l'insertion[5]. En revanche, pour l'anglais, les trois règles sont utilisées. L'insertion d'un modifieur, par exemple *de revascularisation*, au sein du terme complexe *chirurgie coronarienne*, permet d'identifier une relation d'hyperonymie entre les deux termes concernés, *chirurgie coronarienne* et *chirurgie de revascularisation coronarienne*. Dans le cas de la dérivation et de la permutation, nous obtenons très peu de relations avec cette règle, et les relations obtenues sont plus apparentées à des relations de synonymie que d'hyperonymie.

Comme l'approche utilisée ne propose pas de relations typées sémantiquement et comme la plupart des relations sont identifiées grâce à la règle d'insertion, nous avons considéré les relations obtenues comme des relations d'hyperonymie. Les termes hyperonyme et hyponyme sont identifiés à partir du nombre de mots présents dans chaque terme : le terme le plus court correspond alors à l'hyperonyme (*lésion significative*), et le terme le plus long à l'hyponyme (*lésion coronaire significative*).

5.2.4. *Inférence de relations de synonymie*

Pour la normalisation des contextes, nous utilisons également une méthode à base de règles visant l'acquisition de relations sémantiques (Hamon et Nazarenko, 2001). Cette méthode permet d'inférer une relation de synonymie entre des termes complexes si au moins un de leurs composants (têtes) sont synonymes. Pour cela, nous utilisons deux dictionnaires existants. Pour le français, il s'agit du dictionnaire de langue générale *Le Robert*, qui contient des relations de synonymie entre mots. Pour l'anglais, nous avons utilisé les relations de synonymie entre mots proposées par WordNet (Fellbaum, 1998).

5. Nous ne disposions pas de ressources permettant d'identifier des variantes terminologiques par dérivation.

6. Expériences et résultats

Nous présentons dans cette section les expériences que nous avons menées, la méthode d'évaluation utilisée ainsi que les résultats obtenus.

6.1. *Expériences*

L'abstraction des contextes ayant pour objectif la réduction de la dispersion des données, nous avons évalué et caractérisé l'impact de la normalisation et de la généralisation des contextes sur la qualité des regroupements et des relations sémantiques obtenus. Pour cela, nous avons utilisé les règles proposées dans la section 4. Celles-ci s'appuient sur les relations sémantiques acquises automatiquement (cf. section 5.2) pour généraliser et normaliser les contextes séparément et de manière combinée. Afin de cerner la contribution de chaque méthode d'acquisition de relations sémantiques, ainsi que leur complémentarité, nous avons réalisé deux séries d'expériences autour de l'abstraction des contextes : une première série autour de la généralisation des contextes et une seconde pour la normalisation.

Tout d'abord, les règles de généralisation des contextes distributionnels w_i sont appliquées en utilisant séparément les ensembles $\mathbb{H}_{PLS}(w_i)$, relations d'hyperonymie acquises à l'aide des patrons lexico-syntaxiques (AD/PLS), $\mathbb{H}_{IL}(w_i)$, relations d'hyperonymie issues de l'inclusion lexicale (AD/IL), et $\mathbb{H}_{VT}(w_i)$, variantes terminologiques (AD/VT). Toutes les relations d'hyperonymie ont également été prises en compte dans leur ensemble indépendamment de la méthode utilisée pour les acquérir. On considère alors l'union des trois méthodes, c'est-à-dire l'ensemble $H(w_i) = \mathbb{H}_{PLS}(w_i) \cup \mathbb{H}_{IL}(w_i) \cup \mathbb{H}_{VT}(w_i)$ – AD/ALL3, pour appliquer les règles de généralisation sur le contexte w_i.

Nous n'avons réalisé qu'une seule expérience utilisant la normalisation des contextes w_i à l'aide des groupes de synonymes $\mathbb{S}(w_i)$ acquis automatiquement.

6.2. *Évaluation*

L'évaluation des relations acquises par analyse distributionnelle reste aujourd'hui une problématique importante et il est difficile d'évaluer une méthode distributionnelle en raison de la grande variété de relations qu'elle produit (Adam *et al.*, 2013 ; Morlane-Hondère, 2013). En effet, ces ressources contiennent un large spectre de relations lexicales, aussi bien des relations dites classiques que des relations moins bien spécifiées mais qui peuvent être pertinentes dans certaines applications (Morris et Hirst, 2004). Nous présentons tout d'abord les ressources puis les mesures que nous avons utilisées pour l'évaluation de notre approche.

6.2.1. *Références*

Nous faisons le choix d'évaluer notre méthode de manière intrinsèque, c'est-à-dire d'évaluer directement les relations sémantiques produites par la méthode. À l'instar de Curran (2004) et Ferret (2013), nous considérons ici les relations obtenues comme des ensembles de voisins associés à des mots cibles, les voisins étant ordonnés suivant la similarité avec le mot cible.

Aussi, nous comparons les relations sémantiques acquises à celles fournies par des ressources existantes. Nous utilisons les relations sémantiques issues de l'UMLS[6].

Les résultats obtenus sur le corpus Menelas sont évalués par rapport aux relations présentes dans la partie française de l'UMLS, c'est-à-dire 2 434 relations entre les termes du corpus[7]. Les types des relations contenues dans la référence sont majoritairement des co-hyponymes (1 536 relations), mais on dispose également de 333 hyperonymes, de 438 synonymes, et de 128 relations spécifiques du domaine (`expanded_form_of`). Les résultats obtenus sur le corpus de textes cliniques sont comparés aux 53 203 relations de la partie anglaise de l'UMLS restreintes aux termes du corpus. Les types des relations sont majoritairement co-hyponymes (22 680 relations) mais on dispose aussi de 22 939 relations du domaine (`has_sign_or_symptom`, etc.), de 6 505 hyperonymes et de 1 079 synonymes.

6.2.2. *Mesures d'évaluation*

Nous avons utilisé plusieurs métriques habituellement utilisées pour évaluer les résultats d'une analyse distributionnelle : la macro-précision (Sebastiani, 2002), la moyenne des précisions moyennes (MAP) (Buckley et Voorhees, 2005) et la R-précision. Nous utilisons le programme standard trec_eval[8], mis au point lors des campagnes TREC.

La macro-précision est la moyenne des précisions $p(w_i)$ obtenues pour chaque mot cible (w_i) et un ensemble de voisins sémantiques I_i^j, $I_i^{j(+)}$ étant un voisin pertinent pour le mot cible w_i, et n_i le nombre de ses voisins :

$$P = \frac{\sum_{k=1}^{|w_i|} \frac{\sum_{j=1}^{n_i} I_i^{j(+)}}{\sum_{j=1}^{n_i} I_i^j}}{|w_i|}$$

Nous avons considéré quatre sous-ensembles voisins permettant d'obtenir la macro-précision après examen de 1 ($n_i = 1$, P@1), 5 ($n_i = 5$, P@5), 10 ($n_i = 10$, P@10) et 100 voisins ($n_i = 100$, P@100) :

6. http ://www.nlm.nih.gov/research/umls/

7. La partie française de l'UMLS propose 1 735 419 relations mais nous restreignons l'ensemble de référence aux seules relations entre les termes du corpus.

8. `http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz`

$$P@N = \sum_{i=1}^{|w_i|} p(w_i|n_i = N)$$

La R-précision (Buckley et Voorhees, 2005) est une alternative à la précision limitée à un rang n. Elle consiste à utiliser comme seuil n_i le nombre de voisins corrects attendu pour un mot cible w_i, n_i variant alors suivant les mots cibles. La mesure est ainsi plus équitable qu'une précision à seuil fixe, car le seuil de précision varie en fonction du nombre de voisins attendus. Nous utilisons ensuite la moyenne des R-précisions par mot cible.

Nous évaluons également les résultats à l'aide de la *Mean Average Precision* (MAP). Celle-ci est obtenue en considérant la précision non interpolée $UAP(I_i^j)$ des voisins sémantiques I_i^j au rang j, n_i étant le nombre de voisins sémantiques I_i^j du mot cible w_i. La MAP est alors la moyenne de ces précisions non interpolées :

$$MAP = \frac{1}{|w_i|} \sum_{i=1}^{|w_i|} \frac{1}{n_i} \sum_{j=1}^{n_i} UAP(I_i^j)$$

La MAP reflète la qualité du classement : elle valorise le fait que la méthode ordonne tous les voisins sémantiques corrects proches de la tête de liste. Réciproquement, le fait d'ajouter des voisins sémantiques incorrects en fin de liste (après les voisins corrects) ne pénalise pas la méthode. Ainsi, contrairement à la R-précision qui permet d'évaluer également le classement des voisins, la MAP prend en compte tous les voisins, même ceux en fin de classement, alors que la R-précision se limite aux n voisins corrects attendus.

6.3. *Résultats*

Dans cette section, nous présentons et analysons les résultats obtenus sur nos deux corpus, par l'analyse distributionnelle après généralisation ou normalisation des contextes. Nous utilisons comme point de comparaison (*baseline*) les résultats obtenus avec l'analyse distributionnelle seule, c'est-à-dire sans abstraction des contextes. Compte tenu de l'absence de typage sémantique des relations par l'analyse distributionnelle et du nombre de relations proposées par cette approche, la qualité des résultats d'une analyse distributionnelle est généralement faible et difficile à apprécier lorsque ces résultats sont confrontés à des ressources existantes. Aussi, nous nous intéressons surtout à la différence entre les résultats obtenus à l'aide des mesures d'évaluation et à la qualité des regroupements à travers une analyse manuelle.

6.3.1. *Généralisation*

Nous avons analysé les regroupements obtenus après généralisation des contextes distributionnels en utilisant individuellement chaque ensemble de relations séman-

tiques acquises (AD/IL, AD/PLS, AD/VT), en combinant l'ensemble de ces relations (AD/ALL3).

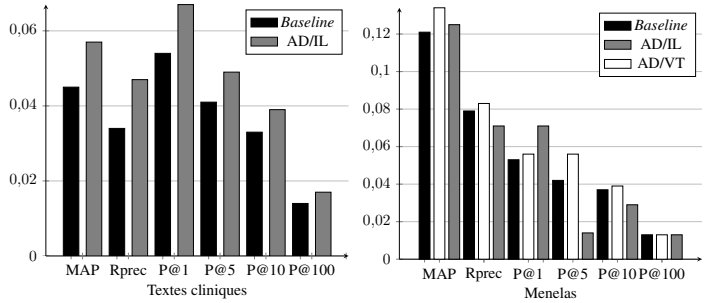

Figure 2. *Résultats obtenus pour la généralisation et comparaison avec la* baseline. *Généralisation avec l'inclusion lexicale (IL) pour le corpus de textes cliniques, et avec la variation terminologique (VT) et l'inclusion lexicale pour le corpus Menelas.*

Les résultats les plus intéressants obtenus après généralisation des contextes sont présentés à la figure 2, pour les deux corpus. Pour le corpus de textes cliniques, la généralisation avec l'inclusion lexicale (AD/IL) permet d'augmenter les résultats quelle que soit la mesure d'évaluation utilisée. L'impact de la généralisation des contextes est beaucoup plus faible lorsque les relations sont acquises à l'aide des patrons lexico-syntaxiques (AD/PLS) ou lorsqu'il s'agit de variantes terminologiques (AD/VT). Parmi les précisions à un rang donné, nous observons que les écarts les plus significatifs sont obtenus avec P@1, c'est-à-dire avec la précision prenant en compte uniquement le premier voisin (+ 0,013). Aussi, la MAP et la R-précision augmentent grâce à la généralisation des contextes avec respectivement + 0,012 et + 0,013. La généralisation des contextes semble ainsi contribuer à l'amélioration de l'ordonnancement des voisins. De plus, des termes pertinents, qui n'apparaissaient pas dans les 10 premiers éléments avec la *baseline* (analyse distributionnelle seule), remontent dans le classement des voisins grâce à la généralisation des contextes.

En ce qui concerne le corpus Menelas, seules les relations acquises à l'aide de la variation terminologique (AD/VT) ont un impact positif sur les résultats lors la généralisation. Les constats sont similaires aux précédents : la MAP et la R-précision augmentent par rapport à la *baseline*, respectivement + 0,013 et + 0,014. Les résultats de l'analyse distributionnelle en utilisant les deux autres ensembles de relations sémantiques (AD/IL et AD/PLS) sont plus mitigés : tandis que la R-précision décroît (– 0,008 et – 0,016), la MAP augmente légèrement (+ 0,004), la précision P@1 est la plus importante avec les relations d'inclusion lexicale (+ 0,018) et un peu plus faible avec les relations issues des patrons lexico-syntaxiques (+ 0,01), mais les précisions aux rangs supérieurs (P@5, P@10 et P@100) sont fortement dégradées (jusqu'à – 0,028 pour la P@5 de AD/IL) ou inchangées (P@100).

Lorsque l'on considère toutes les relations disponibles pour la généralisation des contextes, indépendamment de la méthode utilisée pour les acquérir (AD/ALL3), les résultats sont également améliorés par rapport à la *baseline*, en particulier, sur le corpus de textes cliniques (figure 3). Sur le corpus Menelas, les résultats sont similaires ou légèrement supérieurs à ceux obtenus lorsque les contextes sont généralisés avec l'inclusion lexicale.

Même si les résultats semblent varier selon le corpus, la généralisation des contextes fondée sur les relations d'inclusion lexicale améliore la qualité des résultats obtenus. Si l'on considère les mesures d'évaluation telles que la P@1, la MAP et la R-précision, la généralisation des contextes a un impact positif quelle que soit la méthode des relations employée pour acquérir ces relations. Les résultats bénéficient également de l'union des trois ensembles de relations par rapport aux résultats obtenus par la généralisation individuelle. Enfin, les observations sur le corpus Menelas indiquent que l'inclusion lexicale semble avoir une influence importante sur les relations obtenues lorsqu'elles sont comparées à une référence. Ce constat peut être dû au nombre de relations proposé par l'inclusion lexicale.

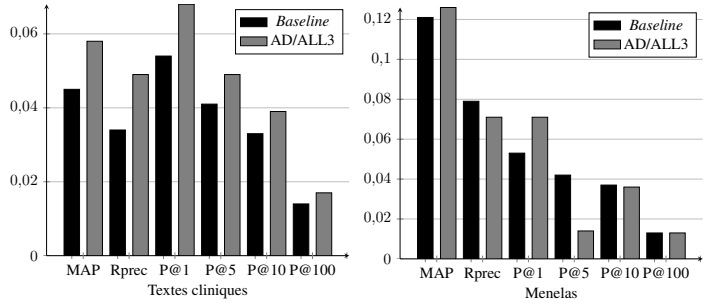

Figure 3. *Résultats obtenus sur les corpus de textes cliniques et Menelas, en généralisant les contextes à l'aide de toutes les relations à disposition (AD/ALL3) et comparaison avec la* baseline.

Étant donné la faible couverture de nos ressources et afin de mieux caractériser les voisins acquis et la qualité des relations sur les corpus, nous avons analysé les 10 premiers voisins de chaque mot cible retrouvés dans l'UMLS, pour les deux corpus.

Nous présentons dans le tableau 2, les 10 premiers voisins du mot cible *cough* (*toux*), obtenus avec la *baseline* et après généralisation des contextes avec l'inclusion lexicale (AD/IL), sur le corpus de textes cliniques. Les voisins soulignés sont ceux présents dans l'UMLS (*pain*, *fever*, *history*, *diarrhea* et *dyspnoea*). La généralisation avec l'inclusion lexicale (AD/IL) permet de mieux classer ces voisins, qui remontent alors dans le classement des 10 premiers voisins, à l'exception de *pain* déjà présent avec la *baseline*. Les 10 premiers voisins obtenus avec la généralisation sont plus pertinents et décrivent mieux le sens du mot cible *cough*. En effet, la *baseline* obtient un groupement sémantique autour de l'évanouissement et de la perte de connaissance

Mot cible : *cough*

	Baseline		**AD/IL**	
Rang	Voisin	Sim	Voisin	Sim
1.	nausea	0,00091	nausea	0,00108
2.	*pain*	0,00063	**fever**	0,00105
3.	*paroxysmal nocturnal dyspnoea*	0,00048	**vomiting**	0,00105
4.	*weakness*	0,00045	**chill**	0,00101
5.	*dizziness*	0,00044	**history**	0,0082
6.	*loss of consciousness*	0,00039	*pain*	0,00080
7.	*abd pain*	0,00036	**patient**	0,00080
8.	*numbness*	0,00036	**diarrhea**	0,00078
9.	*home*	0,00035	**dysuria**	0,00074
10.	*sweat*	0,00035	**dyspnoea**	0,00065

Tableau 2. *Corpus de textes cliniques : exemple de 10 premiers voisins obtenus pour le mot cible « cough », avec la* baseline*, et après généralisation avec l'inclusion lexicale (AD/IL). Les voisins en gras remontent dans le classement, ceux en italique descendent, et les termes soulignés sont ceux présents dans la référence.*

avec les termes *weakness*, *dizziness*, *loss of consciousness*, *numbness*. Ce groupement est sémantiquement homogène, mais sémantiquement moins proche de *cough* que les termes *fever*, *chill*, *dyspnoea*.

Nous avons réalisé une analyse similaire sur le corpus Menelas. Nous présentons ici quelques observations sur les regroupements obtenus lorsque les contextes sont généralisés avec les variantes terminologiques. Le tableau 3 illustre cette analyse avec les 10 premiers voisins du mot cible *cholestérol*. La généralisation des contextes a une influence sur le classement des voisins : 6 des 10 premiers voisins obtenus avec la *baseline* descendent dans le classement (en italique), c'est-à-dire que la généralisation augmente le score de similarité obtenu des voisins classés après les 10 premiers (en gras). Certains voisins restent inchangés du point de vue de leur similarité avec le mot cible même s'ils descendent dans le classement : *oblitération*, et *angiocoronarographie*. D'autres ont leur score de similarité qui est légèrement augmenté (*bilan lipidique*, *triglycéride*) et qui sont maintenus au même rang. Parmi ces 10 voisins, dans les deux cas de figure, aucun voisin n'est retrouvé dans l'UMLS sans que cela signifie qu'il ne s'agit pas de voisins pertinents. Ainsi, pour *cholestérol*, l'analyse distributionnelle après généralisation permet d'acquérir des relations du domaine, avec les voisins *bilan lipidique*, *bilan biologique*, *coronarographie*, *ventriculographie*, *angio-coronarographie*, mais également la relation d'hyperonymie entre *cholestérol* et *cholestérol total*. Globalement, les voisins acquis après généralisation sont un ensemble plus homogène sémantiquement, et correspondent au concept d'examen clinique.

Nous avons également observé que la combinaison AD/IL+PLS a une plus grande influence sur les mots cibles ; cette configuration réduit le nombre de mots cibles re-

trouvés dans la ressource et ceux-ci sont en partie différents des mots cibles obtenus avec la *baseline*. En revanche, en généralisant avec les variantes terminologiques, les mots cibles sont globalement les mêmes qu'avec la *baseline*, la différence se situe plus au niveau des voisins et de leur classement.

Mot cible : *cholestérol*

	Baseline		**AD/VT**	
Rang	Voisin	Sim	Voisin	Sim
1.	bilan lipidique	0,0028	bilan lipidique	0,0029
2.	triglycéride	0,0020	triglycéride	0,0021
3.	*lésion sévère*	0,0011	**cinétique ventriculaire gauche**	0,0012
4.	angio-coronarographie	0,0010	**bilan biologique**	0,0012
5.	oblitération	0,0010	**cholestérol total**	0,0012
6.	*extrasystole ventriculaire*	0,0009	**fonction ventriculaire gauche**	0,0012
7.	*examen clinique*	0,0009	**ventriculographie**	0,0011
8.	*coronaire droite*	0,0008	oblitération	0,0011
9.	*parenchyme pulmonaire*	0,0008	**coronarographie**	0,0010
10.	*pression pulmonaire*	0,0008	angio-coronarographie	0,0010

Tableau 3. *Corpus Menelas, fenêtre restreinte : exemple de 10 premiers voisins obtenus pour le mot cible* cholestérol*, avec la* baseline*, et après généralisation avec les variantes terminologiques (AD/VT). Les voisins en gras remontent dans le classement, ceux en italique descendent.*

6.3.2. *Normalisation*

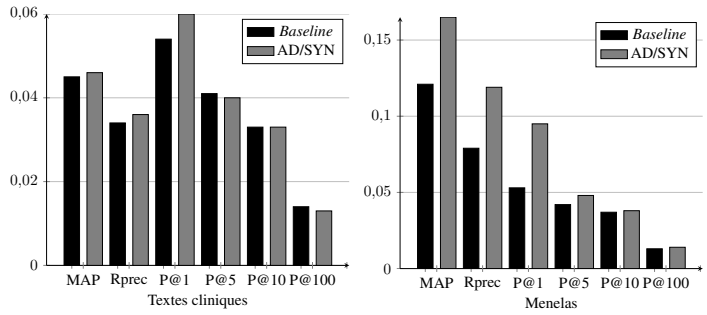

Figure 4. *Résultats obtenus pour la normalisation et comparaison avec la* baseline*. Normalisation avec les relations de synonymie pour les corpus de textes cliniques et Menelas.*

La normalisation est réalisée à l'aide des relations de synonymie acquises en corpus (SYN). Les résultats obtenus sont présentés dans la figure 4. Nous pouvons observer que l'impact de la normalisation est quasiment nul avec le corpus de textes cliniques mais plus important avec le corpus Menelas. Ainsi, pour les textes cliniques,

quelle que soit la mesure d'évaluation utilisée, les résultats sont quasiment identiques à ceux obtenus avec la *baseline*, avec un écart maximal de + 0,006 entre la *baseline* et la généralisation, obtenu en termes de P@1. En revanche, pour le corpus Menelas, l'impact de la normalisation sur la qualité de résultats est positif, quelle que soit la métrique utilisée. À l'instar de la généralisation, l'impact le plus fort de la normalisation est constaté avec la MAP, la R-précision et la P@1, pour lesquelles l'écart avec la *baseline* est respectivement de + 0,044, + 0,040 et + 0,042. Ainsi, la normalisation permet également d'améliorer le classement des termes. Et à l'inverse de la généralisation, son impact semble plus important quand le corpus est de petite taille.

Mot cible : *réseau coronarien*

	Baseline		**AD/SYN**	
Rang	Voisin	Sim	Voisin	Sim
1.	réseau circonflexe	0,00135	réseau circonflexe	0,00163
2.	*examen clinique*	0,00127	**coronaire droite**	0,00141
3.	artère coronaire	0,00121	artère coronaire	0,00140
4.	coronaire droite	0,00117	*examen clinique*	0,00140
5.	*maladie*	0,00091	**lésion**	0,00136
6.	*valve*	0,00088	**hypertension artérielle**	0,00118
7.	*index cardio-thoracique*	0,00088	**sténose**	0,00107
8.	*coeur*	0,00088	**athérome**	0,00104
9.	athérome	0,00085	**réseau coronaire**	0,00099
10.	*hypertrophie ventriculaire gauche pariétale*	0,00083	*maladie*	0,00099

Tableau 4. *Corpus Menelas : exemple de 10 premiers voisins obtenus pour le mot cible* réseau coronarien *, avec la* baseline*, et séparément après normalisation avec les synonymes (AD/SYN).*

De manière similaire à la généralisation, nous avons analysé manuellement les regroupements obtenus. Dans le tableau 4, nous présentons un exemple de regroupement avec les 10 premiers voisins du mot cible *réseau coronarien* obtenu avec la *baseline* et après normalisation (AD/SYN) des contextes. Nous pouvons observer que la normalisation permet de faire remonter dans le classement les termes *coronaire droite*, *lésion*, *hypertension artérielle*, *sténose*, *athérome* et *réseau coronaire*. Le résultat obtenu a une plus grande cohérence sémantique, avec un groupement sémantique autour des *pathologies* et *complications* qui peuvent être liées au réseau coronarien et un autre groupement autour du réseau lui-même. La normalisation permet ainsi de faire remonter la variante *réseau coronaire*.

Dans l'ensemble, nous avons constaté que l'ordre des 10 premiers voisins varie quand la normalisation est appliquée sur les contextes. En revanche, nous avons observé des similarités dans les regroupements obtenus après normalisation et après généralisation à l'aide les variantes terminologiques (AD/VT) : les voisins obtenus sont généralement les mêmes et les valeurs de similarité identiques. De même, l'impact sur le classement des voisins est moins important avec la normalisation qu'après généralisation des contextes.

6.4. *Bilan*

Les expériences présentées ci-dessus montrent que l'abstraction des contextes distributionnels permet d'obtenir des groupements sémantiques plus homogènes et cohérents. C'est essentiellement la pertinence des voisins sémantiques acquis qui est affectée par l'abstraction. L'impact de notre approche est généralement plus important lorsqu'il s'agit d'une généralisation des contextes, notamment à l'aide de l'inclusion lexicale, qu'avec une normalisation. L'importante contribution de l'inclusion lexicale peut être considérée comme un avantage car il s'agit d'informations syntaxiques qui peuvent être facilement fournies en grand nombre par un extracteur de termes. Et, contrairement aux relations acquises grâce à des patrons lexico-syntaxiques ou aux variantes terminologiques, ces relations sont stables formellement (la méthode d'acquisition reste la même et ne dépend pas de marques présentes dans les textes) et sémantiquement (à quelques exceptions près, leur interprétation est très fiable (Dupuch *et al.*, 2012)). De plus, l'analyse manuelle des relations révèle que l'abstraction des contextes distributionnels, et en particulier leur généralisation, permet d'obtenir des groupements sémantiques plus homogènes ainsi que des voisins sémantiques sémantiquement plus proches du mot cible qu'avec l'analyse distributionnelle seule. Les relations obtenues après abstraction des contextes sont majoritairement des co-hyponymes. L'abstraction permet également d'obtenir quelques relations du domaine propres au mot cible, telles que par exemple les relations *maladie*/*examen médical*, *examen médical*/*conséquence*. Cependant, notre méthode possède des limites, car même si elle permet d'identifier des regroupements sémantiques, les relations acquises ne sont pas typées, et notre évaluation manuelle des résultats reste partielle étant donné le très grand nombre de relations acquises.

7. Conclusion

Dans cet article, nous nous sommes intéressés à la réduction de la dispersion des données dans un espace vectoriel, et à la prise en compte des termes dans un modèle vectoriel, afin de pouvoir mettre en œuvre efficacement l'analyse distributionnelle sur des corpus de spécialité. Pour cela, nous avons proposé une méthode d'abstraction des contextes distributionnels s'appuyant sur des relations sémantiques acquises en corpus. Cette adaptation d'une méthode distributionnelle permet (i) de réduire le nombre de dimensions de l'espace vectoriel en diminuant la variation terminologique des contextes distributionnels, tout en conservant leur sémantique, et (ii) de faciliter le regroupement sémantique des termes simples et complexes en augmentant leur nombre de cooccurrences avec des contextes regroupés. Les relations sémantiques utilisées sont calculées en corpus grâce à trois méthodes d'acquisition de relations d'hyperonymie, et une méthode d'acquisition de relations de synonymie. L'abstraction des contextes distributionnels consiste alors à les généraliser grâce à ces relations d'hyperonymie et à les normaliser à l'aide des synonymes. Les résultats des expériences réalisées sur deux corpus du domaine médical en anglais et en français montrent que l'abstraction des contextes distributionnels améliore la qualité des résultats. D'une

part, les groupements sémantiques obtenus sont ainsi plus homogènes et cohérents, et, d'autre part, les termes complexes sont pris en compte comme des mots cibles. Dans l'ensemble, la généralisation, et en particulier les relations fournies par l'inclusion lexicale, ont un impact fort sur les regroupements obtenus. Quant à la normalisation, elle permet surtout d'améliorer le classement des voisins et la qualité des relations obtenues lorsqu'il s'agit d'un corpus de petite taille.

Ce travail ouvre plusieurs perspectives. Tout d'abord, les relations d'hyperonymie et de synonymie que nous avons utilisées ont été exploitées séparément. Or, ces relations acquises automatiquement pourraient être considérées comme une ébauche de taxonomie ou d'un réseau sémantique. L'utilisation du réseau de relations doit nous permettre de réaliser une abstraction des contextes plus précise sémantiquement notamment en prenant en compte cette taxonomie et les distances sémantiques entre les termes. Aussi, l'ensemble des relations acquises en corpus peut être bruité. En effet, les relations générées par les méthodes automatiques contiennent des erreurs ou des relations peu intéressantes en soi, qui peuvent être bénéfiques à l'abstraction des contextes, mais qui pourraient également dégrader les résultats. Pour pallier cette éventuelle dégradation des résultats, nous envisageons d'utiliser d'autres sources de relations comme les terminologies. Il nous sera alors possible d'évaluer l'impact de l'abstraction et des relations lorsque leur statut terminologique est maîtrisé, et ce, avec des relations jugées plus fiables.

8. Bibliographie

Adam C., Fabre C., Muller P., « Évaluer et améliorer une ressource distributionnelle », *Traitement Automatique des Langues*, vol. 54, n° 1, p. 71-97, 2013.

Aubin S., Hamon T., « Improving Term Extraction with Terminological Resources », *Advances in Natural Language Processing*, n° 4139 in *LNAI*, Springer, p. 380-387, 2006.

Baroni M., *Corpus linguistics : An international handbook*, vol. 2, Anke Lüdeling and Merja Kytö, Berlin, chapter Distributions in text, p. 803-821, 2009.

Baroni M., Dinu G., Kruszewski G., « Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 238-247, June, 2014.

Baskaya O., Sert E., Cirik V., Yuret D., « AI-KU : Using Substitute Vectors and Co-Occurrence Modeling For Word Sense Induction and Disambiguation », *Proceedings of SemEval - 2013*, Association for Computational Linguistics, Atlanta, Georgia, USA, p. 300-306, 2013.

Bengio Y., Ducharme R., Vincent P., Janvin C., « A Neural Probabilistic Language Model », *J. Mach. Learn. Res.*, vol. 3, p. 1137-1155, 2003.

Bodenreider O., Rindflesch T. C., Burgun A., « Unsupervised, Corpus-based Method for Extending a Biomedical Terminology », *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3*, BioMed '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 53-60, 2002.

Broda B., Piasecki M., Szpakowicz S., « Rank-Based Transformation in Measuring Semantic Relatedness. », *in* Y. Gao, N. Japkowicz (eds), *Canadian Conference on AI*, vol. 5549, Springer, p. 187-190, 2009.

Buckley C., Voorhees E., « Retrieval System Evaluation », *in* E. Voorhees, D. Harman (eds), *TREC : Experiment and Evaluation in Information Retrieval*, MIT Press, chapter 3, 2005.

Bullinaria J., Levy J., « Extracting semantic representations from word co-occurrence statistics : A computational study », *Behavior Research Methods*, vol. 39, n° 3, p. 510-526, 2007.

Caraballo S. A., « Automatic construction of a hypernym-labeled noun hierarchy from text », *ACL*, p. 120-126, 1999.

Chatterjee N., Mohan S., « Discovering Word Senses from Text Using Random Indexing », *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, Springer-Verlag, Berlin, Heidelberg, p. 299-310, 2008.

Cohen K. B., Demner-Fushman D., *Biomedical Natural Language Processing*, John Benjamins publishing company, 2013.

Curran J. R., From distributional to semantic similarity, PhD thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh, 2004.

Daille B., Morin E., « French-English Terminology Extraction from Comparable Corpora », *Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, p. 707-718, 2005.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391-407, 1990.

Déjean H., Gaussier E., Sadat F., « An approach based on multilingual thesauri and model combination for bilingual lexicon extraction », *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1-7, 2002.

Dupuch M., Dupuch L., Hamon T., Grabar N., « Semantic distance and terminology structuring methods for the detection of semantically close terms », *BioNLP : Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics, Montréal, Canada, p. 20-28, June, 2012.

Fellbaum C. (ed.), *WordNet*, MIT Press, Cambridge, 1998.

Ferret O., « Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel », *TALN 2013*, Les Sables d'Olonne, France, p. 48-61, 2013.

Firth J., *A synopsis of linguistic theory 1930-1955*, Oxford : Blackwell, p. 1-32, 1957.

Généreux M., Hamon T., « Experiments in synonymy : term extraction and mapping to concepts », *Terminologie et Intelligence artificielle (TIA)*, Paris, 2013.

Gorman J., Curran J. R., « Scaling distributional similarity to large corpora », *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 361-368, 2006.

Grabar N., Zweigenbaum P., « Lexically-Based Terminology Structuring », *Terminology*, vol. 10, p. 23-54, 2003.

Grefenstette G., « Corpus-Derived First, Second and Third-Order Word Affinities », *Sixth Euralex International Congress*, p. 279-290, 1994.

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

Hamon T., Nazarenko A., « Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience », *TAL*, vol. 49, nº 2, p. 127-154, 2008.

Harris Z., « Distributional structure », *Word*, vol. 10, nº 23, p. 146-162, 1954.

Hearst M. A., « Automatic acquisition of hyponyms from large text corpora », *International Conference on Computational Linguistics*, Nantes, France, p. 539-545, 1992.

Jacquemin C., *Spotting and discovering terms through natural language processing*, The MIT Press, 2001.

Landauer T., Dumais S., « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. », *Psychological Review ; Psychological Review*, vol. 104, nº 2, p. 211, 1997.

Lund K., Burgess C., « Producing high-dimensional semantic spaces from lexical co-occurrence », *Behavior Research Methods, Instrumentation, and Computers*, vol. 28, p. 203-208, 1996.

McCray A. T., Browne A. C., Bodenreider O., « The Lexical Properties of the Gene Ontology (GO) », *Proceedings of the AMIA 2002 Annual Symposium*, p. 504-508, 2002.

Mikolov T., Yih W., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, p. 746-751, June, 2013.

Morin E., Hazem A., « Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, United States, p. 1284-1293, June, 2014.

Morin E., Jacquemin C., « Automatic Acquisition and Expansion of Hypernym Links », *Computers and the Humanities*, vol. 38, nº 4, p. 363-396, 2004.

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

Morris J., Hirst G., « Non-classical lexical semantic relations », *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS 04, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 46-51, 2004.

Nastase V., Nakov P., Séaghdha D. O., Szpakowicz S., *Semantic Relations Between Nominals*, Morgan and Claypool Publishers, 2013.

Périnet A., Hamon T., « Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité », *Journées d'Analyse des Données Textuelles 2014*, Paris, France, p. 507-518, 2014.

Rapp R., « Word sense discovery based on sense descriptor dissimilarity », *MT Summit'2003*, p. 315-322, 2003.

Sahlgren M., The Word-Space Model : Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces, PhD thesis, Stockholm University, Stockholm, Sweden, 2006.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Sun W., Rumshisky A., Uzuner Ö., « Evaluating temporal relations in clinical text : 2012 i2b2 Challenge », *JAMIA*, vol. 20, n° 5, p. 806-813, 2013.

Tanimoto T., An element mathematical theory of classification, Technical report, I.B.M. Research, New York, NY, USA, 1958.

Tsatsaronis G., Panagiotopoulou V., « A generalized vector space model for text retrieval based on semantic relatedness », *EACL 2009*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 70-78, 2009.

Turney P. D., Pantel P., « From Frequency to Meaning : Vector Space Models of Semantics », *Journal of artificial intelligence research*, vol. 37, p. 141-188, 2010.

van der Plas L., Automatic lexico-semantic acquisition for question answering, Thèse de doctorat, University of Groningen, Groningen, 2008.

Weeds J., Weir D., « Co-occurrence Retrieval : A Flexible Framework for Lexical Distributional Similarity », *Computational Linguistics*, vol. 31, n° 4, p. 439-475, 2005.

Weeds J., Weir D., McCarthy D., « Characterising measures of lexical distributional similarity », *Proceedings of COLING'2004*, Stroudsburg, PA, USA, p. 1015-1022, 2004.

Yuret D., « FASTSUBS : An Efficient and Exact Procedure for Finding the Most Likely Lexical Substitutes Based on an N-Gram Language Model », *IEEE Signal Process. Lett.*, vol. 19, n° 11, p. 725-728, 2012.

Zhitomirsky-Geffet M., Dagan I., « Bootstrapping distributional feature vector quality », *Computational Linguistics*, vol. 35, n° 3, p. 435-461, 2009.

Zweigenbaum P., « Menelas : an access system for medical records using natural language », *Computer Methods and Programs in Biomedicine*, 1994.

Zweigenbaum P., Habert B., « Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. », *Revue Glottopol*, vol. 8, p. 22-44, 2006.