UNIVERSITÉ PARIS 13
LIMICS — UMR U1142
SORBONNE PARIS CITÉ

T H È S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Discipline : Informatique

présentée et soutenue publiquement par

Amandine PÉRINET

le 17 mars 2015

Analyse distributionnelle appliquée aux textes de spécialité : réduction de la dispersion des données par abstraction des contextes

Composition du jury

Mme Cécile FABRE	Professeur, CLLE-ERSS, Univ. Toulouse 2	Rapporteur
M. Emmanuel MORIN	Professeur, LINA-CNRS	Rapporteur
M. Thierry CHARNOIS	Professeur, LIPN, Univ. Paris 13	Président
M. Pierre ZWEIGENBAUM	Directeur de recherche, LIMSI-CNRS	Examinateur
M. Olivier FERRET	Chercheur, CEA LIST	Examinateur
Mme Sylvie DESPRÉS	Professeur, LIMICS-INSERM	Directrice
M. Thierry HAMON	MCF, LIMSI-CNRS et Univ. Paris 13	Encadrant

Remerciements

Je tiens en premier lieu à remercier Cécile Fabre et Emmanuel Morin qui m'ont fait l'honneur d'être rapporteurs de cette thèse.

Je remercie également les autres membres du jury, à commencer par Thierry Charnois qui a accepté d'en être le président. Merci à Olivier Ferret d'avoir pris le temps de relire cette thèse bien que lui-même occupé par la rédaction de son HDR ; ses nombreuses remarques m'ont été très utiles. Et enfin, j'exprime ma profonde reconnaissance à Pierre Zweigenbaum pour sa relecture minutieuse et détaillée qui m'a permis d'améliorer significativement le manuscrit.

Je remercie Sylvie Després d'avoir accepté de diriger cette thèse.

Je remercie chaleureusement Thierry Hamon de m'avoir proposé ce grand défi de thèse en informatique, de par mon cursus *traductique*. Je le remercie pour sa disponibilité (surtout pendant la dernière année), sa grande pédagogie et pour m'avoir transmis sa rigueur dans le travail.

Je remercie l'équipe de l'anciennement Lim&Bio - actuel LIMICS de m'avoir accueillie pendant la durée de ma thèse. Un clin d'œil particulier à mes collègues de bureau, Maïa, Mobin et Romain.

Je remercie Nicolas Grenèche pour son aide et sa disponibilité pour l'utilisation du serveur Magi de l'Université Paris 13.

Merci à Eric de la Clergerie de m'avoir accuieillie dans les locaux de l'INRIA. Mon expérience de thèse est également marquée par ces moments passés avec Corentin, Mikaël, Benjamin, Paul et les stagiaires de passsage. Je garderai en mémoire les nombreuses discussions avec François Barthélémy, dans la navette pour aller ou venir de l'INRIA. Merci ! Enfin, je remercie Claire Lemaire pour ses nombreux conseils.

Merci à l'équipe Pygmalion-FR avec qui j'ai travaillé en parallèle de la thèse pendant la dernière année. Merci à Bruno, Jana, Laurent, Karine, Alice, Romain, Julie, Bénédicte, Gauthier et Asceline.

J'ai été amenée à travailler avec Marie Dupuch et Natalia Grabar. Je les remercie particulièrement de m'avoir fait participer au travail de thèse de Marie. Cette expérience m'a été très utile notamment pour aborder ma thèse sous un angle plus informatique et mathématique.

Lors de mes participations à des conférences ou workshops, j'ai bénéficié de nombreux retours, remarques, questions et commentaires sur mes travaux. Je remercie tous ceux qui m'ont permis de faire avancer ce travail de recherche.

Je remercie tous ceux avec j'ai échangé pendant la thèse, de près ou de loin en rapport avec mon sujet. Merci François (Morlane-Hondère), Mounira, Ornella, et excusez-moi si j'en oublie !

Un très grand merci à ma famille, souvent bien trop loin.

Je terminerai ces remerciements par un merci incommensurable à celui qui m'a accompagnée, encouragée, conseillée, et qui m'a surtout beaucoup aidée quand mes heures de sommeil ne correspondaient qu'au cinquième des heures de travail. Merci Damien pour ta patience et ton soutien. Cette thèse te doit beaucoup.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Pour rechercher une information contenue dans un document, l'utilisateur d'un moteur de recherche soumet une question, ou simplement des mots-clés, au système. Par exemple, dans un domaine de spécialité tel que la médecine, un médecin peut avoir besoin d'utiliser un moteur de recherche pour extraire de sa base documentaire les dossiers des patients ayant présenté un souffle systolique. Lorsque le médecin entre comme requête le terme *souffle systolique*, le système doit alors être capable d'interpréter cette requête et de retrouver parmi les dossiers patients indexés, ceux pour lesquels le patient a présenté un souffle systolique et les renvoyer au médecin. Cependant, dans un domaine de spécialité, le vocabulaire de la requête formulée par l'utilisateur est souvent différent de celui contenu dans le document. Ceci peut entraîner une erreur ou une réponse incomplète à la requête. Par exemple, si les documents ne contiennent pas le terme *souffle systolique*, le système ne sera pas capable de retourner à l'utilisateur les dossiers dans lesquels sont mentionnés uniquement les diagnostics de la maladie, pourtant pertinents pour la requête. Une solution consiste alors à utiliser la paraphrase lexicale, c'est-à-dire associer aux termes de la requête, les termes sémantiquement ou morphologiquement proches, et ainsi étendre la requête par disjonction. La requête étendue serait ainsi, par exemple, *souffle systolique + valvulopathie + rétrécissement aortique + souffle cardiaque*, de manière à retrouver au moins un de ces termes. La requête a alors une plus large couverture lexicale, et le système est capable de capturer la variabilité du langage, améliorant ainsi la performance du moteur de recherche et la qualité de la recherche [Claveau et Sébillot, 2004].

1.1 Contexte

Les termes d'un domaine et les relations liant ces termes pouvant être utilisés pour étendre les requêtes sont généralement recensés dans une ressource terminologique. Cette ressource représente de manière plus ou moins couvrante les connaissances du domaine de spécialité en question. Les termes peuvent être de deux types : il peut s'agir de termes simples, c'est-à-dire des unités monolexicales (composées d'un seul mot, comme *artère*), ou complexes, c'est-à-dire des unités polylexicales (composées de plusieurs mots, comme *artère pulmonaire*). Les termes complexes se caractérisent par un nombre d'occurrences plus faible que les termes simples, de par le fait qu'en combinant des mots faiblement fréquents, leur fréquence est plus faible.
Quant aux relations sémantiques entre les termes, il en existe plusieurs types et plusieurs classifications ont été proposées, avec une granularité plus ou moins fine. Les relations

sémantiques classiques, habituellement contenues dans une ressource terminologique sont l'hyperonymie (*organe - cœur*), la co-hyponymie (*cœur - rein*), la synonymie (*cellule du sang - cellule sanguine*), l'antonymie (*aiguë - chronique*) [Nastase *et al.*, 2013]. Cependant, les relations contenues dans la ressource peuvent ne pas être suffisantes pour l'application, en ne couvrant pas le vocabulaire de la requête. C'est le cas lorsque le terme de la requête n'est pas présent dans la ressource, car la terminologie est inadaptée, incomplète ou non disponible pour des traitements automatiques, notamment dans le cas de néologismes ou de variantes terminologiques. Pour pallier ces insuffisances, la solution consiste généralement à constituer automatiquement un réseau lexical à partir des corpus spécialisés à disposition. Il existe deux types de méthodes automatiques, certaines ont pour but d'acquérir un type précis de relations (patrons lexico-syntaxiques, inclusion lexicale), d'autres visent le regroupement sémantique de termes ayant un sens proche (méthodes de clustering, analyse distributionnelle). Toutes les méthodes présentent des avantages et des inconvénients. Les patrons lexico-syntaxiques [Morin et Jacquemin, 2004, Hearst, 1992] obtiennent une bonne précision mais sont peu couvrants, et la mise au point des patrons peut être une tâche longue et coûteuse. Si l'inclusion lexicale [Grabar et Zweigenbaum, 2003] obtient une bonne précision et un bon rappel, elle est limitée du point de vue des relations acquises.

Ainsi, développer des méthodes automatiques pour l'extraction de relations est nécessaire pour pallier le faible recouvrement des ressources terminologiques avec le vocabulaire des requêtes, et pour éviter la construction manuelle des ressources, coûteuse en temps et en ressources. Les relations acquises automatiquement par ces méthodes permettent ensuite d'améliorer la qualité de l'application.
Ainsi, dans l'exemple précédent, les relations du domaine fournissant un lien entre la maladie et ses symptômes, telles que *souffle systolique - turbulence*, *souffle systolique - valvulopathie* ou *souffle systolique - rétrécissement aortique* peuvent permettre d'améliorer le rappel et de ramener plus de documents pertinents à la requête *souffle systolique* entrée par le médecin, en ramenant également les documents contenant *turbulence*, *valvulopathie* et *rétrécissement aortique*.

1.2 Problématique

Parmi les méthodes permettant l'acquisition de relations sémantiques, l'analyse distributionnelle vise à regrouper les mots supposés sémantiquement proches, mais sans permettre de typer le lien entre ces mots. Ainsi, les méthodes distributionnelles sont fondées sur le contexte d'apparition des mots (l'environnement textuel dans lequel le mot apparaît). Elles définissent la proximité sémantique de deux mots en fonction de la quantité de contextes que ces mots partagent. Plus le nombre de contextes communs est élevé, plus les deux mots sont sémantiquement proches [Harris, 1954, Firth, 1957].

Pour modéliser l'analyse distributionnelle, deux types de modèles existent [Sahlgren, 2006, Morlane-Hondère, 2013] : le modèle géométrique et le modèle probabiliste, même si la représentation reste la même. Dans cette thèse, nous abordons l'analyse distributionnelle à travers le modèle géométrique (ou modèle vectoriel), où les vecteurs représentent à la fois les informations contextuelles mais également des données statistiques distributionnelles [Sahlgren, 2006]. Chaque mot cible d'un texte (mots pour lesquels on cherche à identifier une relation) est représenté comme un point dans un espace mathématique en fonction de ses propriétés distributionnelles dans le texte (nombre et fréquences des contextes) [Turney et Pantel, 2010, Lund et Burgess, 1996]. La similarité sémantique entre deux mots est alors définie comme une proximité dans un espace à n-dimensions où chaque dimension correspond à des contextes partagés possibles. La figure 1.1 représente un espace vectoriel à trois dimensions. Dans cet espace, les trois dimensions correspondent à trois contextes, *auscultation*, *foyer aortique* et *retrouver*. Les mots cibles *rétrécissement aortique*, *souffle systolique*, *insuffisance mitrale* et *éjection*, sont ainsi représentés au sein de cet espace vectoriel, en fonction de leur nombre d'occurrence dans les trois contextes.

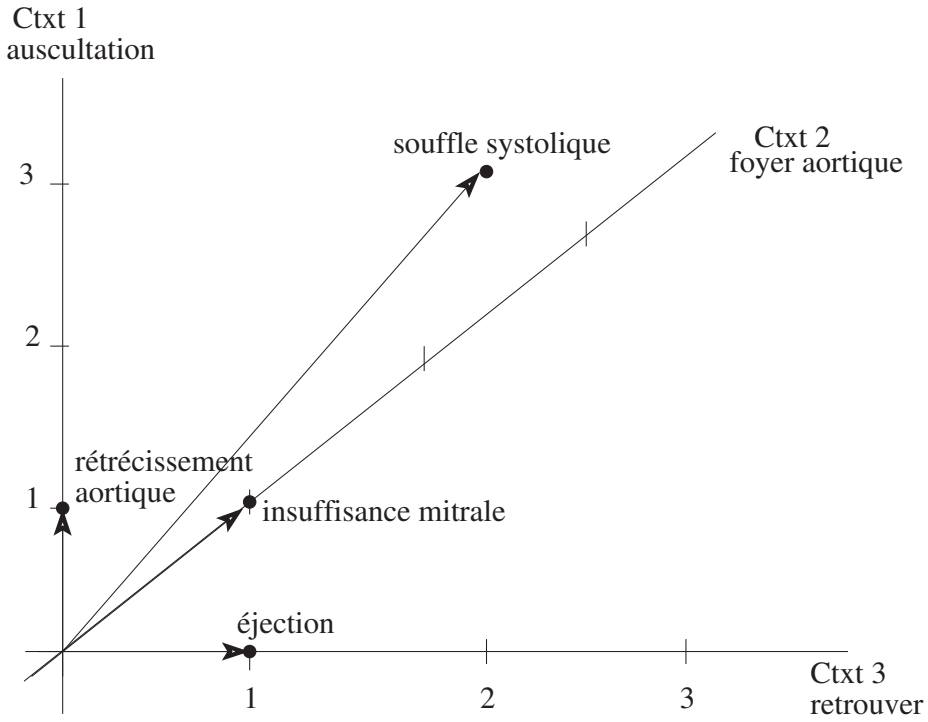

FIGURE 1.1 : Exemple d'espace vectoriel à trois dimensions, les trois dimensions étant les contextes *auscultation*, *foyer aortique* et *retrouver*.

Les modèles vectoriels ont ainsi l'avantage de permettre une quantification facile de la proximité sémantique entre deux mots en mesurant la distance entre deux vecteurs au sein de cet espace (par exemple, le cosinus de leur angle).

Ces modèles vectoriels s'appuient sur une matrice de contextes, qui a pour lignes les mots cibles du texte et pour colonnes les contextes. Cependant, cette matrice a pour inconvénient d'être généralement creuse ou éparse, c'est-à-dire que beaucoup de ses éléments sont à zéro car peu de contextes sont associés à un mot cible. Il s'agit d'un problème de dispersion des données qui est lié essentiellement aux faibles fréquences des mots en corpus. Cet inconvénient des méthodes distributionnelles existe aussi bien pour les corpus en langue générale [1] que pour les corpus en langue de spécialité [2]. Cependant, il est accentué avec les textes de spécialité, caractérisés par des tailles beaucoup plus petites, avec des fréquences de vocabulaire et un nombre de contextes différents plus faibles. De plus, comme nous venons de le voir, les textes de spécialité contiennent des termes simples et complexes. Dans un contexte d'utilisation de l'analyse distributionnelle sur des corpus de spécialité, la prise en compte des termes est essentielle puisque les termes sont porteurs du sens du texte. Cependant, en raison de leurs très faibles fréquences, les termes complexes se retrouvent généralement écartés du calcul de similarité. Récemment, des travaux de recherche sur l'analyse distributionnelle se sont intéressés à la compositionnalité distributionnelle, avec pour objectif la reconnaissance de similarité sémantique pour des unités lexicales allant au-delà du mot, tels que le syntagme, la phrase, le paragraphe, etc. Plusieurs ateliers sur cette thématique ont d'ailleurs été organisés : *Compositionality and Distributional Semantic Models* [3], *Vector Space Models and their compositionality* [4]. Les approches utilisées pour prendre en compte la compositionnalité, sont généralement fondées sur des opérations simples appliquées directement aux vecteurs de contextes, telles que l'addition ou la multiplication des vecteurs [Mitchell et Lapata, 2010]. Mais, à notre connaissance, aucune méthode distributionnelle n'intègre l'identification automatique des termes complexes.

1.3 Proposition

Nous nous intéressons à ce dernier point, le problème de la dispersion des données, en proposant l'adaptation d'une méthode d'analyse distributionnelle pour obtenir de meilleurs résultats sur des textes de spécialité, c'est-à-dire prenant en compte les termes simples et complexes dans le calcul de similarité. Pour cela, de manière à augmenter la fréquence des termes, nous avons cherché à réduire la diversité des contextes en

1. Langage courant, tel qu'utilisé dans les médias. Il s'oppose aux langues de spécialité.
2. Langue spécifique à un domaine (ex : médecine), caractérisée par une terminologie propre, par un vocabulaire technique.
3. http ://clic.cimec.unitn.it/roberto/ESSLLI10-dsm-workshop/
4. https ://sites.google.com/site/cvscworkshop2014/

réalisant une abstraction des contextes (cf. section 3.2). Cette abstraction des contextes est réalisée en les généralisant et en les normalisant à l'aide de relations sémantiques acquises automatiquement à partir de nos corpus de travail. Nous utilisons des relations calculées par trois méthodes visant l'acquisition de relations d'hyperonymie (l'inclusion lexicale, les patrons lexico-syntaxiques et la variation terminologique) et d'une méthode visant l'acquisition de relations de synonymie. Les variations dans les contextes sont ainsi effacées. Nous supposons que sans trop dégrader la sémantique, la fréquence des contextes distributionnels résultant de cette abstraction est augmentée, et la dispersion des données et la dimension de l'espace vectoriel sont réduites.

Prenons l'exemple de deux mots cibles, *coronarographie* et *examen clinique*, dans la figure 1.2. L'analyse distributionnelle s'appuie sur les contextes partagés par deux mots cibles pour déterminer la similarité entre ces mots cibles. Dans l'exemple, *coronarographie* et *examen clinique* ne partagent qu'un seul contexte, le verbe *révéler*, qui a une fréquence égale à 2.

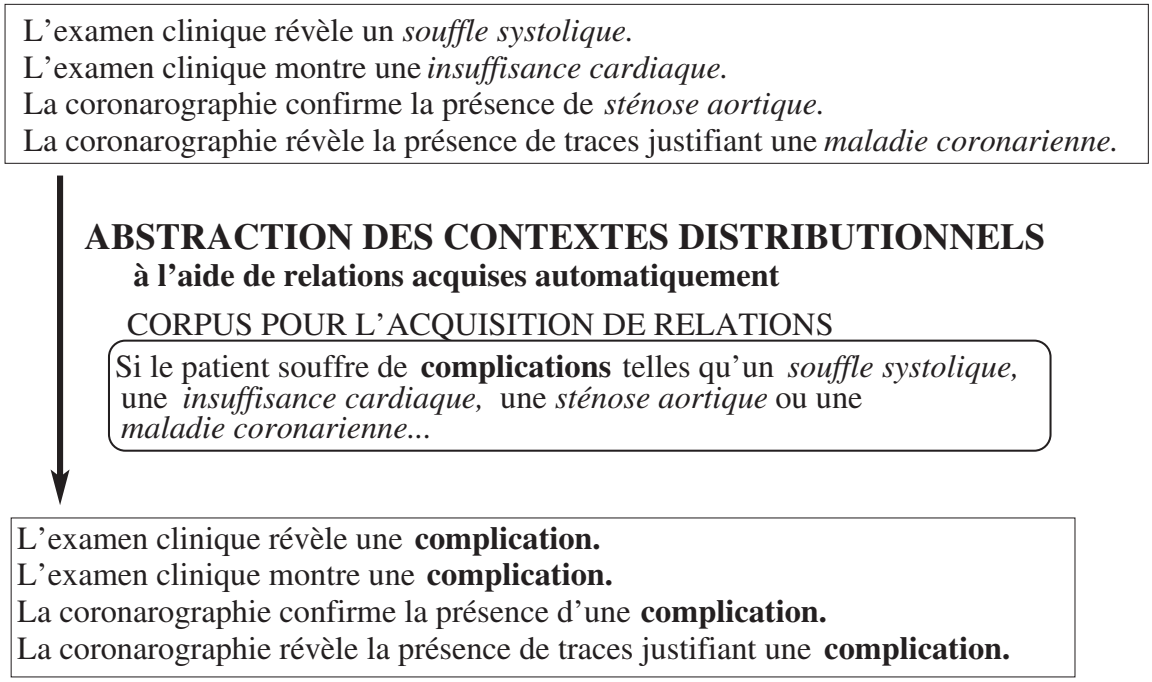

FIGURE 1.2 : Exemple de la méthode d'abstraction des contextes, pour les mots cibles *coronarographie* et *examen clinique*. L'accord du déterminant devant le contexte substitut (*complication*) est ici réalisé artificiellement, pour l'exemple.

Nous proposons de susbtituer les termes présents dans le contexte par leur hyperonyme ou synonyme. Nous exploitons les relations acquises automatiquement à partir de nos corpus de travail, dans l'exemple, les relations d'hyperonymie

– *souffle systolique - complication* ;
– *insuffisance cardiaque - complication* ;
– *sténose aortique - complication* ;
– *maladie coronarienne - complication* ;
pour remplacer les contextes. Ceux-ci sont ainsi remplacés par leur hyperonyme *complication*.

Après l'abstraction des contextes, les mots cibles *coronarographie* et *examen clinique* partagent alors deux contextes : le verbe *révéler* et le nom *complication*, et les fréquences de ces contextes sont plus élevées (une fréquence de 2 et de 4 respectivement).

Afin d'évaluer notre méthode et sa robustesse, nous avons mené des expériences sur quatre corpus de spécialité. Ces corpus diffèrent par leur taille (de l'ordre de 100 000 mots pour les petits corpus et de l'ordre du million de mots pour les plus volumineux), leur domaine de spécialité (médical et alimentaire), et la langue dans laquelle ils ont été rédigés (en français et en anglais). Nous avons réalisé un premier ensemble d'expériences afin de définir les valeurs des paramètres distributionnels adaptés à nos corpus de travail. Puis, nous avons réalisé un autre ensemble d'expériences pour évaluer la méthode d'abstraction des contextes distributionnels. Ceci a été effectué à travers la généralisation (à l'aide de relations d'hyperonymie, cf. section 3.2.1), la normalisation (à l'aide de relations de synonymie, cf. section 3.2.2) et la combinaison de la généralisation et de la normalisation des contextes distributionnels.

1.4 Présentation des chapitres

Outre cette introduction, ce document est organisé en cinq chapitres.
Le **Chapitre 2**, intitulé *Etat de l'art*, présente l'approche d'analyse distributionnelle et l'état de l'art autour des méthodes automatiques mises au point pour réduire la dispersion des données.
Dans le **Chapitre 3**, *Méthode d'abstraction des contextes distributionnels*, nous présentons la méthode d'abstraction proposée, c'est-à-dire les règles de généralisation et la règle de normalisation des contextes distributionnels. Nous décrivons également les méthodes d'acquisition de relations sémantiques qui nous fournissent les relations utiles pour l'abstraction des contextes.
Le **quatrième Chapitre**, *Corpus et évaluation*, présente les quatre corpus sur lesquels nous avons réalisé nos expériences, ainsi que la méthode utilisée pour évaluer l'analyse distributionnelle avec et sans abstraction des contextes. Nous travaillons sur deux corpus médicaux et deux corpus du domaine alimentaire.
Enfin, le **Chapitre 5**, *Expériences et résultats*, est composé de trois parties. Nous décrivons dans une première section les expériences menées autour des paramètres distributionnels et les résultats obtenus. Une fois la méthode distributionnelle définie, nous réalisons dans une deuxième section des expériences autour de l'abstraction des

contextes distributionnels. Enfin, nous comparons la meilleure configuration d'abstraction des contextes aux résultats obtenus avec une méthode par réseaux de neurones. Le document se termine par le **Chapitre 6**, la *Conclusion et les perspectives* de notre travail.

Chapitre 2

Etat de l'art

Les méthodes d'analyse distributionnelle (AD) reposent sur l'hypothèse selon laquelle les mots qui partagent des contextes similaires ont tendance à être sémantiquement proches [Harris, 1954, Firth, 1957]. Elles considèrent le problème des relations sémantiques comme un problème d'identification de similarité sémantique, et sont fondées sur des informations statistiques et des mesures de similarité pour regrouper les mots sémantiquement proches.

Exemple 1

a.	Il ne sera pas réalisé de traitement associant héparine, aspirine et bêta-bloquant.
b.	Le malade reçoit un traitement à l'héparine avant la coronarographie.
c.	Les bêta-bloquants doivent être arrêtés avant la coronarographie.
d.	Il convient de poursuivre le traitement en cours, à base de bêta-bloquants.

Dans l'exemple 1, l'analyse distributionnelle nécessite d'identifier dans le texte des mots cibles. Ces mots cibles sont ceux que l'on souhaite mettre en relation. Ainsi, les mots cibles de l'exemple sont *traitement*, *héparine*, *aspirine*, *bêta-bloquant*, *malade*, *coronarographie*, *cours* et *base*. Nous partons du principe que la méthode distributionnelle utilise un étiqueteur morphosyntaxique pour identifier les mots cibles. Ainsi, *cours* et *base* font partie des mots cibles de l'exemple, même s'ils sont peu pertinents sémantiquement car trop généraux. L'AD calcule ensuite une proximité entre les mots cibles à partir du nombre et de la fréquence des contextes partagés par les mots cibles, et détecte des relations entre ces mots. Pour cela, l'AD s'appuie sur les contextes dans lesquels apparaissent ces mots cibles, et plus particulièrement les contextes qu'ils partagent.
Dans l'exemple 1, si l'on ne considère pas les mots vides, c'est-à-dire les mots non porteurs de sens (déterminants, prépositions, etc.), les mots cibles *bêta-bloquant* et *traitement* partagent les contextes *héparine*, *associer*, *aspirine*, *cours* et *base* (cf. tableau 2.1).

A partir de ces contextes et de ces informations statistiques, les méthodes distributionnelles utilisent une mesure de similarité pour calculer la proximité sémantique entre les deux mots cibles, telle que par exemple, la fréquence des contextes partagés (mesure de similarité définie en section 2.1.3). Cette mesure utilise la fréquence des contextes partagés par les mots cibles pour définir s'ils sont liés sémantiquement. Dans l'exemple 1, les mots cibles *bêta-bloquant* et *traitement* partagent cinq contextes, ayant une fréquence de 2 pour *héparine* et une fréquence de 1 pour les autres (cf. tableau 2.1).

Contextes partagés	Fréquences
$héparine_{\{TERME\}}$	2
$associer_{\{VERBE\}}$	1
$aspirine_{\{TERME\}}$	1
$cours_{\{NOM\}}$	1
$base_{\{NOM\}}$	1

TABLEAU 2.1 : Contextes partagés des mots cibles *bêta-bloquant* et *traitement* et leur nombre d'occurrence, dans l'exemple 1.

Nous obtenons alors pour chaque mot cible une liste de mots en relation, appelés *voisins sémantiques*, supposés être sémantiquement liés au mot cible. Généralement, chaque voisin a un score qui lui est associé et qui représente la proximité entre le voisin et le mot cible.

Le terme *voisins sémantiques* est très générique et englobe l'ensemble des types de relations qui peuvent être acquis par analyse distributionnelle. En effet, l'AD permet de mettre en relation et de regrouper les mots du texte sémantiquement proches. Sans pour autant permettre de distinguer ces mots [Fabre et Bourigault, 2006], elle permet d'acquérir à la fois des relations dites *classiques*, comme la méronymie, la synonymie, l'antonymie, l'hyperonymie ou un lien morphologique [Grefenstette, 1994] et *non classiques* telles que les relations propres au domaine, la co-hyponymie, les relations nom-verbe, les collocations, etc. [Morris et Hirst, 2004].

La distinction entre ces deux types de relations renvoie à la distinction entre les deux notions de *similarité sémantique* et de *proximité contextuelle*, faite notamment par [Budanitsky et Hirst, 2006] et [Zesch et Gurevych, 2010]. Cependant, la limite entre ces deux notions est parfois floue dans les travaux existants, car les deux termes sont souvent employés de manière interchangeable. De plus, la *similarité sémantique* est fréquemment considérée comme incluse dans la *proximité contextuelle*, et les deux problèmes sont souvent abordés avec les mêmes méthodes [Ferret, 2013a]. Dans cette thèse nous parlerons donc de *similarité sémantique* dans son sens général, c'est-à-dire incluant les deux types de relations, sans distinction.

Automatisée depuis les années 1990, l'analyse distributionnelle a d'abord été implémentée dans des systèmes visant la constitution de thesaurus à partir de textes spécialisés. L'AD est ainsi appliquée au domaine médical, notamment pour l'anglais [Grefenstette, 1992] et le français [Bouaud *et al.*, 2000, Nazarenko *et al.*, 1997]. Par ailleurs, l'analyse distributionnelle est souvent utilisée dans une perspective de traduction automatique, pour aider les systèmes de traduction. Actuellement, les tendances sont à l'utilisation de corpus les plus volumineux possibles, avec un nombre de mots dépassant souvent

le milliard. Les petits corpus traités avec des méthodes distributionnelles sont généralement de l'ordre de la centaine de millions de mots ou du milliard de mots. Or, en langue de spécialité, les corpus sont généralement de plus petite taille, plutôt de l'ordre du million de mots, même si néanmoins il existe le corpus des articles de revues médicales PubMed Central Open Access (de plusieurs milliards de mots).

Les concepts et procédures utilisés dans les calculs distributionnels ont été définis [Sahlgren, 2006, Turney et Pantel, 2010, Baroni et Lenci, 2010]. Cependant, ce domaine de recherche présente encore des problématiques autour de la définition de méthodes distributionnelles (calcul de similarité, sélection et pondération des contextes), de l'adaptation au volume des données, de l'évaluation de ces méthodes et de l'exploitation des ressources qu'elles permettent de générer. Plusieurs worshops consacrés à ce sujet ont récemment eu lieu, notamment lors de la conférence ACL 2013 [1], pendant les conférences françaises TALN 2013 [2] et TALN 2014 [3], ou pendant EACL 2014 [4].

Un élément essentiel à toute méthode distributionnelle est le choix des paramètres qu'elle met en œuvre. Ces paramètres sont les mots cibles utilisés, le type de contexte (par exemple, fenêtre graphique ou analyse syntaxique) et les caractéristiques du contexte (par ex., taille, forme et type de fenêtre), et les mesures de similarité et de pondération. Le choix des paramètres à utiliser est essentiel, puisqu'il influence directement le type de relations acquises et leur qualité. Il découle également du type de corpus de travail ; les paramètres ne sont pas les mêmes si l'on travaille avec des corpus en langue générale ou en langue de spécialité, et si les corpus sont de grande taille ou de petite taille [Bullinaria et Levy, 2012]. La mesure de similarité et le type de contexte sont concernés par cette variation. Généralement, les travaux sur corpus de spécialité utilisent l'indice de Jaccard et les dépendances syntaxiques, alors que ceux menés sur de la langue générale préfèrent le cosinus et une fenêtre graphique (de simples co-occurrences) [Bernier-Colborne, 2014]. Aussi, dans le cadre de l'utilisation de l'analyse distributionnelle avec des textes de spécialité, il est nécessaire de prendre en compte la reconnaissance des termes dans la méthode automatique.

Or, à notre connaissance, cela n'a pas beaucoup été réalisé. Sauf erreur, en monolingue, [Bannour *et al.*, 2011] sont les seuls à prendre en compte les termes extraits automatiquement dans un calcul distributionnel.

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). Comme en monolingue, les observations des co-occurrences dans les corpus de spécialité comparables sont moins (voire peu) fiables que les corpus en langue générale, de par leur habituelle petite taille (autour

1. https://sites.google.com/site/cvscworkshop/
2. http://www.taln2013.org/ateliers/appel-atelier-semantique-distributionnelle/
3. http ://www.irit.fr/semdis2014/fr/
4. https://sites.google.com/site/cvscworkshop2014/

d'un million de mot, contre plus des 100 millions de mots pour la langue générale). Pour faire face à ce problème, [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 la fréquence d'occurrence de chaque contexte de manière à rendre ces fréquences plus fiables. Notre problématique est proche de ces travaux, qui s'appuient également sur les travaux fondateurs de [Grefenstette, 1994].

Ainsi, nous commencerons par présenter les paramètres mis en jeu dans une méthode distributionnelle, puis nous nous intéresserons aux modèles vectoriels implémentant cette méthode. Nous décrirons ensuite la réduction de dimensions et nous terminerons par les limites des méthodes distributionnelles.

2.1 Paramètres distributionnels

Les méthodes distributionnelles sont influencées au moins autant par la nature des données que par la nature de la théorie mathématique sous-jacente. Le choix d'un modèle approprié est réalisé en fonction de la nature des données disponibles et de l'objectif de l'application [Weeds *et al.*, 2004]. Ainsi, pour définir un modèle adapté à une application donnée, les paramètres des méthodes distributionnelles doivent être définis de manière à avoir le meilleur impact sur les relations acquises, en termes de quantité, de qualité mais également au niveau des types sémantiques des relations obtenues [van der Plas et Tiedemann, 2010, Curran, 2004, Peirsman et Geeraerts, 2009, Sahlgren, 2006]. La construction d'une méthode distributionnelle adaptée à une tâche définie comprend un très grand nombre de paramètres à adapter [Kiela et Clark, 2014], notamment le choix du contexte et les mesures de similarité.

2.1.1 Définition et sélection des contextes

La première étape de toute méthode distributionnelle consiste à définir le contexte utilisé. En linguistique, le *contexte* d'un mot correspond à l'environnement textuel dans lequel ce mot apparaît [Sahlgren, 2006]. Mais la notion de contexte peut se définir de nombreuses manières et varie d'une approche à l'autre. Le contexte peut être une fenêtre graphique (mot dans le voisinage d'un mot cible), des dépendances syntaxiques [Lin, 1998b, Curran, 2004] (nom argument d'un verbe) ou l'alignement de corpus parallèles, ou des contextes plus riches qui s'appuient sur des liens de dépendances et des préférences de sélection au niveau des positions des arguments [Erk et Padó, 2008]. Le choix du type de contexte contraint la sélection de l'approche d'extraction des contextes. Ces approches diffèrent par leur sophistication linguistique (de la simple fenêtre graphique à la prise en compte de dépendances syntaxiques), leur temps de calcul, leur fiabilité et l'information utilisée (en termes de quantité et qualité) [Curran, 2004]. Pour de plus amples informations sur les types de contextes, nous renvoyons à la thèse de [Sahlgren,

2006] et aux travaux de [Kilgarriff et Yallop, 2000]. Nous décrivons dans ce chapitre les deux types de contexte principalement utilisés en analyse distributionnelle : les fenêtres graphiques, les dépendances syntaxiques. Pour appuyer nos explications autour des types de contextes, nous utilisons la première phrase de l'exemple qui a été présentée en début de chapitre (cf. exemple 1).

2.1.1.1 Fenêtre graphique

La fenêtre graphique recense et compte les co-occurrences, en prenant en compte un nombre n de mots situés dans la proximité du mot cible [Wilks *et al.*, 1990, Schütze, 1998, Lund et Burgess, 1996].

Le contenu des éléments pris en compte dans la fenêtre est variable. La fenêtre peut prendre en compte des informations linguistiques, telles que le lemme ou les catégories morphosyntaxiques. Selon les travaux, l'apport de ces informations dans un modèle vectoriel est variable. Ainsi, [Karlgren et Sahlgren, 2001] augmentent la performance de leur modèle de Random Indexing en utilisant des données lemmatisées. La lemmatisation permet de réduire la dispersion des données, en réduisant le nombre de formes différentes présentes dans le contexte des mots cibles. Les catégories morphosyntaxiques ont tendance à réduire la performance du modèle quelle que soit la taille de la fenêtre graphique, à l'exception de la plus petite fenêtre possible, de taille ± 1 mot centrée sur le mot cible [Sahlgren, 2006]. [Sahlgren, 2006] cite différents travaux dans lesquels l'ajout d'informations linguistiques dans les contextes diminue la performance d'un modèle vectoriel. Ainsi, l'intégration d'une analyse morphosyntaxique pour la prise en compte les catégories morphosyntaxiques dans le modèle augmenterait le nombre de mots uniques dans les données, et ainsi amplifierait le problème de dispersion des données, dégradant ainsi la qualité des résultats. Cependant, si l'on souhaite restreindre le choix des mots dans les contextes à certaines catégories grammaticales, et ainsi écarter les mots sémantiquement vides, la prise en compte des catégories morphosyntaxiques est nécessaire.

Pour illustrer notre propos, considérons par exemple une fenêtre de 5 mots, soit deux mots de chaque côté du mot cible, de laquelle nous excluons les mots vides (déterminants, propositions, etc.). Nous appliquons cette fenêtre à la première phrase de l'exemple (exemple 1), pour les mots cibles *bêta-bloquant* et *traitement*. Les contextes obtenus pour ces mots cibles sont présentés dans le tableau 2.2. Le mot cible *bêta-bloquant* est caractérisé par deux termes [5] dans son contexte, et *traitement* par un terme et quatre verbes.

Les variations autour des types de fenêtres généralement utilisés concernent principalement la taille de la fenêtre. Celle-ci a un impact sur le type, la quantité et la

5. Par *terme*, nous entendons *candidat terme*. Il s'agit des termes extraits automatiquement par YaTeA lors du pré-traitement des corpus (cf. section 4.1.3).

Mots cibles	Contextes
$traitement_{\{TERME\}}$	$être_{\{VERBE\}}$
	$réaliser_{\{VERBE\}}$
	$associer_{\{VERBE\}}$
	$\boldsymbol{héparine}_{\{TERME\}}$
$bêta\text{-}bloquant_{\{TERME\}}$	$\boldsymbol{héparine}_{\{TERME\}}$
	$aspirine_{\{TERME\}}$

TABLEAU 2.2 : Contextes (co-occurrents) des mots cibles *bêta-bloquant* et *traitement* dans la première phrase de l'exemple, avec une fenêtre de 5 mots. Le contexte partagé est indiqué en gras.

qualité des relations acquises. Le choix de la taille correspond généralement à un compromis entre la spécificité et la dispersion des données [Rapp, 2003]. Une fenêtre restreinte génèrera des contextes plus proches des contextes syntaxiques qu'une fenêtre large [Morlane-Hondère, 2013]. La fenêtre de taille restreinte (généralement ±5 mots, centrée sur le mot cible) permet ainsi d'obtenir des contextes plus pertinents pour un mot cible, engendrant des résultats de meilleure qualité pour la similarité sémantique, i.e. pour des relations sémantiques classiques (synonymie, antonymie, hyperonymie, méronymie, etc.). Cependant, en limitant le nombre de contextes sélectionnés, cette taille de fenêtre accentue le problème de dispersion des données. [Rapp, 2003] démontre avec un score de 92.5% de bonnes réponses au test du TOEFL, en utilisant une fenêtre de cinq mots (et sans mots vides), que plus la fenêtre rétrécit, plus la performance augmente. En revanche, les contextes plus larges, tels qu'une phrase, un paragraphe ou un document, sont plus appropriés pour modéliser une similarité plus générale, une proximité contextuelle [Sahlgren, 2006, Peirsman *et al.*, 2008]. Cependant, la fenêtre large peut rapidement devenir difficile à manipuler, car le nombre de dimensions (correspondant au nombre de contextes) est alors très élevé [Curran, 2004].

Pour les textes de spécialité, même si les travaux en analyse distributionnelle sont peu nombreux, des tendances se dessinent. Une fenêtre graphique de cinq mots autour du mot cible est la taille définie comme idéale par plusieurs auteurs [Rapp, 2003, Généreux et Hamon, 2013]. Cette taille semble également convenir en langue générale [Curran, 2004], pouvant s'étendre à trois mots autour du mot cible [Kris Heylen et Speelman, 2008].

Des définitions plus fines des contextes sont possibles avec les fenêtres graphiques. Il est possible de prendre en compte la direction (gauche - droite). Les fréquences des contextes sont calculées en prenant en compte la position du mot dans le contexte par rapport au mot cible : les mots apparaissant à gauche et à droite du mot cible

sont comptabilisés séparément [Clark, 2014, Ferret, 2013a]. Il est également possible de prendre en compte la symétrie de la fenêtre, c'est-à-dire l'étendue de la fenêtre de manière équivalente à droite et à gauche du mot cible [Clark, 2014, Curran, 2004].

Etant donné que l'évaluation des méthodes distributionnelles reste difficile, en raison de la grande variété de relations qu'elles produisent (voir section 4.2), il n'est pas évident de savoir si la prise en compte de ces informations améliore réellement la qualité des relations extraites [Clark, 2014]. Le choix du type de fenêtre graphique et du type d'informations prises en compte dans la fenêtre dépend du type de relations que l'on souhaite extraire.

2.1.1.2 Dépendances syntaxiques

Un autre type de contexte consiste à prendre en compte dans le contexte uniquement les mots liés syntaxiquement au mot cible, c'est-à-dire à considérer les dépendances syntaxiques directes [Lin, 1998a, Lin et Pantel, 2001, Padó et Lapata, 2007, Curran, 2004]. Ce contexte est de fait plus restreint qu'une fenêtre graphique, car seuls les mots qui entrent dans une relation de dépendance syntaxique avec le mot cible font partie du contexte. Par exemple, seuls les verbes gouvernant un nom cible en fonction sujet, ou les adjectifs modifiant le nom cible sont comptabilisés dans le contexte. C'est ce qui est utilisé dans les travaux de [Habert *et al.*, 1998] portant sur les textes de spécialité. La construction des vecteurs de contextes reste semblable à celle des fenêtres graphiques : la fréquence d'un contexte est restreinte au nombre de fois où le mot dans le contexte apparaît en relation de dépendance avec le mot cible.

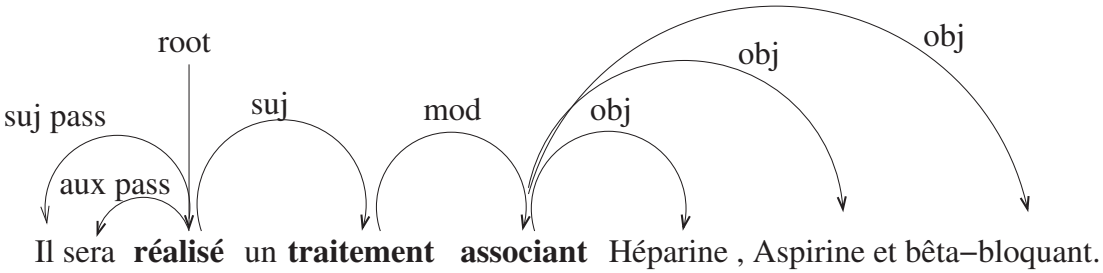

FIGURE 2.1 : Analyse syntaxique de la première phrase de l'exemple 1

Les relations de dépendances sont construites sous la forme de triplets $<$ *gouverneur, relation, dépendant* $>$ (qui correspondent pour nous au triplet $<$ *mot cible, relation syntaxique, contexte* $>$). La figure 2.1 présente l'analyse syntaxique de la première phrase des exemples. Avec cette analyse, le mot cible *traitement* a deux contextes et *bêta-bloquant* a un contexte, mais ils n'ont pas de contexte commun, car les fenêtres syntaxiques sont plus restrictives.

Mots cibles	Relations de dépendance
traitement	< réaliser, suj, traitement > < traitement, mod, associant >
bêta-bloquant	< associant, obj, bêta-bloquant >

TABLEAU 2.3 : Contextes syntaxiques des mots cibles *bêta-bloquant* et *traitement* dans la première phrase de l'exemple 1.

2.1.1.3 Positionnement

Le choix du type de contexte influence la représentation distributionnelle d'un mot cible. Celle-ci peut avoir la forme non structurée d'une fenêtre graphique ou à l'inverse la forme plus structurée d'une relation syntaxique <*mot cible, relation syntaxique, contexte*> [Ferret, 2013a]. Cette différence dans le type de contexte a pour conséquence une différence dans le type d'informations sémantiques capturées. En effet, les représentations les plus élémentaires des contextes ont l'avantage d'être moins coûteuses en temps et en ressources, plus verbeuses, et moins complexes à mettre en œuvre, mais l'information extraite est moins riche. En revanche, une analyse syntaxique en dépendances va fournir un contexte plus précis, avec des chemins de dépendance [Padó et Lapata, 2007] et extraire des informations contextuelles plus riches et plus rares (moins de contextes sont identifiés que lorsqu'une fenêtre graphique est utilisée) au détriment d'une complexité de traitement plus élevée.

Le choix du contexte est également lié à la dispersion des données : les fenêtres graphiques offrent des représentations des contextes plus informatives et bruitées, alors que les relations de dépendance ont des représentations plus dispersées et moins bruitées [Erk et Padó, 2010].

La quantité et la qualité d'informations présentes dans le contexte sont deux éléments indissociables desquelles dépend la précision d'une méthode distributionnelle [Curran et Moens, 2002a]. Pour obtenir une bonne précision, la méthode distributionnelle doit disposer d'un grand nombre de contextes. Mais la quantité seule ne suffit pas. Ainsi, si deux mots cibles partagent un grand nombre de contextes très généraux, ils n'ont donc que peu de contextes discriminants et il est fort probable que cette relation ne soit pas pertinente [Adam *et al.*, 2013b]. A l'inverse, un jeu de contextes très limité, voire la présence d'un seul contexte pour un mot cible risque d'écarter une relation qui peut s'avérer pertinente. Enfin, plus le contexte est structuré, plus le type de relations obtenues est précis. Ainsi, les approches basées sur la syntaxe permettent de générer un thesaurus plus précis que les approches basées sur une fenêtre graphique [van der Plas, 2008, Weeds et Weir, 2005], étant donné qu'elles restreignent les mots distributionnellement similaires à ceux qui sont plausiblement

inter-substituables [Church *et al.*, 1994]. Ces contextes plus précis sont considérés comme de meilleurs indicateurs du sens du mot cible [Clark, 2014].

L'utilisation d'une combinaison de plusieurs types de contextes est démontrée dans la littérature comme apportant de meilleurs résultats [Henestroza Anguiano et Denis, 2011]. Cependant, étant donné que nous ne disposons pas d'analyseur syntaxique adapté, et nous positionnant dans la perspective des textes de spécialité, nous faisons le choix d'utiliser une fenêtre graphique. Ceci à la fois pour le faible coût de traitement des fenêtres graphiques, pour la non-dépendance à la langue traitée, mais surtout pour limiter la dispersion des données. Pour limiter cette dispersion, nous optons également pour une lemmatisation des contextes ainsi que la prise en compte de catégories morphosyntaxiques acquises automatiquement.

2.1.2 Force d'association des contextes

Une fois défini le contexte, les mots dans le contexte peuvent être pondérés de manière à valoriser les contextes les plus significatifs du sens des mots cibles. Des mesures d'association permettent alors de donner plus ou moins de poids aux informations contenues dans le contexte. Il s'agit essentiellement de valoriser les événements redondants, peut-être rares, par rapport à ceux qui sont attendus, et ainsi mesurer l'importance d'un mot dans la définition du contexte d'un mot cible. L'hypothèse est que les événements rares, s'ils sont partagés par deux mots cibles, sont plus représentatifs de la similarité entre deux mots cibles que les événement plus fréquents. Ainsi, parmi les contextes d'un mot, certains sont de meilleurs descripteurs que d'autres [Turney et Pantel, 2010].

Nous présentons deux mesures d'association (cf. tableau 2.4) : la fréquence relative (1) et l'information mutuelle (2). Nous considérons ici les mesures les plus communément utilisées sur des corpus de spécialité. La fréquence relative est généralement utilisée pour pondérer les contextes avant le calcul de l'indice de Jaccard, et l'information mutuelle en amont du calcul avec le cosinus.

Pour définir ces mesures, tout comme pour les mesures de similarité présentées dans la section 2.1.3, nous utilisons la notation en astérisque proposée par [Lin, 1998a]. La méthode distributionnelle calcule une relation entre un mot cible w et et un mot w' dans le contexte de ce mot cible. L'ensemble des contextes d'un mot cible se traduit ainsi par un triplet $(w, (r, w'))$, que l'on peut aplatir pour obtenir (w, r, w'). Dans cette notation, l'astérisque indique un ensemble de valeurs s'étendant sur toutes les valeurs existantes pour ce composant du tuple de la relation. Dans cette notation, tout est défini en termes de l'existance d'instance des contextes, c'est-à-dire des relations de contextes avec une fréquence différente de zéro.

Ainsi, $(w, *, *)$ est un triplet qui représente l'ensemble des contextes pour un mot cible w, et $n(w, *, *)$ (ou $|(w, *, *)|$) correspond au nombre total de contextes pour le mot

(1) Fréquence Relative (FreqRel)	$Fr(w, w') = \frac{\|(w,r,w')\|}{\|(w,*,*)\|}$
(2) Information mutuelle ponctuelle (PMI)	$I(w, w') = log(\frac{p(w,r,w')}{p(w,*,*)p(*,r,w')})$

TABLEAU 2.4 : Mesures d'association.

cible. Enfin, $f(w, *, *)$ représente la fréquence des mots apparaissant dans le contexte du mot cible.

La fréquence relative (1) est la mesure la plus simple à mettre en œuvre :

$$Fr(w, w') = \frac{\|(w,r,w')\|}{\|(w,*,*)\|} \quad (1)$$

La fréquence relative permet de prendre en compte l'importance d'un contexte d'un mot cible, par rapport au nombre total de contextes du mot cible.

L'information mutuelle (IM) (2) est la fonction de pondération la plus utilisée [Bannour *et al.*, 2011].

$$I(w, w') = log(\frac{p(w,r,w')}{p(w,*,*)p(*,r,w')}) \quad (2)$$

Il s'agit d'une information mutuelle qui compare la probabilité d'observer deux événements aléatoires w et w' ensemble (distribution jointe) aux probabilités de les observer indépendemment (distribution indépendante) [Church et Hanks, 1990].

2.1.3 Mesure de la proximité distributionnelle

Après avoir extrait et pondéré les contextes d'apparition de chaque mot cible, les contextes, ou plutôt les vecteurs de contextes (cf. section 2.2), sont comparés afin de calculer un score de similarité sémantique entre chaque couple de mots cibles, et ainsi quantifier dans quelle proportion deux mots cibles sont proches [Pedersen *et al.*, 2004]. De nombreuses mesures de similarité ont été proposées à la fois en recherche d'informations [Jones et Furnas, 1987] pour mesurer la similarité d'un vecteur associé à une requête et d'un vecteur associé à un document, et en sémantique lexicale [Lin, 1998b, Dagan *et al.*, 1999, Lee, 1999, Weeds *et al.*, 2004]. Comme pour la définition des contextes distributionnels, il est difficile de se prononcer sur une seule et meilleure mesure. Le choix de la mesure de similarité dépend de l'application, du type de relations visé, de la dispersion des données, de la distribution de la fréquence des

(3) Nombre de contextes partagés	$S(w_m, w_n) = \lvert(w_m, r, w') \cap (w_n, r, w')\rvert$
(4) Fréquence des contextes partagés	$S(w_m, w_n) = \sum_{w'} min(\lvert(w_m, r, w')\rvert, \lvert(w_n, r, w')\rvert)$
(5) Indice de Jaccard	$S(w_m, w_n) = \frac{\lvert(w_m,*,*)\cap(w_n,*,*)\rvert}{\lvert(w_m,*,*)\cup(w_n,*,*)\rvert)}$
(6) cosinus	$S(w_m, w_n) = \frac{\sum wgt(w_m,*_r,*_{w'})\times wgt(w_n,*_r,*_{w'})}{\sqrt{\sum wgt(w_m,*,*)^2 \times \sum wgt(w_n,*,*)^2}}$

TABLEAU 2.5 : Mesures de similarité.

éléments comparés, et des autres paramètres distributionnels utilisés [Clark, 2014]. Il existe de nombreuses mesures de similarité [Panchenko et Morozova, 2012]. Plusieurs travaux se sont intéressés à une présentation et une évaluation étendue des différentes mesures de similarité utilisées en langue générale [Weeds *et al.*, 2004]. [Curran et Moens, 2002b] réalisent une expérience d'évaluation à large échelle, dans laquelle ils étudient la performance de plusieurs méthodes communément utilisées. [van der Plas et Bouma, 2004] présentent une expérience similaire pour le danois, dans laquelle ils testent la plupart des mesures les plus performantes d'après [Curran et Moens, 2002b]. Dans ces expériences, l'indice de Jaccard et le coefficient de Dice sont les deux mesures qui obtiennent les meilleures performances. Bien que ces deux mesures soient parmi les plus employées et les plus performantes [Manning *et al.*, 2008], le cosinus semble être la mesure la plus utilisée, car elle considère l'angle des vecteurs et non pas leur longueur [Turney et Pantel, 2010, Clark, 2014]. Le cosinus évite ainsi la comparaison de vecteurs trop longs, qui rendent le traitement trop coûteux.

Notre intérêt n'est pas de décrire l'ensemble des mesures de similarité existantes mais d'adapter l'analyse distributionnelle aux corpus de spécialité, en considérant les mesures les plus communément utilisées sur les corpus de spécialité. Nous présentons donc ici uniquement quatre mesures de similarité, les mesures généralement utilisées sur des corpus de spécialité, recensées dans le tableau 2.5, et nous renvoyons vers les travaux cités ci-dessus pour plus de détails sur les autres mesures.

Ces mesures calculent une similarité S entre deux mots cibles w_m et w_n. Chaque mot cible est représenté par un ensemble de triplets (w, r, w'). Nous avons choisi d'utiliser trois mesures utilisées avec des textes de spécialité ; le nombre de contextes partagés, la fréquence des contextes partagés et l'indice de Jaccard, ainsi qu'une mesure qui nous semble être la plus utilisée en langue générale, le cosinus.

Le nombre de contextes partagés (3) calcule le nombre de contextes en commun pour deux mots cibles : deux mots sont considérés similaires s'ils partagent suffisament de types de contextes différents, délimités par un seuil choisi. Il s'agit de la mesure reconnue comme la plus adaptée pour l'analyse distributionnelle aux textes de spécialité :

$$S(w_m, w_n) = |(w_m, r, w') \cap (w_n, r, w')| \tag{3}$$

La fréquence des contextes partagés (4) utilise le nombre d'occurrences de contextes partagés ainsi que le nombre de types de contextes pour déterminer la similarité entre deux mots cibles :

$$S(w_m, w_n) = \sum_{w'} min(|(w_m, r, w')|, |(w_n, r, w')|) \tag{4}$$

L'indice de Jaccard (5) compare le nombre de contextes communs à deux mots cibles à l'ensemble des contextes de ces mots [Tanimoto, 1958]. Nous utilisons dans notre travail 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 minimum et l'union par le poids maximum :

$$S(w_m, w_n) = \frac{|(w_m,*,*) \cap (w_n,*,*)|}{|(w_m,*,*) \cup (w_n,*,*)|)} \tag{5}$$

Le cosinus (6) permet de calculer la similarité entre deux vecteurs à n dimensions en déterminant le cosinus de l'angle entre eux. Ainsi, cette mesure considère l'angle des vecteurs dans un espace indépendemment de la distance ou de la longueur des vecteurs. La valeur du cosinus de l'angle, soit la similarité entre deux mots cibles $S(w_m, w_n)$, est comprise dans l'intervalle [0,1]. La valeur 0 indique des vecteurs indépendants et 1 des vecteurs similaires. Les valeurs intermédiaires permettent d'évaluer le degré de similarité.

$$S(w_m, w_n) = \frac{\sum wgt(w_m,*_r,*_{w'}) \times wgt(w_n,*_r,*_{w'})}{\sqrt{\sum wgt(w_m,*,*)^2 \times \sum wgt(w_n,*,*)^2}} \tag{6}$$

2.1.4 Bilan

Dans cette section, nous avons présenté les différents paramètres distributionnels nécessaires à la mise en œuvre d'une méthode distributionnelle. Ainsi, pour identifier des relations entre les mots cibles au sein d'un corpus, la première étape consiste à choisir un type de contexte (graphique, ou syntaxique), et les éléments à prendre en compte dans ce contexte (restriction à certaines catégories morphosyntaxiques, lemmatisation,

etc.). Le choix des mesures de pondération et de similarité joue également un rôle essentiel et peut influencer la qualité des groupements sémantiques obtenus.

2.2 Modèles vectoriels ou d'espaces sémantiques

L'hypothèse de [Harris, 1954, Firth, 1957] visant à regrouper des mots sémantiquement proches est mise en œuvre grâce à des méthodes *géométriques*, qui consistent généralement à représenter les données par des vecteurs, des matrices ou des tenseurs d'ordre supérieur [Turney et Pantel, 2010]. Le lien entre la méthodologie distributionnelle et la métaphore géométrique du sens des mots est le concept du *vecteur de contexte*, qui permet le passage des statistiques distributionnelles à une représentation vectorielle du sens des mots [Sahlgren, 2006]. Ce qui importe dans le modèle vectoriel n'est pas seulement la représentation du sens, mais surtout comment l'espace est construit par le vecteur de contexte [Sahlgren, 2006].

Dans cette section, nous présentons cette représentation géométrique, ses limites, et les solutions qui ont été proposées à ce jour.

2.2.1 Représentation géométrique du sens des mots

Une manière de calculer la similarité distributionnelle consiste à utiliser des modèles vectoriels (VSM). Le VSM est une représentation spatiale du sens des mots. A partir de l'hypothèse distributionnelle selon laquelle le sens des mots dépend du contexte dans lequel ils apparaissent, chaque dimension de l'espace vectoriel correspond à un contexte. L'idée centrale est de représenter la similarité sémantique comme une proximité dans un espace à n-dimensions, où n peut être un très grand nombre (à titre d'exemple, [Sahlgren, 2006] considère les VSM allant jusqu'à plusieurs millions de dimensions). Afin d'avoir une idée d'une représentation spatiale, un exemple simplifié d'un espace à trois dimensions est présenté dans la figure 2.2, issu de l'exemple 1 donné en début de ce chapître.

Dans ces représentations géométriques, chaque mot cible d'un texte est représenté comme un point dans un espace mathématique en fonction de ses propriétés distributionnelles dans le texte [Turney et Pantel, 2010, Lund et Burgess, 1996]. A la figure 2.2, les mots cibles *malade*, *traitement*, *héparine* et *aspirine* sont représentés au sein d'un espace vectoriel, en fonction de leur fréquence d'occurrence dans les trois contextes *recevoir*, *associer* et *coronarographie*. Ainsi, les dimensions utilisées pour définir l'espace vectoriel d'un corpus sont les mots présents dans le contexte du mot cible. Dans l'exemple, le modèle vectoriel a trois dimensions, correspondant aux trois contextes. Une distance géométrique entre les mots cibles est alors utilisée comme un indice de leur similarité sémantique. Ainsi, les mots ayant des sens proches ou similaires sont

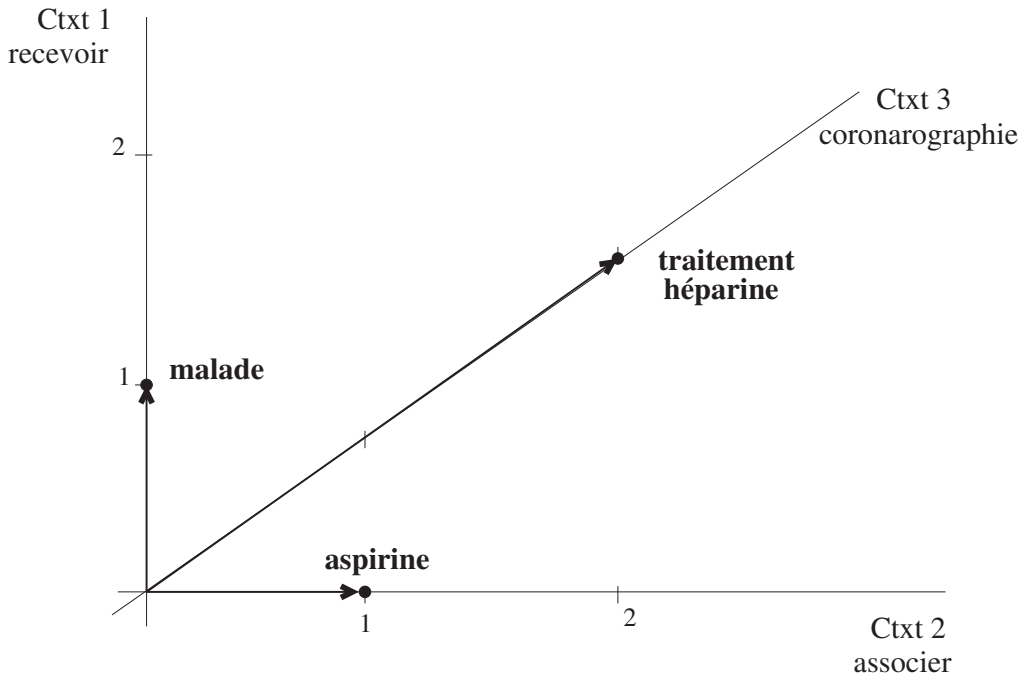

FIGURE 2.2: Exemple d'espace sémantique à trois dimensions, les trois dimensions étant les contextes *associer*, *recevoir* et *coronarographie* - à partir des phrases de l'exemple 1.

MOTS CIBLES	CONTEXTES															
	être	réaliser	traitement	associer	héparine	aspirine	bêta-bloquant	malade	recevoir	coronarographie	devoir	arrêter	convenir	poursuivre	cours	base
traitement	1	1	0	1	2	0	0	1	1	1	0	0	1	1	1	1
héparine	0	0	2	1	0	1	1	0	1	1	0	0	0	0	0	0
aspirine	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
bêta − bloquant	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1
malade	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
coronarographie	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
cours	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
base	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0

FIGURE 2.3 : Matrice *terme-contexte* pour les phrases de l'exemple 1.

proches l'un de l'autre dans cet espace, comme *héparine* et *traitement* et les mots sémantiquement distincts sont éloignés [Widdows et Ferraro, 2008].

2.2.2 Matrice de co-occurrence

L'espace vectoriel de grande dimension peut être construit à travers la collecte des données au sein d'une matrice de co-occurrence. Les éléments de la matrice correspondent au nombre d'occurrences des mots cibles dans différents contextes. Ainsi, les informations contenues dans le texte sont collectées dans la matrice de manière à ce que les lignes correspondent aux mots cibles et les colonnes aux contextes de ces mots cibles. Nous disposons alors sous forme tabulaire des co-occurrents de chaque mot cible.

Nous reprenons l'exemple initial, et nous créons la matrice présentée à la figure 2.3 avec les informations contenues dans ces phrases, en utilisant pour la définition du contexte une fenêtre graphique d'une taille de 5 mots (2 mots avant et 2 mots après le mot cible), de laquelle on retire les mots vides (déterminants, préposition, etc.). Les contextes de $bêta\text{-}bloquant_{\{TERME\}}$ sont alors $être_{\{VERBE\}}$, $héparine_{\{TERME\}}$, $aspirine_{\{TERME\}}$, $cours_{\{NOM\}}$, et $base_{\{NOM\}}$. Ainsi, beaucoup d'éléments de la matrice sont à zéro car peu de contextes sont associés à un mot cible.

Ainsi, la liste de co-occurences du mot cible *bêta-bloquant*$_{\{TERME\}}$ correspond au *vecteur de contexte* [1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1], déterminé par les lignes correspondantes dans la matrice. Les coordonnées décrivent une localisation dans l'espace à n dimensions.

Les VSM ont ainsi l'avantage de permettre une quantification facile de la proximité sémantique entre deux mots en mesurant la distance entre deux vecteurs au sein de cet espace (par exemple, en calculant le cosinus de leur angle). Le problème principal des vecteurs de contextes est que leur nombre de dimensions correspond au nombre de contextes, c'est-à-dire à la taille du vocabulaire du corpus. Les modèles vectoriels souffrent donc d'un nombre important de dimensions, mais également de la dispersion des données dans la matrice de co-occurrences [Chatterjee et Mohan, 2008]. Nous abordons ces limites dans la section 2.3 et présentons les solutions possibles à la section 2.4.

2.3 Limites : dispersion des données

Les VSM sont ainsi limités par la dispersion des données dans la matrice de co-occurrence : beaucoup d'éléments de la matrice sont à zéro car généralement peu de contextes sont associés à un mot cible. La matrice est alors qualifiée de *creuse* [Turney et Pantel, 2010]. [Sahlgren, 2006] constate 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 des fréquences basses, c'est-à-dire 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, en général très volumineux [Weeds et Weir, 2005, van der Plas, 2008], et les textes de spécialité, généralement de plus petite taille et caractérisés par un vocabulaire aux plus petites fréquences. Ainsi, même dans un gros corpus tel que le BNC (100 millions de mots), moins de 14% des mots ont une fréquence 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 fréquences de vocabulaire plus élevées.
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 VSM. Ainsi, même s'il est difficile de saisir la structure sémantique sous-jacente des matrices creuses, plus une matrice est creuse, moins le VSM est performant sur la tâche donnée. Ce rapport est équivalent au rapport liant la fréquence des mots et la qualité des vecteurs [Bullinaria et Levy, 2007] : plus les mots sont fréquents, plus le VSM est performant [Ferret, 2013b, Weeds et Weir, 2005, van der Plas, 2008]. En effet,

la similarité entre les mots cibles à faible fréquence 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 à faible fréquence 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 sont rares, mais peuvent correspondre à des contextes caractéristiques.

Pour construire l'espace vectoriel, la méthode distributionnelle est fondée sur des éléments statistiques. Ainsi, si les données ne sont pas suffisantes, il n'est pas possible de disposer d'informations statistiques suffisantes pour la construction du modèle distributionnel. De plus, la matrice de co-occurrence peut alors devenir extrêmement large quelle que soit la taille du corpus, et l'efficacité de l'algorithme en est alors affecté [Sahlgren, 2006]. Le dilemne 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.

2.4 Solutions aux limites de l'AD

Pour 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 contextes. L'objectif est de réduire l'espace, c'est-à-dire la mémoire occupée, et le temps de traitement.

2.4.1 Influence sur les contextes

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 les fréquences des contextes à l'état brut comme il est d'usage, mais en ordonnant les contextes en fonction de leur fréquence. 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 inter-substituables, c'est-à-dire 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 basant sur les fréquences de co-occurrence 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 co-occurrence, avant d'utiliser un algorithme de clustering. 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'influence sur les contextes peut également être réalisée en y intégrant de l'information sémantique supplémentaire. En effet, il a été démontré que l'intégration de ce type d'information afin de modifier la mise en œuvre classique de l'AD permet d'améliorer sa performance [Tsatsaronis et Panagiotopoulou, 2009]. Cette information sémantique, ou plus précisément les relations sémantiques, peuvent être calculées automatiquement ou provenir d'une ressource existante. Ainsi, avec un amorçage, [Zhitomirsky-Geffet et Dagan, 2009] modifient les poids des éléments au sein des contextes en se basant sur les voisins sémantiques trouvés à l'aide d'une mesure de similarité distributionnelle. En s'appuyant sur ces travaux, [Ferret, 2013b] s'intéresse au problème des mots de faibles fréquences. Afin de mieux prendre en compte ces informations, 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 de faible ou moyenne fréquence.

2.4.2 La réduction de dimensions (par exemple, la projection aléatoire)

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]. Le calcul de la similarité entre toutes les paires de vecteurs est une tâche coûteuse. Or, seuls les vecteurs qui partagent une dimension différente de zéro doivent être comparés (i.e. deux vecteurs ne partageant pas de dimension ne sont pas similaires). La plupart des modèles connus utilisent des méthodes de réduction de dimension, généralement mises au point de manière à conserver les mots faiblement fréquents.

2.4.2.1 Modèles basés sur la Décomposition aux Valeurs Singulières (SVD)

Une solution consiste à projeter les données aux dimensions élevées dans un espace au 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 Singulères (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 basées sur la SVD permettent ainsi de produire des

vecteurs de contexte moins creux et moins affectés par le bruit statistique. A partir des données initiales, c'est-à-dire la matrice de co-occurrences, cette technique divise cette matrice en composants linéaires indépendants. Ces composants sont en quelque sorte une abstraction, des jeux de valeurs qui s'approchent au mieux de la structure sous-jacente des jeux de données le long de chaque dimension prise indépendemment, sans les corrélations bruitées des données initiales. Parce que la majorité de ces composants sont très petits, ils peuvent être ignorés, et le résultat est une approximation des données qui contiennent en grande partie moins de dimensions que la matrice initiale.

La SVD est une méthode de factorisation de matrices coûteuse utilisée dans l'*Analyse Sémantique Latente (LSA)* [Landauer et Dumais, 1997] pour réduire la matrice de co-occurrence. La LSA est une méthode permettant de calculer des vecteurs sémantiques, ou vecteurs de contextes, à grande dimension, à partir des statistiques de co-occurrence des mots cibles. Elle permet :

- l'accès au sens latent des mots et de leurs contextes [Deerwester *et al.*, 1990, Landauer et Dumais, 1997]. Ainsi, en limitant le nombre de dimensions latentes, elle permet aussi une meilleure correspondance entre les mots et leurs contextes. Cette correspondance améliore le calcul de la similarité.
- la réduction du bruit [Rapp, 2003].
- la découverte de co-occurrences d'ordre élevé ou indirectes, c'est-à-dire quand deux mots en contexte apparaissent dans des contextes similaires [Landauer et Dumais, 1997]. Ceci correspond au but de l'analyse distributionnelle.
- la réduction de dispersion. En général, la matrice initiale est très creuse, mais la matrice issue de la SVD tronquée est dense.

En plus d'une réduction du coût de traitement, la réduction de dimension améliore considérablement la précision dans les applications de LSA. Ceci s'explique notamment car avec les mesures de similarité les plus fréquemment employées, les termes sont vus uniquement comme similaires s'ils apparaissent dans le même document dans la matrice terme-document complète. En ce qui concerne les mots de faible fréquence, la SVD tronquée est une manière de simuler le texte manquant, en compensant le manque de données [Vozalis et Margaritis, 2003]. Cependant, un des inconvénients de la SVD est que les dimensions cachées induites sont difficiles à interpréter [Clark, 2014].

Depuis les travaux de [Deerwester *et al.*, 1990], des recherches ont mené à plusieurs alternatives en ce qui concerne le lissage de matrices, avec notamment :
- Factorisation Matricielle Non-négative (NMF) [Zheng *et al.*, 2011, Lee et Seung, 1999],
- Indexation Sémantique Latente Probabiliste (PLSI) [Hofmann, 1999],
- Scaling Itératif (IS) [Ando, 2000],
- Analyse en Composantes Principales à Noyau (KPCA) [Schölkopf *et al.*, 1999],
- Allocation de Dirichlet latente (LDA) [Blei *et al.*, 2003]
- Analyse en Composantes Discrètes [Buntine et Jakulin, 2005].
- Analyse en Composantes Indépendantes (ICA)

– Hyperspace Analogue to Language (HAL) [Lund et Burgess, 1996, Lund *et al.*, 1995]
Ces autres algorithmes de lissage ont tendance à être plus intensifs computationnellement que la SVD, mais ils modélisent mieux les fréquences des mots que la SVD.

Ces méthodes correspondent toutes à un processus d'optimisation généralement non supervisé. Face à ces méthodes, un nouvel ensemble de modèles vectoriels, fondés sur un apprentissage supervisé, a vu le jour ces dernières années et fait l'objet de nombreux travaux. Il s'agit des réseaux de neurones, également appelés modèles prédictifs, dont les travaux fondateurs sont ceux de [Bengio *et al.*, 2003]. 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].

L'évaluation des réseaux de neurones, de manière intrinsèque ou extrinsèque, a démontré que cette méthode obtient de très bons résultats en termes de similarité sémantique [Baroni *et al.*, 2014, Mikolov *et al.*, 2013].

2.4.2.2 Random Indexing (RI) ou projection aléatoire

Face à ces différentes méthodes de lissage de matrice, supervisées ou non supervisées, la méthode du *Random Indexing* a émergé comme une alternative aux modèles de sémantique distributionnelle qui dépendent de la SVD pour l'étape de la réduction de dimensions dans la génération de vecteurs de contexte. Cette méthode permet notamment un gain significatif en temps de traitement et réduction de la mémoire utilisée pour le calcul de similarité sémantique à partir de corpus volumineux [Kanerva *et al.*, 2000, Karlgren et Sahlgren, 2001]. En effet, si la réduction de dimensions facilite le traitement des vecteurs de contextes, cela ne résoud pas le problème initial de construction d'une matrice de co-occurrence potentiellement immense. Même les implémentations telles que la LSA, qui utilisent des réductions de dimensions puissantes, nécessitent une première étape de collecte initiale des données au sein d'une matrice de co-occurrence. Le calcul de la réduction de dimensions est ainsi très lourd. Le *Random Indexing* a pour avantage de ne pas nécessiter, contrairement aux autres modèles vectoriels, de représentation sémantique ni de traitements lourds comme la SVD pour la LSA.

Décrit précisément par [Sahlgren, 2005], le *Random Indexing* est un modèle d'espace de mots incrémental, qui permet ainsi d'éviter la construction d'une trop grande matrice. L'auteur se base sur les travaux de [Kanerva *et al.*, 2000] sur les représentations distribuées éparses. Au lieu de collecter dans une matrice les co-occurrences puis d'extraire les vecteurs de contexte de cette matrice, le RI accumule incrémentalement les vecteurs de contexte qui à leur tour peuvent être assemblés dans une matrice de co-occurrence. Ainsi, un plus petit nombre de dimensions d est choisi *a priori* comme

un paramètre du modèle, et ensuite les vecteurs de contextes de d-dimensions sont construits de manière incrémentale.

L'accumulation des vecteurs de contexte se fait en deux temps :
– Chaque mot dans le contexte est assigné à un vecteur unique et généré aléatoirement, appelé *vecteur d'index*. Un petit nombre de valeurs 1 et -1 sont distribuées aléatoirement, et le reste des éléments sont à zéro. En générant ainsi des vecteurs creux, au nombre de dimensions suffisamment élevé, les représentations des contextes seront avec une très forte probabilité *orthogonaux*.
– Chaque mot dans le contexte est également assigné à un vecteur de contexte au même nombre de dimensions. Les vecteurs de contexte sont alors accumulés avec l'information des contextes en ajoutant les vecteurs d'index aux contextes dans lesquels les mots cibles apparaissent.

Les similarités extraites avec le RI sont de qualité équivalente à la LSA [Kanerva *et al.*, 2000], notamment sur une tâche de sélection de synonymes [Karlgren et Sahlgren, 2001]. Ainsi, [Karlgren et Sahlgren, 2001] ont démontré que le RI produit des résultats similaires à la LSA en utilisant comme évaluation le *Test of English as a Foreign Language (TOEFL)*.

2.5 Bilan

Dans ce chapitre, nous avons présenté les méthodes distributionnelles existantes, leurs paramètres et leurs limites.

Nous nous intéressons dans cette thèse à l'adaptation des méthodes distributionnelles pour les corpus de spécialité. L'étude des paramètres distributionnels nous permet de définir ceux à adapter pour notre méthode. De plus, étant donné que les corpus de spécialité se caractérisent, entre autres, par des faibles fréquences, ils sont touchés par le problème de dispersion des données au sein de la matrice de contextes.

Pour faire face à cette limite, nous proposons d'ajouter des informations sémantiques dans les contextes distributionnels, à l'instar de [Tsatsaronis et Panagiotopoulou, 2009, Ferret, 2013b]. Cependant, notre objectif diffère : nous intégrons des relations sémantiques acquises automatiquement dans les contextes afin de réduire le nombre de contextes et ainsi augmenter leur fréquence. De plus, si les méthodes basées sur la SVD limitent les contextes en supprimant de l'information, nous proposons au contraire de conserver cette information et de regrouper les contextes en les généralisant à l'aide de ces connaissances sémantiques supplémentaires, calculées sur le corpus de travail.

Chapitre 3

Méthode d'abstraction des contextes distributionnels

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, est souvent très creuse (beaucoup d'éléments ont une valeur nulle). Pour tenter de résoudre ce problème, nous proposons une approche consistant à densifier la matrice des contextes. Pour ce faire, nous proposons de réaliser une abstraction des variations superficielles ou des contextes peu significatifs statistiquement ou liés au bruit de la méthode d'identification de ces distributions. Pour cela, nous avons cherché, dans un premier temps, à filtrer les contextes de manière à sélectionner ceux qui semblent les plus pertinents, et surtout, à réaliser une abstraction des contextes, en les généralisant et en les normalisant à l'aide d'informations sémantiques extraites des corpus.

Dans ce chapitre, nous commençons par décrire le processus d'analyse distributionnelle mis en œuvre, puis nous présentons la méthode de généralisation et de normalisation des contextes distributionnels que nous proposons.

3.1 Méthode distributionnelle

Les méthodes d'analyse distributionnelle identifient la similarité entre les mots d'un texte à partir des contextes que ces mots partagent. Les informations statistiques sur les contextes partagés permettent de calculer la proximité distributionnelle de ces mots. Pour y parvenir, plusieurs paramètres entrent en jeu : les mots cibles en relation, le type de contexte utilisé (fenêtres graphiques ou analyse syntaxique), les mesures de similarité et de pondération.

La méthode d'analyse distributionnelle que nous avons mise en œuvre suit le schéma présenté dans la figure 3.1. L'abstraction des contextes se trouve au cœur de la méthode, entre la définition des contextes et des mots cibles et le calcul de similarité sémantique. 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.

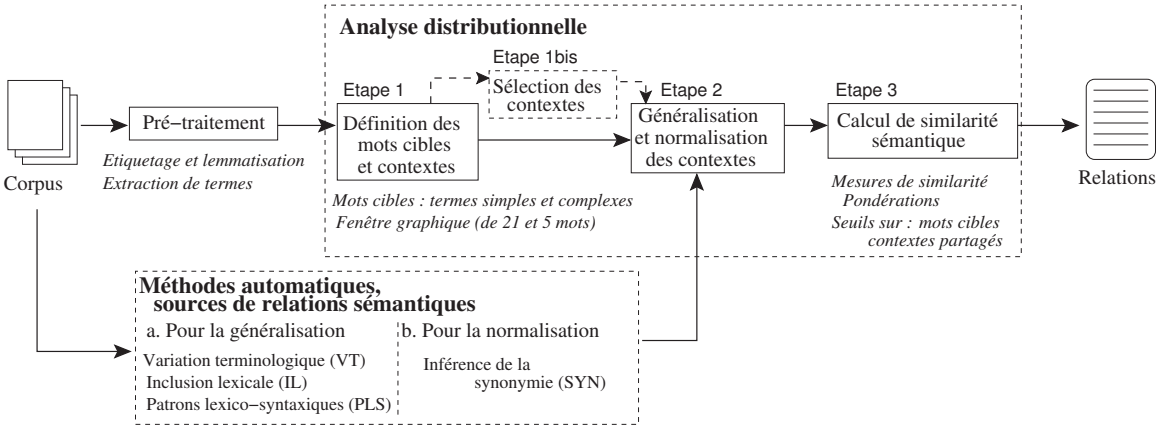

FIGURE 3.1 : Processus d'analyse distributionnelle.

3.1.1 Définition des mots cibles et des contextes (étape 1)

Dans le contexte d'applications en langue de spécialité, l'identification de relations sémantiques entre termes (simples et complexes) est primordiale. Pour cela, nous nous restreignons à l'analyse distributionnelle entre termes, qui constituent pour nous les mots cibles. Nous entendons par *termes* les *candidats termes* extraits par un extracteur de termes (cf. section 4.1.3 pour l'étape de l'extraction de termes). Les termes complexes ne sont pas figés, c'est-à-dire que nous prenons en compte à la fois le terme complexe et les termes simples et complexes qui le composent. Par exemple, nous considérons le terme complexe *souffle systolique d'éjection* identifié automatiquement, tout en prenant également en compte les termes *souffle sytolique*, *éjection*, *souffle* dans les mots cibles. Dans les contextes, la méthode se comporte de la même manière, et intègre également l'adjectif *systolique*.

Comme contextes distributionnels des mots cibles, nous avons choisi d'utiliser des fenêtres graphiques d'une largeur donnée. La notion de *contexte* fait référence pour nous à une unité lexicale qui apparaît dans le voisinage du mot cible. Bien que les contextes soient généralement calculés sur des dépendances syntaxiques, nous avons choisi d'utiliser des contextes graphiques au sein d'une phrase et autour d'un mot cible. Nous avons fait ce choix car il nous était trop coûteux à la fois d'intégrer et de mettre en place l'analyse syntaxique dans l'analyse distributionnelle, mais aussi d'adapter la méthode d'abstraction. De plus, les fenêtres graphiques étant moins restreintes que l'analyse syntaxique, elles permettent de prendre en compte un plus grand nombre de contextes, ce qui facilite la généralisation/normalisation.

Ainsi, les contextes sont composés de mots pleins qui co-occurrent avec le mot cible au sein de la fenêtre graphique. Dans un premier temps, nous avons gardé comme contexte les mots *pleins*, c'est-à-dire les adjectifs, les noms, les verbes, les adverbes

et les termes, en écartant les mots vides (déterminants, conjonctions, etc.). Le choix des mots vides à écarter n'a pas été un choix facile, et nous avons été amenée à revoir les catégories morpho-syntaxiques prises en compte dans les contextes. En effet, nous pensions trouver dans les corpus de travail suffisamment d'adverbes propres aux domaines, de manière à conserver dans les contextes des mots suffisament spécifiques. Or, l'analyse manuelle des adverbes révèle que seuls quelques adverbes sont spécifiques au domaine auquel appartient le corpus : par exemple, pour le français, *nutritionnellement*, *anatomiquement*, *angiographiquement*, *radiologiquement*, *chirurgicalement*, etc., et pour l'anglais, *cardiovascularly*, *clinically*, *hemodynamically*, *neurologically*, etc. Cependant, ces adverbes sont minoritaires dans nos corpus, et sont pour la plupart très généraux ; par exemple, pour le français, *autrement*, *beaucoup*, *éventuellement*, *très*, *majoritairement*, *néanmoins*, etc., et pour l'anglais, *necessarily*, *slightly*, *easily*, *inadvertently*, etc. Ainsi, finalement, nous avons fait le choix d'écarter les adverbes des contextes, étant donné qu'ils sont trop généraux et trop peu spécifiques aux mots cibles avec lesquels ils co-occurrent.
Les mots pris en compte dans les contextes sont donc : les adjectifs, les noms, les verbes et les termes.

Pendant le pré-traitement des corpus (cf. section 4.1.3), l'étiquetage morphosyntaxique et l'extraction de termes nous permettent d'identifier les mots cibles et les contextes. L'étiquetage morphosyntaxique permet de sélectionner les parties du discours utilisées en tant que mots cibles et dans les contextes (pour ne pas prendre en compte les prépositions dans les contextes, par exemple). L'intégration des termes identifiés par un extracteur de termes permet de prendre en compte des termes complexes, utiles dans les textes de spécialité. Nous intégrons également la lemmatisation dans le pré-traitement des corpus. Elle est utile pour l'abstraction des contextes distributionnels, car elle permet de normaliser les mots cibles et contextes, de réduire les variations et constitue une première abstraction lexicale. Enfin, l'intégration de l'étiquetage morphosyntaxique et de la lemmatisation permet l'application de notre méthode à plusieurs langues.

Exemple 2

L'[**examen clinique**$_{\underset{\mathbf{w}}{\_\,terme}}$] *[montre*$_{\underset{w_i,}{\_verbe}}$*]* un *[souffle systolique lombaire*$_{\underset{w_i,}{\_\,terme}}$*]*

L'[**examen clinique**$_{\underset{\mathbf{w}}{\_\,terme}}$] *[retrouve*$_{\underset{w_i,}{\_verbe}}$*]* une *[insuffisance cardiaque*$_{\underset{w_i,}{\_\,terme}}$*]*

La [**coronarographie**$_{\underset{\mathbf{w}}{\_\,terme}}$] *[confirme*$_{\underset{w_i,}{\_\,verbe}}$*]* une *[sténose serrée*$_{\underset{w_i,}{\_\,terme}}$*]*

La [**coronarographie**$_{\underset{\mathbf{w}}{\_\,terme}}$] *[montre*$_{\underset{w_i,}{\_verbe}}$*]* une *[atteinte pluritronculaire*$_{\underset{w_i,}{\_\,terme}}$*]*

Dans l'exemple 2, nous cherchons à mettre en relation les mots cibles w : *examen clinique*$_{\_terme}$ et *coronarographie*$_{\_terme}$. Nous prenons l'exemple de ces deux mots cibles pour expliquer la méthode, mais la méthode fera cette recherche pour tous les couples

de mots. Pour cela, nous prenons en considération leurs contextes $w_{i'}$, c'est-à-dire les mots contenus dans la fenêtre graphique autour de chaque mot cible. Ici, le mot cible *examen clinique*$_{\_terme}$ a pour contextes les verbes *montrer*$_{\_verbe}$ et *retrouver*$_{\_verbe}$, et les termes *souffle systolique lombaire*$_{\_terme}$ et *insuffisance cardiaque*$_{\_terme}$.

3.1.2 Sélection des contextes (étape 1bis)

Lors de la définition des contextes, nous ajoutons une étape facultative de sélection des contextes. Cette étape nous permet de considérer les contextes en fonction de leur caractère discriminant ou non. En effet, parmi les contextes d'un mot, certains sont de meilleurs descripteurs que d'autres [Morlane-Hondère, 2013].

Par exemple, dans le corpus Recettes, avec l'utilisation d'une fenêtre graphique de 5 mots, le nom *vanille*$_{\_NOM}$ co-occurre avec des mots tels que *avoir*$_{\_VER}$ et *exemple*$_{\_NOM}$. Ces contextes très fréquents (respectivement 47 297 et 10 950 occurrences) apparaissent respectivement dans le contexte de 1 478 et 884 lemmes différents : ils sont trop généraux et ne donnent ainsi que peu d'informations sémantiques sur le nom *vanille*$_{\_NOM}$. A l'inverse, des contextes moins fréquents comme *aromatiser*$_{\_VER}$ et *glace*$_{\_NOM}$ (ayant une fréquence respective de 162 et 302), qui co-occurrent avec un ensemble beaucoup plus réduit de lemmes (respectivement 98 et 143), sont plus caractéristiques de *vanille*$_{\_NOM}$ et beaucoup plus pertinents pour sa description sémantique. Ainsi, leur co-occurrence avec *vanille*$_{\_NOM}$ est beaucoup plus significative et porteuse d'informations. Généralement, ce rapport d'exclusivité entre un mot et ses contextes se calcule à l'aide de mesures d'association, comme nous le faisons lors du calcul de similarité (étape 3), une fois que les mots cibles et leurs contextes ont été définis. Nous avons fait le choix d'expérimenter l'élimination des contextes les plus généraux lors de la définition du contexte, en adaptant la méthode standard du Tf.Idf [Jones, 1972] à notre besoin. Tout d'abord, les documents sont les mots cibles (w), et les termes correspondent dans notre cas aux mots dans le contexte (w_i). Nous supprimons le logarithme, car il atténue beaucoup trop les écarts entre les valeurs, ceux-ci étant déjà faibles. Pour nous différencier du Tf.Idf, nous nommons notre adaptation Cf.Itf (1) :

$$Cf.Itf = |(w, r, w_i)| \cdot \frac{n(w)}{n(w)|\exists(w,r,w_i)} \quad (1)$$

où (w, r, w_i) correspond au nombre d'occurrences des mots w et w_i se trouvant en relation r, et $n(w)$ est le nombre de mots cibles.

Une fois calculé le Cf-Itf de chaque contexte d'un mot cible donné, nous procédons à la sélection des contextes, suivant leur Cf-Itf (cf. figure 3.2). Pour chaque mot cible, nous ordonnons les contextes suivant la valeur de Cf-Itf par ordre décroissant. Nous calculons ensuite l'écart entre les valeurs successives de Cf-Itf des contextes. L'objectif est de supprimer les contextes les plus généraux. Il faut donc trouver un bon équilibre,

un écart suffisamment important entre deux contextes, sans pour autant supprimer trop de contextes.

Après observation du comportement des écarts moyens entre deux contextes sur le corpus Recettes, nous avons défini expérimentalement le seuil à "l'écart moyen x 1,5". Ainsi, si l'écart moyen entre le contexte courant et le suivant est inférieur à l'écart moyen x 1,5, nous conservons le contexte. En revanche, dès lors où l'on rencontre un écart supérieur à la moyenne x 1,5, nous ignorons le reste des contextes pour ce mot cible.

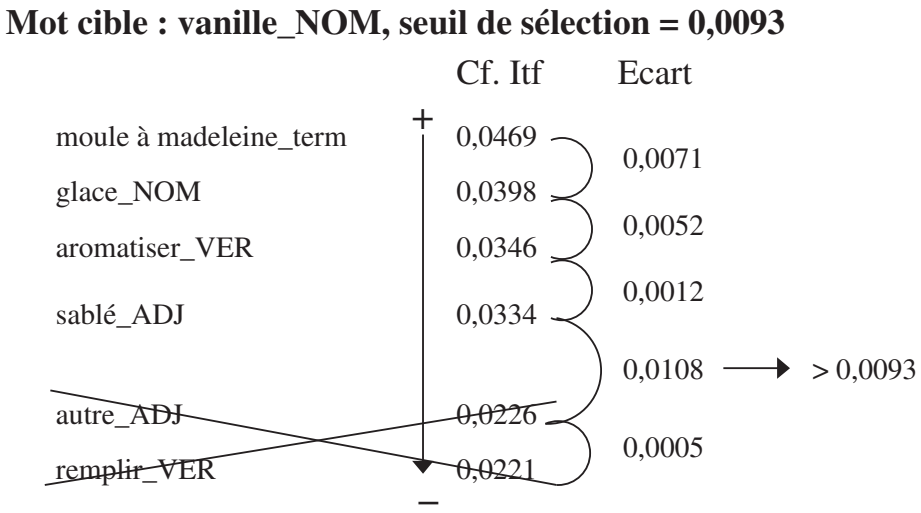

FIGURE 3.2 : Exemple de sélection des contextes à l'aide du Cf-Itf pour le mot cible *vanille* (corpus *Recettes*), avec pour seuil l'écart moyen x 1,5.

Prenons pour exemple le nom *vanille*, mot cible dans le corpus *Recettes* (figure 3.2). Les contextes de *vanille* sont triés par valeur de Cf-Itf, dans l'ordre décroissant. Nous avons calculé l'écart moyen pour ces contextes, qui est de 0,0062. En le multipliant par 1,5 nous obtenons 0,0093, le seuil de sélection des contextes. Ainsi, dès lors où l'écart entre deux contextes est supérieur à 0,0093, nous arrêtons la sélection des contextes.

3.1.3 Calcul de la similarité sémantique (étape 3)

Une fois que 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]. Dans la section 2.1.3, nous avons décrit les principales mesures de similarité. Parmi ces mesures, nous expérimentons : le cosinus, l'indice de Jaccard, la fréquence des contextes partagés ainsi que le nombre des contextes partagés. Nous expérimentons également deux mesures de pondération

pour mesurer le degré d'association entre un contexte et un mot cible, la fréquence relative et l'information mutuelle.

Lors du calcul de similarité entre les mots cibles, si aucune sélection n'a été effectuée à l'aide du Cf.Itf, 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 : la fréquence des mots cibles, la fréquence et le nombre des contextes partagés. Pour chaque paramètre, un seuil est calculé automatiquement en fonction du corpus. Parmi ces trois paramètres, deux d'entre eux sont appliqués aux contextes, et le troisième est appliqué aux mots cibles.

Pour mieux comprendre la définition des seuils, nous nous appuyons sur l'exemple suivant (exemple 3).

Exemple 3

L'[examen clinique$_{_terme}$] ***[montre$_{_verbe}$]*** une ***[sténose serrée$_{_terme}$]***

L'[examen clinique$_{_terme}$] [retrouve$_{_verbe}$] une [insuffisance cardiaque$_{_terme}$]

La [coronarographie$_{_terme}$] [confirme$_{_verbe}$] une ***[sténose serrée$_{_terme}$]***

La [coronarographie$_{_terme}$] ***[montre$_{_verbe}$]*** une *[atteinte pluritronculaire$_{_terme}$]*

- **Nombre de contextes partagés** : nombre de contextes lemmatisés.
  Les deux mots cibles *examen clinique*$_{\_terme}$ et *coronarographie*$_{\_terme}$ partagent les contextes *montrer*$_{\_verbe}$ et *sténose serrée*$_{\_terme}$; le nombre de contextes partagés est 2.
- **Fréquence des contextes partagés** : nombre d'occurrences des contextes lemmatisés.
  Dans l'exemple précédent, les deux contextes partagés *montrer_verbe* et *sténose serrée*$_{\_terme}$ ont chacun une fréquence de 2.
- **Fréquence des mots cibles** : nombre d'occurrences des mots cibles lemmatisés. Les mots cibles *examen clinique*$_{\_terme}$ et *coronarographie*$_{\_terme}$ ont chacun une fréquence de 2.

L'utilisation de seuils sur les contextes et mots cibles devrait nous permettre d'éliminer les relations les moins pertinentes. Les relations dont la valeur d'un des paramètres est inférieure à ce seuil sont écartées. Les valeurs des paramètres pour l'exemple 3 sont présentées dans le tableau 3.1. Pour définir ce seuil, nous utilisons la moyenne des valeurs prises par les paramètres. Nous avons écarté la médiane, après analyse des fréquences des mots cibles et des contextes de nos corpus (cf. section 4.1) : nous avons dans nos corpus un grand nombre de faibles fréquences, égales à 1 qui positionne la médiane à 1 dans la plupart des cas, et rend alors le seuillage inutile.

Mots cibles	Freq. mots cibles	Contextes partagés Nombre = 2	Freq. ctxt part.
coronarographie$_{__terme}$	2	montrer$_{__verbe}$	2
examen clinique$_{__terme}$	2	sténose serrée$_{__terme}$	2

TABLEAU 3.1 : Statistiques distributionnelles pour l'exemple 3.

3.2 Règles d'abstraction des contextes distributionnels

Notre approche d'abstraction des contextes intervient au sein d'un processus de sémantique distributionnelle. Nous adaptons la méthode distributionnelle en définissant des paramètres distributionnels adaptés à nos corpus de travail (cf. section 5.1), notamment en intégrant les contextes généralisés et normalisés. Le processus d'abstraction des contextes intervient après l'étape de définition des contextes. Il correspond à la deuxième étape de la méthode présentée dans la figure 3.1.

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. Ceci permet 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. Ainsi, 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 sémantiquement dans un même cluster ou 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, le même sens. L'objectif est d'une part, de diminuer la diversité des contextes distributionnels (l'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, c'est-à-dire leur fréquence. Le contexte *siège*, s'il remplace ces trois termes, a alors une occurrence de 3.

Dans cette perspective, nous avons choisi d'utiliser des informations sémantiques additionnelles calculées sur le corpus, et fournissant ainsi des indices d'abstraction. Aussi, nous regroupons sous la notion d'*abstraction* deux types d'abstraction, qui se définissent et se différencient par rapport au concept terminologique : la généralisation et la normalisation. La généralisation est réalisée à l'aide de relations d'hyperonymie, qui sont des relations *conceptuelles*, car liant deux concepts. Nous qualifions d'*abstraction conceptuelle* la généralisation. La normalisation est menée avec des relations de synonymie, des relations *lexicales* reliant deux unités lexicales individuelles, au sein d'un même concept. Nous nommons *abstraction lexicale* la normalisation.

Ainsi, nous avons défini deux types de règles. Nous présentons tout d'abord les règles de généralisation puis la règle de normalisation.

3.2.1 Règles de généralisation des contextes

Ainsi, une fois que les mots cibles et les contextes ont été définis, nous généralisons les contextes avec des ensembles de relations d'hyperonymie acquises par des méthodes automatiques à partir du corpus de travail. Ces relations doivent fournir des indices de généralisation. Ces méthodes sont décrites à la section 3.3 : patrons lexico-syntaxiques dédiés à l'hyperonymie (PLS), inclusion lexicale (IL), et variation terminologique (VT).

Les deux premières méthodes proposent principalement des relations d'hyperonymie et seront utilisées pour généraliser les contextes. En revanche, la variation terminologique ne propose pas de relations typées sémantiquement (relations étiquetées par un type sémantique, comme par exemple la *synonymie*, l'*hyperonymie*). Aussi, étant donné que l'opération d'insertion est la seule utilisée pour acquérir des variantes, nous avons considéré que les relations obtenues étaient des relations d'hyperonymie. Le terme hyperonyme et le terme 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*).

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*, *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 la fréquence des hyperonymes $|H_1|, \ldots, |H_n|$ dans le corpus, et nous choisissons l'hyperonyme dont la fréquence est la plus élevée 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 est le plus fréquent est choisi et utilisé pour remplacer *artère* dans le contexte.

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).

3.2.2 Règle de normalisation des contextes

La règle de normalisation vise à réduire les variations sémantiques à l'aide de relations de synonymie. Ces relations sont tout d'abord regroupées sous la forme de clusters de synonymes et un des synonymes du cluster est choisi comme représentant ; il s'agit du mot le plus fréquent au sein de ce cluster.

Alors, à chaque mot w_i dans le contexte du mot cible w, correspond un cluster 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 cluster 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 cluster de synonymes, il est remplacé par le représentant du cluster.

Par exemple, si le terme *altération métabolique* dans le contexte d'un mot cible appartient au cluster de synonymes fourni par la méthode d'acquisition de synonymes (*anomalie métabolique*, *maladie métabolique*, *troubles métaboliques* et *altération métabolique*) celui qui est le plus fréquent est choisi comme représentant et utilisé pour remplacer *altération métabolique* dans le contexte.

3.2.3 Combinaison des règles de normalisation et généralisation

Nous combinons les règles décrites précédemment en appliquant dans un premier temps les règles de normalisation, puis les règles de généralisation sur les mots dans les contextes normalisés $s(w_i)$ (représentés par la fonction composée $h \circ s(x)$).

Ceci nécessite la normalisation de chaque élément de l'ensemble d'hyperonymes $\mathbb{H}_S(w_i) = \{H_1, \ldots, H_n\}$ utilisés pour la généralisation. Les hyponymes sont ainsi normalisés à l'aide des relations de synonymie avant d'être utilisés dans l'étape de généralisation.

3.3 Méthodes d'acquisition de relations sémantiques pour l'abstraction des contextes

Dans cette section, nous présentons les différentes méthodes automatiques d'acquisition de relations que nous avons utilisées pour la phase d'abstraction des contextes.

Pour obtenir ces relations à partir de corpus, nous avons choisi d'utiliser plusieurs approches classiques d'acquisition de relations sémantiques entre termes : les patrons lexico-syntaxiques (PLS), l'hypothèse d'inclusion lexicale (IL), des règles de variation terminologique (VT), et des règles d'inférence de relations de synonymie (SYN). La généralisation des contextes utilise les méthodes visant l'acquisition de relations d'hyperonymie : patrons lexico-syntaxiques, inclusion lexicale et variation terminologique, alors que la normalisation des contextes est basée sur une méthode d'acquisition de relations de synonymie.

3.3.1 Patrons lexico-syntaxiques

Pour le français, nous utilisons les patrons définis par [Morin et Jacquemin, 2004] et pour l'anglais ceux définis par [Hearst, 1992], pour acquérir des relations d'hyperonymie entre termes simples ou complexes, soit par exemple :

Pour le français (ex : *artère coronaire droite - artère*)

1. {quelques | plusieurs} SN : LISTE

2. {autre} ? SN tels que LISTE

où SN est un syntagme nominal et LISTE une liste de syntagmes nominaux.

Pour l'anglais (ex : *disease - diabetes*)

1. SN {, SN}*{,} or other SN
2. SN {,} especially {SN,}*{or|and} SN

où SN est un syntagme nominal.

3.3.2 Inclusion lexicale

Cette approche s'appuie sur l'hypothèse selon laquelle si un terme (ex : *infarctus*) est inclus lexicalement dans un autre (ex : *infarctus du myocarde*) il existe généralement une relation d'hyperonymie entre ces deux termes [Grabar et Zweigenbaum, 2003]. Nous contraignons l'approche en exploitant l'analyse syntaxique des termes fournie par YaTeA [Aubin et Hamon, 2006]. Nous ne considérons ici que les relations syntaxiques entre le terme complexe, par exemple *souffle systolique*, et sa tête, *souffle*.

Si l'inclusion lexicale permet d'acquérir des relations d'hyperonymie avec une bonne précision, elle génère également des erreurs qui peuvent être liées à la qualité et à la rédaction du corpus (exemple 4), ou liées aux outils utilisés pour le pré-traitement du corpus (exemple 5). Ces erreurs liées aux outils correspondent à des erreurs d'étiquetage morphosyntaxique ou à une mauvaise segmentation au moment de l'extraction de termes.

Exemple 4

a.	*aluminium 2 - aluminium 2 feuilles*
b.	*ail persil - ail persil thym*

Exemple 5

a.	*bouillant - coco bouillant*
b.	*bordelais - cannelé bordelais*
c.	*arm last night - left arm last night*
d.	*bed - bed to commode*

De plus, l'inclusion lexicale permet d'obtenir des relations peu utiles d'un point de vue sémantique (exemple 6), mais utiles pour la généralisation en corpus. Ces relations représentent environ un cinquième des relations obtenues par inclusion lexicale. En effet,

certaines relations acquises par inclusion lexicale, comme celles données en exemple 6, sont utiles dans le processus de généralisation car elles permettent d'effacer des variations peu porteuses de sens.

Exemple 6

a.	*bouillabaisse - bouillabaisse facile*
b.	*ail hâchés - ail hâchés 1*

3.3.3 Variation morphosyntaxique

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 variations syntaxiques, morphologiques ou lexicales : principalement la permutation, la dérivation et l'insertion. Dans nos corpus en français, les règles utilisées pour identifier des relations sémantiques entre termes sont essentiellement l'insertion. En revanche, pour l'anglais, les trois règles sont utilisées.

L'insertion d'un modifieur, dans l'exemple 7, *de revascularisation*, au sein d'un terme complexe, ici *chirurgie coronarienne*, permet d'identifier une relation d'hyperonymie entre les deux termes concernés, *chirurgie coronarienne* et *chirurgie de revascularisation coronarienne*.

Exemple 7

a.	chirurgie coronarienne - chirurgie *de revascularisation* coronarienne
b.	bœuf maigre - bœuf *hâché* maigre
c.	abdominal pain - abdominal *muscle* pain

La permutation est une transformation qui associe un terme avec au moins un argument à gauche du nom tête, à une variante où un de ces arguments est à droite du nom tête [Jacquemin, 1997]. Cette variation est assez fréquente en anglais, et est présente uniquement dans notre corpus en anglais (Textes Cliniques). Elle fait passer le terme complexe de la composition germanique à une structure syntaxique. Cela se traduit principalement par le passage d'une construction épithète à une construction attribut, avec comme construction la plus fréquente celle fondée sur une préposition (*of*, *from*, *with*, et *to*). Nous obtenons très peu de relations avec cette règle, et les relations obtenues sont plus apparentées à des relations de synonymie qu'à l'hyperonymie (à l'instar de l'exemple 8).

Exemple 8

a.	*albumin bole - boluses of albumin*
b.	*antibiotic course - courses of antibiotics*

La dérivation concerne les variantes morphosyntaxiques et qui se distinguent en ayant au moins un des composants du terme complexe transformé en un autre mot dérivé de la même racine morphologique. De même que pour la permutation, sur nos corpus de travail, nous n'obtenons que très peu de variantes par dérivation, et ces relations sont également plus proches de la synonymie (comme dans l'exemple 9).

Exemple 9

a.	*artery pressure - arterial pressure*
b.	*biliary drain - biliary drainage*

La méthode que nous avons utilisée ne propose pas de relations typées sémantiquement. Aussi, étant donné que la méthode que nous utilisons met en jeu essentiellement la règle d'insertion qui permet d'identifier des relations d'hyperonymie entre termes complexes, nous avons considéré les relations obtenues comme des relations d'hyperonymie. Le terme hyperonyme et le terme 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*).

3.3.4 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 al.*, 1998]. Cette méthode s'appuie sur le principe de compositionnalité sémantique, pour inférer une relation de synonymie entre des termes complexes si au moins un de leurs composants sont synonymes. Il peut s'agir :

– de têtes identiques et d'expansions synonymes (*tarte aux brimbelles - tarte aux myrtilles*)
– de têtes synonymes et d'expansions identiques (*further problem - further trouble*)
– de têtes synonymes et d'expansions synonymes (*barbecue traditionnel - four classique*)

Pour cela, nous utilisons un dictionnaire existant, le dictionnaire de langue générale, en français, *Le Robert*, qui contient des relations de synonymie entre termes simples.

La qualité des relations fournies par cette méthode est variable. Nous observons tout d'abord des erreurs liées au type de la relation acquise : les relations identifiées ne sont pas seulement des synonymes, mais aussi des relations d'hyperonymie (*agrume à la menthe - citron à la menthe*, *bordeaux rouge - vin rouge*), *fruits rouges - pommes rouges*, des co-hyponymes (*aneth frais - anis frais*, *salade laitue - salade scarole*), de la méronymie (*repas asiatique - soupe asiatique*).

Mises à part ces erreurs de typage des relations, la plupart des erreurs sont dues à un manque d'information sur le contexte des termes pour les mots polysémiques ou aux données bruitées contenues dans le dictionnaire. Ainsi, par exemple, si la relation entre les adjectifs *complet* et *parfait* peut avoir du sens dans certains contextes, dans un contexte alimentaire, la relation *pâtes complètes - pâtes parfaites* qui en est déduite est erronée. De même, si dans la langue générale les termes *lentille* et *verre* peuvent être liés sémantiquement, dans un contexte culinaire ces deux termes ne sont plus synonymes, et la relation *lentilles à l'eau - verre à eau* identifiée par la méthode est erronnée. Les adjectifs trop généraux, tels que *petit*, *gros* et *bon* permettent de mettre en relation des termes qui ne sont pas liés sémantiquement, tels que *sel fin - fines saveurs*, *grand luxe - bonne quantité*. Enfin, certains termes complexes sont des expressions figées et perdent ainsi de leur compositionnalité, ce qui engendre également des erreurs, comme par exemple les relations *huiles essentielles - grosses olives*, et *oranges à jus - mandarines au sirop*.

3.4 Conclusion

Dans ce chapitre, nous avons présenté notre méthode d'abstraction des contextes distributionnels. Cette méthode est intégrée à une méthode distributionnelle dont les paramètres peuvent être adaptés automatiquement. La méthode d'abstraction des contextes comprend deux types de règles, qui correspondent à deux types d'abstraction : les règles de généralisation, qui sont réalisées à l'aide de relations d'hyperonymie permettent une abstraction conceptuelle, et la règle de normalisation qui utilise des relations de synonymie pour normaliser les contextes au niveau lexical. Pour généraliser et normaliser les contextes, nous nous appuyons sur les résultats de méthodes automatiques d'acquisition de relations sémantiques. Nous avons présenté dans ce chapitre les différentes méthodes d'acquisition. Ces méthodes ne produisent pas nécessairement des relations correctes ou les relations attendues. A ce stade, nous ne savons pas si ces erreurs seront gênantes pour l'abstraction des contextes ou si elles pourront permettre d'améliorer le processus d'abstraction.

Chapitre 4

Corpus et évaluation

Dans ce chapitre, les corpus que nous avons utilisés pour nos expérimentations sont présentés. Nous décrivons également les méthodes d'évaluation sélectionnées pour mesurer la qualité des relations produites par l'analyse distributionnelle, l'impact des paramètres, ainsi que l'impact de l'abstraction des contextes distributionnels.

4.1 Corpus

Afin d'évaluer l'indépendance de notre méthode par rapport à la langue et au domaine, nous avons mené nos expériences sur des corpus appartenant à des domaines de spécialité différents, n'ayant pas les mêmes degrés de spécificité, et rédigés dans deux langues différentes : des recettes de cuisine et des guides alimentaires en français, ainsi que des textes cliniques en français et en anglais. De plus, étant donné que les méthodes distributionnelles sont sensibles à la taille du corpus [Bernier-Colborne, 2014], et que la problématique de dispersion des données à laquelle nous nous intéressons est liée aux fréquences du vocabulaire d'un corpus, nous avons également choisi des corpus de tailles différentes, variant de 84 839 mots pour le plus petit corpus (corpus Menelas) à 3 928 658 mots pour le plus volumineux (corpus Recettes). Nous regroupons ces quatre corpus par taille : les plus petits de nos corpus sont deux corpus du domaine médical, et ceux de plus grande taille appartiennent au domaine de l'alimentation.

Nous décrivons dans un premier temps les corpus utilisés dans chaque domaine. Nous présentons ensuite le pré-traitement linguistique des corpus réalisé pour l'ensemble de nos expériences.

4.1.1 Corpus de petite taille : corpus médicaux

Nous présentons dans cette partie les deux corpus du domaine médical que nous avons utilisés. Ces corpus sont des textes médicaux contenant des échanges entre spécialistes et se distinguent ainsi des autres corpus utilisés par un degré de spécialisation plus élevé. Nous avons utilisé deux corpus de langue différente : un premier corpus en langue française (Menelas [Zweigenbaum, 1994]) et un second corpus en anglais, fourni par la compétition I2B2/2012 désigné sous le nom *Textes Cliniques* dans la suite de la thèse [Sun *et al.*, 2013].
Dans les deux cas, les textes sont anonymisés de manière variable. Pour le corpus Menelas, les noms des personnes (médecins, patients) et de lieux (cliniques, hôpitaux, villes,

etc.) sont remplacés par une lettre ou une suite de lettres en majuscules (exemple 10), les dates sont remplacées par d'autres dates (exemple 11). Pour les dates, il s'agit d'une déidentification, c'est-à-dire que les dates sont remplacées par d'autres dates, de manière cohérente.

Exemple 10

a.	*Le patient rejoint donc ce jour la clinique de la MACPRO où il sera pris en charge par le Docteur MADSAV.*
b.	*Patient âgé de 52 ans, adressé pr le Dr. C.*

Exemple 11

a.	*En janvier 1985, l'épreuve d'effort était négative à 210 watts.*
b.	*Le 27.11.86, elle s'avérait positive et symptomatique dès 120 watts.*

Dans le tableau 4.1, nous décrivons ces deux corpus.
Les deux corpus ont des tailles différentes : le corpus français est deux fois plus petit que le corpus anglais. Les longueurs de phrases sont également différentes : les phrases du plus petit corpus, Menelas, sont les plus longues, avec en moyenne environ 17 mots par phrase, contre 11 mots par phrase en moyenne pour les Textes Cliniques. Nous pourrons établir une comparaison entre les deux domaines, mais également entre les deux langues.

	MENELAS	TEXTES CLINIQUES
Nombre de textes	56 textes	311 textes
Nombre de mots	84 839	178 070
Longueur moyenne des phrases	17,55 mots	11,08 mots
Langue	FR	EN
Contenus	Manuel de référence sur les maladies coronariennes + comptes rendus d'hospitalisation	Textes cliniques
Nombre de lemmes	NOMS = 1 733 TERMES = 7 654	NOMS = 3 279 TERMES = 15 028

Tableau 4.1 : Corpus de petite taille : corpus médicaux.

4.1.1.1 Corpus Menelas

Le corpus français Menelas a été constitué dans le cadre du projet du même nom qui visait à réaliser un système de compréhension de comptes rendus d'hospitalisation dans le domaine des maladies coronariennes [Zweigenbaum, 1994]. Le corpus comporte 84 839 mots. Il est constitué de deux grandes parties : un extrait d'un manuel de référence sur la coronographie 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). Nous en présentons un extrait dans la figure 4.1. On y trouve à la fois un extrait du manuel et un extrait des lettres des médecins. Ce corpus est à la fois le plus petit et le plus structuré de nos quatre corpus. Les phrases de ce corpus sont longues, et sont les plus longues de tous nos corpus, 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 l'extrait du 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.

Ciné-ventriculographie gauche.
La ciné-ventriculographie gauche est ensuite pratiquée en position oblique antérieure droite à 30 degrés. C'est le plus souvent la seule incidence obtenue ; toutefois, lorsqu'une dyskinésie ou même un anévrisme de la paroi inférieure et du septum sont suspectés, une seconde ventriculographie en incidence oblique antérieur gauche est également effectuée. La ventriculographie gauche pratiquée au cours du pacing auriculaire et surtout au cours de l'effort dynamique, peut constituer, au besoin, une méthode très sensible pour déceler l'ischémie myocardique.

...

Le patient, chauffeur routier, a présenté au volant de son car un malaise lipothymique associé à une douleur précordiale avec irradiation brachiale survenant dans un contexte d'hyperthermie (syndrome grippal depuis 2 jours).

FIGURE 4.1 : Extraits du corpus Menelas (le premier paragraphe est un extrait du manuel et le second paragraphe est un extrait d'une lettre de médecin).

4.1.1.2 Corpus de textes cliniques

Nous utilisons le corpus de textes cliniques en anglais, fournis par la compétition I2B2/2012 [Sun *et al.*, 2013] (178 070 mots), qui contient 311 documents provenant d'hôpitaux américains, fournis par *Partners HealthCare* et *the Beth Israel Deaconess Medical Center*. Il s'agit de dossiers patients qui ont été anonymisés par les organisateurs de la compétition. Nous présentons un extrait de ce corpus dans la figure 4.2. Ce corpus contient également des abréviations (ex. *b.i.d.*), en plus grand nombre que dans le corpus Menelas. En effet, les documents du corpus sont des comptes-rendus d'hospitalisation alors que le corpus Menelas contient des correspondances entre médecins. 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 ainsi le problème de dispersion des données.

ADMISSION DATE : 02/01/2000
DISCHARGE DATE : 02/08/2000
HISTORY OF PRESENT ILLNESS :
This patient is an seventy eight year old female with history of peripheral vascular disease who is status post above knee to fem pop bypass graft with 6 millimeter PTFE.
This operation was performed for two months of increased rest pain.
Her pain resolved after surgery and she has been doing well since, although at baseline now she is minimally ambulatory from bed to commode.
For the past couple of months the patient has had a non-healing right dorsal foot ulcer which has been increasing in size and started as a pin hole and she does not recollect any trauma, as similar small ulcers developed on the left foot as well around the same time, but that has subsequently healed.
The ulcer was managed conservatively at Har Hospital by Dr. Holes with Silvadene b.i.d. lower extremity non-invasive study obtained at that time showed poor distal right extremity perfusion.

FIGURE 4.2 : Extrait du corpus I2B2.

4.1.2 Corpus de grande taille : corpus alimentaires

Les plus volumineux de nos corpus appartiennent au domaine de l'alimentation : le premier contient des recettes de cuisine (désigné sous le nom de corpus *Recettes*) et le second des guides alimentaires (appelé par la suite corpus *Guides Alimentaires*) (cf. tableau 4.2). Ces deux corpus sont rédigés en français. Ils sont moins spécialisés que les corpus médicaux présentés précédemment (section 4.1.1), et s'adressent à un lecteur profane.

	GUIDES ALIMENTAIRES	RECETTES
Nombre de textes	42 textes	23 121 textes
Nombre de mots	471 463	3 928 658
Longueur des phrases (moyenne)	19,15 mots	8,97 mots
Langue	FR	FR
Contenus	Bonnes pratiques, conseils médicaux, réglementations, comportements des professionnels de santé, bien être, etc.	Recettes de cuisine : titre + ingrédients + préparation
Nombre de lemmes	NOMS = 5 835 TERMES = 20 210	NOMS = 7 066 TERMES = 38 215

TABLEAU 4.2 : Corpus de grande taille : corpus alimentaires.

Nous pouvons tout d'abord constater, à la lecture du tableau 4.2, que le nombre de lemmes différents de chaque corpus n'est pas proportionnel à sa taille. En effet, le corpus Recettes est le plus volumineux, avec presque 4 millions de mots, mais la taille de son vocabulaire est à peine plus élevée que celle du corpus Guides Alimentaires, avec 7 066 noms pour le premier contre 5 835 noms pour le second. Ainsi, nous pouvons en déduire des fréquences plus élevées pour les mots du corpus Recettes, mais une diversité des contextes plus faible. Par contre, en ce qui concerne la longueur des phrases, le corpus Recettes se caractérise par des phrases en moyenne beaucoup moins longues que le corpus Guides Alimentaires, avec une moyenne d'environ 9 mots par phrase pour le premier contre environ 19 mots pour le second. Cette longueur moyenne moins élevée provient en partie de la présence en début de texte de la liste des ingrédients, dans laquelle chaque ingrédient est considéré par les outils de TAL que nous avons utilisés comme une phrase.

Ainsi, nous pouvons déduire de toutes ces informations que le problème de dispersion des données ne sera pas le même pour les deux corpus. Pour le corpus Recettes, les faibles fréquences sont certainement dues à la longueur des phrases : des phrases plus courtes permettront *a priori* de disposer de moins de contextes pour chaque mot cible, et causeront ainsi une faible diversité des contextes. Tandis que pour le corpus Guides Alimentaires, les faibles fréquences sont liées à une plus petite taille de vocabulaire. Nous revenons plus en détail sur le contenu de ces deux corpus dans les deux sous-sections qui leur sont consacrées ci-dessous.

4.1.2.1 Corpus de recettes de cuisine (*Recettes*)

Nous avons utilisé l'ensemble des textes fournis par le challenge DEFT 2013 [1], soit un total de 3 928 658 mots. Ce corpus est le plus volumineux des corpus que nous utilisons et il est composé de recettes de cuisine en langue française. Ces recettes ont été extraites d'un site de cuisine et sont rédigées par des internautes. A l'instar de la figure 4.3 (une recette extraite de notre corpus), chaque texte est composé du titre de la recette, du coût, de la liste d'ingrédients et du corps de la recette. Ces parties apparaissent dans le même ordre pour tous les textes du corpus. En revanche, aucun titre de section ne structure les documents, et la ponctuation n'est pas nécessairement présente.

Tarte au potiron et groseille
Bon marché
50 g de pâte sablée
80 g de beurre
350 g de chair de potiron
5 œufs
300 g fromage blanc à 30%
100 g de crème liquide
100 g de sucre cassonade
2 c à soupe de gelée de groseille
2 pincées de cannelle
Faire fondre le beurre dans une casserole - faites étuver 25 mn feu moyen le potiron coupé en dés jusqu'à ce qu'il soit fondant et laisser tiédir.
Etaler la pâte sablée dans un moule à tarte, piquez le fond et étalez la gelée de groseille.
Fouettez le fromage blanc, la crème liquide, la cassonade, les œufs entiers, saupoudrez de cannelle.
Mettre le potiron égoutté sur la pâte et versez le mélange fromage blanc et œufs.
Cuire à four thermostat 7 pendant 40 mn puis à thermostat 5/6 pendant 25 mn.
muscat (jus de fruits ou grenadine pour les enfants)

FIGURE 4.3 : Exemple d'une recette de cuisine issue du *Corpus Recettes*.

Les textes du corpus Recettes sont rédigés par des internautes à destination d'autres internautes. Ils ne sont pas pour autant dépourvus de spécificité, et se caractérisent par un lexique spécialisé (ingrédients, ustensiles, dosages, etc.) et une structure de phrase propre aux recettes de cuisine. En effet, les recettes, outre l'énumération des ingrédients, sont des textes procéduraux, des textes instructionnels. Ils contiennent généralement des phrases injonctives commençant par un verbe à l'infinitif ou à l'impératif présent,

1. http ://deft.limsi.fr/2013/

comme par exemple la première phrase de la recette 4.3, *Faire fondre le beurre dans une casserole.*

4.1.2.2 Corpus de guides alimentaires (*Guides Alimentaires*)

Le second corpus du domaine de l'alimentation que nous utilisons, également en langue française, est composé de guides alimentaires collectés par nos soins à travers les requêtes Google suivantes : *guide alimentaire*, *guide de nutrition*, *guide nutritionnel* et *guide d'alimentation* (cf. Annexe A). Ce corpus est de plus petite taille que le précédent, avec 21 documents et 471 463 mots.

Nous avons sélectionné les textes provenant de deux origines principales. Un premier ensemble regroupe des textes rédigés par des instances réglementaires qui évoquent les réglementations, règles ou consortiums autour de l'alimentation, aussi bien au niveau régional, gouvernemental ou national (en France et à l'étranger), que dans les cabinets médicaux. Le deuxième ensemble regroupe des guides de bonnes pratiques (aussi bien alimentaires que dans l'hygiène de vie), de conseils médicaux, en lien avec la prévention de pathologies telles que le diabète et l'obésité, mais toujours en lien avec l'alimentation.

4.2. Malabsorption, malnutrition, allergie alimentaire avec manifestations digestives
Les diètes semi-élémentaires (DSE) sont composées de nutriments choisis en raison d'une absorption intestinale aisée : les protéines ont subi une hydrolyse poussée ; les glucides se trouvent sous forme de polymères simples du glucose (dextrines maltose) ; enfin, les lipides consistent en LCT (Triglycérides à longue chaîne) et MCT souvent donnés en association. Leur osmolarité n'excède pas 300mOsm/l.
L'usage des DSE devrait être réservé aux indications suivantes :
- l'allergie aux protéines de lait de vache avec manifestations digestives,
- les diarrhées rebelles (dites intraitables) ou prolongées (> 15 jours),
- les syndromes de malabsorption et situations apparentées (grêle court, atrophie de la muqueuse intestinale...),
- les syndromes inflammatoires du tube digestif (maladie de Crohn et rectocolite ulcéro-hémorragique),
- le régime strict sans déchet (chez le nourrisson notamment),
- les états cataboliques nécessitant un apport calorique élevé (brûlés...).

FIGURE 4.4 : Extrait d'un guide pour les professionnels, *Corpus Guides Alimentaires.*

Nous présentons un extrait dans la figure 4.4. Les phrases du corpus sont plus longues que pour le corpus Recettes et la rédaction est plus standard ; la syntaxe suit le schéma

sujet - verbe - objet. Le degré de technicité est également plus élevé que le corpus Recettes, et les Guides Alimentaires contiennent un nombre important de termes, autour de la nutrition (ex : *malnutrition*, *absorption intestinale*, *diarrhée*, *diabète*),

4.1.3 Pré-traitement des corpus

Les corpus sont analysés à travers la plateforme de TAL Ogmios [Hamon et Nazarenko, 2008]. Cette plateforme permet de réaliser une analyse linguistique des textes, en articulant différents outils de TAL. Le schéma 4.5 présente la configuration que nous avons utilisée. Après le passage des corpus au format XML d'entrée de la plateforme, nous avons configuré le pré-traitement de manière à ce que cette analyse comprenne un étiquetage morphosyntaxique et une lemmatisation du corpus, à l'aide de TreeTagger [Schmid, 1994], et une extraction de termes à l'aide de YaTeA [Aubin et Hamon, 2006].

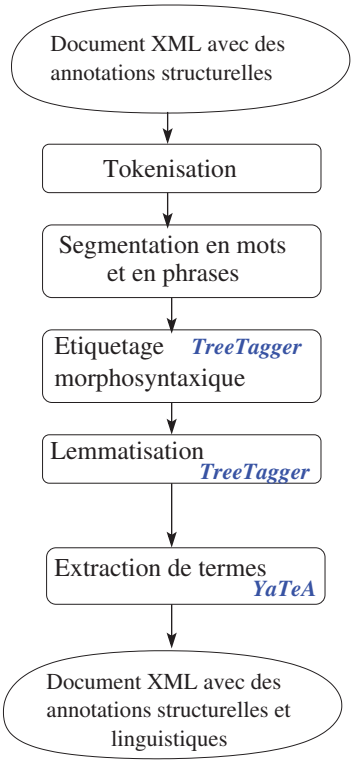

FIGURE 4.5 : Pré-traitement des corpus.

L'étiquetage morphosyntaxique attribue des catégories morphosyntaxiques aux mots. Celles-ci sont utiles à la fois lors de la sélection des mots cibles et lors de la définition des contextes distributionnels. L'identification des noms nous permet d'extraire les termes simples du corpus, utiles en tant que mots cibles et contextes.

L'extraction de termes nous permet d'extraire et d'identifier les termes complexes de nos corpus. Ceux-ci sont ensuite exploités pour l'acquisition des relations sémantiques entre des termes de longueurs différentes présents au sein de nos corpus.

Ces informations linguistiques permettent ensuite, lors de la construction de la matrice de contextes, d'obtenir une matrice linguistiquement riche. Nous avons discuté l'apport de ces informations dans la section 2.1.

4.2 Evaluation

Dans cette section, nous abordons la méthode utilisée pour évaluer les relations acquises par l'analyse distributionnelle avec et sans l'abstraction des contextes. Ainsi, nous présentons l'évaluation menée et les métriques que nous avons utilisées.

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.*, 2013a]. 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].

Les principales approches d'évaluation, peuvent être classées selon l'opposition méthodes extrinsèques ou intrinsèques [Curran, 2004, Poibeau et Messiant, 2008]. Les méthodes extrinsèques, ou évaluation par la tâche, évaluent l'impact de ces ressources dans la performance des systèmes automatiques. Et les méthodes intrinsèques évaluent les relations produites de manière directe, c'est-à-dire du point de vue de leur objectif propre. Pour cela, il est alors possible d'utiliser des ressources existantes ou de mener une évaluation manuelle des relations distributionnelles données. Nous évaluons la qualité des résultats obtenus lors de nos expériences en comparant les relations sémantiques acquises à des ressources existantes, essentiellement pour le coût et la complexité que représentent l'évaluation par la tâche et l'évaluation manuelle. Nous sommes bien consciente des limites de cette évaluation, notamment le fait qu'une telle évaluation est partielle, puisqu'elle permet d'évaluer la méthode uniquement à travers les relations dites classiques qui sont contenues dans les ressources disponibles. De plus, la couverture de la ressource par rapport au corpus est généralement insuffisante [Bodenreider *et al.*, 2002], parce que les résultats produits par des méthodes automatiques d'acquisition de relations reflètent les usages spécifiques aux corpus, et ainsi ne correspondent pas nécessairement aux relations présentes dans les ressources existantes. Ainsi, cette évaluation ne valorise pas l'identification de relations entre les nouveaux termes apparaissant dans un corpus spécialisé, ainsi que l'identification de variantes terminologiques.

Malgré tout, pour évaluer la similarité sémantique, une évaluation intrinsèque est généralement réalisée, en prenant comme référence une ressource, dictionnaire ou thesaurus,

ou en déterminant le degré de corrélation entre les similarités acquises automatiquement et un ensemble de jugements de similarité réalisés par des humains [Ferret, 2014]. Ainsi, l'évaluation par la tâche rendrait difficile l'évaluation des paramètres distributionnels, en raison de la difficulté à démontrer de manière significative leur impact dans la tâche en question. Pour cela, il faut que les paramètres distributionnels y occupent un rôle important, comme par exemple la mesure de similarité dans une tâche d'enrichissement de ressources [Ferret, 2014, Pantel *et al.*, 2009].

À l'instar de [Curran, 2004] et [Ferret, 2013b], 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. L'évaluation des résultats est réalisée avec des métriques d'évaluation utilisées habituellement sur les résultats d'une analyse distributionnelle (macro-précision, moyenne des précisions moyennes - MAP, et R-précision). Nous présentons dans un premier temps les ressources utilisées (sous-section 4.2.1), puis les métriques d'évaluation (sous-section 4.2.2).

4.2.1 Ressources

Afin d'évaluer les relations acquises par notre méthode sur nos corpus de travail, nous avons constitué des références propres à chaque domaine et à chaque corpus, à partir de ressources propres aux domaines.

Afin de calculer la R-précision (cf. section 4.2.2.2), et de manière à permettre une évaluation précise, nous considérons dans la ressource uniquement les relations que notre méthode doit potentiellement retrouver, c'est-à-dire les relations dont les deux éléments sont des mots ou des termes du corpus. Ainsi, nous avons élaboré des références à partir des ressources citées ci-dessous, une pour chaque corpus. L'appariement entre les termes du corpus et les termes contenus dans les ressources est réalisé après lemmatisation de tous les termes.

Nous présentons les ressources utilisées pour l'évaluation des résultats obtenus sur les corpus du domaine médical, puis pour l'évaluation des résultats obtenus avec les corpus du domaine de l'alimentation.

4.2.1.1 Domaine médical

Pour les corpus médicaux, nous utilisons la ressource UMLS [2]. Pour l'évaluation des résultats obtenus avec le corpus Menelas, nous exploitons la partie française de l'UMLS [UMLS], 1 735 419 relations, soit 2 434 relations entre les termes du corpus. Les relations contenues dans la ressource sont majoritairement des co-hyponymes (1 536 relations), mais également des hyperonymes (333), des relations équivalentes parmi lesquelles on

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

	UMLS-FR (Menelas)	UMLS-EN (Textes Cliniques)
Co-hyponymie	artère fémorale - artère brachiale bruits du cœur - débit cardiaque thrombose - embolie	alcohol - smoking atrial pressure - heart rate anxiety - diarrhea
Hyperonymie	artère - aorte examen physique - palpation radiographie - angiographie	activity - exercise antibiotic - ciprofloxacin imaging studies - echocardiogram neurological symptoms - dizziness
Synonymie	angor - angine de poitrine fatigue - épuisement infarctus du myocarde - im	medicine - drug therapy - therapeutic procedure orthopedic - orthopaedics
Rel. du domaine	cancer - tumeur angiocardiographie - cœur souffle cardiaque - bruit du cœur	wound infection - diagnosis vitamin e - blood ovary - secretion

TABLEAU 4.3: Exemples de relations contenues dans les références d'évaluation des corpus médicaux (l'UMLS-FR et l'UMLS-EN).

trouve des synonymes (438), et des relations du domaine (128). Des exemples de ces relations sont donnés dans le tableau 4.3.

Pour les résultats obtenus avec le corpus de textes cliniques, nous avons recours à la partie anglaise de l'UMLS qui contient 40 036 781 relations, soit 53 203 relations entre les termes du corpus. Les types de relations entre les termes du corpus contenues dans la version anglaise de l'UMLS sont une majorité de co-hyponymes (22 680 relations) et de relations du domaine (22 939), des hyperonymes (6 505) et des synonymes (1 079). Des exemples de ces relations sont également données dans le tableau 4.3.

4.2.1.2 Domaine alimentaire

Pour les corpus alimentaires, nous utilisons deux types de ressources : deux ressources terminologiques existantes et une ressource que nous avons construite à partir de sites Web.
Les deux ressources existantes sont la partie française d'Agrovoc [3] [AGRO], 75 222 relations, et la partie française de l'UMLS [4] [UMLS], 1 735 419 relations.

AGROVOC
Les références construites à partir d'Agrovoc contiennent 1 574 relations entre les termes du corpus Recettes et 2 935 relations entre les termes du corpus Guides Alimentaires. Ces relations sont des hyperonymes, des co-hyponymes, des synonymes, et des relations

3. http ://aims.fao.org/standards/agrovoc/about
4. http ://www.nlm.nih.gov/research/umls/

du domaine (cf. tableau 4.4). Pour le corpus Recettes, la référence contient une majorité d'hyperonymes, et pour le corpus Guides Alimentaires une majorité de relations du domaine.

Hyperonymes	agrume - clémentine trouble alimentaire - boulimie volaille - canard récipient - bouteille	Relations du domaine	sucre - miel bouchon - fermeture commerce - achat cuisson - four
Co-hyponymes	réfrigération - congélation cerf - chevreuil endive - salade	Synonymes	sorbet - crème glacée soupe - potage valeur calorique - valeur énergétique

Tableau 4.4: Exemples de relations présentes dans la ressource Agrovoc.

UMLS-FR

Les références construites à partir de l'UMLS contiennent 2 102 relations entre les termes du corpus Recettes et 2 832 relations pour les Guides Alimentaires. Ces relations sont des hyperonymes, des co-hyponymes et des relations du domaine pour les deux corpus (cf. tableau 4.5). Pour le corpus Guides Alimentaires, la référence contient également des synonymes.

Hyperonymes	condiment - épice émotion - peur	Synonymes	agitation - excitation gaz - flatulence malaria - paludisme
Co-hyponymes	apathie - ennui boxe - lutte vomissement - anorexie aidant - pharmacien	Relations du domaine	champignon - microbiologie dent - mastication graphie - tableau laryngectomie - larinx

Tableau 4.5: Exemples de relations présentes dans la ressource UMLS, pour les corpus alimentaires.

Remarques

La raison pour laquelle nous avons choisi de travailler avec ces deux types de ressources repose en premier lieu sur un critère de disponibilité. En ce qui concerne le corpus Recettes, la comparaison avec Agrovoc est justifiée par la présence de relations entre des termes liés à l'alimentation présentes dans les deux ressources. En effet, dans les recettes de cuisine nous estimons que d'autres types de relations peuvent être identifiés, comme la relation entre un terme de nutrition et un terme appartenant à une autre classe sémantique. Pour le corpus Guides Alimentaires, la comparaison avec Agrovoc se justifie de la même manière que pour le corpus Recettes.

Ainsi, la comparaison avec l'UMLS en français est justifiée par la présence de termes médicaux dans le corpus Guides Alimentaires. Même si nous supposons un faible recouvrement entre ces ressources et nos corpus, la comparaison de nos résultats et des relations extraites de ces ressources permet de donner une indication de la contribution de chaque modèle d'abstraction proposé.

Hyperonymes	fruit de mer - pétoncle poisson blanc - merlan viande rouge - kangourou
Méronymes	épinards - vitamine C ananas - coupe de fruits
Co-hyponymes	crèpe nature - gaufre chantilly farine de sarrasin - semoule blanche persil - romarin
Synonymes	ran -bulot pain de son - pain noir pieuvre - poulpe

TABLEAU 4.6 : Exemples de relations présentes dans la ressource issue du Web.

Relations issues du Web

Etant donné qu'Agrovoc et l'UMLS ne sont pas des ressources spécialement dédiées à l'alimentaire ou à la nutrition, ni pour des spécialistes, nous avons choisi de construire une ressource que nous considérons plus adaptée à nos corpus. Cette ressource contient essentiellement des relations entre des noms d'aliments, classes d'aliments et composants alimentaires. Il s'agit de 5 058 relations issues de quatre sites Web.

- Une société fournissant une méthode pour perdre du poids [5],
- Le site Web de Health Canada (Santé Canada), le département du gouvernement du Canada en charge de la santé pulique nationale [6],
- Un centre spécialisé dans la perte de cheveux [7],
- Un site Web fournissant des recettes de cuisine [8].

Nous avons typé manuellement les relations, et la ressource contient 1 570 hyperonymes, 2 809 co-hyponymes, 583 méronymes, et quelques 71 variantes morphosyntaxiques et 25 synonymes (cf. tableau. 4.6).

Nous constituons ensuite à partir de ces 3 ressources, l'UMLS, Agrovoc et la Ressource Adaptée, une seule ressource globale, de manière à couvrir le plus possible l'ensemble des termes des corpus. De même que pour les corpus médicaux, nous définissons une

5. http ://www.bioweight.com/
6. http ://www.hc-sc.gc.ca/fn-an/securit/addit/diction/index-fra.php
7. http ://www.centre-clauderer.com/acides-bases/femme-2.htm
8. http ://www.cuisine-libre.fr/

référence pour chaque corpus. Les références contiennent 3 701 relations entre les termes du corpus Recettes, et 1 825 relations entre les termes du corpus Guides Alimentaires.

4.2.1.3 Bilan

	Ressource	Relations entre mots du corpus
Médical	UMLS-FR	2 434 relations
	UMLS-EN	53 203 relations
Alimentaire	Globale-Recettes	6 015 relations
	Globale-Guides Alim	8 095 relations

TABLEAU 4.7 : Nombre de relations entre les termes du corpus par référence.

Pour l'évaluation de notre méthode, nous avons constitué plusieurs références, adaptées à nos corpus de travail, à partir des ressources précédemment décrites. Nous récapitulons dans le tableau 4.7 les références que nous utilisons pour l'évaluation. Pour les corpus médicaux, ces références sont construites à partir de l'UMLS français pour Menelas, et de l'UMLS anglais pour les Textes Cliniques. Pour les corpus alimentaires, les plus gros corpus, nous avons constituté une référence globale à partir de l'UMLS, d'Agrovoc et d'une ressource que nous avons constituée à partir de données issues du Web.

4.2.2 Métriques d'évaluation

Afin d'évaluer les relations acquises par notre méthode, nous nous comparons aux ressources que nous venons de décrire. Nous utilisons les métriques d'évaluation 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 [9], mis au point lors des campagnes TREC.

4.2.2.1 Macro-précision

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

9. http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz

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

La macro-précision pour l'ensemble des mots cibles est alors : $P = \frac{\sum_{k=1}^{|w_i|} p(w_k)}{|w_i|}$

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) :

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

Afin de mieux expliquer ce calcul, nous reprenons dans la figure 4.6 la représentation faite par Chirag Shah, dans son cours sur les systèmes de recherche d'information [10]. Ainsi, pour chaque mot cible, les voisins sont classés par mesure de similarité, allant de la mesure la plus élevée, à gauche, à la plus faible, à droite. Les voisins pertinents sont colorés. La précision à 1 (P@1), prend donc en compte uniquement le premier voisin, et est égale à 1, la précision à 5 (P@5) prend en compte les cinq premiers voisins, et est égale à 2/5, et ainsi de suite.

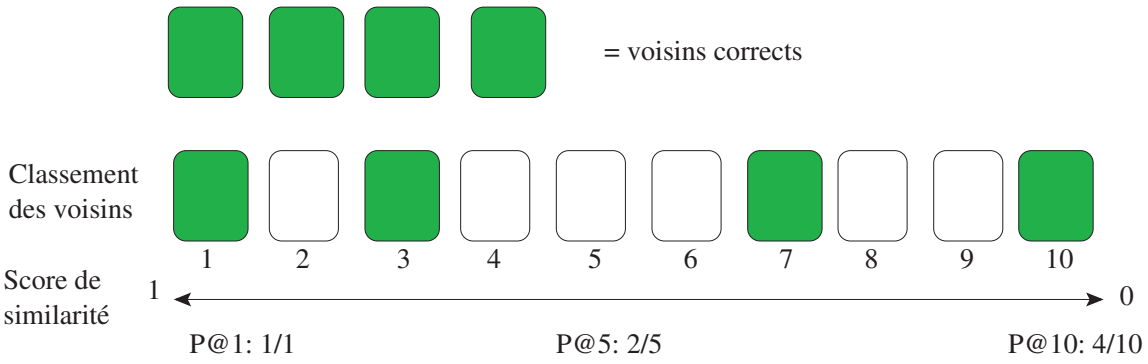

FIGURE 4.6 : Calcul de la macro-précision.

La macro-précision permet d'obtenir une qualité globale des résultats en considérant que tous les mots cibles ont le même poids quel que soit le nombre de voisins. En revanche, la micro-précision a tendance à privilégier les mots-cibles comportant beaucoup de voisins, dont une bonne partie ne sont probablement pas pertinents, au détriment de mots-cibles ayant peu de voisins. Pour P@1, la macro-précision est équivalente à la micro-précision.

10. http://www.inforetrieval.org/2009_fall/inls490_154w/lessons/lesson2_notes.pdf

4.2.2.2 R-précision

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 attendus pour un mot cible, n_i variant alors suivant les mots cibles. La mesure est ainsi plus équitable qu'une précision à un 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.

Pour le calcul de la R-précision, nous comparons nos résultats non plus à l'ensemble des relations contenues dans les ressources, mais à des ensembles de référence constitués à partir de ces ressources. Il s'agit de réduire les relations de référence aux seules relations entre des termes présents dans le corpus de travail et dans chaque expérience. Ainsi, la mesure s'appuie sur les relations présentes dans les références, et le seuil de précision dépend du vocabulaire du corpus.

4.2.2.3 Moyenne des précisions moyennes (Mean Average Precision : MAP)

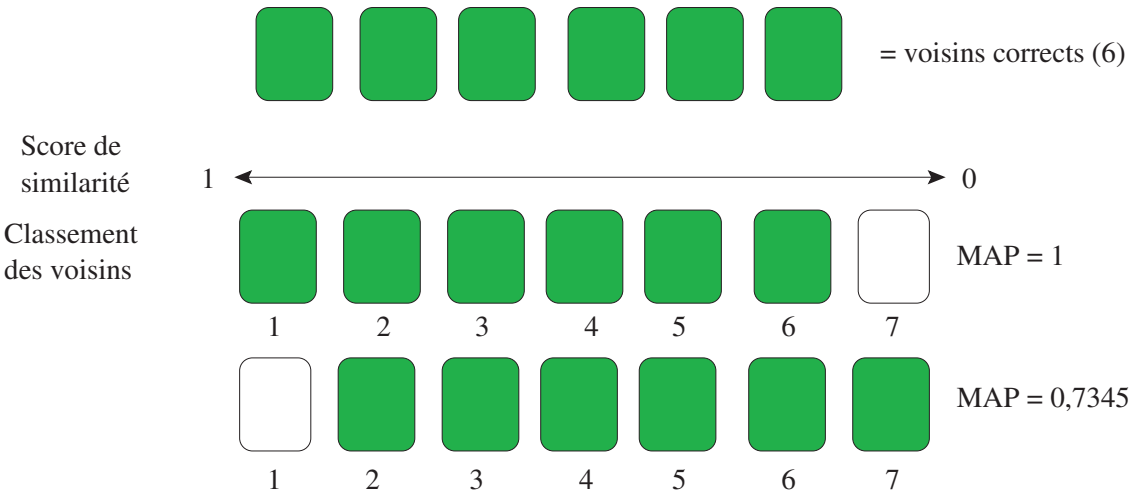

FIGURE 4.7 : Calcul de la MAP.

Nous évaluons également les résultats à l'aide de la Mean Average Precision (MAP) :

La moyenne des précisions moyennes (MAP) est obtenue en considérant la précision non interpolée $UAP(I_i^j)$ des voisins sémantiques I_i^j au rang j, n_i est 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 est le reflet de la qualité du classement et permet d'évaluer la pertinence de la mesure de similarité utilisée. Ainsi, 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. La R-précision, en revanche, se limite aux n voisins corrects attendus.

Par exemple, pour un mot cible ayant six voisins sémantiques dans la référence, si la méthode trouve ces six voisins mais ajoute un voisin incorrect, elle obtiendra un score entre 0,7345 et 1, selon la place occupée par le voisin incorrect entre la première et la septième position (cf. figure 4.7). Réciproquement, avec la même référence, si la méthode ordonne cinq des six voisins en tête de liste et ne fournit pas le sixième voisin, elle obtient une MAP de 0,8333.

Trouver le sixième voisin et l'ordonner après dix voisins bruités augmente le score à 0,8958. Comme mentionné ci-dessus, ajouter des voisins erronés en fin de liste ne change pas le score.
Si la méthode ordonne les six voisins de la référence après dix voisins bruités, le score obtenu est de 0,2471. Enfin, si la méthode n'ordonne que trois des six voisins de la référence en tête de liste et aucun des trois voisins manquants, le score obtenu est de 0,5, car les voisins manquants sont considérés comme obtenant une précision de zéro (la métrique considère qu'ils sont ordonnés à une place infiniment loin dans le classement).

4.3 Conclusion

Dans ce chapitre, nous avons présenté les quatre corpus de spécialité sur lesquels nous évaluons notre méthode d'abstraction des contextes distributionnels. Nous évaluons d'une part la méthode sur de petits corpus, qui sont deux corpus médicaux, et d'autre part sur des corpus de plus grande taille, deux corpus alimentaires.

Nous avons choisi de comparer les relations produites par notre méthode aux relations contenues dans des références construites à partir de ressources existantes ou construites à partir du Web. Pour l'évaluation, nous avons défini les trois métriques utilisées : la macro-précision, la R-précision et la MAP.

Chapitre 5

Expériences et résultats

Dans ce chapitre, nous présentons les expériences que nous avons menées et l'analyse des résultats obtenus. Notre objectif est d'évaluer la méthode d'abstraction des contextes que nous avons mise au point. En préliminaire à cette évaluation et afin de mettre en œuvre une méthode distributionnelle adaptée aux corpus de spécialité, nous avons réalisé un ensemble d'expériences afin d'adapter les valeurs des paramètres distributionnels. Ce premier ensemble d'expériences nous permet d'établir une base d'expérimentation pour l'abstraction des contextes distributionnels.

Une fois les paramètres distributionnels établis, nous avons évalué et caractérisé l'impact de la normalisation et de la généralisation des contextes sur nos quatre corpus de travail (présentés en section 4.1), avec pour objectif la réduction de la dispersion des données. Pour cela, nous avons utilisé les règles proposées dans le chapitre 3. Celles-ci s'appuient sur des relations sémantiques acquises automatiquement (cf. section 3.2) pour généraliser et normaliser les contextes séparément et de manière combinée. Ainsi, nous avons réalisé différentes expériences autour de l'abstraction des contextes, dans lesquelles nous faisons varier les relations utilisées pour l'abstraction (en fonction du type de méthode d'acquisition, et donc du type de relations) et leur ordre d'utilisation pour la substitution des mots dans les contextes distributionnels.

Ces expériences sont menées sur nos quatre corpus de travail (cf. section 4.1). Nous présentons dans un premier temps les expériences et les résultats obtenus pour la définition de l'analyse distributionnelle et ensuite ceux obtenus avec l'abstraction des contextes. Comme nous l'avons présenté dans la section 4.2, 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.

Enfin, nous comparons les résultats obtenus avec la meilleure configuration d'abstraction des contextes aux résultats obtenus sur nos corpus par une méthode de réseaux de neurones. Pour ce faire, nous utilisons l'outil Word2vec [Mikolov *et al.*, 2013].

5.1 Définition de paramètres distributionnels adaptés aux textes de spécialité

La première étape de nos expériences concerne la mise en œuvre de notre méthode distributionnelle à travers la définition de paramètres adaptés aux textes de spécialité. Ceci nous permet de disposer d'une base de comparaison juste, pour pouvoir ensuite

évaluer au mieux les différentes expériences menées autour de l'abstraction des contextes distributionnels.

Pour cela, nous avons analysé le comportement de cinq paramètres pouvant influer sur l'analyse distributionnelle en fonction des valeurs prises par ces paramètres : deux paramètres au niveau de la sélection des contextes (taille de la fenêtre graphique et sélection des contextes les plus discriminants) ainsi que trois paramètres autour du calcul de la similarité sémantique (pondération des contextes, mesure de similarité et seuil sur les contextes et les mots cibles). Ces paramètres ont été définis dans la section 3.1.

Nous présentons dans un premier temps les différentes expériences, et nous analysons ensuite le comportement des paramètres distributionnels, de manière à définir les configurations que nous utiliserons par la suite.

L'objectif de cette première série d'expériences est de choisir la meilleure configuration possible pour l'abstraction des contextes. Pour cela, nous avons défini deux critères de décision :
- minimisation du nombre de relations retournées : il s'agit de valoriser les paramètres qui permettent de produire le moins de relations possible ;
- maximisation des mesures d'évaluation : nous prenons alors en compte les paramètres qui offrent les meilleurs résultats qualitatifs, en termes de précision, R-précision et MAP.

Ainsi, dans cette section, nous présentons dans un premier temps les expériences réalisées pour la définition des paramètres distributionnels, puis les résultats obtenus pour chaque paramètre.

5.1.1 Expériences

Nous réalisons un ensemble d'expériences visant la définition des paramètres distributionnels. Les valeurs des paramètres que nous évaluons sont récapitulées dans le tableau 5.1.

Mots cibles

Nous nous sommes tout d'abord interrogée sur les mots à mettre en relation, c'est-à-dire le choix des mots cibles. Etant donné que nous travaillons à partir de textes de spécialité, nous souhaitons identifier des relations entre termes. Ces termes sont identifiés automatiquement et sont des termes simples et complexes. Nous expérimentons deux ensembles de mots cibles : les termes complexes uniquement, et l'ensemble composé de termes simples, des termes complexes et des noms.

Paramètres	Valeur(s)
MOTS CIBLES	1. Termes complexes 2. Termes simples + termes complexes + noms
CONTEXTES	1. Fenêtre graphique large (21 mots) 2. Fenêtre graphique restreinte (5 mots)
SELECTION DES CONTEXTES	Sélection des contextes les plus discriminants (Cf-Itf)
MESURES DE PONDERATION	1. Information Mutuelle 2. Fréquence Relative
MESURES DE SIMILARITE	1. Nombre de Contextes Partagés 2. Fréquence des Contextes Partagés 3. Indice de Jaccard 4. Cosinus
SEUILS	Combinaison de 3 seuils sur les mots cibles et les contextes (Nombre et Fréquence des contextes partagés + fréquence des mots cibles)

TABLEAU 5.1 : Récapitulatif des paramètres évalués.

Contextes

En ce qui concerne la définition du contexte, nous avons choisi d'avoir recours aux fenêtres graphiques, c'est-à-dire un nombre de mots situés dans les contextes droite et gauche du mot cible (cf. section 2.1.2). Nous testons deux tailles de fenêtres : une fenêtre large de 21 mots ($\pm$10 mots, centrée sur le mot cible) et une fenêtre restreinte de 5 mots ($\pm$2 mots, centrée sur le mot cible).
La fenêtre la plus large est définie de manière à prendre en compte le plus grand nombre de contextes possibles. Nous l'avons ainsi définie à 21 mots, car dans nos corpus de travail, nous avons au maximum 19,15 mots en moyenne par phrase. De plus, nous conservons dans les contextes uniquement les mots qui sont étiquetés comme des verbes, des noms ou des adjectifs, ainsi que les termes identifiés automatiquement.
En ce qui concerne la fenêtre restreinte, nous avons choisi d'utiliser une taille de 5 mots de part et d'autre du mot cible, car il s'agit d'une taille adaptée aux textes de spécialité [Généreux et Hamon, 2013, Rapp, 2003].

Nous n'évaluons pas la taille de la fenêtre indépendamment, mais en combinaison à d'autres paramètres.

Mesures de similarité et de pondération

Pour chaque taille de fenêtre et chaque ensemble de mots cibles, nous testons les quatre mesures de similarité décrites dans la section 3.1.3, c'est-à-dire, l'indice de Jaccard (Jacc), le Cosinus (Cos), la Fréquence des Contextes Partagés (FreqCtxt) et le Nombre de Contextes Partagés (NbCtxt). Nous testons ces mesures utilisées seules, et pour les deux mesures acceptant une pondération (i.e. Cosinus et Jaccard), nous évaluons l'apport de deux mesures de pondération ; la fréquence relative pour l'indice de Jaccard (Jacc-Freq), et l'Information Mutuelle pour la mesure du Cosinus (Cos-IM).

Seuils

Afin de limiter le nombre de relations proposées et d'écarter les relations potentiellement fausses, nous avons défini plusieurs seuils (cf. section 3.1.3) : nous utilisons la combinaison de trois seuils sur les mots cibles et contextes. Ces seuils sont calculés automatiquement et correspondent à la moyenne des valeurs prises par chaque paramètre sur l'ensemble du corpus. Nous évaluons également l'apport de ces seuils.

Sélection des contextes

Enfin, nous évaluons l'apport de la suppression des contextes les moins discriminants à l'aide du Cf-Itf utilisé en amont du calcul de similarité (cf. section 3.1).

Dans la suite de cette section, nous décrivons les résultats obtenus pour les mesures de similarité et de pondération (section 5.1.2), avec et sans l'application des différents seuils (section 5.1.3) et enfin avec et sans sélection des contextes les plus discriminants (section 5.1.3.3).

5.1.2 Mesures de similarité et de pondération

Nous étudions à présent les mesures de similarité et de pondération, décrites en section 3.1.3, à travers l'analyse des résultats obtenus avec les seuils sur les mots cibles et les contextes. En effet, l'utilisation des seuils sur les mots cibles et les contextes ne faisant pas varier le comportement des mesures de similarité par rapport aux résultats, nous avons choisi de présenter et de discuter ici les résultats obtenus avec l'utilisation des seuils.

Nous procédons à l'analyse des résultats en fonction de la taille de nos corpus de travail. Dans un premier temps, nous décrivons les résultats obtenus pour les corpus de petite taille, les deux corpus médicaux, et ensuite nous détaillerons le comportement des mesures de similarité et des pondérations avec les corpus de grande taille.

Indépendamment de la taille des corpus, nous avons observé que le Cosinus utilisé sans pondération obtient dans toutes les configurations des résultats similaires ou légèrement inférieurs à ceux obtenus par le Cosinus pondéré avec l'Information Mutuelle. Nous présentons donc uniquement ces derniers résultats.

5.1.2.1 Corpus de petite taille

Nous présentons dans un premier temps les résultats obtenus avec les corpus de petite taille, c'est-à-dire le corpus Menelas et les Textes Cliniques. Pour les deux corpus, la fenêtre restreinte offre de meilleurs résultats, aussi bien avec les termes complexes (TC) que pour la combinaison des termes simples et des termes complexes (TS+TC).

Les résultats obtenus avec les seuils sur les mots cibles et les contextes, pour les corpus de petite taille, sont présentés dans le tableau 5.2 pour la fenêtre restreinte (5 mots) et dans le tableau 5.3 pour la fenêtre large (21 mots).

Fenêtre restreinte (5 mots)

Pour les deux corpus, quand les termes complexes sont utilisés comme mots cibles, la mesure de similarité la plus adaptée est l'indice de Jaccard pondéré (Jacc-Freq), quelle que soit la métrique d'évaluation, dont notamment une MAP égale à 0,119 pour le corpus Menelas et de 0,052 pour les Textes Cliniques.

Quand les mots cibles sont les termes simples et complexes, pour le corpus Menelas, les résultats obtenus avec NbCtxt, FreqCtxt et JaccFreq sont assez proches, voire identiques (la R-précision est de 0,010 avec ces trois mesures). Malgré tout, l'indice de Jaccard pondéré obtient généralement de meilleures valeurs avec la MAP (0,08), la P@5 (0,029) et la P@10 (0,029). Pour les Textes Cliniques et les termes simples et complexes en mots cibles, l'indice de Jaccard pondéré obtient également les valeurs les plus élevées avec toutes les métriques d'évaluation.

Le Cosinus semble peu adapté aux petites fréquences, et obtient dans l'ensemble des valeurs nulles ou proches de zéro, avec les plus faibles valeurs quand les termes complexes sont les mots cibles. Il est donc préférable de ne pas utiliser le Cosinus si l'on fait le choix d'utiliser la fenêtre graphique restreinte. Malgré tout, le nombre de relations acquises et le nombre de relations retrouvées dans la référence sont identiques pour toutes les mesures de similarité.

Fenêtre large (21 mots)

Les résultats obtenus pour les petits corpus avec la fenêtre large sont présentés dans le tableau 5.3.

		Menelas				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq [1]	Ts+TC	274 447	274 447	274 447	274 447	274 447
	TC	8 118	8 118	8 118	8 118	8 118
Rel dans Ref [2]	Ts+TC	60	60	60	60	60
	TC	96	96	96	96	96
Rprec	TS+TC	0,066	**0,105**	0,053	0,079	0,222
	TC	0	0,045	0,036	**0,080**	0,036
MAP	TS+TC	0,086	**0,171**	0,159	0,121	0,188
	TC	0,014	0,091	0,066	**0,119**	0,084
P@1	TS+TC	0,053	0,105	0	0,053	**0,158**
	TC	0	0,036	0,036	**0,107**	0,036
P@5	TS+TC	0,021	0,042	0,042	0,042	**0,053**
	TC	0,007	0,029	0,029	**0,029**	0,021
P@10	TS+TC	0,016	0,026	0,037	**0,037**	0,037
	TC	0,004	0,021	0,018	**0,025**	0,025
		Textes Cliniques				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq	TS+TC	2 722 289	2 722 289	2 722 289	2 722 289	2 722 289
	TC	1 696 871	1 696 871	1 696 871	1 696 871	1 696 871
Rel dans Ref	TS+TC	13 372	13 372	13 372	13 372	13 372
	TC	11 896	11 896	11 896	11 896	11 896
Rprec	TS+TC	0,001	0,027	0,022	**0,034**	0,028
	TC	0,001	0,039	0,026	**0,046**	0,037
MAP	TS+TC	0,005	0,035	0,027	**0,045**	0,035
	TC	0,007	0,046	0,031	**0,052**	0,046
P@1	TS+TC	0	0,017	0,030	**0,054**	0,017
	TC	0	0,018	0,035	**0,061**	0,020
P@5	TS+TC	0,001	0,035	0,024	**0,041**	0,038
	TC	0	0,046	0,032	**0,048**	0,048
P@10	TS+TC	0,002	0,030	0,020	**0,033**	0,031
	TC	0,001	0,040	0,024	0,041	**0,042**

TABLEAU 5.2 : Scores de similarité pour les corpus de petite taille, avec la fenêtre restreinte (de 5 mots - W5), et l'utilisation des seuils sur les mots cibles et les contextes, pour les termes complexes (TC) et les termes simples et complexes (TS+TC). La valeur la plus élevée par ligne est indiquée en gras.

		Menelas				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq [3]	Ts+TC	19 431	19 431	19 431	19 431	19 431
	TC	2 220 647	2 220 647	2 220 647	2 220 647	2 220 647
Rel dans Ref [4]	Ts+TC	94	94	94	94	94
	TC	196	196	196	196	196
Rprec	TS+TC	0	0,010	0,010	0,010	0,010
	TC	0	0,006	0,005	**0,012**	0,005
MAP	TS+TC	0,031	0,054	0,050	**0,080**	0,065
	TC	0,006	0,021	0,013	**0,024**	0,022
P@1	TS+TC	0	0	0	0	0
	TC	0	0	0,004	**0,022**	0,009
P@5	TS+TC	0,012	0,012	0,024	**0,029**	0,024
	TC	0	0,013	0,007	**0,013**	0,009
P@10	TS+TC	0,006	0,015	0,018	**0,029**	0,015
	TC	0	0,013	0,007	**0,013**	0,009
		Textes Cliniques				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq	TS+TC	4 254 939	4 254 939	4 254 939	4 254 939	4 254 939
	TC	1 696 871	1 696 871	1 696 871	1 696 871	1 696 871
Rel dans Ref	TS+TC	14 632	14 632	14 632	14 632	14 632
	TC	12 980	12 980	12 980	12 980	12 980
Rprec	TS+TC	0	0,021	0,012	**0,025**	0,021
	TC	0,001	0,031	0,014	**0,029**	0,031
MAP	TS+TC	0,004	0,023	0,012	**0,027**	0,021
	TC	0,005	**0,035**	0,014	0,031	0,033
P@1	TS+TC	0	0,015	0,017	**0,044**	0,016
	TC	0	0,020	0,019	**0,041**	0,020
P@5	TS+TC	0	0,025	0,014	**0,029**	0,025
	TC	0	0,034	0,020	**0,031**	0,030
P@10	TS+TC	0	0,025	0,014	**0,023**	0,024
	TC	0	**0,033**	0,018	0,027	0,030

TABLEAU 5.3 : Scores de similarité pour les petits corpus avec la fenêtre large (21 mots), et l'utilisation des seuils sur les mots cibles et les contextes, pour les termes complexes (TC) et les termes simples et complexes (TS+TC). La valeur la plus élevée par ligne est indiquée en gras.

Les résultats sont similaires à ceux obtenus avec la fenêtre restreinte. La même tendance se dessine : l'indice de Jaccard pondéré avec la Fréquence Relative semble être la mesure la plus adaptée, pour les deux corpus, aussi bien pour les termes complexes comme mots cibles que pour la combinaison des termes simples et des termes complexes. De même que pour la fenêtre restreinte, le Cosinus pondéré obtient des résultats nuls ou proches de zéro.

5.1.2.2 Corpus de grande taille

Pour les corpus de grande taille, c'est-à-dire les corpus Recettes et Guides Alimentaires, nous analysons également les résultats obtenus en utilisant des seuils sur les mots cibles et les contextes. De même, nous décrivons dans un premier temps les résultats obtenus avec la fenêtre restreinte (5 mots), puis ceux obtenus avec la fenêtre large (21 mots).

Fenêtre restreinte

Pour les corpus de grande taille et la fenêtre restreinte, les résultats diffèrent selon le corpus. Ainsi, pour le corpus Recettes, nous pouvons observer un comportement des mesures de similarité comparable à celui observé précédemment sur les plus petits corpus. Quels que soient la métrique d'évaluation et les mots cibles utilisés, l'indice de Jaccard pondéré (JACC-FREQ) obtient les valeurs les plus élevées, à l'exception de la P@1 pour les termes simples et termes complexes (TS+TC). La MAP obtenue pour TS+TC s'élève à 0,063 et pour les termes complexes, contre 0,053 pour l'indice de Jaccard non pondéré (JACC), 0,051 pour la Fréquence des Contextes Partagés (FREQCTXT) et 0,047 pour le Nombre de Contextes Partagés (NBCTXT). L'écart entre ces quatre mesures se situe à 0,010 en moyenne. Le Cosinus, en revanche, obtient des valeurs bien plus faibles, même si ces valeurs sont globalement un peu plus élevées que pour les petits corpus, avec une MAP de 0,018 pour les TS+TC. Si le nombre de relations acquises est comparable entre les termes complexes (TC) et la combinaison des termes simples et des termes complexes (TS+TC), avec un nombre de relations un peu plus élevé pour les TC (905 617 relations et 862 078 relations pour les TS+TC), en revanche le nombre de relations retrouvées dans la référence est beaucoup plus faible pour les TC, avec seulement 92 relations.

Pour les Guides Alimentaires, nous pouvons constater un schéma global similaire : quatre mesures proches (NBCTXT, FREQCTXT, JACC et JACCFREQ), et le Cosinus bien plus faible. Cependant, les valeurs les plus élevées sont obtenues essentiellement avec l'indice de Jaccard non pondéré.

		Recettes				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq [5]	Ts+TC	862 078	862 078	862 078	862 078	862 078
	TC	905 617	905 617	905 617	905 617	905 617
Rel dans Ref [6]	Ts+TC	1 860	1 860	1 860	1 860	1 860
	TC	92	92	92	92	92
Rprec	TS+TC	0,001	0,052	0,052	**0,065**	0,041
	TC	0,004	0,019	0,018	**0,034**	0,006
MAP	TS+TC	0,005	0,051	0,053	**0,063**	0,047
	TC	0,006	0,017	0,024	**0,028**	0,009
P@1	TS+TC	0,002	0,059	**0,104**	0,087	0,052
	TC	0,010	0,050	0,059	**0,089**	0,020
P@5	TS+TC	0	0,050	0,061	**0,071**	0,044
	TC	0,012	0,026	0,032	**0,036**	0,018
P@10	TS+TC	0	0,043	0,047	**0,054**	0,039
	TC	0,011	0,016	0,022	**0,024**	0,014
		Guides Alimentaires				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq	TS+TC	1 275 946	1 275 946	1 275 946	1 275 946	1 275 946
	TC	2 801 608	2 801 608	2 801 608	2 801 608	2 801 608
Rel Ress	TS+TC	1 552	1 552	1 552	1 552	1 552
	TC	195	195	195	195	195
Rprec	TS+TC	0,010	0,029	**0,048**	0,027	0,035
	TC	0,002	0,008	0,010	0,003	0,008
MAP	TS+TC	0,018	0,049	0,072	0,027	**0,053**
	TC	0,010	0,017	**0,028**	0,003	0,022
P@1	TS+TC	0,007	0,044	**0,070**	0,054	0,052
	TC	0	0	**0,019**	0,007	0
P@5	TS+TC	0,005	0,035	**0,041**	0,039	0,034
	TC	0	0,008	0,012	0,002	**0,015**
P@10	TS+TC	0,004	0,028	**0,035**	0,030	0,027
	TC	0,004	0,008	0,012	0,003	0,010

TABLEAU 5.4 : Scores de similarité pour les corpus de grande taille avec la fenêtre restreinte (5 mots - W5), utilisation des seuils sur les mots cibles et les contextes, pour les termes complexes (TC) et la combinaison des termes simples et des termes complexes (TS+TC). La valeur la plus élevée par ligne est indiquée en gras.

		Recettes				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq [7]	TS+TC	347 938	347 938	347 938	347 938	347 938
	TC	926 794	926 794	926 794	926 794	926 794
Rel dans Ref [8]	TS+TC	1 350	1 350	1 350	1 350	1 350
	TC	390	390	390	390	390
Rprec	TS+TC	0,007	0,031	0,043	**0,045**	0,030
	TC	0,001	0,005	0,004	**0,016**	0,004
MAP	TS+TC	0,011	0,006	0,007	**0,047**	0,005
	TC	0,002	0,006	0,007	**0,017**	0,005
P@1	TS+TC	0,003	0,034	0,068	**0,068**	0,027
	TC	0	0	0,007	**0,026**	0
P@5	TS+TC	0,005	0,037	0,056	**0,056**	0,031
	TC	0,003	0,005	0,007	**0,009**	0,005
P@10	TS+TC	0,008	0,033	**0,046**	0,044	0,030
	TC	0,001	0,006	0,006	**0,009**	0,004
		Guides Alimentaires				
		COS-IM	FREQCTXT	JACC	JACC-FREQ	NBCTXT
Rel Acq	TS+TC	419 333	419 333	419 333	419 333	419 333
	TC	25 662 142	25 662 142	25 662 142	25 662 142	25 662 142
Rel dans Ref	TS+TC	1 210	1 210	1 210	1 210	1 210
	TC	356	356	356	356	356
Rprec	TS+TC	0,011	0,032	0,054	**0,054**	0,032
	TC	0,003	0,017	**0,019**	0,003	0
MAP	TS+TC	0,025	0,046	**0,086**	0,081	0,049
	TC	0,002	0,016	0,024	**0,033**	0,018
P@1	TS+TC	0,005	0,035	0,079	**0,094**	0,050
	TC	0	0	**0,023**	0,012	0
P@5	TS+TC	0,004	0,035	**0,061**	0,060	0,034
	TC	0	0,009	0,007	**0,014**	0,009
P@10	TS+TC	0,007	0,027	**0,052**	0,047	0,027
	TC	0	0,006	0,006	**0,012**	0,006

TABLEAU 5.5 : Scores de similarité pour les corpus de grande taille, avec la fenêtre large (21 mots - W21), et l'utilisation des seuils, pour les termes complexes (TC) et la combinaison des termes simples et des termes complexes (TS+TC). La valeur la plus élevée par ligne est indiquée en gras.

Fenêtre large

Pour les corpus de grande taille et la fenêtre large, les résultats obtenus avec les seuils sont présentés dans le tableau 5.5.

Pour le corpus Recettes, le comportement des mesures de similarité est comparable à celui observé avec une fenêtre restreinte. Ainsi, pour le corpus le plus volumineux, le corpus Recettes, l'indice de Jaccard pondéré obtient les valeurs les plus élevées, avec toutes les métriques d'évaluation.

Pour les Guides Alimentaires, les résultats sont de meilleure qualité quand l'indice de Jaccard est utilisé, à la fois avec et sans pondération, selon la métrique d'évaluation. Ainsi, quand les mots cibles sont les termes simples et complexes, la MAP est de meilleure qualité avec l'indice de Jaccard non pondéré (JACC), avec une MAP de 0,086, mais la R-précision est plus élevée avec l'indice de Jaccard pondéré (JACC-FREQ), avec une valeur de 0,054.

5.1.3 Seuils et sélection des contextes

Nous présentons et discutons dans cette section l'apport des seuils appliqués aux mots cibles et aux contextes, ainsi que la contribution de la sélection des contextes les plus discriminants. Nous cherchons à définir le meilleur moyen de réduire la dispersion des données, en jouant à la fois sur les contextes, les mots cibles, le score de similarité et les contextes les plus discriminants.

Nous présentons dans un premier temps les valeurs des différents seuils ; le tableau 5.6 récapitule les valeurs des trois seuils sur les contextes et mots cibles.

5.1.3.1 Seuils sur les mots cibles et les contextes

Nous utilisons la combinaison de trois seuils portant sur les mots cibles et sur les contextes partagés. Deux seuils portent sur les contextes : sur le nombre de contextes partagés par deux mots cibles et sur la fréquence des contextes partagés par deux mots cibles, et le troisième seuil porte sur la fréquence des mots cibles.

Comme l'on pouvait s'y attendre, les valeurs des seuils varient en fonction du corpus, avec des seuils plus élevés quand le corpus est plus volumineux. Ainsi, le seuil sur la fréquence des mots cibles est égal à 3 pour les deux plus petits corpus (Menelas et Textes Cliniques) et égal à 22 pour le corpus le plus volumineux (le corpus Recettes). Les seuils sont également plus élevés pour les termes simples et complexes (TS+TC) que pour les termes complexes seuls (TC). Ainsi, pour le corpus Recettes et la fenêtre large, le seuil sur la Fréquence des Contextes Partagés est égal à 6 pour TS+TC et égal à 2 pour les termes complexes seuls. Nous pouvons en déduire que des seuils plus

Corpus	Seuil	TS+TC W5	TS+TC W21	TC W5	TC W21
Recettes	Nombre de contextes partagés	2	4	*1*	2
	Fréquence des contextes partagés	3	6	2	2
	Fréquence des mots cibles	22	22	6	6
Guides alimentaires	Nombre de contextes partagés	*1*	2	*1*	*1*
	Fréquence des contextes partagés	2	3	*1*	*1*
	Fréquence des mots cibles	5	5	2	2
Textes Cliniques	Nombre de contextes partagés	*1*	*1*	*1*	*1*
	Fréquence des contextes partagés	*1*	2	*1*	2
	Fréquence des mots cibles	3	3	3	3
Menelas	Nombre de contextes partagés	*1*	2	*1*	2
	Fréquence des contextes partagés	2	2	*1*	2
	Fréquence des mots cibles	3	3	2	3

TABLEAU 5.6 : Paramètres : valeurs des seuils sur les contextes et mots cibles, pour les termes simples et complexes (TS+TC), pour les termes complexes pris isolément (TC), pour les 4 corpus, avec les deux tailles de fenêtre : 5 mots (W5) et 21 mots (W21).

élevés auront comme conséquence un impact plus important sur la qualité des résultats, avec ainsi un impact beaucoup plus important quand les mots cibles sont les termes simples et complexes.

Parmi les seuils pour le corpus Textes Cliniques, les seuils sur le nombre de contextes partagés sont tous égaux à 1, ce qui équivaut à ne pas utiliser de seuil pour ce paramètre. Pour la fenêtre restreinte, les seuils sur la fréquence des contextes partagés sont également égaux à 1. Ainsi, pour ce corpus et la fenêtre graphique restreinte, les seuils portent uniquement sur les mots cibles.

Le comportement des seuils étant similaire quelle que soit la mesure de similarité utilisée, nous faisons le choix de présenter les résultats obtenus avec l'indice de Jaccard pondéré avec la fréquence relative. Les résultats obtenus avec les autres mesures de similarité sont présentés en annexe (cf. B). Nous présentons les résultats obtenus séparément, selon la taille du corpus. Nous analysons l'impact des seuils à la fois sur le nombre de relations acquises et sur la qualité des résultats (MAP, R-précision et P@1).

5.1.3.2 Impact des seuils sur les mots cibles et les contextes

Nous présentons dans un premier temps l'impact des seuils sur les deux plus petits corpus ; le corpus Menelas et le corpus de Textes Cliniques. Nous décrirons ensuite les résultats obtenus pour les corpus de plus grande taille, les corpus Guides Alimentaires et Recettes.

Pour ces deux ensembles, les résultats sont analysés à travers la comparaison des résultats obtenus sans et avec les seuils sur les mots cibles et contextes, d'un point de vue quantitatif (MAP, R-précision et précision) et du point de vue du nombre de relations acquises. La pertinence des seuils doit se traduire par la réduction du nombre de relations acquises et l'amélioration de la qualité des relations proposées.

Corpus de petite taille

			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	T.Clin.	Avec seuils	0,027	0,031	0,045	**0,052**
		Sans seuils	0,027	0,030	0,036	**0,041**
	Menelas	Avec seuils	**0,080**	0,024	**0,121**	**0,119**
		Sans seuils	**0,026**	0,020	**0,084**	**0,092**
R-prec	Textes Cliniques	Avec seuils	0,025	0,029	0,034	0,046
		Sans seuils	0,023	0,029	0,028	0,035
	Menelas	Avec seuils	0,010	0,012	**0,079**	**0,080**
		Sans seuils	0,009	0,011	**0,035**	**0,064**
P@1	Textes Cliniques	Avec seuils	**0,044**	0,041	0,054	**0,061**
		Sans seuils	**0,034**	0,039	0,046	**0,051**
	Menelas	Avec seuils	0	0	0,053	**0,107**
		Sans seuils	0,007	0	0,049	**0,088**
P@5	Textes Cliniques	Avec seuils	0,029	0,031	**0,041**	0,048
		Sans seuils	0,027	0,031	**0,032**	0,035
	Menelas	Avec seuils	**0,029**	0,022	0,042	0,029
		Sans seuils	**0,017**	0,016	0,039	0,024

TABLEAU 5.7 : Corpus de petite taille : impact des seuils appliqués aux mots cibles et aux contextes sur la qualité des résultats obtenus avec l'indice de Jaccard pondéré avec la Fréquence Relative, pour une fenêtre large (W21) et une fenêtre restreinte (W5), avec les termes simples et complexes (TS+TC) et les termes complexes (TC) comme mots cibles. Les couples de lignes *sans seuils - avec seuils* ayant un écart supérieur ou égal à 0,01 sont indiqués en gras.

Nous discutons dans un premier temps, l'impact des seuils sur les résultats obtenus en termes de MAP, R-précision et précision. Le tableau 5.7 contient les résultats pour les deux corpus de petite taille avec et sans seuils sur les mots cibles et les contextes, pour les deux tailles de fenêtre graphique. Les paires de résultats *avec seuils - sans seuils* caractérisées par un écart important sont en gras (après observation, nous considérons comme écart important tout écart supérieur ou égal à 0,01).

Nous observons que les résultats diffèrent selon le corpus. Ceci s'explique par des seuils définis en fonction des fréquences des contextes et des mots cibles dans chaque corpus.

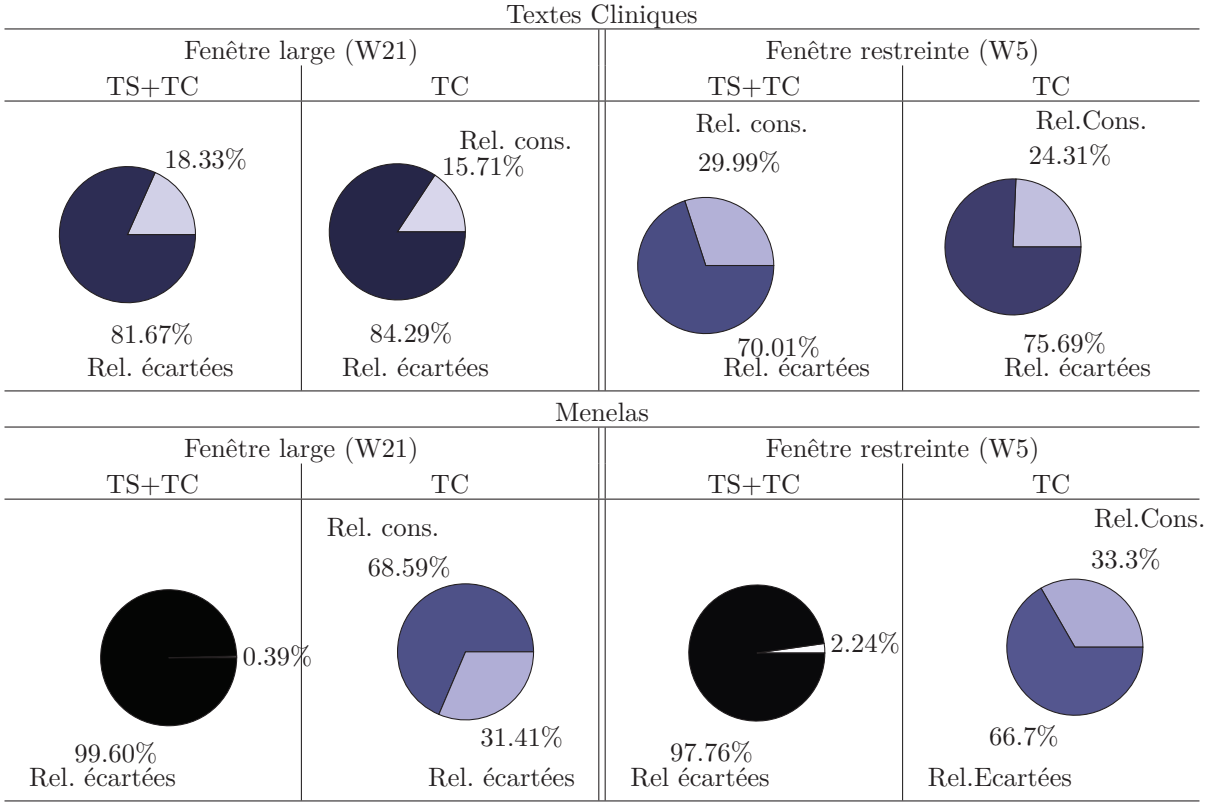

FIGURE 5.1 : Relations acquises et corpus de petite taille (Menelas et les Textes Cliniques) : nombre de relations conservées et écartées lors de l'application des seuils sur les mots cibles et les contextes, avec l'indice de Jaccard pondéré et les deux tailles de fenêtre graphique (21 mots - W21, et 5 mots - W5).

Ainsi, le corpus Menelas est plus impacté par l'application des seuils ; avec la fenêtre restreinte et les termes simples et complexes en mots cibles, la R-précision augmente de 0,044 pour le corpus Menelas et seulement de 0,004 pour les Textes Cliniques. En effet, les seuils sur les mots cibles et les contextes sont plus élevés pour le corpus Menelas (cf. table 5.6), et ont par conséquent plus d'influence sur les résultats.

Globalement, nous observons que l'utilisation des seuils a un impact positif ou nul sur la qualité des résultats, mais jamais négatif, à une seule exception : la précision au rang 1 (P@1) pour le corpus Menelas avec la fenêtre large et les termes simples et complexes. Lorsque la fenêtre restreinte et les termes complexes sont utilisés, soit lorsque les fréquences sont faibles (W5-TC), les seuils permettent avec la plupart des métriques d'évaluation et pour les deux corpus d'améliorer les résultats. Ainsi, la MAP obtenue pour le corpus Menelas augmente de 0,027 et celle des Textes Cliniques augmente de 0,012 avec cette configuration. Lorsque les mots cibles sont les termes simples et complexes, l'impact des seuils est plus important, avec les deux tailles de fenêtre. La MAP pour le corpus Menelas augmente ainsi de 0,054 pour la fenêtre large.

Dans la figure 5.1 nous présentons le nombre de relations acquises avec et sans l'utilisation des seuils sur les mots cibles et les contextes. Nous observons que globalement le nombre de relations écartées par les seuils est le plus élevé quand les mots cibles sont les termes simples et complexes et quand le corpus est petit. Ainsi, pour le corpus Menelas, 99,6% de relations sont écartées par les seuils quand la fenêtre est large (21 mots) et 97,76% quand la fenêtre est restreinte. Nous pouvons constater que pour le corpus Menelas, ces relations écartées sont des relations potentiellement fausses, c'est-à-dire non contenues dans les ressources, car la qualité des résultats est améliorée avec les seuils.

Pour le corpus Textes Cliniques, en revanche, l'impact des seuils est plus important quand les mots cibles sont les termes complexes. Ce comportement propre aux Textes Cliniques et contraire à ceux des autres corpus, est lié aux valeurs des seuils (cf. tableau 5.6). En effet, pour les Textes Cliniques, les seuils sont identiques avec les deux types de mots cibles, TC et TC+TS. Lorsque les fréquences et nombres des contextes et les fréquences des mots cibles sont égales, les seuils ont un impact plus important avec les termes complexes utilisés comme mots cibles.

Corpus de grande taille

Nous discutons à présent l'impact des seuils pour les corpus de grande taille. L'analyse est réalisée à partir des résultats obtenus avec et sans seuils, évalués en termes de MAP, R-précision et précision (cf. tableau 5.8), mais également du nombre de relations écartées lors de l'application des seuils (cf. figure 5.2).

Les seuils sur les mots cibles et les contextes ont un impact fort sur la qualité des résultats, quand la fenêtre large est utilisée et que les mots cibles sont les termes simples

			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Guides Alim.	Avec seuils	**0,081**	*0,033*	0,027	0,003
		Sans seuils	**0,052**	*0,050*	0,030	0,002
	Recettes	Avec seuils	**0,047**	**0,017**	**0,063**	**0,028**
		Sans seuils	**0,031**	**0,007**	**0,036**	**0,010**
R-préc.	Guides Alim.	Avec seuils	**0,054**	0,019	0,027	0,003
		Sans seuils	**0,037**	0,018	0,028	0,002
	Recettes	Avec seuils	**0,045**	**0,016**	**0,065**	**0,034**
		Sans seuils	**0,025**	**0,007**	**0,036**	**0,010**
P@1	Guides Alim.	Avec seuils	**0,094**	0,012	0,054	0,007
		Sans seuils	**0,055**	0,011	0,050	0,003
	Recettes	Avec seuils	**0,068**	**0,026**	**0,087**	**0,089**
		Sans seuils	**0,046**	**0,004**	**0,053**	**0,030**
P@5	Guides Alim.	Avec seuils	**0,060**	0,014	0,039	0,002
		Sans seuils	**0,039**	0,015	0,033	0,002
	Recettes	Avec seuils	**0,056**	0,009	**0,071**	**0,036**
		Sans seuils	**0,030**	0,005	**0,036**	**0,012**

TABLEAU 5.8: Corpus de grande taille : impact des seuils sur les mots cibles et les contextes sur la qualité des résultats obtenus avec l'indice de Jaccard pondéré avec la Fréquence Relative, avec une fenêtre large (W21) et une fenêtre restreinte (W5), avec les termes simples et complexes (TS+TC) et les termes complexes (TC) comme mots cibles. Les couples de lignes *sans seuils* - *avec seuils* ayant un écart supérieur ou égal à 0,01 sont indiquées en gras.

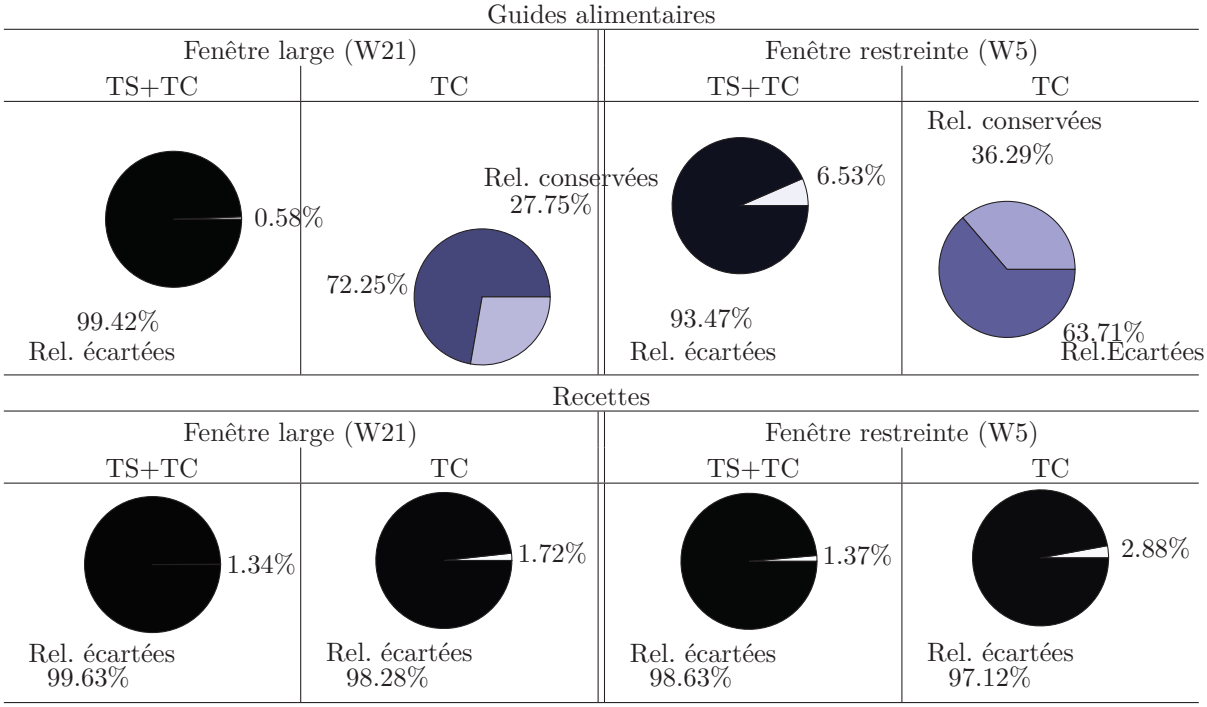

FIGURE 5.2 : Relations acquises et corpus de grande taille (Guides Alimentaires et Recettes) : nombre de relations conservées et écartées lors de l'application des seuils sur les mots cibles et les contextes, avec l'indice de Jaccard pondéré et les deux tailles de fenêtre graphique (21 mots - W21, et 5 mots - W5).

et complexes (TS+TC), quelle que soit la métrique d'évaluation. Ainsi, pour le Guides Alimentaires, la MAP augmente de 0,052 à 0,081, et pour le corpus Recettes de 0,031 à 0,047. L'impact des seuils sur le nombre de relations acquises va de pair avec la qualité ; avec cette même configuration, les mots cibles sont les TS+TC, le nombre de relations acquises diminue fortement, écartant 99,42 % des relations avec la fenêtre large pour les Guides Alimentaires et 99,63 % pour le corpus Recettes.

Quand les termes complexes sont utilisés comme mots cibles (TC), les seuils ont un impact positif sur la qualité des résultats ainsi que sur le nombre de relations acquises pour le corpus Recettes, avec les deux tailles de fenêtre. Ainsi, la P@1 augmente de 0,004 à 0,026 et la R-précision de 0,007 à 0,016. De plus, 98,28 % des relations sont écartées par les seuils quand la fenêtre large est utilisée, et 97,12 % avec la fenêtre restreinte.
En revanche, pour les Guides Alimentaires, l'impact des seuils sur la qualité des résultats est nul ou presque, et le nombre de relations acquises est nettement moins impacté que lorsque les mots cibles sont les termes simples et complexes. En effet, les relations acquises écartées sont de 72,25 % et 63,71 %, contre plus de 90 % pour les TS+TC.

Bilan

Les seuils sur les mots cibles et les contextes ont globalement un impact positif sur la qualité des résultats et sur le nombre de relations acquises qui diminue fortement. Cependant, cet impact est globalement plus fort quand les mots cibles sont les termes simples et complexes, pour lesquels les fréquences sont plus élevées, et par conséquent les seuils sont plus importants (cf. tableau 5.6). Ainsi, pour nos expériences, nous utilisons comme mots cibles les termes simples et complexes, et nous appliquons les seuils sur les mots cibles et les contextes.

5.1.3.3 Sélection des contextes les plus discriminants

Nous décrivons à présent l'impact du Cf-Itf sur les résultats obtenus pour les petits corpus. Les analyses sont similaires pour les corpus volumineux : les résultats obtenus sont donnés en annexe.

Les résultats pour les petits corpus sont donnés dans le tableau 5.9. Les résultats en italique sont ceux pour lesquels nous observons une baisse de la qualité des résultats lors de l'application du Cf-Itf ; diminution de la MAP, R-précision et précision, diminution du nombre de relations retrouvées dans la référence.

Pour le corpus Menelas, le Cf-Itf fait diminuer la qualité des résultats avec toutes les métriques, et toutes les configurations de fenêtre et mots cibles, à l'exception de la combinaison de la fenêtre restreinte et des termes simples et complexes (W5-TS+TC). Ainsi, la MAP baisse de 0,026 à 0,007, quand la fenêtre large et les termes simples

			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Menelas	Sans seuils	*0,026*	*0,020*	0,084	*0,092*
		Avec Cf-Itf	*0,007*	*0,005*	0,330	*0,005*
	Textes Cliniques	Sans seuils	0,027	0,030	0,036	0,041
		Avec Cf-Itf	0,097	0,122	0,078	0,089
R-préc.	Menelas	Sans seuils	*0,009*	*0,011*	0,035	*0,064*
		Avec Cf-Itf	*0*	*0*	*0,208*	*0*
	Textes Cliniques	Sans seuils	0,023	0,029	0,028	0,035
		Avec Cf-Itf	0,053	0,087	0,042	0,049
P@1	Menelas	Sans seuils	*0,007*	0	0,049	*0,088*
		Avec Cf-Itf	*0*	0	0,250	*0*
	Textes Cliniques	Sans seuils	0,034	0,039	*0,046*	*0,051*
		Avec Cf-Itf	0,054	0,077	*0,040*	*0,048*
P@5	Menelas	Sans seuils	*0,017*	*0,016*	0,039	*0,024*
		Avec Cf-Itf	*0,002*	*0*	0,133	*0*
	Textes Cliniques	Sans seuils	*0,027*	0,031	*0,032*	*0,035*
		Avec Cf-Itf	*0,031*	0,036	*0,027*	*0,026*
Rel Acq	Menelas	Sans seuils	4 868 581	3 237 145	362 227	824 204
		Avec Cf-Itf	2 783 187	1 873 747	35 540	531 144
	Textes Cliniques	Sans seuils	23 214 096	6 980 666	9 071 510	6 980 666
		Avec Cf-Itf	1 367 267	1 533 047	1 903 980	1 533 047
Rel Ref	Menelas	Sans seuils	*718*	*210*	*370*	*118*
		Avec Cf-Itf	*338*	*86*	*28*	*38*
	Textes Cliniques	Sans seuils	*26 464*	*23 746*	*17 396*	*15 782*
		Avec Cf-Itf	*794*	*688*	*1 136*	*992*

TABLEAU 5.9 : Corpus de petite taille : impact du Cf-Itf avec l'indice de Jaccard pondéré avec la Fréquence Relative, avec une fenêtre large (W21) et une fenêtre restreinte (W5), avec les termes simples et complexes (TS+TC) et les termes complexes (TC) comme mots cibles. Les cas où le Cf-Itf fait baisser la qualité ont indiqués en italique.

et complexes sont utilisés. Pour la configuration W5-TS+TC, pour laquelle la qualité des résultats est améliorée, le nombre de relations retrouvées dans la référence est en contrepartie fortement diminué, et passe de 370 à 28 relations dans la référence.

Pour les Textes Cliniques, le Cf-Itf provoque une diminution de la précision lorsque la fenêtre restreinte est utilisée : la P@5 pour les termes complexes passe de 0,035 à 0,026. La baisse en qualité est moins importante que pour le corpus Menelas. En revanche, avec les autres métriques d'évaluation, et pour les autres configurations, le Cf-Itf permet d'améliorer la qualité des résultats. Ainsi, la MAP pour les termes complexes avec une fenêtre large (W21-TC) augmente de 0,030 à 0,122. Cependant, si la qualité des résultats est améliorée, le nombre de relations retrouvées dans la référence diminue drastiquement. Pour la configuration citée précédemment (W21-TC), le nombre de relations retrouvées dans la référence passe de 23 746 relations à 688.

Cette baisse de la qualité des relations lors de l'application du Cf-Itf, soit au niveau de la précision et du classement des voisins, soit au niveau du nombre de relations identifiées dans la référence, est similaire avec les corpus de grande taille. Nous avons donc choisi d'exclure ce paramètre de la méthode distributionnelle.

5.1.4 Bilan

Ces expériences autour des paramètres distributionnels nous permettent de déterminer quels sont les paramètres et leurs valeurs les plus adaptés à nos quatre corpus, en fonction des mots cibles et de la taille du corpus. Ce premier ensemble d'expériences nous permet d'établir une base pour nos expériences autour de l'abstraction des contextes distributionnels. L'ensemble des paramètres conservés sont récapitulés dans le tableau 5.10.

Il ressort de ce premier ensemble d'expériences qu'il est préférable d'avoir pour mots cibles à la fois les termes simples et complexes identifiés automatiquement, et de calculer des relations entre ces deux catégories de mots prises comme un ensemble. L'étude des mesures de similarité montre qu'il est préférable d'utiliser l'indice de Jaccard pondéré par la fréquence relative et le Nombre de Contextes Partagés avec les petits corpus (autour de 100 000 mots), et l'indice de Jaccard avec et sans pondération lorsqu'il s'agit de corpus plus volumineux (autour du million de mots).

Par ailleurs, l'utilisation de l'ensemble des seuils sur les différents paramètres améliore les résultats en termes de qualité tout en réduisant le nombre de relations acquises, en écartant par contre généralement la moitié des relations identifiées dans la ressource. Malgré cet aspect négatif, il est préférable d'utiliser les seuils sur les mots cibles et les contextes. Enfin, la sélection des contextes les plus discriminants par le Cf-Itf dégrade les résultats et ne sera finalement pas conservée dans la suite des expériences.

Paramètre	Valeur(s)
MOTS CIBLES	Termes simples et complexes
CONTEXTES	1. Fenêtre graphique large (21 mots) 2. Fenêtre graphique restreinte (5 mots)
MESURES DE SIMILARITE ET PONDERATION	1. indice de Jaccard + fréquence relative 2. Cosinus + IM
SEUILS	Combinaison de 3 seuils sur les mots cibles et les contextes (Nombre et Fréquence des contextes partagés + fréquence des mots cibles)

TABLEAU 5.10 : Récapitulatif des paramètres choisis pour les expériences autour de l'abstraction des contextes.

5.2 Abstraction des contextes

Après avoir défini les paramètres adaptés à l'analyse distributionnelle de nos corpus, nous évaluons et discutons maintenant l'impact de la normalisation et de la généralisation des contextes distributionnels. Nous utilisons comme point de comparaison (*baseline*) les résultats obtenus avec l'analyse distributionnelle seule, c'est-à-dire sans normalisation. L'ensemble des expériences est réalisé en suivant les configurations identifiées précédemment (cf. section 5.1). Les mots cibles sont les termes simples et complexes, et nous utilisons comme contextes deux tailles de fenêtre graphique : la fenêtre large (de 21 mots) et la fenêtre restreinte (de 5 mots). Les mesures de similarité choisies sont le Cosinus pondéré avec l'Information Mutuelle et l'indice de Jaccard pondéré avec la Fréquence Relative. Et enfin, nous appliquons trois seuils sur les mots cibles et sur les contextes (Fréquence et Nombre des Contextes Partagés, Fréquence des mots cibles).

Les règles d'abstraction, définies en section 3.2, sont appliquées individuellement et séparément avec les trois ensembles de relations d'hyperonymie $\mathbb{H}_{\mathrm{PLS}}(w_i)$ - acquises grâce aux patrons lexico-syntaxiques, $\mathbb{H}_{\mathrm{IL}}(w_i)$ - acquises grâce à l'inclusion lexicale, $\mathbb{H}_{\mathrm{VT}}(w_i)$ - acquises à l'aide des variantes terminologiques, et un ensemble de relations de synonymie $\mathbb{S}(w_i)$, pour chaque contexte w_i - acquises par compositionnalité sémantique.

5.2.1 Expériences

Afin de cerner la contribution de chaque méthode linguistique décrite à la section 3.3, ainsi que la complémentarité de chaque source de relations, nous avons réalisé trois séries d'expériences autour de l'abstraction des contextes : une première série autour

de la généralisation des contextes, une seconde pour la normalisation, et enfin un dernier ensemble d'expériences concerne la combinaison de la généralisation et de la normalisation.

Généralisation
Tout d'abord, pour généraliser les contextes (cf. sous-section 3.2.1), nous avons réalisé un ensemble d'expériences où les règles de généralisation exploitent chaque méthode individuellement. Les règles de généralisation des contextes distributionnels w_i sont alors 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) acquises avec Fastr.
Puis, de manière séquentielle, nous avons appliqué les règles de généralisation à deux des trois ensembles de relations d'hyperonymie ($\mathbb{H}_{PLS}(w_i)$ puis $\mathbb{H}_{IL}(w_i)$ – AD/PLS+IL, $\mathbb{H}_{VT}(w_i)$ puis $\mathbb{H}_{PLS}(w_i)$ – AD/VT+PLS, etc.). Tous les contextes sont alors généralisés en utilisant les relations proposées par l'un des ensembles (par exemple $\mathbb{H}_{PLS}(w_i)$), puis les contextes généralisés ou non sont à nouveau généralisés en utilisant un autre ensemble de relations (par exemple $\mathbb{H}_{IL}(w_i)$). De même, nous combinons les trois ensembles de relations (par exemple, $\mathbb{H}_{PLS}(w_i)$ puis $\mathbb{H}_{IL}(w_i)$ puis $\mathbb{H}_{VT}(w_i)$ – AD/-PLS+IL+VT).
Nous avons également considéré toutes les relations d'hyperonymie 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.

Normalisation
En ce qui concerne la normalisation, définie en sous-section 3.2.2, nous considérons une seule expérience ; la normalisation des contextes w_i à l'aide des clusters de synonymes $\mathbb{S}(w_i)$ acquis automatiquement avec SynoTerm.

Combinaison de la normalisation et de la généralisation
Enfin, nous normalisons dans un premier temps les contextes w_i, puis nous appliquons à ces contextes normalisés $\mathbb{S}(w_i)$ la configuration précédemment décrite pour généraliser les contextes. Pour cela, les termes en relation d'hyperonymie dans $\mathbb{H}$ sont également normalisés avant d'être utilisés pour la généralisation.

Nous décrivons dans les deux sous-sections suivantes les résultats obtenus pour la généralisation des contextes basée sur les ensembles de relations d'hyperonymie pris séparément, tous ensemble et quelques combinaisons séquentielles qui renseignent sur l'utilité et l'impact de l'ordre de généralisation. Nous présentons également les résultats obtenus pour la normalisation des contextes réalisée seule et en la combinant à la généralisation.

5.2.2 Généralisation des contextes distributionnels

Nous présentons dans un premier temps, les résultats obtenus avec la généralisation des contextes, avec les corpus de petite taille ; le corpus Menelas (84 839 mots) et le corpus Textes Cliniques (178 070 mots), puis avec les corpus de plus grande taille, les Guides Alimentaires (471 463 mots) et le corpus Recettes (3 928 658 mots).

En ce qui concerne la qualité des résultats et des limites de l'évaluation, nous souhaitons souligner un fait important, même si cela peut sembler évident : nous évaluons la qualité des voisins distributionnels uniquement pour les mots cibles retrouvés dans la ressource.

Corpus Textes Cliniques

	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	**2 722 289**	**13 372**	**0,045**	**0,034**	**0,054**	**0,041**	**0,033**	**0,014**
AD/VT	2 220 105	12 184	0,047	0,037	0,060	0,040	0,033	0,013
AD/IL	**763 415**	*9 180*	**0,057**	**0,047**	**0,067**	**0,049**	**0,039**	**0,017**
AD/PLS	2 291 724	12 474	0,046	0,036	0,060	0,040	0,033	0,013
AD/VT+IL	**752 768**	9 166	0,057	**0,048**	0,067	0,049	0,039	0,017
AD/IL+PLS	**763 221**	9 126	0,057	0,047	0,067	0,049	0,038	0,017
AD/PLS+IL	**763 295**	9 168	0,057	0,047	0,067	0,049	0,039	0,017
AD/ALL3	**777 264**	9 262	**0,058**	**0,049**	**0,068**	**0,049**	**0,039**	**0,017**

Corpus Menelas

	REL ACQ	REL RES	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	**8 118**	**60**	**0,121**	**0,079**	**0,053**	**0,042**	**0,037**	**0,013**
AD/VT	6 915	56	**0,134**	**0,083**	**0,056**	**0,056**	**0,039**	0,013
AD/IL	**3 100**	38	0,125	0,071	**0,071**	0,014	0,029	0,013
AD/PLS	6 341	50	0,126	0,063	**0,063**	0,038	0,025	0,014
AD/VT+IL	3 100	38	0,125	0,071	**0,071**	0,014	0,029	0,013
AD/IL+PLS	**2 714**	32	**0,133**	*0,077*	**0,077**	*0,031*	**0,039**	0,012
AD/PLS+IL	3 007	38	0,125	0,071	**0,071**	0,014	0,029	0,013
AD/ALL3	3 285	38	**0,126**	0,071	**0,071**	0,014	0,036	0,013

TABLEAU 5.11 : Résultats obtenus pour les corpus de petite taille, avec la fenêtre restreinte (W5) et l'indice de Jaccard pondéré.

5.2.2.1 Corpus de petite taille

Les plus petits de nos corpus de travail correspondent aux corpus médicaux (décrits en section 4.1.1). La qualité des résultats obtenus avec ces corpus est évaluée en comparant les relations sémantiques acquises avec les relations fournies par l'UMLS. Pour le corpus Menelas, la référence est constituée à partir de la partie française de l'UMLS, et comprend 2 343 relations entre les termes du corpus. Pour le corpus de Textes Cliniques, la référence est constituée à partir de l'UMLS anglais, et contient 53 203 relations.

Les deux corpus s'opposent par rapport à la taille de la fenêtre graphique. En effet, pour les Textes Cliniques, le comportement de la généralisation des contextes est similaire avec les deux tailles de fenêtre (de 5 mots et 21 mots), même si les valeurs sont un peu plus élevées pour la fenêtre restreinte. En revanche, pour le corpus Menelas, les résultats obtenus sont différents selon la taille de fenêtre utilisée ; la qualité des résultats est meilleure avec la fenêtre restreinte qu'avec la fenêtre large, mais la généralisation a plus d'impact sur la fenêtre large que sur la fenêtre restreinte. Ces différences sont certainement liées aux valeurs des seuils, qui sont plus élevées avec le corpus Menelas.

En ce qui concerne la mesure de similarité, globalement, pour les petits corpus et les petites fréquences, les résultats obtenus pour le Cosinus sont très faibles [Périnet et Hamon, 2014]. Même si dans cette partie nous commenterons occasionnellement certains comportements particuliers observés pour le Cosinus, les résultats discutés et représentés dans les tableaux sont ceux obtenus avec l'indice de Jaccard pondéré (Jacc-Freq).

5.2.2.1.1 Fenêtre graphique restreinte

Les résultats obtenus pour les corpus de petite taille avec la fenêtre restreinte sont présentés dans le tableau 5.11. Avec une fenêtre graphique restreinte (W5), les mesures de similarité n'ont pas le même comportement pour les deux corpus. Ainsi, pour le corpus Textes Cliniques, les résultats obtenus sont de meilleure qualité avec l'indice de Jaccard, qui obtient une P@1 entre 0,054 et 0,068, contre une P@1 entre 0 et 0,002 avec le Cosinus. En revanche, pour le corpus Menelas, l'écart entre les deux mesures de similarité est nettement moins important, et le Cosinus permet même d'obtenir dans certaines configurations de généralisation de meilleurs résultats que l'indice de Jaccard. Ainsi, pour Menelas, les voisins retrouvés dans l'UMLS et classés en première position sont identiques quelle que soit la mesure de similarité, et ceci dans toutes les configurations (avec et sans généralisation). Notamment la P@1 est égale à 0,071 pour les deux mesures dans la généralisation avec l'inclusion lexicale. Si l'on considère les cinq premiers voisins retournés, la généralisation avec les patrons puis avec l'inclusion lexicale (AD/PLS+IL) est de meilleure qualité pour la mesure de Cosinus, avec une P@5 de 0,029, pour une P@1 de 0,014 avec l'indice de Jaccard.

Valeurs de précision

Les valeurs de précision évoluent différemment pour les deux corpus, en fonction du nombre de voisins pris en compte dans l'évaluation. Pour les Textes Cliniques, la mesure de Jaccard permet de prendre en compte plus de voisins lors de l'évaluation, mais la précision diminue. Pour le Cosinus c'est l'inverse. En revanche pour Menelas, avec les deux mesures de similarité, plus on prend en compte de voisins dans l'évaluation, plus la précision diminue. Ainsi, pour les plus petits corpus, l'indice de Jaccard permet de mieux classer les voisins, alors que le Cosinus permet de récupérer plus de voisins pertinents lorsque plusieurs voisins sont considérés.

Ordonnancement des voisins

Sur le corpus de Textes Cliniques, nous constatons que la généralisation améliore essentiellement l'ordonnancement des voisins, avec une MAP et une R-précision en hausse quand la généralisation est utilisée. De plus, la précision est nettement améliorée quand peu de voisins sont pris en compte dans l'évaluation, c'est-à-dire surtout au premier rang ou dans les cinq premiers voisins, quelle que soit la configuration utilisée. La configuration de généralisation la plus adaptée à ce corpus, avec la fenêtre restreinte, est l'inclusion lexicale utilisée individuellement, avec les deux mesures de similarité. En effet, celle-ci permet de retrouver plus de voisins corrects que la baseline, aussi bien parmi les cinq, les dix que les cent premiers voisins. Avec cette généralisation, la P@5 augmente ainsi de 0,008 et la P@10 de 0,006. La P@1 est la précision qui démontre la plus grande augmentation après généralisation, avec un gain de 0,013.

Sur le corpus Menelas, les variations de la MAP montrent que l'ordonnancement des voisins est amélioré quand les variantes terminologiques (AD/VT) sont utilisées pour généraliser les contextes, et également avec la combinaison de l'inclusion lexicale suivie des patrons lexico-syntaxiques (AD/IL+PLS). En revanche, la R-précision est améliorée uniquement avec les variantes terminologiques. La généralisation avec les variantes permet d'augmenter le nombre de voisins corrects en rang 1, mais également parmi les cinq et les dix premiers voisins, tandis que la généralisation avec l'inclusion lexicale et les patrons (AD/IL+PLS) améliore essentiellement les voisins corrects retrouvés en rang 1.

A la vue des mesures, la qualité des résultats semble bénéficier de la combinaison, aussi bien séquentielle que de l'union des trois sources de relations d'hyperonymie, et ce avec les deux corpus. Toutefois, la combinaison séquentielle réalisée à partir des trois méthodes d'acquisition de relations n'apporte pas de plus-value par rapport à la combinaison à deux méthodes. L'union des trois méthodes permet par contre d'améliorer le classement des voisins, par rapport aux autres configurations de généralisation, en termes de MAP et de R-précision pour les Textes Cliniques, et seulement en termes de MAP pour le corpus Menelas. La précision au premier voisin est également améliorée pour les deux corpus.

Nombre de relations

De même, le nombre de relations retournées diminue avec toutes les configurations de généralisation, par rapport à l'AD seule. Cette diminution permet de diviser au mieux par 2,5 le nombre de relations acquises pour Menelas, et au mieux par 3 le nombre de relations acquises pour les Textes Cliniques, dans les deux cas quand l'inclusion lexicale est utilisée pour la généralisation (soit seule, soit en combinaison). En contrepartie, malheureusement, le nombre de relations retrouvées dans l'UMLS diminue également avec la généralisation, jusqu'à être divisé par 2 pour le corpus Menelas (60 relations obtenues avec l'AD seule, 32 avec la combinaison offrant la meilleure précision). C'est également le cas pour le corpus Textes Cliniques, mais dans une moindre mesure.

Analyse des relations

Etant donné la faible couverture de nos ressources et afin de mieux caractériser les voisins acquis ainsi que la qualité des relations, pour les deux petits corpus, nous avons analysé les dix premiers voisins de chaque mot cible retrouvé dans l'UMLS. Cette analyse est menée à la fois pour la baseline (AD sans normalisation) et pour les meilleures configurations de généralisation (AD/VT et AD/IL+PLS pour Menelas, et AD/IL pour les Textes Cliniques).

Nous observons tout d'abord que pour Menelas, sur les dix premiers voisins, uniquement un seul ou deux voisins au maximum sont retrouvés dans la ressource. Cela ne signifie pas pour autant que les autres voisins ne sont pas pertinents. La combinaison AD/IL+PLS a une plus grande influence sur les mots cibles ; cette configuration réduit le nombre de mots cibles retrouvés dans la ressource et ceux retrouvés sont en partie différents de ceux 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. La généralisation avec la combinaison est certainement plus importante que celle avec les variantes.
Enfin, les contextes et les mots cibles ne sont pas tous affectés de la même manière par la généralisation. Dans certains cas, l'ensemble des dix premiers voisins obtenus après généralisation est totalement modifié par rapport à ceux obtenus avec la baseline, alors que dans d'autres cas le changement est très faible.

Dans le tableau 5.12, nous présentons les dix premiers voisins du mot cible *cholestérol*, obtenus avec la baseline et après généralisation des contextes avec les variantes terminologiques. La généralisation des contextes a une influence sur le classement des voisins : six des dix premiers voisins obtenus avec la baseline descendent dans le classement, c'est-à-dire que la généralisation augmente le score de similarité obtenu par certains voisins classés après les dix premiers. Ces six voisins sont représentés en italique. Les six voisins que la généralisation fait remonter dans la liste des dix premiers sont représentés

Mot cible ***cholestérol***				
	Baseline		**AD/VT**	
Rang	Voisin	Jacc	Voisin	Jacc
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 5.12 : 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), avec l'indice de Jaccard. Les voisins qui remontent dans le classement sont en gras, ceux qui descendent sont en italique.

en gras. Ainsi, certains voisins restent inchangés du point de vue de leur similarité avec le mot cible, tels que *oblitération*, et *angio-coronarographie*, même s'ils descendent dans le classement. D'autres ont leur score de similarité qui est légèrement augmenté (*bilan lipidique*, *triglycéride*) et sont maintenus au même rang.
Parmi ces dix voisins, dans les deux configurations, aucun des voisins n'est retrouvé dans l'UMLS. Cependant, cela ne signifie pas que les voisins identifiés et classés parmi les dix premiers ne sont pas des voisins pertinents. 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 *cholestérol - cholestérol total*. Globalement, les voisins acquis après généralisation sont un ensemble plus homogène sémantiquement, et correspondent au concept d'examen médical.

	Mot cible *cough*					
	Baseline		**AD/IL**		**AD/ALL3**	
Rang	Voisin	Sim	Voisin	Sim	Voisin	Sim
1.	nausea	0,00091	nausea	0,00108	nausea	0,00108
2.	*pain*	0,00063	**fever**	0,00105	**fever**	0,00105
3.	*paroxysmal nocturnal dyspnoea*	0,00048	**vomiting**	0,00105	**vomiting**	0,00105
4.	*weakness*	0,00045	**chill**	0,00101	**chill**	0,00101
5.	*dizziness*	0,00044	**history**	0,0082	**history**	0,00082
6.	*loss of consciousness*	0,00039	*pain*	0,00080	*pain*	0,00081
7.	*abd pain*	0,00036	**patient**	0,00080	**patient**	0,00080
8.	*numbness*	0,00036	**diarrhea**	0,00078	**diarrhea**	0,00078
9.	*home*	0,00035	**dysuria**	0,00074	**dysuria**	0,00074
10.	*sweat*	0,00035	**dyspnoea**	0,00065	**dyspnoea**	0,00066

TABLEAU 5.13 : Corpus Textes Cliniques, fenêtre restreinte : exemple de 10 premiers voisins obtenus pour le mot cible *cough*, avec la baseline, et après généralisation avec l'inclusion lexicale (AD/IL) et l'union des trois méthodes (AD/ALL3). Les termes soulignés sont les termes présents dans la référence.

Dans le tableau 5.13, nous présentons un exemple extrait du corpus Textes Cliniques. Il s'agit des dix premiers voisins obtenus pour le mot cible *cough* (toux), avec la baseline et les deux meilleures configurations de généralisation pour la fenêtre restreinte (AD/IL et AD/ALL3). Les voisins soulignés sont ceux présents dans l'UMLS (*pain*, *fever*, *history*,

diarrhea et *dyspnoea*). La généralisation avec l'inclusion lexicale (AD/IL) et l'union des trois méthodes (AD/ALL3) permet de mieux classer ces voisins, qui remontent alors dans le classement des dix premiers voisins, à l'exception de *pain* déjà présent avec la baseline. Les dix 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 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 observons très peu de différences entre la généralisation avec l'inclusion lexicale (AD/IL) et la généralisation à l'aide des trois méthodes (AD/ALL3). En effet, l'union des trois méthodes permet d'acquérir globalement les mêmes relations que celles obtenues avec l'inclusion lexicale. Cette généralisation (AD/IL) est la plus importante des généralisations réalisées.

5.2.2.1.2 Fenêtre graphique large

Lorsqu'une fenêtre graphique large est utilisée, on observe moins de variation dans les résultats. Dans les deux cas, la généralisation avec l'inclusion lexicale permet d'améliorer la qualité des résultats, avec toutes les métriques d'évaluation (MAP, R-précision et précision).

Corpus Menelas

Pour le corpus Menelas, l'utilisation individuelle des patrons lexico-syntaxiques pour généraliser les contextes permet d'améliorer le classement des voisins, en termes de MAP et de R-précision, en augmentant essentiellement le nombre de voisins corrects classés parmi les dix premiers voisins. Utilisés en combinaison avec l'inclusion lexicale (AD/PLS+IL), les patrons permettent d'améliorer la qualité des résultats obtenus avec la généralisation avec l'inclusion lexicale individuellement, en termes de MAP et de P@10. Pour ce corpus, les variantes n'ont que très peu, voire aucun impact sur la généralisation, aussi bien utilisées individuellement qu'en combinaison.

Textes Cliniques

Le comportement de la généralisation d'un côté avec les patrons, de l'autre avec les variantes, est différent pour les Textes Cliniques. Avec ce corpus, même si l'amélioration est faible, les variantes utilisées en combinaison avec l'inclusion lexicale (AD/VT+IL) permettent d'augmenter le nombre de voisins corrects retrouvés en rang 1 et parmi les cent premiers. En contrepartie, le nombre de relations retrouvées dans l'UMLS

	Corpus Textes Cliniques							
	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	**8 502 244**	**14 632**	**0,027**	**0,025**	**0,044**	**0,029**	**0,023**	**0,010**
AD/VT	7 949 956	14 234	0,028	0,024	0,040	0,029	0,024	0,010
AD/IL	**2 022 626**	*9 870*	**0,037**	**0,032**	**0,058**	**0,034**	**0,028**	**0,013**
AD/PLS	8 342 338	14 460	0,027	0,024	0,041	0,029	0,024	0,010
AD/VT+IL	**1 982 470**	*9 816*	**0,038**	**0,033**	**0,059**	0,034	**0,029**	**0,014**
AD/IL+PLS	2 017 020	9 774	0,037	0,031	0,057	0,034	0,028	0,013
AD/PLS+IL	2 022 640	9 872	0,037	0,032	0,058	0,034	0,028	0,013
AD/ALL3	**2 038 278**	*9 960*	0,037	0,032	**0,060**	**0,035**	**0,029**	0,013
	Corpus Menelas							
	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	**19 431**	**94**	**0,080**	**0,010**	**0**	**0,029**	**0,029**	**0,009**
AD/VT	18 585	94	0,082	0,010	0	0,029	0,029	0,009
AD/IL	**8 401**	*56*	**0,133**	**0,018**	**0,053**	**0,053**	**0,032**	**0,013**
AD/PLS	17 034	84	0,083	**0,027**	0	0,032	**0,036**	0,009
AD/VT+IL	8 402	56	0,133	0,018	0,053	0,053	0,032	0,013
AD/IL+PLS	**7 644**	*52*	**0,135**	0,018	0,053	0,053	**0,037**	0,013
AD/PLS+IL	8 270	58	0,129	0,017	0	0,050	0,030	0,014
AD/ALL3	8 552	58	0,121	0,017	0,050	0,040	0,035	0,014

TABLEAU 5.14 : Résultats obtenus pour les corpus de petite taille, avec la fenêtre large (W21) et l'indice de Jaccard pondéré.

diminue de 9 870 (AD/IL) à 9 816 (AD/VT+IL). Les patrons n'ont pas d'impact sur les résultats avec la fenêtre large et ce corpus.

Fusion des trois approches (AD/ALL3)

En ce qui concerne la fusion des trois approches, elle permet d'améliorer les résultats avec les deux corpus, par rapport à l'AD seule. Pour les Textes Cliniques, la fusion permet d'obtenir des résultats équivalents à ceux obtenus par la combinaison séquentielle offrant la meilleure précision (AD/VT+IL), avec également une P@5 plus élevée que celles obtenues avec les combinaisons séquentielles. Pour Menelas, même si cette fusion des trois approches améliore les résultats par rapport à la baseline, elle obtient par contre des résultats plus faibles que la combinaison ordonnée des méthodes qui offre la meilleure précision (AD/IL+PLS).

	Mot cible *headache*			
	Baseline		**AD/VT+IL**	
Rang	Voisin	Jacc	Voisin	Jacc
1.	*pain*	0,00037	**vomiting**	0,00058
2.	*dysuria*	0,00030	**cough**	0,00054
3.	*weakness*	0,00029	**fever**	0,00054
4.	*hospital*	0,00022	**chill**	0,00053
5.	*episode*	0,00019	**nausea**	0,00051
6.	*tachycardia*	0,00018	**diarrhea**	0,00050
7.	*work*	0,00017	*pain*	0,00045
8.	*diabetes mellitus*	0,00017	**patient**	0,00044
9.	*numbness*	0,00017	*dysuria*	0,00042
10.	*hospitalization*	0,00017	**symptom**	0,00038

TABLEAU 5.15 : Corpus Textes Cliniques, fenêtre large (W21) : exemple de 10 premiers voisins obtenus pour le mot cible *headache*, avec la baseline, et après généralisation avec les variantes et l'inclusion lexicale (AD/VT+IL).

Analyse des relations

Nous présentons dans le tableau 5.15 les dix premiers voisins obtenus pour le terme *headache*, à partir du corpus Textes Cliniques, avec la baseline et après généralisation en combinant l'inclusion lexicale et les variantes terminologiques (AD/VT+IL). Dans cet exemple, les dix premiers voisins obtenus après généralisation avec l'union des trois méthodes (AD/ALL3) sont les mêmes que ceux obtenus avec AD/VT+IL. Les voisins soulignés sont ceux présents dans la ressource. Ainsi, la généralisation permet de faire remonter dans le classement un très grand nombre de voisins présents dans la ressource pour le mot cible *headache*. Comme pour la fenêtre restreinte, les différences entre l'union des trois méthodes et la généralisation avec l'inclusion lexicale (seule ou combinée) sont faibles. En ce qui concerne la qualité des relations acquises, la baseline génère des groupements sémantiques moins homogènes que ceux obtenus avec la généralisation AD/VT+IL, avec notamment des termes tels que *work*, *hospital*, et *hospitalization*.

	Mot cible *coronaire droite*			
	Baseline		**AD/IL**	
Rang	Voisin	Jacc	Voisin	Jacc
1.	circonflexe	0,00450	**sténose**	0,00656
2.	artère	0,00297	**IVA**	0,00563
3.	patient	0,00292	circonflexe	0,00503
4.	*traitement médical*	0,00204	**lésion**	0,00454
5.	*infarctus*	0,00183	**coronarographie**	0,00392
6.	*lit d'aval*	0,00131	artère	0,00332
7.	*thallium*	0,00118	patient	0,00292
8.	*tabagisme*	0,00114	**traitement**	0,00275
9.	*ECG*	0,00094	**angioplastie**	0,00264
10.	*artère circonflexe*	0,00092	**angor**	0,00257

TABLEAU 5.16 : Corpus Menelas, fenêtre large (W21) : exemple de 10 premiers voisins obtenus pour le mot cible *coronaire droite*, avec la baseline, et après généralisation avec l'inclusion lexicale (AD/IL).

Comme lors de l'utilisation de la fenêtre restreinte, lorsque l'on considère le nombre de relations retournées par l'analyse distributionnelle, nous constatons que la généralisation avec l'inclusion lexicale permet de diviser le nombre de relations acquises par 2 avec les deux corpus, tout comme le nombre de relations identifiées dans l'UMLS.

Dans le tableau 5.16, nous présentons les dix premiers voisins obtenus pour le mot cible *coronaire droite*, dans le corpus Menelas, avec la baseline et après généralisation avec l'inclusion lexicale (AD/IL). Ce mot cible est absent de la référence, il n'est donc pas pris en compte lors de l'évaluation quantitative. Les dix premiers voisins de *coronaire droite* obtenus avec la baseline sont moins homogènes sémantiquement que ceux obtenus avec AD/IL. Les voisins obtenus avec la généralisation sont plus proches sémantiquement du mot cible *coronaire droite*, et correspondent à des co-hyponymes, des maladies, des examens et des traitements de *coronaire droite*.

5.2.2.1.3 Bilan

A travers les expériences menées sur les corpus de petite taille, il ressort qu'il est préférable d'utiliser avec ces corpus l'indice de Jaccard, avec une fenêtre restreinte. La généralisation avec l'inclusion lexicale permet une plus grande généralisation qu'avec les autres méthodes utilisées, et représente un bon moyen pour réduire la dispersion des données et améliorer les résultats avec les faibles fréquences [Périnet et Hamon, 2014]. Pour les plus petits corpus, le recours aux patrons lexico-syntaxiques après l'inclusion lexicale lors de la généralisation, permet d'améliorer le classement des voisins parmi les dix premiers voisins. Pour le plus volumineux des corpus de petite taille (environ 178 000 mots), c'est la combinaison des variantes terminologiques avec l'inclusion lexicale, qui améliore la qualité des résultats.

L'analyse manuelle des relations, bien que partiellement menée, révèle que pour les petits corpus, quelle que soit la taille de la fenêtre, généraliser les contextes distributionnels permet d'obtenir des groupements sémantiques plus homogènes, des voisins sémantiques sémantiquement plus proches du mot cible qu'avec une méthode distributionnelle sans généralisation.

5.2.2.2 Corpus de grande taille

Nous présentons maintenant les résultats obtenus sur nos deux corpus de plus grande taille, c'est-à-dire les corpus alimentaires décrits en section 4.1.2. Notons que si ces corpus sont pour nous les plus volumineux, ils sont tout de même de taille bien différente ; le corpus Recettes contient environ 4 millions de mots, alors que le corpus Guides Alimentaires contient environ 500 000 mots.

Pour l'évaluation des résultats, nous comparons les relations acquises avec l'analyse distributionnelle aux relations contenues dans une référence construite à partir de

trois ressources existantes : l'UMLS français, Agrovoc, et une ressource construite à partir de sites Web spécialisés (cf. section 4.2.1 pour une description de ces ressources). Pour le corpus Recettes, la référence contient 6 015 relations, et pour le corpus Guides alimentaires, la référence comprend 8 095 relations.

5.2.2.2.1 Fenêtre graphique restreinte

Corpus Guides Alimentaires

	ACQ	RESS	MAP		Rprec		P@1		P@5	
			COS	JAC	COS	JAC	COS	JAC	COS	JAC
Baseline	**1 275 949**	**1 552**	**0,018**	**0,027**	**0,010**	**0,027**	**0,007**	**0,054**	**0,005**	**0,039**
AD/VT	1 140 753	1 546	0,022	0,081	0,009	0,057	0,007	0,071	0,006	0,054
AD/IL	**752 135**	1 290	**0,030**	**0,092**	**0,015**	0,054	**0,012**	0,066	**0,007**	**0,059**
AD/PLS	929 695	929	0,012	0,058	0	0,042	0	0,061	0,002	0,041
AD/VT+IL	752 135	1 290	0,030	0,092	0,015	0,054	0,012	0,066	0,007	0,059
AD/IL+PLS	593 496	714	0,014	0,074	0	0,054	0	0,070	0,003	0,048
AD/PLS+IL	834 407	1 208	0,013	0,067	0,002	0,040	0,005	0,055	0,003	0,038
AD/ALL3	843 482	1 053	0,011	0,064	0	0,037	0	0,050	0,003	0,040

Corpus Recettes

	ACQ	RESS	MAP		Rprec		P@1		P@5	
			COS	JAC	COS	JAC	COS	JAC	COS	JAC
Baseline	**862 075**	**1 860**	**0,005**	**0,063**	**0,001**	**0,065**	**0,002**	**0,087**	**0**	**0,071**
AD/VT	811 309	1 837	0,006	0,056	0	0,041	0	0,049	0,001	0,037
AD/IL	450 502	1 682	**0,007**	**0,101**	0	**0,086**	0	**0,089**	0	0,068
AD/PLS	653 969	728	0,005	0,058	0	0,041	0	0,028	0	0,023
AD/VT+IL	450 551	1 682	0,007	0,101	0	0,086	0	0,089	0	0,068
AD/IL+PLS	303 160	660	**0,008**	0,079	**0,002**	0,046	0	0,053	**0,002**	0,035
AD/PLS+IL	423 741	1 166	0,008	0,081	0,001	0,052	0	0,068	0,001	0,048
AD/ALL3	434 510	1 224	0,007	**0,083**	0	0,055	0	0,067	0,001	0,051

TABLEAU 5.17 : Résultats obtenus pour les corpus de grande taille, avec la fenêtre restreinte (W5). Les valeurs en gras correspondent à la ou aux deux valeurs les plus élevées de chaque colonne.

Corpus Recettes

Pour le corpus Recettes, le plus volumineux de nos corpus, avec la fenêtre graphique restreinte, les résultats obtenus avec le Cosinus sont faibles, même après généralisation. Quelle que soit l'origine des relations utilisées, la P@1 est égale à zéro, et la R-précision et les autres précisions obtiennent des résultats compris entre 0 et 0,002. Avec l'indice de Jaccard, les résultats obtenus sur ce même corpus soulignent globalement un faible impact de la généralisation. L'augmentation de la MAP de 0,01 à 0,04 montre que la généralisation a un impact positif sur le classement des voisins. L'amélioration la plus importante est observée quand l'inclusion lexicale est utilisée individuellement pour généraliser les contextes. Cette généralisation permet également d'augmenter la R-précision et la P@1. Avec le calcul de précision, quel que soit le nombre de voisins pris en compte, la généralisation avec les patrons dégrade les résultats quelle que soit la configuration de généralisation. Il en est de même pour les variantes, sauf quand elles sont utilisées après l'inclusion lexicale (AD/VT+IL) : elles n'ont alors aucun impact.

Corpus Guides Alimentaires

Sur le corpus Guides Alimentaires et pour la fenêtre restreinte, la généralisation a un impact positif sur les résultats en termes de MAP, de R-précision, de P@1, de P@5 et de P@10 dans quasiment toutes les configurations de généralisation. L'inclusion lexicale utilisée individuellement permet d'améliorer la MAP, c'est-à-dire le classement des voisins, mais également la précision quand elle est calculée pour cinq, dix et cent voisins. La P@1 et la R-précision sont les plus élevées lorsque la généralisation est réalisée à l'aide des variantes morphosyntaxiques, mais l'écart avec l'inclusion lexicale est faible : 0,057 de R-précision pour les variantes, 0,054 pour l'inclusion, et 0,071 de P@1 pour les variantes, avec 0,066 pour l'inclusion. En revanche, si les variantes sont combinées à une autre méthode, leur impact sur la qualité des résultats est nul. Quant aux patrons, ils font baisser la qualité des résultats quand ils sont utilisés en combinaison avec une autre méthode (AD/PLS+IL, AD/IL+PLS, AD/PLS+VT, etc.). Enfin, on peut également remarquer qu'avec le Cosinus, pour le corpus Guides Alimentaires, l'ensemble des généralisations diminue la qualité des résultats, sauf si cette généralisation est réalisée avec l'inclusion lexicale utilisée individuellement.

Nombre de relations

En ce qui concerne le nombre de relations acquises, la généralisation avec l'inclusion lexicale permet de réduire le nombre de relations acquises qui se trouve alors divisé par deux, sans que le nombre de relations retrouvées dans la référence ne soit trop affecté (par rapport aux autres corpus) : 1 860 relations sont retrouvées dans la référence avec la baseline, 1 682 avec la généralisation avec inclusion lexicale. Les patrons combinés à l'inclusion lexicale divisent le nombre de relations acquises par 3, mais il en va de même pour les relations identifiées dans la référence. Avec le corpus Guides Alimentaires, la généralisation avec les patrons réduit également le nombre de relations acquises,

surtout quand ils sont utilisés en généralisation séquentielle après l'inclusion lexicale, où le nombre de relations est divisé par deux, comme le nombre de relations retrouvées dans la référence.

Analyse des résultats

Nous procédons à présent à l'analyse manuelle des résultats. Pour le corpus Recettes, nous présentons un exemple dans le tableau 5.18.

	Mot cible *courgette*			
	Baseline		**AD/IL**	
Rang	Voisin	Jacc	Voisin	Jacc
1.	*poivron*	0.00103	**tomate**	0.00193
2.	*légume*	0.00098	**oignon**	0.00185
3.	*rondelle*	0.00096	**pomme de terre**	0.00166
4.	*champignon*	0.00090	**tranche**	0.00144
5.	*viande*	0.00086	**carotte**	0.00141
6.	*temps*	0.00083	*poivron*	0.00140
7.	*cube*	0.00079	*légume*	0.00140
8.	*eau*	0.00079	*rondelle*	0.00134
9.	*minute*	0.00075	**pomme**	0.00131
10.	*plat*	0.00069	*champignon*	0.00126

TABLEAU 5.18: Corpus Recettes : exemple des 10 premiers voisins obtenus pour le mot cible *courgette*, avec la baseline, et après généralisation avec l'inclusion lexicale (AD/IL), avec l'indice de Jaccard et une fenêtre restreinte (W5). Les voisins qui remontent dans le classement sont en gras, ceux qui descendent en italique.

Il s'agit des dix premiers voisins sémantiques du mot cible *courgette*, obtenus avec la baseline et avec la généralisation à l'aide de l'inclusion lexicale. Avec ces deux configurations, pour ce mot cible, seul le voisin *légume* est retrouvé dans la référence. La généralisation permet globalement d'acquérir un ensemble de voisins plus homogène et cohérent autour du mot cible *courgette*. Les voisins obtenus avec l'AD/IL sont un ensemble de co-hyponymes, d'autres légumes (*tomate, oignon, pomme de terre*, etc.), l'hyperonyme *légume*, et deux termes liés au mot cible par une relation du domaine

(*rondelle* et *tranche*). En faisant remonter dans le classement ces termes proches sémantiquement, la généralisation permet d'écarter des voisins tels que *minute*, *temps*, *plat*, plus en lien avec la recette en elle-même que sémantiquement proche de *courgette*.

	Mot cible *anchois*			**Mot cible *grossesse***			
	Baseline	**AD/IL**		**Baseline**		**AD/IL**	
Rang	Voisin	Voisin	Jacc	Voisin	Jacc	Voisin	Jacc
1.	Pas de voisin	**hareng**	0,00052	mère	0,00040	mère	0,00051
2.		**maquereau**	0,00050	enfant	0,00040	**âge**	0,00051
3.		**thon**	0,00015	*obésité*	0,00033	*enfant*	0,00051
4.		**truite**	0,00011	*année*	0,00033	**risque**	0,00046
5.		**saumon**	0,00008	*traitement*	0,00031	**maladie**	0,00045
6.				*activité physique*	0,00031	**alimentation**	0,00043
7.				*activité*	0,00030	**femme**	0,00043
8.				*surpoids*	0,00030	**poids**	0,00042
9.				*fer*	0,00029	*obesité*	0,00042
10.				*questionnaire*	0,00029	*activité physique*	0,00041

TABLEAU 5.19 : Corpus Guides Alimentaires : exemple des 10 premiers voisins obtenus pour les mots cibles *grossesse* et *anchois* pour la baseline, et après généralisation avec l'inclusion lexicale (AD/IL), avec l'indice de Jaccard et une fenêtre restreinte (W5). Les voisins qui remontent dans le classement sont en gras, ceux qui descendent en italique.

Pour le corpus Guides Alimentaires, un exemple est présenté dans le tableau 5.19, pour la fenêtre restreinte, avec également la baseline et la généralisation avec inclusion lexicale. Nous présentons les résultats obtenus pour deux mots cibles : *anchois* et *grossesse*, le premier étant absent de la référence, et le second présent mais pour lequel notre méthode ne retrouve aucun voisin de la référence.
Pour *anchois*, quand la baseline est utilisée, le terme n'apparait pas dans les mots cibles en sortie, certainement écarté par l'utilisation des seuils. Ces seuils sont les mêmes pour les différentes généralisations, car ils ont été définis d'après la baseline. Cependant, une fois les contextes généralisés avec l'inclusion lexicale, le mot cible *anchois* apparait dans les résultats et est caractérisé par un cluster homogène, contenant cinq voisins co-hyponymes : *hareng*, *maquereau*, *thon*, *truite* et *saumon*.
Pour le mot cible *grossesse*, la généralisation avec l'inclusion lexicale permet d'obtenir également un cluster plus homogène, en faisant remonter dans le classement les termes

liés à la grossesse, tels que *femme*, *poids* et *alimentation*, et en faisant descendre les voisins sémantiquement moins proches tels qu'*obésité*, *activité* et *surpoids*.

5.2.2.2.2 Fenêtre graphique large

Nous présentons dans le tableau 5.20 les résultats obtenus pour les corpus de grande taille avec la fenêtre large. Nous ne reportons pas dans ce tableau les résultats obtenus en termes de P@10 et P@100, car la généralisation dégrade l'ensemble des résultats dès qu'un plus grand nombre de voisins est pris en compte dans l'évaluation.

Ainsi, avec la fenêtre graphique large, pour nos deux corpus de grande taille, c'est-à-dire le cas où les fréquences sont les plus élevées, avec l'indice de Jaccard, la généralisation a un impact nul ou négatif sur toutes les précisions, quelle que soit la ou les méthodes utilisée(s) pour la généralisation. En revanche, avec cette mesure de similarité, la généralisation améliore le classement des voisins, en termes de MAP pour le corpus Guides Alimentaires, et en termes de MAP et de R-précision pour les Recettes. Pour les Guides Alimentaires, l'impact est même quasiment nul, puisque la MAP augmente de 0,001 quand la généralisation est réalisée avec la combinaison de l'inclusion lexicale et des patrons (AD/IL+PLS). Pour le corpus Recettes, la MAP est améliorée avec toutes les configurations de généralisation, mais l'impact le plus important est réalisé avec l'inclusion lexicale. Quant à la R-précision, elle est améliorée uniquement quand la généralisation est réalisée avec l'inclusion lexicale.

Avec le Cosinus, les résultats sont de meilleure qualité qu'avec la fenêtre restreinte, même si la différence est faible, contrairement aux petits corpus. Avec le corpus Guides Alimentaires, la généralisation des contextes améliore les résultats avec toutes les mesures d'évaluation, et pour cinq voisins maximum pris en compte dans le calcul de la précision, lorsqu'elle est réalisée avec la combinaison inclusion lexicale et patrons (AD/IL+PLS). Le Cosinus offre ainsi des résultats de meilleure qualité avec la fenêtre graphique large, et permet quand il est utilisé avec l'inclusion lexicale d'obtenir une P@1 plus élevée qu'avec l'indice de Jaccard. Les variantes et les patrons utilisés indépendamment améliorent la P@5. Et les variantes utilisées en combinaison ont un impact nul sur la qualité des résultats. Pour les Recettes, les patrons utilisés seuls dans la généralisation des contextes permettent d'améliorer les résultats avec toutes les métriques d'évaluation, avec également jusqu'à cinq voisins pris en compte dans la précision. Cependant, en contrepartie, le nombre de relations identifiées dans la référence est divisé par deux. La généralisation avec l'inclusion lexicale permet de réduire d'un tiers le nombre de relations acquises, comme avec les patrons, sans que le nombre de relations identifiées dans la référence ne soit trop pénalisé (1 350 relations avec la *baseline*, 1 299 avec l'inclusion lexicale). Cette généralisation permet d'augmenter la MAP, la P@1 et la P@5. L'amélioration de la qualité est moins importante qu'avec les patrons, mais le nombre de relations correctes est moins pénalisé.

Corpus Guides Alimentaires

	ACQ	RESS	MAP		Rprec		P@1		P@5	
			COS	JAC	COS	JAC	COS	JAC	COS	JAC
Baseline	**419 336**	**1 210**	**0,025**	**0,081**	**0,011**	**0,054**	**0,005**	**0,094**	**0,004**	**0,060**
AD/VT	418 049	1 210	0,014	0,079	0,003	0,048	0,005	0,089	0,006	0,060
AD/IL	369 017	1 052	0,011	0,081	0,004	0,050	**0,007**	0,085	0,005	0,059
AD/PLS	464 365	740	0,024	0,079	0,010	0,051	0	0,064	**0,007**	0,049
AD/VT+IL	369 047	1 052	0,011	0,081	0,004	0,050	0	0,085	0,005	0,059
AD/IL+PLS	360 655	593	**0,027**	**0,082**	**0,016**	0,052	**0,024**	0,056	**0,008**	0,051
AD/IL+PLS	597 667	1 116	0,013	0,069	0,002	0,041	0	0,055	0,003	0,044
AD/ALL3	595 209	1 145	0,012	0,077	0,001	0,051	0	0,063	0,004	0,046

Corpus Recettes

	ACQ	RESS	MAP		Rprec		P@1		P@5	
			COS	JAC	COS	JAC	COS	JAC	COS	JAC
Baseline	**347 938**	**1 350**	**0,011**	**0,047**	**0,007**	**0,045**	**0,003**	**0,068**	**0,005**	**0,056**
AD/VT	347 708	1 350	0,016	0,070	0,004	0,051	0	0,044	0,004	0,050
AD/IL	**232 405**	1 299	**0,022**	**0,071**	0,006	**0,055**	0,006	0,057	**0,008**	0,046
AD/PLS	277 695	571	0,023	0,064	**0,014**	0,022	**0,034**	0,011	0,007	0,043
AD/VT+IL	232 082	1 299	0,022	0,071	0,006	0,055	0,006	0,057	0,008	0,046
AD/IL+PLS	152 318	517	0,016	0,056	0,007	0,012	0	0,012	0,005	0,032
AD/PLS+IL	248 908	1 049	0,015	0,066	0,004	0,039	0,007	0,036	0,003	0,036
AD/ALL3	254 386	1 071	0,016	0,069	0,010	0,043	0,007	0,036	0,004	0,043

TABLEAU 5.20 : Résultats obtenus pour les corpus de grande taille, avec la fenêtre large (W21). Les résultats en gras correspondent à la ou aux deux valeurs les plus élevées de chaque colonne.

Analyse des résultats

Le tableau 5.21 rassemble les dix premiers voisins obtenus pour le mot cible *palette*, dans le corpus Guides Alimentaires. Dans cet exemple, les voisins obtenus avec la baseline sont sémantiquement homogènes, et entrent dans la classe sémantique de la *viande*, mais uniquement trois voisins sont obtenus. En revanche, la généralisation permet d'étendre le nombre de voisins obtenus, tout en conservant une cohérence sémantique ; les dix premiers voisins obtenus pour *palette* sont des termes désignant différents types de viande et des produits laitiers.

	Mot cible *palette*			
	Baseline		**AD/PLS+IL**	
Rang	Voisin	COS	Voisin	COS
1.	*côte de porc*	0,99999	**grillade**	0,99688
2.	**grillade**	0,99982	**escalope de veau**	0,98108
3.	andouillette	0,99978	andouillette	0,97293
4.			**poisson**	0,93832
5.			*côte de porc*	0,93580
6.			**viande**	0,92294
7.			**lapin**	0,91918
8.			**fromage**	0,91383
9.			**volaille**	0,90658
10.			**lait**	0,90148

TABLEAU 5.21 : Corpus Guides Alimentaires, fenêtre large : exemple de 10 premiers voisins obtenus pour le mot cible *palette*, avec la baseline, et après généralisation avec les patrons et l'inclusion lexicale (AD/PLS+IL), avec le Cosinus.

5.2.2.3 Bilan sur la généralisation des contextes

Suite à l'analyse des résultats, nous faisons à présent le bilan sur la généralisation des contextes distributionnels. Généralement, la généralisation permet de diviser par deux le nombre de relations acquises. De plus, la généralisation des contextes distributionnels a un comportement différent en fonction de la taille du corpus et en fonction de la taille de la fenêtre graphique.

Ainsi, il ressort de ces expériences qu'avec les plus petits corpus, (de l'ordre de 100 000 mots), il est préférable d'utiliser une fenêtre graphique restreinte et l'indice de Jaccard. Le Cosinus ne semble pas adapté aux corpus de petite taille. Enfin, l'utilisation de la généralisation à préférer pour ce type de paramètres est celle basée sur l'inclusion lexicale. Pour le très petit corpus, les variantes terminologiques pour la généralisation permettent d'améliorer le classement des voisins distributionnels.

Lorsqu'elle est appliquée à des corpus de spécialité plus importants, d'une taille de l'ordre du million de mots (entre 500 000 et 4 millions de mots), la généralisation a un impact plus important quand le Cosinus est utilisé, mais les résultats obtenus avec cette mesure de similarité restent inférieurs à ceux obtenus avec l'indice de Jaccard. Généralement, il est préférable d'utiliser la fenêtre restreinte, l'Indice de Jaccard et de généraliser avec l'inclusion lexicale. Avec le corpus moyen (un million de mots), l'utilisation de la fenêtre large avec la généralisation réalisée à l'aide des patrons lexico-syntaxiques permet d'obtenir des résultats comparables à la configuration précédemment citée (W5 + AD/IL).

Généralement, sur tous les corpus, lorsque les configurations les plus adaptées sont prises en compte, la généralisation des contextes distributionnels permet l'amélioration quantitative et qualitative des regroupements sémantiques. Ceux-ci sont alors plus homogènes et plus pertinents. Si avec l'analyse distributionnelle le type de relations acquises est varié, la généralisation permet d'acquérir principalement des co-hyponymes.

5.2.3 Normalisation des contextes distributionnels

Nous présentons dans cette section les résultats obtenus lors de la normalisation des contextes, dans un premier temps réalisée seule (section 5.2.3.1), puis la normalisation réalisée en combinaison avec la généralisation (section 5.2.3.2).

5.2.3.1 Normalisation des contextes

En général, sur nos quatre corpus, la normalisation des contextes a très peu d'impact sur la qualité des résultats obtenus. Pour les quatre corpus, la normalisation permet une faible réduction du nombre de relations acquises, mais également du nombre de relations identifiées dans la référence.

Afin d'analyser l'impact de la normalisation des contextes sur les résultats de l'analyse distributionnelle, nous procédons comme pour la généralisation à une comparaison des résultats obtenus après normalisation à ceux obtenus avec la baseline, l'AD seule. De plus, les résultats obtenus après normalisation sont comparés avec les résultats obtenus après généralisation. Comme précédemment, nous décrivons dans un premier temps les résultats obtenus pour les corpus de petite taille (Menelas et Textes Cliniques),

puis les résultats obtenus pour les corpus de plus grande taille (Recettes et Guides Alimentaires).

	Corpus Menelas							
	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	8 118	60	0,121	0,079	0,053	0,042	0,037	0,013
AD/SYN	**7 058**	*56*	**0,165**	**0,119**	**0,095**	**0,048**	**0,038**	**0,014**
AD/VT	**6 915**	56	0,134	0,083	0,056	**0,056**	**0,039**	0,013
	Corpus Textes Cliniques							
	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	2 722 289	13 372	0,045	0,034	0,054	0,041	0,033	0,014
AD/SYN	**1 126 829**	12 508	**0,046**	**0,036**	**0,060**	0,040	0,033	0,013
AD/VT	2 220 105	12 176	**0,047**	**0,037**	0,060	0,040	0,033	0,013

TABLEAU 5.22: Résultats obtenus séparément après normalisation des contextes et après généralisation avec les variantes terminologiques pour les corpus de petite taille, avec l'indice de Jaccard pondéré et la fenêtre restreinte (W5).

5.2.3.1.1 Corpus de petite taille

Nous décrivons tout d'abord le comportement de la normalisation des contextes avec les corpus Menelas et les Textes Cliniques (cf. tableau 5.22).

Le comportement de la normalisation est comparable pour le corpus Menelas et le corpus Textes Cliniques. Quand le Cosinus est utilisé, la normalisation a soit un impact négatif soit un impact nul sur la qualité des résultats, avec les deux tailles de fenêtres. Avec l'indice de Jaccard pondéré, quand la fenêtre large est utilisée, la normalisation n'a également aucun impact sur la qualité des résultats. En revanche, avec la fenêtre restreinte, l'impact de la normalisation sur la qualité de résultats est positif, quelle que soit la métrique utilisée, avec notamment une augmentation de la MAP de 0,044 pour Menelas et proche de 0,001 pour les Textes Cliniques, et une augmentation de la P@1 de 0,042 et de 0,006 respectivement. La normalisation a un impact plus important quand le corpus est de plus petite taille.

Si l'on compare ces résultats positifs obtenus avec la fenêtre restreinte et l'indice de Jaccard à ceux obtenus avec les différentes configurations de généralisation et les mêmes

paramètres, nous remarquons que la normalisation a un comportement comparable à la généralisation réalisée avec les variantes terminologiques.

Analyse des relations

Nous avons procédé à l'analyse manuelle des relations obtenues après la normalisation des contextes. Etant donné le très grand nombre de relations obtenues, il est difficile de réaliser une analyse manuelle de l'ensemble des relations. Nous avons donc analysé et comparé les dix premiers voisins obtenus après normalisation aux dix premiers voisins obtenus avec la baseline pour un sous-ensemble de mots cibles. Nous présentons un exemple dans le tableau 5.23, pour le mot cible *électrocardiogramme*, pour lequel nous observons le plus de variations entre la normalisation et la généralisation avec les variantes (AD/VT). Dans cet exemple, nous présentons les dix premiers voisins obtenus pour ce mot cible, avec la baseline, après normalisation (AD/SYN) et après généralisation avec les variantes (AD/VT). Globalement nous pouvons constater que les voisins classés parmi les dix premiers varient peu, ou tout du moins varient moins quand la normalisation est utilisée. La normalisation modifie le classement des dix premiers voisins, sans trop faire remonter de voisins pertinents classés plus bas. Ainsi, la normalisation fait apparaître dans les dix premiers les voisins *cholestérol total* et *cholestérol*, et en contre-partie fait sortir des dix premiers voisins *hypertension* et *hypertrophie ventriculaire gauche*.

L'impact sur le classement des voisins est donc nettement moins important avec la normalisation qu'après généralisation des contextes réalisée avec l'inclusion lexicale, AD/IL (cf. section 5.2.2.1). En revanche, nous observons que la normalisation a globalement un impact similaire à la généralisation réalisée avec les variantes terminologiques (AD/VT) : les voisins obtenus sont généralement les mêmes et les valeurs de similarité identiques.

5.2.3.1.2 Corpus de grande taille

Nous décrivons à présent l'impact de la normalisation sur les corpus de plus grande taille, le corpus Recettes et les Guides Alimentaires.

Comme pour les corpus de petite taille, quand le Cosinus est utilisé, la normalisation a un impact nul ou négatif sur la qualité des résultats obtenus avec les deux corpus.

Avec l'indice de Jaccard, les résultats varient selon le corpus et la taille de la fenêtre. Sur le plus petit corpus, les Guides Alimentaires, le comportement de la normalisation est semblable à celui observé avec les corpus de petite taille : la normalisation a un impact positif sur la qualité des résultats, quelle que soit la métrique d'évaluation utilisée. Ainsi, la MAP augmente de 0,065 et la P@1 de 0,017. Avec le corpus le plus volumineux, le corpus Recettes, l'impact varie en fonction de la taille de la fenêtre :

	Mot cible *électrocardiogramme*		
	Baseline	**AD/SYN**	**AD/VT**
Rang	Voisins	Voisins	Voisins
1.	*rythme régulier sinusal*	**cholestérol total**	**cholestérol total**
2.	rythme sinusal	*rythme régulier sinusal*	rythme sinusal
3.	*examen clinique*	*rythme sinusal*	**surcharge pondérale**
4.	surcharge pondérale	surcharge pondérale	*examen clinique*
5.	ECG	*examen clinique*	ECG
6.	bilan lipidique	*ECG*	bilan lipidique
7.	triglycéride	*bilan lipidique*	triglycéride
8.	*hypertrophie ventriculaire gauche*	*triglycéride*	**cholestérol**
9.	hypertension	**cholestérol**	**cinétique ventriculaire gauche**
10.	artère coronaire	artère coronaire	*hypertrophie ventriculaire gauche*

TABLEAU 5.23 : Corpus Menelas, fenêtre restreinte (W5) : exemple de 10 premiers voisins obtenus pour le mot cible *électrocardiogramme*, avec la baseline, et séparément après normalisation avec les synonymes (AD/SYN) et après généralisation avec les variantes (AD/VT), avec l'indice de Jaccard.

l'impact est nul avec la fenêtre restreinte, mais positif avec la fenêtre large, en termes de MAP et de R-précision, avec une augmentation respective de 0,023 et de 0,006.

Corpus Guides Alimentaires - fenêtre restreinte (W5)

	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	2 551 882	1 552	0,027	0,027	0,054	0,039	0,030	0,008
AD/SYN	**2 253 642**	*1 256*	**0,082**	**0,057**	**0,071**	**0,055**	**0,039**	**0,010**
AD/VT	2 281 490	1 546	0,081	0,057	0,071	0,054	0,039	0,010

Corpus Recettes - fenêtre large (W21)

	REL ACQ	RESS	MAP	Rprec	P@1	P@5	P@10	P@100
Baseline	347 938	1 350	0,047	0,045	0,068	0,056	0,044	0,018
AD/SYN	347 872	1 350	**0,070**	**0,051**	0,044	0,050	0,043	0,017
AD/VT	347 704	1 350	0,070	0,051	0,044	0,050	0,043	0,017

TABLEAU 5.24 : Résultats obtenus pour la normalisation des contextes avec les synonymes (AD/SYN) et pour la généralisation avec les variantes terminologiques (AD/VT) pour les corpus de grande taille, avec la fenêtre restreinte (W5) pour les Guides Alimentaires et la fenêtre large (W21) pour les Recettes, avec l'indice de Jaccard pondéré.

Analyse des relations

Avec les corpus volumineux, comme pour les corpus de petite taille, les relations obtenues après normalisation sont proches de celles obtenues avec la baseline (l'AD seule). Nous présentons dans le tableau 5.25 l'exemple du mot cible *sucre glace*, présent dans le corpus Recettes, avec les dix premiers voisins obtenus avec la baseline et après normalisation. Globalement la normalisation a un impact positif sur le classement des voisins. Dans l'exemple, huit sur dix voisins restent les mêmes après normalisation, seul le classement des derniers voisins est légèrement altéré.

On observe que la normalisation réalisée sur les corpus de grande taille n'a qu'un très faible voire aucun impact sur les relations acquises, comme le montre l'exemple 5.25.

5.2.3.2 Normalisation combinée à la généralisation

Afin d'évaluer les résultats obtenus avec la combinaison de la normalisation et de la généralisation, nous comparons ces résultats avec ceux obtenus non seulement avec la baseline, mais surtout avec les résultats obtenus avec la généralisation seule. L'objectif

	Mot cible *sucre glace*			
	Baseline		**AD/SYN**	
Rang	Voisins	Jacc	Voisins	Jacc
1.	sucre	0,00051	sucre	0,00060
2.	chocolat	0,00046	chocolat	0,00054
3.	crème	0,00045	crème	0,00054
4.	pâte	0,00045	pâte	0,00054
5.	mélange	0,00045	mélange	0,00053
6.	beurre	0,00044	œuf	0,00053
7.	préparation	0,00044	farine	0,00053
8.	sel	0,00042	beurre	0,00052
9.	four	0,00042	préparation	0,00052
10.	saladier	0,00042	sel	0,00050

TABLEAU 5.25 : Corpus Recettes, fenêtre large (W21) ; exemple de 10 premiers voisins obtenus pour le mot cible *sucre glace*, avec la baseline, et après normalisation avec les synonymes (AD/SYN), avec l'indice de Jaccard.

est de déterminer s'il est utile de normaliser les contextes avant de les généraliser, et quelles sont les configurations les plus adaptées. Nous analysons les résultats pour les corpus de petite taille (Menelas et les Textes Cliniques) dans un premier temps, puis ceux obtenus pour les corpus de grande taille (Recettes et Guides Alimentaires).

	Corpus Textes Cliniques - fenêtre large (W21)							
	MAP		Rprec		P@1		P@5	
Baseline	**0,027**		**0,025**		**0,044**		**0,029**	
	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**
VT	0,028	0,028	0,024	**0,025**	0,040	**0,043**	0,029	0,029
IL	0,037	0,037	0,032	0,032	0,058	**0,068**	0,034	0,034
PLS	0,027	0,027	0,024	0,024	0,041	**0,045**	0,029	0,029
VT+IL	0,038	0,038	0,033	0,033	0,059	**0,069**	0,034	0,034
IL+PLS	0,037	0,037	0,031	0,031	0,057	**0,067**	0,034	0,034
PLS+IL	0,037	0,037	0,032	0,032	0,058	**0,068**	0,034	0,034
ALL3	0,037	**0,038**	0,032	0,032	0,060	**0,068**	0,035	0,035

TABLEAU 5.26 : Textes Cliniques : résultats obtenus pour la généralisation réalisée à partir de la baseline, l'AD seule (AD), et pour la généralisation réalisée après normalisation des contextes à l'aide des synonymes (ADSYN). Résultats obtenus avec la fenêtre large (W21) avec l'indice de Jaccard pondéré.

5.2.3.2.1 Corpus de petite taille

Avec les corpus Menelas et les Textes Cliniques, et les deux tailles de fenêtre, quand le Cosinus est utilisé, l'impact de la normalisation avant la généralisation est nul ou négatif.

Avec l'indice de Jaccard, les résultats diffèrent selon le corpus et la taille de fenêtre. Nous présentons dans un premier temps les résultats obtenus pour les Textes Cliniques, puis ceux obtenus pour le corpus Menelas.

Les résultats obtenus avec l'indice de Jaccard sont présentés dans les tableaux 5.26 et 5.27. Nous nous intéressons aux résultats obtenus avec la combinaison de la normalisation (ADSYN) et de la généralisation, et nous comparons ces résultats à la fois à ceux obtenus après généralisation des contextes (AD) et à ceux obtenus avec la baseline, l'analyse distributionnelle (AD seule), comme précédemment. Les résultats en gras

signifient que ces résultats sont supérieurs à la généralisation seule (sans normalisation). Lorsqu'ils sont également soulignés, cela signifie qu'ils sont également supérieurs à la baseline (l'AD seule).

Textes Cliniques

Pour les Textes Cliniques, avec la fenêtre restreinte, l'utilisation de la normalisation avant la généralisation n'a pas d'impact sur les résultats. En revanche, quand la fenêtre large est utilisée (cf. tableau 5.26), la généralisation des contextes réalisée après normalisation permet d'augmenter la P@1 quelle que soit la configuration de généralisation, par rapport à la généralisation réalisée seule (sans normalisation). L'impact le plus important de cette combinaison est constaté quand la généralisation est réalisée avec l'inclusion lexicale (ADSYN/IL). En effet, la combinaison de la généralisation et de la normalisation (ADSYN/IL) permet d'augmenter la P@1 de 0,010, par rapport à la généralisation seule (AD/IL).

Corpus Menelas

	Corpus Menelas - fenêtre restreinte (W5)							
	MAP		Rprec		P@1		P@5	
Baseline	**0,121**		**0,079**		**0,053**		**0,042**	
	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**
VT	0,134	**0,166**	0,083	**0,119**	0,056	**0,095**	0,056	**0,057**
IL	0,125	**0,182**	0,071	**0,125**	0,071	**0,125**	0,014	**0,025**
PLS	0,126	**0,169**	0,063	**0,111**	0,063	**0,111**	0,038	**0,044**
VT+IL	0,125	**0,182**	0,071	**0,125**	0,071	**0,125**	0,014	**0,025**
IL+PLS	0,133	**0,147**	0,077	**0,071**	0,077	0,071	0,031	0,043
PLS+IL	0,125	**0,193**	0,071	**0,125**	0,071	**0,125**	0,014	**0,038**
ALL3	0,126	**0,193**	0,071	**0,125**	0,071	**0,125**	0,014	**0,038**

TABLEAU 5.27: Corpus Menelas : résultats obtenus pour la généralisation réalisée à partir de la baseline, l'AD seule (AD), et pour la généralisation réalisée après normalisation des contextes à l'aide des synonymes (ADSYN). Résultats obtenus avec la fenêtre restreinte (W5) avec l'indice de Jaccard pondéré.

Pour le corpus Menelas, nous présentons dans le tableau 5.27 les résultats obtenus avec la fenêtre restreinte (W5). En effet, quand la fenêtre restreinte est utilisée en combinaison avec l'indice de Jaccard, pour ce corpus, l'impact de la combinaison de la normalisation et de la généralisation est le plus important des quatre corpus.

Avec quasiment toutes les configurations de généralisation, la normalisation réalisée en amont de la généralisation améliore la qualité des résultats en termes de MAP, de R-précision, et pour les précisions prenant en compte jusqu'à 10 voisins dans l'évaluation. La qualité des résultats est plus élevée que le baseline mais également que la généralisation seule. L'impact le plus élevé est constaté quand la normalisation précède la généralisation avec inclusion lexicale (ADSYN/IL). Ainsi, par rapport à la baseline, la MAP augmente de 0,061 (alors que l'augmentation est de 0,013 uniquement quand la généralisation n'est pas précédée de normalisation) et la P@1 est multipliée par deux, en passant de 0,053 à 0,125.

Analyse des résultats

Dans le tableau 5.28, nous présentons l'exemple du mot cible *angioplastie* et des dix premiers voisins. La combinaison de la normalisation et de la généralisation par inclusion lexicale améliore essentiellement la qualité des relations acquises. Ainsi, pour *angioplastie*, les dix premiers voisins forment un groupement sémantique plus cohérent après normalisation et généralisation qu'avec la baseline. En effet, parmi les dix premiers, les voisins sont tous des termes plus techniques et plus proches sémantiquement du mot cible. Ainsi, nous observons l'apparition de co-hyponyme *coronarographie*, synonyme *dilatation*, et de relations du domaine pour *angioplastie* : les conséquences (la *revascularisation*), la cause (*sténose*), où elle est réalisée (*coronaire droite*). En revanche, pour la baseline, les voisins obtenus sont moins proches sémantiquement du mot cible.

5.2.3.2.2 Corpus de grande taille

Comme pour les petits corpus, la combinaison de la normalisation et de la généralisation a un impact nul ou négatif sur les deux corpus de grande taille (les Recettes et les Guides Alimentaires), quand le Cosinus est utilisé, quelle que soit la taille de la fenêtre, par rapport à la baseline.

Sur le corpus Recettes, le plus volumineux de nos quatre corpus, la combinaison de la normalisation et de la généralisation n'a aucun impact sur les résultats qui restent identiques à ceux obtenus avec la généralisation seule, et ceci avec toutes les configurations de généralisation (voir la sous section 5.2.2.2 pour l'analyse des résultats avec la généralisation seule).

	Mot cible *angioplastie*	
	Baseline	**ADSYN/IL**
Rang	Voisins	Voisins
1.	*patient*	**coronarographie**
2.	*tabagisme*	**sténose**
3.	*bilan angio-coronarographique*	**lésion**
4.	*récidive douloureuse*	**IVA**
5.	*fibrinolyse*	**coronaire droite**
6.	*thallium*	*patient*
7.	*réseau circonflexe*	**circonflexe**
8.	*facteur de risque*	**revascularisation**
9.	*athérome*	**dilatation**
10.	*docteur*	**dissection**

TABLEAU 5.28 : Corpus Menelas, fenêtre restreinte (W5) ; exemple de 10 premiers voisins obtenus pour le mot cible *angioplastie*, avec la baseline, et après normalisation et généralisation par inclusion lexicale, avec l'indice de Jaccard.

	Corpus Guides Alimentaires - Fenêtre restreinte (W5)							
	MAP		Rprec		P@1		P@5	
Baseline	**0,027**	**0,027**	**0,027**	**0,027**	**0,054**	**0,054**	**0,039**	**0,039**
	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**	AD	**ADSYN**
VT	0,081	**0,082**	0,057	0,057	0,071	0,071	0,054	**0,055**
IL	0,092	0,089	0,054	0,050	0,066	0,062	0,059	0,058
PLS	0,058	**0,060**	0,042	0,042	0,061	0,060	0,041	0,042
VT+IL	0,092	0,089	0,054	0,050	0,066	0,062	0,059	0,058
IL+PLS	0,074	**0,075**	0,054	**0,056**	0,070	0,070	0,048	0,048
PLS+VT	0,057	**0,060**	0,042	0,042	0,060	0,060	0,041	**0,042**
ALL3	0,064	0,063	0,037	0,036	0,050	0,050	0,040	0,039

TABLEAU 5.29 : Résultats obtenus pour le corpus Guides Alimentaires : généralisation réalisée à partir de la baseline, l'AD seule (AD), et généralisation réalisée après normalisation des contextes à l'aide des synonymes (ADSYN). Résultats obtenus avec la fenêtre restreinte (W5) et l'indice de Jaccard pondéré.

Pour le corpus Guides Alimentaires, avec l'indice de Jaccard, quand la fenêtre large est utilisée, l'impact de la combinaison de la généralisation et de la normalisation est nul ou très faiblement supérieur à la généralisation seule (une augmentation de 0,001), mais les résultats obtenus restent dans tous les cas inférieurs à la baseline. En revanche, avec la fenêtre restreinte, la normalisation a un impact positif sur les résultats obtenus. Ces résultats sont présentés dans le tableau 5.29.

Ainsi, l'impact de la combinaison de la normalisation et de la généralisation est positif quand la généralisation est réalisée avec les patrons lexico-syntaxiques, soit seuls (AD/PLS) soit combinés avec les variantes terminologiques (AD/PLS+VT). Cette combinaison a un impact sur le classement des voisins, avec une augmentation de la MAP de 0,002 et 0,003 respectivement, par rapport à la généralisation seule, et de 0,033 et 0,048 respectivement par rapport à la baseline. La R-précision obtenue pour l'AD/PLS+VT est également augmentée de 0,002 par rapport à la généralisation seule (soit une augmentation de 0,029 par rapport à la baseline). Pour les autres configurations de généralisation, l'impact de la normalisation par rapport à la généralisation seule est généralement nul, mais les résultats restent globalement supérieurs à la baseline. A l'exception des valeurs de précision à un et cinq voisins (P@1 et P@5) obtenues avec la normalisation et la généralisation réalisée avec l'union des trois méthodes (ADSYN/ALL3) qui sont en baisse par rapport à la baseline (-0,004 pour la P@1 et -0,001 pour la P@5).

	Mot cible *état nutritionnel*		
	Baseline	**AD/VT+PLS**	**ADSYN/VT+PLS**
Rang	Voisins	Voisins	Voisins
1.	nutrition	alimentation	alimentation
2.	santé	santé	santé
3.	alimentation	enquête	enquête
4.	habitude alimentaire	nutrition	état de santé
5.	cholestérol	enfant	nutrition
6.	moyen	état de santé	enfant
7.	pratique	moyen	moyen
8.	équilibre	but	but
9.	état de santé	comportement	habitude alimentaire
10.	médicament	habitude alimentaire	comportement

TABLEAU 5.30 : Corpus Guides Alimentaires, fenêtre restreinte (W5) : exemple de 10 premiers voisins obtenus pour le mot cible *état nutritionnel*, avec la baseline, et après normalisation et généralisation, avec l'indice de Jaccard.

Analyse des relations obtenues

La normalisation combinée a essentiellement un impact sur le classement des voisins sémantiques. L'analyse manuelle vient confirmer ce que montrent les résultats obtenus en termes de MAP sur le corpus Guides Alimentaires. Nous présentons dans le tableau 5.30, les dix premiers voisins obtenus pour le mot cible *état nutritionnel*, avec la baseline, après généralisation avec les variantes et les patrons (AD/VT+PLS), mais aussi après normalisation et généralisation (ADSYN/VT+PLS). La plus importante modification des voisins est observée lorsque la généralisation des contextes est réalisée (dans l'exemple, 4 des 10 voisins diffèrent entre la baseline et l'AD/VT+PLS). En revanche, la normalisation n'influence que très légèrement le classement, sans modifier suffisamment les dix premiers voisins qui restent identiques entre la généralisation seule (AD/VT+PLS) et la généralisation combinée avec la normalisation (ADSYN/VT+PLS). La normalisation a tout de même un léger impact, mais faible, qui permet de faire remonter dans le classement les termes *état de santé* et *habitude alimentaire*.

5.2.3.2.3 Bilan

Ainsi, la combinaison de la généralisation et de la normalisation offre de meilleurs résultats quand les fréquences sont très faibles ; avec la fenêtre restreinte et corpus de plus petite taille.

Pour les corpus de grande taille, avec l'indice de Jaccard et la fenêtre restreinte, même si nous pouvons constater une légère hausse des résultats par rapport à la généralisation seule, en termes de MAP, de P@5 et de P@10 lorsque les patrons ou les variantes sont utilisés individuellement ou combinés ensemble, les résultats obtenus restent inférieurs à ceux obtenus sans normalisation, avec la généralisation grâce à l'inclusion lexicale.

5.3 Comparaison à une approche par réseaux de neurones

Afin de compléter l'évaluation de notre méthode, nous l'avons comparée aux méthodes actuelles comme les réseaux de neurones. Celles-ci ont en commun avec notre méthode de réduire la taille de l'espace vectoriel. Nous avons utilisé l'outil Word2vec [Mikolov *et al.*, 2013] permettant de construire des vecteurs de contextes pour chaque mot d'un corpus. Cet outil est conçu pour les corpus de langue générale, mais surtout pour les corpus très volumineux, de plusieurs milliards de mots (et avec plusieurs millions de lemmes dans le vocabulaire). L'outil Word2vec implémente deux architectures permettant de calculer des représentations vectorielles : le modèle en sac de mots et le modèle Skip-Gram. Nous avons choisi d'utiliser le second modèle, qui a l'inconvénient

d'être plus lent mais qui est adapté aux mots de faible fréquence [9]. Pour le calcul de similarité, nous utilisons l'outil fourni avec Word2vec par les auteurs, qui comporte une seule mesure de similarité : le Cosinus.

Nous présentons dans un premier temps les paramètres qu'offre Word2vec et les choix que nous avons opérés autour de ces paramètres, afin d'avoir des résultats comparables. Puis nous discutons la qualité des résultats obtenus sur nos quatre corpus, en comparaison à notre méthode.

5.3.1 Word2vec : choix des paramètres

Les paramètres de Word2vec sont en partie similaires aux paramètres que nous utilisons et que nous avons évalués dans la section 5.1. Les paramètres que nous avons en commun sont les mots cibles et les contextes. En revanche, quatre autres paramètres sont propres aux réseaux de neurones : le nombre de dimensions des vecteurs, le *softmax* hiérarchique, l'échantillonnage négatif et le sous-échantillonage des mots fréquents. Nous discutons ci-dessous le nombre de dimensions. Pour les trois autres paramètres, nous utilisons les valeurs données par défaut par les auteurs [Mikolov *et al.*, 2013].

Mots cibles et contextes
Nous fournissons comme entrée au système le texte brut, sans réaliser de pré-traitement. La prise en compte d'informations linguistiques issues du pré-traitement, telles que le lemme ou l'étiquette morphosyntaxique exige une adaptation au format de l'outil Word2vec, comme l'a fait [Ferret, 2014]. Notre choix est essentiellement lié au temps dont nous disposions. Par conséquent, tous les mots du corpus sont pris en compte en tant que mots cibles et contextes. Le résultat obtenu correspond donc à des groupements sémantiques indépendants de la catégorie morphosyntaxique, c'est-à-dire regroupant des adjectifs, verbes, noms et prépositions.

De même que nous prenons en compte les termes, Word2vec intègre la reconnaissance d'expressions composées de plusieurs mots. Cependant, l'outil est conçu pour des corpus de langue générale, et par conséquent n'intègre pas d'extracteur de termes. Il utilise l'information mutuelle en se basant sur les co-occurrences de deux mots et leur fréquence d'apparition ensemble pour les regrouper. Ainsi, deux mots sont considérés comme une *expression* s'ils apparaissent ensemble à une fréquence suffisamment élevée, en opposition à de faibles fréquences avec d'autres co-occurrents. Pour la comparaison, nous choisissons d'utiliser la reconnaissance des expressions composées.

Enfin, le contexte correspond, comme nous le faisons, à une fenêtre graphique, dont le centre est le mot cible. La taille de fenêtre est définie par l'utilisateur. Ainsi, nous utilisons Word2vec avec la taille de fenêtre offrant les résultats de meilleure qualité avec notre méthode : la fenêtre restreinte, de 5 mots.

9. voir https://code.google.com/p/word2vec/

Nombre de dimensions

Word2vec permet de définir la taille des vecteurs de contextes, c'est-à-dire le nombre de dimensions des vecteurs. Etant donné que l'outil est mis au point pour les corpus de langue générale, nous avons adapté proportionnellement les tailles des vecteurs à la taille de nos corpus (tous inférieurs à quatre millions de mots). En effet, pour une taille de corpus de plusieurs centaines de millions de mots, les vecteurs ont généralement entre 50 et 100 dimensions [Mikolov *et al.*, 2013]. Nous avons donc expérimenté 50 dimensions, de manière à couvrir le nombre maximum de contextes.

5.3.2 Qualité des groupements sémantiques obtenus

Nous comparons les résultats obtenus par Word2vec aux résultats obtenus par notre méthode d'abstraction des contextes. Pour cette dernière, nous utilisons la configuration obtenant les meilleurs résultats : la généralisation avec l'inclusion lexicale (AD/IL), la fenêtre restreinte (5 mots) et l'indice de Jaccard pondéré avec la Fréquence Relative.

	MAP		Rprec		P@1		P@5	
	W2V	AD/IL	W2V	AD/IL	W2V	AD/IL	W2V	AD/IL
Textes Cliniques	0,001	0,057	0,001	0,047	0,003	0,067	0,002	0,049
Menelas	0	0,125	0	0,071	0,002	0,071	0,002	0,014

TABLEAU 5.31 : Résultats obtenus avec Word2vec (W2V) et après généralisation des contextes avec l'inclusion lexicale (AD/IL) pour les corpus Textes Cliniques et Menelas.

Pour nos quatre corpus, nous avons évalué les relations obtenues avec Word2vec par rapport aux relations contenues dans l'UMLS, et à l'aide des mesures d'évaluation décrites en section 4.2. Nous présentons dans le tableau 5.31 les résultats obtenus pour les corpus médicaux, avec Word2vec et avec notre méthode. Les résultats obtenus avec Word2vec sont bien plus faibles et sont semblables aux résultats que l'on obtient avec notre méthode, lorsque le Cosinus est utilisé. En effet, la MAP est de 0,001 pour les Textes Cliniques contre 0,057 pour notre méthode. Ces résultats sont toutefois peu significatifs, étant donné que nous n'avons pas réalisé de pré-traitement avec Word2vec. Le filtrage des mots cibles et des contextes par catégorie morphosyntaxique ainsi qu'une lemmatisation devraient nettement améliorer les résultats de Word2vec sur les corpus de spécialité.

	Mot cible ***débit cardiaque***		
	Word2vec	**Baseline (AD seule)**	**Abstraction (AD/IL)**
Rang	Voisins	Voisins	Voisins
1.	onde T	hypertension artérielle pulmonaire	ventricule gauche
2.	point	gêne fonctionnelle	hypertension artérielle
3.	dr DDD		instabilité coronarienne
4.	anévrisme		tabagisme
5.	accident		hypercholestérolémie
6.	dilaté		posologie
7.	examens		angor
8.	sévère		infarctus
9.	fonctionnel		patient
10.	bilan biologique		sténose

Tableau 5.32 : Corpus Menelas ; exemple de 10 premiers voisins obtenus pour le mot cible *débit cardiaque*, avec Word2vec, la baseline, et après généralisation (avec l'inclusion lexicale).

Pour se faire une idée plus précise des voisins obtenus avec Word2vec et du parallèle avec notre méthode, nous avons ensuite comparé pour une sélection de mots cibles (des termes simples et des termes complexes), les voisins obtenus avec ces deux méthodes.

Nous présentons dans les tableaux 5.32 et 5.33, deux exemples de mots cibles et des voisins obtenus en fonction de la méthode utilisée : Word2vec, l'analyse distributionnelle (AD seule) et l'analyse distributionnelle après abstraction des contextes (AD/IL). L'analyse manuelle révèle un comportement similaire des réseaux de neurones avec les deux langues de nos corpus : les corpus rédigés en anglais obtiennent des résultats équivalents aux corpus en français.

	Mot cible ***cough***		
	Word2vec	**Baseline (AD seule)**	**Abstraction (AD/IL)**
Rang	Voisins	Voisins	Voisins
1.	infections		nausea
2.	most likely		fever
3.	headache		vomiting
4.	palpitations		chill
5.	decline		history
6.	experienced		pain
7.	ros		patient
8.	complaining		diarrhea
9.	coughing		dysuria
10.	fever		dyspnae

TABLEAU 5.33 : Corpus Textes Cliniques ; exemple de 10 premiers voisins obtenus pour le mot cible *cough*, avec Word2vec, la baseline et après généralisation (avec l'inclusion lexicale).

En ce qui concerne les voisins sémantiques obtenus avec Word2vec, ces voisins sont généralement plus bruités que les voisins obtenus avec l'abstraction des contextes, et plus généraux. Par exemple, avec Word2vec, pour le corpus Menelas et le mot cible *cholestérol*, le voisin obtenu au deuxième rang est la préposition *à*. Une des raisons à cela est l'absence de pré-traitement linguistique dans Word2vec. En effet, la prise en compte des catégories morphosyntaxiques nous permet de conserver uniquement une sélection de catégories, aussi bien pour les mots cibles que pour les contextes. Les catégories que nous écartons des contextes sont essentiellement les mots non porteurs

de sens, tels que les prépositions et les adverbes. Ces mots peuvent être écartés par Word2vec, lors du sous-échantillonage des mots très fréquents, car ces mots vides se caractérisent généralement par une fréquence très élevée.

Les expressions identifiées par Word2vec coïncident généralement avec des termes complexes, par exemple *bilan lipidique*, *surcharge pondérale*, mais contiennent également beaucoup de bruit, comme par exemple des groupements verbaux (*sont représentés* ou *nous avons retenu l'indication*). Une grande partie des termes complexes n'est pas identifiée par Word2vec, indépendamment de la langue. C'est le cas notamment lorsque le terme complexe contient plus de deux mots comme par exemple *insulin-dependent diabetes mellitus*.

Nous pouvons enfin constater que les voisins obtenant le score de similarité le plus élevé sont généralement de bons voisins : ici dans nos exemples, les termes *onde T* et *infection* sont pertinents.

Là où l'écart entre les deux méthodes est le plus important, en termes de qualité, concerne les groupements sémantiques et les relations obtenues. La restriction par catégories morphosyntaxiques et l'abstraction des contextes distributionnels permettent d'obtenir des groupements sémantiques nettement plus riches et plus homogènes sémantiquement, comme le montrent les tableaux 5.33 et 5.32. L'avantage principal de notre méthode est d'être adaptée aux textes de spécialité, et d'être capable d'en extraire une connaissance plus approfondie et fine du domaine, tout en étant indépendante du domaine.

5.4 Bilan sur les expériences

Dans ce chapitre, nous avons présenté les expériences que nous avons menées afin d'évaluer notre méthode d'abstraction des contextes distributionnels. Autour de l'abstraction des contextes, nous avons réalisé trois types d'expériences : les expériences autour de la généralisation des contextes (abstraction réalisée avec des hyperonymes), de la normalisation des contextes (abstraction réalisée à l'aide de synonymes) et de la combinaison de la normalisation et de la généralisation. Dans l'ensemble et quel que soit le corpus utilisé, la généralisation, l'*abstraction conceptuelle* a un impact beaucoup plus fort que la normalisation, l'*abstraction lexicale*. La normalisation utilisée seule apporte peu, mais combinée à la généralisation permet d'affiner les résultats avec les plus petits corpus.

Ainsi, avec les corpus de petite taille (de l'ordre de 100 000 mots), il est préférable d'utiliser l'indice de Jaccard, la fenêtre restreinte (de 5 mots) et de combiner la normalisation et la généralisation des contextes distributionnels réalisée avec l'inclusion lexicale. Avec les corpus de plus grande taille (de l'ordre du million de mots), il est préférable d'utiliser également la fenêtre graphique restreinte (5 mots) et l'indice de Jaccard, et de généraliser avec l'inclusion lexicale. Avec ces corpus plus volumineux,

la généralisation des contextes a plus d'impact quand le Cosinus est utilisé, mais les résultats restent inférieurs à ceux obtenus avec l'indice de Jaccard.

Dans l'ensemble, 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. Ainsi, les relations obtenues après abstraction des contextes sont majoritairement des co-hyponymes. L'abstraction permet également d'obtenir quelques relations du domaine et propres au mot cible, telles que par exemple les relations *maladie - examen médical*, *examen médical - conséquence*. Notre méthode est cependant limitée, 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.

La comparaison avec une méthode actuelle de réseaux de neurones montre que les groupements obtenus avec notre méthode sont sémantiquement plus cohérents et obtiennent des résultats de meilleure qualité. Toutefois, pour approfondir cette évaluation, nous envisageons de réaliser un pré-traitement aux corpus avant de les traiter avec Word2vec. Nous pourrons ainsi sélectionner les catégories morphosyntaxiques des mots cibles et des contextes, mais également augmenter les fréquences du vocabulaire à l'aide de la lemmatisation. Nous envisageons également d'utiliser l'abstraction des contextes en amont de Word2vec.

Chapitre 6

Conclusion et perspectives

6.1 Conclusion

Les méthodes distributionnelles ont l'avantage de regrouper les mots sémantiquement proches. Actuellement, ces méthodes sont généralement utilisées sur des corpus en langue générale, très volumineux (de plusieurs centaines de millions de mots). Ces corpus se caractérisent par des fréquences de vocabulaire (nombre d'occurrences) élevées. L'application de ces méthodes à des textes de spécialité nécessite l'adaptation des paramètres distributionnels (type de contexte utilisé, mesure de similarité, etc.). De plus, les corpus de spécialité sont généralement de plus petite taille et se caractérisent par de faibles fréquences du vocabulaire. L'application de l'analyse distributionnelle à ces corpus est ainsi pénalisée par le problème de la dispersion des données dans la matrice de contextes.

La dispersion des données se traduit par un faible nombre de contextes associés aux mots cibles. Dans le cadre de l'application de l'analyse distributionnelle sur des corpus de spécialité, nous sommes confrontés à la prise en compte des termes simples et des termes complexes. Les termes se caractérisent par de très faibles fréquences, et n'ont que très peu d'occurrences dans les textes de spécialité. Les termes complexes ont une fréquence encore plus faible, car ils combinent des termes simples aux faibles fréquences. Les faibles fréquences sont ainsi une limite des méthodes distributionnelles et rendent difficile la capture de relations sémantiques pertinentes.

Pour répondre à ce problème de dispersion des données, nous avons proposé une méthode distributionnelle adaptée aux textes de spécialité. Cette méthode prend en compte les termes identifiés automatiquement et vise le regroupement sémantique des termes simples et des termes complexes. La méthode proposée réalise également une abstraction des contextes distributionnels, à travers leur généralisation et leur normalisation. La généralisation consiste à remplacer les contextes par leur hyperonyme, celui-ci étant identifié dans le corpus à l'aide soit des patrons lexico-syntaxiques, de l'inclusion lexicale ou de la variation terminologique. Quant à la normalisation, il s'agit de remplacer les contextes par le représentant de clusters de synonymes générés automatiquement à partir des corpus de travail à l'aide d'une méthode d'inférence de relations de synonymie.

Afin d'évaluer la robustesse de notre méthode, nous avons réalisé des expériences sur quatre corpus de spécialité. Ceux-ci diffèrent par leur taille, leur domaine de spécialité et la langue dans laquelle ils sont rédigés. Avant d'évaluer cette méthode d'abstraction, nous avons analysé le comportement de plusieurs paramètres distributionnels. Cette

première étape nous a permis d'adapter les valeurs des paramètres de la méthode distributionnelle à nos corpus de textes de spécialité. Ces paramètres sont la taille de la fenêtre graphique (large - 21 mots, et restreinte - 5 mots), les mesures de similarité (Nombre de Contextes Partagés, Fréquence des Contextes Partagés, l'indice de Jaccard et le Cosinus), de pondération (Information Mutuelle et Fréquence Relative), la sélection des contextes les plus discriminants que nous avons proposée (avec le Cf-Itf), l'utilisation de seuils sur les mots cibles et contextes (Fréquence des mots cibles, Fréquence et Nombre des Contextes Partagés), et le seuil sur le score de similarité. Cette première série d'expériences nous a permis de définir les paramètres que nous utilisons lors de l'abstraction des contextes : les deux tailles de fenêtre, l'indice de Jaccard pondéré avec la Fréquence Relative et le Cosinus pondéré avec l'Information Mutuelle, et l'utilisation de trois seuils sur les mots cibles et contextes partagés (Fréquence et Nombre des Contextes Partagés, Fréquence des mots cibles).

Concernant l'abstraction des contextes, nous avons réalisé trois types d'expériences : des expériences de généralisation des contextes (abstraction réalisée avec des hyperonymes), de normalisation des contextes (abstraction réalisée à l'aide de synonymes) et de combinaison de la normalisation et de la généralisation. Les résultats de ces expériences sont évalués par comparaison aux relations contenues dans des ressources existantes (Agrovoc, UMLS) et dans une ressource issue du Web. Pour cela, nous utilisons des mesures de précision, de MAP et de R-précision, tout en discutant également le nombre de relations acquises et retrouvées dans les ressources.

De ces expériences, il ressort que l'abstraction des contextes distributionnels améliore la qualité des résultats. Les groupements sémantiques obtenus sont ainsi plus homogènes et cohérents, et les termes complexes sont pris en compte dans les mots cibles. La configuration la plus adaptée est la généralisation des contextes avec les relations acquises par inclusion lexicale. Pour que celle-ci soit efficace, il est préférable d'utiliser l'indice de Jaccard pondéré par la Fréquence Relative comme mesure de similarité, et de définir le contexte à l'aide d'une fenêtre graphique restreinte (de 5 mots).

Dans l'ensemble, la généralisation a un impact beaucoup plus fort que la normalisation. La normalisation utilisée seule apporte peu, mais combinée à la généralisation elle permet d'affiner les résultats pour les corpus de petite taille (de l'ordre de 100 000 mots). Avec les corpus les plus volumineux, la généralisation des contextes a plus d'impact quand le Cosinus est utilisé, mais les résultats restent inférieurs à ceux obtenus avec l'indice de Jaccard.

Enfin, nous nous sommes comparés de façon rapide à une méthode par réseaux de neurones. L'évaluation montre que dans les conditions de notre expérience, notre méthode obtient des groupements sémantiques plus cohérents et de meilleurs résultats, grâce à la prise en compte dans notre méthode d'une analyse linguistique.

6.2 Perspectives

Ce travail de thèse 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. Pour cela, nous envisageons de construire un réseau de relations acquises en corpus à partir de ces relations. Il nous faudra alors déterminer le niveau de la hiérarchie à prendre en compte afin de substituer les termes dans les contextes, en prenant en compte, par exemple, la distance sémantique dans la hiérarchie. Par exemple, si le niveau 3 est généralisé au niveau 1 ou 2, selon le degré d'abstraction sémantique souhaité. Une abstraction au niveau 1 permettrait une abstraction d'un plus grand nombre de contextes, mais effacerait également un plus grand nombre de traits sémantiques qui peuvent être nécessaires dans la constitution des groupements sémantiques distributionnels. La méthode d'abstraction des contextes devra alors être adaptée de manière à prendre en compte ce réseau de relations. Globalement, l'utilisation du réseau de relations devrait nous permettre de réaliser une abstraction des contextes plus importante ou plus précise sémantiquement.

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 pourraient 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 celles proposées par des terminologies. Afin d'adapter les terminologies aux corpus de travail, nous sélectionnerons dans la ressource les relations contenant les termes du corpus. Il nous sera alors possible d'évaluer l'impact de l'abstraction et des relations lorsque leur statut terminologique est maîtrisé, et avec des relations jugées un peu plus fiables.

Enfin, l'évaluation des résultats reste une étape difficile lorsque l'on travaille sur des ressources distributionnelles, contenant un très grand nombre de relations aux types très variés. Nous avons réalisé une évaluation manuelle d'une partie des relations acquises par notre méthode d'abstraction. Il serait intéressant de réaliser également une évaluation par la tâche, à travers l'évaluation de la performance d'un moteur de recherche ou d'un outil de traduction. Ceci permettrait d'évaluer l'impact des groupements acquis sur les documents retournés par le moteur ou sur la qualité des traductions.

Enfin, la comparaison de notre méthode à d'autres méthodes de réduction de dispersion des données, telles que le Random Indexing et la LSA, serait utile pour évaluer plus précisément la méthode d'abstraction du point de vue de cette problématique.

Annexes

A Références pour les textes du corpus Guides Alimentaires

Corpus Guides Alimentaires : liste des URL correspondant aux document PDF ayant servis à la constitution du corpus.

- http://www.exobiologie.info/diabete/33%20alimentation.pdf
- http://nosobase.chu-lyon.fr/recommandations/cclin/cclinParisNord/1999_alimentation_CCLIN.pdf
- http://qualiteofficine.univ-lille2.fr/fileadmin/user_upload/memoires_2005_a_2006/DU2005-6_chap1.pdf
- http://kid.pasteur-lille.fr/ateliers/nutrition/imp_nutrition.pdf
- http://agriculture.gouv.fr/IMG/pdf/avis_cna_53.pdf
- http://agriculture.gouv.fr/IMG/pdf/avis_cna_54.pdf
- http://www.guidesurlediabete.com/downloads/fvol202004.pdf
- http://reseaudiabete41.fr/dossier_adhesion_version_3.pdf
- http://www.cap-sciences.net/upload/equilibre_alim.pdf
- http://www.observatoiredupain.com/images/produits/da36410f-1c08-406b-ad35-a17301ec43ee.pdf
- http://www.inpes.sante.fr/10000/themes/nutrition/etude_aidealim/Etude0604.pdf
- http://www.spondylarthrite-ankylosante.info/alimentation3.pdf
- http://www.fleurymichonsports.fr/medias/pdf/Fiche-Conseils-Natation-pr-pdf2--1.pdf
- http://host-13.celuga.net/247/Images/Produits/A95FAAB5-63AE-4826-9E95-984EFEE137B5.PDF
- http://www.en-mouvement.ca/opFichier/-LYDFgDAgitra-13223.pdf
- http://www.milupa.ca/assets/Documents/fr/guidefeedingyourbabyfr.pdf
- http://www.inpes.sante.fr/CFESBases/catalogue/pdf/567.pdf
- http://www.cancer.be/sites/default/files/guide-oncodieteticien.pdf
- http://www.gov.mb.ca/healthyschools/foodinschools/documents/handbook.fr.pdf
- http://www.one.be/fileadmin/user_upload/coin-emploi/SELOR_AFG14180/enfant_et_nutrition_one.pdf
- http://www.urps-ml-paysdelaloire.fr/APIMED/uploads/pdf/Le%20patient%20diab%C3%A9tique/DNID_ald8_guidemedecin_diabetetype2_revunp_vucd.pdf
- http://www.inpes.sante.fr/CFESBases/catalogue/pdf/747.pdf
- https://www.creditmutuel.fr/cmse/fr/banques/telechargements/guide-nutrition-et-sante.pdf
- http://www.fnors.org/uploadedFiles/publicationsFnors/guide_nutrition_fnors2009.pdf
- http://stomatonantes.free.fr/Resources/guide.pdf
- http://www.inpes.sante.fr/CFESBases/catalogue/pdf/715.pdf

- https://www.isostar.com/share/guide-de-la-nutrition.pdf
- http://www.uprt.fr/nutrition_ville_pnns_2005.pdf
- http://www.againstpain.org/pdf/Dlrs_10.pdf
- http://www.fantaproject.org/sites/default/files/resources/Haiti_NutHIV_Guidelines_2010.pdf
- http://www.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---ilo_aids/documents/legaldocument/wcms_127595.pdf
- http://www.em-consulte.com/em/emm/numspecial/Int-Guide_therapeutique.pdf
- http://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCoQFjAA&url=http%3A%2F%2Fwww.generale-de-sante.fr%2Fhopital-prive-residence-du-parc-marseille%2Fcontent%2Fdownload%2F5772%2F67249%2Ffile%2FLivret%2520%25C3%25A9ducation%2520th%25C3%25A9rapeutique%2520diab%25C3%25A9tologie.pdf&ei=quFrVPGyDofnaqmUgYAB&usg=AFQjCNHAAMdJXI9BeM8-WeHClHNFkiKeKw&siuTpyvkIiT20F2AyXGZ5grw&cad=rja
- http://www.ffcd.fr/DOC/PATIENT/GENERAL/livret_conseils_alimentaires.pdf
- http://www.inpes.sante.fr/CFESBases/catalogue/pdf/1059.pdf
- http://bef.novonordisk.be/Images/Images_dvra/Nutrition_FR_0605.pdf
- http://www.cap-sciences.net/upload/nutrition_nutriments.pdf
- http://www.intercomsante57.fr/html/profsante/pdf/Alimentation-et-prevention-du-diabete.pdf
- http://www.dietandcancerreport.org/cancer_resource_center/downloads/summary/french.pdf
- http://www.sfdiabete.org/sites/sfd.prod.saegir.cyim.com/files/files/Pdf/Recos-R%C3%A9f%C3%A9rentiels/Education_dietetique.pdf
- http://www.economie.gouv.fr/files/directions_services/daj/marches_publics/oeap/gem/nutrition/nutrition.pdf
- http://www.rsfs.ca/opFichier/bien_manger_avec_le_guide_alimentaire_canadien_oCK8VLZJHra9_6659.pdf
- http://www.editions-dangles.fr/bibliotheque/documents/9782703303152.pdf

B Résultats : Impact des seuils en fonction de la mesure de similarité utilisée (corpus de grande taille)

B.1 Cosinus

Cosinus

			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Guides Alim.	Avec seuils	0,015	0,001	0,018	0,009
		Sans seuils	0,001	0,002	0,007	0,005
	Recettes	Avec seuils	0,002	0,002	0,005	0,007
		Sans seuils	0,001	0,001	0,001	0,005
R-précision	Guides Alim.	Avec seuils	0,006	0	0,010	0,002
		Sans seuils	0	0	0,004	0,001
	Recettes	Avec seuils	0,004	0	0,001	0,006
		Sans seuils	0	0	0	0,005
P@1	Guides Alim.	Avec seuils	0,005	0	0,007	0
		Sans seuils	0	0	0,004	0
	Recettes	Avec seuils	0,003	0	0	0,008
		Sans seuils	0	0	0	0,006
P@5	Guides Alim.	Avec seuils	0,005	0	0,005	0
		Sans seuils	0,001	0	0,002	0
	Recettes	Avec seuils	0,004	0	0,002	0,017
		Sans seuils	0	0	0	0,003

B.2 Cosinus pondéré avec l'information mutuelle

			Cosinus + IM			
			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Guides Alim.	Avec seuils	0,025	0,002	0,018	0,010
		Sans seuils	0,001	0,002	0,006	0,006
	Recettes	Avec seuils	0,011	0,002	0,005	0,004
		Sans seuils	0	0	0,001	0,004
R-précision	Guides Alim.	Avec seuils	0,011	0	0,010	0,002
		Sans seuils	0	0	0,004	0
	Recettes	Avec seuils	0,007	0,001	0,001	0,006
		Sans seuils	0	0	0	0,004
P@1	Guides Alim.	Avec seuils	0,005	0	0,007	0
		Sans seuils	0	0	0,004	0
	Recettes	Avec seuils	0,003	0	0,002	0,010
		Sans seuils	0	0	0	0,003
P@5	Guides Alim.	Avec seuils	0,005	0	0,005	0
		Sans seuils	0,001	0	0,001	0
	Recettes	Avec seuils	0,005	0,003	0	0,012
		Sans seuils	0	0	0	0,001

B.3 Indice de Jaccard non pondéré

			Jaccard			
			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Guides Alim.	Avec seuils	0,086	0,024	0,072	0,028
		Sans seuils	0,047	0,027	0,061	0,019
	Recettes	Avec seuils	0,007	0,007	0,053	0,024
		Sans seuils	0,007	0,007	0,026	0,005
R-précision	Guides Alim.	Avec seuils	0,054	0,017	0,048	0,010
		Sans seuils	0,034	0,016	0,046	0,009
	Recettes	Avec seuils	0,043	0,004	0,052	0,018
		Sans seuils	0,023	0,006	0,025	0,003
P@1	Guides Alim.	Avec seuils	0,079	0,023	0,070	0,019
		Sans seuils	0,049	0,022	0,066	0,018
	Recettes	Avec seuils	0,068	0,007	0,104	0,059
		Sans seuils	0,044	0,009	0,048	0,003
P@5	Guides Alim.	Avec seuils	0,061	0,007	0,041	0,012
		Sans seuils	0,034	0,009	0,041	0,007
	Recettes	Avec seuils	0,056	0,007	0,061	0,032
		Sans seuils	0,028	0,008	0,027	0,004

B.4 Nombre de contextes partagés

			NbCtxt			
			W21-TS+TC	W21-TC	W5-TS+TC	W5-TC
MAP	Guides Alim.	Avec seuils	0,049	0,018	0,053	0,022
		Sans seuils	0,032	0,016	0,033	0,019
	Recettes	Avec seuils	0,005	0,005	0,047	0,009
		Sans seuils	0,003	0,003	0,026	0,009
R-précision	Guides Alim.	Avec seuils	0,032	0,003	0,035	0,008
		Sans seuils	0,022	0,003	0,020	0,007
	Recettes	Avec seuils	0,030	0,004	0,041	0,006
		Sans seuils	0,019	0,002	0,022	0,008
P@1	Guides Alim.	Avec seuils	0,050	0	0,052	0
		Sans seuils	0,026	0	0,037	0
	Recettes	Avec seuils	0,027	0	0,052	0,020
		Sans seuils	0,023	0	0,021	0,020
P@5	Guides Alim.	Avec seuils	0,005	0,009	0,034	0,015
		Sans seuils	0,001	0,009	0,023	0,014
	Recettes	Avec seuils	0,031	0,005	0,044	0,018
		Sans seuils	0,012	0,002	0,014	0,009

Bibliographie

[Adam *et al.*, 2013a] Adam, C., Fabre, C. et Muller, P. (2013a). Évaluer et améliorer une ressource distributionnelle. *Traitement Automatique des Langues*, 54(1):71–97.

[Adam *et al.*, 2013b] Adam, C., Fabre, C. et Tanguy, L. (2013b). Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle. *In Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, pages 140–153, Les Sables d'Olonne, France.

[Ando, 2000] Ando, R. K. (2000). Latent semantic space : Iterative scaling improves precision of inter-document similarity measurement. *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 216–223, New York, NY, USA. ACM.

[Aubin et Hamon, 2006] Aubin, S. et Hamon, T. (2006). Improving term extraction with terminological resources. *In Advances in Natural Language Processing*, numéro 4139 de LNAI, pages 380–387. Springer.

[Bannour *et al.*, 2011] Bannour, S., Audibert, L. et Nazarenko, A. (2011). Mesures de similarité distributionnelle entre termes. *In 22es journées francophones d'ingénierie des connaissances*, pages 523–538, Chambéry, France.

[Baroni, 2009] Baroni, M. (2009). *Corpus linguistics : An international handbook*, volume 2, chapitre Distributions in text, pages 803–821. Anke Lüdeling and Merja Kytö, Berlin.

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

[Baroni et Lenci, 2010] Baroni, M. et Lenci, A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

[Baskaya *et al.*, 2013] Baskaya, O., Sert, E., Cirik, V. et Yuret, D. (2013). Ai-ku : Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. *In Proceedings of SemEval - 2013*, pages 300–306, Atlanta, Georgia, USA. Association for Computational Linguistics.

[Bengio *et al.*, 2003] Bengio, Y., Ducharme, R., Vincent, P. et Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

[Bernier-Colborne, 2014] Bernier-Colborne, G. (2014). Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques. *In Actes de SemDis 2014*, pages 238–251, Marseille, France.

[Blei *et al.*, 2003] Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Bodenreider *et al.*, 2002] Bodenreider, O., Rindflesch, T. C. et Burgun, A. (2002). Unsupervised, corpus-based method for extending a biomedical terminology. *In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3*, BioMed '02, pages 53–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Bouaud *et al.*, 2000] Bouaud, J., Habert, B., Nazarenko, A. et Zweigenbaum, P. (2000). Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. *In* Charlet, J., Zacklad, M., Kassel, G. et Bourigault, D., éditeurs : *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, chapitre 17, pages 275–290. Eyrolles, Paris.

[Broda *et al.*, 2009] Broda, B., Piasecki, M. et Szpakowicz, S. (2009). Rank-based transformation in measuring semantic relatedness. *In* Gao, Y. et Japkowicz, N., éditeurs : *Canadian Conference on AI*, volume 5549, pages 187–190. Springer.

[Buckley et Voorhees, 2005] Buckley, C. et Voorhees, E. (2005). Retrieval system evaluation. *In* Voorhees, E. et Harman, D., éditeurs : *TREC : Experiment and Evaluation in Information Retrieval*, chapitre 3. MIT Press.

[Budanitsky et Hirst, 2006] Budanitsky, A. et Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

[Bullinaria et Levy, 2007] Bullinaria, J. et Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior Research Methods*, 39(3):510–526.

[Bullinaria et Levy, 2012] Bullinaria, J. A. et Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD. *Behav Res Methods*, pages 890–907.

[Buntine et Jakulin, 2005] Buntine, W. L. et Jakulin, A. (2005). Discrete component analysis. *In SLSFS*, pages 1–33.

[Caraballo, 1999] Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. *In ACL*, pages 120–126.

[Chatterjee et Mohan, 2008] Chatterjee, N. et Mohan, S. (2008). Discovering word senses from text using random indexing. *In Proceedings of the 9th International*

Conference on Computational Linguistics and Intelligent Text Processing, CICLing'08, pages 299–310, Berlin, Heidelberg. Springer-Verlag.

[Church *et al.*, 1994] Church, K., Gale, W., Hanks, P., Hindle, D. et Moon, R. (1994). *Lexical substituability.* Computational Approaches to the Lexicon. Oxford University Press, Oxford.

[Church et Hanks, 1990] Church, K. W. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

[Clark, 2014] Clark, S. (2014). *Handbook of Contemporary Semantics*, chapitre Vector Space Models of Lexical Meaning. Shalom Lappin and Chris Fox, Oxford.

[Claveau et Sébillot, 2004] Claveau, V. et Sébillot, P. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *In 11e conférence sur le traitement automatique des langues naturelles, TALN'04*, Fès, Maroc.

[Curran et Moens, 2002a] Curran, J. et Moens, M. (2002a). Scaling context space. *In 40th Annual Meeting of the ACL*, pages 231–238, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Curran, 2004] Curran, J. R. (2004). *From distributional to semantic similarity.* Thèse de doctorat, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

[Curran et Moens, 2002b] Curran, J. R. et Moens, M. (2002b). Improvements in automatic thesaurus extraction. *In Workshop on Unsupervised lexical acquisition*, volume 9, pages 59–66, Morristown, NJ, USA. Association for Computational Linguistics.

[Dagan *et al.*, 1999] Dagan, I., Lee, L. et Pereira, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Mach. Learn.*, 34(1-3):43–69.

[Daille et Morin, 2005] Daille, B. et Morin, E. (2005). French-english terminology extraction from comparable corpora. *In Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, pages 707–718.

[Deerwester *et al.*, 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Déjean *et al.*, 2002] Déjean, H., Gaussier, E. et Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. *In Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Erk et Padó, 2008] Erk, K. et Padó, S. (2008). A structured vector space model for word meaning in context. *In Proceedings of the Conference on Empirical Methods*

in Natural Language Processing, EMNLP '08, pages 897–906, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Erk et Padó, 2010] Erk, K. et Padó, S. (2010). Exemplar-based models for word meaning in context. *In Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 92–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Fabre et Bourigault, 2006] Fabre, C. et Bourigault, D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *In TALN 2006*, pages 121–129, Leuven.

[Ferret, 2013a] Ferret, O. (2013a). Identifying bad semantic neighbors for improving distributional thesauri. *In ACL 2013*, pages 561–571.

[Ferret, 2013b] Ferret, O. (2013b). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. *In TALN 2013*, pages 48–61, Les Sables d'Olonne, France.

[Ferret, 2014] Ferret, O. (2014). Utiliser un modèle neuronal générique pour la substitution lexicale. *In Actes de SemDis 2014*, pages 218–227, Marseille, France.

[Firth, 1957] Firth, J. (1957). *A synopsis of linguistic theory 1930-1955*, pages 1–32. Oxford : Blackwell.

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

[Gorman et Curran, 2006] Gorman, J. et Curran, J. R. (2006). Scaling distributional similarity to large corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 361–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Grabar et Zweigenbaum, 2003] Grabar, N. et Zweigenbaum, P. (2003). Lexically-based terminology structuring. *In Terminology*, volume 10, pages 23–54.

[Grefenstette, 1992] Grefenstette, G. (1992). Sextant : Exploring unexplored contexts for semantic extraction from syntactic analysis. *In Proceedings of the 30th Annual Meeting on Association for Computational Linguistics*, ACL '92, pages 324–326, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Grefenstette, 1994] Grefenstette, G. (1994). Corpus-derived first, second and third-order word affinities. *In Sixth Euralex International Congress*, pages 279–290.

[Habert *et al.*, 1998] Habert, B., Nazarenko, A., Zweigenbaum, P. et Bouaud, J. (1998). Extending an existing specialized semantic lexicon. *In Language Resources and Evaluation (LREC)*, pages 663–668, Grenade.

[Hamon et Nazarenko, 2008] Hamon, T. et Nazarenko, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *TAL*, 49(2):127–154.

[Hamon *et al.*, 1998] Hamon, T., Nazarenko, A. et Gros, C. (1998). A step towards the detection of semantic variants of terms in technical documents. *In International Conference on Computational Linguistics (COLING-ACL'98)*, pages 498–504, Université de Montréal, Montréal, Quebec, Canada.

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

[Henestroza Anguiano et Denis, 2011] Henestroza Anguiano, E. et Denis, P. (2011). Fredist : Automatic construction of distributional thesauri for French. *In Actes de la 18ème conférence TALN*, volume 2, pages 119–124, Montpellier, France, France.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA. ACM.

[Jacquemin, 1997] Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'HDR en informatique, Université de Nantes.

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

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

[Jones et Furnas, 1987] Jones, W. P. et Furnas, G. W. (1987). Pictures of relevance : A geometric analysis of similarity measures. *J. Am. Soc. Inf. Sci.*, 38(6):420–442.

[Kanerva *et al.*, 2000] Kanerva, P., Kristofersson, J. et Holst, A. (2000). Random indexing of text samples for latent semantic analysis. *In* Gleitman, L. et Josh, A., éditeurs : *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, volume 1036, pages 103–106, Erlbaum, New Jersey.

[Karlgren et Sahlgren, 2001] Karlgren, J. et Sahlgren, M. (2001). From words to understanding. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 294–308. Foundations of Real-World Intelligence.

[Kiela et Clark, 2014] Kiela, D. et Clark, S. (2014). A systematic study of semantic vector space model parameters. *In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL 2014*, pages 21–30, Gothenburg, Sweden.

[Kilgarriff et Yallop, 2000] Kilgarriff, A. et Yallop, C. (2000). What's in a thesaurus. *In IN PROCEEDINGS OF THE SECOND CONFERENCE ON LANGUAGE RESOURCE AN EVALUATION*, pages 1371–1379.

[Kris Heylen et Speelman, 2008] Kris Heylen, Yves Peirsman, D. G. et Speelman, D. (2008). Modelling word similarity : an evaluation of automatic synonymy extraction algorithms. *In Language Resources and Evaluation (LREC'08)*, pages 3243–3249, Marrakech, Morocco. European Language Resources Association (ELRA).

[Landauer et Dumais, 1997] Landauer, T. et Dumais, S. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review ; Psychological Review*, 104(2):211.

[Lee et Seung, 1999] Lee, D. D. et Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791.

[Lee, 1999] Lee, L. (1999). Measures of distributional similarity. *In Proceedings of ACL-1999*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. *In Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lin, 1998b] Lin, D. (1998b). An information-theoretic definition of similarity. *In In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

[Lin et Pantel, 2001] Lin, D. et Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

[Lund et Burgess, 1996] Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208.

[Lund *et al.*, 1995] Lund, K., Burgess, C. et Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *In Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

[Manning *et al.*, 2008] Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Mikolov *et al.*, 2013] Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Mitchell et Lapata, 2010] Mitchell, J. et Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

[Morin et Hazem, 2014] Morin, E. et Hazem, A. (2014). Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1284–1293, Baltimore, United States.

[Morin et Jacquemin, 2004] Morin, E. et Jacquemin, C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38(4):363–396.

[Morlane-Hondère, 2013] Morlane-Hondère, F. (2013). *Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique*. Thèse de doctorat, Université de Toulouse.

[Morris et Hirst, 2004] Morris, J. et Hirst, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS 04, pages 46–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Nastase *et al.*, 2013] Nastase, V., Nakov, P., Séaghdha, D. O. et Szpakowicz, S. (2013). *Semantic Relations Between Nominals*. Morgan and Claypool Publishers.

[Nazarenko *et al.*, 1997] Nazarenko, A., Zweigenbaum, P., Bouaud, J. et Habert, B. (1997). Corpus-based identification and refinement of semantic classes. *Proc AMIA Annu Fall Symp*, pages 585–9.

[Padó et Lapata, 2007] Padó, S. et Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

[Panchenko et Morozova, 2012] Panchenko, A. et Morozova, O. (2012). A study of hybrid similarity measures for semantic relation extraction. *In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 10–18, Avignon, France. Association for Computational Linguistics.

[Pantel *et al.*, 2009] Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M. et Vyas, V. (2009). Web-scale distributional similarity and entity set expansion. *In EMNLP '09 : Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 938–947, Morristown, NJ, USA. Association for Computational Linguistics.

[Pedersen *et al.*, 2004] Pedersen, T., Patwardhan, S. et Michelizzi, J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. *In Demonstration Papers at HLT-NAACL 2004*, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Peirsman et Geeraerts, 2009] Peirsman, Y. et Geeraerts, D. (2009). Predicting strong associations on the basis of corpus data. *In EACL-2009*, pages 648–654, Athens, Greece.

[Peirsman *et al.*, 2008] Peirsman, Y., Kris, H. et Dirk, G. (2008). Size matters. tight and loose context definitions in english word space models. *In ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41, Hamburg, Germany.

[Périnet et Hamon, 2014] Périnet, A. et Hamon, T. (2014). Generalising and normalising distributional contexts to reduce data sparsity : application to medical corpora. *In CompuTerm 2014 : 4th International Workshop on Computational Terminology*, pages 39–46, Dublin,Ireland. COLING.

[Périnet et Hamon, 2014] Périnet, A. et Hamon, T. (2014). Réduction de la dispersion des données par généralisation des contextes distributionnels : application aux textes de spécialité. *In Proceedings of TALN 2014 (Volume 1 : Long Papers)*, pages 232–243. Association pour le Traitement Automatique des Langues.

[Poibeau et Messiant, 2008] Poibeau, T. et Messiant, C. (2008). Do we still need gold standards for evaluation ? *In Proceedings of LREC 2008*, pages 1–6. European Language Resources Association.

[Rapp, 2003] Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. *In MT Summit'2003*, pages 315–322.

[Sahlgren, 2005] Sahlgren, M. (2005). An introduction to random indexing. *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, pages 1–9.

[Sahlgren, 2006] Sahlgren, M. (2006). *The Word-Space Model : Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Thèse de doctorat, Stockholm University, Stockholm, Sweden.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In New Methods in Language Processing*, pages 44–49, Manchester, UK.

[Schölkopf *et al.*, 1999] Schölkopf, B., Smola, A. J. et Müller, K.-R. (1999). Kernel principal component analysis. *In* Schölkopf, B., Burges, C. J. C. et Smola, A. J., éditeurs : *Advances in Kernel Methods*, pages 327–352, Cambridge, MA, USA. MIT Press.

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

[Sun *et al.*, 2013] Sun, W., Rumshisky, A. et Uzuner, Ö. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *JAMIA*, 20(5):806–813.

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

[Tsatsaronis et Panagiotopoulou, 2009] Tsatsaronis, G. et Panagiotopoulou, V. (2009). A generalized vector space model for text retrieval based on semantic relatedness. *In EACL 2009*, pages 70–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Turney et Pantel, 2010] Turney, P. D. et Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

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

[van der Plas et Bouma, 2004] van der Plas, L. et Bouma, G. (2004). Syntactic contexts for finding semantically related words. *In* van der Wouden, T., Poss, M., Reckman, H. et Cremers, C., éditeurs : *Computational Linguistics in the Netherlands 2004, Selected Papers from the Fifteenth CLIN Meeting, December 17, Leiden Centre for Linguistics*, pages 173–186. LOT Utrecht.

[van der Plas et Tiedemann, 2010] van der Plas, L. et Tiedemann, J. (2010). Finding medical term variations using parallel corpora and distributional similarity. *In Proceedings of the 6th Workshop on Ontologies and Lexical Resources (OntoLex 2010*, pages 28–37, Beijing, China.

[Vozalis et Margaritis, 2003] Vozalis, E. et Margaritis, K. G. (2003). Analysis of recommender systems' algorithms. *In The 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA), Athens, Greece*, pages 1–14.

[Weeds et Weir, 2005] Weeds, J. et Weir, D. (2005). Co-occurrence retrieval : A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

[Weeds *et al.*, 2004] Weeds, J., Weir, D. et McCarthy, D. (2004). Characterising measures of lexical distributional similarity. *In Proceedings of COLING'2004*, pages 1015–1022, Stroudsburg, PA, USA.

[Widdows et Ferraro, 2008] Widdows, D. et Ferraro, K. (2008). Semantic vectors : a scalable open source package and online technology management application. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation)*, pages 1183–1190, Marrakech, Morocco. European Language Resources Association (ELRA).

[Wilks *et al.*, 1990] Wilks, Y. A., Dan, Mcdonald, J. E., Plate, T. et Slator, B. M. (1990). Providing machine tractable dictionary tools. *Journal of Machine Translation*, 2:750–755.

[Yuret, 2012] Yuret, D. (2012). Fastsubs : An efficient and exact procedure for finding the most likely lexical substitutes based on an n-gram language model. *IEEE Signal Process. Lett.*, 19(11):725–728.

[Zesch et Gurevych, 2010] Zesch, T. et Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists ; measuring the semantic relatedness of words. *Nat. Lang. Eng.*, 16(1):25–59.

[Zheng *et al.*, 2011] Zheng, W., Qian, Y. et Tang, H. (2011). Dimensionality reduction with category information fusion and non-negative matrix factorization for text categorization. *In* Deng, H., Miao, D., Lei, J. et Wang, F. L., éditeurs : *AICI*, volume 7004 de *Lecture Notes in Computer Science*, pages 505–512. Springer.

[Zhitomirsky-Geffet et Dagan, 2009] Zhitomirsky-Geffet, M. et Dagan, I. (2009). Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.

[Zweigenbaum, 1994] Zweigenbaum, P. (1994). Menelas : an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45.

[Zweigenbaum et Habert, 2006] Zweigenbaum, P. et Habert, B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *Revue Glottopol*, 8:22–44.

Index

Résumé. Dans les domaines de spécialité, les applications telles que la recherche d'information ou la traduction automatique, s'appuient sur des ressources terminologiques pour prendre en compte les termes, les relations sémantiques ou les regroupements de termes. Pour faire face au coût de la constitution de ces ressources, des méthodes automatiques ont été proposées. Parmi celles-ci, l'analyse distributionnelle s'appuie sur la redondance d'informations se trouvant dans le contexte des termes pour établir une relation. Alors que cette hypothèse est habituellement mise en œuvre grâce à des modèles vectoriels, ceux-ci souffrent du nombre de dimensions considérable et de la dispersion des données dans la matrice des vecteurs de contexte. En corpus de spécialité, ces informations contextuelles redondantes sont d'autant plus dispersées et plus rares que les corpus ont des tailles beaucoup plus petites. De même, les termes complexes sont généralement ignorés étant donné leur faible nombre d'occurrence. Dans cette thèse, nous nous intéressons au problème de la limitation de la dispersion des données sur des corpus de spécialité et nous proposons une méthode permettant de densifier la matrice des contextes en réalisant une abstraction des contextes distributionnels. Des relations sémantiques acquises en corpus sont utilisées pour généraliser et normaliser ces contextes. Nous avons évalué la robustesse de notre méthode sur quatre corpus de tailles, de langues et de domaines différents. L'analyse des résultats montre que, tout en permettant de prendre en compte les termes complexes dans l'analyse distributionnelle, l'abstraction des contextes distributionnels permet d'obtenir des groupements sémantiques de meilleure qualité mais aussi plus cohérents et homogènes.

Mots clés. : Traitement Automatique des Langues, textes de spécialité, terminologie, analyse distributionnelle, modèle vectoriel, groupements sémantiques, termes complexes, relations sémantiques, abstraction de contextes.

Abstract. In specialised domains, the applications such as information retrieval for machine translation rely on terminological resources for taking into account terms or semantic relations between terms or groupings of terms. In order to face up to the cost of building these resources, automatic methods have been proposed. Among those methods, the distributional analysis uses the repeated information in the contexts of the terms to detect a relation between these terms. While this hypothesis is usually implemented with vector space models, those models suffer from a high number of dimensions and data sparsity in the matrix of contexts. In specialised corpora, this contextual information is even sparser and less frequent because of the smaller size of the corpora. Likewise, complex terms are usually ignored because of their very low number of occurrences. In this thesis, we tackle the problem of data sparsity on specialised texts. We propose a method that allows making the context matrix denser, by performing an abstraction of distributional contexts. Semantic relations acquired from corpora are used to generalise and normalise those contexts. We evaluated the method robustness on four corpora of different sizes, different languages and different domains. The analysis of the results shows that, while taking into account complex terms in distributional analysis, the abstraction of distributional contexts leads to defining semantic clusters of better quality, that are also more consistent and more homogeneous.

Keywords. : Natural Language Processing, specialised corpora, terminology, distributional analysis, vector space model, semantic cluster, complex terms, semantic relations, context abstraction.