Université de Montréal

Aide à l'identification de relations lexicales au moyen de la sémantique distributionnelle et son application à un corpus bilingue du domaine de l'environnement

par
Gabriel Bernier-Colborne

Département de linguistique et de traduction
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en traduction, option terminologie

Août, 2016

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

Aide à l'identification de relations lexicales au moyen de la sémantique distributionnelle et son application à un corpus bilingue du domaine de l'environnement

présentée par :

Gabriel Bernier-Colborne

a été évaluée par un jury composé des personnes suivantes :

Marie-Claude L'Homme,	président-rapporteur
Patrick Drouin,	directeur de recherche
Dominic Forest,	codirecteur
Philippe Langlais,	membre du jury
Vincent Claveau,	examinateur externe
Timothée Poisot,	représentant du doyen de la FESP

Thèse acceptée le : 3 novembre 2016

RÉSUMÉ

L'analyse des relations lexicales est une des étapes principales du travail terminologique. Cette tâche, qui consiste à établir des liens entre des termes dont les sens sont reliés, peut être facilitée par des méthodes computationnelles, notamment les techniques de la sémantique distributionnelle. En estimant la similarité sémantique des mots à partir d'un corpus, ces techniques peuvent faciliter l'analyse des relations lexicales.

La qualité des résultats offerts par les méthodes distributionnelles dépend, entre autres, des nombreuses décisions qui doivent être prises lors de leur mise en œuvre, notamment le choix et le paramétrage du modèle. Ces décisions dépendent, à leur tour, de divers facteurs liés à l'objectif visé, tels que la nature des relations lexicales que l'on souhaite détecter ; celles-ci peuvent comprendre des relations paradigmatiques classiques telles que la (quasi-)synonymie (p. ex. *conserver* → *préserver*), mais aussi d'autres relations telles que la dérivation syntaxique (p. ex. *conserver* → *conservation*).

Cette thèse vise à développer un cadre méthodologique basé sur la sémantique distributionnelle pour l'analyse des relations lexicales à partir de corpus spécialisés. À cette fin, nous vérifions comment le choix, le paramétrage et l'interrogation d'un modèle distributionnel doivent tenir compte de divers facteurs liés au projet terminologique envisagé : le cadre descriptif adopté, les relations ciblées, la partie du discours des termes à décrire et la langue traitée (en l'occurrence, le français ou l'anglais).

Nous montrons que deux des relations les mieux détectées par l'approche distributionnelle sont la (quasi-)synonymie et la dérivation syntaxique, mais que les modèles qui captent le mieux ces deux types de relations sont très différents. Ainsi, les relations ciblées ont une influence importante sur la façon dont on doit paramétrer le modèle pour obtenir les meilleurs résultats possibles.

Un autre facteur à considérer est la partie du discours des termes à décrire. Nos résultats indiquent notamment que les relations entre verbes sont moins bien modélisées par cette approche que celles entre adjectifs ou entre noms.

Le cadre descriptif adopté pour un projet terminologique est également un facteur important à considérer lors de l'application de l'approche distributionnelle. Dans ce travail, nous comparons deux cadres descriptifs, l'un étant basé sur la sémantique lexicale et l'autre, sur la sémantique des cadres. Nos résultats indiquent que les méthodes distributionnelles détectent les termes qui évoquent le même cadre sémantique moins bien que certaines relations lexicales telles que la synonymie. Nous montrons que cet écart est attribuable au fait que les termes qui évoquent des cadres sémantiques comprennent une proportion importante de verbes et aux différences importantes entre les modèles qui produisent les meilleurs résultats pour la dérivation syntaxique et les relations paradigmatiques classiques telles que la synonymie.

En somme, nous évaluons deux modèles distributionnels différents, analysons systématiquement l'influence de leurs paramètres et vérifions comment cette influence varie en fonction de divers aspects du projet terminologique. Nous montrons de nombreux exemples de voisinages distributionnels, que nous explorons au moyen de graphes, et discutons les sources d'erreurs. Ce travail fournit ainsi des balises importantes pour l'application de méthodes distributionnelles dans le cadre du travail terminologique.

Mots clés : terminologie, sémantique distributionnelle, sémantique lexicale, sémantique des cadres, relations lexicales, évaluation.

ABSTRACT

Identifying semantic relations is one of the main tasks involved in terminology work. This task, which aims to establish links between terms whose meanings are related, can be assisted by computational methods, including those based on distributional semantics. These methods estimate the semantic similarity of words based on corpus data, which can help terminologists identify semantic relations.

The quality of the results produced by distributional methods depends on several decisions that must be made when applying them, such as choosing a model and selecting its parameters. In turn, these decisions depend on various factors related to the target application, such as the types of semantic relations one wishes to identify. These can include typical paradigmatic relations such as (near-)synonymy (e.g. *preserve* → *protect*), but also other relations such as syntactic derivation (e.g. *preserve* → *preservation*).

This dissertation aims to further the development of a methodological framework based on distributional semantics for the identification of semantic relations using specialized corpora. To this end, we investigate how various aspects of terminology work must be accounted for when selecting a distributional semantic model and its parameters, as well as those of the method used to query the model. These aspects include the descriptive framework, the target relations, the part of speech of the terms being described, and the language (in this case, French or English).

Our results show that two of the relations that distributional semantic models capture most accurately are (near-)synonymy and syntactic derivation. However, the models that produce the best results for these two relations are very different. Thus, the target relations are an important factor to consider when choosing a model and tuning it to obtain the most accurate results.

Another factor that should be considered is the part of speech of the terms that are being worked on. Among other things, our results suggest that relations between verbs are not captured as accurately as those between nouns or adjectives by distributional

semantic models.

The descriptive framework used for a given project is also an important factor to consider. In this work, we compare two descriptive frameworks, one based on lexical semantics and another based on frame semantics. Our results show that terms that evoke the same semantic frame are not captured as accurately as certain semantic relations, such as synonymy. We show that this is due to (at least) two reasons: a high percentage of frame-evoking terms are verbs, and the models that capture syntactic derivation most accurately are very different than those that work best for typical paradigmatic relations such as synonymy.

In summary, we evaluate two different distributional semantic models, we analyze the influence of their parameters, and we investigate how this influence varies with respect to various aspects of terminology work. We show many examples of distributional neighbourhoods, which we explore using graphs, and discuss sources of noise. This dissertation thus provides important guidelines for the use of distributional semantic models for terminology work.

Keywords: terminology, distributional semantics, lexical semantics, frame semantics, semantic relations, evaluation.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

AD	analyse distributionnelle
ANTI	antonymes
AP	*average precision* (précision moyenne)
DRV	dérivés syntaxiques
EN	anglais
FR	français
HYP	relations hiérarchiques (hyperonymie et hyponymie)
JJ	adjectif
MAP	*mean average precision* (moyenne des précisions moyennes)
NN	nom
PDD	partie du discours
PPV	plus proche voisin
QSYN	quasi-synonymes
TAL	traitement automatique de la langue
UL	unité lexicale
VV	verbe
W2V	`word2vec`

REMERCIEMENTS

Je tiens d'abord à exprimer ma gratitude à mon directeur de recherche, Patrick Drouin, sans qui je n'aurais sans doute pas poursuivi mes études après avoir complété mon baccalauréat en traduction. Je le remercie de m'avoir donné la chance de travailler à l'Observatoire de linguistique Sens-Texte (OLST), de m'avoir donné le goût de la recherche et de m'avoir guidé et encouragé tout au long de mon doctorat. Je remercie aussi Dominic Forest, qui a gracieusement accepté de codiriger ce travail, et dont les commentaires m'ont permis d'y apporter des améliorations significatives.

J'aimerais également témoigner ma reconnaissance à Marie-Claude L'Homme, avec qui j'ai eu la chance de collaborer et que j'ai eu le plaisir de côtoyer dans quelques conférences. Je la remercie également de l'appui qu'elle m'a fourni tout au long de mon séjour à l'OLST.

Je remercie aussi Philippe Langlais et Vincent Claveau, qui ont gentiment accepté de faire partie du jury.

Je suis redevable au Conseil de recherches en sciences humaines (CRSH) pour l'octroi d'une bourse doctorale qui m'a permis de me consacrer entièrement à ma recherche. Je remercie également le Département de linguistique et de traduction et la Faculté des études supérieures et postdoctorales de l'Université de Montréal pour les bourses qu'ils m'ont accordées, ainsi que le gouvernement du Canada pour celles qui m'ont été offertes dans le cadre du Programme de renforcement du secteur langagier au Canada.

Je remercie enfin mes parents, mes amis et tous mes collègues à l'OLST.

CHAPITRE 1

INTRODUCTION

1.1 Problématique générale

Cette thèse concerne l'utilisation d'outils informatiques dans le cadre du travail terminologique. L'évolution de plus en plus rapide des connaissances dans les divers domaines de l'activité humaine, qui s'accompagne d'une évolution constante de la terminologie de ces domaines, fait en sorte qu'il est désormais nécessaire de faire appel à des méthodes computationnelles pour répondre aux besoins en matière de terminologie. Heureusement, l'accès à des quantités toujours plus importantes de textes spécialisés en format électronique et à des ordinateurs de plus en plus puissants, ainsi que les progrès réalisés dans le domaine du traitement automatique de la langue (TAL) et des domaines connexes, favorisent l'utilisation de méthodes computationnelles en terminologie.

Ainsi, de nos jours, le travail terminologique s'appuie sur des outils informatiques à chacune de ses étapes, comme le souligne L'Homme [111, p. 47]. Par exemple,

- la collecte de la documentation peut exiger que l'on fasse appel à un logiciel de reconnaissance optique des caractères ou à un aligneur, entre autres ;
- les extracteurs de termes sont utilisés pour faciliter le repérage des termes à décrire ;
- les concordanciers sont utilisés lors de la collecte des données terminologiques pour récupérer des contextes, des éléments de définition et d'autres données terminologiques ;
- les systèmes de gestion de bases de données sont utilisés pour créer et gérer les fiches terminologiques ; en outre, les mémoires de traduction comprennent souvent des outils qui facilitent la gestion terminologique dans le cadre de la traduction.

Les problématiques abordées dans cette thèse concernent la collecte et l'analyse des données terminologiques, en particulier :

- l'identification des relations lexicales auxquelles participent les termes d'un domaine de spécialité (p. ex. *bioéthanol* est un hyponyme de *biocarburant*) ;
- la recherche de cadres sémantiques caractéristiques du domaine (p. ex. un Patient [*écosystème, population,* ...] est exposé à une Menace [*changement, perturbation,* ...]).

Il existe des techniques susceptibles de faciliter ces tâches, notamment celles de la sémantique distributionnelle, mais l'utilisation de ces techniques ne semble pas très répandue dans le domaine de la terminologie. Du moins, si certains y font appel, il n'existe pas à l'heure actuelle un cadre méthodologique clairement défini pour l'utilisation des méthodes distributionnelles en terminologie. Ainsi, il n'est pas aisé de savoir de quels facteurs dépend la réussite de l'approche distributionnelle. Par exemple, dépend-elle des relations que l'on souhaite identifier ? de la partie du discours des termes à décrire ? de la langue traitée ? du cadre descriptif adopté pour un projet terminologique donné ? Par ailleurs, comment doit-on paramétrer les modèles distributionnels pour obtenir les meilleurs résultats possibles dans le cadre d'un projet terminologique ? Est-ce que ces décisions doivent tenir compte des autres facteurs mentionnés ci-dessus ?

En abordant ces questions, cette thèse fournit des balises importantes pour la mise en œuvre de l'approche distributionnelle dans le cadre de l'élaboration de ressources terminologiques basées sur des cadres descriptifs particuliers. Ce travail pourra servir, entre autres, à l'exploitation ou à la mise au point d'outils pouvant assister non seulement l'élaboration de ressources terminologiques, mais aussi la mise à jour et l'enrichissement de telles ressources au moyen de l'exploitation de nouvelles données textuelles. Cette dernière possibilité découle du fait que les méthodes distributionnelles induisent des liens de similarité sémantique directement d'un corpus et n'exigent aucune ressource lexicale externe.

Sur le plan applicatif, cette thèse concerne également l'analyse des textes spécialisés du domaine de l'environnement et la description des termes utilisés dans ce domaine. En effet, les résultats de l'application des méthodes distributionnelles sont évalués, entre autres, en comparant les liens de similarité détectés aux relations décrites dans des dictionnaires spécialisés du domaine de l'environnement. Ce domaine présente des enjeux qui rendent cette application particulièrement importante. Il présente par ailleurs des caractéristiques pouvant rendre cette application difficile :

- il s'agit d'un domaine multidisciplinaire faisant appel à des concepts issus de disciplines nombreuses et parfois très différentes les unes des autres (la chimie, la biologie, la climatologie, la géographie, les sciences économiques, etc.) ;
- les textes produits dans ce domaine sont destinés à des publics hétérogènes, présentent différents niveaux de spécialisation et appartiennent à une variété de genres.

Un cadre méthodologique qui faciliterait la collecte et l'analyse des données terminologiques, en l'occurrence l'identification de relations lexicales et de cadres sémantiques, serait donc particulièrement utile dans ce domaine. Cette thèse, qui vise à développer ce cadre méthodologique en exploitant les techniques de la sémantique distributionnelle, aura ainsi des retombées importantes.

1.2 Structure de la thèse

La suite de ce travail est organisée de la façon suivante :

- Nous expliquons au chapitre 2 les deux cadres descriptifs que nous utilisons pour caractériser différentes relations lexicales que l'on peut chercher à identifier au moyen de l'approche distributionnelle.
- Nous présentons au chapitre 3 une revue de la littérature sur la sémantique distributionnelle et son utilisation en terminologie.

- Au chapitre 4, nous énonçons les problématiques spécifiques que nous abordons dans cette thèse, puis nous formulons deux hypothèses qui guideront notre méthodologie par la suite. Nous y expliquons également les objectifs de ce travail.

- Notre méthodologie est décrite au chapitre 5. Celle-ci repose sur la construction de modèles sémantiques distributionnels et l'évaluation de ces modèles au moyen de données de référence que nous avons extraites de dictionnaires spécialisés.

- Enfin, l'analyse des résultats de cette expérience est présentée aux chapitres 6 et 7. Le chapitre 6 porte sur l'évaluation des modèles distributionnels et l'influence des différents facteurs dont nous avons tenu compte dans ce travail, tels que le cadre descriptif adopté ou les relations que l'on souhaite identifier. Le chapitre 7 porte plus spécifiquement sur une façon particulière d'interroger les modèles distributionnels, à savoir les graphes de voisinage distributionnel ; nous y présentons une analyse des paramètres de ces graphes ainsi que des exemples visuels de voisinages distributionnels qui illustrent bien, à notre avis, l'utilité de l'approche distributionnelle dans le cadre du travail terminologique.

CHAPITRE 2

CADRES DESCRIPTIFS

Introduction

Ce chapitre porte sur les deux cadres descriptifs que nous utilisons pour caractériser différents types de relations lexicales que l'on peut chercher à identifier au moyen des méthodes distributionnelles. Nous exploitons dans ce travail deux cadres descriptifs utilisés dans le domaine de la terminologie :

1. l'approche lexico-sémantique à la terminologie ;
2. la sémantique des cadres.

De plus, nous utilisons des ressources terminologiques reflétant ces deux cadres descriptifs, à savoir le DiCoEnviro et le Framed DiCoEnviro respectivement, afin d'évaluer des modèles sémantiques distributionnels. Ainsi, l'application et l'évaluation des méthodes distributionnelles sont balisées par ces deux cadres descriptifs.

Avant de décrire ces deux cadres descriptifs et les deux ressources correspondantes, soulignons d'abord que, sur le plan théorique, ce travail présuppose une conception de la terminologie qui est ancrée dans les textes plutôt que dans la structure conceptuelle d'un domaine. Il suppose donc que l'on adhère aux principes d'approches contemporaines à la terminologie telles que la terminologie textuelle [29] plutôt qu'à l'optique conceptuelle au cœur de la théorie générale de la terminologie (TGT) élaborée par Eugen Wüster [193].

2.1 L'approche lexico-sémantique

2.1.1 Description de l'approche

L'approche lexico-sémantique à la terminologie [112] est un cadre descriptif ancré dans le travail pratique des terminologues et basé, entre autres, sur l'approche descriptive à la sémantique lexicale proposée par Cruse [44] et sur la lexicologie explicative et combinatoire [129]. Il fournit des outils concrets qui permettent de répondre à des questions fondamentales que soulève tout travail de description terminologique, notamment :

- Comment choisit-on les termes qui feront l'objet d'une description ?
- Comment doit-on décrire ces termes et encoder les descriptions résultantes ?

Cette approche a émergé d'une période de remise en question des principes de la TGT, période qui a vu apparaître de nouvelles approches telles que la socioterminologie [78], la terminologie textuelle [29] et l'approche sociocognitive [179]. Ces approches ont remis en cause de nombreux aspects de la TGT, y compris :

- sa visée normalisatrice ;
- l'importance et le caractère immuable qu'elle attribuait aux concepts ;
- le fait qu'elle privilégiait (en théorie) une démarche onomasiologique, qui part du concept, auquel on associe une étiquette linguistique *a posteriori*.

Selon l'approche lexico-sémantique, le terme n'est pas une simple étiquette linguistique apposée à un concept posé *a priori*, mais une unité lexicale, telle que la définit Cruse [44], dotée d'un sens spécialisé. L'unité lexicale (UL), objet d'étude de la sémantique lexicale, est une association entre une forme linguistique[1] et un sens. Une UL peut

[1] Plus précisément, on parle ici d'une *forme lexicale*, qui représente un mot ainsi que ses variantes flexionnelles : « a lexical unit is the union of a single sense with a lexical form ; a lexical form is an abstraction from a set of word forms (or alternatively – it is a family of word forms) which differ only in respect of inflections » [44, p. 80].

correspondre à un mot appartenant à l'une des classes ouvertes (noms, verbes, adjectifs et adverbes) ou à une expression non compositionnelle telle qu'une locution [44, ch. 1–2] ; les mots des classes fermées (p. ex. les prépositions) seraient plutôt des unités grammaticales.

Ainsi, l'approche lexico-sémantique à la terminologie s'intéresse non seulement aux noms et aux syntagmes nominaux, unités privilégiées par la TGT, mais à toute UL ayant un sens spécialisé, y compris les verbes, les adjectifs et les adverbes. De plus, les syntagmes nominaux ne sont décrits que si leur sens n'est pas compositionnel (c'est-à-dire qui peut être déduit du sens de leurs composantes) « puisque les unités lexicales individuelles qui les composent font l'objet de descriptions à part entière » [112, p. 1126].

L'approche lexico-sémantique est une terminologie textuelle (ou une terminologie basée sur corpus), qui dégage le sens des termes à partir de leurs contextes linguistiques[2]. En effet, l'approche lexico-sémantique propose d'adopter une démarche sémasiologique plutôt qu'onomasiologique, c'est-à-dire une démarche qui prend les formes linguistiques comme point de départ plutôt que les concepts, et de considérer le terme comme un construit résultant de l'analyse des occurrences d'une UL dans un corpus.

De plus, cette approche fournit des critères pour l'identification des termes qui feront l'objet d'une description, ainsi que des tests lexico-sémantiques pour distinguer les différents sens d'un terme et pour identifier les relations lexicales entre les termes[3].

Suivant cette approche, une relation lexicale telle qu'un lien de synonymie est une relation entre deux UL, c'est-à-dire des acceptions particulières de deux termes. Par ailleurs, deux termes sont considérés comme des synonymes (absolus) seulement s'ils

[2]L'approche descriptive à la sémantique lexicale proposée par Cruse [44] est également contextuelle. Par exemple, les relations lexicales sont identifiées en fonction des relations ensemblistes (telles que l'identité et l'inclusion) entre les contextes dans lesquels les mots sont (ou peuvent être) utilisés : si l'ensemble des contextes possibles de deux mots participent à une relation d'identité, les deux mots sont des synonymes (absolus) ; l'hyponymie correspondrait plutôt à une relation d'inclusion.

[3]Il est intéressant de noter que Cruse [44, ch. 1] exploite des tests qu'il appelle *diagnostic frames* pour vérifier s'il existe une relation lexicale particulière entre deux UL ; ces tests correspondent à ce que d'autres appellent des *patrons lexico-syntaxiques* [93]. Par exemple, le patron *Xs and other Ys* permettrait de déterminer si le mot *X* est un hyponyme du mot *Y*. Les patrons lexico-syntaxiques sont d'ailleurs utilisés pour détecter automatiquement des relations lexicales, comme nous l'expliquons au chapitre 3.

peuvent se combiner avec les mêmes mots en contexte ; les termes ayant le même sens, mais des collocations différentes, ne sont pas considérés comme des synonymes absolus, mais des quasi-synonymes.

L'approche lexico-sémantique s'intéresse à une variété de relations lexicales. Celles-ci peuvent être de type paradigmatique ou syntagmatique. Les relations paradigmatiques classiques (p. ex. la synonymie, l'antonymie, l'hyperonymie et l'hyponymie) peuvent être considérées comme des relations de *substituabilité* ; ainsi, il est possible de remplacer, dans un contexte donné, un mot par un de ses (quasi-)synonymes ou par un mot dont le sens est plus générique ou spécifique. En revanche, les relations syntagmatiques relient des UL qui apparaissent typiquement ensemble dans un même contexte (p. ex. les collocations).

En décrivant, à partir de données textuelles, une variété de relations paradigmatiques et syntagmatiques entre termes, cette approche mène à la description d'un *réseau lexical*.

2.1.2 Le DiCoEnviro

L'approche lexico-sémantique à la terminologie a été mise en pratique dans le cadre de l'élaboration de ressources lexicales telles que le DiCoEnviro[4] [115], un dictionnaire spécialisé du domaine de l'environnement élaboré à l'Observatoire de linguistique Sens-Texte (Université de Montréal). Ce dictionnaire multilingue décrit des termes appartenant à divers sous-domaines du domaine de l'environnement, tels que les changements climatiques, les énergies renouvelables, les transports électriques et la gestion des matières résiduelles.

Suivant l'approche lexico-sémantique, ce dictionnaire vise à décrire le sens et le comportement linguistique des termes et à expliciter les relations lexicales auxquelles ils participent. Il décrit des termes appartenant aux catégories du nom, du verbe, de l'adjectif et de l'adverbe, ainsi que certains termes complexes (p. ex. *effet de serre*).

[4] `http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi` (page consultée le 20 novembre 2016).

Le développement du DiCoEnviro (qui est actuellement en cours) se décline en plusieurs étapes : l'extracteur de termes TermoStat [52] est utilisé afin d'obtenir des candidats-termes à partir d'un corpus du domaine de l'environnement[5] ; les candidats-termes sont ensuite validés au moyen de critères d'identification de termes et leurs différentes acceptions sont distinguées au moyen de tests lexico-sémantiques [111, ch. 2]. Chaque acception d'un terme (c'est-à-dire chaque UL) fait l'objet d'un article séparé.

La description d'un terme, basée sur une analyse de ses occurrences dans le corpus, contient sa structure actancielle, des phrases annotées illustrant son comportement dans des contextes linguistiques réels (la méthodologie d'annotation étant basée sur celle utilisée dans le projet FrameNet, que nous décrivons à la section 2.2), ainsi que des relations lexicales. Les termes reliés décrits dans un article du DiCoEnviro peuvent comprendre :

- des hyponymes[6] (p. ex. *carburant→biogaz*) ;
- des hyperonymes[7] (p. ex. *diesel→carburant*) ;
- des synonymes (p. ex. *diesel→gazole*) ;
- des quasi-synonymes et autres sens voisins (p. ex. *emprisonner→piéger*) ;
- des antonymes ou contraires[8] (p. ex. *réchauffement→refroidissement*) ;
- des méronymes (p. ex. *éolienne→nacelle*, *éolienne→parc d'~*) ;
- des dérivés syntaxiques[9] (p. ex. *accélération→accélérer*, *océan→marin*) ;

[5] Ce corpus contient à la fois des textes spécialisés et de vulgarisation.

[6] Dans le DiCoEnviro, les hyponymes sont appelés des *sortes de*.

[7] Dans le DiCoEnviro, les hyperonymes sont appelés des *génériques*.

[8] Selon Cruse [44, ch. 9–10], les antonymes sont une sous-catégorie de la famille des contraires, qui comprend également d'autres relations, telles que les complémentaires et les réversifs.

[9] Dans le cadre descriptif de la lexicologie explicative et combinatoire, deux mots ayant le même sens, mais appartenant à des parties du discours différentes, sont appelés des *dérivés syntaxiques* [129, p. 133]. Ce sont souvent des dérivés morphologiques, mais pas toujours (p. ex. *combustion→brûler*). Soulignons que les fonctions lexicales qui encodent ces relations (p. ex. S_0 et V_0) font partie des fonctions lexicales paradigmatiques [155, p. 131] ; en revanche, selon Cruse [44, ch. 1], les affinités paradigmatiques concernent seulement les mots appartenant à la même catégorie grammaticale. Ainsi, les dérivés syntaxiques ne font pas partie de ce que nous appelons les relations paradigmatiques *classiques*.

- des collocations (p. ex. *carbone→piéger*).

L'équipe qui assure le développement de cette ressource est dirigée par Marie-Claude L'Homme, qui valide chaque article, entre autres. Cette validation, ainsi qu'un système de statuts qui indiquent l'état d'avancement des articles, assurent la cohérence et la validité de l'ensemble des données encodées par les multiples annotateurs participant au projet. Ce dictionnaire constitue ainsi une excellente référence pour la détection de relations lexicales dans le domaine de l'environnement.

2.2 La sémantique des cadres

2.2.1 Description de l'approche

La sémantique des cadres (*frame semantics*) [71] est un cadre de recherche empirique sur la sémantique qui fournit une méthodologie de description du lexique basée sur la notion de « cadre sémantique » (*frame*).

Un cadre sémantique est un scénario conceptuel qui décrit une situation, un objet ou un événement ainsi que ses participants et accessoires[10]. En effet, les cadres sémantiques peuvent représenter des événements (p. ex. Social_event[11]), des actions (p. ex. Putting_out_fire), des propriétés (p. ex. Being_dry) ou des processus (p. ex. Becoming_dry), entre autres.

Concrètement, la description d'un cadre sémantique comprend une ou plusieurs unités lexicales (au sens de Cruse [44]) qui *évoquent* ce cadre et un ou plusieurs *éléments de cadre*[12] qui représentent les participants du scénario en question [7]. On peut ainsi décrire le sens des mots en décrivant le cadre sémantique qu'ils évoquent ou le rôle qu'ils jouent au sein d'un cadre sémantique particulier.

[10] Cette définition est adaptée de celle de Ruppenhofer et al. [163, p. 5] : « a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props ».

[11] Les exemples de cadres présentés ici proviennent de FrameNet, ressource que nous décrivons dans cette section.

[12] Certains éléments de cadre sont *centraux* (en anglais, *core frame elements*) et d'autres, *périphériques* (*non-core frame elements*) ; ces équivalents français sont ceux utilisés par L'Homme [114]. Des critères précis sont utilisés pour distinguer ces deux types d'éléments de cadre [4].

Prenons un exemple donné par Fillmore [71], celui d'une transaction commerciale (Commercial_transaction). Ce cadre sémantique est évoqué par des UL telles que les verbes *acheter*, *vendre* et *payer*, et les éléments de ce cadre comprennent l'Acheteur, le Vendeur, les Biens et l'Argent. Il serait impossible, selon Fillmore, de comprendre le sens des UL (*acheter, vendre, payer*) sans comprendre le cadre sémantique qu'elles évoquent (en l'occurrence, celui d'un échange commercial) ainsi que ses participants [71, p. 111] ; il en serait de même dans les domaines spécialisés de l'activité humaine [71, p. 120].

La sémantique des cadres était d'abord axée sur les verbes [7], mais elle permet également de décrire des UL appartenant à d'autres parties du discours. Ainsi, un même cadre peut comprendre des UL de différentes parties du discours. Par exemple, le cadre Commercial_transaction est évoqué non seulement par des verbes tels que *vendre*, mais aussi par des noms tels que *vente*.

Le cadre de recherche proposé par Fillmore a mené à l'élaboration de la ressource FrameNet[13] [8], un dictionnaire de la langue anglaise basé sur la sémantique des cadres.

La méthodologie utilisée dans le cadre du projet FrameNet pour décrire un cadre sémantique est un processus qui commence par l'élaboration d'une représentation schématique de ce cadre et l'énumération d'UL qui l'évoquent [73]. L'étape suivante consiste à annoter des phrases qui contiennent ces UL, en identifiant les éléments de cadre qui sont réalisés dans chaque phrase et en indiquant pour chacune de ces réalisations le type de construction syntaxique dont il s'agit et sa fonction grammaticale. Ces phrases annotées, qui sont extraites d'un corpus [72], illustrent la réalisation des UL et des éléments de cadre dans des contextes linguistiques réels.

Ainsi, FrameNet encode la similarité de groupes d'UL sur les plans sémantique et syntaxique : « Frames are generalizations over groups of words which describe similar states of affairs and which could be expected to share similar sets of roles, and (to some extent) similar syntactic patterns for them. » [7, p. 2]

La méthodologie de FrameNet comprend également l'établissement de relations entre

[13] `https://framenet.icsi.berkeley.edu/fndrupal/` (page consultée le 21 juin 2016).

les différents cadres sémantiques. Par exemple, le cadre Commercial_transaction est relié au cadre Transfer. C'est qu'une transaction commerciale est en fait un processus complexe qui comporte deux « sous-cadres » correspondant au cadre Transfer : l'Acheteur donne l'Argent au Vendeur et le Vendeur donne les Biens à l'Acheteur [72].

Le Tableau 2.I présente une partie de la description du cadre sémantique Emitting[14] (en français, *émettre*) que l'on retrouve dans FrameNet, qui comprend la définition du cadre, les UL, les éléments de cadre, des phrases annotées et les relations auxquelles le cadre participe.

Soulignons enfin que FrameNet ne décrit pas explicitement des relations lexicales telles que la synonymie ou l'antonymie, bien que l'on puisse déduire une partie de cette information en analysant les UL qui évoquent un cadre ou les relations entre les cadres [72, p. 248].

2.2.2 Le Framed DiCoEnviro

Bien qu'elle contienne des cadres sémantiques liés à des domaines spécialisés, la ressource lexicale FrameNet vise une description générale de la langue anglaise : « the FrameNet team have assumed that technical vocabulary, whose definitions are established by domain experts, will be handled in terminologies for each domain » [7, p. 3].

Des chercheurs ont donc proposé d'appliquer les principes de la sémantique des cadres ou la méthodologie de FrameNet à des domaines spécialisés. Par exemple, la méthodologie de FrameNet a été appliquée au domaine biomédical [50] et à celui du droit [185] ; Schmidt [170] a exploité la sémantique des cadres et la sémantique lexicale afin de créer une ressource lexicale multilingue sur le soccer ; la sémantique des cadres a été utilisée pour établir des équivalences interlinguistiques dans le domaine du droit [154] et pour caractériser les usages de verbes dans des corpus médicaux de différents niveaux de spécialisation [187] ; Faber [60] a proposé un cadre sémantique général caractérisant le domaine de l'environnement, *l'événement environnemental*, afin de dé-

[14] `https://goo.gl/M1nsNQ` (page consultée le 26 avril 2016).

Emitting	
Definition	In this frame a Source_emitter discharges its Emission along a Path or to a Goal.
Examples	(1) [The factory]$_{\text{Source\_emitter}}$ EXUDES [toxic fumes]$_{\text{Emission}}$ [from three large chimneys]$_{\text{Sub-source}}$. (2) [Light]$_{\text{Emission}}$ is EMITTED [from the tower]$_{\text{Source\_emitter}}$. (3) [Noxious odors]$_{\text{Emission}}$ are EMITTED [from his feet]$_{\text{Sub-source}}$. In (3) we would tag 'his' as Source_emitter on a second layer because the inalienable possession on the Sub-source indicates the Source_emitter
Core Frame Elements	Emission : The Emissions are discharged out of the Source_emitter. Source_emitter : The Source_emitter is the item from which the Emission are [*sic*] discharged.
Non-Core Frame Elements	Area : The Area is used for expressions which describe a general area in which emission takes place when the motion is understood to be irregular or not to consist of a single, linear path. Carrier : This is the means of conveyance of the Emission. [The acorn seeds]$_{\text{Emission}}$ are EXCRETED [in the feces of squirrels]$_{\text{Carrier}}$. [. . .] Sub-source : The Sub-source is the subregion of the Source_emitter which emits the Emission. Time (Semantic Type : Time) : This FE identifies the Time when the Source_emitter emits the Emission. [. . .]
Frame-frame Relations	Is Causative of : Emanating
Lexical Units	*discharge.v, emanate.v, emit.v, excrete.v, exhale.v, exude.v, give off.v, radiate.v, secrete.v, void.v*

Tableau 2.I – Un extrait du cadre Emitting dans FrameNet.

crire les différents processus et événements propres à ce domaine.

La sémantique des cadres est une approche tout à fait adaptée à la terminologie, comme le soulignent L'Homme et Robichaud [116]. Les approches classiques à la terminologie, qui sont axées sur les termes dénotant des entités, n'offrent pas de moyen de représenter adéquatement les propriétés linguistiques des termes qui évoquent plutôt des processus, des événements ou des propriétés, notamment le fait qu'ils appellent des arguments (p. ex. le verbe *menacer* a deux arguments : X *menace* Y). Ces termes occupent une place importante dans de nombreux domaines spécialisés, y compris le domaine de l'environnement. Beaucoup de terminologues s'entendent pour dire qu'il est important de décrire non seulement les connaissances ou concepts représentés par les termes, mais aussi leurs propriétés linguistiques. La sémantique des cadres et la méthodologie de FrameNet fournissent une solution à ce problème, en permettant de décrire à la fois les connaissances nécessaires à la compréhension des termes et leurs propriétés linguistiques.

La sémantique des cadres a d'ailleurs le potentiel de faciliter plusieurs applications du traitement automatique de la langue (TAL) ; en fait, un des objectifs principaux du développement de FrameNet était de montrer que ce genre de descriptions pouvait servir des applications du TAL [72]. Ainsi, l'exploitation de la sémantique des cadres en terminologie est un terrain fertile à l'heure actuelle.

Une étude visant à découvrir des cadres sémantiques à partir d'une ressource lexicale spécialisée existante (le DiCoEnviro, que nous avons décrit à la section 2.1.2) a été réalisée récemment par L'Homme et al. [117], l'hypothèse étant que les UL qui partagent des propriétés lexico-sémantiques selon cette ressource évoquent le même cadre :

> In this work, we hypothesize that terms sharing argument structures in specific subject fields (along with other lexico-semantic properties) evoke semantic frames. [...] The data used in this analysis is extracted from a database that was not compiled with a view to describing terms as lexical units that evoke frames ; however, we believe that these can be discovered afterwards based

> on the lexico-semantic properties of lexical units. Hence, we assume that our descriptions can lend themselves to a frame-like analysis. [117, p. 1365]

L'étude a confirmé qu'il est possible de découvrir *a posteriori* des cadres sémantiques dans une ressource décrivant les propriétés lexico-sémantiques des termes. Ainsi, contrairement à d'autres travaux appliquant le cadre descriptif de la sémantique des cadres à des domaines spécialisés, cette étude propose d'adopter une approche dite *ascendante* (*bottom-up*) à la description de cadres sémantiques [116] : on décrit d'abord les termes, puis on découvre des cadres à partir de ces descriptions.

L'analyse au cœur de cette étude consistait à établir des correspondances entre certains des termes décrits dans le DiCoEnviro (à savoir des verbes et des noms dénotant des processus ou des activités, ainsi que quelques adjectifs) et les cadres sémantiques décrits dans FrameNet. Des associations entre un terme et un cadre sémantique ont été établies lorsqu'il était possible de mettre en correspondance l'ensemble des éléments centraux du cadre (*core frame elements*) et l'ensemble des actants sémantiques du terme. Dans les cas où aucun cadre équivalent n'existait dans FrameNet, de nouveaux cadres ont été définis.

Par la suite, des relations entre les cadres sémantiques ont été établies [116], les cadres formant ainsi des scénarios complexes représentant divers aspects du domaine de l'environnement.

Ces descriptions ont été encodées dans une ressource lexicale spécialisée appelée le *Framed DiCoEnviro*[15]. À l'heure actuelle, le Framed DiCoEnviro décrit environ 120 cadres sémantiques du domaine de l'environnement. Chaque cadre (à l'exception de quelques cadres dits *non lexicaux*) contient ou contiendra les éléments suivants :

- une définition ;
- les UL (en trois langues, à savoir le français, l'anglais et l'espagnol) qui évoquent

[15] `http://olst.ling.umontreal.ca/dicoenviro/framed/index.php` (page consultée le 25 avril 2016).

le cadre ;

- les éléments de cadre ;
- des phrases annotées ;
- les relations auxquelles le cadre participe ;
- des notes indiquant notamment si le cadre a été adapté de FrameNet ou défini spécifiquement pour le domaine de l'environnement ;
- des liens vers la description des UL dans le DiCoEnviro.

Le Tableau 2.II présente une partie de la description du cadre sémantique Emitting que l'on retrouve dans le Framed DiCoEnviro.

2.3 Synthèse

Dans ce chapitre, nous avons présenté les deux cadres descriptifs que nous exploitons dans cette thèse afin de caractériser les différentes relations lexicales que l'on peut chercher à détecter au moyen des méthodes distributionnelles : l'approche lexico-sémantique à la terminologie et la sémantique des cadres. En effet, nous vérifierons non seulement si les méthodes distributionnelles permettent de détecter des relations paradigmatiques classiques telles que la synonymie ou l'antonymie, mais aussi si elles permettent de détecter des termes qui évoquent le même cadre sémantique, comme nous l'expliquerons au chapitre 4. Ainsi, nous considérons que l'appartenance au même cadre sémantique constitue en soi une relation lexicale, celle-ci pouvant correspondre, si on la considère du point de vue de l'approche lexico-sémantique, à diverses relations telles que la synonymie ou l'antonymie ; à ce sujet, nous analysons à la section 5.6.1.2 les différentes relations auxquelles participent les termes évoquant le même cadre sémantique.

Emitting	
Definition	An Agent or a Source discharges a Patient into a Destination.
Examples	when flourishing, [forests]$_{\text{Source}}$ absorb about as much [carbon]$_{\text{Patient}}$ as [they]$_{\text{Source}}$ EMIT [Aircrafts]$_{\text{Agent}}$ EMIT [gases and particles]$_{\text{Patient}}$ directly [into the upper troposphere and lower stratosphere]$_{\text{Destination}}$ [...] Currently, [warm tropical ocean waters brought north eastwards by this system]$_{\text{Source}}$ RELEASE [enough heat]$_{\text{Patient}}$ [to the air]$_{\text{Destination}}$ to keep temperatures over much of Europe some 10°C warmer than land areas at similar latitudes in other parts of the Northern Hemisphere [...] D' autres mesures consistent à [...] réduire les ÉMISSIONS [de méthane]$_{\text{Patient}}$ [en provenance des décharges]$_{\text{Source}}$ [...]
Notes	This frame in based on Emitting in FrameNet. The number of actants vs. core FEs differs. There is an Agent/Source alternation.
Participants (core)	Agent, Source, Patient, Destination
Participants (non-core)	Manner, Direction, Time, Substitute, Method, Frequency, Purpose, Instrument, Duration, Location, Value, Result, Reason, Path, Degree, Beginning, End
Frame Relations	Precedes : Soaking_up, Reflecting Is Preceded by : Soaking_up
English Lexical Units	emission$_1$, emission$_{2.1}$, emit$_1$, emit$_2$, inject$_1$, radiate$_1$, release$_{1a}$, release$_{1b}$, release$_{1b.1}$
French Lexical Units	injecter$_1$, libération$_1$, libérer$_{1a}$, libérer$_{1b}$, produire$_1$, émettre$_1$, émettre$_2$, émission$_1$, émission$_{2.1}$
Spanish Lexical Units	emitir$_2$

Tableau 2.II – Un extrait du cadre Emitting dans le Framed DiCoEnviro.

Nous avons également présenté des ressources terminologiques basées sur ces deux cadres descriptifs, le DiCoEnviro et le Framed DiCoEnviro, que nous utiliserons comme références afin d'évaluer des modèles sémantiques distributionnels.

Au chapitre 3, nous présenterons une revue de la littérature sur la détection automatique de relations lexicales, et les méthodes de la sémantique distributionnelle en particulier. La sémantique distributionnelle joue un rôle primordial dans ce travail, mais il est important de se rappeler que ce rôle est d'ordre méthodologique, plutôt que théorique ou descriptif : l'approche distributionnelle est considérée comme un moyen de faciliter l'élaboration ou l'enrichissement d'une ressource terminologique basée sur l'approche lexico-sémantique ou sur la sémantique des cadres.

CHAPITRE 3

ÉTAT DE LA QUESTION

Introduction

Dans ce chapitre, nous présentons une revue de la littérature sur la sémantique distributionnelle et son utilisation en terminologie. L'accent est placé sur l'utilisation de la sémantique distributionnelle pour l'identification de relations lexicales, en particulier à partir de corpus spécialisés, ainsi que l'évaluation de ces techniques, qui joue un rôle très important dans cette thèse ; les trois premières sections de ce chapitre sont consacrées à ces sujets. Puis, nous présentons à la section 3.4 une façon particulière d'interroger les modèles distributionnels, que nous exploitons dans ce travail. Enfin, nous faisons un bref survol de la recherche sur l'acquisition automatique de cadres sémantiques à la section 3.5.

3.1 L'identification de relations lexicales

De nombreuses méthodes computationnelles ont été proposées pour faciliter l'identification de relations lexicales entre termes ou, autrement dit, la structuration de la terminologie d'un domaine.

Les relations que ces méthodes visent à détecter peuvent être de différents types, et les méthodes privilégiées dépendent du type de relations ciblées. Premièrement, les relations lexicales peuvent être divisées en relations paradigmatiques et syntagmatiques, comme nous l'avons expliqué au chapitre 2. Les différentes méthodes que nous décrivons dans cette section sont généralement utilisées pour détecter des relations de type paradigmatique[1], telles que la synonymie, l'antonymie, l'hyponymie, l'hyperonymie, la

[1] Soulignons que, selon Cruse [44, p. 86], les relations paradigmatiques occupent une place plus importante dans la recherche sur la sémantique lexicale que les relations syntagmatiques.

cohyponymie et différentes formes de variation (p. ex. la variation morphosyntaxique).

Si on peut diviser les relations lexicales en relations paradigmatiques et syntagmatiques, d'autres classifications sont possibles. Dans les travaux en sémantique distributionnelle, on distingue parfois la *similarité sémantique* et la *proximité sémantique* (en anglais, *semantic similarity* et *semantic relatedness*) [70]. Dans ce contexte, la similarité sémantique correspond aux relations liant des termes qui partagent des attributs sémantiques, en particulier la synonymie et la cohyponymie ; p. ex. *cyclone* et *ouragan* ont une similarité sémantique élevée parce qu'ils ont beaucoup d'attributs sémantiques en commun. En revanche, la proximité sémantique relie des termes dont les sens sont proches même s'ils n'ont pas beaucoup d'attributs sémantiques en commun, tels que les méronymes (p. ex *voiture* et *moteur*) [62].

Les relations que l'on peut chercher à détecter automatiquement comprennent également des relations qui sont propres à des domaines particuliers, que Manser [125] appelle *transversales* (ou *inter-hiérarchiques*), telles que la relation *a-pour-symptôme* dans le domaine médical.

Pour en revenir aux méthodes computationnelles utilisées pour détecter ces relations, celles-ci exploitent soit la structure interne des termes, soit les contextes dans lesquels les termes apparaissent dans un corpus[2] [125] ; Grabar et Zweigenbaum [80] appellent ces deux types de méthodes *internes* et *externes* respectivement.

Les méthodes internes comprennent :

- Les méthodes basées sur l'inclusion lexicale [26, 98], qui peut indiquer une relation hiérarchique (hyponymie ou hyperonymie). Par exemple, le terme *énergie solaire* contient le terme *énergie*, et il s'agit d'un hyponyme de ce dernier.

- Les méthodes basées sur la parenté morphologique [99, 196, 197]. Ces méthodes visent à détecter des dérivés morphologiques (p. ex. *polluer* et *pollueur*) ou des

[2] Il existe également des méthodes qui n'exploitent ni la structure interne des termes, ni un corpus. Par exemple, on peut estimer la similarité de deux mots en fonction des mots que partagent leur définition dans une ressource lexicale existante [25].

variantes morphosyntaxiques (p. ex. *changement du climat* et *changement climatique*).

- L'analyse compositionnelle, qui peut détecter un lien de synonymie entre deux termes complexes si toutes leurs composantes sont soit synonymiques, soit identiques [88, 89]. Si on prend, par exemple, les termes *mât d'éolienne* et *tour d'éolienne*, leur première composante sont des synonymes et leur deuxième composante sont identiques ; cette information permet de détecter le lien de synonymie entre les deux termes complexes.

En ce qui concerne les méthodes externes, qui exploitent les contextes dans lesquels les termes apparaissent dans un corpus, on peut en distinguer au moins deux types. Le premier exploite des indices linguistiques appelés *patrons lexico-syntaxiques* [5, 93, 126], c'est-à-dire des patrons constitués de mots reliant des groupes syntaxiques particuliers, qui suggèrent en contexte une relation spécifique entre deux ou plusieurs termes ; par exemple, le patron <syntagme nominal> *tels que* <syntagme nominal> suggère une relation d'hyponymie (p. ex. *les biocarburants tels que le bioéthanol*). Cette méthode est exploitée surtout pour la détection de relations hiérarchiques et de la méronymie [92]. Les patrons eux-mêmes peuvent être identifiés manuellement, mais il existe également des méthodes pour les détecter automatiquement [46, 58].

Le deuxième type de méthode externe exploite la distribution des contextes dans lesquels les mots apparaissent dans un corpus. Le cadre méthodologique dans lequel s'inscrivent ces techniques s'appelle la *sémantique distributionnelle* ; nous appellerons les modèles représentant les mots en fonction de la distribution de leurs contextes des *modèles sémantiques distributionnels*[3].

[3]Nous utilisons le terme *modèle sémantique distributionnel* pour désigner tout modèle associant aux mots des représentations produites à partir de la distribution de leurs contextes dans un corpus. Ces modèles peuvent être produits au moyen de techniques classiques telles que l'analyse distributionnelle, mais aussi d'autres techniques telles que les modèles de langue neuronaux. En anglais, on utilise de plus en plus le terme *distributional semantic model* dans ce sens ; voir, par exemple, les travaux de Baroni et al. [10] ou de Gyllensten et Sahlgren [85]. Nous utilisons donc le terme *modèle sémantique distributionnel*,

La sémantique distributionnelle est un champ de recherche très actif à l'heure actuelle, et les méthodes qui en sont issues sont utilisées dans de nombreuses applications [189, p. 2249]. La popularité des méthodes distributionnelles découle du fait qu'elles permettent d'obtenir, à partir d'un corpus, des représentations d'unités linguistiques qui rendent compte de la similarité sémantique entre les unités et qui peuvent être lues et traitées par un ordinateur dans le cadre de diverses applications. Cohen et Widdows [42] fournissent un bon survol des différentes applications dans lesquelles les méthodes distributionnelles ont été utilisées, qui comprennent la construction et l'enrichissement de thésaurus et d'ontologies, l'extraction d'information (en particulier dans le domaine biomédical), la recherche d'information, la désambiguisation sémantique, l'induction des usages de mots (*word sense induction*) et la correction automatique d'examens. L'approche distributionnelle peut également servir à identifier des équivalences interlinguistiques à partir de corpus multilingues [49, 131, 158].

Ainsi, la sémantique distributionnelle « tend à s'imposer comme un mode de représentation et d'exploitation incontournable dans les travaux sur le lexique et la sémantique lexicale » [61, p. 266].

Les méthodes distributionnelles sont avantageuses à de nombreux égards. Comme le soulignent Moskowich et Caplan [140, p. 147], leur mise en œuvre peut être reproduite à souhait (p. ex. sur un nouveau corpus) ; elles révèlent différents aspects du sens des mots tels qu'ils sont utilisés dans le corpus[4] ; et elles ne reposent pas sur des jugements subjectifs, mais seulement sur les données linguistiques contenues dans le corpus. En effet, les méthodes distributionnelles font un minimum de présuppositions sur la langue et constituent donc un excellent outil pour une approche descriptive et empirique à la sémantique : « By grounding the representations in actual usage data, it only represents what is *really there* in the current universe of discourse. » [165, p. 11]. En ce qui concerne

bien qu'il ne soit pas encore très courant en français. Soulignons également que le terme *modèle* désigne parfois la méthode utilisée pour construire un modèle.

[4] Il est intéressant de noter que, selon l'analyse distributionnelle réalisée par Bullinaria [32], une tomate est un légume plutôt qu'un fruit.

le fait qu'elles captent les spécificités des corpus, Pantel et Lin [144] soulignent que les méthodes distributionnelles permettent de combler les lacunes des dictionnaires compilés manuellement, notamment le fait qu'ils ne décrivent pas systématiquement les sens spécialisés, propres à des domaines particuliers. Elles permettent également de dégager des phénomènes que l'on ne peut observer qu'en s'appuyant sur un corpus de taille importante [94, p. 53].

Par contre, ces méthodes ne peuvent pas produire des descriptions sémantiques complètes et de bonne qualité sans l'intervention d'un humain [140, p. 147]. En outre, les méthodes distributionnelles classiques ne permettent pas de déterminer la nature de la relation à laquelle deux mots participent, seulement leur degré de similarité, bien qu'il existe des techniques permettant de traiter ce problème dans une certaine mesure [189, 191]. Fabre et Lenci [62] affirment que ces méthodes captent la similarité des mots à un niveau peu élevé de granularité, les paires de mots distributionnellement similaires pouvant correspondre à diverses relations de similarité ou de proximité sémantique.

Nous tenons à souligner une autre critique souvent formulée à l'endroit des modèles distributionnels, bien que nous ne soyons pas tout à fait d'accord avec celle-ci, à savoir qu'ils ne permettent pas de traiter la polysémie, puisqu'ils associent généralement à chaque mot une seule représentation, comme le soulignent Reisinger et Mooney [161, p. 109] : « A single "prototype" vector is simply incapable of capturing phenomena such as homonymy and polysemy. » Pour combler cette lacune, ils proposent une méthode qui consiste à construire plusieurs représentations différentes pour chaque mot, cette méthode permettant notamment de prendre en compte le contexte lorsqu'il s'agit de calculer la similarité de deux mots. Sahlgren [165] souligne aussi le problème de la polysémie, qui serait exacerbé dans le cas des corpus traitant plusieurs sujets ou domaines :

> This points to one of the most serious weaknesses in word-space methodology : it cannot differentiate between several simultaneous distributions. When the data is topically dispersed, word-space representations may become hybrids of different, possibly unrelated, meanings. In the worst-case

> scenario, the context vector fails to properly represent any one of the meanings involved. [165, p. 105]

En fait, comme nous le montrerons à la section 3.4, entre autres, il existe des façons de traiter la polysémie même si une seule représentation est associée à chaque mot[5]. Quoi qu'il en soit, il est vrai que l'approche distributionnelle suppose généralement que la similarité de deux mots peut être représentée par une mesure unique, alors qu'il est possible de comparer et de classifier les éléments du lexique en fonction de différents critères [160].

En somme, parmi les différentes méthodes computationnelles utilisées pour détecter des relations lexicales, nous optons pour les méthodes distributionnelles, pour les raisons suivantes, entre autres :

- Elles s'adaptent relativement facilement à de nouvelles langues et à de nouveaux domaines (comparativement aux méthodes qui exigent de faire appel à des ressources lexicales existantes, par exemple).
- Elles permettent de détecter des relations entre termes simples (contrairement aux méthodes basées sur l'inclusion lexicale et à l'analyse compositionnelle), y compris entre les termes simples qui ne sont pas reliés sur le plan morphologique.
- Elles permettent de détecter une variété de relations lexicales, contrairement aux patrons lexico-syntaxiques, qui sont plutôt utilisés pour cibler des relations particulières.

Nous expliquerons l'approche distributionnelle à la section suivante. Avant de nous concentrer sur cette approche, soulignons qu'il est possible de combiner différentes méthodes pour l'identification de relations lexicales. Par exemple, Hazem et Daille [92]

[5] Nous montrerons à la section 3.4 que les graphes de voisinage permettent d'identifier différents usages d'un mot dans une certaine mesure. Il existe également d'autres façons d'utiliser les modèles distributionnels à cette fin. Par exemple, Heylen et Bertels [94] appliquent une technique de visualisation à un modèle distributionnel afin de visualiser en deux dimensions les similarités entre les cooccurrents d'un mot et ainsi identifier ses différents usages.

proposent une variante de l'analyse compositionnelle [89] qui exploite un modèle distributionnel plutôt qu'un dictionnaire existant pour déterminer si les composantes des termes complexes sont reliées sémantiquement. Dupuch et al. [53–57] exploitent, entre autres, l'inclusion lexicale, la variation morphosyntaxique et l'analyse compositionnelle afin de détecter automatiquement des ensembles de termes reliés dans le domaine de la pharmacovigilance. Si les méthodes distributionnelles détectent une variété de relations sans faire de distinctions entre celles-ci, il est possible de faire appel à d'autres méthodes afin de filtrer ou de classifier les relations détectées [42, 194]. En outre, on peut combiner l'analyse distributionnelle et les patrons lexico-syntaxiques de différentes façons [9, 35, 139, 174].

3.2 La sémantique distributionnelle

Les méthodes distributionnelles sont basées entièrement sur corpus. Elles consistent à observer, pour chaque unité linguistique (généralement des mots), les contextes dans lesquels elle apparaît dans un corpus et à construire une représentation de cette unité basée sur la distribution des contextes dans lesquels on l'a observée ; ce processus sera illustré à la section 3.2.3. Une fois ces représentations produites, elles peuvent être comparées entre elles afin d'estimer la similarité sémantique des unités linguistiques.

3.2.1 Bref historique de la sémantique distributionnelle

La sémantique distributionnelle est basée sur des travaux remontant aux années 1950 [95, p. 103], en particulier ceux de Harris [91], de Firth [75] et de Weaver [188].

Les méthodes distributionnelles reposent sur *l'hypothèse distributionnelle*, d'abord formulée par Harris [91]. Cette hypothèse, telle qu'elle est mise en œuvre dans les modèles sémantiques distributionnels, stipule essentiellement que les unités linguistiques apparaissant dans des contextes similaires ont tendance à avoir des sens similaires. Cette

hypothèse, qui a été formulée de nombreuses façons[6], est basée sur le principe selon lequel il existe des régularités dans la façon dont les éléments de la langue sont combinés et ordonnés. Elle sous-tend les modèles sémantiques distributionnels, mais aussi d'autres techniques du TAL, telles que les modèles de langue [157, p. 27]. L'hypothèse distributionnelle est illustrée dans l'exemple suivant :

> If we consider *oculist* and *eye-doctor* we find that, as our corpus of actually-occurring utterances grows, these two occur in almost the same environments [. . .] If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms : *oculist* and *eye-doctor*. If A and B have some environments in common and some not (e.g. *oculist* and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. [91, p.156–157].

Ainsi, deux mots apparaissant dans les mêmes contextes (ou *environnements*) ont tendance à être sémantiquement similaires.

Harris définit le terme *distribution* de la façon suivante : « The distribution of an element will be understood as the sum of all its environments. » [91, p. 146] Ainsi, la distribution d'un élément (une unité linguistique) est l'ensemble des contextes dans lesquels il apparaît. Les notions « d'élément » et de « contexte » peuvent être définies de différentes façons ; il peut s'agir de phonèmes, de morphèmes, de mots, et ainsi de suite. Si on s'intéresse au sens des mots [7], on peut formuler l'hypothèse distributionnelle de la façon suivante : les mots ayant des distributions similaires ont tendance à avoir des sens similaires ; plus leur distribution est similaire, plus ils sont sémantiquement similaires.

[6] QasemiZadeh [157, p. 24] énumère 17 façons différentes dont elle a été formulée, par des auteurs tels que Harris [91] et Firth [75], mais aussi Cruse [44].

[7] L'approche distributionnelle de Harris a une portée très large et n'est pas limitée à la similarité sémantique des mots. Par exemple, on pourrait identifier la tête d'un syntagme en se basant sur sa distribution et celle des mots qui le constituent : « if A occurs in environment X, and AB does too, but B does not, then A is the head of AB » [91, p. 162].

Selon Clark [40], parmi les premiers travaux reliés à la sémantique distributionnelle, ce sont peut-être ceux de Firth [75] qui reflètent le mieux l'état actuel de la sémantique distributionnelle, puisque Firth modélisait le comportement des mots en fonction des contextes dans lesquels ils apparaissent et s'intéressait particulièrement aux collocations. En effet, selon Firth, le sens des mots peut être caractérisé par leurs collocations, comme l'indique sa célèbre citation[8] : « You shall know a word by the company it keeps ! » [75, p. 11]. Soulignons que la notion de « collocation » chez Firth correspond à ce que l'on pourrait appeler des *collocations empiriques* [59], c'est-à-dire des combinaisons fréquentes de mots, plutôt que des *collocations lexicales*, qui sont des expressions lexicalisées dont le sens n'est pas compositionnel (p. ex. *heavy smoker*[9]).

Ainsi, un des concepts fondamentaux de la sémantique distributionnelle, que l'on retrouve à la fois chez Harris et chez Firth, est que le sens des mots peut être modélisé en fonction des mots avec lesquels ils ont tendance à se combiner[10].

On dit souvent que les méthodes distributionnelles, bien qu'elles remontent aux années 1950, ont seulement été mises en œuvre à partir des années 1990, parce qu'il a fallu attendre la disponibilité de gros corpus électroniques et d'ordinateurs suffisamment puissants [95]. Dans les faits, des méthodes automatisées basées sur l'hypothèse distributionnelle ont été conçues et mises à l'épreuve dès les années 1960. À partir de 1958, Spärck Jones et ses collègues exploraient des façons de construire des thésaurus, c'est-à-dire des classifications constituées d'ensembles de mots reliés aux mêmes concepts, ces

[8] Ceux qui ont souvent vu cette citation, mais n'ont jamais lu le travail de Firth, trouveront cela enrichissant (et peut-être amusant) de l'observer dans son contexte original : « a text [...] may contain sentences such as 'Don't be such an ass !', 'You silly ass !', 'What an ass he is !' In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly –*, *he is a silly –*, *don't be such an –*. You shall know a word by the company it keeps ! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted. » [75, p. 11]

[9] Le sens de *heavy* est modifié lorsqu'il se combine avec certaines unités lexicales liées à la notion de consommation, tels que *smoker* et *drinker* [44, p. 40–41].

[10] On retrouve également cette notion chez Cruse [44, p. 16] : « A syntagmatic affinity is established by a capacity for normal association in an utterance: there is a syntagmatic affinity, for instance, between *dog* and *barked*, since *The dog barked* is normal [...] Paradigmatically, a semantic affinity between two grammatically identical words is the greater the more congruent their patterns of syntagmatic normality. So, for instance, *dog* and *cat* share far more normal and abnormal contexts than, say, *dog* and *lamp-post*. »

ressources étant perçues comme importantes à la fois pour la recherche d'information et pour la traduction automatique [176]. Les premières expériences visant à exploiter à cette fin de l'information distributionnelle extraite d'un corpus ont été réalisées dans les années 1960. Harper [90] a montré que l'on pouvait estimer la similarité des noms (en russe) en fonction des relations syntaxiques auxquelles ils participent. Ainsi, la distribution d'un mot était définie dans ce travail comme l'ensemble des mots auxquels il est relié syntaxiquement ; ce type de contexte demeure l'un des plus courants dans le domaine de la sémantique distributionnelle[11].

Il est intéressant de noter que Harper utilisait un corpus spécialisé, et affirmait que cela atténue les difficultés posées par la polysémie : « Our audacity in attempting the experiment at all is based on three factors : the possession of a text in a limited field (physics), the foreknowledge that the multiple-meaning problem is minimal, and the capability for the automatic processing of text. » [90, p. 2–3]

Ainsi, il existe déjà en 1965 un champ de recherche appelé la *sémantique distributionnelle* : « One of the goals of Distributional Semantics is the establishment of word classes on the basis of the observed behaviour of words in written texts. » [90, p. 1]

De la recherche empirique sur la sémantique distributionnelle a également été réalisée à l'est du rideau de fer à cette époque. En fait, selon Moskowich et Caplan [140], ce seraient les chercheurs soviétiques qui ont exploité en premier la sémantique distributionnelle du point de vue de la linguistique : « Though the development of DSA [distributive-statistical analysis] is an international achievement, it is in the USSR that the experiments and supporting theoretical work were carried out as a purely linguistic enterprise. » [140, p. 113]

Si la recherche empirique en sémantique distributionnelle remonte aux années 1960,

[11] On retrouve dans le travail de Harper plusieurs outils méthodologiques utilisés aujourd'hui. Par exemple, il exploite une mesure de similarité qui est pondérée afin de rendre compte du fait que certains mots sont beaucoup plus fréquents que d'autres, et qui ressemble beaucoup à celles utilisées aujourd'hui. De plus, Harper énumère les relations syntaxiques dont on pourrait tenir compte dans le calcul distributionnel et discute le problème de la normalisation des relations syntaxiques. Il propose d'ailleurs que l'utilisation de corpus de taille plus importante permettrait de réduire la quantité de bruit qu'il observe dans les résultats.

peu de travaux ont été réalisés à ce sujet dans les années 1970 et 1980, bien que l'on retrouve tout de même quelques travaux exploitant des méthodes distributionnelles, par exemple pour caractériser la macrostructure des textes [153] ou pour l'analyse de concepts [127].

Selon Spärck Jones [176], la recherche sur les approches statistiques au TAL a été interrompue à la fin des années 1960, mais elle a connu un nouvel essor au début des années 1990, grâce à la disponibilité de gros corpus électroniques et d'ordinateurs beaucoup plus puissants. Ainsi, c'est surtout à partir du début des années 1990 que les méthodes distributionnelles ont été développées. Ces méthodes ont suscité de plus en plus d'intérêt dans le domaine du TAL et ont été appliquées à de nouvelles tâches, telles que la désambiguisation sémantique [171] et l'induction des usages de mots [173], alors qu'elles avaient surtout servi à détecter des classes sémantiques et à construire des thésaurus auparavant.

L'approche distributionnelle intéressait non seulement la communauté du TAL, mais aussi des chercheurs dans le domaine des sciences cognitives, qui ont commencé à utiliser des méthodes distributionnelles pour modéliser l'acquisition de la langue. Par exemple, Landauer et Dumais [104] ont montré qu'une méthode appelée *analyse sémantique latente* [48] modélise bien l'acquisition du vocabulaire chez les êtres humains. De même, Li et al. [118] ont montré qu'une méthode basée sur la cooccurrence graphique, comme celles que nous utilisons dans ce travail, modélise bien l'apprentissage du sens des mots par les enfants. Ces résultats sont d'autant plus impressionnants que ces méthodes n'exploitent pas la similarité aux niveaux morphologique, phonologique ou orthographique, ni d'autres sources d'information disponibles aux êtres humains, telles que la vision[12].

Ainsi, la recherche réalisée dans les sciences cognitives a montré que l'information captée par les modèles distributionnels est non seulement utile, mais aussi valide d'un

[12]Plus récemment, des modèles dits *multimodaux*, qui exploitent à la fois la distribution des mots dans un corpus et de l'information visuelle, ont été proposés [31].

point de vue cognitif :

> These methods have been successful in a number of applications, including the emulation of human performance on cognitive tasks [...] the associations derived by distributional semantics have been shown to correspond to psychological studies of human estimates of similarity between terms, and as such capture additional information that is both cognitively valid and practically useful. [42, p. 403]

Des travaux de synthèse récents [165, 182] ont permis de consolider et de clarifier les concepts et méthodes propres à la sémantique distributionnelle. Ils ont également permis de caractériser l'apport de la sémantique distributionnelle à l'étude du sens des mots. À ce sujet, Sahlgren [165] affirme que l'utilité de ces méthodes dépend de la théorie sémantique à laquelle on adhère. Selon la théorie structuraliste de Saussure [168], le sens d'un mot est défini en fonction de sa position relative par rapport aux autres mots sur les plans syntagmatique et paradigmatique. De même, dans *l'espace sémantique*[13] que représente un modèle distributionnel, les coordonnées d'un mot ne sont pas forcément significatives en soi ; c'est plutôt sa position relative par rapport aux autres mots qui importe, car elle indique quels autres mots ont un sens similaire. En outre, la distribution des contextes des mots peut servir à détecter non seulement des relations paradigmatiques, mais aussi des affinités syntagmatiques. Donc, les méthodes distributionnelles sont tout à fait compatibles avec la théorie structuraliste de Saussure. Par contre, si l'on adhère à une théorie qui accorde beaucoup d'importance à la fonction référentielle de la langue, l'approche distributionnelle ne modéliserait pas aussi bien le sens des mots :

[13] Les modèles distributionnels (ou plutôt certains de ceux-ci) reposent sur une métaphore géométrique du sens des mots : « Meanings are locations in a semantic space, and semantic similarity is proximity between the locations. » [165, p. 19] Soulignons que d'autres modèles géométriques du sens des mots ont été proposés, notamment les *espaces conceptuels* de Gärdenfors [77], un modèle sociocognitif de la sémantique qui repose sur le même principe : « the meanings that we use in communication can be described as organized in abstract spatial structures that are expressed in terms of *dimensions, distances, regions,* and other geometric notions » [77, p. 21].

> Does the word-space model constitute a complete model of the full spectrum of meaning, or does it only convey specific aspects of meaning ? The word-space model constitutes a complete model of meaning, if what we mean by “meaning” is a structuralist dichotomy of syntagma and paradigm. The answer to this question thoroughly depends on our meaning theory ; if we believe that meaning is essentially referential, then the answer will be very different. [165, p. 133]

À l'heure actuelle, la sémantique distributionnelle est un champ de recherche très dynamique, et les modèles distributionnels intéressent des chercheurs dans plusieurs domaines, notamment celui de l'intelligence artificielle. Si de nombreux travaux tendent à confirmer l'hypothèse distributionnelle, celle-ci continue à être mise à l'épreuve. De nouveaux axes de recherche ont émergé, notamment la modélisation de la composition sémantique [13, 30, 135, 192], et des avancées importantes continuent à être réalisées, bien que les méthodes distributionnelles existent depuis longtemps : « the continuing progress with VSMs [vector space models of semantics] suggests we are far from reaching their limits » [182, p. 175].

3.2.2 Types de contextes

Comme nous l'avons expliqué à la section précédente, l'hypothèse distributionnelle stipule que les mots qui apparaissent dans des contextes similaires tendent à avoir des sens similaires, la notion de « contexte » pouvant être définie de différentes façons. Les contextes utilisés pour construire des modèles sémantiques distributionnels appartiennent généralement à l'une des deux catégories suivantes :

1. les segments textuels ;
2. les cooccurrents.

Si on utilise les segments textuels comme contextes, les mots sont modélisés en fonction des segments dans lesquels ils apparaissent ; les segments peuvent être les documents, les paragraphes, les phrases ou tout autre découpage d'un corpus. Ainsi, les mots apparaissant dans les mêmes segments (ou des segments similaires) auront des représentations similaires. C'est ce type de contexte qu'exploite une méthode appelée *analyse sémantique latente* (LSA) [48] ainsi que des modèles probabilistes appelés *topic models*, tels que l'allocation de Dirichlet latente (LDA) [24].

L'autre type de contexte couramment utilisé en sémantique distributionnelle est la cooccurrence, qui peut à son tour être sous-divisée en plusieurs catégories. Evert [59] distingue trois types de cooccurrence, qui ont tous été utilisés en sémantique distributionnelle[14] :

1. La cooccurrence textuelle : deux mots sont cooccurrents s'ils apparaissent dans un même segment textuel, qu'il s'agisse de la phrase, du paragraphe ou du document.

2. La cooccurrence syntaxique : deux mots sont cooccurrents s'ils apparaissent dans la même phrase et sont liés par une relation de dépendance syntaxique. Par exemple, dans la phrase « Jean aime Marie », le mot *Jean* est un cooccurrent syntaxique du mot *aime* parce qu'il est le sujet de ce verbe.

3. La cooccurrence graphique (ou cooccurrence de surface) : deux mots sont cooccurrents s'ils apparaissent près l'un de l'autre. Les cooccurrents graphiques d'un mot sont déterminés au moyen d'une *fenêtre de contexte* qui couvre un certain nombre de mots à gauche ou à droite du mot, ou dans les deux directions.

Ce dernier type de cooccurrence mérite que l'on s'y attarde un peu, d'autant plus que c'est ce type de contexte que nous utilisons dans ce travail. Prenons, par exemple, la phrase suivante, un contexte extrait de la fiche du terme *forêt* dans le DiCoEnviro[15] :

[14] On pourrait ajouter d'autres types de cooccurrence à cette liste, tels que la cooccurrence basée sur les patrons lexico-syntaxiques [174].

[15] `http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi` (site consulté le 4 septembre 2015).

> *Ce qui reste de la **forêt** amazonienne est menacé par la combinaison de perturbations d'origine humaine, l'augmentation de la fréquence et de l'ampleur des incendies, et les baisses de précipitations dues à des pertes d'évapotranspiration, au réchauffement du globe et à El Niño.*

Si on utilise une fenêtre de contexte couvrant trois mots à gauche et trois mots à droite, les cooccurrents graphiques de cette occurrence de *forêt* seraient : {*reste*, *de*, *la*, *amazonienne*, *est*, *menacé*}.

Parmi les différents types de contextes utilisés en sémantique distributionnelle, les cooccurrents graphiques ou syntaxiques sont ceux que l'on utilise le plus souvent pour mesurer la similarité sémantique et réaliser des tâches connexes [182]. Dans ce travail, nous utilisons les cooccurrents graphiques comme contextes.

3.2.3 Construction et interrogation d'un modèle distributionnel

Dans cette section, nous illustrons la construction d'un modèle distributionnel[16] ainsi que la méthode que l'on utilise habituellement pour interroger ce genre de modèle. Ces explications sont nécessaires notamment pour comprendre la question du paramétrage et de l'évaluation des modèles distributionnels, qui occupe une place primordiale dans ce travail.

Un modèle sémantique distributionnel modélise des mots au moyen de représentations numériques construites automatiquement à partir d'un corpus. Nous appellerons les mots représentés par le modèle les *mots-cibles* du modèle. Ceux-ci peuvent comprendre tous les mots dans le corpus, ou un sous-ensemble de ceux-ci ; ils peuvent aussi comprendre d'autres unités linguistiques, telles que des termes complexes.

Les représentations que le modèle distributionnel associe aux mots-cibles, que nous appellerons parfois *représentations lexicales*[17], sont généralement des vecteurs, que l'on

[16] Une explication semblable à celle que nous présentons ici est fournie par Heylen et Bertels [94].

[17] Nous les appelons ainsi parce qu'elles sont des représentations d'éléments du lexique. On pourrait également les appeler des *représentations de mots* ou des *représentations sémantiques*, puisqu'elles modélisent le sens des mots et permettent de capter des affinités sémantiques entre les mots.

peut se représenter simplement comme des listes de nombres. Ainsi, un mot *w* pourrait être représenté par le vecteur suivant :

$$\vec{w} = \begin{pmatrix} 0.1 & 0.5 & 1.0 \end{pmatrix}$$

Les représentations lexicales sont généralement beaucoup plus longues que ce vecteur, qui ne comporte que trois *composantes* (ou *coordonnées*) ; elles comportent souvent des centaines de composantes, voire des centaines de milliers de composantes, selon les méthodes utilisées et le paramétrage de ces méthodes.

Ces représentations lexicales peuvent être construites de différentes façons. En effet, il existe de nombreuses méthodes pour construire des modèles distributionnels, que l'on peut classifier en fonction de différents critères [94, 182]. La méthode particulière que nous illustrons ci-dessous est une méthode distributionnelle classique généralement appelée *analyse distributionnelle* (AD) ; cette méthode est l'une des deux méthodes que nous utilisons dans ce travail pour construire des modèles distributionnels.

L'AD consiste essentiellement à produire les représentations des mots-cibles en comptant leurs cooccurrents. Il peut s'agir de cooccurrents textuels, syntaxiques ou graphiques, comme nous l'avons expliqué à la section 3.2.2 ; dans ce travail, nous utilisons les cooccurrents graphiques.

La première étape de l'AD consiste à calculer une *matrice de cooccurrence* indiquant la fréquence de cooccurrence des mots-cibles et de leurs cooccurrents ; les mots qui sont pris en compte comme cooccurrents sont parfois appelés les *mots-contextes*. Si on utilise la cooccurrence graphique, cette matrice est calculée en observant toutes les occurrences de chaque mot-cible, en plaçant la fenêtre de contexte autour de chacune de ces occurrences, et en incrémentant, pour chaque mot-contexte dans la fenêtre, la fréquence de cooccurrence du mot-cible et du mot-contexte. Les cooccurrents graphiques ne faisant pas partie des mots-contextes ne sont pas comptabilisés[18].

[18]Une autre option consiste à supprimer les mots qui ne font pas partie des mots-contextes, ce qui a pour effet d'élargir la fenêtre de contexte dans certains cas.

Par exemple, supposons que l'on ait un corpus constitué seulement de la phrase « Jean aime Marie » et que l'on construise une matrice de cooccurrence en utilisant une fenêtre de contexte couvrant seulement le mot immédiatement à gauche de chaque mot ; supposons également que les mots-cibles et les mots-contextes contiennent tous les mots dans le corpus. On obtiendrait alors la matrice de cooccurrence présentée dans la Figure 3.1. Dans cette matrice, les lignes (ou rangées) correspondent aux mots-cibles et les colonnes, aux mots-contextes. Dans la ligne correspondant à un mot-cible donné, les nombres dans chaque colonne indiquent le nombre de fois que le mot-contexte correspondant à cette colonne a été observé comme cooccurrent du mot-cible (c'est-à-dire le nombre de fois qu'il apparaît à l'intérieur d'une fenêtre de contexte centrée sur une occurrence du mot-cible). Autrement dit, la matrice de cooccurrence encode, pour chaque combinaison d'un mot-cible et d'un mot-contexte, la fréquence à laquelle le mot-contexte apparaît comme cooccurrent du mot-cible.

$$\begin{array}{c} \\ \textit{Jean} \\ \textit{aime} \\ \textit{Marie} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{Jean} & \textit{aime} & \textit{Marie} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

Figure 3.1 – Matrice de cooccurrence.

La matrice de cooccurrence présentée dans la Figure 3.1 peut être considérée comme un modèle distributionnel contenant les représentations lexicales suivantes :

- le mot *Jean* est représenté par le vecteur ligne $\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$, qui indique qu'aucun mot-contexte n'apparait comme cooccurrent de ce mot dans le corpus (du moins lorsqu'on observe seulement le mot immédiatement à gauche) ;
- le mot *aime* est représenté par le vecteur ligne $\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, qui indique que le mot-contexte *Jean* apparait une fois comme cooccurrent de ce mot dans le corpus ;

- le mot *Marie* est représenté par le vecteur ligne $\begin{pmatrix}0 & 1 & 0\end{pmatrix}$, qui indique que le mot-contexte *aime* apparait une fois comme cooccurrent de ce mot dans le corpus.

Si on ajoutait au corpus la phrase « Jean adore Marie » (et que l'on ajoutait le mot *adore* aux mots-cibles et aux mot-contextes), on obtiendrait alors la matrice de cooccurrence présentée dans la Figure 3.2. Dans cette matrice, le vecteur du mot-cible *adore* est identique à celui du mot-cible *aime*.

$$\begin{array}{c} \\ Jean \\ aime \\ adore \\ Marie \end{array} \begin{array}{c} \begin{array}{cccc} Jean & aime & adore & Marie \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Figure 3.2 – Matrice de cooccurrence après l'ajout d'une 2e phrase.

Ainsi, des mots ayant des cooccurrents en commun, comme *aime* et *adore* dans cet exemple, auront des représentations similaires. C'est de cette façon que ce modèle met en œuvre l'hypothèse distributionnelle, à savoir que les mots qui apparaissent dans des contextes similaires ont tendance à avoir des sens similaires. Cette *similarité distributionnelle* (similarité des contextes dans lesquels apparaissent les mots) est calculée en comparant les représentations lexicales entre elles ; nous illustrons ce processus ci-dessous.

Il existe de nombreuses mesures permettant de calculer la similarité distributionnelle. Le cosinus de l'angle des vecteurs [167] est la mesure la plus courante [182, p. 160] ; c'est donc cette mesure que nous utilisons dans ce travail.

Pour illustrer comment le cosinus de l'angle des vecteurs permet de mesurer la similarité distributionnelle, supposons que l'on ait le modèle présenté dans la Figure 3.3, qui contient les quatre représentations lexicales (artificielles) suivantes : $\vec{w_1} = \begin{pmatrix}1 & 10\end{pmatrix}$, $\vec{w_2} = \begin{pmatrix}2 & 10\end{pmatrix}$, $\vec{w_3} = \begin{pmatrix}9 & 2\end{pmatrix}$ et $\vec{w_4} = \begin{pmatrix}10 & 1\end{pmatrix}$. Il est facile de voir que les représentations de w_1 et de w_2 sont similaires, puisque la différence entre les valeurs dans la première

$$\begin{array}{c} w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{pmatrix} 1 & 10 \\ 2 & 10 \\ 9 & 2 \\ 10 & 1 \end{pmatrix}$$

Figure 3.3 – Matrice contenant 4 représentations lexicales de dimension 2.

colonne est petite et que les valeurs dans la deuxième colonne sont identiques. En revanche, les représentations de w_1 et de w_4 sont très différentes, puisque les différences entre les valeurs dans chaque colonne sont élevées.

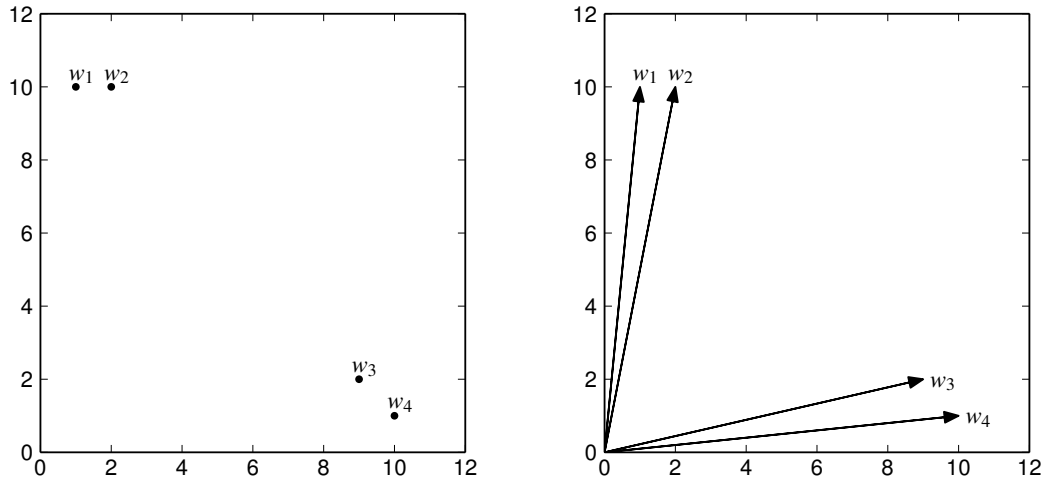


Figure 3.4 – Quatre représentations lexicales de dimension 2 illustrées sous la forme **(a)** de points et **(b)** de flèches.

Puisque ces représentations artificielles ont seulement 2 composantes[19] (ou coordonnées), on peut les illustrer au moyen de points ou de flèches dans un plan cartésien, comme le montre la Figure 3.4, ce plan représentant un espace sémantique à deux dimensions.

[19]Comme nous l'avons signalé ci-dessus, les représentations lexicales distributionnelles ont généralement une dimension beaucoup plus élevée. Nous utilisons ici des vecteurs de dimension 2 seulement pour illustrer le calcul de la similarité distributionnelle.

Dans la Figure 3.4b, on peut voir que les vecteurs $\vec{w}_1$ et $\vec{w}_2$ forment un angle beaucoup plus petit que celui formé par $\vec{w}_1$ et $\vec{w}_4$ (angles de 5.6° et de 78.6° respectivement). On peut donc se servir de l'angle formé par deux vecteurs pour mesurer leur proximité[20]. Dans la pratique, on se sert plutôt du cosinus de l'angle des vecteurs. Formellement, on peut définir le cosinus de la façon suivante :

$$\text{cosinus}(x,y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

où x et y sont deux vecteurs[21] de dimension n. Le cosinus donne 0 pour deux vecteurs formant un angle de 90° et 1 pour des vecteurs ayant exactement la même direction (c'est-à-dire qui forment un angle de 0°). Ainsi, un cosinus de 0 indique une similarité minimale[22] et un cosinus de 1 indique une similarité maximale. Par exemple, le cosinus de l'angle entre $\vec{w}_1$ et $\vec{w}_2$ est élevé (0.9952), ce qui indique une similarité élevée, tandis que celui de l'angle entre $\vec{w}_1$ et $\vec{w}_4$ est faible (0.1980).

Ce calcul de la similarité distributionnelle peut être utilisé pour obtenir les *voisins distributionnels* d'un mot. Comme le soulignent Gyllensten et Sahlgren [85], la façon la plus courante d'interroger un modèle sémantique distributionnel consiste à calculer, pour une requête donnée, ses *plus proches voisins* (PPV), c'est-à-dire les mots-cibles dont la représentation est la plus similaire à celle de la requête.

Par exemple, pour obtenir les PPV de w_1, on calcule la similarité distributionnelle entre w_1 et tous les autres mots-cibles (w_2, w_3 et w_4), ce qui indiquerait que le mot le plus similaire à w_1 est w_2, le deuxième est w_3, et ainsi de suite.

Si on calcule la similarité entre les quatre représentations lexicales prises deux à deux, on obtient la *matrice de similarité* présentée dans la Figure 3.5. Si on observe la ligne (ou la colonne, ce qui revient au même parce que la matrice de similarité est

[20] Soulignons que cette mesure ne tient pas compte de la norme des vecteurs (autrement dit, la longueur des flèches), seulement de leur direction.

[21] Nous omettons ici les flèches pour alléger la notation.

[22] Le cosinus minimal est 0 si les coordonnées des vecteurs sont toutes non négatives, car l'angle maximal formé par ces vecteurs est 90°.

symétrique) correspondant au mot w_2, on voit que le mot le plus similaire à w_2 est w_2 lui-même (similarité de 1) ; le deuxième mot le plus similaire est w_1 (similarité de 0.9952), suivi de w_3 (0.4042) et de w_4 (0.2927). On peut ainsi obtenir les PPV de tous les mots-cibles en triant les valeurs dans la matrice de similarité.

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 \end{array} \\ \begin{pmatrix} 1.0000 & 0.9952 & 0.3130 & 0.1980 \\ 0.9952 & 1.0000 & 0.4042 & 0.2927 \\ 0.3130 & 0.4042 & 1.0000 & 0.9929 \\ 0.1980 & 0.2927 & 0.9929 & 1.0000 \end{pmatrix} \end{array}$$

Figure 3.5 – Matrice de similarité des 4 représentations lexicales.

Dans cette section, nous avons présenté une version simplifiée de l'AD. Dans les faits, d'autres étapes s'ajoutent au calcul de la matrice de cooccurrence (notamment la pondération des fréquences de cooccurrence, comme nous l'expliquerons à la section 3.2.5), mais en somme, l'AD produit des représentations lexicales basées essentiellement sur la fréquence de cooccurrence des mots-cibles et de leurs cooccurrents. L'ensemble des représentations lexicales construites au moyen de cette technique constitue un modèle sémantique distributionnel. Une fois le modèle construit, on peut calculer la similarité distributionnelle des mots au moyen d'une mesure de similarité telle que le cosinus de l'angle des vecteurs. Cette similarité est souvent utilisée pour interroger le modèle en calculant, pour une requête donnée, les mots-cibles les plus similaires à celle-ci. Les représentations lexicales peuvent également être utilisées d'autres façons, selon l'application envisagée. Par exemple, on peut :

- classifier les mots-cibles en ensembles de mots similaires au moyen d'une technique de partitionnement de données (*clustering*) ;
- cartographier les représentations lexicales au moyen d'une technique de réduction de dimension ;
- exploiter les représentations lexicales au sein d'une autre technique de TAL.

Ces différentes façons d'exploiter les représentations lexicales ne sont pas propres aux représentations produites par l'AD ; elles peuvent être appliquées quelle que soit la méthode utilisée pour construire le modèle distributionnel, les différentes méthodes constituant simplement différentes approches à l'acquisition de représentations lexicales.

3.2.4 Les modèles de langue neuronaux

Comme nous l'avons souligné à la section 3.2.3, il existe différentes méthodes pour construire un modèle sémantique distributionnel. Nous avons expliqué une de ces méthodes, l'AD, qui consiste essentiellement à compter les cooccurrents des mots-cibles.

Il existe une autre famille de méthodes qui produisent des représentations lexicales basées sur la cooccurrence des mots, à savoir les *modèles de langue neuronaux*. Ces modèles permettent d'estimer la probabilité d'une séquence quelconque de mots [14] et peuvent ainsi servir différentes applications du TAL, telles que la reconnaissance automatique de la parole [134, 175] ; des modèles neuronaux pour la traduction automatique ont d'ailleurs été conçus à partir de ces modèles [6]. Les modèles de langue neuronaux apprennent, à partir d'un corpus, des représentations lexicales qui peuvent servir à estimer la similarité sémantique des mots, de la même façon que les représentations construites au moyen de l'AD. Bien que la méthode utilisée pour apprendre ces représentations soit différente, elle repose sur la même hypothèse, à savoir que les mots qui apparaissent dans des contextes similaires ont tendance à être sémantiquement proches.

Parmi les différents modèles de langue neuronaux qui ont été proposés dans les dernières années [43, 136, 147, 175], le mieux connu est implémenté dans un outil appelé `word2vec` [130, 132]. Pour expliquer d'une façon générale la différence entre `word2vec` et l'AD, on peut dire, comme Baroni et al. [10], que l'AD modélise les mots en comptant leurs contextes, tandis que `word2vec` apprend à prédire les mots en fonction de leur contexte (ou vice-versa, selon l'architecture utilisée).

Mikolov et al. [133] ont montré que ce modèle capte des régularités linguistiques aux niveaux syntaxique et sémantique. Par exemple, les directions relatives des vecteurs

des différentes formes fléchies d'un même mot sont similaires d'un mot à l'autre, de sorte qu'il est possible de résoudre des analogies telles que « *pris* est à *prendre* ce que *tenu* est à . . . » en identifiant le mot manquant (en l'occurrence, *tenir*), au moyen d'opérations simples d'algèbre linéaire. Il est également possible de résoudre des analogies sémantiques telles que « *roi* est à *homme* ce que *reine* est à . . . ».

Par la suite, Levy et Goldberg [109] ont montré que les régularités linguistiques (ou similarités relationnelles) captées par `word2vec` sont aussi captées par l'AD, et qu'il est possible de résoudre des analogies avec un modèle distributionnel classique et d'obtenir une précision semblable à celle que l'on obtient avec `word2vec`.

Il est important de noter que `word2vec` exploite les mêmes genres de contextes que l'AD, à savoir les cooccurrents graphiques ou syntaxiques. En effet, Levy et Goldberg [108] ont montré que l'on peut utiliser des dépendances syntaxiques plutôt que la cooccurrence graphique pour entraîner `word2vec`, et que l'influence du type de contexte est semblable à celle que l'on observe dans le cas de l'AD (nous reviendrons sur ce point à la section 3.3).

Ainsi, l'AD et les modèles de langue neuronaux exploitent les mêmes genres de contextes et semblent capter des régularités linguistiques similaires. En fait, il n'est pas tout à fait clair pour l'instant s'il existe des différences importantes entre l'information sémantique captée par les deux types de modèles. Des analyses qualitatives des voisinages sémantiques captés par ces modèles ont commencé à être réalisées [85] et nous aideront éventuellement à comprendre les différences entre ces modèles.

3.2.5 (Hyper)paramètres des modèles distributionnels

Toutes les méthodes permettant de construire un modèle distributionnel exigent que l'on prenne un certain nombre de décisions lors de la construction d'un modèle. Ces décisions concernent les *paramètres* du modèle (ou, plus exactement, de la méthode utilisée pour le construire). Par exemple, dans l'illustration du calcul de l'AD que nous avons présentée à la section 3.2.3, nous utilisions une fenêtre de contexte couvrant un

mot à gauche ; la taille et la direction de la fenêtre de contexte sont deux des paramètres de l'AD (si on exploite les cooccurrents graphiques comme contextes). Pour le corpus constitué seulement de la phrase « Jean aime Marie », on obtenait ainsi la matrice de cooccurrence présentée dans la Figure 3.1. Dans cette matrice, le vecteur ligne du mot-cible *aimer* était $\begin{pmatrix}1 & 0 & 0\end{pmatrix}$, indiquant que le mot-contexte *Jean* avait été observé une fois comme cooccurrent de *aimer*. Il serait possible d'observer plutôt le mot à droite ; dans ce cas, la représentation du mot *aimer* serait $\begin{pmatrix}0 & 0 & 1\end{pmatrix}$, indiquant que le mot-contexte *Marie* apparaît une fois comme cooccurrent. Il est donc nécessaire de préciser la *direction* de la fenêtre de contexte.

Dans la pratique, on n'utilise généralement pas une fenêtre de contexte couvrant seulement les mots à gauche ou seulement ceux à droite, mais une fenêtre couvrant un certain nombre de mots dans les deux directions. Or, il existe plus d'une façon de prendre en compte les deux directions. On peut prendre la somme des vecteurs obtenus en observant les cooccurrents à gauche d'une part et ceux à droite d'autre part ; dans ce cas, le vecteur du mot *aimer* serait :

$$\begin{pmatrix}1 & 0 & 0\end{pmatrix} + \begin{pmatrix}0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 1\end{pmatrix}$$

On parle alors d'une fenêtre *gauche+droite* (ou *symétrique*). Une autre option consiste à concaténer les deux vecteurs plutôt que prendre leur somme, ce qui donne une représentation deux fois plus longue, dans laquelle les fréquences de cooccurrence à gauche et à droite du mot-cible sont encodées séparément[23]. Pour le mot-cible *aimer*, on obtiendrait ainsi la représentation $\begin{pmatrix}1 & 0 & 0 & 0 & 0 & 1\end{pmatrix}$. On parle alors d'une fenêtre *gauche&droite* (ou *directionnelle*) plutôt que gauche+droite.

Un deuxième paramètre lié à la fenêtre de contexte est sa taille, c'est-à-dire le nombre de mots à gauche et à droite du mot-cible qui sont considérés comme cooccurrents.

Enfin, la fenêtre peut avoir différentes *formes*. La forme de la fenêtre définit la valeur

[23] Il est aussi possible d'encoder séparément les fréquences observées à chaque position dans la fenêtre de contexte [16, 172].

que l'on ajoute dans la matrice de cooccurrence (*l'incrément*) pour chaque cooccurrent dans la fenêtre. Dans le cas d'une fenêtre dite *rectangulaire*, l'incrément est 1 pour tous les cooccurrents, mais il est possible de faire en sorte qu'il varie en fonction de la distance entre le mot-cible et le cooccurrent. Par exemple, dans le cas d'une fenêtre dite *triangulaire*, l'incrément est inversement proportionnel à la distance entre les deux mots : l'incrément est donc 1 pour les cooccurrents à une distance de 1 mot du mot-cible, $\frac{1}{2}$ pour les cooccurrents à une distance de 2 mots, $\frac{1}{3}$ pour les cooccurrents à une distance de 3 mots, et ainsi de suite. Les cooccurrents les plus près du mot-cible ont ainsi un effet plus important sur sa représentation.

La contribution des différents mots-contextes aux représentations des mots-cibles peut effectivement être ajustée de différentes façons, notamment en appliquant une *pondération* aux valeurs dans la matrice de cooccurrence. La raison pour laquelle on pondère les fréquences de cooccurrence est liée au fait que les mots très fréquents apparaissent fréquemment près de beaucoup de mots, contrairement aux mots peu fréquents. Par exemple, l'article *le* apparaîtra près de beaucoup de mots dans le corpus, donc sa fréquence de cooccurrence avec les mots-cibles sera relativement élevée ; autrement dit, les valeurs dans la colonne correspondant au mot-contexte *le* seront élevées, en moyenne. Ainsi, deux mots-cibles choisis au hasard pourraient apparaître plus similaires qu'ils ne le sont vraiment, simplement parce que l'un et l'autre apparaissent souvent près du mot *le*.

L'application d'une pondération aux fréquences de cooccurrence peut diminuer l'influence des cooccurrents peu discriminants tels que *le* et augmenter l'influence de cooccurrents plus discriminants ; autrement dit, cela permet de « donner plus de poids aux cooccurrents qui apparaissent significativement plus souvent avec le mot-cible qu'attendu par le hasard. Ces cooccurrents avec un poids plus élevé fournissent plus d'informations sur le sens du mot-cible que les autres cooccurrents, quelle que soit leur fréquence absolue. » [94, p. 57]. On peut ainsi améliorer de façon significative la qualité des résultats que l'on obtient au moyen de l'AD [106].

Plusieurs pondérations utilisées pour l'AD sont des *mesures d'association*. Comme leur nom l'indique, les mesures d'association estiment le degré d'association entre deux (ou plusieurs) mots, et peuvent donc servir à détecter automatiquement des collocations. Une mesure d'association bien connue, et souvent utilisée comme pondération pour l'AD, est *l'information mutuelle* [39]. Comme l'explique Evert [59, ch. 4] dans un excellent travail de synthèse sur les collocations et la détection automatique de celles-ci, les mesures d'association simples telles que l'information mutuelle comparent la fréquence à laquelle deux mots apparaissent ensemble (la fréquence de cooccurrence *observée*) à la fréquence à laquelle on s'attendrait de les observer ensemble s'il n'y avait aucune association entre les deux mots (*l'espérance* de la fréquence de cooccurrence)[24]. Si la fréquence de cooccurrence observée de deux mots est beaucoup plus élevée que l'espérance de cette fréquence, cela indique une association forte entre les deux mots. Ces mesures peuvent donc servir à détecter des collocations ou des termes complexes, mais aussi à pondérer les valeurs dans la matrice de cooccurrence pour mieux estimer la similarité distributionnelle.

Prenons, par exemple, les mots *effet* et *serre*. Dans les textes du domaine de l'environnement, le mot *serre* apparaît très fréquemment près du mot *effet*, puisque les deux mots font partie du terme complexe *effet de serre*. Le mot *serre* sera donc un cooccurrent très fréquent du mot-cible *effet* (à condition que l'on utilise une fenêtre de contexte qui couvre au moins deux mots à droite). Or, si l'on choisissait un autre mot-cible au hasard, le mot *serre* ne serait probablement pas un cooccurrent très fréquent de ce mot, car le mot *serre* apparaît presque exclusivement dans l'expression *effet de serre* dans les textes de ce domaine. Ainsi, dans la colonne correspondant au mot-contexte *serre* dans la matrice de cooccurrence, la plupart des valeurs seront nulles ou faibles, mais la valeur dans la ligne correspondant au mot-cible *effet* sera élevée. Puisqu'il y a une association forte entre les deux mots, la valeur correspondant au mot-contexte *serre* (autrement dit,

[24]Pour saisir ce que signifie l'espérance de la fréquence de cooccurrence, on peut se la représenter comme la fréquence à laquelle on observerait les deux mots ensemble si tous les mots dans le corpus étaient réordonnés de manière aléatoire [59, p. 17].

le *poids* de ce mot-contexte) dans la représentation du mot-cible *effet* devrait être élevée (et vice-versa). En revanche, les mots-contextes qui apparaissent fréquemment près de *effet*, mais qui apparaissent fréquemment près de beaucoup d'autres mots, devraient avoir un poids moins élevé dans la représentation de ce mot. À cette fin, on peut pondérer les fréquences de cooccurrence au moyen d'une mesure d'association. Le choix de la pondération constitue donc un autre paramètre important de l'AD.

En somme, les paramètres de l'AD, qui consiste essentiellement à calculer, puis à pondérer les fréquences de cooccurrence[25], comprennent des paramètres liés à la fenêtre de contexte (si on exploite la cooccurrence graphique) ainsi que la pondération. Ces paramètres, qui doivent être fixés avant de construire le modèle[26], exercent une influence sur les relations qui sont captées par ce dernier. Ainsi, un des objectifs importants de l'évaluation des modèles distributionnels consiste à évaluer l'influence de leurs paramètres et à déterminer les paramétrages optimaux, comme nous l'expliquerons à la section 3.3.2.

Il est important de noter que les paramètres des modèles distributionnels sont parfois appelés des *hyperparamètres*, selon le type de modèle. Dans le cas de l'AD, on les appelle généralement des *paramètres*, mais dans le cas de l'autre méthode que nous utilisons dans ce travail pour produire des modèles, à savoir `word2vec`, il est plus exact d'utiliser le terme *hyperparamètre*. Nous utiliserons parfois le terme *(hyper)paramètre* pour désigner ce que l'on appelle soit *paramètre*, soit *hyperparamètre* selon le type de modèle.

En ce qui concerne les hyperparamètres de `word2vec`, nous les décrirons lorsque nous expliquerons notre méthodologie, au chapitre 5. Pour l'instant, il suffit de savoir que la sélection des (hyper)paramètres constitue une étape obligatoire de la construction d'un modèle distributionnel et que le paramétrage du modèle a une influence sur les

[25] D'autres techniques peuvent être appliquées, telles que la réduction de dimension, mais elles sont facultatives. En revanche, le choix de la fenêtre de contexte est obligatoire (si on utilise les cooccurrents graphiques comme contextes), et il est très important de pondérer les fréquences de cooccurrence, comme le montrent les résultats que nous présentons au chapitre 6.

[26] Plus précisément, la pondération peut être choisie après le calcul de la matrice de cooccurrence initiale, et on peut éventuellement changer la pondération si on conserve cette matrice ; si on désire changer les paramètres liés à la fenêtre de contexte, on doit recalculer la matrice de cooccurrence.

mesures de similarité et les voisinages que l'on obtient à partir de ce dernier.

3.2.6 La sémantique distributionnelle et la terminologie

Au début des années 1990, les techniques distributionnelles ont d'abord été appliquées à des corpus spécialisés [138]. La tendance actuelle dans le domaine du TAL, en ce qui concerne l'utilisation des modèles sémantiques distributionnels, consiste à traiter de très gros corpus couvrant de nombreux domaines, dans le but de créer des ressources sémantiques à large couverture, pouvant servir différentes applications. Par exemple, Baroni et Lenci [11] exploitent un corpus de 12.8 milliards de mots. Agirre et al. [3] utilisent un corpus énorme contenant 4 milliards de pages Web totalisant 1.6 billions (1600 milliards) de mots, mais l'utilisation de corpus de cette taille demeure rare. En général, la taille des corpus utilisés en sémantique distributionnelle est de l'ordre du milliard de mots. Par contre, certains chercheurs choisissent de limiter la taille des corpus, même s'ils traitent la langue générale, notamment parce qu'il n'est pas possible de compiler de très gros corpus dans toutes les langues ni dans tous les domaines [1, 66].

En effet, s'il est possible (et avantageux) d'utiliser de très gros corpus lorsqu'on traite la langue générale, il n'est pas toujours possible de compiler des corpus de taille importante lorsqu'il s'agit de traiter un domaine spécialisé. Par exemple, les corpus utilisés dans le domaine biomédical comprennent des collections de comptes-rendus cliniques, et il n'est pas possible d'obtenir des grandes quantités de ces textes [42].

Depuis l'émergence de la sémantique distributionnelle, les travaux exploitant des corpus spécialisés (de taille relativement petite) demeurent comparativement peu nombreux, comme le soulignent Périnet et Hamon [148]. Il existe tout de même un certain nombre de travaux qui méritent d'être soulignés ici. Ci-dessous, nous nous penchons sur les travaux exploitant des corpus spécialisés, mais nous tenons à souligner que des corpus de taille relativement petite ont également été utilisés dans les sciences cognitives, le choix d'un petit corpus étant motivé dans ce cas par un objectif différent, à savoir modéliser l'apprentissage du lexique par les êtres humains.

Nous ne tenterons pas d'énumérer de manière exhaustive les travaux dans lesquels des méthodes distributionnelles sont appliquées à des corpus spécialisés ; notre objectif est plutôt d'illustrer quelques tendances que l'on peut observer dans ces travaux, notamment quant aux domaines traités et aux méthodes utilisées.

Premièrement, on constate un intérêt particulier pour le domaine biomédical, comme l'observent Tanguy et al. [178], les applications biomédicales de la sémantique distributionnelle comprenant l'extraction d'information et l'enrichissement d'ontologies. En effet, une proportion importante des travaux dont nous sommes au courant traitent (au moins) un corpus de ce domaine [27, 28, 82, 84, 86, 120, 137, 142, 150, 151, 157, 190]. Soulignons que les travaux traitant le domaine de l'environnement, que nous traitons dans cette thèse, sont très peu nombreux : nous ne connaissons que celui de Hazem et Daille [92], qui exploitent un corpus du domaine des énergies renouvelables (compilé automatiquement, comme ceux que nous utilisons, mais de taille beaucoup plus petite) afin de découvrir des synonymes de termes complexes au moyen d'une technique qui combine l'AD et l'analyse compositionnelle.

Une deuxième tendance concerne le type de contexte exploité pour calculer la similarité distributionnelle, beaucoup des travaux dans ce courant de recherche exploitant la cooccurrence syntaxique [27, 28, 76, 82, 84, 86, 120, 137, 142, 190]. Certains ont observé que la cooccurrence syntaxique offre de meilleurs résultats que la cooccurrence graphique (nous reviendrons sur ce point à la section 3.3), particulièrement lorsque la taille du corpus est petite, ce qui expliquerait éventuellement la tendance à exploiter ce type de contexte lorsqu'il s'agit de traiter des corpus spécialisés. Tanguy et al. [178] comparent ces deux types de cooccurrence sur un corpus spécialisé de petite taille, et leurs résultats indiquent que la cooccurrence syntaxique offre de meilleurs résultats que la cooccurrence graphique si on l'exploite correctement, et suggèrent que l'amélioration est d'autant plus importante que la taille du corpus est petite. Pour sa part, Grefenstette [83] observe que le meilleur type de cooccurrence varie selon la fréquence des mots que l'on utilise pour interroger le modèle, la cooccurrence graphique offrant de meilleurs

résultats pour les termes à faible fréquence.

En effet, les corpus de petite taille posent des difficultés d'un point de vue distributionnel, parce que la plupart des termes qu'ils contiennent ont un nombre très faible d'occurrences ; par conséquent, on dispose d'un nombre trop faible de contextes pour bien modéliser ces termes. Une façon simple d'obtenir de meilleurs résultats est donc d'augmenter la taille du corpus, mais comme nous l'avons souligné, cela n'est pas toujours possible :

> Une façon évidente d'améliorer la qualité des thésaurus [distributionnels] est ainsi d'augmenter la taille de leur corpus source, tendance que l'on observe clairement dans les travaux récents où les corpus sont rarement inférieurs au milliard de mots. Cette approche est concevable en domaine général où les corpus de grandes tailles ne sont pas rares. Elle est plus difficile à mettre en œuvre dans beaucoup de domaines spécialisés, même si ce n'est pas le cas de tous. [70, p. 45]

Ainsi, on retrouve plusieurs travaux visant à mieux modéliser les termes peu fréquents ou à améliorer la qualité des résultats que l'on obtient sur des petits corpus [148, 150–152, 190].

Si la cooccurrence syntaxique semble être plus souvent utilisée dans les travaux exploitant des corpus spécialisés que la cooccurrence graphique, cette dernière a tout de même été utilisée dans plusieurs travaux, parmi lesquels on peut mentionner ceux de Périnet et Hamon [148–152], de Thanopoulos et al. [180, 181], de Bertels et Speelman [21–23] et de QasemiZadeh [157].

Soulignons enfin que l'évaluation occupe souvent une place importante dans les travaux sur la sémantique distributionnelle en corpus spécialisé, tout comme dans les travaux exploitant des corpus généraux (nous reviendrons sur ce point à la section 3.3).

Si la tendance actuelle en sémantique distributionnelle consiste à traiter de très gros corpus, Fabre et Lenci [62] soulignent, dans l'introduction d'un numéro spécial de la re-

vue *TAL* consacré à la sémantique distributionnelle, que dans 3 des 4 articles, on exploite un corpus spécialisé. Cela signalerait éventuellement un regain d'intérêt pour l'étude des méthodes distributionnelles dans le contexte du travail terminologique.

3.3 L'évaluation des modèles distributionnels

L'évaluation occupe une place très importante dans cette thèse, comme dans les travaux en sémantique distributionnelle en général. Dans cette section, nous décrivons les méthodes utilisées pour évaluer les modèles distributionnels et les différents objectifs visés par cette tâche. Toutes les problématiques abordées dans cette thèse sont reliées à la question de l'évaluation des modèles distributionnels ; nous évoquons donc dans cette section la plupart de ces problématiques, que nous énumérerons au chapitre 4.

3.3.1 Méthodes d'évaluation

Les méthodes utilisées pour évaluer les modèles distributionnels peuvent être divisées en deux catégories : les méthodes intrinsèques et les méthodes extrinsèques. Les méthodes extrinsèques consistent à évaluer le modèle d'une manière indirecte, en l'exploitant pour réaliser une tâche particulière et en évaluant la qualité des résultats obtenus sur cette tâche. Par exemple, Schütze [171] réalise une évaluation extrinsèque en utilisant un modèle distributionnel afin de désambiguiser le sens des mots.

Les méthodes intrinsèques consistent à évaluer directement les mesures de similarité que l'on obtient à partir du modèle, au moyen d'une comparaison exploitant une ressource qui sert de référence.

Une façon courante de réaliser une évaluation intrinsèque consiste à comparer ces mesures de similarité à des jugements humains sur la similarité des mots. Des jeux de données ont été créés spécifiquement à cette fin, tels que SimLex-999 [96].

Une évaluation intrinsèque peut également être réalisée en comparant les paires de mots similaires selon le modèle distributionnel aux relations recensées dans une res-

source lexicale. Les travaux exploitant des corpus spécialisés utilisent souvent cette méthode d'évaluation, qui remonte à Grefenstette [83], sinon à des travaux antérieurs. C'est d'ailleurs cette méthode d'évaluation que nous utilisons dans ce travail.

Une des limites des évaluations intrinsèques est liée à la couverture des données de référence :

> Intrinsic evaluation involves comparing the resource directly against a manually created gold standard, such as an existing thesaurus. The problem with this approach is that automatic methods are designed to overcome some of the limitations of manually created resources, such as lack of coverage. So it may be that the automatic system correctly posits a synonym for some target word which was not considered by the lexicographer creating the gold standard. [40, p. 19]

En effet, si on prend en compte toutes les paires de mots ayant une similarité distributionnelle relativement élevée[27], la proportion de ces couples qui seront recensés dans une ressource lexicale quelconque sera généralement faible [2].

Le problème de la couverture des données de référence est attribuable à plusieurs facteurs. Fabre et Lenci [62] mettent en valeur le fait que les méthodes distributionnelles captent une variété de relations, alors que les ressources lexicales sont souvent axées sur des relations particulières, ce qui engendre des problèmes de couverture. De plus, la similarité distributionnelle reflète les spécificités du corpus utilisé pour la calculer, donc elle peut indiquer des relations qui sont effectivement pertinentes, mais qui n'ont pas été observées ou encodées lors de la construction de la ressource lexicale. Ainsi, les évaluations intrinsèques réalisées de cette façon seraient surtout utiles à des fins de comparaison [152].

Il existe de nombreux jeux de données qui sont couramment utilisés pour l'évaluation

[27] On peut utiliser à cette fin un seuil de similarité minimale ou prendre en compte un nombre fixe de PPV pour chaque mot-cible.

des modèles distributionnels[28], qu'il s'agisse de jugements de similarité, de tests de synonymie à choix multiple, de problèmes de résolution d'analogies ou de classifications sémantiques. Un répertoire de ces jeux de données a notamment été créé par Faruqui et Dyer [63]. Soulignons que ces jeux sont presque tous en anglais, bien qu'il existe quelques jeux de données pour d'autres langues (plusieurs d'entre eux ayant été adaptés de données anglaises). En ce qui concerne le français, il existe une version française du jeu de données de Rubenstein et Goodenough [162], créée par Joubarne et Inkpen [101], mais les jeux de données utilisés pour le français sont souvent construits sur mesure pour une évaluation particulière ; ces jeux *ad hoc* prennent généralement la forme d'un thésaurus, c'est-à-dire une liste de mots accompagnés de mots sémantiquement reliés à ceux-ci.

Par ailleurs, les jeux de données couramment utilisés en sémantique distributionnelle ne ciblent généralement pas des relations lexicales particulières (à l'exception de la synonymie) et ne font généralement pas de distinctions entre différentes relations lexicales. En effet, beaucoup de ces jeux de données représentent des liens de similarité ou de proximité sémantique plus ou moins clairement définis. Une exception notable est le jeu BLESS [12], qui permet d'évaluer la détection de relations lexicales (ou conceptuelles) particulières (cohyponymie, hyperonymie, méronymie, concept-attribut et concept-événement).

Parmi les relations lexicales prises en compte dans cette thèse, certaines sont représentées dans d'autres jeux de données utilisés en sémantique distributionnelle ; c'est notamment le cas de la (quasi-)synonymie. En revanche, nous ne connaissons aucun jeu de données qui représente la dérivation syntaxique, qui fait partie des relations que nous avons prises en compte. Le jeu d'analogies sémantiques et syntaxiques présenté par Mikolov et al. [130] comprend des relations de dérivation morphologique adjectif-adverbe (p. ex. *apparently* est à *apparent* ce que *rapidly* est à *rapid*), mais nous ne connaissons

[28] Voir, par exemple, les données utilisées par Bullinaria et Levy [34], Baroni et al. [10] ou Kiela et Clark [102].

aucun jeu de données utilisé en sémantique distributionnelle qui couvre de manière plus exhaustive la dérivation morphologique[29] ou la dérivation syntaxique[30]. De plus, nous ne connaissons aucun travail visant à déterminer si les modèles distributionnels détectent l'appartenance au même cadre sémantique aussi bien que d'autres relations lexicales.

En somme, il existe de nombreuses méthodes pour évaluer les modèles distributionnels, et la méthode d'évaluation devrait être choisie en fonction de l'application envisagée [164]. D'ailleurs, la diversité des méthodologies fait en sorte qu'il est difficile d'établir des comparaisons entre les résultats présentés dans différents travaux [70].

3.3.2 Objectifs

L'évaluation des modèles distributionnels peut avoir différents objectifs. Certaines évaluations visent à comparer différentes méthodes pour construire des modèles distributionnels ; d'autres visent à déterminer le paramétrage optimal d'une méthode particulière.

De nombreux travaux ont comparé les différents types de contextes que peuvent exploiter les méthodes distributionnelles, que nous avons décrits à la section 3.2.2. Certains ont comparé les cooccurrents et les segments textuels [107, 165]. D'autres ont comparé les cooccurrents graphiques et syntaxiques. À ce sujet, Baroni et Lenci [11] affirment que les travaux antérieurs suggèrent que les modèles basés sur la cooccurrence syntaxique ont un léger avantage par rapport aux modèles exploitant la cooccurrence graphique, bien qu'ils soient plus difficiles à mettre en œuvre, à cause du prétraitement linguistique plus complexe. En effet, plusieurs évaluations comparatives indiquent que la cooccurrence syntaxique donne de meilleurs résultats [70, 96, 143], bien que certaines suggèrent le contraire [102] et d'autres produisent des résultats mitigés [183]. Levy et

[29] *JeuxDeMots* [103], une ressource créée de manière collaborative au moyen d'un jeu et distribuée librement, contient des relations de dérivation morphologique ainsi que des mots de la même famille lexicale ; voir `http://www.jeuxdemots.org/jdm-about-detail-relations.php` (page consultée le 12 novembre 2016). À notre connaissance, ces données n'ont pas été utilisées pour évaluer des modèles distributionnels.

[30] Rappelons que les dérivés syntaxiques sont souvent des dérivés morphologiques, mais pas toujours.

Goldberg [108] montrent que les voisins distributionnels que l'on obtient pour une requête donnée en utilisant la cooccurrence syntaxique (en l'occurrence, dans un modèle de langue neuronal) sont plus souvent reliés à la requête par une relation de similarité sémantique à proprement parler, et contiennent plus souvent des cohyponymes et des mots de la même partie du discours que la requête. Par contre, Rapp [159, p. 4] observe que les cooccurrents graphiques produisent d'aussi bons résultats que les cooccurrents syntaxiques, et conclut qu'il n'est pas nécessaire de recourir à l'analyse syntaxique sous le prétexte que les cooccurrents graphiques contiennent beaucoup de bruit. Quoi qu'il en soit, les travaux comparant les cooccurrents graphiques et syntaxiques suggèrent, dans l'ensemble, que ces derniers offrent une précision légèrement supérieure pour la détection de relations paradigmatiques, bien qu'ils exigent un prétraitement plus complexe.

D'autres travaux ont comparé différentes façons de construire des modèles distributionnels, telles que l'AD et les modèles de langue neuronaux. Baroni et al. [10] réalisent une évaluation comparative entre l'AD et `word2vec` sur plusieurs jeux de données et observent que ce dernier offre systématiquement de meilleurs résultats. Par contre, Ferret [69] observe le contraire lorsqu'il compare deux thésaurus construits à partir des mêmes représentations lexicales qu'utilisaient Baroni et al. [10] ; Ferret affirme au sujet des modèles de langue neuronaux que « l'utilisation de ce type de représentations distribuées n'est pas encore une option intéressante pour la construction de thésaurus distributionnels » [70, p. 31]. Levy et al. [110] montrent que lorsqu'on optimise correctement les (hyper)paramètres de l'AD et de `word2vec`[31], la méthode qui produit les meilleurs résultats dépend de la tâche utilisée pour l'évaluation, et les deux produisent souvent des résultats d'une qualité similaire. Soulignons que nous ne connaissons aucun travail comparant ces deux types de modèles sur des corpus spécialisés et des ressources lexicales spécialisées.

Un autre objectif important de l'évaluation des modèles distributionnels est l'analyse

[31] Plus précisément, ils utilisent l'architecture *skip-gram*, entraînée par échantillonnage d'exemples négatifs.

de l'influence de leurs (hyper)paramètres, tels que la taille de la fenêtre de contexte dans le cas des modèles qui exploitent la cooccurrence graphique. L'analyse de l'influence des (hyper)paramètres est primordiale parce que le paramétrage des modèles distributionnels a une influence importante sur la qualité des résultats qu'ils offrent et sur le genre de relations lexicales qu'ils captent. Levy et al. [110] montrent que le paramétrage des méthodes basées sur la cooccurrence graphique a une influence plus importante sur la qualité des résultats que le choix de la méthode (en l'occurrence, l'AD ou un modèle de langue neuronal) ; ils montrent d'ailleurs qu'il peut être plus profitable d'optimiser davantage les (hyper)paramètres que d'utiliser un plus gros corpus (en l'occurrence, un corpus de 10.5 milliards de mots plutôt que 1.5 milliards).

Dans le cas de l'AD, des études réalisées dès les années 1960 auraient montré, entre autres, que la taille de la fenêtre de contexte a une influence sur le genre de relations qui sont captées par le modèle [140], les fenêtres étroites captant des relations plutôt paradigmatiques et les fenêtres larges, des relations plutôt thématiques, par exemple. Elles auraient également montré l'importance de pondérer les fréquences de cooccurrence.

Plus récemment, plusieurs études systématiques de l'influence des paramètres de l'AD ont été réalisées [33, 34, 95, 102, 105]. L'influence des hyperparamètres des modèles de langue neuronaux a reçu comparativement peu d'attention, ces modèles étant relativement nouveaux.

Les évaluations visant à optimiser le paramétrage des modèles distributionnels prennent parfois en compte les types de relations que l'on souhaite identifier. Sahlgren [165, ch. 12 et 13] examine l'influence des relations ciblées sur le paramétrage optimal d'une méthode distributionnelle, et observe que les paramétrages qui produisent les meilleurs résultats pour les antonymes sont très différents de ceux pour les synonymes. Lapesa et al. [106] analysent l'influence des paramètres de l'AD sur sa capacité à capter une variété de relations, au moyen d'un jeu de données qui distingue différentes relations paradigmatiques (synonymie, antonymie, cohyponymie) et syntagmatiques. En ce qui concerne les modèles de langue neuronaux, nous ne connaissons aucun travail traitant

systématiquement l'influence de leurs hyperparamètres sur leur capacité à détecter différentes relations lexicales.

L'analyse du paramétrage peut également prendre en compte la partie du discours (PDD) des mots utilisés pour interroger le modèle [178]. Par exemple, Hill et al. [96] montrent que les fenêtres de contexte étroites (en l'occurrence, de 2 mots) produisent de meilleurs résultats pour les verbes et les adjectifs que les fenêtres larges (de 10 mots), mais n'observent pas la même tendance en ce qui concerne les noms.

La prise en compte de la PDD permet de déterminer quelles méthodes et quels paramétrages produisent les meilleurs résultats pour chaque PDD, mais aussi quelles PDD sont le mieux modélisées par l'approche distributionnelle. Par exemple, Hill et al. [96] montrent que les adjectifs sont mieux modélisés que les noms et les verbes ; en revanche, Fabre et al. [61] observent, sur un jeu de données différent, que l'adjectif est la PDD la plus difficile à modéliser (bien que cette catégorie était la plus facile à annoter lors de la création du jeu de données). En comparant les résultats obtenus sur trois tâches de classification sémantique, Van de Cruys [183] et Bullinaria [32] montrent que les verbes sont plus difficiles à classifier que les noms, du moins lorsqu'on exploite l'approche distributionnelle. De même, la prise en compte des relations ciblées permet de déterminer les paramétrages qui produisent les meilleurs résultats pour chaque relation, mais aussi les relations que les modèles distributionnels captent le mieux. L'analyse du paramétrage pourrait également tenir compte d'autres facteurs liés à l'application envisagée, tels que la langue traitée, mais nous ne connaissons aucun travail où cet aspect a été traité systématiquement.

Ainsi, l'analyse du paramétrage occupe une place importante dans les travaux en sémantique distributionnelle, y compris ceux exploitant des corpus spécialisés. Entre autres, Ferret [66], Périnet et Hamon [149] et QasemiZadeh [157] analysent l'influence de plusieurs des paramètres de l'AD en utilisant des ressources terminologiques comme données de référence.

L'évaluation des modèles distributionnels peut être utilisée pour explorer non seule-

ment l'influence de la façon dont ils sont paramétrés, mais aussi d'autres facteurs qui influent sur la qualité des résultats qu'ils produisent, tels que la taille des corpus. En effet, plusieurs travaux indiquent que la qualité des résultats augmente en fonction de la taille des corpus [3, 34, 106], comme nous l'avons mentionné à la section 3.2.6.

3.4 Les graphes de voisinage distributionnel

Comme nous l'avons signalé à la section 3.2.3, la façon la plus courante d'interroger un modèle sémantique distributionnel consiste simplement à calculer les PPV d'une requête particulière. Gyllensten et Sahlgren [85] soulignent que ce mode d'interrogation a plusieurs désavantages. Premièrement, consulter une liste de PPV ne fournit aucune information sur la structure interne du voisinage de la requête. En effet, une liste de PPV nous montre quels mots sont distributionnellement similaires à la requête, mais ne nous permet pas de savoir, parmi les voisins, lesquels sont particulièrement similaires entre eux. De plus, si la requête a plusieurs sens dans le corpus, une liste de PPV ne fournit aucun moyen de distinguer les voisins correspondant à chacun de ses sens. Ainsi, le fait de ne pas distinguer les différents sens d'un mot ne serait pas un défaut des modèles distributionnels eux-mêmes (contrairement à ce qu'affirmait Sahlgren [165], mais aussi plusieurs autres auteurs), mais de la façon de les interroger.

Gyllensten et Sahlgren [85] proposent donc d'utiliser un graphe pour interroger les modèles distributionnels plutôt que simplement consulter des listes de PPV.

Toute ressource décrivant des relations lexicales peut être considérée comme un réseau lexical, c'est-à-dire un graphe dont les nœuds sont les unités du lexique et les arêtes sont les relations auxquelles elles participent. De même, les liens de similarité entre les mots que l'on détecte au moyen d'un modèle sémantique distributionnel peuvent être représentés au moyen d'un graphe, comme nous le montrons ci-dessous. Il est intéressant de noter que ces *graphes de voisinage distributionnel* et les réseaux lexicaux ont des propriétés en commun. En effet, Steyvers et Tenenbaum [177] ont montré que les deux

types de graphes, bien qu'ils présentent des différences importantes, ont une structure semblable à certains égards (ce sont des graphes de *petits mondes*).

La construction de réseaux lexicaux peut donc s'appuyer sur des graphes de voisinage distributionnel. Par exemple, Morardo et Villemonte de La Clergerie [137] présentent une plateforme de production de ressources lexicales qui repose sur un modèle distributionnel et une interface Web permettant de valider et de visualiser les relations au moyen de graphes. C'est dans cette optique que nous concevons l'apport de la sémantique distributionnelle au travail terminologique, c'est-à-dire comme un moyen d'assister la création de réseaux lexicaux afin de rendre compte des relations entre les termes.

Plusieurs types de graphes peuvent servir à interroger un modèle sémantique distributionnel. Gyllensten et Sahlgren [85] utilisent un type de graphe appelé *graphe de voisinage relatif*. Dans ce travail, nous faisons appel à un type de graphe plus simple (et moins coûteux d'un point de vue computationnel) appelé *graphe de k plus proches voisins* (ou *graphe de k PPV*).

Comme son nom l'indique, la construction du graphe de k PPV repose sur le calcul des PPV, que nous avons illustré à la section 3.2.3. Ce calcul est souvent utilisé pour créer un *thésaurus distributionnel*, ressource dans laquelle chaque entrée est accompagnée d'une liste de ses PPV, comme dans les exemples suivants[32] :

- *absorb : emit, sequester, convert, produce, accumulate, store, radiate, consume, remove, reflect, ...*
- *extreme : severe, intense, harsh, catastrophic, unusual, seasonal, mild, cold, dramatic, increase, ...*
- *precipitation : rainfall, snowfall, temperature, rain, evaporation, runoff, moisture, snow, weather, deposition, ...*

Un thésaurus distributionnel qui associe un nombre fixe de voisins à chaque entrée peut être considéré comme un graphe de k PPV, c'est-à-dire un graphe dans lequel

[32]Ces exemples sont extraits d'un article que nous avons présenté à l'atelier *SEM [16].

chaque terme est relié à ses k plus proches voisins. Par exemple, on peut représenter l'entrée *extreme* et ses 5 PPV au moyen du graphe présenté dans la Figure 3.6.

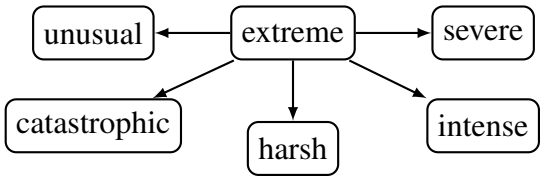


Figure 3.6 – Graphe de 5 PPV du terme *extreme*.

Plutôt qu'utiliser seulement des listes de PPV pour interroger les modèles distributionnels, nous exploitons de manière explicite[33] le graphe de k PPV.

Les graphes de voisinage comme le graphe de k PPV se prêtent bien à l'interrogation de modèles distributionnels pour plusieurs raisons. Premièrement, ils fournissent un moyen simple de visualiser les similarités captées par le modèle [42]. La dimension élevée des représentations lexicales peut constituer un obstacle à leur visualisation : si elles avaient seulement 2 ou 3 composantes, il serait simple de les représenter visuellement au moyen d'une sorte de carte sémantique, mais elles ont généralement des centaines de composantes ou plus[34]. Or, les graphes de voisinage représentent les liens de similarité entre les représentations lexicales plutôt que les représentations elles-mêmes, donc leur dimension élevée ne pose aucune difficulté. Par ailleurs, il existe de nombreuses méthodes pour dessiner des graphes et de nombreux outils qui mettent en œuvre ces méthodes. Les graphes de voisinage sont donc un moyen simple d'obtenir une interface visuelle pour la découverte et la validation de relations lexicales à partir d'un modèle distributionnel, comme nous l'illustrons au chapitre 7.

Un autre avantage des graphes de voisinage est qu'ils offrent au terminologue un moyen efficace d'analyser les voisins d'un *ensemble* de termes, tel qu'un ensemble d'unités lexicales qui évoquent un cadre sémantique. Supposons que l'on ait identifié

[33] Si un thésaurus distributionnel représente un graphe de k PPV, on n'exploite pas la structure de ce graphe lorsqu'on consulte la liste triée des PPV d'une requête.

[34] Il est possible de cartographier les représentations lexicales dans un espace de dimension 2 ou 3 en faisant appel à une technique de réduction de dimension.

un cadre sémantique ainsi que deux termes qui évoquent ce cadre : *warm* et *cool*. Pour enrichir cette liste, on pourrait prendre les PPV de chaque terme et valider chaque voisin en déterminant s'il évoque le même cadre sémantique. Mais il risque d'y avoir beaucoup de recoupement entre les deux listes de PPV : si *warm* et *cool* apparaissent dans des contextes similaires, la liste des termes distributionnellement similaires à *warm* (ses PPV) risque de ressembler beaucoup à la liste des termes distributionnellement similaires à *cool*, puisque ces deux termes ont eux-mêmes des distributions semblables ; cette redondance sera d'autant plus importante que la liste de termes dont on observe les PPV est longue. On pourrait éliminer le recoupement entre les listes de PPV en prenant leur union ou en supprimant au fur et à mesure les voisins déjà validés, mais on peut aussi exploiter un graphe de voisinage pour résoudre ce problème.

Par exemple, supposons que les 5 PPV de ces deux termes soient :

- *cool : cooling, cooler, mild, warming, warm.*
- *warm : warming, cool, warmer, hot, mild.*

Le graphe de k PPV que l'on obtient en reliant ces deux termes à leurs 5 PPV est illustré dans la Figure 3.7.

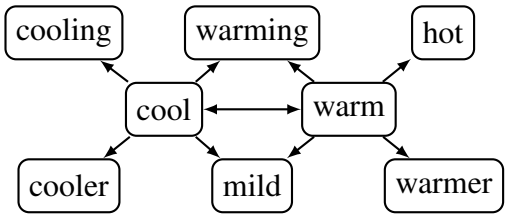


Figure 3.7 – Graphe de 5 PPV des termes *warm* et *cool*.

Si l'on construisait le graphe en prenant en compte les PPV de tous les mots, non seulement ceux de la requête qui nous intéresse, on obtiendrait une perspective plus complète de la structure du voisinage de cette requête ; il pourrait effectivement y avoir des connexions supplémentaires entre les voisins de la requête. C'est cette méthode que nous utilisons pour illustrer des exemples de voisinages au chapitre 7.

Le graphe nous permet donc d'analyser le voisinage distributionnel d'un ensemble de termes en évitant la redondance que l'on risquerait d'observer dans les listes de PPV de chacun des termes, tout en rendant compte des différents liens entre ces termes et leurs voisins.

En effet, le fait que le graphe de k PPV rend compte des différents liens de voisinage au sein d'un ensemble de mots constitue en soi un avantage. Cette structure interne du voisinage peut notamment être exploitée afin d'identifier les différents sens d'un mot, comme le soulignent Gyllensten et Sahlgren [85]. Prenons un exemple tiré d'un réseau lexical construit manuellement (plutôt qu'un graphe de voisinage distributionnel). Le DiCoEnviro[35] décrit deux acceptions différentes du terme *stockage*, une liée au sous-domaine des véhicules électriques et une autre à celui des changements climatiques[36]. Les liens de (quasi-)synonymie et d'antonymie auxquels participent ce terme et ses termes reliés sont présentés au moyen d'un graphe dans la Figure 3.8. Il est important de noter que les nœuds de ce graphe correspondent à des formes linguistiques, non pas à des sens ; ainsi, certains nœuds reliés au nœud *stockage* correspondent à un sens de ce terme, et d'autres, à son autre sens.

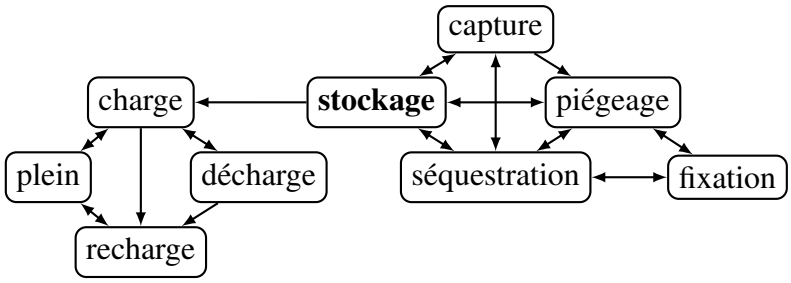


Figure 3.8 – Graphe représentant les relations de (quasi-)synonymie et d'antonymie autour du terme *stockage*.

Dans cette figure, on peut observer à gauche et à droite du terme *stockage* deux sous-ensembles de nœuds entre lesquels il existe de nombreuses connexions, ces deux

[35]Les données présentées dans cet exemple ont été récupérées le 13 juillet 2015.

[36]Le dictionnaire distingue en fait deux sens de ce terme dans le sous-domaine des changements climatiques, mais nous ne tenons pas compte de cette distinction.

ensembles correspondant aux deux acceptions de ce terme :

1. *charge*, *décharge*, *recharge*, *plein* ;

2. *séquestration*, *capture*, *piégeage*, *fixation*.

De la même façon, un graphe de voisinage distributionnel peut éventuellement rendre compte des différents sens d'une même forme linguistique, même si une seule représentation lexicale est associée à chaque forme linguistique et qu'un seul nœud est utilisé pour représenter chaque forme linguistique dans le graphe.

Soulignons enfin que les graphes se prêtent à l'identification d'ensembles de termes sémantiquement reliés (ou classes sémantiques), que celle-ci soit réalisée manuellement (par inspection visuelle) ou automatiquement[37]. Dans les deux cas, l'objectif est d'identifier des ensembles de nœuds qui participent entre eux à un nombre élevé de connexions, mais qui participent à un nombre relativement faible de connexions avec les autres nœuds du graphe. Les ensembles identifiés de cette façon constituent des ensembles de mots apparaissant dans des contextes similaires, et les mots qui apparaissent dans des contextes similaires ont tendance à avoir des sens similaires, comme le stipule l'hypothèse distributionnelle.

En somme, les graphes de k PPV constituent un moyen simple et efficace d'interroger un modèle distributionnel, qui offre plus de possibilités que la consultation de listes de PPV.

S'ils sont parfois utilisés pour visualiser les liens de similarité distributionnelle, les graphes de voisinage distributionnel ont également été utilisés à d'autres fins, notamment pour l'induction des usages de mots ; on a d'abord utilisé des graphes de cooccurrence à cette fin [51, 186], mais des graphes de voisinage distributionnel ont également été utilisés [65, 141]. En outre, Claveau et al. [41] considèrent les thésaurus distributionnels

[37] De nombreuses techniques de partitionnement de données (*clustering*) basées sur graphe peuvent être utilisées à cette fin ; pour une synthèse, voir l'étude de Chen et Ji [37]. En outre, nous avons montré que l'identification des composantes connexes d'un graphe à 1 PPV symétrique est un moyen simple et efficace d'identifier automatiquement des classes de termes distributionnellement similaires [17].

comme des graphes de k PPV, et exploitent la structure de ces graphes pour améliorer la qualité des thésaurus.

En ce qui concerne l'évaluation des graphes de voisinage distributionnel, nous ne connaissons aucun travail comparant différents types de graphes dans le cadre de l'identification de relations lexicales.

3.5 L'acquisition automatique de cadres sémantiques

Cette thèse est également liée à la recherche sur la sémantique des cadres. Au chapitre 2, nous avons énuméré quelques travaux portant sur l'utilisation de la sémantique des cadres dans le domaine de la terminologie. Cette thèse est reliée à ces travaux, puisque nous utilisons la sémantique des cadres comme cadre descriptif, mais elle est également liée aux travaux sur l'acquisition automatique de cadres sémantiques. Elle ne vise pas à identifier automatiquement des cadres sémantiques complets (qui comprennent non seulement des unités lexicales, mais aussi des éléments de cadre), mais un de ses objectifs est de vérifier si la sémantique distributionnelle permet de détecter des termes qui évoquent le même cadre sémantique. Il nous semble donc important de souligner les travaux visant à automatiser l'identification de cadres sémantiques, pour mieux situer notre recherche et mettre en valeur ses retombées. Or, ces travaux sont peu nombreux, à notre connaissance.

En effet, les travaux en TAL liés à la sémantique des cadres visent surtout à reproduire automatiquement le genre d'annotations que l'on retrouve dans la ressource FrameNet [79, 122] ou à exploiter les données de FrameNet dans différentes applications du TAL, telles que la traduction automatique ou les systèmes de question-réponse [45]. D'autres travaux tentent de créer des ressources semblables à FrameNet dans d'autres langues ou à augmenter la couverture de FrameNet [47].

Un des seuls travaux que nous connaissons sur l'acquisition automatique de cadres sémantiques, celui de Green et al. [81], vise à identifier des classes de verbes semblables

aux ensembles d'unités lexicales dans FrameNet, et à étiqueter ces classes ainsi que leur structure argumentale, en exploitant une ressource lexicale existante, WordNet [64].

Church [38] se demande quand il sera possible d'apprendre des cadres sémantiques à partir de corpus, et croit que la disponibilité de corpus volumineux facilitera cette tâche. Soulignons qu'il existe des travaux visant à apprendre des schémas d'événements semblables à des cadres sémantiques, tels que celui de Chambers et Jurafsky [36] ; d'ailleurs, les auteurs évaluent leur méthode en comparant les descriptions qu'elle produit à celles que l'on trouve dans FrameNet. Soulignons également le travail de Van de Cruys [184], qui propose une méthode pour modéliser les préférences sélectionnelles des verbes, qui permet également d'induire des sortes de cadres sémantiques.

3.6 Synthèse

Dans ce chapitre, nous avons présenté une revue de la littérature sur les méthodes utilisées pour détecter automatiquement des relations lexicales, en particulier la sémantique distributionnelle. L'accent a été placé sur les travaux exploitant des corpus spécialisés et sur l'évaluation des méthodes distributionnelles, puisque nous utilisons des corpus spécialisés et que l'évaluation joue un rôle important dans cette thèse. Nous avons également expliqué comment les modèles distributionnels sont construits et interrogés. Enfin, nous avons fait un bref survol des travaux sur l'acquisition automatique de cadres sémantiques.

Au chapitre 4, nous énoncerons les problématiques que nous abordons dans cette thèse, que nous avons évoquées dans ce chapitre, ainsi que nos hypothèses et objectifs.

CHAPITRE 4

PROBLÉMATIQUES, HYPOTHÈSES ET OBJECTIFS

Introduction

Au chapitre 3, nous avons fait un survol de la littérature sur la sémantique distributionnelle et son utilisation pour la détection de relations lexicales, en particulier à partir de corpus spécialisés. Dans ce chapitre, nous présentons les problématiques spécifiques que nous abordons dans cette thèse, ainsi que nos hypothèses et nos objectifs.

4.1 Problématiques

Bien qu'il existe une littérature abondante sur la sémantique distributionnelle, nous avons identifié plusieurs questions qui, à notre connaissance, n'ont pas de réponses définitives à l'heure actuelle, et que nous abordons dans cette thèse. Toutes ces questions sont reliées à des facteurs dont devrait tenir compte un cadre méthodologique pour l'application de l'approche distributionnelle dans le contexte du travail terminologique. Certains de ces facteurs sont liés à l'objectif visé, d'autres sont plutôt liés aux méthodes distributionnelles elles-mêmes.

Nous énumérons donc ci-dessous les problématiques abordées dans cette thèse ; celles-ci ont toutes été évoquées à la section 3.3 sur l'évaluation des modèles distributionnels, à l'exception de la dernière, qui porte sur la méthode utilisée pour interroger les modèles distributionnels et que nous avons évoquée à la section 3.4.

Le cadre descriptif Dans quelle mesure l'apport des méthodes distributionnelles au travail de description terminologique dépend-il du cadre descriptif adopté ? Plus précisément, est-ce que ces méthodes permettent de détecter des termes évoquant le même cadre sémantique aussi bien que des relations lexicales telles que la synonymie ?

Les relations ciblées Il n'est pas tout à fait clair dans quelle mesure la qualité des résultats qu'offrent les méthodes distributionnelles dépend des relations lexicales ciblées. Par exemple, est-ce que les modèles distributionnels captent la dérivation syntaxique aussi bien que des relations paradigmatiques classiques telles que la synonymie et l'antonymie ? Par ailleurs, comment l'influence des (hyper)paramètres des modèles distributionnels varie-t-elle en fonction des relations lexicales ciblées ?

La langue traitée Est-ce que la qualité des modèles distributionnels dépend de la langue traitée ? Est-ce que l'on doit tenir compte de la langue traitée lorsqu'on fixe les (hyper)paramètres d'un modèle distributionnel ?

Le choix de la méthode distributionnelle Comment se comparent les modèles de langue neuronaux à la méthode classique d'AD si on exploite des corpus spécialisés et des données de référence extraites de dictionnaires spécialisés ?

L'influence des (hyper)paramètres L'influence des paramètres de l'AD a fait l'objet de nombreux travaux, mais celle des hyperparamètres du modèle de langue neuronal implémenté dans `word2vec` n'a pas été étudiée aussi systématiquement. Quelle est l'influence des hyperparamètres de `word2vec` sur sa capacité à détecter différentes relations lexicales ?

Le mode d'interrogation Si on interroge un modèle distributionnel au moyen d'un graphe de voisinage, quelle est l'influence des paramètres du graphe sur la précision des voisinages qu'il représente ?

Cette dernière problématique sera abordée spécifiquement dans la deuxième partie de l'analyse des résultats, au chapitre 7.

4.2 Hypothèses

Nous abordons les problématiques spécifiques énumérées à la section 4.1 en réalisant une expérience qui consiste à construire des modèles sémantiques distributionnels à partir de corpus spécialisés et à les évaluer au moyen de données de référence extraites de dictionnaires spécialisés. D'un point de vue général, cette expérience vise à valider deux hypothèses principales, chacune étant reliée à un des deux cadres descriptifs que nous avons présentés au chapitre 2 :

- **Hypothèse 1** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme donné, des termes reliés sur le plan paradigmatique, tels que des (quasi-)synonymes, des antonymes, des hyperonymes, des hyponymes et des dérivés syntaxiques.

- **Hypothèse 2** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme ou un ensemble de termes donné, des termes évoquant le même cadre sémantique.

En ce qui concerne la première hypothèse, des hypothèses semblables ont déjà fait l'objet de nombreux travaux, comme nous l'avons expliqué au chapitre 3, mais cette thèse se distingue de ceux-ci en abordant les problématiques présentées à la section 4.1.

Pour ce qui est de la seconde hypothèse, nous ne connaissons aucun travail qui a exploré cette hypothèse particulière, bien que des questions reliées aient été abordées dans des travaux portant sur l'acquisition automatique de cadres sémantiques, que nous avons évoqués à la section 3.5.

Une précision d'ordre méthodologique mérite d'être soulignée en ce qui concerne la seconde hypothèse : nous supposons que des requêtes ont été choisies au préalable (une requête étant un terme ou un ensemble de termes évoquant un cadre sémantique), et que la tâche à réaliser consiste à identifier des termes qui évoquent le même cadre sémantique que ces requêtes. Il serait possible de définir cette tâche autrement, notam-

ment en choisissant un ensemble complet de termes qu'il s'agirait de partitionner en sous-ensembles de termes évoquant le même cadre sémantique[1]. Nous avons choisi de ne pas supposer que tous les termes à regrouper sont connus *a priori*. Nous supposons plutôt que l'on connait un certain nombre de termes évoquant des cadres sémantiques du domaine de l'environnement, et qu'il s'agit d'enrichir les ensembles de termes évoquant chaque cadre (ou de découvrir des termes évoquant des cadres reliés). En définissant la tâche de cette façon, nous pouvons utiliser les mêmes méthodes pour évaluer nos deux hypothèses (nous reviendrons sur ce point à la section 5.1).

4.3 Objectifs

Ce travail comporte plusieurs objectifs. D'une manière générale, il vise à vérifier si les modèles sémantiques distributionnels constituent un moyen efficace d'assister le terminologue dans l'analyse des relations lexicales entre termes à partir d'un corpus spécialisé, et ainsi de faciliter une description terminologique basée sur un cadre descriptif particulier. En effet, la mise en œuvre et l'évaluation des méthodes distributionnelles seront encadrées par deux cadres descriptifs différents, et nous avons formulé deux hypothèses générales, chacune étant liée à une de ces approches.

D'une manière plus spécifique, cette thèse vise à caractériser les différents facteurs dont dépend la qualité des résultats qu'offrent les méthodes distributionnelles dans le contexte du travail terminologique. Ainsi, un des objectifs de ce travail consiste à identifier les paramétrages optimaux des modèles distributionnels pour l'identification de relations lexicales en corpus spécialisé. Les différents (hyper)paramètres de ces modèles peuvent avoir une influence importante sur leur capacité à capter la similarité sémantique des termes et sur le genre de relations qui sont détectées (voir section 3.3.2, entre autres). Il s'agit donc d'identifier les paramétrages qui offrent les meilleurs résultats pour cette application particulière.

[1] En l'occurrence, on priviliégierait éventuellement un partitionnement dit *flou*, puisqu'une même forme linguistique peut évoquer plus d'un cadre sémantique.

Un autre objectif consiste à déterminer quels facteurs doivent être pris en compte lors du choix de la méthode distributionnelle et de la sélection de ses (hyper)paramètres. Par exemple, nous vérifions quels paramètres doivent être fixés en tenant compte de la langue traitée (le français ou l'anglais) et si ces paramètres peuvent être fixés de façon à assurer de bons résultats dans les deux langues. De même, nous vérifions si le cadre descriptif adopté, les relations ciblées ou la partie du discours des termes sont déterminants quant à la qualité des résultats que l'on obtient et à la configuration optimale des méthodes distributionnelles.

Cette thèse vise également à explorer différentes façons d'interroger un modèle distributionnel pour assister l'identification des relations lexicales entre termes. La méthode de base utilisée pour interroger ces modèles consiste à obtenir, pour un terme donné, une liste triée de termes similaires. Nous évaluons cette méthode, mais aussi différents types de graphes de voisinage que l'on peut utiliser pour interroger ces modèles, et nous vérifions comment les résultats varient en fonction du type de graphe utilisé.

Pour mettre en valeur les liens entre ces différents objectifs, on peut formuler l'objectif global de cette thèse, dans lequel s'imbriquent les objectifs mentionnés ci-dessus, de la façon suivante : *cette thèse vise à développer un cadre méthodologique basé sur la sémantique distributionnelle pour l'analyse des relations lexicales auxquelles participent les termes d'un domaine de spécialité*.

4.4 Synthèse

Dans ce chapitre, nous avons présenté les problématiques spécifiques abordées dans cette thèse. Nous avons ensuite formulé deux hypothèses générales et énoncé nos objectifs.

Cette thèse se distingue des travaux qui lui sont reliés notamment par le nombre de facteurs pris en compte dans l'expérience que nous avons réalisée : deux langues ; deux cadres descriptifs ; deux méthodes pour construire les modèles distributionnels,

comportant chacune plusieurs (hyper)paramètres dont nous analysons l'influence ; deux modes d'interrogation, à savoir les listes de PPV et les graphes de voisinage ; trois parties du discours ; et quatre types de relations lexicales paradigmatiques. Tous ces aspects de notre méthodologie seront décrits au chapitre 5.

CHAPITRE 5

MÉTHODOLOGIE

Introduction

Les objectifs énoncés au chapitre 4 sont réalisés au moyen d'une expérience qui consiste à construire et à évaluer des modèles sémantiques distributionnels. Cette expérience fait appel à différentes techniques servant à estimer la similarité sémantique, à exploiter cette information afin de faciliter l'identification de relations lexicales et de cadres sémantiques dans le cadre du travail terminologique, et à évaluer la qualité des résultats. Avant de décrire en détail les techniques et outils utilisés, nous ferons un survol de notre méthodologie afin de fournir une vue d'ensemble de ses différentes composantes. Nous décrivons ensuite les corpus que nous utilisons à la section 5.2. La section 5.3 porte sur la sélection des mots-cibles pour lesquels nous produisons des représentations. Nous présentons les deux méthodes que nous exploitons pour construire des modèles distributionnels à la section 5.4 et les graphes que nous utilisons pour interroger ces modèles à la section 5.5. Enfin, la section 5.6 porte sur notre méthodologie d'évaluation.

5.1 Survol de la méthode

Notre méthodologie comporte essentiellement trois composantes : le calcul de la similarité distributionnelle, la construction de graphes de voisinage distributionnel et l'évaluation des résultats. Elle comporte également une étape préliminaire qui consiste à créer ou à adapter les ressources nécessaires pour réaliser ces tâches, à savoir un corpus et des données de référence.

À la section 3.2.3, nous avons montré comment construire un modèle sémantique distributionnel et calculer la similarité des mots représentés par le modèle. La méthode que nous avons utilisée dans cet exemple est l'analyse distributionnelle (AD). Dans le

cadre de ce travail, nous faisons appel à deux méthodes distributionnelles différentes, à savoir l'AD et le modèle de langue neuronal implémenté dans l'outil `word2vec` ; nous expliquerons comment nous construisons et paramétrons ces deux types de modèles à la section 5.4.

Une fois les modèles construits, nous utilisons le cosinus de l'angle des vecteurs pour mesurer la similarité distributionnelle entre les mots-cibles, ce qui nous permet d'obtenir pour chaque mot-cible une liste triée de ses voisins distributionnels, comme nous l'avons expliqué à la section 3.2.3.

La qualité des listes ordonnées de voisins que l'on obtient de cette façon est évaluée de manière quantitative en les comparant à des données de référence que nous avons extraites de dictionnaires spécialisés.

Cette méthode d'évaluation a plusieurs objectifs. Un premier objectif est de mettre à l'épreuve nos deux hypothèses :

- **Hypothèse 1** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme donné, des termes reliés sur le plan paradigmatique, tels que des (quasi-)synonymes, des antonymes, des hyperonymes, des hyponymes et des dérivés syntaxiques.

- **Hypothèse 2** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme ou un ensemble de termes donné, des termes évoquant le même cadre sémantique.

Autrement dit, cette méthode d'évaluation vise à vérifier dans quelle mesure les voisinages distributionnels correspondent à des voisinages *sémantiques*.

Cette évaluation vise également à optimiser les (hyper)paramètres des méthodes utilisées et à caractériser leur influence sur les résultats obtenus. Comme nous l'avons expliqué au chapitre 3, lorsqu'on contruit un modèle distributionnel, certains paramètres doivent être fixés au préalable, tels que la taille de la fenêtre de contexte dans le cas des modèles qui exploitent la cooccurrence graphique. Ces paramètres peuvent avoir une

influence importante sur les résultats que l'on obtient. Une recherche systématique des valeurs optimales de chaque paramètre peut être réalisée de la manière suivante :

- On définit l'ensemble des valeurs possibles ou plausibles pour chaque paramètre (p. ex. 1 à 10 mots dans le cas de la taille de la fenêtre de contexte).
- On construit des modèles correspondant à chaque paramétrage possible, c'est-à-dire chaque combinaison possible des différentes valeurs des paramètres.
- On évalue chaque modèle en calculant des mesures d'évaluation basées sur les données de référence.

En évaluant de manière quantitative différents paramétrages, nous pouvons déterminer la valeur optimale de chacun des (hyper)paramètres et déterminer de quelle façon ils influencent la qualité des résultats. Nous décrirons notre méthodologie d'évaluation en détail à la section 5.6.

Nous évaluons non seulement la qualité des listes ordonnées de PPV que l'on obtient à partir des modèles distributionnels, mais aussi différentes façons d'interroger ces modèles. En effet, nous utilisons ces listes de PPV afin de construire des graphes représentant les liens de voisinage distributionnel entre les termes. En interrogeant les modèles au moyen de différents types de graphes de voisinage, nous pouvons vérifier comment la qualité des résultats varie en fonction du type de graphe utilisé.

Les graphes que nous utilisons dans ce travail sont des graphes de k PPV. Au chapitre 3, nous avons illustré la construction de ces graphes, et nous avons souligné plusieurs avantages qu'offrent les graphes de voisinage par rapport aux listes de PPV en ce qui concerne l'interrogation des modèles distributionnels. Entre autres, le fait qu'ils offrent un moyen efficace d'analyser le voisinage d'un ensemble de termes, et qu'ils se prêtent bien à l'identification d'ensembles de termes reliés, les rend particulièrement utiles pour l'identification ou l'enrichissement de cadres sémantiques.

Ces graphes sont également évalués au moyen des données de référence afin de déterminer de manière quantitative la qualité des voisinages qu'ils représentent, et pour

analyser l'influence de leurs paramètres. En effet, les graphes de k PPV ont également quelques paramètres, tels que la valeur de k ; nous les optimisons au moyen de la même méthode que nous utilisons pour optimiser les (hyper)paramètres des modèles distributionnels, à savoir une évaluation comparative de différents paramétrages du graphe, mais en utilisant des mesures d'évaluation adaptées aux voisinages (non ordonnés) que contiennent ces graphes.

Nous présenterons également, au chapitre 7, une évaluation plutôt qualitative de ces graphes, en analysant des exemples visuels de voisinages extraits de ces derniers.

Comme nous l'avons souligné à la section 4.2, il serait possible de traiter l'identification d'ensembles de termes qui évoquent le même cadre sémantique comme une tâche de partitionnement, ce qui nous amènerait à adopter une méthodologie différente et des mesures d'évaluation différentes. En effet, si on connaissait *a priori* tous les termes que l'on voudrait inclure dans une description basée sur la sémantique des cadres, une approche basée sur le partitionnement pourrait être envisagée. Nous avons choisi de traiter ce problème non pas en appliquant une technique de partitionnement à un ensemble de termes choisis *a priori*, mais en explorant un graphe de voisinage distributionnel contenant tous les mots-cibles pour lesquels des représentations ont été produites ; nous considérons l'exploration de ce graphe comme un moyen de découvrir, à partir d'un terme ou d'un ensemble de termes donné, des termes évoquant le même cadre sémantique ou des cadres reliés, cette méthode pouvant ainsi servir à identifier ou à enrichir des cadres sémantiques. Cela nous permet notamment d'utiliser les mêmes méthodes d'évaluation pour valider nos deux hypothèses, en changeant simplement le jeu de données de référence utilisé pour calculer les mesures d'évaluation.

5.2 Corpus et prétraitement

Les deux corpus que nous utilisons afin de construire des modèles distributionnels sont des corpus distribués librement à des fins de recherche par l'Agence pour l'évalua-

tion et la distribution des ressources linguistiques (ELDA) :

1. le corpus monolingue français PANACEA – domaine de l'environnement (code de catalogue ELRA-W0065)[1] ;

2. le corpus monolingue anglais PANACEA – domaine de l'environnement (code de catalogue ELRA-W0063)[2].

Ces corpus ont été compilés au moyen d'un outil de construction automatique de corpus spécialisés [156]. Ils sont constitués de pages Web portant sur divers aspects du domaine de l'environnement, totalisant environ 50 millions de mots dans chaque langue. Ces documents proviennent de différentes sources : sites d'organismes gouvernementaux et non gouvernementaux, sites de vulgarisation scientifique, encyclopédies, journaux, blogues et répertoires de sites Web, entre autres [15, p. 240]. Ainsi, ces corpus sont caractérisés par un degré d'hétérogénéité plus élevé qu'un corpus spécialisé typique. À notre avis, cette caractéristique du corpus reflète la nature hétérogène du domaine de l'environnement lui-même, que nous avons soulignée au chapitre 1. Quoi qu'il en soit, la taille de ces deux corpus est telle qu'ils contiennent de nombreux contextes pour beaucoup de termes, ce qui les rend propices à l'utilisation de techniques distributionnelles. En outre, le fait que ces corpus soient distribués librement constitue un avantage important, les expériences que nous avons réalisées pouvant être reproduites plus facilement.

Dans les sections suivantes, nous décrirons les opérations de prétraitement que nous avons appliquées aux corpus avant de les utiliser pour construire des modèles distributionnels. Ces opérations, qui ont été réalisées au moyen d'un programme que nous avons écrit en Python, comprennent l'extraction du contenu textuel, la normalisation des caractères et la lemmatisation.

[1] `http://catalog.elra.info/product_info.php?products_id=1186&language=fr` (page consultée le 3 août 2015).

[2] `http://catalog.elra.info/product_info.php?products_id=1184&language=en` (page consultée le 3 août 2015).

5.2.1 Extraction du contenu textuel

Les documents constituant les corpus PANACEA sont en format XML, c'est-à-dire qu'ils sont structurés au moyen de balises, et contiennent beaucoup de métadonnées. Nous ne prenons pas en compte les balises ni les métadonnées lors de la construction des modèles distributionnels, seulement le contenu textuel des documents. Il est donc nécessaire de repérer les éléments XML qui contiennent le contenu textuel de chaque document et d'extraire ce contenu.

Le programme de prétraitement identifie, dans chaque document, chacun des éléments contenant un paragraphe. Ces éléments ont un attribut qui indique si le paragraphe est dans une autre langue, s'il est considéré comme trop court ou s'il s'agit d'un paragraphe passe-partout (*boilerplate*). Le programme extrait tous les paragraphes qui n'ont aucun de ces attributs, ainsi que le titre du document si celui-ci est présent.

5.2.2 Normalisation des caractères

L'étape suivante est la normalisation des caractères. Pour décrire cette opération, nous devons d'abord expliquer la notion de « codage de caractères ». Un codage de caractères associe des codes numériques à des caractères. Par exemple, dans le codage ASCII, la lettre *a* est représentée par le code 97. Le codage ASCII définit 128 caractères, qui comprennent des signes de ponctuation, les chiffres 0 à 9 et les lettres *a* à *z* en minuscules et en majuscules, mais aucune lettre accentuée. Le codage ISO-8859-1, aussi appelé *Latin-1*, définit 191 caractères, y compris presque toutes les lettres utilisées en français, les seules lettres manquantes étant *œ* et *Œ*.

Les corpus PANACEA sont encodés en UTF-8, un codage qui offre une couverture beaucoup plus complète que ASCII et Latin-1. Une des particularités d'UTF-8 est que certains caractères peuvent être encodés de plus d'une façon. Par exemple, le code (hexadécimal) 0301 correspond à un accent aigu qui est ajouté au caractère précédent ; ainsi, la suite de codes 0065 0301 (*e* suivi de l'accent aigu) représente le même caractère que le

code 00E9 (la lettre *é*). Cela peut poser un problème lors de la construction de modèles distributionnels : on ne voudrait pas que certaines occurrences d'une forme linguistique ne soient pas prises en compte parce qu'elles sont encodées différemment.

La stratégie de normalisation de caractères que nous avons adoptée consiste à éliminer tout code de caractère ne faisant pas partie des codages ASCII (en anglais) ou Latin-1 (en français), en remplaçant ces codes par un code correspondant au même caractère dans la mesure du possible. La méthode de normalisation est décrite en détail à l'annexe I.

5.2.3 Lemmatisation

La dernière opération est la lemmatisation, qui consiste à remplacer les différentes formes fléchies d'un mot par le lemme correspondant (p. ex *pommes* devient *pomme*). La lemmatisation fait partie des opérations de normalisation qui sont fréquemment appliquées à un corpus avant la construction d'un modèle distributionnel, qui comprennent aussi la mise en minuscules. Dans les deux cas, l'objectif est de remplacer les différentes formes correspondant au même mot par une forme unique ; de cette façon, toutes les formes correspondant à un même mot sont prises en compte lors de la construction de sa représentation (et une seule représentation est construite pour ce mot, plutôt qu'une représentation par forme). Le choix des opérations de normalisation que l'on applique au corpus dépend de plusieurs facteurs [182, p. 154–155], notamment de l'application envisagée. Dans le cadre de ce travail, il ne serait pas utile de détecter des affinités sémantiques entre les différentes formes fléchies d'un même mot ; d'ailleurs, les données de référence que nous utilisons à des fins d'évaluation (voir section 5.6) ne comprennent pas de formes fléchies. Pour ces raisons, entre autres, il nous semblait préférable de lemmatiser le corpus. En outre, Bullinaria et Levy [34] ont montré que la lemmatisation offre parfois une légère amélioration des résultats que l'on obtient au moyen de l'AD.

Pour lemmatiser le corpus, le programme fait appel à l'étiqueteur morphosyntaxique

TreeTagger [169] et à la bibliothèque TreeTaggerWrapper [3]. TreeTagger réalise à la fois la segmentation en mots, l'étiquetage morphosyntaxique et la lemmatisation, associant à chaque occurrence de mot sa catégorie grammaticale et sa forme lemmatisée. Le programme de prétraitement conserve la forme lemmatisée si TreeTagger est arrivé à l'identifier, sinon il conserve la forme non lemmatisée ; il conserve également la forme non lemmatisée dans les cas où TreeTagger est incertain et fournit deux lemmes possibles.

Tous les mots sont ensuite mis en minuscules, et si le document contient au moins 50 mots, le programme l'ajoute au corpus prétraité ; sinon, le document est exclu.

Pour illustrer cette séquence d'opérations de prétraitement, supposons qu'un des documents du corpus contienne l'élément XML suivant :

```
<p>Jean aime Marie.</p>
```

L'extraction du contenu textuel retourne le paragraphe suivant :

```
Jean aime Marie.
```

La normalisation des caractères n'engendre aucun changement dans ce cas. Puis, lorsque le programme appelle TreeTagger, celui-ci retourne les triplets suivants :

```
Jean    NAM      Jean
aime    VER:pres aimer
Marie   NAM      Marie
.       PUN      .
```

Enfin, la lemmatisation et la mise en minuscules produisent la chaîne suivante :

```
jean aimer marie .
```

[3] `https://pypi.python.org/pypi/treetaggerwrapper` (page consultée le 31 août 2015).

5.3 Sélection des mots-cibles

Les mots-cibles représentés par un modèle distributionnel peuvent comprendre tous les mots dans le corpus, mais pour différentes raisons, dans la pratique, on choisit souvent d'exclure certains mots des mots-cibles (ou d'exclure les représentations de certains mots-cibles lors de l'évaluation). Une raison est que les méthodes distributionnelles sont sensibles à la fréquence des mots[4]. Une autre raison est que le traitement des représentations d'un très grand nombre de mots peut exiger beaucoup de mémoire et de temps de calcul. Or, dans la pratique, on utilise souvent un nombre très élevé de mots-cibles, en excluant seulement les mots dont la fréquence dans le corpus est jugée trop faible.

En effet, un critère très courant pour la sélection des mots-cibles est la fréquence. Par exemple, on peut définir un nombre minimal d'occurrences et exclure les mots dont la fréquence est sous ce seuil, le principe sous-jacent étant que si on observe un nombre trop faible de contextes pour un mot donné, on ne dispose pas de suffisamment d'information sur la distribution de ce mot pour pouvoir estimer convenablement la similarité distributionnelle entre ce mot et les autres. Une autre façon d'exploiter la fréquence pour réaliser la sélection des mots-cibles consiste à prendre un certain nombre des mots les plus fréquents dans le corpus, ce qui revient au même. En théorie, il serait possible de remplacer la fréquence par un autre critère, tel que la spécificité ou le potentiel terminologique des mots, mais dans la pratique, on utilise généralement la simple fréquence.

Si on connaît d'avance des mots particuliers qui ne devraient pas faire partie des mots-cibles (p. ex. les mots vides, tels que les conjonctions et les prépositions), on peut utiliser une liste d'exclusion, aussi appelée un *antidictionnaire*. On peut aussi filtrer les mots-cibles en fonction de leur partie du discours ou des caractères qu'ils contiennent (p. ex. en excluant les mots qui contiennent des caractères non alphabétiques). Ces critères peuvent également être combinés.

Nous avons choisi d'utiliser comme mots-cibles les 10 000 formes les plus fréquentes

[4] Nous reviendrons sur ce point à la section 6.6.2.4.

dans le corpus lemmatisé, en excluant les formes suivantes :

- les mots vides ;
- les chaînes contenant un signe de ponctuation (à l'exception du trait d'union) ou tout caractère ne correspondant ni à une lettre ni à un chiffre[5] ;
- les formes commençant ou se terminant par un trait d'union[6].

En ce qui concerne les mots vides, nous avons utilisé les antidictionnaires exploités par le logiciel libre de racinisation (*stemming*) Snowball[7]. Ces listes contiennent 180 mots en anglais et 166 en français.

Il est important de noter que les mots-cibles ne comprennent pas de termes complexes. Il serait tout à fait possible d'inclure des termes complexes (voir, par exemple, le travail de Périnet et Hamon [152]), mais nous avons choisi de modéliser seulement les termes simples. Cette décision a plusieurs motivations. D'abord, la prise en compte de termes complexes nécessiterait l'ajout de modules supplémentaires à notre chaîne de traitement (un extracteur de termes pour la sélection des mots-cibles, un module de reconnaissance de termes pour le calcul des cooccurrences) ou l'utilisation de méthodes distributionnelles de composition sémantique [13, 30, 135, 192]. Par ailleurs, l'approche lexico-sémantique à la terminologie (que nous avons présentée à la section 2.1) ne décrit les termes complexes que si leur sens n'est pas compositionnel. Ainsi, dans les ressources résultant de cette démarche, telles que le DiCoEnviro, que nous utilisons à des fins d'évaluation, une proportion élevée des entrées sont des termes simples, donc ces

[5] Nous permettons la présence de chiffres pour que les symboles chimiques, tels que CO2, puissent être inclus parmi les mots-cibles.

[6] Cette règle sert à exclure des formes telles que *-ci* qui résultent de la façon dont TreeTagger réalise la segmentation en mots.

[7] Voir `http://snowballstem.org/algorithms/french/stop.txt` et `http://snowballstem.org/algorithms/english/stop.txt` (pages consultées le 14 juillet 2016). En anglais, nous avons exclu de l'antidictionnaire la liste de mots fréquents fournie à titre indicatif ; nous avons également exclu *us*, parce que cette forme est éventuellement un mot-cible utile (le sigle *US* mis en minuscules). En français, nous avons ajouté à l'antidictionnaire la forme infinitive des verbes *avoir* et *être*, parce que toutes leurs formes fléchies y apparaissaient déjà, et parce que *have* et *be* font partie de l'antidictionnaire en anglais.

ressources nous semblent tout à fait adaptées à l'évaluation de modèles distributionnels représentant seulement des termes simples. En somme, nous jugeons qu'il n'est pas nécessaire d'aborder le problème du traitement des termes complexes pour réaliser les objectifs de cette thèse.

Mot-cible	Fréquence
plus	237 393
pouvoir	172 119
tout	163 671
faire	133 404
eau	129 152
comme	95 678
autre	94 016
devoir	91 882
développement	85 438
environnement	85 006
...	...
tirage	157
sous-catégorie	157
conserve	157
rehausser	157
urdos	157
immobile	157
subdivision	157
jan	157
immigrant	157
chlordécone	157

(a) FR

Mot-cible	Fréquence
change	192 649
climate	192 601
use	131 032
forest	129 326
water	115 288
global	115 134
year	113 266
one	95 182
energy	92 427
also	91 655
...	...
sensagent	183
stanley	182
icelandic	182
timeframe	182
therapy	182
long-lasting	182
daly	182
nutrient-rich	182
anxiety	182
plough	182

(b) EN

Tableau 5.I – Mots-cibles les plus fréquents et les moins fréquents **(a)** en français et **(b)** en anglais.

Le Tableau 5.I présente les mots-cibles les plus fréquents et les moins fréquents obtenus de cette façon. Signalons que les mots-cibles comprennent beaucoup de mots qui présentent peu d'intérêt d'un point de vue terminologique, tels que *plus*, *tout* et *comme*. Si la présence de ces mots était problématique, ils pourraient être ajoutés à l'antidictionnaire, ou les critères de sélection des mots-cibles pourraient être ajustés de sorte que

ces mots ne seraient pas inclus. La présence de ces mots ne nous semble pas problématique *a priori*, tant qu'ils n'apparaissent pas dans le voisinage distributionnel des termes qui nous intéressent, ce que nous vérifierons dans l'analyse des résultats. Ainsi, nous avons choisi d'exploiter des antidictionnaires existants et d'utiliser la simple fréquence comme critère de sélection, plutôt qu'élaborer nos propres listes d'exclusion ou utiliser des critères plus complexes.

Quant au nombre de mots-cibles que nous avons retenus, il a été fixé de sorte à assurer un bon compromis entre la couverture des mots présents dans le corpus et des termes décrits dans les dictionnaires utilisés à des fins d'évaluation d'une part, et la mémoire et le temps de calcul exigés pour réaliser l'expérience d'autre part. En ce qui concerne la couverture des termes décrits dans les dictionnaires dont nous avons extrait les données de référence (voir section 5.6.1), soulignons à titre indicatif que les relations qui ont été extraites du DiCoEnviro, mais qui ont été exclues parce qu'elles contenaient au moins un terme ne faisant pas partie des mots-cibles, représentent environ 23% des relations extraites en français et 16% en anglais.

5.4 Méthodes utilisées pour construire les modèles

Dans cette section, nous décrivons les deux méthodes que nous utilisons pour construire des modèles distributionnels représentant les mots-cibles décrits à la section 5.3. Nous décrivons d'abord ces méthodes ainsi que certains de leurs (hyper)paramètres, puis les paramétrages que nous évaluons pour chaque méthode afin d'optimiser ces paramètres, ensuite les outils que nous utilisons pour la construction des modèles.

5.4.1 Description des méthodes

Les deux méthodes exploitent le même type de contexte, à savoir les cooccurrents graphiques. Si nous avons choisi d'utiliser les cooccurrents graphiques plutôt que syntaxiques, c'est parce qu'ils exigent moins de prétraitement en amont, ce qui facilite l'ap-

plication des méthodes distributionnelles à des corpus dans différentes langues. En effet, l'utilisation de cooccurrents syntaxiques exige une analyse syntaxique et une normalisation des dépendances qu'elle produit, tandis que le calcul des cooccurrents graphiques n'exige aucun prétraitement particulier, mis à part la segmentation en mots ; nous avons choisi de lemmatiser le corpus, mais cette étape est facultative.

La première méthode est l'AD. Comme nous l'avons expliqué au chapitre 3, l'AD consiste essentiellement à calculer et à pondérer la fréquence de cooccurrence des mots-cibles et de leurs cooccurrents. Les paramètres de l'AD, que nous avons décrits à la section 3.2.5, comprennent quelques paramètres liés à la fenêtre de contexte (taille, direction, forme) ainsi que la pondération. Il serait possible de prendre en compte d'autres paramètres, tels que le choix d'une technique de réduction de dimension, mais pour limiter le temps de calcul et la complexité des analyses, nous avons choisi de retenir seulement ces paramètres, qui nous semblent fondamentaux.

La deuxième méthode est le modèle de langue neuronal implémenté dans l'outil `word2vec` [130, 132]. En ce qui concerne les hyperparamètres de `word2vec`, certains sont également des paramètres de l'AD, tandis que d'autres sont propres à cette méthode. La documentation de `word2vec`[8] énumère cinq hyperparamètres qui ont une influence importante sur les résultats. Nous analysons donc l'influence de chacun de ces hyperparamètres[9] comme nous le faisons pour l'AD. Nous fournissons ci-dessous une brève description de ces cinq hyperparamètres.

Le premier hyperparamètre est l'architecture du modèle. Deux architectures neuronales ont été implémentées dans `word2vec` : *continuous bag-of-words* (CBOW) et *continuous skip-gram*. Pour simplifier, on peut expliquer la différence entre ces deux architectures de la façon suivante : CBOW vise à prédire chaque mot en fonction de son contexte (ses cooccurrents), tandis que skip-gram vise à prédire le contexte en fonction

[8] Voir `https://code.google.com/p/word2vec/#Performance` (page consultée le 27 juillet 2015).

[9] Pour les autres hyperparamètres de `word2vec`, nous utilisons les valeurs par défaut.

du mot[10].

Le deuxième est l'algorithme d'entraînement : quelle que soit l'architecture, le modèle peut être entraîné au moyen d'un *softmax hiérarchique* ou par échantillonnage d'exemples négatifs[11].

Le troisième est un seuil de fréquence utilisé par une fonction de sous-échantillonnage de mots fréquents. Cette fonction supprime du corpus des occurrences de mots dont la fréquence relative est supérieure au seuil, la probabilité qu'une occurrence soit supprimée étant proportionnelle à la fréquence du mot.

Les deux derniers hyperparamètres sont la taille de la fenêtre de contexte et la dimension des représentations lexicales.

Soulignons que la fenêtre de contexte implémentée dans `word2vec` a une forme particulière[12], semblable à celle d'une fenêtre triangulaire ; ainsi, les cooccurrents les plus près du mot-cible contribuent le plus à sa représentation.

Comme nous l'avons mentionné ci-dessus, certains (hyper)paramètres s'appliquent aux deux méthodes utilisées pour construire des modèles distributionnels ; parmi les (hyper)paramètres que nous avons pris en compte, c'est seulement le cas de la taille de la fenêtre de contexte. En revanche, les architectures et les algorithmes d'entraînement implémentés dans `word2vec` sont propres à cette méthode. En ce qui concerne le sous-échantillonnage de mots fréquents, cette technique pourrait être appliquée lors du calcul de l'AD aussi [110]. Donc, il aurait été possible d'observer l'influence de cette technique non seulement sur la qualité des modèles produits par `word2vec`, mais aussi sur celle des modèles produits par AD (ainsi que l'influence de la forme de la fenêtre dans le cas de `word2vec`, par exemple). Nous avons plutôt choisi de nous pencher sur un ensemble restreint d'(hyper)paramètres typiques pour chacune des deux méthodes ; par exemple, le sous-échantillonnage n'est généralement pas appliqué lors du calcul de l'AD, et le programme `word2vec` ne permet pas de choisir la forme de la fenêtre, bien que ce soit

[10] Pour plus d'information sur ces deux architectures, voir le travail de Mikolov et al. [130].
[11] Pour plus d'information sur ces algorithmes d'entraînement, voir le travail de Mikolov et al. [132].
[12] Voir le travail de Levy et al. [110] pour plus d'information à ce sujet.

possible de modifier le programme pour permettre ce choix, puisque le code source est libre. En outre, dans le cas des (hyper)paramètres qui s'appliquent aux deux méthodes, il nous semble probable que leur influence serait semblable dans les deux cas, comme le suggèrent les résultats que nous avons obtenus en ce qui concerne la taille de fenêtre (voir chapitre 6).

5.4.2 Paramétrages évalués

Dans le cas de l'AD, les paramètres que nous optimisons sont la taille, la direction et la forme de la fenêtre de contexte, ainsi que la pondération appliquée aux fréquences de cooccurrence. Les valeurs testées pour chacun de ces paramètres sont :

- Taille de la fenêtre : 1 à 10 mots.
- Forme de la fenêtre : rectangulaire ou triangulaire.
- Direction de la fenêtre : gauche+droite ou gauche&droite.
- Pondération[13] : Aucune, log, MI, MI^2, MI^3, local-MI, z-score, t-score et simple-LL.

Nous évaluons toutes les combinaisons possibles de ces valeurs, ce qui nous donne $10 \times 2 \times 2 \times 9 = 360$ paramétrages à évaluer.

Quant à `word2vec`, le choix des valeurs à tester pour chacun de ses hyperparamètres a été déterminé en utilisant :

- soit toutes les valeurs possibles (dans le cas de l'architecture et de l'algorithme d'entraînement) ;
- soit des valeurs recommandées dans la documentation de `word2vec` (dans le cas du seuil pour le sous-échantillonnage) ;

[13] Toutes les formules sont fournies dans l'annexe II.

- soit des valeurs utilisées dans d'autres travaux exploitant `word2vec` (dans le cas de la dimension des représentations et du nombre d'exemples négatifs).

Les valeurs testées pour chacun des hyperparamètres sont :

- Dimension des représentations lexicales : 100 ou 300.
- Taille de la fenêtre : 1 à 10 mots.
- Architecture : skip-gram ou CBOW.
- Nombre d'exemples négatifs pour l'entraînement du modèle : 0 (on utilise alors un algorithme d'entraînement différent, le *softmax* hiérarchique), 5 ou 10.
- Seuil pour le sous-échantillonnage de mots fréquents : seuil élevé (10^{-3}), seuil faible (10^{-5}) ou aucun sous-échantillonnage (0).

Nous évaluons donc $2 \times 10 \times 2 \times 3 \times 3 = 360$ paramétrages différents.

5.4.3 Outils utilisés

L'AD est calculée au moyen d'un programme que nous avons écrit en Python. Quelques détails concernant notre implémentation de cette méthode méritent d'être soulignés. Premièrement, lors du calcul de la matrice de cooccurrence, nous permettons à la fenêtre de contexte de chevaucher les frontières de phrases ; ainsi, un mot peut être considéré comme un cooccurrent d'un autre mot même s'il se trouve dans la phrase précédente ou suivante. Deuxièmement, nous avons choisi d'utiliser le même ensemble de mots pour les mots-cibles et les mots-contextes, plutôt que définir séparément ces deux ensembles ; ainsi, chaque mot-cible est aussi un mot-contexte, et à chacun de ces mots correspond à la fois une ligne et une colonne dans la matrice de cooccurrence, comme dans les Figures 3.1 et 3.2. Enfin, les mots hors-vocabulaire (les mots dans le corpus qui ne font pas partie des mots-cibles ou des mots-contextes) ne sont pas supprimés du corpus ; ils ne sont simplement pas comptés.

En ce qui concerne le modèle de langue neuronal, le programme de construction et d'évaluation de modèles fait appel au programme `word2vec`[14]. Il est important de noter que, bien que ce programme apprenne par défaut des représentations pour tous les mots dans le corpus ayant une fréquence supérieure ou égale à un certain seuil[15] (5 occurrences par défaut), nous ne conservons que les représentations des 10 000 mots-cibles.

5.5 Graphes

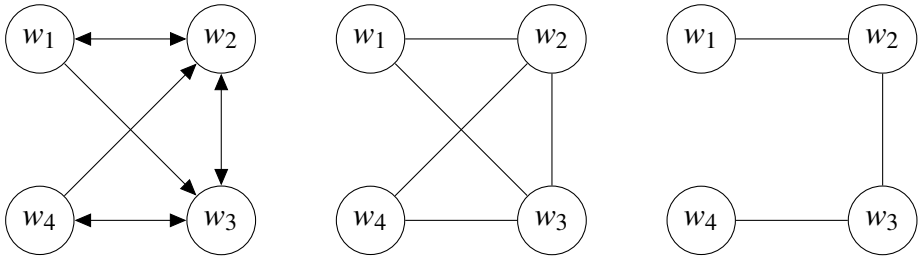
Comme nous l'avons expliqué à la section 5.1, nous utilisons des graphes de k PPV afin d'interroger les modèles distributionnels, en plus d'évaluer les listes ordonnées de PPV que l'on obtient pour chaque mot-cible. Les graphes de k PPV, qui contiennent un nœud pour chaque mot-cible, sont simples à construire : il suffit de créer des connexions (appelées *arêtes* ou *arcs*) entre chaque mot-cible et ses k PPV. L'ensemble des arêtes que contient le graphe dépend de la valeur de k, mais aussi du type de graphe que l'on utilise.

5.5.1 Types de graphes

En effet, il existe plusieurs types de graphes de k PPV. Premièrement, le graphe peut être un graphe *orienté* ou un graphe *non orienté*. Dans la théorie des graphes, un graphe orienté est un graphe dans lequel les arêtes (appelées *arcs* dans ce cas) ont un sens ; ainsi, deux nœuds x et y peuvent être reliés par un arc (x,y) allant de x à y ou par un arc (y,x) allant dans le sens contraire ; ils peuvent aussi être reliés par les deux arcs. En revanche, dans un graphe non orienté, les arêtes n'ont pas de direction. Ainsi, on représente les arcs dans un graphe orienté au moyen de flèches, et les arêtes dans un graphe non orienté au moyen de traits, comme l'illustrent la Figure 5.1a d'une part et les Figures 5.1b et 5.1c d'autre part.

[14] Voir `https://code.google.com/p/word2vec/` (page consultée le 21 janvier 2016).

[15] Les mots dont la fréquence est sous ce seuil sont supprimés du corpus avant l'entraînement du modèle, ce qui a pour effet d'augmenter, dans les contextes où des mots ont été supprimés, la taille effective de la fenêtre de contexte [110].



(a) Orienté (b) Symétrique (c) Mutuel

Figure 5.1 – Graphes de k PPV orienté, symétrique et mutuel ($k = 2$).

Dans le cas particulier des graphes de k PPV, un graphe orienté contient un arc reliant un mot w_i à un autre mot w_j si w_j est parmi les k PPV de w_i. Les Figures 3.6 et 3.7, présentées au chapitre 3, illustrent des graphes de k PPV orientés.

En ce qui concerne les graphes de k PPV non orientés, on peut en distinguer (au moins) deux sous-types : le graphe de k PPV symétrique et le graphe de k PPV mutuel [123]. Dans un graphe de k PPV symétrique, deux mots w_i et w_j sont reliés par une arête si w_j est parmi les k PPV de w_i ou si w_i est parmi les k PPV de w_j. Dans un graphe de k PPV mutuel, w_i et w_j sont reliés seulement si ces deux conditions sont vraies.

Les graphes illustrés dans la Figure 5.1 sont en fait des graphes de k PPV (avec $k = 2$) construits à partir des 4 représentations lexicales artificielles utilisées comme exemples au chapitre 3 et représentées visuellement dans la Figure 3.4. Le graphe dans la Figure 5.1a est un graphe orienté ; il montre que les deux PPV du mot w_1 sont w_2 et w_3, les deux PPV de w_2 sont w_1 et w_3, et ainsi de suite. Le graphe dans la Figure 5.1b est un graphe de k PPV symétrique ; dans ce graphe, l'arête entre les nœuds w_1 et w_2 indique soit que w_2 est un des deux PPV de w_1, soit que w_1 est un des deux PPV de w_2, soit que ces deux propositions sont vraies (ce qui est effectivement le cas). Enfin, le graphe dans la Figure 5.1c est un graphe de k PPV mutuel ; dans ce graphe, l'arête entre w_1 et w_2 indique à la fois que w_2 est un des deux PPV de w_1 et que w_1 est un des deux PPV de w_2.

L'utilisation d'un graphe mutuel plutôt que symétrique reflète le principe suivant : si

deux mots sont sémantiquement reliés, on s'attendrait non seulement à ce qu'un des deux mots soit parmi les PPV de l'autre, mais que chacun des deux mots soit parmi les PPV de l'autre. Ce principe de réciprocité des liens de similarité distributionnelle a été exploité dans quelques travaux [41, 67, 119]. Par ailleurs, d'un point de vue linguistique, utiliser un graphe non orienté plutôt qu'un graphe orienté revient à supposer que si un mot x est sémantiquement relié à un mot y, alors l'inverse (que y est relié à x) doit aussi être vrai. Cette hypothèse nous semble raisonnable, bien que la nature de la relation puisse être différente dans un sens que dans l'autre, notamment dans le cas de l'hyperonymie. Quoi qu'il en soit, nous vérifions dans ce travail lequel de ces trois types de graphes est le plus à même de faciliter l'identification de relations lexicales. Le choix du graphe résulte donc d'une démarche empirique plutôt que de justifications théoriques.

Une des différences entre les graphes de k PPV orientés et non orientés concerne la définition de la notion de « voisinage » ainsi que le nombre de voisins par nœud. Dans les graphes non orientés, le voisinage d'un nœud w contient tous les nœuds reliés à w par une arête (appelés les nœuds *adjacents* à w). Dans le graphe orienté, le voisinage d'un nœud w désigne (dans le cadre de ce travail) les nœuds reliés à w par un arc partant de w, appelés les *successeurs* de w. En ce qui concerne le nombre de voisins par nœud, dans le graphe orienté, tous les nœuds ont le même nombre de voisins, chaque nœud ayant exactement k voisins (successeurs). En revanche, dans les graphes non orientés, le nombre de voisins (nœuds adjacents) varie d'un nœud à l'autre. Par exemple, dans la Figure 5.1, les nœuds dans le graphe orienté ont tous deux voisins (successeurs). En revanche, dans le graphe symétrique, les mots w_1 et w_4 ont deux voisins (nœuds adjacents) et les mots w_2 et w_3 en ont trois ; dans le graphe mutuel, w_1 et w_4 ont un seul voisin et w_2 et w_3 en ont deux. La taille du voisinage d'un mot est donc variable dans le cas des graphes de k PPV non orientés, puisqu'elle dépend non seulement des PPV du mot, mais aussi de ceux des autres mots, tandis qu'elle vaut toujours k dans le cas du graphe orienté. Cette propriété des graphes de k PPV symétriques et mutuels permet éventuellement de rendre compte du fait que les termes ne participent pas tous au même

nombre de relations lexicales.

5.5.2 Paramétrages évalués

En ce qui concerne le graphe de k PPV, nous optimisons deux paramètres : k et le type de graphe. Nous testons les valeurs suivantes :

- k : 1, 2, 4, 8, 16, 32, 64 ou 128.
- Type de graphe : orienté, symétrique ou mutuel.

Ainsi, pour chaque modèle distributionnel, nous construisons et évaluons $8 \times 3 = 24$ graphes de voisinage différents.

5.5.3 Outils utilisés

La mesure de similarité utilisée pour le calcul des PPV est le cosinus de l'angle des vecteurs, mesure que nous avons expliquée au chapitre 3. Pour calculer cette mesure, le programme de construction et d'évaluation de modèles fait appel à la bibliothèque Scikit-Learn [145]. Une fois la matrice de similarité calculée pour un modèle particulier, le programme construit et évalue les différents graphes de k PPV. Il fait appel à la bibliothèque NetworkX [87] pour créer la structure de données correspondant au graphe et pour identifier les voisins (nœuds adjacents ou successeurs) des termes lors de l'évaluation ; nous avons également utilisé cette bibliothèque afin d'extraire les sous-graphes illustrés au chapitre 7.

5.6 Évaluation

La principale méthode d'évaluation que nous utilisons dans cette thèse est une évaluation quantitative des voisinages distributionnels réalisée en comparant les listes ordonnées de PPV que l'on obtient à partir des modèles distributionnels, ainsi que les

graphes de *k* PPV, à des données de référence. Ces données ont été extraites de dictionnaires spécialisés que nous avons décrits au chapitre 2. Nous utilisons une même méthode d'évaluation pour tester nos deux hypothèses, à savoir l'évaluation des listes ordonnées de PPV au moyen de la MAP (voir section 5.6.2.2), mais en utilisant deux jeux différents de données de référence, chacun reflétant un des cadres descriptifs que nous avons adoptés. Nous évaluons par la suite les graphes de *k* PPV, en nous concentrant dans ce cas sur le cadre descriptif de la sémantique des cadres.

Comme nous l'avons souligné à la section 5.1, cette évaluation a plusieurs objectifs. Premièrement, elle sert à tester nos deux hypothèses. Si notre première hypothèse est vraie, alors les PPV des termes du domaine de l'environnement devraient contenir des termes reliés paradigmatiquement à ceux-ci selon le DiCoEnviro. De même, si notre seconde hypothèse est vraie, alors les termes qui évoquent le même cadre sémantique devraient être proches selon le modèle distributionnel.

Deuxièmement, en appliquant cette méthode d'évaluation à différents paramétrages des modèles et du graphe, nous pouvons analyser l'influence des (hyper)paramètres et déterminer leurs valeurs optimales. Cette tâche serait impossible à réaliser en évaluant les modèles ou les graphes manuellement, puisque le nombre de paramétrages que l'on doit évaluer pour pouvoir analyser adéquatement l'influence des (hyper)paramètres peut être très élevé. Par exemple, dans le cadre de ce travail, nous construisons 360 modèles par AD et 360 modèles au moyen de `word2vec`, et 24 graphes de voisinage sont construits pour chaque modèle, ce qui donne un total de 720 modèles et 17 280 graphes à évaluer dans chaque langue. Il faut d'ailleurs que la méthode d'évaluation puisse être appliquée de manière identique à chaque modèle et à chaque graphe pour que les résultats soient comparables.

Dans cette section, nous décrivons les données de référence et les méthodes utilisées pour les extraire, puis nous expliquons les mesures utilisées pour évaluer les graphes de voisinage et les modèles distributionnels sous-jacents.

5.6.1 Données de référence

Les données de référence ont été extraites au moyen de programmes dont nous décrivons le fonctionnement aux sections 5.6.1.1 et 5.6.1.2. Ceux-ci prennent en entrée les dictionnaires dont ils extraient les données de référence ainsi que la liste des mots-cibles. La chaîne de traitement que nous avons utilisée pour l'extraction des données de référence et la sélection des mots-cibles est illustrée dans la Figure 5.2.

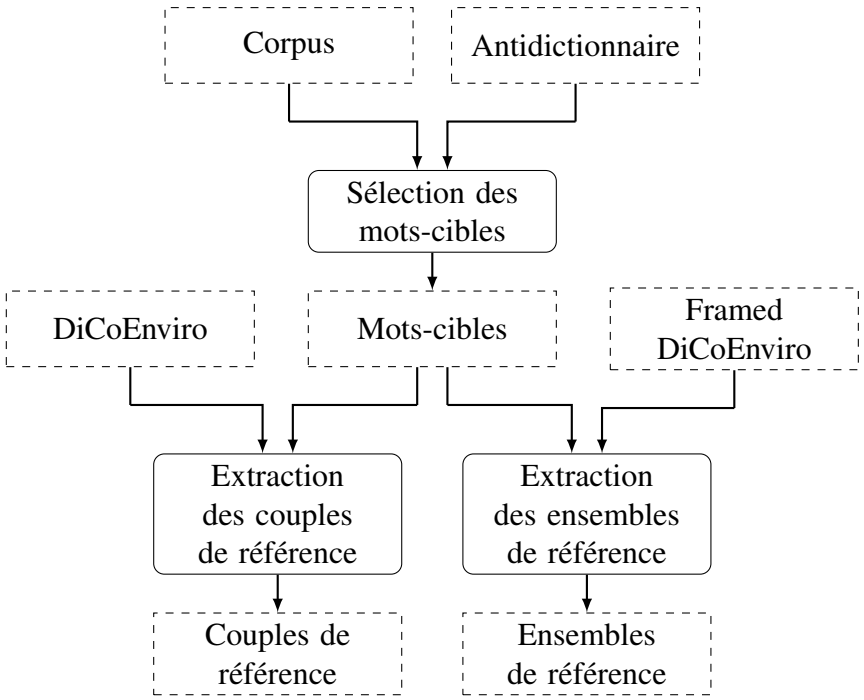


Figure 5.2 – Chaîne de traitement pour la sélection des mots-cibles et l'extraction des données de référence.

Le premier jeu de données est constitué de paires de termes reliés par une relation lexicale paradigmatique ; il sert à évaluer notre première hypothèse. Le deuxième est constitué d'ensembles de termes évoquant le même cadre sémantique et sert à évaluer notre seconde hypothèse. Nous appellerons ces deux jeux de données les *couples de référence* et les *ensembles de référence*.

5.6.1.1 Couples de référence

Les données de référence utilisées pour tester notre première hypothèse sont des couples de termes reliés extraits du DiCoEnviro. Ces couples sont constitués d'une entrée du DiCoEnviro et d'un terme relié décrit dans l'article consacré à l'entrée.

Les termes faisant partie des couples de référence sont tous des noms (NN), des verbes (VV) ou des adjectifs (JJ) ; les adverbes ont été exclus parce qu'ils sont trop peu nombreux dans le dictionnaire. En ce qui concerne l'inclusion des verbes, soulignons que l'approche lexico-sémantique à la terminologie accorde une place importante au verbe, contrairement à d'autres approches :

> l'analyse du verbe peut mener à la découverte de caractéristiques conceptuelles ou lexicales propres aux domaines spécialisés. Le verbe joue un rôle important dans l'expression de faits spécialisés et, s'il est moins présent dans les réseaux conceptuels, il est au cœur de réseaux lexicaux. [113, p. 105–106]

De plus, les unités lexicales qui évoquent des cadres sémantiques comprennent souvent des verbes, donc il nous semblait essentiel d'inclure des verbes dans les couples de référence, de sorte que les résultats obtenus sur les verbes puissent informer ceux obtenus sur les ensembles de référence.

Lors de l'extraction des couples de référence, nous avons ciblé un certain nombre de relations paradigmatiques, que nous regroupons en quatre catégories. La première, que nous appelons les *quasi-synonymes* (QSYN) comprend les relations lexicales suivantes[16] :

- synonyme (p. ex. *diesel→gazole*, *défrichement→défrichage*, *méthane→ch4*) ;
- quasi-synonyme (p. ex. *diminuer→réduire*, *batterie→accumulateur*) ;

[16] Tous les exemples présentés dans cette section proviennent du DiCoEnviro (données récupérées le 19 août 2015).

- sens voisin (p. ex. *estimer→calculer*, *gaspillage→surconsommation*) ;
- variante (p. ex. *autopartage→auto-partage*, *modelling→modeling*).

La deuxième catégorie, que nous appelons les *liens hiérarchiques* (HYP), comprend les relations suivantes :

- hyponyme (p. ex. *énergie→bioénergie*, *arbre→bouleau*) ;
- hyperonyme (p. ex. *azote→gaz*, *diesel→carburant*).

La troisième catégorie, que nous appelons les *antonymes* (ANTI), comprend tous les antonymes ou contraires (p. ex. *diminuer→augmenter*, *chaud→froid*, *faune→flore*).

La quatrième catégorie, que nous appelons les *dérivés syntaxiques*[17] (DRV), est constituée de paires de termes ayant le même sens, mais appartenant à des parties du discours différentes :

- nom dérivé[18] (p. ex. *absorber→absorption*, *urbain→ville*) ;
- verbe dérivé (p. ex. *absorption→absorber*, *combustion→brûler*) ;
- adjectif dérivé (p. ex. *agriculture→agricole*, *vent→éolien*).

Il serait possible de réaliser une analyse plus fine en considérant séparément chaque relation, mais cela pourrait être problématique en raison du nombre très variable d'exemples correspondant à chacune de ces relations dans le DiCoEnviro ; par exemple, les sens voisins et quasi-synonymes sont beaucoup plus nombreux que les synonymes et les variantes.

[17] Rappelons que ces relations ne font pas partie de ce que nous appelons les relations paradigmatiques *classiques* (voir note à la section 2.1.2). Par ailleurs, les dérivés syntaxiques ne sont pas toujours des dérivés morphologiques, mais la grande majorité des dérivés syntaxiques que nous avons extraits du DiCoEnviro sont morphologiquement apparentés.

[18] Nous empruntons à Mel'čuk et al. [128, p. 51] l'appellation *nom dérivé* ; on pourrait également utiliser le terme *nominalisation* [129, p. 133]. De même, nous utilisons les termes *verbe dérivé* et *adjectif dérivé* pour désigner les autres dérivés syntaxiques. Dans la lexicologie explicative et combinatoire, ces trois relations sont encodées au moyen des fonctions lexicales S_0, V_0 et A_0 respectivement.

L'extraction des couples de référence a été réalisée au moyen d'un programme que nous avons écrit en Python, dont le fonctionnement est expliqué dans l'annexe III.

Les couples extraits du dictionnaire ont été filtrés afin de ne retenir que ceux dont les deux termes font partie des mots-cibles, puisqu'un couple ne peut servir à évaluer un modèle que si le modèle contient une représentation pour les deux termes. En effet, si l'entrée ne fait pas partie des mots-cibles, on ne peut pas calculer ses PPV, et si le terme relié ne fait pas partie des mots-cibles, il est certain qu'il ne fera pas partie des PPV de l'entrée ; dans les deux cas, le couple n'est pas utile pour l'évaluation du modèle.

Puisque les mots-cibles ne comprennent pas de termes complexes, les couples de référence sont tous constitués de deux termes simples. Comme nous l'avons souligné à la section 5.3, nous ne nous intéressons qu'aux termes simples dans le cadre de ce travail.

Les termes simples extraits du dictionnaire mais ne faisant pas partie des mots-cibles comprennent non seulement des termes qui sont effectivement absents du corpus (p. ex. *herbicyclage*, *éco-cammionage*, *auto-décharger*, *e-scrap*, *landfilling*, *depollute*) ou très rares (p. ex. *électromobilité*, *biodégrader*, *éco-centre*, *depollution*, *e-bike*, *electromobility*), mais aussi des termes dont l'absence ou la faible fréquence est causée par la façon dont TreeTagger opère la lemmatisation. Par exemple, TreeTagger interprète systématiquement l'adjectif *menacé* comme un verbe au participe passé, et retourne la forme infinitive du verbe *menacer* comme forme lemmatisée ; ainsi l'adjectif a-t-il trop peu d'occurrences dans le corpus lemmatisé pour être inclus dans les mots-cibles. Signalons finalement que certains termes extraits du dictionnaire ne font pas partie des mots-cibles parce qu'ils sont dans l'antidictionnaire (voir section 5.3). En effet, les symboles chimiques *N* (azote), *C* (carbone) et *S* (soufre), une fois mis en minuscules, correspondent à trois formes que l'on retrouve dans l'antidictionnaire en français (les formes élidées de *ne*, *ce* et *se*). À titre indicatif, soulignons que la proportion des termes simples qui ont été extraits du dictionnaire, mais qui ne faisaient pas partie des mots-cibles, s'élevait à environ 25% en français et 15% en anglais.

La liste des couples de référence contient, pour chaque couple, l'entrée, le terme

relié, la PDD des deux termes et leur relation. Ces trois derniers renseignements permettent au programme de construction et d'évaluation de modèles de calculer les mesures d'évaluation (voir section 5.6.2) non seulement sur tous les couples, mais aussi sur les sous-ensembles de couples correspondant à chaque PDD et à chaque catégorie de relations.

En effet, nous utilisons 4 sous-ensembles de couples correspondant à chacune des catégories de relations (QSYN, ANTI, HYP et DRV) et 3 sous-ensembles de couples appartenant à chacune des trois PDD que nous avons prises en compte (NN, VV et JJ), ces 3 sous-ensembles contenant seulement les couples constitués de deux termes appartenant à la même PDD (autrement dit, ils ne contiennent pas de dérivés). La raison pour laquelle nous utilisons ces différents sous-ensembles de couples pour l'évaluation est que cela nous permet d'aborder, dans l'analyse des résultats, des questions telles que :

- Est-ce que les modèles distributionnels captent mieux certaines relations paradigmatiques que d'autres ?
- Captent-ils mieux les relations entre les termes appartenant à certaines PDD qu'à d'autres ?
- Captent-ils mieux les relations entre termes de la même PDD ou entre termes de PDD différentes ?
- Les paramétrages optimaux des modèles distributionnels dépendent-ils du type de relations que l'on souhaite identifier ?
- Dépendent-ils de la PDD des termes entre lesquels on cherche à identifier des relations ?

Le programme que nous avons développé nous a permis d'obtenir 1314 couples de référence en français et 888 en anglais[19]. Ces jeux de données ont donc une taille plus

[19]Les couples de référence ont été extraits le 19 août 2015. Ils sont disponibles à l'adresse `https://github.com/gbcolborne/exp_phd`.

Sous-ensemble	Nb couples	Nb entrées	Nb termes différents	Exemple
NN	700	274	345	*terre* → {*atmosphère, eau, écosystème, globe, mer, monde, planète, soleil* }
VV	225	98	121	*éliminer* → {*enfouir, jeter, récupérer, recycler*}
JJ	130	71	88	*lacustre* → {*fluvial, marin, terrestre*}
QSYN	689	357	456	*écologique* → {*écolo, propre, vert, environnemental, biotique*}
ANTI	173	126	154	*absorber* → {*réfléchir, émettre, libérer*}
HYP	193	100	126	*biocarburant* → {*carburant, biogaz, biodiesel, bioéthanol, éthanol*}
DRV	259	259	307	*absorber* → {*absorption*}
Tous	1314	490	624	*reboisement* → {*boisement, déboisement, plantation, reboiser, reforestation, régénération*}

Tableau 5.II – Caractéristiques des sous-ensembles de couples de référence (FR).

importante que plusieurs jeux de données couramment utilisés pour l'évaluation de modèles distributionnels, tels que le test de synonymie TOEFL [104], qui comprend 80 questions à choix multiple, ou le jeu WordSim-353 [74], qui contient 353 paires de mots. Leur taille est comparable à celle du jeu SimLex-999 [96], qui comprend 999 paires de mots, mais beaucoup plus petite que celle du jeu BLESS [12], qui comprend plus de 14 000 relations pour 200 noms (ainsi qu'un nombre comparable de relations factices).

Les Tableaux 5.II et 5.III présentent, pour le français et l'anglais respectivement, des caractéristiques des sous-ensembles de couples correspondant à chaque PDD et à chaque relation ; ils présentent le nombre de couples dans chaque sous-ensemble, ainsi que le nombre d'entrées différentes et le nombre de formes différentes (entrées ou termes reliés). Dans le cas des dérivés, toutes les entrées ont exactement un dérivé en français ; en anglais, il existe une entrée qui a deux dérivés : *adaptive*→{*adaptation, adapt*}. En ce qui concerne la répartition des couples en fonction de la PDD, soulignons qu'elle est

Sous-ensemble	Nb couples	Nb entrées	Nb termes différents	Exemple
NN	404	190	272	*drought* → {*desertification, event, flood*}
VV	187	84	95	*reflect* → {*absorb, emit, radiate, release, trap*}
JJ	122	67	89	*extinct* → {*endangered, threatened, vulnerable*}
QSYN	517	282	381	*hazardous* → {*dangerous, harmful, toxic*}
ANTI	109	77	97	*dispose* → {*recover, recycle, reuse*}
HYP	87	61	79	*fuel* → {*biodiesel, biofuel, diesel, gasoline*}
DRV	175	174	225	*adaptive* → {*adapt, adaptation*}
Tous	888	381	523	*deforestation* → {*afforestation, clearing, deforest, desertification, reforestation*}

Tableau 5.III – Caractéristiques des sous-ensembles de couples de référence (EN).

semblable à celle que l'on observe dans le jeu SimLex-999, qui a été conçu de sorte que les verbes sont deux fois plus nombreux que les adjectifs, et les noms, trois fois plus nombreux que les verbes [96].

Le Tableau 5.IV présente, pour les sous-ensembles de couples appartenant à chaque PDD[20], la répartition des couples en fonction de la relation. Pour les lecteurs qui s'étonneraient de la présence de deux liens hiérarchiques parmi les couples verbaux[21] (en français), soulignons qu'il s'agit de :

1. *consommer→gaspiller* (« consommer de manière exagérée ») ;

2. *exploiter→surexploiter* (« exploiter de manière trop importante »).

[20]Rappelons que ces sous-ensembles ne contiennent pas de dérivés, car ceux-ci appartiennent à des PDD différentes.

[21]Cruse [44, ch. 6] souligne que les verbes participent moins souvent à des relations hiérarchiques que les noms. Il propose d'ailleurs un patron lexico-syntaxique (*diagnostic frame*) pour identifier les relations d'hyponymie (plus précisément, de taxonymie) entre verbes : *X-ing is a way of Y-ing*.

	QSYN	ANTI	HYP
NN	442	67	191
VV	166	57	2
JJ	81	49	0

(a) FR

	QSYN	ANTI	HYP
NN	293	24	87
VV	138	49	0
JJ	86	36	0

(b) EN

Tableau 5.IV – Répartition des couples en fonction de la relation pour les sous-ensembles de couples appartenant à chaque PDD.

5.6.1.2 Ensembles de référence

Les données de référence utilisées pour tester notre seconde hypothèse sont des ensembles de termes qui évoquent le même cadre sémantique, que nous avons extraits du Framed DiCoEnviro. Par exemple, le cadre Adding_trees_in_location est évoqué par les UL *boiser*, *boisement*, *reboiser* et *reboisement* en français, et par les UL *afforest*, *afforestation*, *reforest* et *reforestation* en anglais[22].

La méthode utilisée pour extraire les ensembles de référence est plus simple que celle utilisée pour extraire les couples de référence, notamment parce que nous n'avons pas jugé nécessaire de filtrer les termes en fonction des relations auxquelles ils participent ni de leur PDD. Le programme qui extrait les ensembles de référence fonctionne de la façon suivante :

- Il extrait de chaque cadre sémantique les UL qui évoquent ce cadre en français et en anglais.
- Il supprime de ces deux listes les UL qui ne font pas partie des mots-cibles dans la langue correspondante.
- Dans chacune des deux langues, s'il reste au moins deux UL dans la liste, il l'ajoute aux ensembles de référence pour cette langue.

[22]Tous les exemples présentés dans cette section proviennent du Framed DiCoEnviro (données récupérées le 19 août 2015).

Nous avons ainsi obtenu 69 ensembles de référence en français et 57 en anglais[23]. Ces ensembles sont énumérés dans les annexes IV et V pour le français et l'anglais respectivement.

Les ensembles de référence contiennent entre 2 et 10 UL. La Figure 5.3 présente la répartition du nombre d'UL par ensemble. La majorité des ensembles (41 des 69 ensembles en français, 37 des 57 ensembles en anglais) contiennent 2 ou 3 UL, mais beaucoup d'ensembles contiennent 4 UL ou plus.

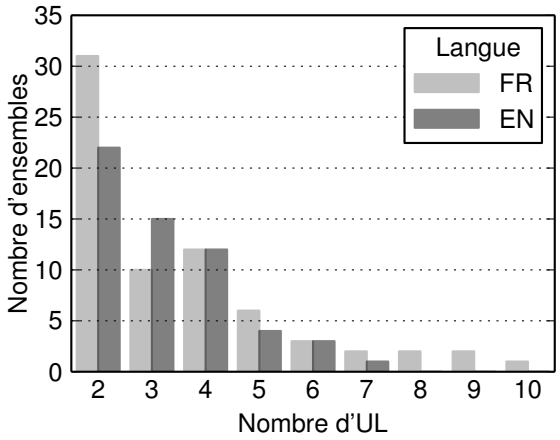


Figure 5.3 – Répartition du nombre d'unités lexicales par ensemble de référence.

En français, les 69 ensembles contiennent 246 UL ; si on compte seulement une fois les formes qui évoquent plus d'un cadre, on obtient 223 formes différentes. En effet, une même forme linguistique peut évoquer plus d'un cadre sémantique ; par exemple, la forme *réchauffement* peut évoquer le cadre Change_of_temperature ou le cadre Cause_temperature_change. En anglais, les 57 ensembles contiennent 182 UL, dont 168 formes différentes.

Le Tableau 5.V montre la répartition des UL en fonction de la PDD[24]. On observe ici que les ensembles de référence sont constitués principalement de noms et de verbes.

[23]Les ensembles de référence ont été extraits le 19 août 2015.

[24]Dans certains cas, deux UL correspondant à la même forme mais appartenant à des PDD différentes (p. ex. *retreat*) évoquent le même cadre sémantique. Dans ces cas, puisque la forme n'apparait qu'une fois dans l'ensemble de référence, nous comptons seulement une des PDD, celle-ci étant choisie au hasard.

PDD	Nb UL
NN	122
VV	111
JJ	13

(a) FR

PDD	Nb UL
NN	79
VV	87
JJ	16

(b) EN

Tableau 5.V – Répartition des UL en fonction de la PDD.

Le Tableau 5.VI illustre les relations auxquelles participent les termes dans les ensembles de référence. Ces données ont été obtenues en vérifiant, pour chaque paire différente de termes faisant partie du même ensemble de référence, si cette paire fait partie des couples de référence et, en l'occurrence, à quelle relation elle correspond. Nous comptons ainsi, parmi les paires de termes évoquant le même cadre sémantique, le nombre de paires appartenant à chacune des quatre catégories de relations lexicales. Ce tableau montre que les ensembles de référence contiennent beaucoup de quasi-synonymes et de dérivés, ces deux catégories représentant, dans les deux langues, environ 87% des relations identifiées dans les ensembles de référence. Il est important de noter que les QSYN et les DRV représentent 72% des couples de référence en français et 78% en anglais, donc la proportion élevée de QSYN et de DRV parmi les paires de termes faisant partie du même ensemble de référence est attribuable en partie à la répartition des couples de référence en fonction de la relation. Nous fournissons ces données à titre indicatif afin

Relation	Nb couples
QSYN	192 (49%)
DRV	147 (38%)
ANTI	40 (10%)
HYP	10 (3%)
Total	389 (100%)

(a) FR

Relation	Nb couples
QSYN	142 (55%)
DRV	83 (32%)
ANTI	25 (10%)
HYP	7 (3%)
Total	257 (100%)

(b) EN

Tableau 5.VI – Répartition des relations identifiées dans les ensembles de référence.

de montrer que l'appartenance au même cadre sémantique représente un mélange de dérivation et de relations paradigmatiques classiques telles que la quasi-synonymie.

5.6.2 Mesures d'évaluation

Les données de référence décrites à la section 5.6.1 servent à évaluer les graphes de voisinage distributionnel et les modèles sous-jacents. Les listes ordonnées de PPV que l'on obtient à partir d'un modèle distributionnel sont évaluées en calculant la précision de ces classements par rapport aux données de référence (couples et ensembles) ; l'évaluation des graphes de voisinage est réalisée en calculant dans quelle mesure les termes appartenant aux mêmes ensembles de référence sont reliés dans les graphes. Nous décrivons les mesures d'évaluation utilisées à cette fin dans les sections suivantes. Nous commençons par expliquer les mesures utilisées pour évaluer les graphes, car la compréhension de ces mesures facilite, à notre avis, la compréhension de la mesure utilisée pour évaluer les listes ordonnées de PPV.

5.6.2.1 Mesures d'évaluation du graphe

La comparaison entre un graphe de voisinage et les données de référence est réalisée au moyen de mesures d'évaluation utilisées dans les domaines de la recherche d'information et du TAL, entre autres, à savoir la précision et le rappel.

Ces mesures d'évaluation sont calculées pour chacun des termes dans les ensembles de référence ; nous les avons également calculées pour chacune des entrées dans les couples de référence, mais dans l'analyse des résultats de l'évaluation des graphes présentée au chapitre 7, nous nous concentrons sur les résultats obtenus sur les ensembles de référence. Pour chacun de ces termes, nous comparons l'ensemble des termes auxquels il est relié dans le graphe de voisinage distributionnel à l'ensemble de ses termes reliés selon les données de référence, au moyen des mesures d'évaluation. On peut donc considérer cette évaluation comme une comparaison de graphes, car nous comparons les graphes de voisinage distributionnel au graphe que forment les données de référence. En

effet, les ensembles de référence peuvent être considérés comme un graphe non orienté dans lequel chaque ensemble forme une clique (un sous-graphe où chaque nœud est relié à chaque autre nœud) ; les couples de référence forment plutôt un graphe orienté.

Pour expliquer comment les mesures d'évaluation sont calculées, nous utiliserons la notation suivante :

- $Q = \{q_1, \ldots, q_n\}$ est l'ensemble des requêtes que l'on utilise pour l'évaluation. Puisqu'on utilise les ensembles de référence pour évaluer les graphes, Q contient tous les termes dans les ensembles de référence ; le nombre de requêtes n est donc 223 en français et 168 en anglais. Si on utilisait plutôt les couples de référence pour l'évaluation, Q contiendrait chacune des entrées dans les couples de référence ; n vaudrait alors 490 en français et 381 en anglais.

- Pour chaque requête q_i, l'ensemble $T_i = \{t_i^1, \ldots, t_i^{m_i}\}$ contient les termes reliés à q_i selon les données de référence. T_i est donc l'union des ensembles de référence qui contiennent q_i, sans compter q_i ; autrement dit, T_i est l'ensemble des termes qui évoquent un des cadres sémantiques évoqués par q_i selon le Framed DiCoEnviro. Si on utilisait plutôt les couples de référence, T_i serait l'ensemble des termes qui forment avec l'entrée q_i un couple de référence.

- $V(q)$ est une fonction qui retourne le voisinage distributionnel de la requête q, c'est-à-dire les mots reliés à q dans le graphe de voisinage distributionnel[25]. Comme nous l'avons expliqué à la section 5.5.1, dans le cas du graphe orienté, le voisinage du nœud q désigne l'ensemble des successeurs de q, c'est-à-dire les nœuds reliés à q par un arc partant de ce dernier.

En utilisant cette notation, on peut formuler la précision (p) et le rappel (r) pour le voisinage d'un terme q_i de la façon suivante :

[25]La façon dont nous calculons les mesures d'évaluation tient seulement compte des mots reliés directement aux requêtes, mais il serait possible de modifier la méthode d'évaluation pour prendre en compte les mots que l'on peut atteindre en faisant plus d'un saut dans le graphe.

$$p(q_i) = \frac{|T_i \cap V(q_i)|}{|V(q_i)|} \tag{5.1}$$

$$r(q_i) = \frac{|T_i \cap V(q_i)|}{|T_i|} \tag{5.2}$$

où $|T_i|$ est la taille de l'ensemble T_i, $|V(q_i)|$ est la taille du voisinage $V(q_i)$, et $|T_i \cap V(q_i)|$ est la taille de l'intersection de T_i et $V(q_i)$. La précision correspond donc à la proportion des voisins qui sont parmi les termes reliés, et le rappel, la proportion des termes reliés qui sont parmi les voisins. La précision est définie comme étant 0 si $|V(q_i)| = 0$.

Pour illustrer comment ces mesures sont calculées, supposons que l'on évalue le voisinage de la requête $q_i = extreme$. Dans les couples de référence, l'entrée *extreme* a deux termes reliés :

$$T_i = \{severe,\ intense\}$$

Dans la Figure 3.6, les 5 PPV de *extreme* étaient :

$$V(q_i) = \{\textbf{\textit{severe}},\ \textbf{\textit{intense}},\ harsh,\ catastrophic,\ unusual\}$$

Donc, cette requête aurait 5 voisins et 2 termes reliés, dont tous les deux sont parmi ses voisins. La précision et le rappel pour le voisinage de *extreme* seraient donc :

$$p(q_i) = \frac{2}{5} = 0.4$$

$$r(q_i) = \frac{2}{2} = 1.0$$

Si on calculait ces mesures sur les ensembles de référence plutôt que les couples de référence, le résultat serait différent dans ce cas-ci. Les quatre termes suivants font partie

d'un ensemble de référence qui contient *extreme* :

$$T_i = \{\textit{severe, severity, intense, intensity}\}$$

Ainsi, la précision du voisinage de *extreme* serait toujours de 0.4, mais le rappel serait de 0.5 plutôt que 1.0, puisque seulement 2 de ses 4 termes reliés font partie de son voisinage.

On obtient des mesures d'évaluation globales pour le graphe en calculant la précision et le rappel (équations 5.1 et 5.2) pour toutes les requêtes et en prenant la moyenne de chaque mesure :

$$P(Q) = \frac{1}{n}\sum_{i=1}^{n} p(q_i) \tag{5.3}$$

$$R(Q) = \frac{1}{n}\sum_{i=1}^{n} r(q_i) \tag{5.4}$$

Il est important de noter que plus le nombre de voisins par mot est élevé, plus le rappel tendra à être élevé ; dans le cas où le graphe contient une arête entre chaque paire de mots, le rappel est toujours de 1. À l'inverse, une précision élevée peut seulement être atteinte si le nombre de voisins par nœud est relativement petit, car les termes participent généralement à un petit nombre de relations lexicales, ou du moins un nombre beaucoup plus petit que le nombre maximal de voisins qu'ils peuvent avoir dans le graphe de voisinage ; si une requête q_i a un nombre très élevé de voisins, la précision $p(q_i)$ sera forcément faible.

Dans le cas du graphe de k PPV, la taille des voisinages augmente en fonction du paramètre k. Donc, plus k est élevé, plus le rappel tendra à être élevé, et plus la précision tendra à être faible. Si on cherche un graphe qui offre à la fois un rappel élevé et une bonne précision, on peut faire appel à une mesure qui combine ces deux mesures. Nous calculons à cette fin une troisième mesure appelée *mesure* F_1 (ou *F-mesure*). La mesure

F_1 est la moyenne harmonique de la précision et du rappel[26] :

$$F_1(Q) = \frac{2 \times P(Q) \times R(Q)}{P(Q) + R(Q)} \tag{5.5}$$

La mesure F_1 accorde un poids égal à la précision et au rappel. D'une façon plus générale, on peut formuler la mesure F_β de la façon suivante :

$$F_\beta(Q) = (1 + \beta^2) \times \frac{P(Q) \times R(Q)}{(\beta^2 \times P(Q)) + R(Q)} \tag{5.6}$$

où β détermine l'importance relative accordée au rappel par rapport à la précision. Ainsi, si on accordait deux fois plus d'importance au rappel qu'à la précision, on fixerait β à 2 ; inversement, on utiliserait $\beta = 0.5$ pour accorder deux fois plus d'importance à la précision qu'au rappel.

Puisque nous n'avons aucun présupposé sur l'importance relative de la précision et du rappel, nous utilisons la mesure F_1 par défaut, mais nous montrerons au chapitre 7 que la valeur de β nous offre un moyen efficace d'ajuster la densité du graphe de voisinage.

Soulignons que, dans le cas où la précision et le rappel sont nuls, la mesure F_β est nulle par définition.

Nous utilisons donc la précision, le rappel et la mesure F_β pour vérifier dans quelle mesure les voisins distributionnels correspondent à des termes évoquant le même cadre sémantique ou à des paires de termes participant à une relation paradigmatique. En effet, ces mesures ont été calculées sur les deux jeux de données de référence, mais dans le cadre de l'analyse des résultats de l'évaluation des graphes présentée au chapitre 7, nous nous concentrons sur les résultats obtenus sur les ensembles de référence. Par ailleurs, la méthode d'évaluation quantitative présentée ci-dessus est accompagnée d'une analyse plutôt qualitative portant sur le voisinage des ensembles de référence, afin d'explorer davantage notre seconde hypothèse (voir section 4.2), qui concerne non seulement le

[26] Il serait possible de calculer cette mesure différemment, notamment en calculant la mesure F_1 pour chaque requête et en prenant la moyenne sur toutes les requêtes. Nous discutons ce point à l'annexe VI.

voisinage de termes individuels, mais également le voisinage d'ensembles de termes.

5.6.2.2 Mesure d'évaluation des listes ordonnées de PPV

En plus de calculer la précision et le rappel pour évaluer les graphes de voisinage distributionnel, nous calculons une autre mesure pour évaluer directement les listes ordonnées de voisins que l'on obtient à partir du modèle distributionnel, en faisant abstraction du paramétrage du graphe que l'on utilise pour l'interroger. Rappelons que ces listes ordonnées de voisins sont obtenues en calculant la similarité distributionnelle entre tous les mots-cibles deux à deux, au moyen du cosinus de l'angle des vecteurs.

La mesure d'évaluation que nous utilisons à cette fin est la moyenne des précisions moyennes (*mean average precision* ou *MAP*), une mesure couramment utilisée pour l'évaluation des moteurs de recherche[27], mais aussi des modèles distributionnels.

Nous calculons la MAP sur les deux jeux de données de référence (les ensembles et les couples), et nous la calculons également sur les sous-ensembles de couples correspondant à chaque PDD et à chaque catégorie de relations lexicales (voir Tableaux 5.II et 5.III), ce qui nous permet de vérifier comment les résultats varient en fonction de ces deux facteurs.

Par ailleurs, étant donné que nous calculons la même mesure d'évaluation sur les deux jeux de données de référence, il sera possible de comparer les résultats obtenus sur les deux jeux, qui servent à tester nos deux hypothèses.

Dans le cadre de l'évaluation d'un modèle distributionnel, on calcule la MAP de la façon suivante : pour chaque requête, on calcule la précision de sa liste ordonnée de PPV à chacun des rangs où se trouvent ses termes reliés, puis on calcule la moyenne de ces précisions (*average precision* ou AP) ; la MAP est alors la moyenne des précisions moyennes obtenues pour chaque requête.

[27] Pour plus d'information sur la MAP dans le contexte de la recherche d'information, voir les explications de Manning et al. [124] à ce sujet, à l'adresse suivante : `http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html` (page consultée le 12 août 2015).

Reprenons l'exemple de la requête $q_i = \textit{extreme}$ ayant les termes reliés suivants :

$$T_i = \{\textit{severe, intense}\}$$

Supposons que la liste ordonnée de ses PPV commence par :

$$V(q_i) = \{\textit{harsh, } \textbf{\textit{intense}}\textit{, catastrophic, unusual, } \textbf{\textit{severe}}, \ldots\}$$

Il est facile de voir que la précision de cette liste est de 0.5 au rang du terme relié *intense*, car le rang de ce terme est 2, et 1 des 2 premiers voisins est un terme relié ; de même, la précision au rang du terme relié *severe* est de 0.4, puisque 2 des 5 premiers voisins sont des termes reliés. Ainsi, la précision moyenne est

$$AP(q_i) = \frac{0.5 + 0.4}{2} = 0.45$$

et la MAP serait la moyenne des précisions moyennes calculées de cette façon pour toutes les requêtes.

Pour fournir une définition formelle de cette mesure, supposons que la fonction $V(q,t)$ retourne la liste ordonnée des PPV de la requête q jusqu'à ce qu'on obtienne le terme t. On peut alors formuler la précision moyenne (AP) pour la requête q_i de la façon suivante :

$$AP(q_i) = \frac{1}{|T_i|} \sum_{j=1}^{|T_i|} \frac{|T_i \cap V(q_i, t_i^j)|}{|V(q_i, t_i^j)|} \tag{5.7}$$

Et la MAP est la moyenne des précisions moyennes (AP) pour toutes les requêtes :

$$MAP(Q) = \frac{1}{n} \sum_{i=1}^{n} AP(q_i) \tag{5.8}$$

En somme, nous utilisons la MAP pour évaluer la qualité des listes ordonnées de PPV que l'on obtient à partir des modèles distributionnels. Plus les voisins « corrects »

apparaissent près du début des listes de PPV, plus la MAP est élevée. Soulignons également que les voisins « incorrects » qui apparaissent après le dernier voisin correct d'une requête donnée ne diminuent pas la MAP. La MAP nous fournit donc une mesure globale de la qualité du classement des voisins distributionnels basée sur les données de référence, qui tient compte à la fois de la précision et du rappel.

Plusieurs mesures d'évaluation peuvent être utilisées pour évaluer les modèles distributionnels au moyen de données de référence, telles que la précision moyenne à des rangs fixes (précision au rang 1, 10, 100, etc.) et la R-précision. Dans le cadre de ce travail, nous avons choisi d'utiliser une seule mesure pour évaluer les modèles distributionnels (en plus des mesures utilisées pour évaluer les graphes). Puisque nous évaluons deux types différents de modèles sur plusieurs jeux de données et analysons en profondeur l'influence de leurs (hyper)paramètres, il serait très difficile de tenir compte de plusieurs mesures d'évaluation dans l'analyse des résultats. En outre, selon Manning et al. [124], la R-précision et la MAP sont fortement corrélées[28]. En somme, nous ne jugeons pas nécessaire d'utiliser plusieurs mesures pour évaluer les modèles distributionnels dans le cadre de ce travail ; nous reviendrons sur le choix de cette mesure d'évaluation dans l'analyse des résultats, à la section 6.6.3.

5.7 Synthèse

Notre méthodologie est axée sur la construction et l'évaluation de modèles sémantiques distributionnels. Nous construisons ces modèles sur des corpus du domaine de l'environnement dans deux langues, le français et l'anglais. À partir de ces modèles, nous calculons les PPV de chaque mot-cible, puis nous construisons des graphes de k PPV à partir de ces listes. Les listes triées de PPV ainsi que les graphes de k PPV sont évalués au moyen d'une comparaison entre les voisinages (ordonnés et non ordonnés respectivement) qu'ils représentent et deux jeux de données de référence que nous avons

[28] `http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html` (page consultée le 13 août 2015).

extraits de dictionnaires spécialisés du domaine de l'environnement, ces deux jeux de données servant à tester nos deux hypothèses.

Comme le montre la Figure 5.4, le programme qui construit et évalue les modèles distributionnels prend en entrée le corpus, la liste des mots-cibles et les deux jeux de données de référence ; il produit un fichier de résultats qui contient les mesures d'évaluation (précision et rappel calculés sur les graphes, MAP calculée directement à partir du modèle) pour chaque paramétrage testé. Les mêmes mesures d'évaluation sont utilisées pour les deux jeux de données de référence.

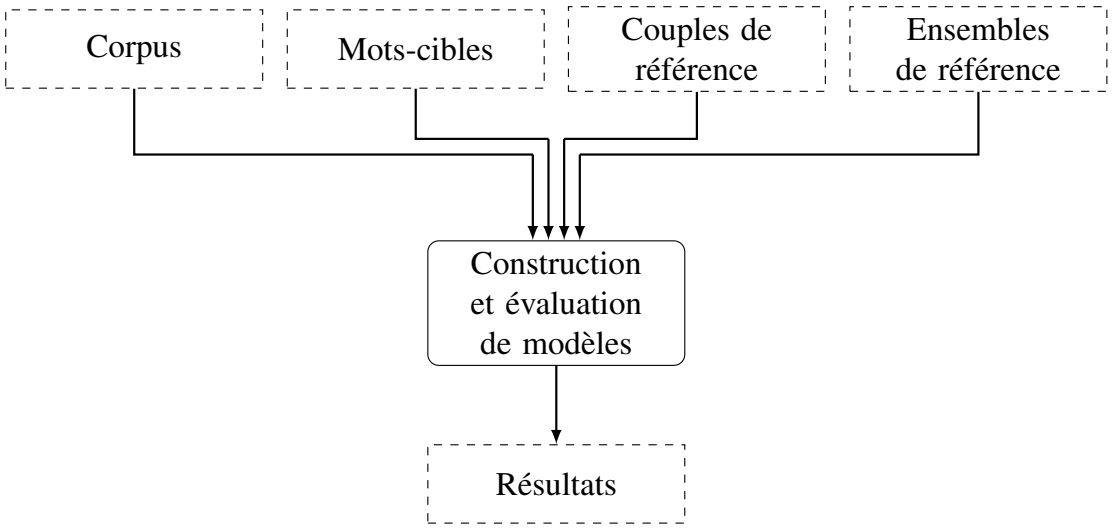


Figure 5.4 – Entrées et sortie du programme de construction et d'évaluation de modèles.

Le programme contient une liste de valeurs à tester pour les différents (hyper)paramètres des deux méthodes utilisées pour construire les modèles (AD et `word2vec`) et pour les paramètres du graphe. En appliquant la méthode d'évaluation automatique à différents paramétrages des modèles et du graphe, nous pouvons analyser l'influence des (hyper)paramètres sur les résultats et déterminer les valeurs optimales de chacun. On peut résumer le fonctionnement de ce programme de la façon suivante :

- Pour chaque combinaison possible des valeurs à tester pour les (hyper)paramètres de la méthode distributionnelle (AD ou `word2vec`), il construit un modèle.

- Il calcule la matrice de similarité à partir du modèle et évalue le modèle en calculant la MAP sur les données de référence.

- À partir de la matrice de similarité calculée à l'étape précédente, pour chaque combinaison possible des valeurs à tester pour les paramètres du graphe, il construit un graphe et l'évalue en calculant la précision et le rappel sur les données de référence ; la mesure F_β peut être calculée par la suite à partir de la précision et du rappel.

Les résultats de cette expérience sont analysés aux chapitres 6 et 7. Le chapitre 6 porte sur l'évaluation des listes ordonnées de PPV et prend en compte les deux jeux de données de référence. Au chapitre 7, nous analysons les résultats de l'évaluation des graphes de voisinage, en nous concentrant sur le cadre descriptif de la sémantique des cadres, que reflètent les ensembles de référence.

Soulignons enfin que les données de référence et les programmes que nous avons développés afin de réaliser cette expérience sont disponibles à l'adresse `https://github.com/gbcolborne/exp_phd`, y compris le programme de prétraitement du corpus et le programme de construction et d'évaluation de modèles distributionnels. Il est important de noter que, dans la version de ce programme que nous avons rendue disponible, la sélection des mots-cibles se fait à la volée, plutôt que par la lecture d'une liste de mots-cibles créée au préalable, contrairement à ce qui est indiqué dans la Figure 5.4. Les mesures d'évaluation obtenues pourraient être légèrement différentes des résultats présentés dans cette thèse en raison des différences quant aux mots-cibles utilisés. Par ailleurs, nous fournissons deux versions du programme de construction et d'évaluation de modèles, une pour l'AD et l'autre pour `word2vec`.

CHAPITRE 6

RÉSULTATS DE L'ÉVALUATION DES MODÈLES

Introduction

Dans ce chapitre, nous présentons les résultats de l'évaluation des listes ordonnées de voisins que l'on obtient pour chaque mot-cible à partir d'un modèle distributionnel. Comme nous l'avons expliqué au chapitre 5, l'évaluation des listes ordonnées de PPV est réalisée en les comparant, au moyen d'une mesure d'évaluation appelée *MAP*, à des jeux de données de référence constitués de requêtes (termes) et de termes sémantiquement reliés à ces requêtes ; plus les termes reliés à chaque requête sont près du début de la liste de ses PPV, plus la MAP est élevée.

Ces résultats sont analysés en fonction de la méthode utilisée pour construire le modèle, de son paramétrage, de la langue traitée et du jeu de données de référence utilisé pour l'évaluation. Nous présentons ensuite au chapitre 7 les résultats de l'évaluation des graphes de voisinage que nous utilisons pour interroger les modèles.

Dans la première partie de ce chapitre, nous présentons une analyse quantitative des résultats, suivie d'une discussion de nos observations. Après avoir expliqué les méthodes d'analyse utilisées, nous présentons à la section 6.2 une vue d'ensemble des résultats afin de fournir une première perspective sur la précision des voisinages distributionnels, que nous analysons en fonction de la partie du discours (PDD) des requêtes et des relations lexicales ciblées. Puis, nous comparons les deux méthodes que nous utilisons pour construire les modèles, l'analyse distributionnelle (AD) et `word2vec` (W2V). Par la suite, nous analysons l'influence des (hyper)paramètres de ces deux méthodes. Enfin, nous présentons une discussion des résultats à la section 6.6 : nous discutons la question de l'interprétation des mesures d'évaluation, nous résumons nos observations quant aux facteurs dont dépend la qualité des résultats et nous analysons quelques facteurs pouvant

expliquer la faible précision de certains voisinages distributionnels.

6.1 Méthodes d'analyse utilisées

L'analyse présentée dans la première partie de ce chapitre repose sur l'observation des valeurs moyennes ou maximales de la MAP, mesure d'évaluation que nous utilisons pour estimer la qualité du classement des PPV que l'on obtient à partir de chaque modèle.

Par exemple, lorsque nous comparons l'AD et W2V, nous observons pour chaque méthode la MAP maximale, c'est-à-dire celle du paramétrage qui atteint la MAP la plus élevée.

Lorsque nous analysons l'influence des (hyper)paramètres des deux méthodes, nous déterminons les valeurs optimales de chaque paramètre en observant la MAP moyenne pour chaque valeur du paramètre, c'est-à-dire la MAP moyenne de tous les paramétrages où ce paramètre a cette valeur. Cette méthode d'analyse nous permet non seulement de déterminer la valeur optimale de chaque paramètre, mais aussi d'avoir une idée de l'importance de l'influence exercée par ce paramètre sur la MAP. En effet, si les MAP moyennes que l'on obtient pour chaque valeur d'un paramètre sont très différentes, cela indique que celui-ci a une influence importante sur la précision des voisinages que l'on obtient à partir du modèle ; et plus les moyennes sont différentes, plus le paramètre exerce une influence importante. Une analyse de la variance permettrait de quantifier l'importance relative des paramètres [106], mais nous jugeons qu'une telle analyse n'est pas nécessaire pour réaliser les objectifs de cette thèse.

Soulignons que nous analysons seulement l'influence des (hyper)paramètres considérés séparément, et ne traitons pas les interactions entre les paramètres. Il est possible que la valeur optimale d'un paramètre dépende de la valeur utilisée pour un autre, mais nous ne le vérifions pas dans le cadre de ce travail. Étant donné que nous tenons compte de nombreux facteurs dans l'analyse des résultats (la langue traitée, la PDD des requêtes, le cadre descriptif adopté, les relations ciblées, la méthode utilisée pour construire le mo-

dèle distributionnel et son paramétrage), nous ne pouvons pas examiner en profondeur toutes ces dimensions dans cette thèse. En effet, l'analyse du paramétrage tient compte des résultats obtenus dans les deux langues et sur les différents jeux de données de référence, ce qui nous permet d'aborder des questions telles que :

- Est-ce que les paramétrages optimaux dépendent de la langue traitée ?
- Est-ce qu'ils dépendent du cadre descriptif adopté ?
- Est-ce qu'ils dépendent du type de relations lexicales que l'on souhaite identifier ?
- Est-ce qu'ils dépendent de la PDD des requêtes ?

En somme, en ce qui concerne l'analyse du paramétrage des modèles, nous nous limitons à une étude de l'influence des paramètres considérés séparément, qui vise à déterminer les paramétrages optimaux et les facteurs dont on doit tenir compte pour fixer chaque paramètre. Les interactions entre les paramètres pourraient faire l'objet d'un travail en soi, et ont d'ailleurs été abordées dans d'autres travaux [105, 106].

Nous tenons à souligner que la MAP, bien qu'elle soit très utile pour comparer les résultats obtenus sur différents jeux de données ou en utilisant différents modèles, peut être difficile à interpréter. Nous tenterons de faciliter l'interprétation de cette mesure dans la discussion des résultats, à la section 6.6.3. D'ailleurs, nous fournirons une perspective très concrète de la qualité de voisinages distributionnels en présentant des exemples visuels de ces voisinages à la section 7.4. En outre, nous montrerons, dans l'analyse des sources d'erreurs présentée à la section 6.6.2, que l'évaluation automatique basée sur des données de référence tend à sous-estimer la précision. Quoi qu'il en soit, dans cette partie de l'analyse des résultats, la MAP nous servira surtout à comparer différents modèles et à comparer les résultats obtenus sur différents jeux de données.

6.2 Résultats globaux sur les différents jeux de données

Avant de comparer les deux méthodes utilisées pour construire des modèles distributionnels et d'analyser l'influence de leurs (hyper)paramètres, nous présentons dans cette section une vue d'ensemble des résultats produits par tous les modèles évalués.

La Figure 6.1 présente la MAP obtenue sur les deux jeux de données principaux, à savoir les couples et les ensembles de référence, au moyen d'un diagramme de quartiles ; cette figure montre la dispersion de la MAP de tous les modèles, dans les deux langues. Rappelons qu'un diagramme de quartiles illustre les quartiles d'un ensemble de données, c'est-à-dire les valeurs qui séparent l'ensemble (trié) de données en quatre sous-ensembles de taille égale : les deux extrémités des boîtes rectangulaires indiquent les quartiles supérieur et inférieur, et la ligne au milieu des boîtes indique la médiane ; les *moustaches* situées à l'extérieur des boîtes s'étendent de la valeur minimale à la valeur maximale et les points situés à l'extérieur des moustaches représentent des valeurs aberrantes (p. ex. dans la Figure 6.1, les points situés en dessous des moustaches représentent des valeurs anormalement basses).

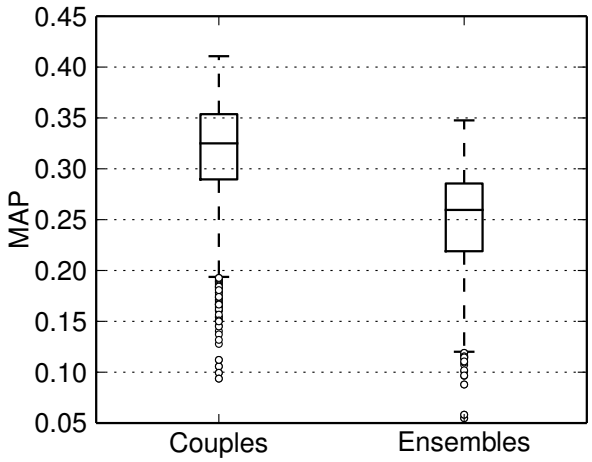


Figure 6.1 – Dispersion de la MAP de tous les modèles évalués, calculée sur les couples de référence et les ensembles de référence.

La Figure 6.1 montre que l'on atteint une MAP plus élevée sur les couples de réfé-

rence que sur les ensembles de référence. Cela signifie que les termes que l'on souhaite repérer se trouvent plus près du début des listes ordonnées de PPV, en moyenne, lorsqu'on cherche des termes reliés paradigmatiquement (des synonymes, des dérivés, etc.) que lorsqu'on cherche des termes évoquant le même cadre sémantique. Nous proposerons des explications ci-dessous.

La Figure 6.1 montre également que les résultats obtenus présentent beaucoup de variation. Par exemple, la MAP obtenue sur les couples de référence varie entre 0.0937 et 0.4107. Nous montrerons aux sections 6.4 et 6.5 que la MAP que l'on obtient au moyen de l'AD et de W2V dépend beaucoup des valeurs utilisées pour les différents (hyper)paramètres de ces deux méthodes.

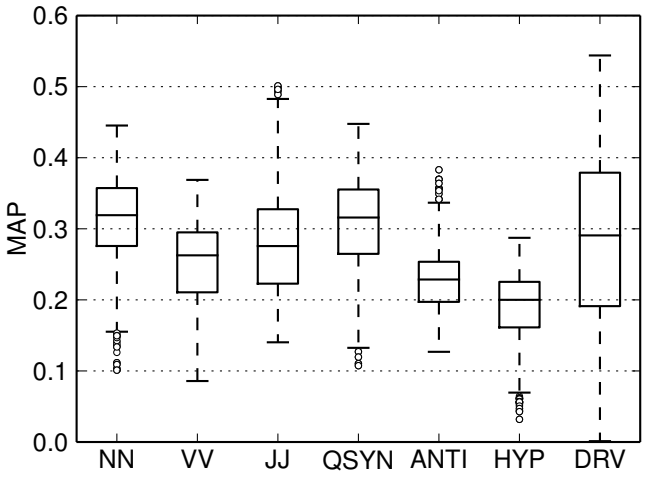


Figure 6.2 – Dispersion de la MAP, calculée sur chaque sous-ensemble de couples.

La Figure 6.2 présente les résultats obtenus sur chacun des sous-ensembles de couples de référence, toute approche confondue (AD et W2V) et dans les deux langues. Ces résultats indiquent notamment que les relations entre verbes sont moins bien détectées que celles entre les noms ou celles entre les adjectifs, puisque la MAP atteint des valeurs moins élevées lorsque l'on la calcule sur ce sous-ensemble de couples. Ainsi, si la MAP est moins élevée sur les ensembles de référence que sur les couples de référence, comme le montrait la Figure 6.1, une des raisons est que les ensembles de référence contiennent

une proportion élevée de verbes (voir la section 5.6.1.2).

La Figure 6.2 montre également que les relations les mieux captées sont les QSYN et les DRV, les relations les moins bien captées étant les HYP. La MAP élevée que l'on obtient sur les DRV indique que les modèles distributionnels captent aussi bien les relations entre termes appartenant à des PDD différentes, et ayant donc des comportements syntaxiques différents, qu'entre termes appartenant à la même PDD.

Par ailleurs, ces données indiquent que les antonymes sont moins bien détectés que les (quasi-)synonymes. Les antonymes ont tendance à avoir des distributions similaires[1], mais apparaissent plus souvent ensemble, dans le même contexte, que les synonymes [106, 139]. Cela expliquerait éventuellement le fait que les antonymes sont moins bien détectés que les synonymes, du moins en partie.

Les ensembles de référence contiennent beaucoup de quasi-synonymes, mais aussi beaucoup de dérivés syntaxiques (voir la section 5.6.1.2), et puisque ces deux relations sont celles que les modèles distributionnels captent le mieux, on pourrait s'attendre à ce que la précision sur les ensembles de référence soit plus élevée que celle que nous avons observée. Nous verrons aux sections 6.4 et 6.5 que les paramétrages de l'AD et de W2V qui captent le mieux les DRV sont très différents des paramétrages optimaux pour les autres relations lexicales, notamment pour les QSYN. La précision relativement faible que l'on obtient sur les ensembles de référence est attribuable en partie à ce facteur, à savoir que les relations entre les termes dans les ensembles de référence comprennent un mélange de dérivation syntaxique et de relations paradigmatiques classiques telles que la quasi-synonymie, et que les paramétrages optimaux pour ces deux types de relations sont très différents.

Soulignons enfin que la MAP obtenue sur les DRV a une dispersion très élevée par rapport aux autres relations. Nous verrons ci-dessous qu'elle est particulièrement sensible au paramétrage du modèle.

[1] « The closeness of opposites [...] manifests itself, for instance, in the fact that the members of a pair have almost identical distributions, that is to say, very similar possibilities of normal and abnormal occurrence. » [44, p. 197]

6.3 Évaluation comparative de l'AD et de word2vec

À la section 6.2, nous avons montré, entre autres, que la précision du classement des PPV présente beaucoup de variation. Dans cette section, nous comparons les deux méthodes que nous avons utilisées pour construire des modèles distributionnels afin de vérifier si le choix de la méthode a une influence importante sur la précision que l'on obtient, et pour déterminer quel type de modèle offre les meilleurs résultats. Cette comparaison est réalisée en observant pour chaque type de modèle (AD et W2V) la MAP maximale, c'est-à-dire celle du paramétrage qui obtient la MAP la plus élevée.

Modèle	Couples	Ensembles
AD	**0.4107**	0.3258
W2V	0.3930	**0.3476**

Tableau 6.I – MAP maximale en fonction du type de modèle, sur les deux jeux de données.

Le Tableau 6.I montre la MAP maximale pour chaque type de modèle, les deux langues confondues (nous vérifierons comment les résultats varient en fonction de la langue traitée ci-dessous) ; les meilleurs résultats obtenus sur chaque jeu de données sont en caractères gras. Ces résultats indiquent que l'AD atteint une MAP maximale plus élevée que W2V sur les couples de référence[2], mais que c'est W2V qui obtient la MAP la plus élevée sur les ensembles de référence.

Le Tableau 6.II présente les résultats obtenus sur chacun des sous-ensembles de couples. On observe ici que l'AD obtient une MAP maximale plus élevée que W2V sur tous les sous-ensembles de couples sauf les DRV ; sur ces couples, constitués de termes appartenant à des PDD différentes, W2V atteint une MAP maximale beaucoup plus élevée que l'AD. Autrement dit, l'AD modélise mieux la similarité sémantique des

[2]Dans le cadre d'une étude reliée à cette thèse [19], nous avons souligné que les différences de MAP que nous avons observées entre l'AD et W2V sont statistiquement significatives ($p < 0.05$) selon un test de Wilcoxon [97], du moins dans le cas des jeux de données pris en compte dans cette étude, c'est-à-dire les couples de référence en français, ainsi que les 4 sous-ensembles correspondant aux différentes relations lexicales.

Modèle	NN	VV	JJ	QSYN	ANTI	HYP	DRV
AD	**0.4453**	**0.3688**	**0.5011**	**0.4476**	**0.3830**	**0.2873**	0.4578
W2V	0.4160	0.3425	0.4537	0.4147	0.3213	0.2653	**0.5439**

Tableau 6.II – MAP maximale en fonction du type de modèle, sur chaque sous-ensemble de couples.

termes ayant des comportements syntaxiques similaires, tandis que W2V capte mieux la similarité des termes qui ont des comportements syntaxiques différents, mais le même sens[3].

Ainsi, si W2V produit de meilleurs résultats sur les ensembles de référence, cela est attribuable au fait qu'il détecte mieux les relations de dérivation syntaxique.

Langue	Modèle	Couples	Ensembles
FR	AD	**0.4107**	**0.3183**
	W2V	0.3819	0.3046
EN	AD	**0.3953**	0.3258
	W2V	0.3537	**0.3476**

Tableau 6.III – MAP maximale en fonction du type de modèle et de la langue, sur les deux jeux de données.

Les Tableaux 6.III et 6.IV présentent la MAP maximale des deux types de modèles en fonction de la langue traitée, sur les deux jeux de données et sur chacun des sous-ensembles de couples respectivement. Sur la plupart des jeux de données, c'est l'AD qui offre les meilleurs résultats, du moins lorsqu'elle est paramétrée correctement : W2V atteint une MAP plus élevée que l'AD seulement sur les DRV (dans les deux langues) et sur les ensembles et les verbes en anglais. Bien que nous ne présentions pas ces données ici, nous tenons à souligner que si on observait la MAP moyenne plutôt que la MAP maximale, on observerait des tendances similaires, W2V offrant de meilleurs résultats en moyenne sur les DRV et les ensembles (dans les deux langues) et l'AD offrant de

[3]Cette différence serait éventuellement attribuable à la réduction de dimension opérée par W2V, comme nous l'avons souligné ailleurs [19].

meilleurs résultats sur tous les autres jeux de données.

Langue	Modèle	NN	VV	JJ	QSYN	ANTI	HYP	DRV
FR	AD	**0.4453**	**0.3688**	**0.4575**	**0.4476**	**0.2958**	**0.2873**	0.3960
	W2V	0.4160	0.3425	0.4214	0.4147	0.2810	0.2653	**0.4567**
EN	AD	**0.3982**	0.3264	**0.5011**	**0.4175**	**0.3830**	**0.2524**	0.4578
	W2V	0.3727	**0.3287**	0.4537	0.3962	0.3213	0.1990	**0.5439**

Tableau 6.IV – MAP maximale en fonction du type de modèle et de la langue, sur chaque sous-ensemble de couples.

La Figure 6.3 illustre la dispersion de la MAP en fonction du type de modèle au moyen d'un diagramme de quartiles, la MAP étant calculée ici sur les couples de référence. Si l'AD atteint une MAP plus élevée, la MAP que l'on obtient avec W2V a une dispersion moins élevée que celle que l'on obtient au moyen de l'AD. Le fait que la MAP varie plus dans le cas de l'AD indique que la qualité des résultats dépend davantage du paramétrage ; il est donc d'autant plus important de choisir avec soin la valeur des différents paramètres de l'AD.

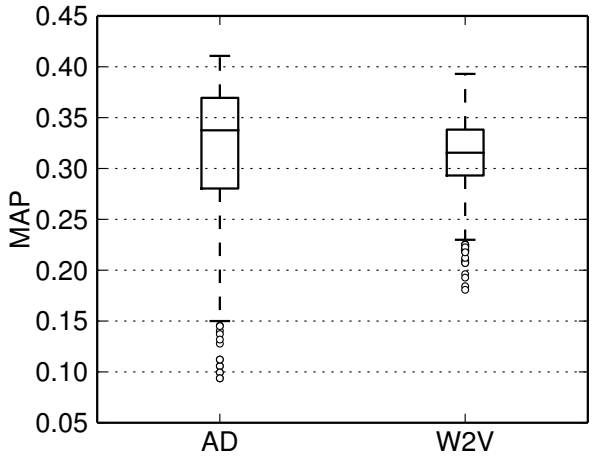


Figure 6.3 – Dispersion de la MAP en fonction du type de modèle, calculée sur les couples de référence.

6.4 Influence des paramètres de l'AD

À la section 6.3, nous avons montré que l'AD offre de meilleurs résultats que W2V sur la plupart des jeux de données utilisés dans ce travail. Nous avons également montré que la MAP atteinte par chaque méthode varie beaucoup d'un paramétrage à l'autre, en particulier dans le cas de l'AD. Dans cette section, nous analysons l'influence des paramètres de l'AD sur la qualité des résultats, puis dans la suivante, nous analyserons celle des hyperparamètres de W2V.

Rappelons que pour analyser l'influence d'un paramètre, nous comparons la MAP moyenne que l'on obtient pour chaque valeur de ce paramètre, c'est-à-dire la MAP moyenne de tous les paramétrages où ce paramètre a cette valeur.

6.4.1 Taille de la fenêtre de contexte

La taille de la fenêtre de contexte est le nombre de mots à gauche et à droite des occurrences du mot-cible qui sont considérés comme cooccurrents. Par exemple, si la taille de la fenêtre est 1, on observe un mot à gauche et un mot à droite. Plusieurs travaux suggèrent que les fenêtres étroites, d'un ou deux mots, captent le mieux les relations paradigmatiques [66, 165, entre autres].

La Figure 6.4 montre la MAP moyenne en fonction de la taille de la fenêtre de contexte, calculée sur les couples et les ensembles de référence. Cette figure montre que la taille optimale est de 3 mots dans le cas des couples et de 3 ou 4 mots dans le cas des ensembles ; la MAP diminue lentement à mesure que l'on augmente la taille de la fenêtre au-delà de ces valeurs. Elle montre également que la taille de la fenêtre exerce une influence importante sur la MAP, les fenêtres très étroites produisant une MAP beaucoup plus basse, en moyenne, que les fenêtres de 3 ou 4 mots.

La Figure 6.5 montre l'influence de la taille de fenêtre en fonction de la PDD des couples. On observe ici que sur les noms et les verbes, la taille optimale est 2, mais sur les adjectifs, c'est plutôt 1. Cette différence est simple à expliquer : les collocations des

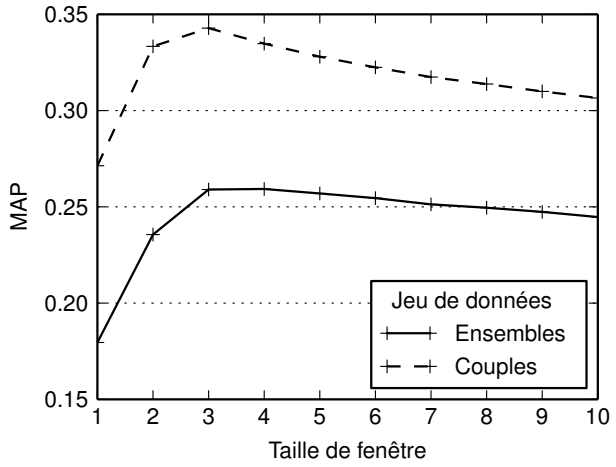


Figure 6.4 – MAP moyenne des modèles produits par AD en fonction de la taille de fenêtre et du jeu de données.

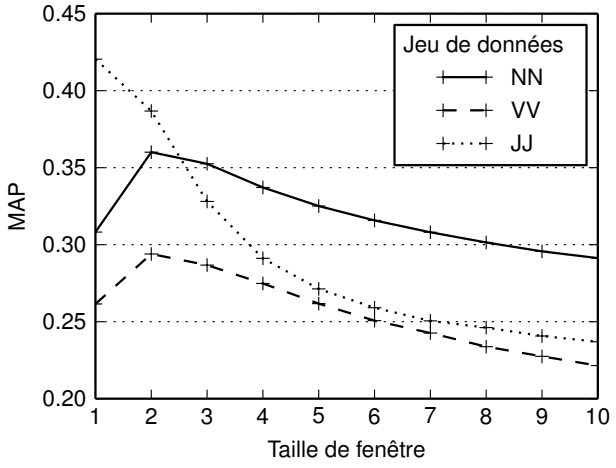


Figure 6.5 – MAP moyenne des modèles produits par AD en fonction de la taille de fenêtre et de la PDD des couples.

adjectifs qualificatifs ont tendance à apparaître immédiatement à côté de l'adjectif (du moins dans le cas des adjectifs épithètes) ; dans le cas des verbes et des noms, leurs collocations sont souvent à une distance de plus d'un mot, surtout en français. On observe également que l'influence de la taille de fenêtre est particulièrement importante dans le cas des adjectifs, la MAP diminuant rapidement à mesure que l'on élargit la fenêtre.

La Figure 6.6 montre l'influence de la taille de fenêtre en fonction des relations ciblées. La taille optimale est 1 ou 2 pour les ANTI, 2 pour les QSYN et 3 pour les HYP. Dans le cas des DRV, plus la fenêtre est large, plus la MAP est élevée ; la MAP moyenne ne semble pas avoir atteint un maximum même avec une taille de 10, mais soulignons que la MAP maximale que nous avons observée sur les DRV a été atteinte avec une fenêtre de 8 ou 9 mots, selon la langue. Ainsi, la valeur de ce paramètre qui donne les meilleurs résultats pour les DRV est très différente de la valeur optimale pour les autres relations, notamment les QSYN. Si on utilise la valeur optimale pour les QSYN (2 mots), la précision sur les DRV est beaucoup plus faible que si on utilisait une fenêtre plus large. Si on cherche une valeur qui donne de bons résultats sur ces deux types

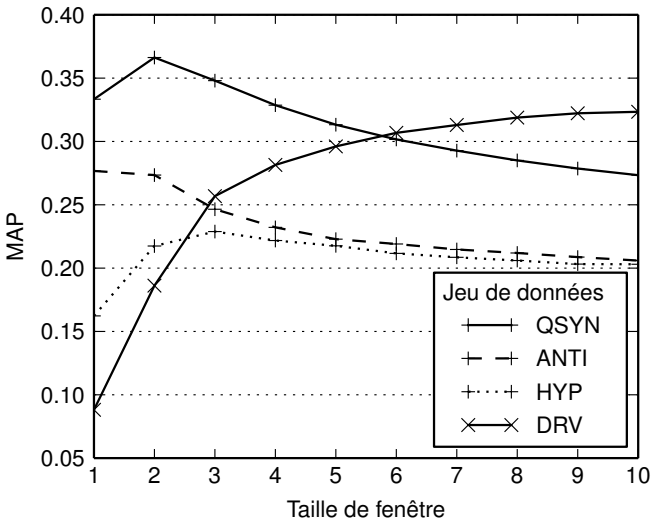


Figure 6.6 – MAP moyenne des modèles produits par AD en fonction de la taille de fenêtre et des relations ciblées.

de relations lexicales, ce qui est notamment le cas si on souhaite détecter des termes évoquant le même cadre sémantique, on utiliserait une valeur plus élevée. Les résultats que nous avons obtenus sur les ensembles de référence (voir Figure 6.4) suggèrent que cette valeur serait 3 ou 4.

En somme, ces résultats confirment que la taille de fenêtre optimale pour les relations paradigmatiques classiques telles que la synonymie est relativement petite. En revanche, ils indiquent que la dérivation syntaxique est une relation que l'on détecte mieux avec une fenêtre plus large ; nous proposerons une explication pour cela à la section 6.6.2.3.

En ce qui concerne l'influence de la langue traitée, la taille optimale sur les couples et les ensembles est la même dans les deux langues, mais on observe des petites différences lorsqu'on décompose les résultats en fonction de la PDD et des relations ciblées, la taille optimale étant parfois un peu plus élevée en français qu'en anglais :

- ANTI et QSYN : la taille optimale est 1 pour l'anglais, 2 pour le français.
- NN : la taille optimale est 1 ou 2 pour l'anglais, 2 pour le français.
- VV : la taille optimale est 2 pour l'anglais, 2 ou 3 pour le français.

Soulignons qu'il serait possible de définir la fenêtre de contexte de sorte qu'elle contienne un nombre fixe de mots pleins (ou de mots faisant partie des mots-contextes), plutôt qu'un nombre fixe de mots tout court, que ceux-ci soient pleins ou vides (et qu'ils fassent partie ou non des mots-contextes). Si on observait un nombre fixe de mots pleins, les différences observées entre l'anglais et le français quant à la taille optimale de la fenêtre de contexte seraient probablement encore moins importantes.

6.4.2 Direction de la fenêtre de contexte

Comme nous l'avons expliqué au chapitre 3, la direction de la fenêtre de contexte détermine si on observe les fréquences de cooccurrence à gauche (G) ou à droite (D) des occurrences d'un mot-cible, ou les deux ; dans ce cas, on peut prendre la somme des

deux fréquences (G+D) ou les encoder séparément (G&D). Dans le cadre de ce travail, nous avons testé les fenêtres G+D et G&D. Selon Bullinaria et Levy [34], la fenêtre G+D est plus courante que la fenêtre G&D, bien que cette dernière offre parfois une légère amélioration des résultats.

Fenêtre	Couples	Ensembles
G+D	**0.3217**	**0.2527**
G&D	0.3144	0.2349

Tableau 6.V – MAP moyenne des modèles produits par AD en fonction de la direction de la fenêtre, sur les deux jeux de données.

Le Tableau 6.V présente l'influence de ce paramètre sur la MAP, calculée sur les couples et les ensembles de référence. Ces résultats indiquent que la fenêtre G+D donne, en moyenne, de meilleurs résultats que la fenêtre G&D. Par contre, lorsqu'on observe les résultats obtenus sur chaque sous-ensemble de couples, présentés dans le Tableau 6.VI, on observe que la fenêtre G&D produit les meilleurs résultats sur tous les sous-ensembles, à l'exception des DRV, pour lesquels la fenêtre G+D donne une MAP beaucoup plus élevée, en moyenne. Si l'influence de ce paramètre est très importante en ce qui concerne les DRV, elle est moins importante pour les autres relations. Le gain que l'on obtient sur les DRV en utilisant une fenêtre G+D est tellement important que c'est cette fenêtre qui donne la meilleure MAP sur l'ensemble des couples.

Fenêtre	NN	VV	JJ	QSYN	ANTI	HYP	DRV
G+D	0.3116	0.2461	0.2794	0.3009	0.2256	0.2063	**0.3060**
G&D	**0.3274**	**0.2649**	**0.3069**	**0.3233**	**0.2368**	**0.2097**	0.2326

Tableau 6.VI – MAP moyenne des modèles produits par AD en fonction de la direction de la fenêtre, sur chaque sous-ensemble de couples.

Cette observation mérite que l'on s'y attarde un peu, d'autant plus qu'elle illustre bien pourquoi le paramétrage optimal pour les DRV est très différent de celui pour les autres relations lexicales. Qu'est-ce qui expliquerait qu'une fenêtre G+D produise des ré-

sultats plus précis pour les DRV, alors que c'est la fenêtre G&D qui donne les meilleurs résultats pour les autres relations ? La caractéristique principale qui distingue les DRV des autres relations prises en compte dans ce travail est qu'ils sont constitués de termes appartenant à des PDD différentes, et ayant donc des comportements syntaxiques différents. Parmi les couples de dérivés syntaxiques dans les données de référence, beaucoup sont constitués d'un nom et d'un verbe, les deux termes ayant le même sens et partageant des actants sémantiques. Par exemple, les termes *disparaître* et *disparition* ont un patient, dont les réalisations comprennent des termes tels que *écosystème* et *espèce*. Si cet actant a tendance à se réaliser d'un côté de l'un des deux termes, mais de l'autre côté de l'autre terme (p. ex. comparer *un écosystème disparaît* et *disparition d'un écosystème*), on ne modélisera pas bien le fait que *disparaître* et *disparition* partagent un actant (réalisé, en l'occurrence, par *écosystème*) si on utilise une fenêtre G&D, c'est-à-dire si les fréquences de cooccurrence observées à gauche et à droite des occurrences de ces deux termes sont encodées séparément. Cela explique vraisemblablement le fait que la fenêtre G+D capte mieux les dérivés syntaxiques que la fenêtre G&D.

Soulignons enfin que l'influence de ce paramètre est la même dans les deux langues à tous les égards, donc nous ne présentons pas séparément les résultats obtenus en français et en anglais.

6.4.3 Forme de la fenêtre de contexte

Le troisième paramètre lié à la fenêtre de contexte est la forme de la fenêtre. Celle-ci détermine l'incrément que l'on ajoute aux fréquences de cooccurrence pour chaque mot dans la fenêtre de contexte. Dans le cadre de ce travail, nous avons testé une fenêtre rectangulaire (l'incrément est 1 pour tous les cooccurrents dans la fenêtre) et une fenêtre triangulaire (l'incrément est inversement proportionnel à la distance entre le mot-cible et le cooccurrent).

Le Tableau 6.VII présente l'influence de ce paramètre sur la MAP, calculée sur les deux jeux de données. Ces résultats indiquent que la fenêtre triangulaire produit

Fenêtre	Couples	Ensembles
Rect.	0.3053	0.2386
Tri.	**0.3308**	**0.2489**

Tableau 6.VII – MAP moyenne des modèles produits par AD en fonction de la forme de la fenêtre, sur les deux jeux de données.

les meilleurs résultats. Par contre, si on observe les résultats obtenus sur chaque sous-ensemble de couples, présentés dans le Tableau 6.VIII, on observe de nouveau que la valeur optimale de ce paramètre pour les DRV est différente de celle pour toutes les autres relations, car c'est la fenêtre rectangulaire qui produit les meilleurs résultats sur les DRV. Nous pouvons proposer une explication pour cette différence à la lumière de notre analyse de l'influence de la taille de fenêtre. Nous avons montré à la section 6.4.1 que l'on détecte mieux les DRV en utilisant une fenêtre de contexte large plutôt qu'étroite, ce qui indique qu'il est plus important de prendre en compte des contextes (cooccurrents) éloignés dans le cas des DRV. Or, une fenêtre rectangulaire accorde plus de poids aux contextes éloignés qu'une fenêtre triangulaire, la différence étant d'autant plus importante que la distance est élevée. Cela expliquerait pourquoi la fenêtre rectangulaire produit de meilleurs résultats pour les DRV, du moins en partie.

Fenêtre	NN	VV	JJ	QSYN	ANTI	HYP	DRV
Rect.	0.3011	0.2349	0.2620	0.2890	0.2130	0.2060	**0.2861**
Tri.	**0.3379**	**0.2761**	**0.3243**	**0.3352**	**0.2494**	**0.2101**	0.2525

Tableau 6.VIII – MAP moyenne des modèles produits par AD en fonction de la forme de la fenêtre, sur chaque sous-ensemble de couples.

Quoi qu'il en soit, le gain obtenu sur les DRV au moyen de la fenêtre rectangulaire n'est pas suffisant pour que ce soit cette forme de fenêtre qui produise les meilleurs résultats sur l'ensemble des couples (étant donné la proportion de dérivés dans les couples de référence utilisés dans ce travail) ; globalement, c'est la fenêtre triangulaire qui produit les meilleurs résultats.

On observe également que ce paramètre exerce une influence plutôt importante sur la MAP, la différence entre la MAP moyenne des deux formes se situant généralement autour de 3 ou 4 points, excepté sur les HYP, où l'influence de la forme est très faible, et sur les QSYN et les JJ, où elle est particulièrement forte.

Soulignons enfin que, comme dans le cas de la direction de la fenêtre, l'influence de la forme de la fenêtre est la même dans les deux langues.

6.4.4 Pondération

La pondération est une fonction appliquée aux fréquences de cooccurrence. La pondération la plus courante en sémantique distributionnelle est sans doute l'information mutuelle ; plusieurs travaux indiquent que, parmi les différentes pondérations et mesures de similarité, la combinaison offrant les meilleurs résultats est l'information mutuelle et le cosinus de l'angle des vecteurs [33, 61, 66].

Nous avons testé huit pondérations dans le cadre de ce travail, en plus de la fonction d'identité (aucune pondération) ; les formules sont fournies dans l'annexe II. Les Figures 6.7 et 6.8 montrent la MAP moyenne en fonction de la pondération, sur les couples et les ensembles de référence respectivement ; dans chacune de ces figures, les résultats obtenus en français et en anglais sont présentés séparément.

La conclusion la plus évidente que l'on peut tirer en observant ces résultats est qu'il est très important d'appliquer une pondération aux fréquences de cooccurrence, puisque n'importe quelle pondération permet d'obtenir une MAP plus élevée que celle que l'on obtient sans pondération. Le choix de la pondération exerce également une influence importante sur la MAP, certaines pondérations offrant une MAP beaucoup plus élevée que d'autres.

Dans les deux langues, la pondération optimale pour les couples est simple-LL, suivie de local-MI, puis du z-score ; pour les ensembles, les deux meilleures pondérations sont simple-LL et le z-score. Ainsi, nous ne constatons pas de différences très importantes entre les langues en ce qui concerne l'influence de la pondération, comme le

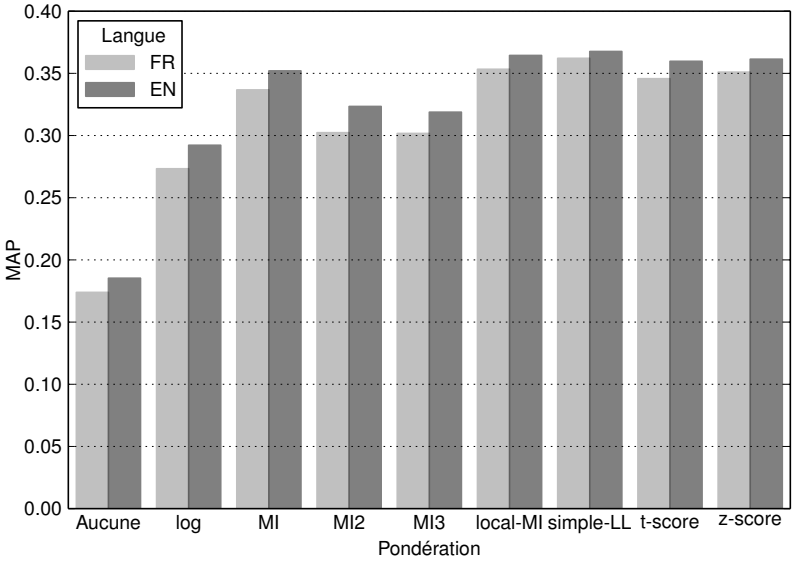


Figure 6.7 – MAP moyenne des modèles produits par AD sur les couples de référence, en fonction de la pondération.

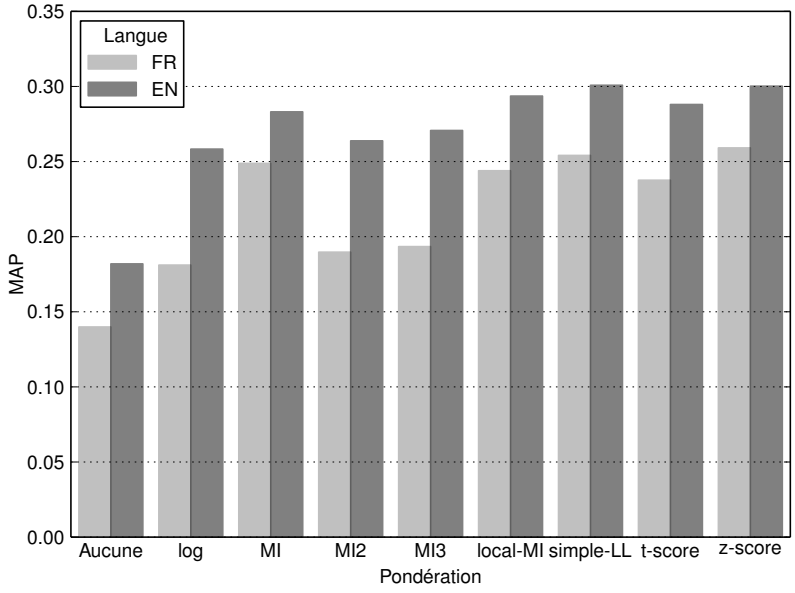


Figure 6.8 – MAP moyenne des modèles produits par AD sur les ensembles de référence, en fonction de la pondération.

montrent les Figures 6.7 et 6.8.

Le Tableau 6.IX montre la pondération qui obtient la meilleure MAP en moyenne sur chacun des sous-ensembles de couples, dans les deux langues. On observe ici que c'est souvent la pondération simple-LL qui offre les meilleurs résultats. Les trois meilleures pondérations sont souvent simple-LL, local-MI et le z-score, les différences entre celles-ci étant généralement petites. Il y a un seul sous-ensemble de couples pour lesquels une pondération différente donne des résultats clairement supérieurs : les DRV. En effet, dans le cas des DRV, l'information mutuelle (MI) produit en moyenne une MAP nettement plus élevée que les autres pondérations. L'information mutuelle accorde plus de poids aux mots-contextes dont la fréquence est faible que d'autres mesures d'association telles que simple-LL[4] ; nos résultats indiquent que cela a un effet favorable dans le cas des DRV, mais pas pour les autres relations.

Jeu	Pondération optimale (FR)	Pondération optimale (EN)
NN	simple-LL	simple-LL
VV	local-MI	simple-LL
JJ	simple-LL	z-score
QSYN	simple-LL	simple-LL
ANTI	simple-LL	z-score
HYP	simple-LL	MI^2
DRV	MI	MI
Tous	simple-LL	simple-LL

Tableau 6.IX – Pondération qui produit la MAP la plus élevée en moyenne, en fonction du sous-ensemble de couples et de la langue.

En somme, il est crucial de pondérer les fréquences de cooccurrence, et en général, simple-LL, local-MI et le z-score sont tous des bons choix, bien que l'information mutuelle soit plus souvent utilisée en sémantique distributionnelle ; cette conclusion concorde avec celles de Lapesa et al. [106]. Par contre, si on souhaite surtout détecter

[4]Cette propriété de l'information mutuelle peut être ajustée en appliquant une technique de lissage à la distribution des mots-contextes [110].

des dérivés syntaxiques, c'est l'information mutuelle qui produit les meilleurs résultats.

6.4.5 Synthèse

Nous résumons ci-dessous, pour chacun des paramètres de l'AD que nous avons pris en compte, nos observations principales au sujet de leur valeur optimale et de leur importance relative par rapport aux autres paramètres.

La valeur optimale de tous les paramètres dépend des relations lexicales que l'on souhaite identifier, la différence la plus importante étant celle entre les DRV et les autres relations lexicales paradigmatiques que nous avons prises en compte. La taille de fenêtre optimale dépend également de la PDD. En ce qui concerne l'influence de la langue traitée sur les paramétrages optimaux, nous n'avons pas observé de différences importantes entre le français et l'anglais, bien que la taille de fenêtre optimale varie un peu en fonction de ce facteur.

Taille de la fenêtre Ce paramètre est probablement celui qui exerce le plus d'influence sur la précision des voisinages que l'on obtient à partir du modèle (si on suppose qu'une pondération quelconque est appliquée aux fréquences de cooccurrence). La taille de fenêtre optimale est de 3 ou 4 mots sur les deux jeux de données principaux. La valeur optimale de ce paramètre varie un peu selon la PDD des termes ; en particulier, la précision sur les adjectifs est maximisée avec une fenêtre de 1 mot, et diminue rapidement à mesure que la taille de fenêtre augmente. Il y a également une différence importante entre les DRV et les autres relations : dans le cas des DRV, on obtient les meilleurs résultats avec une fenêtre de 8 ou 9 mots, tandis que pour les autres relations, la MAP commence à diminuer dès que la taille de fenêtre dépasse 2 ou 3 mots.

Pondération Il est très important de pondérer les fréquences de cooccurrence, et le choix de la pondération a une influence moyennement importante sur la précision des voisinages. La meilleure pondération est simple-LL en général, mais deux ou trois autres

pondérations (notamment le z-score) donnent des résultats semblables, et parfois meilleurs. La pondération optimale pour les DRV, et la plus courante en sémantique distributionnelle, à savoir l'information mutuelle, n'est pas parmi celles qui produisent les meilleurs résultats pour les autres relations paradigmatiques prises en compte.

Direction de la fenêtre Ce paramètre n'a pas une influence très importante sur la précision par rapport aux autres paramètres, sauf en ce qui concerne les DRV. La fenêtre G+D produit les meilleurs résultats sur les deux jeux de données. Elle produit des résultats beaucoup plus précis que la fenêtre G&D sur les DRV, tandis que la fenêtre G&D donne des résultats légèrement meilleurs sur les autres sous-ensembles de couples. Si les DRV font partie des relations que l'on souhaite identifier, on favoriserait une fenêtre G+D.

Forme de la fenêtre Ce paramètre a une influence moyennement importante (excepté pour les HYP) ; pour les QSYN et les ANTI, elle est plus importante que celle de la direction de la fenêtre, mais pour les DRV, la direction est plus importante. La fenêtre triangulaire donne de meilleurs résultats que la fenêtre rectangulaire sur les deux jeux de données et sur tous les sous-ensembles de couples, excepté les DRV, mais la différence dans ce cas n'est pas assez importante pour que l'on favorise une fenêtre rectangulaire, à moins que l'on s'intéresse surtout aux DRV.

6.5 Influence des hyperparamètres de word2vec

À la section 6.3, nous avons montré que l'AD offre de meilleurs résultats que W2V sur la plupart des jeux de données utilisés dans ce travail, mais que W2V détecte mieux les relations de dérivation syntaxique, et atteint une MAP maximale plus élevée que l'AD sur les ensembles de référence en anglais. Dans cette section, nous examinons l'influence des hyperparamètres de W2V, comme nous l'avons fait pour les paramètres de l'AD à la section précédente.

6.5.1 Taille de la fenêtre de contexte

La Figure 6.9 montre la MAP moyenne en fonction de la taille de la fenêtre de contexte. Sur les couples, la MAP atteint un maximum avec une taille de 7 mots. Sur les ensembles, plus la taille est élevée, plus la MAP est élevée, mais les gains que l'on obtient en élargissant la fenêtre sont très petits à partir de 8 mots environ.

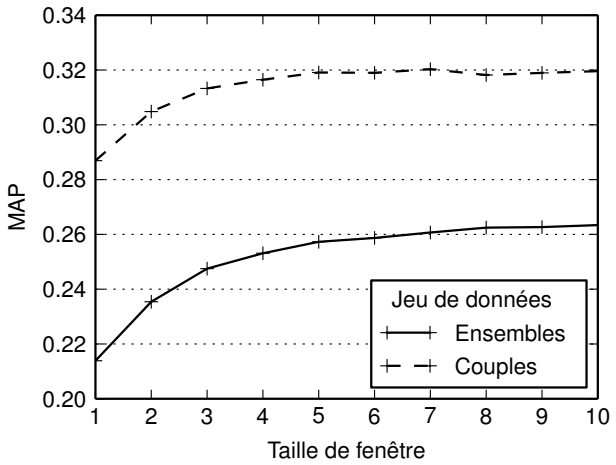


Figure 6.9 – MAP moyenne des modèles W2V en fonction de la taille de fenêtre.

Les Figures 6.10 et 6.11 présentent les résultats en fonction de la PDD et des relations ciblées respectivement. Comme pour l'AD, la valeur optimale de la taille de fenêtre varie en fonction de la PDD, et les tailles optimales pour les trois PDD sont semblables à celles que nous avons observées dans le cas de l'AD. En effet, la taille optimale est 2 ou 3 pour les noms, 3 pour les verbes et 1 pour les adjectifs. Ainsi, pour tous les couples où les deux termes ont la même PDD, la taille optimale ne dépasse pas 3 mots.

La valeur optimale de la taille de fenêtre varie également en fonction du type de relation lexicale que l'on souhaite identifier : la taille optimale est 1 pour les ANTI, 2 pour les QSYN, entre 3 et 6 pour les HYP et 10 (ou plus) pour les DRV. Comme dans le cas de l'AD (voir section 6.4.1), la MAP moyenne sur les DRV ne semble pas avoir atteint son maximum même avec une de fenêtre de 10 mots, mais soulignons que la MAP la plus élevée que nous avons observée sur les DRV a été atteinte avec une fenêtre

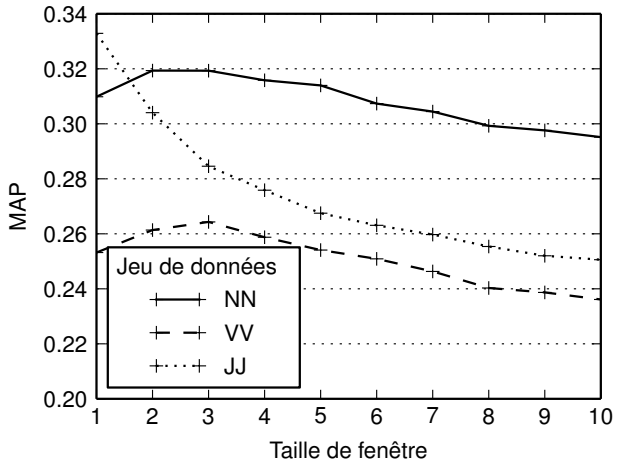


Figure 6.10 – MAP moyenne des modèles W2V en fonction de la taille de fenêtre et de la PDD des couples.

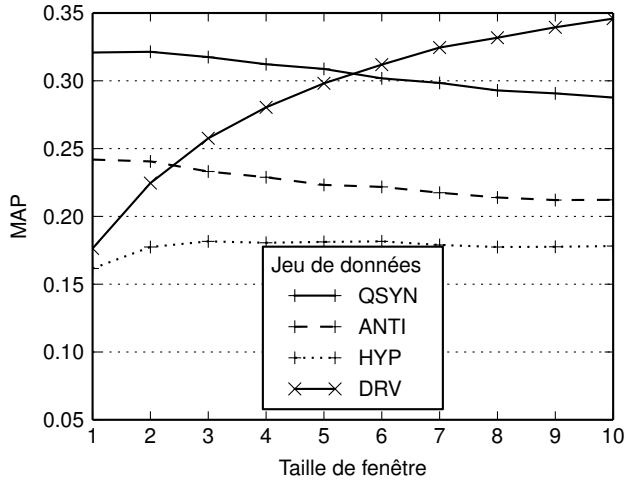


Figure 6.11 – MAP moyenne des modèles W2V en fonction de la taille de fenêtre et des relations ciblées.

de 9 mots (dans les deux langues). Et comme dans le cas de l'AD, les tailles de fenêtre optimales pour les DRV et pour les QSYN sont très différentes. Si on utilise la valeur optimale pour les QSYN (2 mots), la précision sur les DRV est beaucoup plus faible que si on utilisait une fenêtre plus large. Si on se fie aux résultats obtenus sur les ensembles (Figure 6.9), on obtient le meilleur compromis en utilisant une fenêtre large (8-10 mots).

Ainsi, la taille de fenêtre exerce une influence importante sur la MAP, particulièrement dans le cas des DRV et des adjectifs.

L'influence de cet hyperparamètre est semblable dans les deux langues, mais on observe des différences quant à la valeur optimale :

- Couples (tous) : la taille optimale est 5 ou 6 en français, entre 7 et 10 en anglais.
- NN : la taille optimale est 2 en français, 3 en anglais.
- VV : la taille optimale est 3 en français, 1 en anglais.
- QSYN : la taille optimale est 2 en français, 1 en anglais.
- ANTI : la taille optimale est 2 en français, 1 en anglais.
- HYP : la taille optimale est entre 3 et 5 mots en français, entre 6 et 10 en anglais.

6.5.2 Dimension des représentations lexicales

La dimension des représentations lexicales est le nombre de composantes que comportent les vecteurs qui représentent chacun des mots-cibles.

Le Tableau 6.X montre l'influence de cet hyperparamètre sur la MAP, calculée sur les deux jeux de données, et le Tableau 6.XI montre les résultats que l'on obtient sur chacun des sous-ensembles de couples. La dimension plus élevée (300) donne toujours de meilleurs résultats, sauf sur les ANTI, pour lesquels la MAP moyenne est un peu plus élevée lorsqu'on utilise une dimension de 100. L'influence de cet hyperparamètre

Dimension	Couples	Ensembles
100	0.2943	0.2337
300	**0.3330**	**0.2693**

Tableau 6.X – MAP moyenne des modèles W2V en fonction de la dimension des représentations lexicales, sur les deux jeux de données.

est moyennement importante, le gain que l'on obtient en utilisant une dimension de 300 se situant généralement autour de 2 ou 3 points.

Notons également que l'influence de cet hyperparamètre est la même dans les deux langues, sauf en ce qui concerne les ANTI : en moyenne, les deux dimensions produisent le même résultat en anglais ; en français, la faible dimension donne des résultats légèrement meilleurs.

Dimension	NN	VV	JJ	QSYN	ANTI	HYP	DRV
100	0.2942	0.2413	0.2673	0.2907	**0.2281**	0.1686	0.2604
300	**0.3222**	**0.2595**	**0.2818**	**0.3198**	0.2210	**0.1866**	**0.3177**

Tableau 6.XI – MAP moyenne des modèles W2V en fonction de la dimension des représentations lexicales, sur chaque sous-ensemble de couples.

6.5.3 Sous-échantillonnage des mots fréquents

Cette fonction, qui vise à diminuer l'influence des mots très fréquents lors de l'entraînement du modèle, prend un seuil de fréquence ; les mots dont la fréquence relative est supérieure à ce seuil sont sous-échantillonnés en supprimant[5] certaines de leurs occurrences du corpus avant l'entraînement du modèle. Chacune de leurs occurrences a une certaine probabilité d'être supprimée, cette probabilité augmentant en fonction de la fréquence du mot. Plus le seuil de fréquence est bas, plus le nombre de mots supprimés est élevé. Nous avons testé un seuil faible (10^{-5}) et un seuil élevé (10^{-3}), en plus de vérifier les résultats que l'on obtient sans sous-échantillonnage.

[5]Le fait que certains mots soient supprimés du corpus a pour effet d'augmenter, dans les contextes où des mots ont été supprimés, la taille effective de la fenêtre de contexte [110].

Cette fonction aurait un effet semblable à celui d'une technique souvent utilisée dans le cadre de l'AD, et que nous avons utilisée dans ce travail, qui consiste à exclure les mots vides des mots-contextes [110], du moins si on utilise un seuil élevé ; elle offrirait une amélioration de la qualité des résultats selon Baroni et al. [10].

Seuil	Couples	Ensembles
aucun	0.2947	0.2248
faible	0.3190	**0.2700**
élevé	**0.3273**	0.2597

Tableau 6.XII – MAP moyenne des modèles W2V en fonction du seuil pour le sous-échantillonnage, sur les deux jeux de données.

Le Tableau 6.XII montre la MAP moyenne en fonction du seuil pour le sous-échantillonnage des mots fréquents. Ces résultats indiquent que, en moyenne, les résultats sont effectivement meilleurs lorsqu'on applique la fonction de sous-échantillonnage, mais la valeur optimale du seuil dépend du jeux de données.

Seuil	NN	VV	JJ	QSYN	ANTI	HYP	DRV
aucun	0.3241	0.2691	**0.3145**	0.3295	**0.2451**	0.1760	0.1785
faible	0.2705	0.2062	0.2203	0.2562	0.1945	0.1705	**0.4141**
élevé	**0.3301**	**0.2759**	0.2889	**0.3300**	0.2339	**0.1862**	0.2746

Tableau 6.XIII – MAP moyenne des modèles W2V en fonction du seuil pour le sous-échantillonnage, sur chaque sous-ensemble de couples.

Le Tableau 6.XIII montre les résultats obtenus sur chaque sous-ensemble de couples. On observe ici que sur la plupart des sous-ensembles, on n'améliore pas beaucoup (ou pas du tout) les résultats en appliquant le sous-échantillonnage, sauf sur les DRV. De plus, un seuil faible (10^{-5}), qui produit un sous-échantillonnage plus intensif, donne des résultats beaucoup moins précis que les deux autres valeurs de cet hyperparamètre sur tous les sous-ensembles sauf les HYP et les DRV ; au contraire, sur les DRV, un sous-échantillonnage intensif donne une précision beaucoup plus élevée. Encore une fois, la valeur optimale pour les DRV est différente de la valeur optimale pour tous les autres

sous-ensembles de couples. En outre, la MAP que l'on obtient sur les DRV est très sensible à cet hyperparamètre : on obtient une MAP plus de deux fois plus élevée en utilisant la fonction de sous-échantillonnage (de manière intensive) ; ce gain est d'environ 80% en anglais et 250% en français.

Ainsi, quand on observe les résultats obtenus sur tous les couples, le fait que le seuil faible (sous-échantillonnage intensif) donne une MAP plus élevée que le seuil de 0 (aucun sous-échantillonnage) est un peu trompeur : ce seuil donne de très bons résultats sur les DRV, mais diminue considérablement la qualité des résultats sur les autres couples (sauf les HYP). Et si le sous-échantillonnage intensif produit les meilleurs résultats sur les ensembles de référence, c'est sans doute attribuable à l'amélioration des résultats sur les DRV.

L'influence de cet hyperparamètre varie un peu selon la langue, car la valeur optimale est différente sur trois des sous-ensembles de couples : sur les NN, les VV et les QSYN, la valeur optimale est le seuil élevé (10^{-3}) pour le français, mais 0 pour l'anglais. Ainsi, pour l'anglais, sur 5 des 7 sous-ensembles de couples, les meilleurs résultats sont obtenus sans appliquer de sous-échantillonnage, mais on obtient des résultats légèrement plus précis sur les HYP avec un seuil élevé (sous-échantillonnage modéré) et beaucoup plus précis sur les DRV avec un seuil faible (sous-échantillonnage intensif). Sur l'ensemble des couples et sur les ensembles de référence, les valeurs optimales sont les mêmes dans les deux langues.

En somme, la fonction de sous-échantillonnage ne semble pas avoir un effet très favorable sur la qualité des résultats[6], à moins que les relations que l'on cherche à identifier comprennent les dérivés syntaxiques. Cet hyperparamètre exerce une influence très importante sur la MAP, et il est important de fixer le seuil avec soin.

[6]Rappelons que nous avons seulement testé deux valeurs du seuil, celles-ci étant les bornes de l'intervalle des valeurs recommandées.

6.5.4 Architecture

Deux architectures neuronales différentes sont implémentées dans `word2vec` : *continuous skip-gram* et *continuous bag-of-words* (CBOW). Pour simplifier, on peut dire que l'architecture CBOW vise à prédire chaque mot à partir de son contexte (les mots dans la fenêtre de contexte) tandis que l'architecture skip-gram fait le contraire : elle prédit le contexte à partir du mot.

Architecture	Couples	Ensembles
CBOW	0.3068	0.2424
Skip-gram	**0.3206**	**0.2606**

Tableau 6.XIV – MAP moyenne des modèles W2V en fonction de l'architecture, sur les deux jeux de données.

L'influence de l'architecture est présentée dans les Tableaux 6.XIV et 6.XV. La plus grande différence de précision que l'on observe entre les deux architectures est sur les DRV, pour lesquels l'architecture skip-gram offre des résultats beaucoup plus précis. Pour cette raison, cette architecture donne globalement les meilleurs résultats, bien que CBOW produise une MAP plus élevée (de 1 ou 2 points) sur 4 des sous-ensembles de couples. En ce qui concerne les ANTI et les HYP, les deux architectures donnent des résultats très semblables.

Architecture	NN	VV	JJ	QSYN	ANTI	HYP	DRV
CBOW	**0.3158**	**0.2607**	**0.2805**	**0.3165**	0.2236	0.1759	0.2465
Skip-gram	0.3006	0.2401	0.2686	0.2939	**0.2254**	**0.1793**	**0.3316**

Tableau 6.XV – MAP moyenne des modèles W2V en fonction de l'architecture, sur chaque sous-ensemble de couples.

Notons que l'influence du choix de l'architecture est très similaire dans les deux langues. La valeur optimale de cet hyperparamètre est la même dans les deux langues sur tous les jeux de données sauf les ANTI, pour lesquels l'architecture skip-gram donne

des résultats supérieurs en français, mais inférieurs en anglais, la différence étant petite (moins d'un point) dans les deux cas.

6.5.5 Algorithme d'entraînement

Deux algorithmes peuvent être utilisés pour entraîner un modèle avec `word2vec` (quelle que soit l'architecture utilisée) : l'échantillonnage d'exemples négatifs ou un *softmax* hiérarchique. L'échantillonnage d'exemples négatifs offrirait de meilleurs résultats selon Baroni et al. [10]. Si on utilise cet algorithme, il faut aussi choisir le nombre d'exemples négatifs qui sont échantillonnés lors de l'entraînement. Dans le cadre de ce travail, nous avons testé trois valeurs de cet hyperparamètre : 10, 5 et 0 (dans ce cas, on utilise le *softmax* hiérarchique).

Nb exemples	Couples	Ensembles
0	0.2998	0.2469
5	0.3171	0.2508
10	**0.3241**	**0.2568**

Tableau 6.XVI – MAP moyenne des modèles W2V en fonction du nombre d'exemples négatifs, sur les deux jeux de données.

Les Tableaux 6.XVI et 6.XVII montrent la MAP moyenne en fonction du nombre d'exemples négatifs utilisés pour l'entraînement du modèle. Ces résultats indiquent que l'échantillonnage d'exemples négatifs produit de meilleurs résultats que le *softmax* hiérarchique sur tous les jeux de données, le choix de l'algorithme d'entraînement ayant une influence moyennement importante sur la MAP, mais moins importante que celle

Nb exemples	NN	VV	JJ	QSYN	ANTI	HYP	DRV
0	0.2926	0.2353	0.2695	0.2890	0.2192	0.1728	0.2808
5	0.3139	0.2578	0.2742	0.3110	0.2264	**0.1803**	0.2876
10	**0.3182**	**0.2581**	**0.2800**	**0.3157**	**0.2280**	0.1797	**0.2988**

Tableau 6.XVII – MAP moyenne des modèles W2V en fonction du nombre d'exemples négatifs, sur chacun des sous-ensembles de couples.

d'autres hyperparamètres. De plus, on obtient généralement une amélioration en utilisant 10 exemples plutôt que 5, mais ce gain n'est pas très important.

L'influence de cet hyperparamètre est très similaire dans les deux langues : la valeur optimale est 10 exemples sur tous les jeux de données sauf les VV et les HYP en français ; sur ces deux jeux, on obtient des résultats légèrement supérieurs en utilisant 5 exemples plutôt que 10. À titre indicatif, soulignons que sur trois des jeux (les ensembles et les DRV en français, les JJ en anglais), le *softmax* hiérarchique donne des résultats légèrement supérieurs à ceux que l'on obtient en échantillonnant 5 exemples négatifs, mais inférieurs à ceux que l'on obtient avec 10 exemples négatifs.

6.5.6 Synthèse

Nous résumons ci-dessous, pour chacun des hyperparamètres de `word2vec` que nous avons pris en compte, nos observations principales au sujet de leur valeur optimale et de leur importance relative.

Tous les hyperparamètres ont une influence identique ou très similaire dans les deux langues traitées ; nous avons observé quelques différences quant à leurs valeurs optimales, mais ces différences sont peu importantes.

En ce qui concerne la PDD des requêtes et les relations lexicales que l'on souhaite détecter, l'influence de deux des hyperparamètres ne semble pas dépendre de ces facteurs : la dimension des représentations et l'algorithme d'entraînement.

L'influence des trois autres hyperparamètres varie beaucoup en fonction des relations ciblées, la différence la plus importante étant celle entre les DRV et les autres relations. L'influence de deux des hyperparamètres varie (mais dans une moindre mesure) selon la PDD des requêtes : la taille de la fenêtre de contexte et le seuil pour le sous-échantillonnage.

Seuil pour le sous-échantillonnage de mots fréquents Cet hyperparamètre a une influence très importante sur la précision des voisinages que l'on obtient à partir du mo-

dèle. La valeur optimale varie un peu en fonction de la PDD, mais surtout en fonction des relations ciblées, l'influence de cet hyperparamètre étant très différente pour les DRV que pour les autres relations. Sur presque tous les sous-ensembles de couples, la valeur optimale est soit 0 (aucun sous-échantillonnage), soit un seuil élevé (sous-échantillonnage modéré), et le gain que l'on obtient en appliquant le sous-échantillonnage dans ces cas est généralement petit ; la précision est beaucoup moins élevée si on utilise un seuil faible (sous-échantillonnage intensif). En revanche, pour les DRV, c'est le seuil faible qui produit les meilleurs résultats, la précision étant beaucoup plus élevée que si on ne fait pas de sous-échantillonnage.

Taille de la fenêtre Cet hyperparamètre a une influence importante dans certains cas, mais peu importante dans d'autres (p. ex. pour les HYP). La valeur optimale varie en fonction de la relation : sur les DRV, la MAP augmente rapidement à mesure que la taille de fenêtre augmente ; en revanche, sur les QSYN et les ANTI, elle atteint un maximum avec une fenêtre de 1 ou 2 mots selon la langue, puis diminue lentement à mesure que la taille de fenêtre augmente. La valeur optimale dépend également de la PDD, mais dans une moindre mesure : comme dans le cas de l'AD, la taille optimale pour les adjectifs est 1, et la MAP diminue rapidement à mesure que la taille augmente ; la taille optimale est 2 ou 3 pour les noms selon la langue, et 1-3 pour les verbes.

Architecture L'architecture optimale dépend des relations ciblées : skip-gram donne des résultats largement supérieurs pour les DRV, tandis que c'est CBOW qui produit les meilleurs résultats pour les QSYN. Si les relations que l'on souhaite détecter comprennent les DRV, on favoriserait l'architecture skip-gram.

Algorithme d'entraînement L'entraînement par échantillonnage d'exemples négatifs avec 10 exemples donne toujours de meilleurs résultats que le *softmax* hiérarchique. La qualité des résultats que l'on obtient en échantillonnant 5 exemples plutôt que 10 est généralement plus faible, mais la différence n'est pas très importante.

Dimension des représentations lexicales La MAP est toujours plus élevée lorsqu'on utilise une dimension de 300 plutôt que 100, sauf pour les ANTI.

6.6 Discussion

Ci-dessous, nous discutons les résultats présentés dans les sections précédentes. Nous résumons d'abord ce que ces résultats indiquent quant à divers facteurs qui expliquent éventuellement la qualité des résultats qu'offrent les modèles sémantiques distributionnels et l'influence de leurs (hyper)paramètres, ces facteurs étant liés à l'application envisagée, en l'occurrence une description terminologique basée sur un cadre descriptif particulier. Nous explorons ensuite plusieurs sources potentielles d'erreurs, c'est-à-dire des facteurs qui peuvent expliquer la faible précision de certains voisinages distributionnels. Enfin, nous discutons la mesure d'évaluation que nous avons utilisée afin de faciliter l'interprétation de la qualité des résultats.

6.6.1 Influence des facteurs liés au projet terminologique

L'expérience que nous avons réalisée tient compte de plusieurs facteurs qui peuvent influencer la qualité des résultats qu'offrent les modèles distributionnels et la façon dont cette qualité varie en fonction de leur paramétrage. Dans cette section, nous résumons nos observations quant à l'influence de quatre aspects du projet terminologique : la langue traitée, la PDD des requêtes, le cadre descriptif adopté et les relations ciblées.

La langue traitée Nos résultats suggèrent que la langue traitée n'a pas une influence très importante sur la qualité des résultats ni sur les paramétrages optimaux des modèles distributionnels, du moins dans le cas des langues que nous avons prises en compte, l'anglais et le français. On obtient une MAP semblable dans les deux langues sur les couples et les ensembles de référence (voir Tableau 6.III), la différence entre les deux langues étant limitée à quelques points de MAP, bien que des différences plus importantes aient été observées sur certains sous-ensembles de couples, tels que les ANTI et les DRV (voir

Tableau 6.IV). De plus, nous avons observé très peu de différences entre les deux langues quant au paramétrage optimal des modèles distributionnels.

La partie du discours Nos résultats suggèrent que la PDD des requêtes a une influence plus importante que la langue traitée sur la qualité des résultats et sur les paramétrages optimaux. En effet, on modélise mieux les adjectifs et les noms que les verbes, quelle que soit la langue traitée (voir Tableau 6.IV). De plus, la valeur optimale de certains (hyper)paramètres dépend de ce facteur dans une certaine mesure, notamment la taille de la fenêtre de contexte, tant pour l'AD que pour W2V (voir Figures 6.5 et 6.10 respectivement), et le seuil pour le sous-échantillonnage de mots fréquents dans le cas de W2V (voir Tableau 6.XIII).

Les relations ciblées Les relations ciblées ont une influence très importante sur la qualité des résultats et sur les paramétrages optimaux, la différence la plus importante à cet égard étant celle entre la dérivation syntaxique et les relations paradigmatiques classiques telles que la (quasi-)synonymie. Nous avons observé que la (quasi-)synonymie et la dérivation syntaxique sont les relations les mieux captées et que les relations les moins bien captées sont les liens hiérarchiques (hyperonymie et hyponymie). C'est sur les dérivés que l'on atteint la MAP la plus élevée, mais c'est aussi sur cette relation que la qualité des résultats varie le plus (voir Figure 6.2), ceux-ci étant particulièrement sensibles au paramétrage du modèle. De plus, l'influence des (hyper)paramètres sur la précision que l'on obtient pour les dérivés est souvent très différente de leur influence sur celle que l'on obtient pour les quasi-synonymes ; c'est le cas pour tous les paramètres de l'AD que nous avons pris en compte, et 3 des 5 hyperparamètres de W2V (à l'exception de la dimension des représentations lexicales et de l'algorithme d'entraînement).

Le cadre descriptif Le cadre descriptif adopté, un facteur étroitement relié aux relations ciblées, a également une influence importante sur la qualité des résultats. Dans ce travail, nous avons pris en compte deux cadres descriptifs : l'approche lexico-sémantique

à la terminologie et la sémantique des cadres. Les résultats que nous avons obtenus sur les couples de référence et les ensembles de référence, qui reflètent respectivement ces deux cadres descriptifs, indiquent que l'on arrive à mieux détecter certaines relations lexicales spécifiques, à savoir la quasi-synonymie et la dérivation syntaxique, que l'appartenance au même cadre sémantique. En effet, les résultats que nous avons obtenus sur les couples de référence sont meilleurs que ceux obtenus sur les ensembles de référence (voir Tableau 6.I), et cette différence est encore plus importante si on tient seulement compte des relations que l'on détecte le mieux, c'est-à-dire les QSYN et les DRV (voir Tableau 6.II). Cela est attribuable, du moins en partie, au fait que les termes qui évoquent le même cadre sémantique comprennent à la fois des QSYN et des DRV (voir Tableau 5.VI), ce qui les rend plus difficiles à détecter que les QSYN d'une part et les DRV d'autre part, puisque les paramétrages optimaux pour ces deux relations sont très différents. Soulignons que la MAP relativement faible sur les ensembles de référence est également attribuable à la proportion élevée de verbes qu'ils contiennent.

6.6.2 Sources d'erreurs

Dans cette section, nous explorons divers facteurs qui expliquent éventuellement pourquoi les voisins distributionnels d'un terme ne correspondent pas toujours aux voisins sémantiques de ce terme selon les données de référence. Cette question pourrait en fait être formulée de deux façons différentes (selon qu'on l'aborde du point de vue de la précision ou du rappel), à savoir :

1. Pourquoi les termes reliés sémantiquement selon les données de référence ne sont-ils pas toujours similaires selon le modèle distributionnel ?

2. Pourquoi les voisins distributionnels d'un terme comprennent-ils des mots qui ne sont pas reliés sémantiquement à ce terme ?

Nous avons montré à la section 6.3 que l'AD produit de meilleurs résultats que W2V sur la plupart des jeux de données utilisés dans ce travail, les DRV étant l'exception

la plus notable. Nous avons ensuite analysé l'influence des paramètres de l'AD à la section 6.4. Cette analyse suggère qu'un modèle construit au moyen d'une fenêtre G+D triangulaire de 3 mots et pondéré au moyen de simple-LL offrirait de bons résultats sur la plupart des jeux de données de référence (quoique l'on pourrait augmenter la précision sur les DRV en optant pour une fenêtre plus large et de forme rectangulaire, ainsi qu'une pondération différente).

En français, ce modèle atteint une MAP de 0.3979 sur les couples de référence, ce qui indique que la précision moyenne (AP) que l'on obtient pour chacune des requêtes (les entrées dans les couples de référence) est de 0.3979 en moyenne.

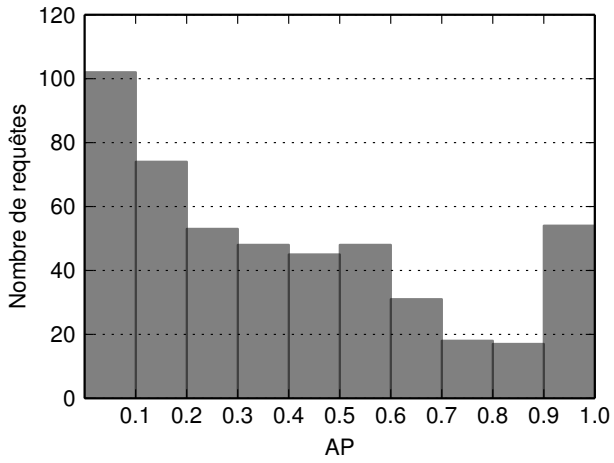


Figure 6.12 – Répartition des précisions moyennes (AP) obtenues au moyen d'un modèle produit par AD.

La répartition des précisions moyennes (AP) obtenues avec ce modèle pour chacune des requêtes est illustrée dans la Figure 6.12. Cette figure montre que l'AP est entre 0 et 0.1 pour une centaine de requêtes, entre 0.1 et 0.2 pour un peu plus de 70 requêtes, et ainsi de suite. Afin d'identifier des sources d'erreurs, nous analyserons notamment les requêtes pour lesquelles l'AP est la plus près de 0, en observant leurs PPV, leurs termes reliés et leurs cooccurrents.

L'analyse présentée dans cette section porte surtout sur le modèle décrit ci-dessus, produit par AD, mais nous ferons également appel à un modèle produit par W2V à quelques occasions, pour vérifier si nos observations s'appliquent également à ce modèle. Le modèle W2V que nous avons retenu, en nous basant sur l'analyse présentée à la section 6.5 de l'influence des hyperparamètres de W2V[7], est paramétré de la façon suivante : fenêtre de 6 mots, architecture skip-gram entraînée par échantillonnage de 10 exemples négatifs, représentations de dimension 300, sous-échantillonnage de mots fréquents avec seuil élevé (10^{-3}). Ce modèle obtient une MAP d'environ[8] 0.3752 sur les couples de référence en français.

Une méthode simple pour explorer les sources d'erreurs consiste à comparer, pour les requêtes dont le voisinage a une faible précision, les termes reliés à ces requêtes selon les données de référence et leurs voisins selon le modèle distributionnel. Les 15 requêtes pour lesquelles la précision moyenne (AP), calculée à partir du modèle retenu (produit par AD), est la plus près de 0 sont présentées dans le Tableau 6.XVIII, en ordre décroissant de précision. Dans ce tableau, chacune de ces requêtes est accompagnée de ses 10 PPV et de ses termes reliés selon les couples de référence. Dans l'analyse d'erreurs présentée dans cette section, nous ferons référence à ce tableau à plusieurs reprises.

[7]Plus précisément, pour chaque hyperparamètre, nous avons utilisé la valeur qui maximise, en moyenne, la MAP sur les couples de référence en français.

[8]L'initialisation aléatoire du modèle (et la fonction de sous-échantillonnage s'il y a lieu) fait en sorte que, pour un paramétrage donné, les résultats que l'on obtient varient un peu chaque fois que l'on entraîne un modèle. Lors de l'expérience que nous avons réalisée, ce paramétrage a obtenu une MAP de 0.3752. Lorsque nous avons entraîné de nouveau un modèle pour l'analyse d'erreurs, en utilisant le même paramétrage, nous avons obtenu une MAP légèrement supérieure, de 0.3796.

Requête	PPV	Termes reliés
voyageur	passager, visiteur, touriste, fret, ferroviaire, personne, naturaliste, gens, marchandise, voyage, ...	utilisateur
flotte	navire, port, bateau, brise-glace, pêche, baleinier, navigation, propulsion, plaisance, compagnie, ...	parc
delta	embouchure, fleuve, bassin, vallée, plaine, estuaire, saint-laurent, rive, lacs, île, ...	écosystème
stable	homogène, constant, abondant, inchangé, élever, chaud, incertain, instable, dense, préoccupant, ...	variable
prévision	prédiction, estimation, projection, scénario, simulation, hypothèse, prédire, donnée, giec, évaluation, ...	prévoir
industriel	agricole, industrie, agroalimentaire, commercial, agro-alimentaire, minier, secteur, artisanal, déchet, urbain, ...	naturel
conduire	entraîner, mener, aboutir, provoquer, déboucher, traduire, impliquer, entrainer, amener, accélérer, ...	conduite
hydrologique	hydrogéologique, météorologique, climatologique, hydrique, pédologique, biogéochimiques, topographique, climatique, biophysique, hydrographique, ...	eau
plage	côte, littoral, île, rivage, falaise, dune, rivière, lac, estuaire, sable, ...	écosystème
reproduire	trouver, déplacer, disparaître, observer, nourrir, nicher, voir, vivre, retrouver, survivre, ...	simuler
prévoir	mentionner, fixer, viser, relatif, définir, applicable, soumettre, énoncer, prescrire, application, ...	prévision, prédire
développé	industrialiser, membres, développer, signataire, coopérer, émergent, ue, membre, scandinave, africains, ...	riche
parc	réserve, national, aire, pnr, espace, territoire, sanctuaire, situer, jasper, centre, ...	flotte
embarquer	bateau, équipage, tara, bord, transporter, navire, équiper, acheminement, installer, capitaine, ...	intégrer
plein	beau, durant, pendant, froid, coeur, quand, passer, premier, plonger, tout, ...	recharge, charge

Tableau 6.XVIII – PPV et termes reliés des 15 requêtes pour lesquelles l'AP est la plus faible (modèle produit par AD).

Nous avons également observé les 15 requêtes pour lesquelles la précision est la plus faible lorsqu'on utilise le modèle W2V que nous avons retenu pour cette analyse (plutôt que le modèle produit par AD). Ces requêtes, leurs PPV et leurs termes reliés sont présentés dans le Tableau 6.XIX. Il est intéressant de noter que la liste de requêtes est similaire à celle obtenue au moyen du modèle produit par AD, 9 des 15 requêtes apparaissant dans les deux listes. De plus, lorsqu'on compare ce tableau au Tableau 6.XVIII, on observe que les listes de PPV des requêtes qui apparaissent dans les deux tableaux ont un degré plutôt élevé de recoupement, ces requêtes ayant entre 3 et 6 voisins en commun parmi leurs 10 PPV, à l'exception de *plage*, dont les deux voisinages ont seulement un mot en commun. Il semble donc que les voisinages que l'on obtient avec les deux modèles soient similaires. Une analyse systématique des différences entre les voisinages permettrait éventuellement d'identifier les facteurs pouvant expliquer ces différences, mais cela n'entre pas dans les objectifs de cette thèse.

6.6.2.1 La couverture des données de référence

Une conclusion que l'on peut tirer en observant les données présentées dans le Tableau 6.XVIII est que la précision des voisinages est souvent plus élevée que le suggère notre procédure d'évaluation automatique basée sur des données de référence. Il est évident que les 10 PPV de chacune des requêtes présentées dans le tableau comprennent beaucoup de termes qui leur sont reliés sémantiquement, mais la précision moyenne (AP) de leur voisinage est très faible (presque nulle) parce que les rangs de leurs termes reliés parmi leurs PPV sont très élevés.

En effet, toutes les listes de PPV présentées dans le Tableau 6.XVIII contiennent des termes dont le sens (ou un des sens) est paradigmatiquement relié au sens (ou à un des sens) de la requête correspondante, la proportion des PPV qui sont reliés à la requête étant généralement élevée, exception faite de la requête *plein*. Ainsi, dans certains cas, si la précision moyenne (AP) des PPV d'une requête est faible, ce n'est pas forcément parce que son voisinage immédiat contient beaucoup de bruit, et ce n'est pas seulement parce

Requête	PPV	Termes reliés
station	épuration, step, ski, pompage, fosse, balnéaire, échantillon, effluent, résiduaire, septique, ...	infrastructure, borne
substance	toxique, composé, biocide, chimique, micropolluant, métabolite, mutagène, polluant, volatil, contaminants, ...	matière
alternatif	alternative, solution, innovateur, envisageable, encouragement, respectueux, mode, innovants, innovantes, promouvoir, ...	traditionnel, continu
ligne	ligner, url, tramway, lgv, tracé, internaute, twitter, distance, téléchargement, électronique, ...	parcours, itinéraire, trajet, infrastructure
activité	pratique, action, exploitation, extractif, manufacturier, aquacole, touristique, secteur, opération, récréatif, ...	agriculture
reproduire	survivre, nicher, nourrir, bernache, reproduction, esturgeon, descendance, reconstituer, pondre, hiverner, ...	simuler
bois	grume, sciage, sciure, copeau, bûche, ligneux, bois-énergie, scierie, résineux, charpente, ...	ressource, matière
flotte	navire, bateau, brise-glace, embarcation, patrouille, baleinier, véhicule, plaisance, cargaison, avion, ...	parc
parc	ecrins, mercantour, vanoise, jasper, park, iroise, écrin, pnr, aire, yellowstone, ...	flotte
delta	mackenzie, fleuve, danube, mékong, estuaire, beaufort, embouchure, golfe, aral, lagune, ...	écosystème
industriel	industrie, minier, agroalimentaire, agro-alimentaire, commercial, technologique, entreprise, transformateur, manufacturier, aquacole, ...	naturel
embarquer	voilier, bateau, débarquer, embarcation, acheminer, navire, bord, tara, hélicoptère, pirogue, ...	intégrer
plein	fouet, capsule, lune, dedans, vigoureux, tout, gagner, relancer, pique-nique, promener, ...	recharge, charge
voyageur	fret, automobiliste, ferroviaire, autobus, passager, marchandise, gare, activiste, naturaliste, voyager, ...	utilisateur
plage	rivage, dune, côte, vasière, estran, sable, littoral, falaise, lagune, sableux, ...	écosystème

Tableau 6.XIX – PPV et termes reliés des 15 requêtes pour lesquelles l'AP est la plus faible (modèle W2V).

que ses termes reliés ont un rang élevé parmi ses PPV : c'est aussi parce que certains de ses PPV qui lui sont effectivement reliés sont absents des données de référence ; nous avons souligné cette lacune des méthodes d'évaluation intrinsèques à la section 3.3.1. Dans le cadre de l'analyse des résultats présentée dans ce chapitre et le suivant, il faut donc se rappeler que les procédures d'évaluation automatiques que nous utilisons tendent à sous-estimer la précision en raison du fait que les données de référence dont nous disposons n'offrent pas une couverture complète des relations lexicales paradigmatiques qui existent effectivement entre les mots-cibles. Une évaluation manuelle systématique des voisinages nous indiquerait sans doute que leur précision est plus élevée que le suggèrent les procédures d'évaluation automatiques basées sur des données de référence ; c'est effectivement ce que nous avons observé lorsque nous avons réalisé une évaluation manuelle de cette sorte [15].

6.6.2.2 La polysémie

Si la précision des voisinages distributionnels est plus élevée que le suggère l'évaluation automatique, il n'en demeure pas moins que les termes reliés que l'on souhaiterait détecter ont parfois un rang très élevé parmi les PPV des requêtes correspondantes.

Un des facteurs qui expliquent la similarité distributionnelle faible de certaines paires de termes paradigmatiquement reliés est la polysémie. On peut illustrer ce phénomène en comparant les PPV de certains termes sémantiquement reliés mais dont la similarité distributionnelle est faible. Prenons, par exemple, la requête *prévoir* (voir Tableau 6.XVIII). Certains PPV de cette requête (*fixer, définir, prescrire*, etc.) sont reliés à un de ses sens (« disposer pour l'avenir »), tandis que ses termes reliés selon la référence (*prédire* et *prévision*) ont un sens proche d'un autre sens de *prévoir* (« déterminer qu'une chose se produira à l'avenir »). Puisque *prévision* fait aussi partie des 15 requêtes pour lesquelles la précision est la plus faible, on peut voir dans le Tableau 6.XVIII que ses PPV (*prédiction, estimation, projection*, etc.) correspondent à ce deuxième sens ; on observe d'ailleurs que *prévision*, *prédiction* et *prédire* sont très proches selon le modèle distribu-

tionnel, mais que *prévoir* est éloigné de ces mots. Ainsi, les voisinages distributionnels des termes *prévoir* et *prédire* sont éloignés parce que le modèle capte un sens différent de *prévoir* que celui correspondant à *prédire*, sans doute parce que c'est ce sens de *prévoir* qui est utilisé le plus fréquemment dans le corpus. Les requêtes *conduire*, *parc* et *reproduire* illustrent également des voisinages correspondant à des sens différents de ceux des termes reliés correspondants.

Il est simple de vérifier pourquoi deux termes ont une similarité distributionnelle faible (ou élevée), en comparant les cooccurrents les plus fortement associés aux deux termes ; plus les deux termes partagent des cooccurrents forts, plus ils sont distributionnellement similaires (et plus ils ont tendance à être sémantiquement reliés). Un modèle produit par AD rend cette analyse très simple à réaliser, puisqu'on peut obtenir les cooccurrents les plus discriminants d'un terme (plus précisément, ses cooccurrents les plus discriminants parmi les mots-contextes du modèle) en cherchant les valeurs les plus élevées dans le vecteur (la représentation) de ce terme[9]. Reprenons l'exemple de la requête *prévoir*. Selon le modèle que nous avons retenu, les cooccurrents les plus fortement associés à ce terme sont :

- *prévoir : article, condition, procédure, modalité, disposition, décret, amende, cas, mesure, peine, ...*

Ces cooccurrents correspondent tous au sens juridique ou administratif de *prévoir*, ce qui explique pourquoi ses PPV (voir Tableau 6.XVIII) ont tous un sens proche de celui-ci. En revanche, les cooccurrents les plus discriminants du terme *prévision* sont :

- *prévision : météorologique, modèle, météo, rendement, climatique, selon, alarmiste, crue, pessimiste, saisonnier, ...*

Ces cooccurrents correspondent à un autre sens de *prévoir* (« déterminer qu'une chose se produira à l'avenir »). On voit ainsi pourquoi *prévoir* et *prévision* ont une si-

[9]Nous supposons ici qu'aucune réduction de dimension n'a été appliquée au modèle.

milarité distributionnelle faible : ils ont des cooccurrents différents dans ce corpus, et ils ont effectivement des sens différents dans ce corpus.

Prenons un deuxième exemple, celui de la requête *industriel* et de son terme relié *naturel*. Les cooccurrents les plus discriminants de ces deux termes sont :

- *industriel : révolution, déchet, écologie, activité, secteur, ère, propriété, procédé, commercial, société, ...*
- *naturel : ressource, milieu, réserve, patrimoine, espace, parc, catastrophe, habitat, gaz, histoire, ...*

Le fait que les deux termes ont des cooccurrents très différents fait en sorte que leur similarité distributionnelle est faible et que *naturel* est éloigné du voisinage de *industriel*.

6.6.2.3 Présence de cooccurrents parmi les PPV

Un type particulier de bruit que l'on observe dans les voisinages distributionnels est la présence de liens qui sont plutôt syntagmatiques que paradigmatiques. Par exemple, dans le Tableau 6.XVIII, on peut voir que le 2e PPV du terme *parc* est *national*. Ces deux mots apparaissent fréquemment ensemble (*parc national*), mais ils ne sont pas paradigmatiquement reliés et n'apparaissent pas (séparément) dans des contextes similaires.

En effet, les méthodes distributionnelles n'arrivent pas toujours à distinguer les mots apparaissant dans des contextes similaires des mots apparaissant souvent ensemble[10]. Ainsi, les paires de mots distributionnellement similaires comprennent des mots qui sont reliés sur le plan syntagmatique (des collocations, ou du moins des mots qui ont une fréquence de cooccurrence élevée) plutôt que sur le plan paradigmatique. Si on cherche à détecter des relations paradigmatiques seulement, ces paires constituent du bruit.

Il est facile, du moins dans le cas de l'AD[11], de comprendre pourquoi les voisins

[10]Cette distinction est difficile à faire, puisque les mots paradigmatiquement reliés, qui apparaissent séparément dans des contextes similaires, peuvent aussi apparaître souvent ensemble.

[11]Le même phénomène se produit avec W2V, mais nous utilisons l'AD dans cet exemple parce qu'il est plus simple d'illustrer ce phénomène de cette façon.

distributionnels d'une requête comprennent parfois des mots qui sont en fait des cooccurrents de la requête. Reprenons l'exemple des mots *parc* et *national*. Chaque fois que ces deux mots apparaissent ensemble (*parc national*) dans le corpus, ils sont comptés comme cooccurrents l'un de l'autre, mais ils partagent également des cooccurrents, du moins si la taille de la fenêtre de contexte est supérieure à 1. Par exemple, si la taille de la fenêtre de contexte est 2, alors le mot apparaissant à gauche de *parc* sera compté comme cooccurrent de ce dernier, mais aussi de *national* ; de même, le mot apparaissant à droite de *national* sera compté comme cooccurrent des deux mots.

Ainsi, la façon dont on compte les fréquences de cooccurrence lors du calcul de l'AD (basée sur la cooccurrence graphique) fait en sorte que les mots qui partagent des cooccurrents, et qui ont par conséquent des représentations similaires, comprennent non seulement des mots qui apparaissent (séparément) dans des contextes similaires, mais aussi des mots qui apparaissent ensemble dans les mêmes contextes. Plus la fenêtre de contexte est large, plus les voisins distributionnels d'une requête donnée tendront à contenir des mots qui sont en fait des cooccurrents de la requête. Cela expliquerait éventuellement pourquoi on détecte mieux les DRV avec une fenêtre large (voir sections 6.4.1 et 6.5.1), du moins si les DRV ont une plus forte tendance à apparaître près les uns des autres que les synonymes et les autres relations que nous avons prises en compte. Il serait intéressant à cet égard de vérifier si les DRV ont effectivement une fréquence de cooccurrence plus élevée que les termes qui participent à une des relations paradigmatiques classiques telles que la synonymie ; il serait aussi intéressant de vérifier si l'on obtiendrait une précision plus élevée sur les DRV au moyen d'une mesure d'association plutôt qu'une mesure de similarité distributionnelle.

Plusieurs stratégies de réduction de bruit ont été proposées pour augmenter la précision des PPV que l'on obtient à partir d'un modèle distributionnel [1, 69, entre autres], notamment pour augmenter la proportion de voisins qui sont reliés sur le plan paradigmatique plutôt que syntagmatique. Par exemple, Rapp [159] souligne que l'on peut filtrer les voisins distributionnels pour ne retenir que ceux qui ont une association (syntagma-

tique) faible (ou vice-versa). Par ailleurs, l'utilisation de cooccurrents syntaxiques plutôt que graphiques peut réduire ce type de bruit [108, p. 304], si on exploite correctement ce type de contexte distributionnel.

6.6.2.4 Influence de la fréquence

Un facteur qui pose une difficulté pour l'approche distributionnelle est la fréquence extrêmement variable des mots. En effet, plusieurs travaux en sémantique distributionnelle indiquent que cette approche modélise mieux les mots fréquents, et que la qualité des résultats que l'on obtient pour les mots peu fréquents est plus faible [70, 178].

On peut donc se demander si les méthodes distributionnelles que nous avons utilisées modélisent moins bien le sens des mots dont elles observent un nombre faible de contextes, ou d'une façon plus générale, si elles modélisent le sens des mots plus ou moins bien selon leur fréquence.

Une méthode simple pour vérifier l'influence de la fréquence consiste à observer la fréquence des mots dont le voisinage est le moins précis. Dans le cas du modèle (produit par AD) que nous avons retenu pour cette analyse, les 15 requêtes dont le voisinage est le moins précis ont des fréquences très variables, la requête la moins fréquente (*développé*) ayant 211 occurrences et la plus fréquente (*parc*) ayant 27 733 occurrences. Cela ne suggère pas un lien entre la fréquence et la précision.

Pour vérifier d'une façon plus systématique s'il existe un lien entre la fréquence des requêtes et la précision de leur voisinage, plutôt qu'observer simplement la fréquence d'un petit nombre de requêtes pour lesquelles la précision est faible, on peut prendre en compte toutes les requêtes et analyser de manière quantitative la relation entre leur fréquence et la précision de leur voisinage. À cette fin, nous commençons par trier toutes les requêtes en fonction de leur fréquence, puis nous divisons la liste triée de requêtes en 10 *bandes de fréquences*, chacune contenant un nombre approximativement égal de requêtes, ensuite nous calculons la moyenne des précisions moyennes (MAP) pour chaque bande de fréquences. Cette technique a été utilisée dans quelques travaux en sémantique

distributionnelle afin d'analyser l'influence de la fréquence [68, 146].

Le résultat de ce calcul est illustré dans la Figure 6.13. Cette figure montre que la précision que l'on obtient pour les requêtes dont la fréquence est faible n'est pas plus basse, en moyenne, que celle pour les mots très fréquents. Au contraire, on observe que la première bande de fréquences (le sous-ensemble des termes les plus fréquents) a la MAP la plus faible, et que la MAP de la dernière bande de fréquences (0.4160) est beaucoup plus élevée que celle de la première (0.3085) ; en outre, la bande de fréquences dont la MAP est la plus élevée (0.5186) est l'avant-dernière.

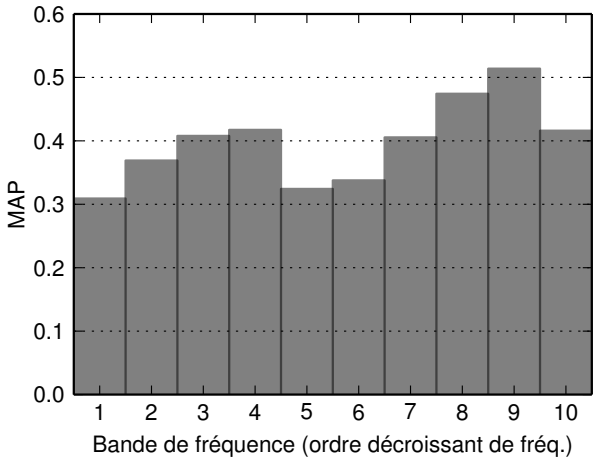


Figure 6.13 – MAP du modèle produit par AD sur les requêtes dans les couples de référence, réparties en 10 bandes de fréquences.

Ces résultats indiquent donc que les voisinages distributionnels des mots-cibles relativement peu fréquents sont en fait plus précis que ceux des mots-cibles très fréquents, du moins dans le cas du modèle que nous avons utilisé ici. Signalons que cette observation ne concorde pas avec les résultats de Ferret [70], qui indiquent que les mots peu fréquents ont des voisinages distributionnels très peu précis, mais les mots-cibles utilisés dans ce travail comprennent des mots ayant aussi peu que 11 occurrences, tandis que les mots-cibles les moins fréquents que nous avons utilisés ont plus de 150 occurrences.

Une explication possible pour le fait que la précision est plus faible, en moyenne,

pour les requêtes à fréquence très élevée est que les mots fréquents ont tendance à être plus polysémiques [195], et comme nous l'avons montré ci-dessus, la polysémie (ainsi que la couverture des données de référence) est un facteur qui explique la faible précision de certains voisinages distributionnels. Par exemple, si une requête est polysémique et que son voisinage comprend des mots correspondant à différents sens de cette requête, la précision de ce voisinage sera faible si les données de référence rendent seulement compte d'un de ses sens.

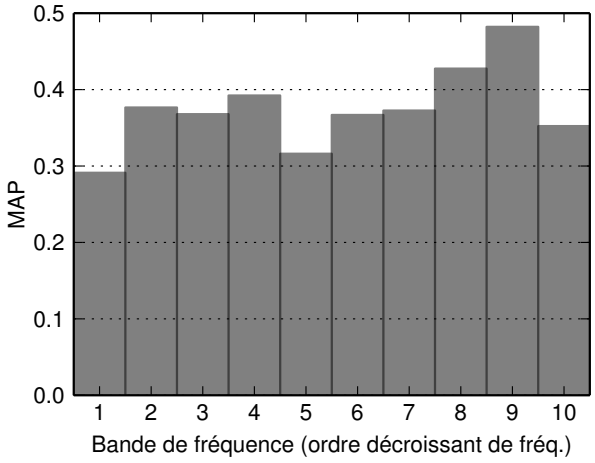


Figure 6.14 – MAP du modèle W2V sur les requêtes dans les couples de référence, réparties en 10 bandes de fréquences.

Il est intéressant de noter que nous observons une tendance presque identique dans le cas du modèle W2V, comme le montre la Figure 6.14.

6.6.2.5 Autres sources d'erreurs

Certaines erreurs que l'on observe dans les voisinages distributionnels sont causées par les prétraitements appliqués au corpus. Par exemple, le Tableau 6.XVIII montre que le 1er PPV de l'adjectif *développé* est le verbe *industrialiser*. Cela est attribuable au fait que nous utilisons un corpus lemmatisé, et que le programme que nous avons utilisé pour la lemmatisation a systématiquement remplacé les occurrences de l'adjectif *indus-*

trialisé, qui partage des cooccurrents avec l'adjectif *développé* (tels que *pays*), par le verbe infinitif *industrialiser*, ainsi la représentation de ce verbe est-elle similaire à celle de l'adjectif *développé*.

Une autre source possible d'erreurs est le fait que les cooccurrents graphiques d'un mot-cible comprennent des mots qui ne sont pas informatifs quant à son sens, que Levy et Goldberg [108] appellent des *contextes accidentels* (*accidental contexts*) ; d'ailleurs, ces contextes accidentels expliquent éventuellement pourquoi les fenêtres étroites produisent les meilleurs résultats (excepté pour les DRV). Si les cooccurrents syntaxiques offrent des résultats légèrement meilleurs que les cooccurrents graphiques dans le cadre de l'AD, comme le suggèrent plusieurs études que nous avons soulignées au chapitre 3, c'est attribuable au fait que les cooccurrents syntaxiques contiennent moins de bruit que les cooccurrents graphiques, tout en captant des cooccurrents éloignés, mais pertinents.

6.6.3 Interprétation de la qualité des résultats

L'analyse quantitative présentée dans la première partie de ce chapitre porte sur la moyenne des précisions moyennes (MAP) des voisinages distributionnels, celle-ci étant calculée en utilisant les relations encodées dans des dictionnaires spécialisés comme référence. Par exemple, on atteint une MAP maximale de 0.4107 sur les couples de référence en français et de 0.3953 en anglais. Or, comme nous l'avons souligné à la section 6.1, bien qu'elle soit très utile à des fins de comparaison, la MAP elle-même ne nous donne pas une idée très claire de la qualité du classement des PPV que l'on obtient à partir d'un modèle distributionnel, pour au moins deux raisons : nous avons déjà souligné le fait que la couverture incomplète des données de référence fait en sorte que l'évaluation automatique sous-estime la précision et surestime le bruit (voir section 6.6.2) ; une autre raison, sur laquelle nous ne nous sommes pas penchés jusqu'à maintenant, est le fait que les valeurs de la MAP peuvent être difficiles à interpréter, du moins plus difficiles que d'autres mesures plus simples telles que la précision et le rappel. Nous tenterons ci-dessous de faciliter l'interprétation de la MAP, pour fournir une idée plus claire de la

qualité du classement des PPV que l'on obtient à partir des modèles distributionnels.

Rappelons d'abord que la MAP est la moyenne des précisions moyennes (AP) pour un ensemble de requêtes (termes), et que l'AP pour une requête particulière est la précision moyenne de sa liste ordonnée de PPV au rang de chacun de ses termes reliés selon les données de référence.

Supposons que l'on observe les PPV d'une requête pour laquelle on connait deux termes reliés[12] et vérifions quels rangs ces deux termes doivent avoir dans la liste de PPV pour que l'AP soit d'au moins 0.4 pour cette requête. La Figure 6.15 montre l'AP en fonction du rang des deux termes reliés.

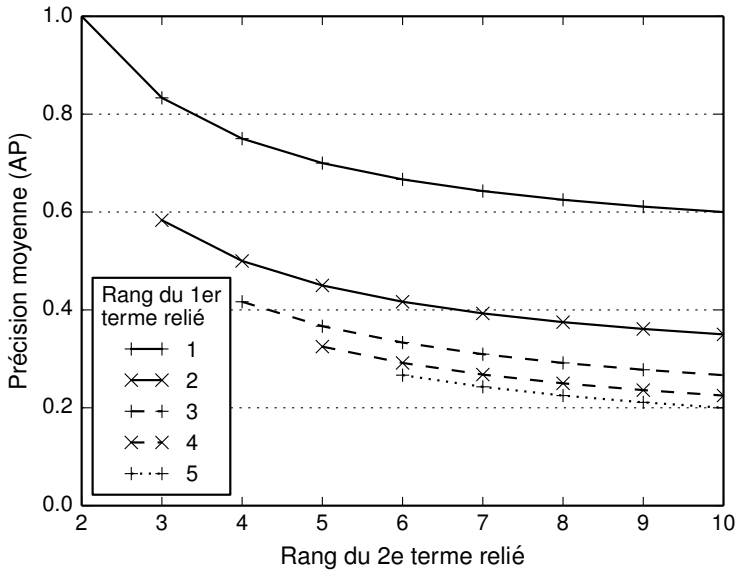


Figure 6.15 – AP pour une requête ayant deux termes reliés, en fonction du rang des deux termes reliés.

[12]Nous supposons qu'il y a seulement deux termes reliés pour pouvoir illustrer graphiquement la relation entre l'AP et le rang des termes reliés parmi les PPV de la requête. Si le nombre de termes reliés peut être plus élevé que 2, soulignons que les couples de référence contiennent en moyenne 2.68 termes reliés par requête en français (1314/490), et 2.33 en anglais (888/381).

Cette figure montre que l'AP est égale ou supérieure à 0.4 seulement dans les situations suivantes :

- Un des deux rangs est 1.
- Un des deux rangs est 2 et l'autre est inférieur ou égal à 6.
- Un des deux rangs est 3 et l'autre est 4.

Dans tous les autres cas, l'AP est inférieure à 0.4.

Ainsi, le fait que l'on atteint une AP moyenne (MAP) autour de 0.4 sur l'ensemble des requêtes suggère que le classement des PPV que l'on obtient à partir d'un modèle distributionnel est très bon, voire remarquablement bon, surtout lorsqu'on tient compte du nombre élevé de mots-cibles à classer (10000), ceux-ci ayant été choisis simplement en fonction de la fréquence, et du fait que les corpus utilisés ont été compilés automatiquement, entre autres.

Pour mieux caractériser la qualité du classement des PPV, on peut aussi faire appel à des mesures d'évaluation plus faciles à interpréter que la MAP, telles que la précision au rang k ($P@k$) ou le rappel au rang k ($R@k$). Le calcul de ces mesures consiste simplement à prendre un nombre fixe (k) de voisins pour chaque requête et à calculer la précision ou le rappel de ces listes de k voisins, de la même façon que nous calculons la précision ou le rappel sur les graphes, que nous avons expliquée à la section 5.6.2.1. Un des avantages de cette approche est qu'elle reflète le fait qu'un utilisateur typique (en l'occurrence, un terminologue) examinerait un petit nombre de voisins pour une requête donnée ; il ne parcourrait probablement pas la liste complète, qui contiendrait 10000 termes dans ce cas-ci.

Si on prend le modèle (produit par AD) que nous avons retenu pour cette discussion (voir section 6.6.2), et que l'on calcule la précision et le rappel des k PPV sur les couples de référence pour différentes valeurs de k, on obtient les données présentées dans le Tableau 6.XX. Ces données sont effectivement plus faciles à interpréter que la MAP : par

k	$P@k$	$R@k$
1	0.4163	0.1944
10	0.1467	0.5862
100	0.0232	0.8734

Tableau 6.XX – $P@k$ et $R@k$ du modèle retenu (produit par AD) en fonction de k.

exemple, la $P@10$ de 0.1467 indique que les 10 PPV d'une requête donnée contiennent environ 1.5 termes qui sont effectivement reliés à la requête selon les données de référence, en moyenne ; de même, le $R@10$ de 0.5862 indique que, pour une requête donnée, un peu moins de 60% de ses termes reliés se retrouvent parmi ses 10 PPV, en moyenne.

Si nous avons utilisé la MAP dans ce travail plutôt que la précision ou le rappel au rang k, c'est parce qu'elle nous fournit une mesure unique de la qualité du classement des PPV pour un modèle donné, qui tient compte à la fois de la précision et du rappel. S'il fallait observer séparément la précision et le rappel, il serait beaucoup plus difficile de réaliser une analyse comme celle que nous avons présentée dans ce chapitre, d'autant plus qu'il faudrait prendre en compte plusieurs valeurs de k pour avoir une bonne idée de la qualité des classements.

6.7 Synthèse

Pour réaliser les objectifs de ce travail, que nous avons énoncés au chapitre 4, nous avons fait deux hypothèses générales, à savoir :

- **Hypothèse 1** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme donné, des termes reliés sur le plan paradigmatique, tels que des (quasi-)synonymes, des antonymes, des hyperonymes, des hyponymes et des dérivés syntaxiques.

- **Hypothèse 2** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme ou un ensemble de termes donné, des

termes évoquant le même cadre sémantique.

Dans ce chapitre, nous avons analysé les résultats de l'évaluation du classement des PPV que l'on obtient à partir des modèles distributionnels. La MAP élevée que l'on obtient sur les couples de référence confirme notre première hypothèse, les relations les mieux détectées étant la (quasi-)synonymie et la dérivation syntaxique. Bien que certaines paires de termes reliés selon le DiCoEnviro aient une similarité distributionnelle très faible (cette faible similarité étant parfois liée à la polysémie, comme nous l'avons montré à la section 6.6.2), et bien que les paires de termes distributionnellement similaires comprennent des relations syntagmatiques, le classement des PPV des entrées de cette ressource demeure très bon en général.

Si on se base sur la précision du classement des PPV, nos résultats appuient la seconde hypothèse un peu moins que la première, puisqu'on obtient une MAP moins élevée sur les ensembles de référence que sur les couples de référence. Notre analyse des résultats en fonction de la PDD des termes et des relations ciblées nous a permis de proposer des explications pour cette différence. Premièrement, les ensembles de référence contiennent beaucoup de verbes, qui sont moins bien modélisés que les noms ou les adjectifs par l'approche distributionnelle. Deuxièmement, les ensembles de référence contiennent à la fois des dérivés syntaxiques et des relations paradigmatiques classiques telles que la (quasi-)synonymie, et les modèles qui captent le mieux ces deux types de relations sont très différents, ce qui limite la capacité d'un modèle particulier à détecter les termes qui évoquent le même cadre sémantique. À ce sujet, nous avons montré ailleurs que l'on peut mieux modéliser cette propriété en combinant deux modèles différents (un modèle produit par AD et un modèle produit par W2V) qui captent chacun cette propriété relativement bien [18].

La MAP que l'on obtient sur les ensembles de référence demeure élevée, ce qui indique que les modèles distributionnels permettent effectivement de détecter des termes évoquant le même cadre sémantique.

Un des objectifs de ce travail était d'identifier les paramétrages optimaux des modèles distributionnels pour l'identification de relations lexicales paradigmatiques et de termes appartenant au même cadre sémantique. Nos observations principales quant à l'influence des (hyper)paramètres ont été résumées aux sections 6.4.5 et 6.5.6. Par exemple, nous avons montré que, dans le cas de W2V, le seuil pour le sous-échantillonnage des mots fréquents exerce une influence très importante sur la précision des voisinages distributionnels. Nous avons également montré que l'influence des (hyper)paramètres varie en fonction des relations ciblées, et dans une moindre mesure, de la PDD des requêtes ; en revanche, elle varie très peu en fonction de la langue traitée.

En ce qui concerne le choix de la méthode utilisée pour construire le modèle, nos résultats indiquent que l'AD offre de meilleurs résultats que W2V sur la plupart des jeux de données utilisés dans ce travail (voir section 6.3), du moins sur les corpus que nous avons utilisés, mais que W2V détecte mieux les dérivés syntaxiques.

À la section 6.6, nous avons présenté une discussion des résultats analysés d'une manière quantitative dans les sections précédentes. Nous avons résumé nos observations quant aux différents facteurs liés à l'application envisagée qui peuvent éventuellement expliquer la qualité des résultats qu'offrent les modèles distributionnels et l'influence de leurs (hyper)paramètres. Par la suite, nous avons exploré différentes sources d'erreurs, c'est-à-dire des facteurs expliquant la faible précision de certains voisinages distributionnels. En observant les PPV de requêtes pour lesquelles la précision moyenne (AP) est faible, nous avons notamment montré que la polysémie est un facteur qui explique pourquoi certaines paires de termes qui participent effectivement à une relation lexicale paradigmatique ne sont pas distributionnellement similaires. Ainsi, les voisinages peu précis selon l'évaluation automatique ne contiennent pas forcément beaucoup de bruit ; dans certains cas, c'est qu'ils contiennent des termes reliés à un autre sens de la requête que celui décrit (ou ceux décrits) dans la référence, les voisins distributionnels de la requête dépendant du (ou des) sens qu'elle a dans le corpus. Par ailleurs, en analysant la relation entre l'AP des requêtes et leur fréquence relative, nous avons montré que la précision du

voisinage des requêtes à faible fréquence n'est pas moins élevée que celle des mots très fréquents, et qu'en fait, ce sont les termes les plus fréquents qui ont, en moyenne, les voisinages distributionnels les moins précis, du moins dans le cas des modèles que nous avons analysés. Nous avons proposé une explication pour cette observation, à savoir que les mots fréquents ont tendance à être plus polysémiques.

Jusqu'ici, notre analyse des résultats a porté sur la qualité du classement des PPV que l'on obtient pour chaque terme au moyen d'un modèle sémantique distributionnel. Nous avons ainsi abordé la plupart des problématiques que nous avons énoncées au chapitre 4, et sur lesquelles nous reviendrons dans la conclusion de cette thèse. La dernière de ces problématiques, qui porte sur le mode d'interrogation des modèles, est abordée au chapitre suivant, dans lequel nous analysons les résultats de l'évaluation de différents types de graphes que l'on peut utiliser pour interroger les modèles distributionnels.

CHAPITRE 7

RÉSULTATS DE L'ÉVALUATION DES GRAPHES

Introduction

Au chapitre 6, nous avons analysé les résultats de l'évaluation des listes ordonnées de PPV que l'on obtient à partir des modèles sémantiques distributionnels. Dans ce chapitre, nous nous intéressons à une autre façon d'interroger les modèles distributionnels, à savoir la construction et la visualisation d'un graphe de voisinage. Nous y présentons à cette fin les résultats de l'évaluation des différents graphes de k PPV que nous avons construits à partir de chacun des modèles évalués.

Cette évaluation a été réalisée au moyen de la précision, du rappel et de la F-mesure. Comme nous l'avons expliqué au chapitre 5, ces mesures nous permettent de vérifier dans quelle mesure les ensembles (non ordonnés) de voisins distributionnels que contient le graphe de k PPV pour chaque requête correspondent aux voisins sémantiques de cette requête selon les données de référence.

Nous limiterons dans certains cas le nombre de facteurs pris en compte dans cette analyse en nous concentrant sur une seule langue (le français), un seul jeu de données (les ensembles de référence) et un seul type de modèle distributionnel (produit par AD), paramétré d'une façon particulière. Nous expliquerons comment ces décisions ont été prises à la section 7.3.

Comme le chapitre précédent, ce chapitre comporte une analyse quantitative, suivie d'une partie plutôt qualitative, dans laquelle nous présentons et analysons des exemples visuels de voisinages distributionnels et discutons quelques aspects de l'utilisation des graphes de k PPV pour l'interrogation des modèles distributionnels. Nous analysons d'abord, aux sections 7.1 et 7.2, l'influence des paramètres du graphe sur la taille des voisinages et sur les mesures d'évaluation. À la section 7.3, nous montrons comment

nous avons choisi un graphe particulier pour produire les exemples présentés à la section 7.4. Enfin, la section 7.5 porte sur la question des mots n'ayant aucun voisin dans le graphe de *k* PPV mutuel, et présente une stratégie pour ajuster la densité des graphes.

7.1 Influence des paramètres du graphe sur la taille des voisinages

Les graphes de voisinage que nous utilisons dans ce travail, les graphes de *k* PPV, ont deux paramètres, à savoir le type de graphe (orienté, symétrique ou mutuel) et *k*, le nombre de PPV pris en compte pour chaque mot-cible lors de la construction du graphe. Comme nous l'avons expliqué au chapitre 5, le nombre d'arêtes dans le graphe dépend de ces deux paramètres, entre autres ; et plus le nombre d'arêtes dans le graphe est élevé, plus la taille du voisinage de chaque mot est élevée, en moyenne. Ci-dessous, nous vérifions comment les deux paramètres du graphe influencent la taille des voisinages.

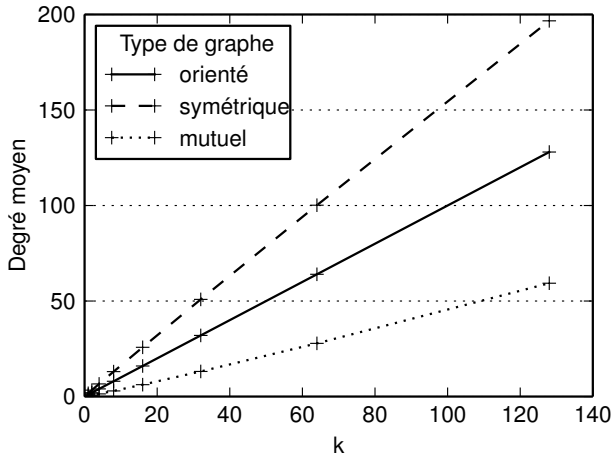


Figure 7.1 – Densité (degré moyen) des graphes en fonction de *k* et du type de graphe.

La Figure 7.1 montre l'influence du type de graphe et de la valeur de *k* sur la densité des graphes, c'est-à-dire sur le nombre moyen de voisins par mot. Dans la théorie des graphes, on appelle le nombre de voisins d'un nœud le *degré* de ce nœud. Ce que nous appelons la *densité* d'un graphe correspond donc au degré moyen de ce graphe. Pour chaque valeur de *k* et chaque type de graphe, la Figure 7.1 montre la moyenne (sur tous

les modèles évalués, dans les deux langues) de la densité des graphes correspondants. Comme nous l'avons souligné au chapitre 5, le nombre de voisins par mot vaut toujours k dans le cas des graphes orientés, chaque mot ayant exactement k voisins (ou successeurs). La Figure 7.1 illustre le fait que la densité du graphe (et la taille des voisinages) est plus élevée[1], pour une valeur de k donnée, lorsqu'on utilise un graphe symétrique plutôt qu'orienté, et moins élevée lorsqu'on utilise un graphe mutuel.

Il sera important de se rappeler cette propriété des graphes de k PPV lorsque nous analyserons, à la section suivante, l'influence des paramètres du graphe sur la qualité des voisinages, que nous estimons au moyen des mesures d'évaluation.

7.2 Influence des paramètres du graphe sur les mesures d'évaluation

Les paramètres du graphe influencent non seulement le nombre d'arêtes qu'il contient, mais aussi la précision du voisinage de chacun des mots-cibles. Dans cette section, nous analysons l'influence des paramètres du graphe sur les mesures d'évaluation, en nous concentrant sur un jeu de données particulier, à savoir les ensembles de référence[2].

La Figure 7.2 montre l'influence de k sur les mesures d'évaluation. Les mesures affichées sont la précision moyenne, le rappel moyen et la mesure F_1, calculés sur les ensembles de référence ; chaque point est une moyenne sur les trois types de graphe et sur tous les modèles (dans les deux langues) qui ont servi à construire les graphes. Cette figure montre clairement l'influence de k sur la précision et le rappel, mais on voit mieux comment la mesure F_1 varie en fonction de k en utilisant une échelle logarithmique pour l'abscisse, comme le montre la Figure 7.3.

Ces figures illustrent bien la relation entre k et le rappel. Comme nous l'avons expliqué au chapitre 5, plus le nombre d'arêtes dans le graphe est élevé, plus le rappel tend à

[1] Plus précisément, la densité des graphes symétriques que nous avons construits dans le cadre de ce travail se situe, en moyenne, entre $1.54k$ et $1.71k$, selon la valeur de k. Celle des graphes mutuels se situe entre $0.29k$ et $0.46k$. La densité relative des graphes symétriques par rapport à k diminue à mesure que k aumente, et celle des graphes mutuels augmente.

[2] Nous justifions ce choix à la section 7.3.

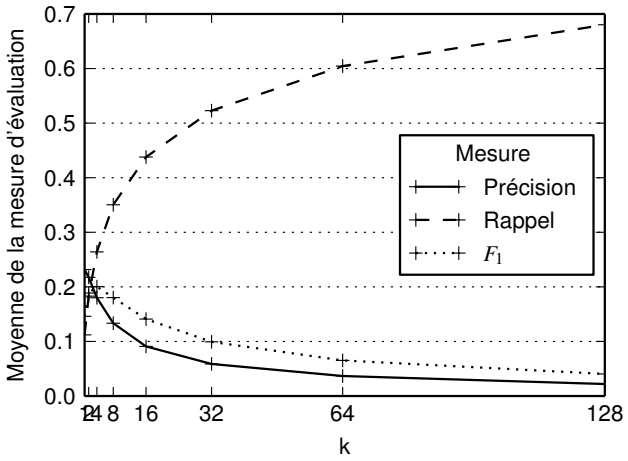


Figure 7.2 – Mesures d'évaluation moyennes en fonction de k.

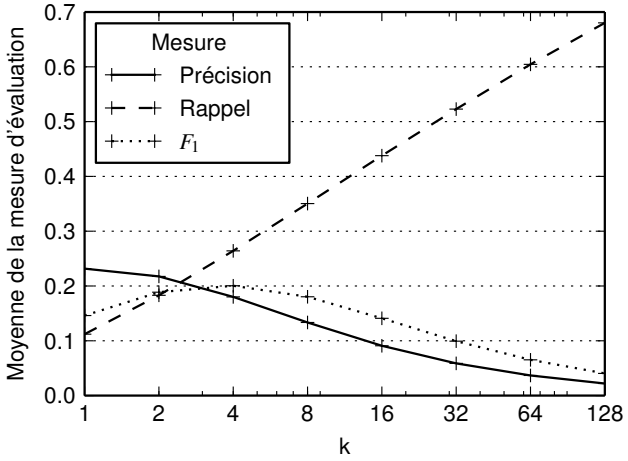


Figure 7.3 – Mesures d'évaluation moyennes en fonction de k (abscisse en échelle logarithmique).

être élevé ; en effet, pour une requête donnée, plus la taille de son voisinage est élevée, plus on a de chances d'y retrouver les termes reliés à cette requête selon les données de référence. Plus précisément, la Figure 7.3, montre que le rappel augmente de façon approximativement linéaire en fonction du logarithme de k.

Inversement, la précision tend à baisser en fonction de k, comme le montrent les Figures 7.2 et 7.3. Étant donné que les données de référence contiennent un nombre relativement petit de voisins sémantiques pour chaque requête, plus le voisinage distributionnel d'une requête est vaste, plus il tend à contenir du bruit, c'est-à-dire des mots qui ne font pas partie de ses voisins sémantiques.

Quant à la mesure F_1, qui combine la précision et le rappel, la Figure 7.3 montre qu'elle atteint un maximum lorsque $k = 4$, en moyenne. Donc, les voisinages qui offrent le meilleur compromis entre la précision et le rappel sont relativement petits, du moins si on accorde une importance égale à la précision et au rappel[3].

Type de graphe	Précision	Rappel	F_1
Orienté	0.1331	0.3957	0.1386
Symétrique	0.0930	**0.4683**	0.1177
Mutuel	**0.1380**	0.3191	**0.1417**

Tableau 7.I – Mesures d'évaluation moyennes en fonction du type de graphe.

En ce qui concerne le type de graphe, l'influence de ce paramètre sur les mesures d'évaluation (calculées sur les ensembles de référence) est présentée dans le Tableau 7.I ; chaque valeur est une moyenne sur tous les modèles, dans les deux langues, et toutes les valeurs de k. Ces résultats indiquent que les graphes symétriques ont la précision la plus faible et le rappel le plus élevé, en moyenne ; cela s'explique par le fait que la taille des voisinages est plus élevée, pour une valeur de k donnée, dans les graphes symétriques que dans les graphes orientés et mutuels, comme nous l'avons illustré à la section 7.1. Inversement, les graphes mutuels ont la précision la plus élevée et le rappel

[3]Nous reviendrons sur ce point à la section 7.5.

le plus faible, en moyenne. En ce qui concerne la mesure F_1, ce sont les graphes mutuels qui produisent les meilleurs résultats, en moyenne ; la mesure F_1 est un peu moins élevée pour les graphes orientés, et beaucoup moins élevée pour les graphes symétriques.

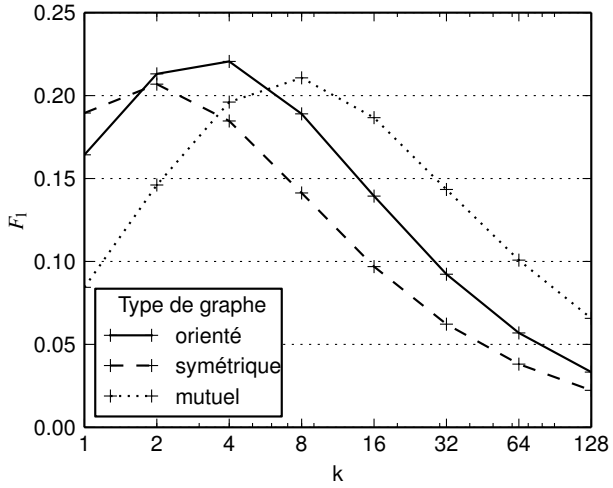


Figure 7.4 – Mesure F_1 moyenne sur les ensembles de référence en fonction de k et du type de graphe (abscisse en échelle logarithmique).

Nous avons constaté qu'il y a une interaction importante entre le type de graphe et k en ce qui concerne leur influence sur la mesure F_1, la valeur optimale de k dépendant du type de graphe (ou vice-versa). Cette relation est illustrée dans la Figure 7.4, qui montre la mesure F_1 moyenne (pour tous les modèles, dans les deux langues) en fonction de k et du type de graphe, calculée sur les ensembles de référence. Cette figure montre que la valeur optimale de k varie en fonction du type de graphe, cette valeur étant 2 pour les graphes symétriques, 4 pour les graphes orientés et 8 pour les graphes mutuels.

Type de graphe	Précision	Rappel	F_1
Orienté	**0.4226**	0.8119	0.2968
Symétrique	0.3490	**0.8693**	0.3052
Mutuel	0.3535	0.7785	**0.3343**

Tableau 7.II – Mesures d'évaluation maximales en fonction du type de graphe.

Le Tableau 7.II montre les mesures d'évaluation maximales atteintes par chaque type

de graphe. Que l'on observe les mesures d'évaluation moyennes (Tableau 7.I) ou maximales (Tableau 7.II), ce sont les graphes mutuels qui offrent le meilleur compromis entre précision et rappel (bien que les graphes orientés obtiennent une mesure F_1 moyenne légèrement plus élevée si on observe seulement, pour chaque type de graphe, les résultats obtenus avec la valeur optimale de k pour ce type de graphe, comme le montre la Figure 7.4).

7.3 Choix du graphe à visualiser

Nous avons tenu compte de nombreux facteurs dans l'analyse des résultats présentée au chapitre 6 et dans les premières sections de ce chapitre-ci : deux langues, deux méthodes pour construire des modèles distributionnels, plusieurs (hyper)paramètres pour chaque méthode, deux cadres descriptifs, trois parties du discours, quatre types de relations lexicales et différents paramétrages du graphe de voisinage. Or, si on désire approfondir l'analyse des résultats en observant des exemples visuels de voisinages distributionnels, comme nous le ferons à la section 7.4, on ne peut pas tenir compte de tous ces facteurs ; en effet, une analyse qualitative systématique de la façon dont le contenu des voisinages varie en fonction de tous ces facteurs dépasse largement les objectifs de cette thèse. Nous choisirons donc un graphe particulier, et tous les exemples que nous présenterons par la suite proviendront de ce graphe, sauf indication contraire.

Pour choisir ce graphe, nous stipulons d'abord que nous nous intéressons particulièrement au cas du français et au cadre descriptif de la sémantique des cadres. Ce deuxième choix est motivé par le fait que le graphe de voisinage nous semble particulièrement utile lorsqu'il s'agit d'observer le voisinage d'un ensemble de termes, tel qu'un ensemble d'unités lexicales qui évoquent un même cadre sémantique, plutôt que le voisinage d'un seul terme[4]. Ainsi, la sélection du modèle et du graphe sera basée sur des mesures d'évaluation calculées sur les ensembles de référence en français.

[4]Le graphe de voisinage présente des avantages par rapport aux listes ordonnées de PPV même lorsqu'il s'agit d'observer le voisinage d'un seul terme, comme nous l'avons souligné au chapitre 3.

En ce qui concerne la construction du modèle, nous optons pour l'AD. Nous avons montré à la section 6.3 que l'AD offre de meilleurs résultats que W2V sur les ensembles de référence en français en ce qui concerne le classement des PPV. On peut montrer que l'AD offre aussi de meilleurs résultats en ce qui concerne les mesures d'évaluation du graphe. Le Tableau 7.III, qui présente la précision, la mesure F_1 et le rappel maximaux atteints par chaque méthode sur les ensembles de référence en français, montre que l'AD atteint effectivement une mesure F_1 plus élevée que W2V.

Modèle	Précision	Rappel	F_1
AD	**0.4081**	**0.8387**	**0.3343**
W2V	0.4036	0.8122	0.3088

Tableau 7.III – Mesures d'évaluation maximales en fonction du type de modèle, calculées sur les ensembles de référence en français.

Quant au paramétrage du modèle, l'analyse de l'influence des paramètres de l'AD présentée à la section 6.4 indique que les valeurs qui maximisent, en moyenne, la MAP sur les ensembles de référence en français sont :

- Direction de la fenêtre : G+D.
- Forme de la fenêtre : triangulaire.
- Taille de la fenêtre : 4.
- Pondération : z-score.

On peut se demander si l'influence de ces paramètres sur les mesures d'évaluation du graphe est la même que leur influence sur la MAP. Le Tableau 7.IV montre les valeurs de chaque paramètre qui maximisent, en moyenne, la précision, le rappel et la mesure F_1 sur les ensembles de référence en français. Ces données ont été produites au moyen de la même méthode que nous avons utilisée à la section 6.4, c'est-à-dire en comparant, pour chaque paramètre, la mesure d'évaluation que l'on obtient en moyenne pour

chaque valeur du paramètre, mais nous prenons ici la moyenne sur tous les paramétrages du modèle et du graphe. Ces résultats indiquent que la pondération optimale est soit l'information mutuelle (MI), soit le z-score, selon qu'on favorise la précision ou le rappel ; puisque MI offre une mesure F_1 (légèrement) plus élevée, en moyenne, nous optons pour cette pondération. En ce qui concerne les trois autres paramètres, la valeur qui maximise les trois mesures d'évaluation du graphe est la même que celle qui maximise la MAP.

Mesure d'évaluation	Direction optimale	Forme optimale	Taille optimale	Pondération optimale
Précision	G+D	Triangulaire	4	MI
Rappel	G+D	Triangulaire	4	z-score
F_1	G+D	Triangulaire	4	MI

Tableau 7.IV – Valeurs des paramètres de l'AD qui maximisent, en moyenne, chacune des mesures d'évaluation du graphe.

Le modèle étant paramétré, il ne nous reste qu'à fixer les paramètres du graphe, à savoir le type de graphe et k. La mesure d'évaluation que nous utilisons pour réaliser ce choix est la mesure F_1, parce que celle-ci représente un compromis entre la précision et le rappel. L'analyse de l'influence des paramètres du graphe présentée à la section 7.2 indique que ce sont les graphes mutuels qui offrent la mesure F_1 la plus élevée, et que la valeur optimale de k pour ce type de graphe est 8.

Pour nous assurer de choisir le meilleur graphe possible, nous avons observé la mesure F_1 de tous les graphes qui ont été construits sur le modèle que nous avons retenu. Le Tableau 7.V présente les mesures d'évaluation des graphes construits sur ce modèle, calculées sur les ensembles de référence en français. Ces résultats montrent que, étant donné le modèle que nous avons choisi, le graphe qui maximise la mesure F_1 est un graphe mutuel avec $k = 8$, ce qui confirme notre choix.

Par ailleurs, il est intéressant de noter, en observant les données présentées dans le Tableau 7.V, que la valeur de k qui maximise la précision n'est pas 1 dans le cas des graphes mutuels, contrairement aux graphes orientés et symétriques. Cela s'explique par

k	P	R	F_1
1	**0.3543**	0.1188	0.1779
2	0.3318	0.2219	0.2659
4	0.2455	0.3136	**0.2754**
8	0.1654	0.4102	0.2357
16	0.1045	0.4967	0.1727
32	0.0625	0.5821	0.1129
64	0.0370	0.6667	0.0701
128	0.0208	**0.7405**	0.0405

(a) Orienté

k	P	R	F_1
1	**0.2821**	0.1704	0.2125
2	0.2648	0.3311	**0.2943**
4	0.1702	0.4091	0.2404
8	0.1040	0.4993	0.1721
16	0.0617	0.5759	0.1115
32	0.0360	0.6483	0.0682
64	0.0212	0.7189	0.0412
128	0.0120	**0.8017**	0.0236

(b) Symétrique

k	P	R	F_1
1	0.1794	0.0661	0.0966
2	0.2466	0.1016	0.1439
4	**0.3408**	0.2142	0.2631
8	0.3195	0.3169	**0.3182**
16	0.2095	0.4141	0.2782
32	0.1216	0.5114	0.1965
64	0.0682	0.6114	0.1227
128	0.0334	**0.6741**	0.0636

(c) Mutuel

Tableau 7.V – Précision (P), rappel (R) et mesure F_1 des graphes construits sur le modèle retenu, en fonction de k et du type de graphe.

le fait que le graphe mutuel peut contenir des nœuds n'ayant aucun voisin[5] et que la précision vaut 0 par définition pour une requête n'ayant aucun voisin.

En somme, les paramètres utilisés pour construire le graphe dont sont extraits les exemples de voisinages distributionnels présentés à la section 7.4 sont :

- Direction de la fenêtre : G+D.
- Forme de la fenêtre : triangulaire.
- Taille de la fenêtre : 4.
- Pondération : MI.
- Type de graphe : graphe de k PPV mutuel.
- Nombre de PPV (k) : 8.

7.4 Visualisation de voisinages

Dans cette section, nous présentons des exemples de voisinages extraits du graphe que nous avons choisi à la section 7.3. Étant donné le nombre de nœuds (10 000) et d'arêtes (11 883) dans ce graphe, il ne serait pas aisé de représenter visuellement le

[5] Nous reviendrons sur ce point à la section 7.5.

graphe au complet au moyen d'une seule image. Nous explorons donc le graphe en observant des sous-ensembles de nœuds et d'arêtes, c'est-à-dire des *sous-graphes*. Dans la pratique, il serait possible d'explorer le graphe de manière dynamique, en choisissant au fur et à mesure les nœuds dont on veut explorer le voisinage davantage.

La méthode que nous utilisons pour explorer le graphe consiste à extraire le sous-graphe correspondant au voisinage de requêtes particulières et à produire des représentations visuelles[6] de ces sous-graphes. Les requêtes sont des ensembles de référence, c'est-à-dire des ensembles d'unités lexicales (UL) qui évoquent un cadre sémantique. Nous commençons par visualiser uniquement les UL ainsi que les arêtes les reliant dans le graphe. Par la suite, nous ajouterons à ces sous-graphes les voisins des UL, c'est-à-dire les mots qui sont reliés aux UL dans le graphe (leurs nœuds adjacents). On pourrait appeler ce type de voisinage un *voisinage à 1 saut*, puisqu'il contient tous les mots que l'on peut atteindre à partir des UL dans la requête en faisant un saut (ou moins) dans le graphe ; on appellerait alors les voisinages que nous avons décrits précédemment, qui ne contiennent que les UL, des *voisinages à 0 saut*. Dans un troisième temps, nous irons plus loin en ajoutant non seulement les voisins des UL, mais aussi les voisins de leurs voisins, ce qui produit un *voisinage à 2 sauts*. À des fins de concision, nous appellerons les voisinages à 0, 1 et 2 sauts le *0-voisinage*, le *1-voisinage* et le *2-voisinage* respectivement.

Nous montrons donc dans les sections suivantes des exemples de voisinages distributionnels, et nous vérifions comment ils se comparent à des voisinages sémantiques, en nous appuyant sur les données encodées dans le DiCoEnviro et le Framed DiCoEnviro.

7.4.1 0-voisinages

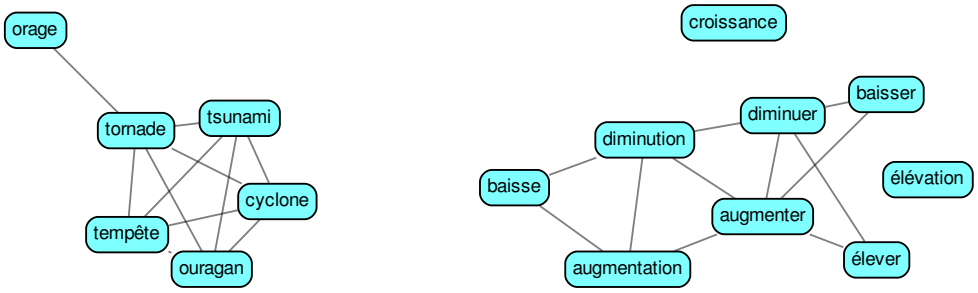
Nous commençons par vérifier comment les UL dans les ensembles de référence sont reliées entre elles dans le graphe de voisinage distributionnel, en observant des

[6] Nous utilisons à cette fin Graphviz (`http://www.graphviz.org/`) et la bibliothèque PyGraphviz pour Python (`https://pygraphviz.github.io/`) ; pages consultées le 21 novembre 2015.

exemples de 0-voisinages. Les 0-voisinages ne contiennent que les mots de la requête (en l'occurrence, les UL d'un ensemble de référence) et les liens entre ceux-ci. Ils peuvent être utilisés pour identifier les relations qui existent au sein d'un ensemble de mots (la requête), mais ne permettent pas de découvrir d'autres mots reliés à la requête. Si notre objectif était d'enrichir la liste d'UL qui évoquent un cadre sémantique, par exemple, ce genre de voisinage ne serait pas utile. Néanmoins, les 0-voisinages nous montrent de quelles façons les UL sont reliées entre elles dans le graphe de voisinage distributionnel. Considérés du point de vue des mesures d'évaluation, ils nous donnent une idée du rappel que l'on obtient pour chacune des UL.

Puisque le 0-voisinage ne contient que les mots de la requête, nous avons choisi comme requêtes, pour les exemples présentés dans cette section, des ensembles de référence contenant un nombre relativement élevé d'UL (6 ou plus).

En observant les 0-voisinages, on constate notamment que leur densité est très variable, certains voisinages contenant un nombre élevé d'arêtes (et offrant donc un rappel élevé), et d'autres contenant très peu d'arêtes (rappel faible).

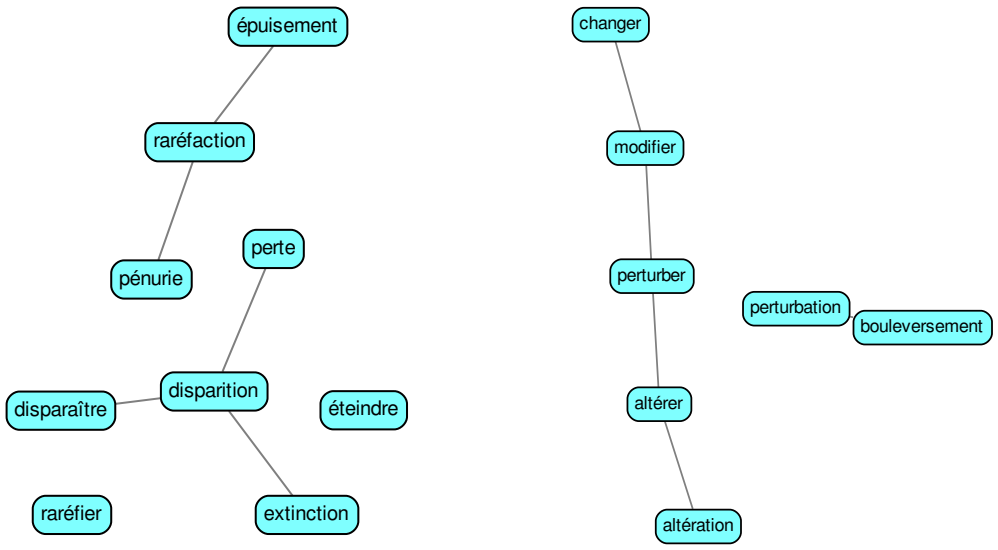


(a) Storm (b) Change_position_on_a_scale

Figure 7.5 – 0-voisinages à rappel élevé.

La Figure 7.5 montre deux exemples de 0-voisinages à rappel élevé, correspondant aux cadres sémantiques Storm et Change_position_on_a_scale. On observe que 5 des

6 UL du cadre Storm sont toutes reliées entre elles, formant ainsi un sous-graphe dit *complet*. Le rappel est donc élevé pour ces 5 UL (rappel de 1 pour *tornade* et de $\frac{4}{5}$ pour les autres UL[7]), mais faible ($\frac{1}{5}$) pour l'autre UL, *orage*. De même, 7 des 9 UL du cadre Change_position_on_a_scale participent entre elles à beaucoup de liens de voisinage distributionnel. En revanche, les UL *croissance* et *élévation* ne sont pas reliées aux autres UL de ce cadre. Le 0-voisinage ne permet pas de savoir à quels autres mots ces deux UL sont reliées, ce qui pourrait nous indiquer pourquoi elles ne sont pas reliées aux autres UL, comme nous le montrerons aux sections 7.4.2 et 7.4.3 lorsque nous observerons des voisinages à 1 saut et à 2 sauts.



(a) Ceasing_to_be

(b) Cause_change_of_state

Figure 7.6 – 0-voisinages à rappel moyen.

La Figure 7.6 montre deux exemples de 0-voisinages à rappel moyen. Ces sous-

[7]La manière dont nous calculons le rappel sur les ensembles de référence est expliquée au chapitre 5. Soulignons que l'on peut calculer cette mesure en observant le 0-voisinage d'un ensemble de référence seulement si les UL de cet ensemble ne sont pas incluses dans d'autres ensembles de référence. Les 6 UL du cadre Storm évoquent seulement ce cadre, ce qui nous permet de calculer exactement le rappel pour chacune de ces UL en observant le 0-voisinage.

graphes sont composés de plusieurs *composantes connexes*, c'est-à-dire des sous-ensembles de nœuds reliés entre eux. Par exemple, les UL du cadre Cause_change_of_state forment deux composantes connexes :

- *changer, modifier, perturber, altérer, altération* ;
- *perturbation, bouleversement*.

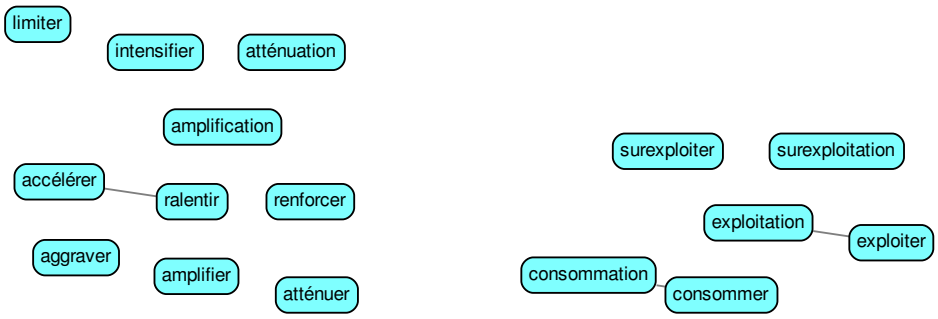
En ce qui concerne le cadre Ceasing_to_be, le 0-voisinage est constitué de quatre composantes connexes, dont deux sont constituées d'un seul nœud :

- *pénurie, raréfaction, épuisement* ;
- *disparaître, disparition, perte, extinction* ;
- *éteindre* ;
- *raréfier*.

Ainsi, ces 0-voisinages ne suggèrent pas que les termes contenus dans ces requêtes sont tous reliés entre eux, mais ils suggèrent tout de même que ces requêtes contiennent des sous-ensembles de termes reliés.

La Figure 7.7 montre deux exemples de 0-voisinages à rappel faible, c'est-à-dire qui contiennent très peu d'arêtes. Dans le cas du cadre Cause_change_of_impact, seuls les antonymes *accélérer* et *ralentir* sont reliés. Quant aux cadre Using_resource, seules les paires de dérivés syntaxiques {*exploiter, exploitation*} et {*consommer, consommation*} sont reliées.

Ces exemples montrent que le rappel que l'on obtient sur les ensembles de référence au moyen de ce graphe varie beaucoup d'un ensemble à l'autre, la densité des 0-voisinages étant très variable. Dans tous les cas où les UL d'un cadre sémantique ne sont pas reliées entre elles dans le graphe, on pourrait observer les cooccurrents des UL afin de déterminer pourquoi certaines ne sont pas distributionnellement similaires aux



(a) Cause_change_of_impact (b) Using_resource

Figure 7.7 – 0-voisinages à rappel faible.

autres, comme nous l'avons montré à la section 6.6.2. Dans cette section, nous nous concentrons plutôt sur les PPV et sur la structure des voisinages dans le graphe de k PPV. Nous ne nous attarderons pas davantage aux 0-voisinages, et passerons plutôt aux 1-voisinages, qui permettent de découvrir des mots sémantiquement reliés à ceux qui composent la requête.

7.4.2 1-voisinages

Les voisinages à 1 saut ou 1-voisinages comprennent la requête (en l'occurrence, un ensemble de référence) ainsi que les voisins distributionnels des termes de la requête, c'est-à-dire les nœuds adjacents à ces termes dans le graphe de k PPV. La visualisation de ces sous-graphes est une façon efficace d'interroger un modèle distributionnel, particulièrement lorsque la requête est composée de plusieurs termes. Ils permettent non seulement d'identifier des relations entre les termes de la requête, mais aussi de découvrir d'autres mots qui sont distributionnellement similaires, et éventuellement reliés sémantiquement, à ces termes ; ils seraient donc plus utiles que les 0-voisinages pour l'enrichissement de cadres sémantiques, par exemple. Considérés du point de vue des mesures d'évaluation, ces sous-graphes nous permettent d'observer non seulement le

rappel que l'on obtient pour chaque UL, mais aussi la précision de son voisinage.

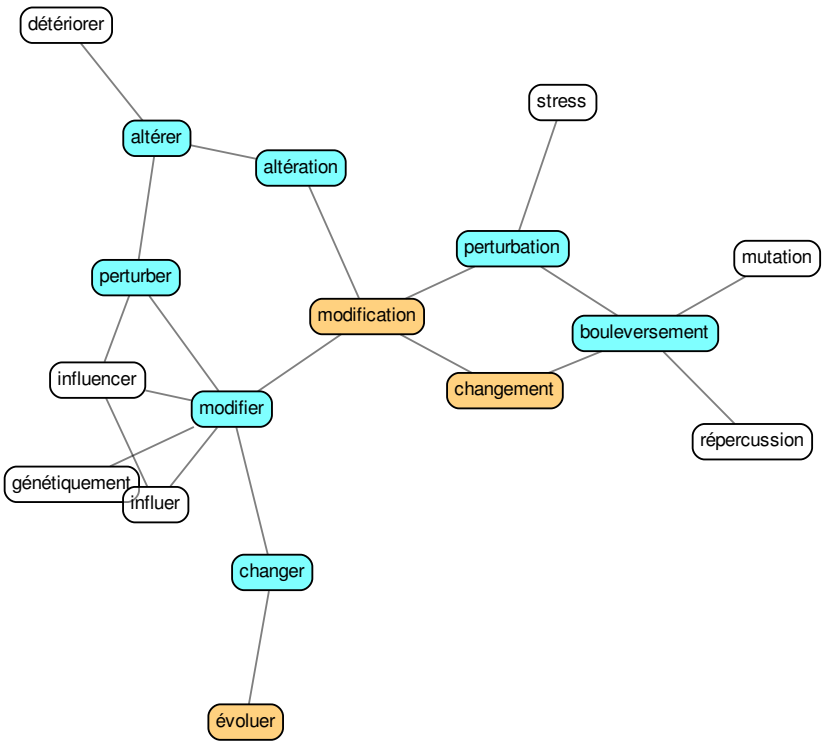


Figure 7.8 – 1-voisinage du cadre Cause_change_of_state.

La Figure 7.8 montre le 1-voisinage du cadre Cause_change_of_state, dont le 0-voisinage était illustré dans la Figure 7.6. Les UL qui évoquent ce cadre correspondent aux nœuds dessinés en bleu ; les nœuds oranges représentent des termes qui évoquent un autre cadre évoqué par une de ces UL. Par exemple, *modification* n'évoque pas le cadre Cause_change_of_state selon le Framed DiCoEnviro, mais il évoque un autre cadre, Undergo_change_of_state, dont les UL comprennent *modifier* et *changer* ; ce cadre est d'ailleurs relié à Cause_change_of_state (par une relation de type causatif/inchoatif). Ainsi, lorsqu'on calcule la précision du voisinage de *modifier*, *modification* fait partie de ses voisins de référence, et n'est pas considéré comme du bruit.

Bien que le 0-voisinage de ce cadre montre que *perturbation* et *bouleversement* ne

sont pas directement reliées aux autres UL du cadre, le 1-voisinage montre qu'elles le sont par le biais du terme *modification*. On observe également que le 1-voisinage comprend plusieurs termes qui évoquent d'autres cadres évoqués par les UL de ce cadre : *changement*, *modification* et *évoluer* évoquent toutes le cadre Undergo_change_of_state. Les autres nœuds dans cette figure (nœuds blancs) représentent du bruit en ce qui concerne le calcul de la précision, car seuls les termes évoquant un des cadres évoqués par un terme donné sont considérés comme corrects[8], mais cela ne veut pas forcément dire que ces mots ne sont pas pertinents ni sémantiquement reliés aux UL. Par exemple, les termes *répercussion*, *influer* et *influencer* évoquent un autre cadre, Objective_influence, qui est relié au cadre Cause_change_of_state (par une relation de type *voir aussi*). De plus, le terme *détériorer* évoque les cadres Damaging et Deteriorating, qui sont indirectement reliés à Cause_change_of_state. Enfin, le terme *stress* n'est pas encodé dans le Framed DiCoEnviro[9], mais selon le DiCoEnviro, il est paradigmatiquement relié à *perturbation*, ces deux termes ayant des sens voisins.

Ainsi, presque tous les termes dans le 1-voisinage de ces UL évoquent soit un autre cadre évoqué par une des UL, soit un cadre relié, ou participent à une relation paradigmatique avec une des UL, les seules exceptions étant *génétiquement* et *mutation*. Donc, la précision réelle du voisinage de ces UL est sans doute plus élevée que le suggère l'évaluation automatique, qui indique qu'elle est de 0.5357 en moyenne pour ces UL. En effet, une évaluation manuelle des 1-voisinages d'un ensemble de cadres sémantiques a montré que la précision réelle de ces voisinages est plus élevée que la précision calculée automatiquement au moyen de données de référence, ces voisinages contenant une proportion élevée d'UL évoquant les mêmes cadres sémantiques ou des cadres reliés [20].

La Figure 7.9 montre le 1-voisinage du cadre Change_position_on_a_scale. Le 0-voisinage de ce cadre (Figure 7.5) montrait que les UL *élévation* et *croissance* n'étaient

[8]On pourrait éventuellement modifier le calcul des mesures d'évaluation pour que les termes évoquant des cadres reliés aux cadres évoqués par la requête ne soient pas considérés comme complètement incorrects.

[9]Rappelons que cette ressource était toujours en développement lors de la rédaction de cette thèse. Cette partie de l'analyse des résultats a été réalisée au mois de novembre 2015.

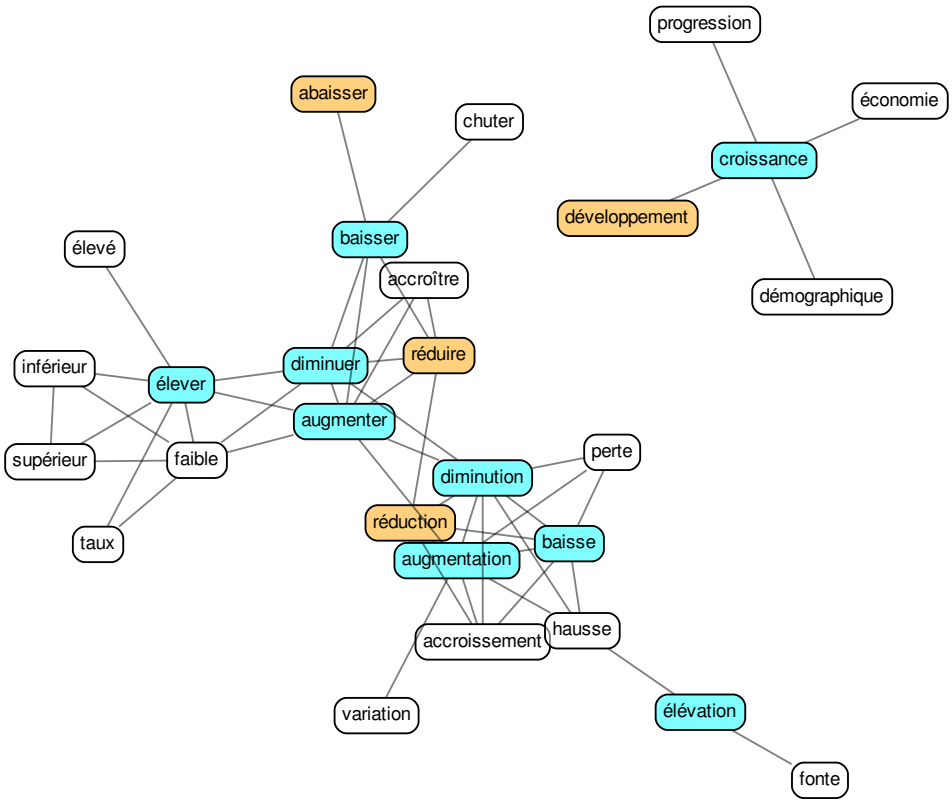


Figure 7.9 – 1-voisinage du cadre Change_position_on_a_scale.

pas reliées aux autres UL de ce cadre. Dans le 1-voisinage, on observe que *élévation* est reliée à *baisse*, à *diminution* et à *augmentation* par le biais du terme *hausse*. En revanche, *croissance* n'est pas reliée aux autres UL, ce terme ayant un voisinage différent, qui n'est pas relié au voisinage des autres UL, ce qui suggère que cette forme a un sens différent dans ce corpus, plus proche de *développement* que de *augmentation*.

Comme dans l'exemple précédent, les voisins des UL de ce cadre comprennent plusieurs termes qui évoquent un autre cadre évoqué par une de ces UL : *abaisser*, *réduire* et *réduction* évoquent tous le cadre Cause_change_of_position_on_a_scale, et *développement* évoque le cadre Expansion.

En ce qui concerne tous les autres nœuds (en blanc), soulignons que quelques-uns

de ces mots évoquent d'autres cadres du domaine de l'environnement : *perte*, *variation* et *fonte* évoquent les cadres Ceasing_to_be, Undergo_change_of_state et Change_of_phase respectivement. Aucun des autres mots n'est encodé dans le Framed DiCoEnviro actuellement. Ces autres mots ne sont pas encodés dans le DiCoEnviro non plus ; en fait, les termes *accroître*, *accroissement* et *hausse* le sont, mais leur description n'est pas suffisamment avancée pour être consultée par le public à l'heure actuelle ; ces termes seront sans doute reliés aux autres UL du cadre Change_position_on_a_scale.

Il est aussi intéressant de noter que dans la composante que l'on observe au centre de la Figure 7.9, on peut identifier deux sous-ensembles de nœuds participant entre eux à un nombre élevé de liens (des *clusters*), l'un étant constitué de verbes (*baisser*, *accroître*, *réduire*, *diminuer*, *augmenter*) et l'autre, de noms (*réduction*, *diminution*, *perte*, *augmentation*, *baisse*, *accroissement*, *hausse*).

Reprenons l'exemple d'un cadre dont le 0-voisinage avait une très faible densité (donc, pour lequel le rappel était faible). La Figure 7.10 montre le 1-voisinage du cadre Using_resource, dont le 0-voisinage était illustré dans la Figure 7.7. Dans cette figure, on observe que les termes *consommer* et *consommation* sont indirectement reliés aux termes *exploiter* et *exploitation*, par le biais des termes *produire* et *production* (ou encore *utilisation*). Ce sous-graphe indique donc que *consommation* et *exploitation* sont tous les deux distributionnellement similaires à *production* (et à *utilisation*), et qu'ils sont en fait plus similaires, sur le plan distributionnel, à ce terme qu'ils ne sont similaires entre eux. Soulignons que *produire*, *production* et *utilisation* ne sont pas encodés dans le Framed DiCoEnviro à l'heure actuelle, donc nous ne pouvons pas nous appuyer sur cette ressource pour déterminer s'ils évoquent un cadre sémantique du domaine de l'environnement ; ils ne sont pas dans la version publique du DiCoEnviro non plus, mais les deux premiers sont en cours d'ajout (au sens de *émettre*).

Par ailleurs, les termes *surexploiter* et *surexploitation* ne sont pas reliés aux autres UL de ce cadre, ni entre eux, malgré la relation de dérivation syntaxique qui existe entre ces deux termes.

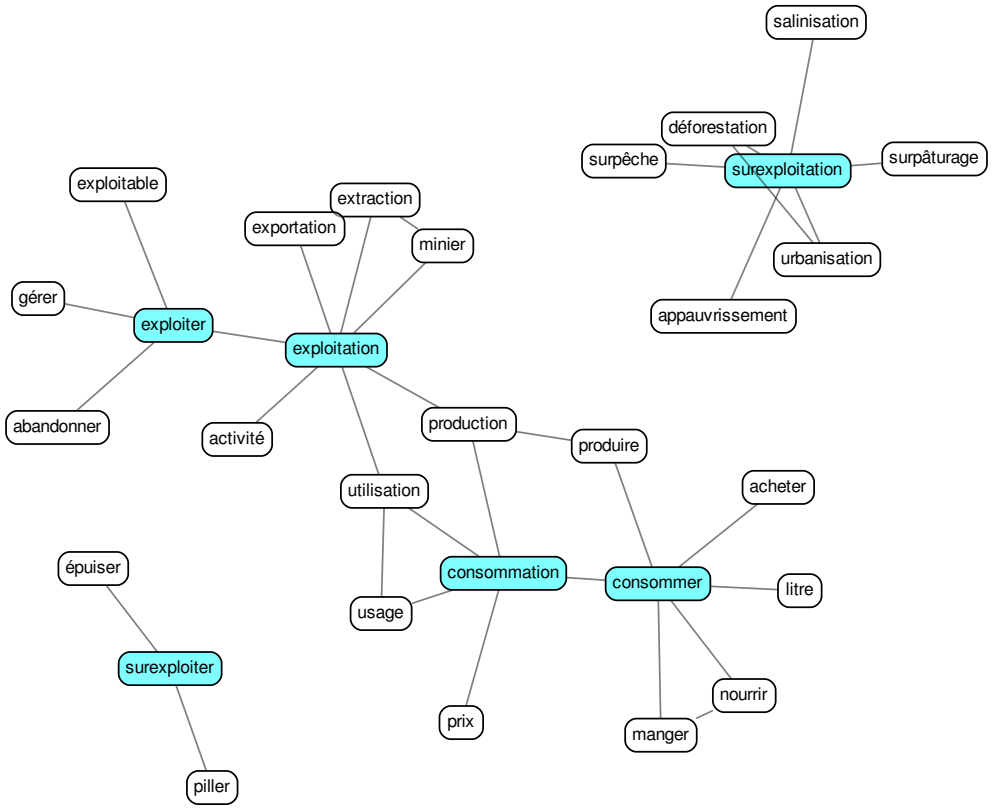


Figure 7.10 – 1-voisinage du cadre Using_resource.

Parmi les voisins des UL de ce cadre, soulignons que les termes *salinisation*, *urbanisation*, *appauvrissement*, *extraction* et *activité* sont tous encodés comme des UL évoquant d'autres cadres dans le Framed DiCoEnviro. Certains de ceux-ci évoquent des cadres reliés, directement ou indirectement, au cadre Using_resource : *urbanisation* (cadre Cause_change_into_organized_society), *extraction* (Mining) et *activité* (Human_activity).

7.4.3 2-voisinages

Dans certains cas, le 1-voisinage d'une requête (un terme ou un ensemble de termes) pourrait ne pas contenir suffisamment de mots pour fournir une bonne idée du voisinage distributionnel de la requête. Même si le 1-voisinage contient un bon nombre de termes, il peut être utile de pousser la recherche plus loin en observant non seulement les voisins de la requête, mais aussi les voisins de ses voisins[10], c'est-à-dire son 2-voisinage.

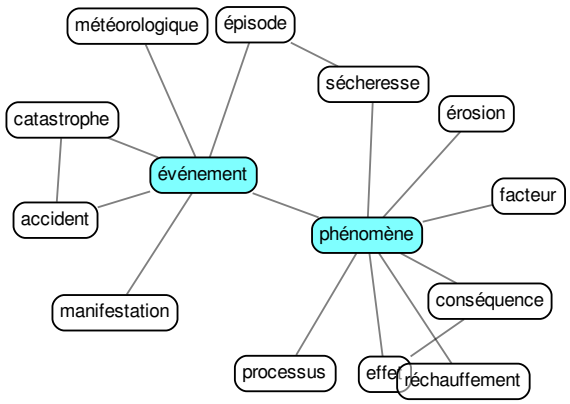


Figure 7.11 – 1-voisinage du cadre Weather_event.

Prenons par exemple le cadre Weather_event. Le 1-voisinage des UL qui évoquent ce cadre, *événement* et *phénomène*, est illustré dans la Figure 7.11. Ce voisinage contient plusieurs UL qui évoquent d'autres cadres du domaine de l'environnement, dont certains sont reliés à Weather_event (tels que Drought et Catastrophe) :

- *effet* (Objective_influence)
- *conséquence* (Causation)
- *catastrophe* (Catastrophe)

[10]Comme nous l'avons déjà souligné, l'exploration du graphe peut aussi se faire de façon dynamique, en ajoutant de nouveaux voisins à mesure que les termes et les relations sont validés.

- *sécheresse* (Drought)
- *érosion* (Change_natural_feature)
- *réchauffement* (Change_of_temperature et Cause_temperature_change)

Puisque le 1-voisinage de ce cadre contient beaucoup de termes qui évoquent d'autres cadres de ce domaine, il pourrait être intéressant d'explorer davantage ce voisinage en ajoutant les voisins des voisins, comme le montre le 2-voisinage illustré dans la Figure 7.12.

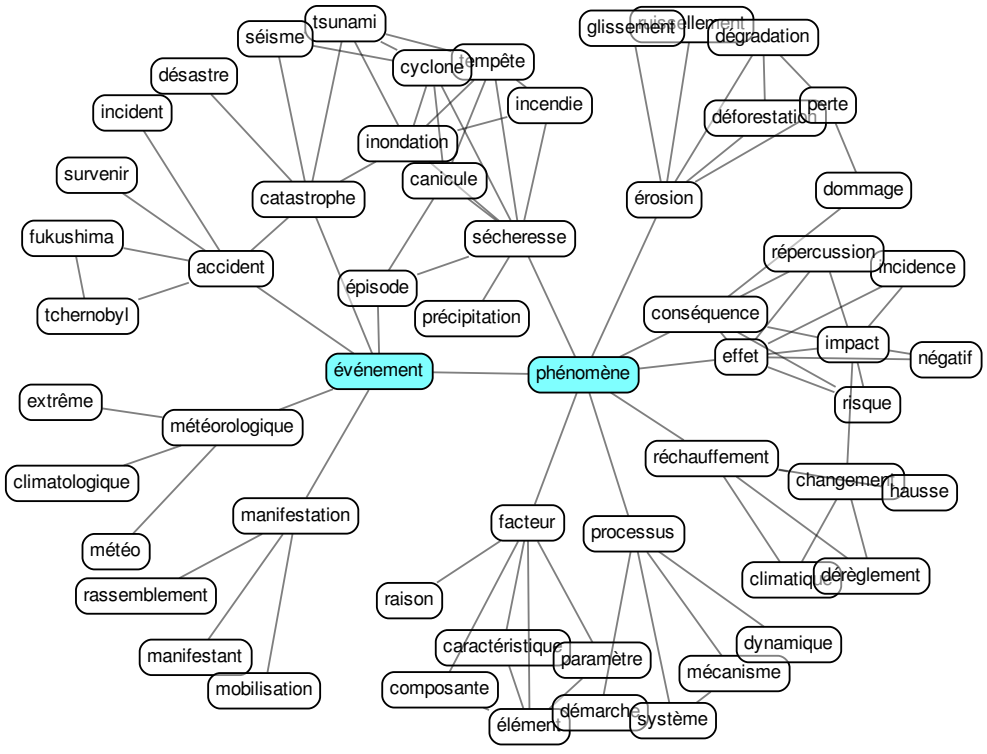


Figure 7.12 – 2-voisinage du cadre Weather_event.

Dans cette figure, on peut voir que beaucoup des voisins à 2 sauts sont des termes qui participent avec les UL ou leurs voisins à des relations lexicales paradigmatiques. En effet, selon le DiCoEnviro[11],

- *séisme*, *tsunami* et *incendie* sont des sortes de *catastrophe* ;
- *cyclone* est un *événement* ;
- *tempête* est un sens voisin de *cyclone* et de *séisme* ;
- et ainsi de suite.

De plus, certains de ces voisins à deux sauts évoquent les mêmes cadres sémantiques que les UL ou leurs voisins. Par exemple, les voisins du terme *effet* comprennent d'autres termes qui évoquent le cadre Objective_influence : *répercussion*, *incidence* et *impact*.

Ainsi, explorer les voisinages distributionnels de cette façon nous fournit des pistes pour la définition de nouveaux cadres sémantiques ou l'enrichissement du contenu lexical de cadres existants.

[11]Consulté le 12 novembre 2015.

7.4.4 Retour sur les sources d'erreurs

On peut observer dans le graphe de voisinage certaines des sources d'erreurs que nous avons soulignées à la section 6.6.2, outre le problème de la couverture des données de référence, que nous avons évoqué ci-dessus. Par exemple, la Figure 7.13, qui présente le 2-voisinage du cadre Predicting, illustre l'exemple de polysémie que nous avons expliqué dans cette section : les voisinages de *prévoir* et de *prévision* montrent clairement que ce sont deux sens différents qui ont été captés, à savoir le sens juridique ou administratif de *prévoir* et le sens plutôt scientifique ou technique de *prévision*.

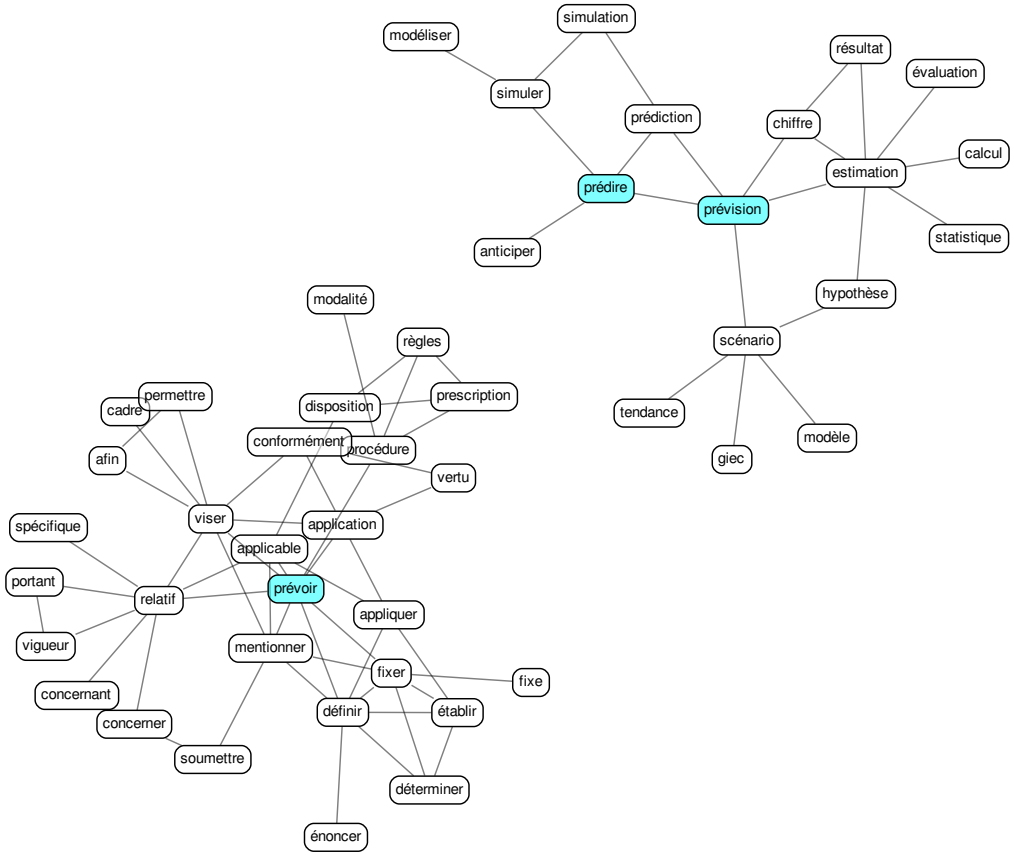


Figure 7.13 – 2-voisinage du cadre Predicting.

En outre, le 2-voisinage du cadre Cause_balance, présenté dans la Figure 7.14, montre l'influence des prétraitements, en l'occurrence la lemmatisation : si le verbe *équilibrer* est relié à des adjectifs, c'est attribuable (en partie) au fait que l'adjectif *équilibré* a systématiquement été remplacé par le verbe *équilibrer* lors de la lemmatisation ; cela explique pourquoi ce verbe est relié à des adjectifs tels que *harmonieux* plutôt que des verbes tels que *stabiliser*.

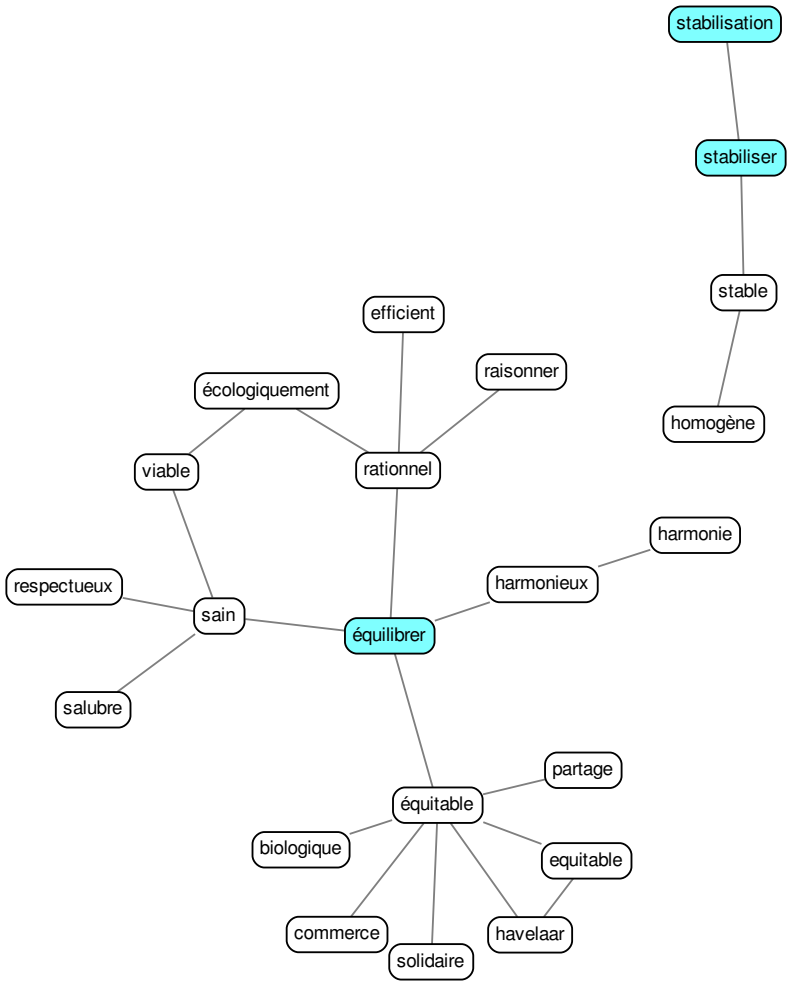


Figure 7.14 – 2-voisinage du cadre Cause_balance.

Enfin, les exemples de voisinages présentés dans ce chapitre contiennent des liens qui sont d'ordre syntagmatique plutôt que paradigmatique, tels que le lien entre *commerce* et *équitable* dans la Figure 7.14.

7.5 Les nœuds isolés

La procédure de sélection de graphe présentée à la section 7.3 nous a amené à opter pour un graphe de k PPV mutuel plutôt qu'un graphe orienté ou symétrique ; en outre, l'analyse de l'influence des paramètres du graphe présentée à la section 7.2 montre que ce sont les graphes mutuels qui produisent la mesure F_1 la plus élevée. Or, les graphes de k PPV mutuels ont une propriété particulière dont on doit tenir compte : alors que tous les nœuds ont au moins un voisin dans les graphes orienté et symétrique (à moins que $k = 0$, ce qui produit un graphe sans arêtes), le graphe mutuel peut contenir des nœuds qui n'ont aucun voisin, appelés *nœuds isolés*.

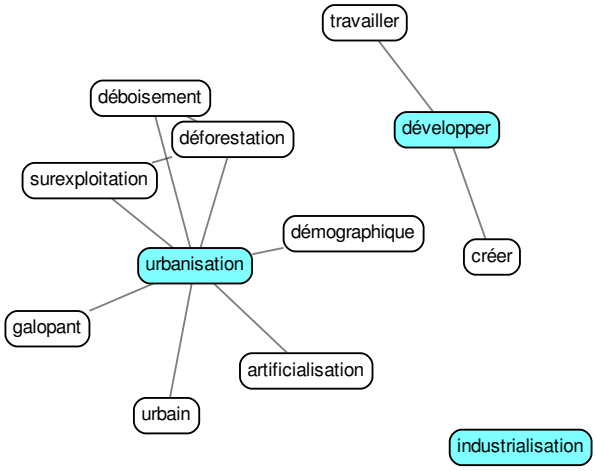


Figure 7.15 – 1-voisinage du cadre Cause_change_into_organized_society.

Par exemple, la Figure 7.15, qui illustre le 1-voisinage du cadre Cause_change_into_organized_society, montre que l'UL *industrialisation* n'a aucun voisin dans le graphe.

Parmi les 223 différentes UL dans les ensembles de référence en français, 11 n'ont aucun voisin dans le graphe que nous avons retenu[12]. Au total, 2366 des 10000 nœuds dans le graphe sont des nœuds isolés.

Le nombre de nœuds isolés dans le graphe de k PPV mutuel dépend, entre autres, de la valeur de k : plus cette valeur est élevée, plus le nombre de nœuds isolés tend à être faible. La Figure 7.16 illustre cette relation en montrant comment le nombre de nœuds isolés dans le graphe varie en fonction de k, pour le modèle que nous avons retenu.

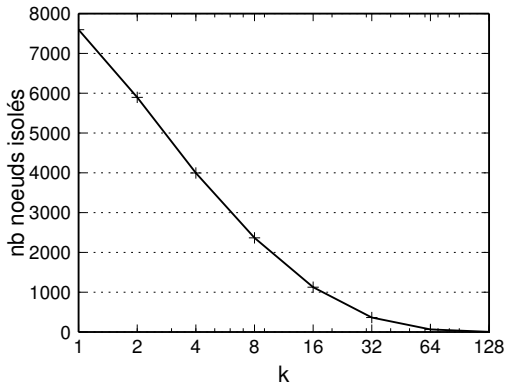


Figure 7.16 – Nombre de nœuds isolés dans le graphe de k PPV mutuel en fonction de k (abscisse en échelle logarithmique).

Selon l'application envisagée, la présence de nœuds isolés, ou une proportion trop importante de nœuds isolés, peut être problématique. Si tel est le cas, on peut bien sûr augmenter la valeur de k, ou encore opter pour un graphe orienté ou symétrique. Si la présence de nœuds isolés n'est pas un problème en soi, mais que l'on souhaite tout de même réduire leur nombre (ou que l'on juge la densité du graphe trop faible pour quelque raison que ce soit), une autre stratégie, éventuellement plus efficace[13], peut être

[12] À titre indicatif, ces UL sont *amplifier, amplification, raréfier, reproduire, transformation, industrialisation, microclimat, propulser, décharger, conduite* et *recul*.

[13] Cette stratégie peut donner de meilleurs résultats que les deux autres stratégies que nous avons mentionnées parce qu'elle garantit théoriquement que l'on obtiendra le meilleur compromis possible entre la précision et le rappel étant donné l'importance relative accordée à chaque mesure. En outre, elle peut indiquer qu'un graphe orienté ou symétrique donnerait de meilleurs résultats, ce qui n'est pas le cas si on augmente simplement la valeur de k jusqu'à ce que la densité du graphe ou le nombre de nœuds isolés soit jugé acceptable.

envisagée (si l'on dispose de données de référence, bien entendu) : optimiser la mesure F_β de sorte que l'on accorde plus d'importance au rappel. Comme nous l'avons expliqué au chapitre 5, plus le graphe est dense, plus le rappel tend à être élevé et la précision, faible. Donc, un graphe qui offre un rappel élevé aura une densité relativement élevée, et un nombre relativement faible de nœuds isolés. Et si on pondère la mesure F_β de sorte qu'elle accorde plus d'importance au rappel qu'à la précision (en donnant à β une valeur supérieure à 1, comme nous l'avons expliqué au chapitre 5), le graphe qui maximise cette mesure tendra à être plus dense et à contenir moins de nœuds isolés que si on utilisait une mesure F_β qui accorde autant d'importance à la précision qu'au rappel.

Au départ, nous avons fixé β à 1 parce que nous n'avions aucun présupposé par rapport à l'importance relative de la précision et du rappel. Étant donné que le graphe que nous avons retenu contient une proportion importante de nœuds isolés, nous vérifierons maintenant quels résultats on obtient en optimisant une mesure F_β où β a une valeur plus élevée, à savoir 2 ; nous supposons donc maintenant que le rappel est deux fois plus important que la précision, pour obtenir un graphe plus dense et contenant moins de nœuds isolés.

k	F_2
1	0.1370
2	0.2376
4	0.2971
8	**0.3165**
16	0.2837
32	0.2186
64	0.1514
128	0.0935

(a) Orienté

k	F_2
1	0.1851
2	0.3153
4	**0.3194**
8	0.2837
16	0.2160
32	0.1473
64	0.0948
128	0.0566

(b) Symétrique

k	F_2
1	0.0757
2	0.1151
4	0.2314
8	0.3174
16	**0.3464**
32	0.3116
64	0.2358
128	0.1394

(c) Mutuel

Tableau 7.VI – Mesure F_2 des graphes construits sur le modèle retenu, en fonction de k et du type de graphe.

Étant donné le modèle que nous avons retenu, le graphe qui atteint la mesure F_2

la plus élevée est un graphe mutuel[14] avec $k = 16$, comme le montre le Tableau 7.VI. Soulignons qu'il serait possible d'optimiser non seulement le graphe, en choisissant celui qui maximise la nouvelle mesure F_β (avec $\beta = 2$), mais aussi le modèle distributionnel utilisé pour construire le graphe, mais ici, nous supposons que le modèle utilisé demeure le même.

Dans le graphe qui obtient la meilleure mesure F_2 (graphe mutuel avec $k = 16$), le nombre de nœuds isolés est de 1126 plutôt que 2366, et seulement 4 des UL dans les ensembles de référence sont sans voisin, plutôt que 11.

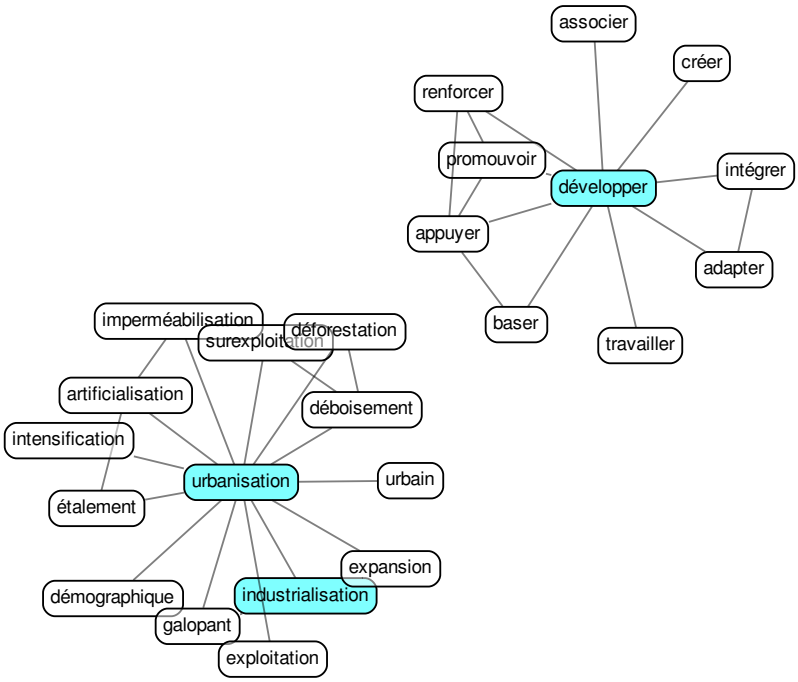


Figure 7.17 – 1-voisinage du cadre Cause_change_into_organized_society ($k = 16$).

La Figure 7.17 montre le 1-voisinage du cadre Cause_change_into_organized_society dans le graphe mutuel avec $k = 16$. Si on compare ce voisinage à celui illustré dans la Fi-

[14]Il est intéressant de noter que, pour le modèle que nous avons retenu, les graphes qui maximisent les mesures F_β pour $\beta \in \{1,2,3,4,5\}$ sont tous des graphes mutuels.

gure 7.15, on observe que la densité du voisinage a effectivement augmenté, et que l'UL *industrialisation* n'est plus un nœud isolé, étant maintenant reliée à l'UL *urbanisation*. On observe également que le voisinage de *développer* demeure séparé de celui des deux autres UL.

7.6 Synthèse

Dans ce chapitre, nous nous sommes penchés sur l'évaluation des graphes de voisinage que nous utilisons pour interroger les modèles distributionnels. Un de nos objectifs était d'identifier le paramétrage optimal de ces graphes. Nous avons d'abord montré comment les paramètres du graphe de k PPV, à savoir la valeur de k et le type de graphe, influencent la taille des voisinages, puis comment ils influencent leur qualité, celle-ci étant évaluée automatiquement au moyen de la précision, du rappel et de la mesure F_1. Cette analyse a montré que la valeur de k qui maximise la mesure F_1 est peu élevée ($k = 4$, en moyenne[15]), et que ce sont les graphes mutuels qui offrent le meilleur compromis entre précision et rappel.

Par la suite, nous avons présenté des exemples visuels de voisinages distributionnels. Pour choisir le graphe dont nous avons extrait ces exemples, nous avons choisi de nous concentrer sur le français et sur le cadre descriptif de la sémantique des cadres. Après avoir paramétré le modèle distributionnel en fonction de ces décisions, nous avons choisi le graphe qui offrait, pour ce modèle, la meilleure mesure F_1, à savoir un graphe de k PPV mutuel avec $k = 8$.

Ce graphe a obtenu un rappel de 0.3169 et une précision de 0.3195 sur les ensembles de référence en français. Cela signifie que si on compare le 1-voisinage de chaque requête (chaque UL dans les ensembles de référence) à l'ensemble des termes qui évoquent un des cadres évoqués par la requête, un peu moins d'un tiers de ces termes reliés se retrouvent parmi les voisins de la requête, et un peu moins d'un tiers des voisins sont

[15] Cette valeur est celle qui maximise la mesure F_1, en moyenne, pour les trois types de graphe. Rappelons qu'il y a une interaction entre le type de graphe et la valeur optimale de k.

effectivement des termes reliés, en moyenne. Il serait facile d'obtenir un rappel parfait : il suffit d'utiliser une valeur suffisamment élevée de k. De même, il est possible d'augmenter la précision, notamment en utilisant une valeur plus faible de k, la précision maximale que nous avons observée étant de 0.4081 sur les ensembles de référence en français. Or, l'objectif ici n'était pas d'obtenir une précision parfaite, mais un bon compromis entre la précision et le rappel, d'où l'utilisation de la mesure F_1. Étant donné que nous savons *a priori* que les données de référence n'offrent qu'une couverture partielle des relations lexicales entre les termes du domaine de l'environnement (en l'occurrence, la relation d'appartenance au même cadre sémantique), l'objectif n'est pas d'atteindre une précision de 100% : on s'attend en fait à ce que le voisinage des requêtes contienne des termes qui n'ont pas encore été encodés dans la ressource, mais qui pourraient y être ajoutés.

Ainsi, l'équilibre entre précision et rappel que l'on arrive à obtenir avec un graphe de voisinage distributionnel nous semble très bon, bien qu'il demeure sans doute possible d'obtenir une mesure F_1 plus élevée. De plus, les exemples de voisinages que nous avons présentés montrent bien qu'ils peuvent servir à faciliter et à enrichir l'analyse des relations lexicales dans le cadre du travail terminologique.

Nous avons d'abord montré le 0-voisinage de plusieurs ensembles de référence, ces exemples montrant, entre autres, que le rappel varie beaucoup d'un ensemble à l'autre. Nous avons ensuite montré des exemples de 1-voisinages, ce qui nous a permis d'observer non seulement le rappel que l'on obtient pour chacun des cadres sémantiques utilisés comme exemples, mais aussi la précision de leur voisinage ; du coup, cela nous a donné une meilleure idée de l'utilité des graphes de voisinage distributionnel pour la construction et l'enrichissement de ressources lexicales. Dans ces exemples, le voisinage distributionnel d'un ensemble de termes évoquant un cadre sémantique contenait des termes évoquant d'autres cadres évoqués par les termes dans la requête, ou d'autres cadres reliés à ceux-ci. Les exemples de 1-voisinages ont également montré que la visualisation de ces sous-graphes constitue un moyen efficace d'interroger un modèle distributionnel, car

elle nous permet d'observer la structure du voisinage d'un terme ou d'un ensemble de termes. En somme, les 0-voisinages sont un moyen efficace d'identifier les relations existant au sein d'un ensemble de termes connu *a priori*, ou des sous-ensembles de termes reliés entre eux, tandis que les 1-voisinages permettent d'identifier des termes sémantiquement reliés à une requête (un terme ou un ensemble de termes) en plus de permettre l'identification des relations existant entre tous ces termes (ceux de la requête et leurs voisins).

Nous avons ensuite montré que l'on peut explorer davantage le voisinage d'une requête en faisant deux sauts dans le graphe plutôt qu'un seul. Puis, nous avons montré que l'on peut optimiser une mesure F_β pondérée de sorte que l'on accorde plus d'importance au rappel ou à la précision afin de favoriser un graphe plus ou moins dense. Nous avons vérifié quel graphe maximisait, sur le modèle que nous avons retenu, la mesure F_2, qui accorde deux fois plus d'importance au rappel qu'à la précision, et cela nous a amené à opter pour le même type de graphe (mutuel), mais une valeur plus élevée de k (16 plutôt que 8). Ce graphe offrait un rappel plus élevé, des voisinages plus denses, et un nombre moins élevé de nœuds isolés.

CHAPITRE 8

CONCLUSION

L'objectif principal de cette thèse était de développer un cadre méthodologique basé sur la sémantique distributionnelle pour l'analyse des relations lexicales paradigmatiques auxquelles participent les termes d'un domaine de spécialité. Ce cadre méthodologique concerne le choix d'une méthode pour construire le modèle distributionnel et la sélection de ses (hyper)paramètres, ainsi que le choix d'une méthode pour interroger le modèle. Ces différentes décisions dépendent à leur tour de facteurs liés à l'application envisagée, tels que le cadre descriptif adopté et les relations que l'on souhaite identifier.

Au chapitre 2, nous avons décrit deux cadres descriptifs utilisés dans le domaine de la terminologie que nous avons exploités pour encadrer l'application et l'évaluation de l'approche distributionnelle. Nous avons également décrit deux ressources terminologiques qui reflètent ces deux cadres descriptifs, dont nous avons extrait des données de référence que nous avons utilisées afin d'évaluer des modèles distributionnels.

Au chapitre 3, nous avons présenté une revue de la littérature sur la détection automatique de relations lexicales, et particulièrement sur la sémantique distributionnelle, l'accent étant placé sur l'évaluation des méthodes distributionnelles et les travaux exploitant des corpus spécialisés. Nous avons également illustré comment les modèles distributionnels sont construits et interrogés.

Au chapitre 4, nous avons d'abord énoncé les problématiques que nous avons identifiées en ce qui concerne la mise en œuvre de l'approche distributionnelle dans le cadre du travail terminologique, et que nous avons abordées dans la suite de ce travail. Ces problématiques concernent divers aspects de l'application envisagée (en l'occurrence, une description terminologique basée sur un cadre descriptif particulier) et la façon dont ceux-ci influencent la qualité des résultats qu'offrent les modèles distributionnels ainsi que la façon optimale de construire et d'interroger ces modèles ; ces facteurs comprennent

le cadre descriptif adopté, les relations ciblées et la langue traitée. Nous avons ensuite énoncé nos objectifs, après avoir formulé deux hypothèses générales, chacune étant liée à un des deux cadres descriptifs que nous avons exploités :

- **Hypothèse 1** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme donné, des termes reliés sur le plan paradigmatique, tels que des (quasi-)synonymes, des antonymes, des hyperonymes, des hyponymes et des dérivés syntaxiques.

- **Hypothèse 2** : Un modèle sémantique distributionnel construit à partir d'un corpus spécialisé permet d'obtenir, pour un terme ou un ensemble de termes donné, des termes évoquant le même cadre sémantique.

Au chapitre 5, nous avons présenté notre méthodologie. Celle-ci consiste à construire des modèles distributionnels à partir de corpus spécialisés du domaine de l'environnement et à évaluer ces modèles au moyen des données de référence que nous avons extraites de dictionnaires spécialisés.

Au chapitre 6, nous avons présenté les résultats de l'évaluation des modèles distributionnels, celle-ci visant à déterminer la qualité du classement des voisins distributionnels que l'on obtient pour chaque mot-cible à partir de ces modèles. La précision élevée de ces classements par rapport aux données de référence tend à confirmer nos deux hypothèses, particulièrement la première : nos résultats indiquent que les modèles distributionnels captent l'appartenance au même cadre sémantique moins bien que certaines relations paradigmatiques spécifiques, mais la précision que nous avons observée sur les cadres sémantiques demeure élevée.

Nos résultats indiquent notamment que l'analyse distributionnelle (AD) offre de meilleurs résultats que `word2vec` sur la plupart des jeux de données utilisés dans ce travail. Si le choix de la méthode utilisée pour construire le modèle détermine dans une certaine mesure la qualité des résultats, le paramétrage de la méthode a aussi une influence importante sur celle-ci, et parfois plus importante, selon l'(hyper)paramètre pris

en considération. En analysant l'influence des (hyper)paramètres de ces deux méthodes, nous avons pu identifier leurs valeurs optimales et déterminer lesquels ont une influence particulièrement importante sur la qualité des résultats.

L'analyse présentée dans ce chapitre a montré que des facteurs tels que le cadre descriptif adopté et les relations ciblées ont une influence importante sur la qualité des résultats et sur les modèles qui offrent les meilleurs résultats, tandis que la langue traitée ne joue pas un rôle très important à cet égard, du moins dans le cas du français et de l'anglais. Ainsi, si nous n'avons pas discuté davantage l'influence de la langue traitée, ce n'est pas parce que nous ne nous sommes pas intéressé aux différences entre les deux langues ; c'est seulement que l'analyse des résultats n'a pas révélé de différences importantes à cet égard. D'ailleurs, la similarité des résultats obtenus dans les deux langues indique que les tendances que nous avons observées ne sont pas propres à un corpus particulier et à un ensemble particulier de données de référence, mais ont une portée plus générale.

Au chapitre 7, nous avons présenté les résultats de l'évaluation de différents types de graphes que nous avons utilisés afin d'interroger les modèles distributionnels. Nous avons d'abord analysé l'influence des paramètres de ces graphes sur la taille et la précision des voisinages distributionnels qu'ils représentent. Après avoir identifié un paramétrage offrant de bons résultats, nous avons présenté des exemples visuels de voisinages distributionnels et montré que ces graphes offrent un moyen efficace d'analyser le voisinage d'un terme ou d'un ensemble de termes. Dans ces exemples (plus précisément, dans les exemples de voisinages à 1 ou 2 sauts), le voisinage distributionnel d'un ensemble donné de termes qui évoquent un cadre sémantique contenait des termes évoquant d'autres cadres évoqués par les termes dans cet ensemble, ou des cadres reliés à ceux-ci. Nous avons ensuite exploré un moyen d'ajuster la densité du graphe tout en maximisant la qualité des voisinages qu'il représente.

8.1 Contributions

Plusieurs aspects de cette thèse constituent, à notre connaissance, des contributions originales par rapport aux problématiques identifiées au chapitre 4. Nous résumons ces contributions ci-dessous, celles-ci étant regroupées en fonction des différentes dimensions du cadre méthodologique que cette thèse visait à développer.

Le cadre descriptif Nous avons présenté une évaluation de méthodes distributionnelles encadrée par deux cadres descriptifs spécifiques du domaine de la terminologie, à savoir l'approche lexico-sémantique et l'approche basée sur les cadres sémantiques. Nos résultats indiquent que l'appartenance au même cadre sémantique est moins bien détectée par les modèles distributionnels que certaines relations paradigmatiques spécifiques, telles que la (quasi-)synonymie et la dérivation syntaxique. Cela est attribuable au fait que les verbes, qui sont moins bien modélisés par cette approche que les noms et les adjectifs, occupent une place importante dans les cadres sémantiques, et au fait qu'il existe des différences importantes entre les modèles qui captent le mieux les différentes relations auxquelles participent les termes qui évoquent le même cadre sémantique.

Les relations ciblées Nous avons évalué systématiquement la capacité des modèles distributionnels à capter, à partir de corpus spécialisés, quatre types de relations lexicales paradigmatiques, y compris la dérivation syntaxique. Nos résultats indiquent que les modèles distributionnels captent la dérivation syntaxique aussi bien que la (quasi-)synonymie, mais les résultats obtenus sur les dérivés sont particulièrement sensibles au paramétrage du modèle, et les paramétrages optimaux pour cette relation sont très différents de ceux pour la (quasi-)synonymie et les autres relations paradigmatiques classiques.

La langue traitée En évaluant les résultats obtenus dans deux langues, le français et l'anglais, nous avons montré que la qualité des résultats est tout à fait comparable, et que

l'influence du paramétrage des modèles est très similaire dans les deux langues.

Le choix de la méthode distributionnelle Nous avons comparé un modèle de langue neuronal à la technique classique d'AD sur des corpus spécialisés et des données de référence extraites de dictionnaires spécialisés. Nos résultats indiquent que l'AD offre une précision plus élevée que `word2vec` sur la plupart des jeux de données utilisés dans ce travail, mais que `word2vec` détecte mieux la dérivation syntaxique. Ils indiquent également que la différence de précision entre les deux méthodes, lorsqu'elles sont paramétrées correctement, se limite généralement à quelques points de MAP.

L'influence des (hyper)paramètres Nous avons analysé systématiquement l'influence des hyperparamètres d'un modèle de langue neuronal sur sa capacité à détecter différentes relations lexicales. Nous avons notamment observé que le seuil pour le sous-échantillonnage des mots fréquents dans `word2vec` exerce une influence très importante sur la précision du classement des voisins distributionnels. L'importance du sous-échantillonnage est une observation intéressante en soi, d'autant plus que cette technique peut être utilisée non seulement avec `word2vec`, mais aussi lors du calcul de l'AD [110]. Nous avons également analysé l'influence de certains des paramètres de l'AD. Dans les deux cas, nous avons vérifié comment l'influence des (hyper)paramètres varie en fonction des différents aspects du projet terminologique mentionnés ci-dessus.

Le mode d'interrogation Nous avons interrogé des modèles distributionnels non seulement en obtenant la liste triée des voisins de chaque requête, mais aussi au moyen de graphes de voisinage ; nous avons d'ailleurs comparé trois types de graphes et analysé l'influence du nombre de voisins pris en compte.

Ces contributions nous ont permis, à notre avis, de faire avancer de façon significative le développement d'un cadre méthodologique basé sur la sémantique distributionnelle pour l'identification des relations auxquelles participent les termes d'un domaine spécialisé, bien que nous n'ayons pas couvert tous les aspects possibles d'un tel cadre,

comme nous le soulignons à la section 8.2. Ainsi, nous avons pu réaliser l'objectif principal de ce travail et, du coup, les sous-objectifs que nous avons énumérés au chapitre 4.

Soulignons enfin que le domaine que nous avons traité était celui de l'environnement, un domaine qui a reçu très peu d'attention dans les travaux en sémantique distributionnelle.

8.2 Limites et perspectives

Les facteurs que nous avons pris en compte dans ce travail comprennent certains (hyper)paramètres des modèles distributionnels et des méthodes utilisées pour les interroger, ainsi que certains facteurs liés à l'application envisagée, tels que le cadre descriptif adopté et les relations ciblées. Nous avons ainsi pris en compte un ensemble de facteurs qui nous semblaient particulièrement importants pour réaliser notre objectif principal, mais nous tenons à souligner, parmi les limites de ce travail, que nous n'avons pas inclus tous les facteurs possibles.

Premièrement, nous n'avons pas examiné l'influence de tous les (hyper)paramètres des deux méthodes utilisées pour construire les modèles. Par exemple, dans le cas de `word2vec`, nous n'avons pas pris en compte certains hyperparamètres qui n'ont généralement pas besoin d'être optimisés selon la documentation de cet outil, tels que le taux d'apprentissage. Soulignons également que plusieurs des hyperparamètres de `word2vec` sont aussi des paramètres de l'AD ou peuvent être adaptés à celle-ci [110], et vice-versa. Ainsi, il est possible d'analyser l'influence d'un même (hyper)paramètre sur la qualité des deux types de modèles, comme nous l'avons fait pour la taille de la fenêtre de contexte. On pourrait donc analyser un même ensemble d'(hyper)paramètres pour les deux méthodes. Or, nos résultats montrent que l'influence de la taille de fenêtre est très similaire pour les deux types de modèles. On pourrait supposer que ce serait le cas pour tout (hyper)paramètre qui s'applique à différentes méthodes pour construire des modèles distributionnels.

Une autre limite de ce travail est liée aux valeurs testées pour chacun des (hyper)paramètres pris en compte. Pour certains de ceux-ci, tels que l'architecture du modèle dans le cas de `word2vec`, nous avons testé toutes les valeurs possibles, mais dans d'autres cas, nous avons restreint le nombre de valeurs testées pour différentes raisons, notamment pour limiter le temps de calcul et la complexité des analyses. On aurait une meilleure idée de l'influence de ces (hyper)paramètres si on testait un nombre plus élevé de valeurs. Il serait notamment intéressant d'évaluer d'autres types de graphes pour l'interrogation des modèles distributionnels, tels que la ε*-method* utilisée par Steyvers et Tenenbaum [177] ou le graphe de voisinage relatif utilisé par Gyllensten et Sahlgren [85].

Un autre facteur dont nous n'avons pas analysé l'influence est la taille des corpus. L'utilisation d'un corpus plus petit ou plus gros nous amènerait éventuellement à opter pour une méthode distributionnelle différente ou à paramétrer le modèle d'une façon différente, par exemple. Plusieurs études ont abordé l'influence de la taille des corpus, comme nous l'avons souligné au chapitre 3. Ces travaux suggèrent que l'approche distributionnelle n'offre pas de bons résultats pour les mots très peu fréquents. Pour notre part, nous avons observé que les mots-cibles les moins fréquents étaient en fait ceux pour lesquels on obtenait les meilleurs résultats, mais la fréquence minimale des mots-cibles utilisés dans ce travail est relativement élevée.

De plus, nous n'avons pas examiné l'influence du nombre de mots-contextes (dans le cas de l'AD) ou du seuil de fréquence minimale (dans le cas de `word2vec`). Les résultats de Bullinaria et Levy [34] suggèrent qu'il est inutile d'utiliser plus de 50 000 mots-contextes, et qu'il n'est généralement pas très bénéfique d'utiliser plus de 10 000 mots-contextes, mais il aurait peut-être été possible d'augmenter légèrement la précision en utilisant plus de 10 000 mots-contextes (ou en modifiant le seuil de fréquence minimale dans le cas de `word2vec`).

Les facteurs dont nous n'avons pas examiné l'influence comprennent aussi la mesure de similarité. Dans ce travail, nous avons utilisé une seule mesure de similarité, le cosinus

de l'angle des vecteurs, mais il existe beaucoup d'autres mesures qui peuvent servir à estimer la similarité distributionnelle.

Par ailleurs, nous n'avons pas traité toutes les interactions possibles entre les différents facteurs dont dépend la qualité des résultats qu'offre l'approche distributionnelle. Par exemple, dans le cas de l'évaluation des graphes de voisinage, nous nous sommes concentré sur une langue et un cadre descriptif.

En outre, nous avons seulement évalué deux méthodes pour construire les modèles, afin d'explorer la différence entre la technique classique d'AD et les modèles de langue neuronaux. On pourrait enrichir ce travail en évaluant d'autres techniques distributionnelles, telles que celles visant à tenir compte de l'ordre des mots [100, 166] ou d'autres modèles de langue neuronaux. Une méthode qui nous semble particulièrement intéressante pour l'application que nous avons envisagée est la méthode basée sur les patrons lexico-syntaxiques proposée par Schwartz et al. [174], puisqu'elle capte très bien la similarité entre les verbes (celle-ci étant moins bien captée par les modèles que nous avons utilisés que celle entre les noms ou les adjectifs), et qu'elle peut être adaptée pour cibler des relations lexicales particulières.

En ce qui concerne l'implémentation de l'approche distributionnelle, soulignons enfin que nous avons seulement exploité un type de contexte, à savoir les cooccurrents graphiques, et que l'on pourrait analyser l'influence de ce facteur en comparant ceux-ci à d'autres types de contextes, tels que les cooccurrents syntaxiques. Cette question a fait l'objet de nombreux travaux, comme nous l'avons souligné au chapitre 3.

Quant aux méthodes d'évaluation utilisées, la principale consiste à évaluer de manière quantitative les voisinages distributionnels (listes triées de PPV et graphes de k PPV) au moyen de données de référence. Nous considérons donc que les données de référence représentent une sorte d'approximation du jugement du terminologue, mais il serait enrichissant de demander à un terminologue d'évaluer manuellement les résultats produits par les méthodes distributionnelles, afin d'en apprendre davantage sur l'utilité de ces méthodes ainsi que leurs faiblesses. Rappelons que des évaluations manuelles ont

été réalisées dans le cadre de deux études reliées à cette thèse [15, 20]. Ces études ont confirmé que l'évaluation automatique au moyen de données de référence sous-estime la précision des voisinages distributionnels, et nous ont fourni une perspective plus complète des relations existant au sein de ces voisinages.

Par ailleurs, dans cette thèse, nous avons évalué directement la précision des voisinages distributionnels, mais nous n'avons pas abordé d'autres façons d'exploiter les modèles distributionnels. Il serait intéressant à cet égard de vérifier si la classification automatique de représentations distributionnelles permettrait d'identifier des ensembles de termes évoquant le même cadre sémantique.

De plus, il serait intéressant de comparer les résultats obtenus dans différents domaines. Un plus grand nombre d'évaluations enrichirait le cadre méthodologique proposé et augmenterait les chances d'obtenir de bons résultats lors de son application dans de nouveaux contextes, comme le soulignent Bullinaria et Levy [34, p. 901] : « It is clear that one should be wary of relying on a single task to optimize methods for use on other tasks. »

Une autre possibilité qui nous semble intéressante consisterait à identifier, dans l'espace sémantique que représente un modèle distributionnel, des vecteurs correspondant à des relations lexicales particulières, à partir d'exemples de paires de termes participant à ces relations, comme le font Mikolov et al. [133] pour résoudre des analogies syntaxiques et sémantiques. Ces vecteurs pourraient ensuite être utilisés pour détecter d'autres paires de termes participant à ces relations.

Comme nous l'avons montré au chapitre 6, les modèles qui détectent le mieux la dérivation syntaxique sont très différents de ceux qui produisent les meilleurs résultats pour les relations paradigmatiques classiques telles que la (quasi-)synonymie. Il serait donc intéressant de vérifier si on peut combiner ces deux types de modèles pour mieux détecter les termes qui évoquent le même cadre sémantique ou d'autres relations lexicales, une possibilité que nous avons explorée ailleurs [18].

Une autre perspective à explorer consisterait à exploiter les résultats présentés dans

cette thèse afin de développer des techniques pour adapter automatiquement les méthodes distributionnelles en fonction des caractéristiques d'un projet terminologique donné.

Enfin, il serait intéressant de vérifier si on peut combiner les représentations distributionnelles et d'autres sources d'information sur la similarité des termes pour mieux assister la description des relations auxquelles ils participent. À ce sujet, Hill et al. [96] soulignent que pour mieux modéliser la similarité sémantique, il pourrait être nécessaire de prendre en compte des phénomènes linguistiques autres que la cooccurrence, voire d'autres modalités de la perception humaine.

BIBLIOGRAPHIE

[1] Clémentine Adam, Cécile Fabre et Philippe Muller. Évaluer et améliorer une ressource distributionnelle : Protocole d'annotation de liens sémantiques en contexte. *TAL*, 54(1):71–97, 2013.

[2] Clémentine Adam, Cécile Fabre et Ludovic Tanguy. Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle. Dans *Actes de la 20e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 140–153, 2013.

[3] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca et Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. Dans *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. ACL, 2009.

[4] Sue Atkins, Charles J. Fillmore et Christopher R. Johnson. Lexicographic relevance : Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3):251–280, 2003.

[5] Alain Auger et Caroline Barrière. Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, 14(1):1–19, 2008.

[6] Dzmitry Bahdanau, Kyunghyun Cho et Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

[7] Collin F. Baker. FrameNet : A knowledge base for natural language processing. Dans *Proceedings of Frame Semantics in NLP : A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, 2014. ACL.

[8] Collin F. Baker, Charles J. Fillmore et John B. Lowe. The Berkeley FrameNet project. Dans *Proceedings of the 36th Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*, volume 1, pages 86–90, Montréal, 1998. ACL.

[9] A. Baneyx, V. Malaisé, J. Charlet, P. Zweigenbaum et B. Bachimont. Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. Dans *Proceedings of Terminology and Artificial Intelligence (TIA)*, Rouen, 2005.

[10] Marco Baroni, Georgiana Dinu et Germán Kruszewski. Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, Baltimore, 2014. ACL. URL `http://www.aclweb.org/anthology/P14-1023`.

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

[12] Marco Baroni et Alessandro Lenci. How we BLESSed distributional semantic evaluation. Dans *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 1–10. ACL, 2011.

[13] Marco Baroni et Roberto Zamparelli. Nouns are vectors, adjectives are matrices : Representing adjective-noun constructions in semantic space. Dans *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193. ACL, 2010. URL `http://dl.acm.org/citation.cfm?id=1870658.1870773`.

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent et Christian Jauvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3: 1137–1155, 2003.

[15] Gabriel Bernier-Colborne. Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques. Dans *Actes de SemDis : Enjeux actuels de la sémantique distributionnelle*, pages 238–251, Marseille, 2014.

[16] Gabriel Bernier-Colborne. Identifying semantic relations in a specialized corpus through distributional analysis of a cooccurrence tensor. Dans *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 57–62, Dublin, 2014. ACL/DCU. URL `http://www.aclweb.org/anthology/S14-1007`.

[17] Gabriel Bernier-Colborne. Exploration de modèles distributionnels au moyen de graphes 1-PPV. Dans *Actes de la 22e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 593–599, Caen, 2015. ATALA.

[18] Gabriel Bernier-Colborne et Patrick Drouin. Combiner des modèles sémantiques distributionnels pour mieux détecter les termes évoquant le même cadre sémantique. Dans *Actes de la 23e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 381–388, Paris, 2016.

[19] Gabriel Bernier-Colborne et Patrick Drouin. Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique. Dans *Actes de la 23e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 125–138, Paris, 2016.

[20] Gabriel Bernier-Colborne et Marie-Claude L'Homme. Using a distributional neighbourhood graph to enrich semantic frames in the field of the environment. Dans *Proceedings of the conference Terminology and Artificial Intelligence (TIA)*, pages 9–16, Granada, 2015.

[21] Ann Bertels et Dirk Speelman. La contribution des cooccurrences de deuxième ordre à l'analyse sémantique. *Corpus*, 11:147–165, 2012.

[22] Ann Bertels et Dirk Speelman. Analyse de positionnement multidimensionnel sur le corpus spécialisé TALN. Dans *Actes de SemDis : Enjeux actuels de la sémantique distributionnelle*, pages 252–265, Marseille, 2014.

[23] Ann Bertels et Dirk Speelman. Analyse exploratoire des cooccurrents de premier ordre dans un corpus technique. Dans *Actes des 12es Journées d'analyse statistique des données textuelles (JADT)*, pages 67–78, 2014.

[24] David M. Blei, Andrew Y. Ng et Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[25] Vincent D. Blondel et Pierre Senellart. Automatic extraction of synonyms in a dictionary. Dans *Proceedings of TMW*, pages 7–13, Arlington, 2002.

[26] Olivier Bodenreider, Anita Burgun et Thomas C. Rindflesch. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. Dans *Proceedings of Terminology and Artificial Intelligence (TIA)*, pages 11–21, Nancy, 2001.

[27] Jacques Bouaud, Benoît Habert, Adeline Nazarenko et Pierre Zweigenbaum. Regroupements issus de dépendances syntaxiques en corpus : Catégorisation et confrontation à deux modélisations conceptuelles. Dans *Actes des journées Ingénierie des connaissances*, pages 207–223, 1997.

[28] Didier Bourigault. UPERY : Un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Dans *Actes de la 9e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 75–84, Nancy, 2002.

[29] Didier Bourigault et Monique Slodzian. Pour une terminologie textuelle. *Terminologies nouvelles*, 19:29–32, 1999.

[30] Antoine Bride, Tim Van de Cruys et Nicolas Asher. Une évaluation approfondie de différentes méthodes de compositionalité sémantique. Dans *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 149–160, Marseille, 2014. ATALA. URL `http://www.aclweb.org/anthology/F14-1014`.

[31] Elia Bruni, Nam-Khanh Tran et Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

[32] John A. Bullinaria. Semantic categorization using simple word co-occurrence statistics. Dans *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 1–8, 2008.

[33] John A. Bullinaria et Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior research methods*, 39(3):510–526, 2007.

[34] John A. Bullinaria et Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD. *Behavior research methods*, 44(3):890–907, 2012.

[35] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. Dans *Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL '99)*, pages 120–126, College Park, 1999. ACL.

[36] Nathanael Chambers et Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. Dans *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 602–610. ACL, 2009.

[37] Zheng Chen et Heng Ji. Graph-based clustering for computational linguistics : A survey. Dans *Proceedings of the 2010 Workshop on Graph-based Methods for*

Natural Language Processing, pages 1–9. Association for Computational Linguistics, 2010.

[38] Kenneth Church. The case for empiricism (with and without statistics). Dans *Proceedings of Frame Semantics in NLP : A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 6–9, Baltimore, 2014. ACL.

[39] Kenneth Ward Church et Patrick Hanks. Word association norms, mutual information, and lexicography. Dans *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, 1989.

[40] Stephen Clark. Vector space models of lexical meaning (draft chapter). Dans Shalom Lappin et Chris Fox, éditeurs, *The Handbook of Contemporary Semantic Theory, 2nd Edition*. Wiley-Blackwell, 2015. URL `http://www.cl.cam.ac.uk/~sc609/pubs/sem_handbook.pdf`.

[41] Vincent Claveau, Ewa Kijak et Olivier Ferret. Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. Dans *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 220–231, Marseille, 2014. ATALA. URL `http://www.aclweb.org/anthology/F14-1020`.

[42] Trevor Cohen et Dominic Widdows. Empirical distributional semantics : Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.

[43] Ronan Collobert et Jason Weston. A unified architecture for natural language processing : Deep neural networks with multitask learning. Dans *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM, 2008.

[44] D.A. Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, 1986.

[45] Dipanjan Das. Statistical models for frame-semantic parsing. Dans *Proceedings of Frame Semantics in NLP : A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 26–29, Baltimore, 2014. ACL.

[46] Dmitry Davidov et Ari Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 297–304. ACL, 2006.

[47] Diego De Cao, Danilo Croce, Marco Pennacchiotti et Roberto Basili. Combining word sense and usage for modeling frame semantics. Dans *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 85–101. ACL, 2008.

[48] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer et Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[49] Georgiana Dinu et Marco Baroni. How to make words with vectors : Phrase generation in distributional semantics. Dans *Proceedings of ACL*, pages 624–633, 2014.

[50] Andrew Dolbey, Michael Ellsworth et Jan Scheffczyk. BioFrameNet : A domain-specific FrameNet extension with links to biomedical ontologies. Dans *Proceedings of KR-MED*, 2006.

[51] Beate Dorow et Dominic Widdows. Discovering corpus-specific word senses. Dans *Proceedings of the tenth conference of the European chapter of the Association for Computational Linguistics*, volume 2, pages 79–82. ACL, 2003.

[52] Patrick Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115, 2003.

[53] Marie Dupuch, Laëtitia Dupuch, Thierry Hamon et Natalia Grabar. Inferring semantic relations between pharmacovigilance terms with the NLP methods. Dans *Proceedings of Computational Methods in Pharmacovigilance*, pages 27–31, 2012.

[54] Marie Dupuch, Laëtitia Dupuch, Thierry Hamon et Natalia Grabar. Semantic distance and terminology structuring methods for the detection of semantically close terms. Dans *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 20–28. ACL, 2012.

[55] Marie Dupuch, Laëtitia Dupuch, Amandine Périnet, Thierry Hamon et Natalia Grabar. Structuration de terminologies pour la création de groupements de termes en pharmacovigilance. Dans *Actes des 11es Journées internationales d'analyse statistique des données textuelles (JADT)*, pages 431–444, 2012.

[56] Marie Dupuch, Thierry Hamon et Natalia Grabar. Cross-language detection of linguistic and semantic regularities in pharmacovigilance terms. Dans *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis*, 2013.

[57] Marie Dupuch, Amandine Périnet, Thierry Hamon et Natalia Grabar. Utilisation de méthodes de structuration de terminologies pour la création de groupements de termes de pharmacovigilance. Dans *Long papers of the 9th International Conference on Terminology and Artificial Intelligence (TIA)*, pages 3–9, 2011.

[58] Mehdi Embarek et Olivier Ferret. Learning patterns for building resources about semantic relations in the medical domain. Dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008.

[59] Stefan Evert. Corpora and collocations (extended manuscript). Dans Anke Lüdeling et Merja Kytö, éditeurs, *Corpus Linguistics. An International Handbook*, volume 2. Walter de Gruyter, Berlin/New York,

2007. URL http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf.

[60] Pamela Faber, éditeur. *A Cognitive Linguistics View of Terminology and Specialized Language*. De Gruyter, Berlin/Boston, 2012.

[61] Cécile Fabre, Nabil Hathout, Franck Sajous et Ludovic Tanguy. Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. Dans *Actes de SemDis : Enjeux actuels de la sémantique distributionnelle*, pages 266–279, Marseille, 2014.

[62] Cécile Fabre et Alessandro Lenci. Distributional semantics today : Introduction to the special issue. *TAL*, 56(2):7–20, 2015.

[63] Manaal Faruqui et Chris Dyer. Community evaluation and exchange of word vectors at wordvectors.org. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Baltimore, 2014. ACL.

[64] Christiane Fellbaum. *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

[65] Olivier Ferret. Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales. Dans *Actes de la 11e conférence sur le traitement automatique des langues naturelles (TALN)*, Fès, 2004.

[66] Olivier Ferret. Similarité sémantique et extraction de synonymes à partir de corpus. Dans *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN)*, Montréal, 2010.

[67] Olivier Ferret. Combining bootstrapping and feature selection for improving a distributional thesaurus. Dans *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pages 336–341, Montpellier, 2012.

[68] Olivier Ferret. Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. Dans *Actes de la 20e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 48–61, 2013.

[69] Olivier Ferret. Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels. Dans *Actes de la 22e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 146–157, Caen, 2015. ATALA.

[70] Olivier Ferret. Réordonnancer des thésaurus distributionnels en combinant différents critères. *TAL*, 56(2):21–49, 2015.

[71] Charles J. Fillmore. Frame semantics. Dans The Linguistic Society of Korea, éditeur, *Linguistics in the Morning Calm : Selected Papers from SICOL-1981*, pages 111–137. Hanshin Publishing Co., Seoul, 1982.

[72] Charles J. Fillmore, Christopher R. Johnson et Miriam R.L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003.

[73] Charles J. Fillmore, Miriam R.L. Petruck, Josef Ruppenhofer et Abby Wright. FrameNet in action : The case of attaching. *International Journal of Lexicography*, 16(3):297–332, 2003.

[74] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman et Eytan Ruppin. Placing search in context : The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

[75] John Rupert Firth. A synopsis of linguistic theory 1930–1955. Dans The Philological Society, éditeur, *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, 1957.

[76] Pablo Gamallo, Gabriel P. Lopes et Alexandre Agustini. Inducing classes of terms from text. Dans Václav Matoušek et Pavel Mautner, éditeurs, *Text, Speech and*

Dialog : 10th International Conference TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings, pages 31–38. Springer, 2007.

[77] Peter Gärdenfors. *The Geometry of Meaning : Semantics Based on Conceptual Spaces*. MIT Press, Cambridge, 2014.

[78] François Gaudin. *Socioterminologie : Des problèmes sémantiques aux pratiques institutionnelles*. Publications de l'Université de Rouen, Rouen, 1993.

[79] Daniel Gildea et Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[80] Natalia Grabar et Pierre Zweigenbaum. Lexically-based terminology structuring : Some inherent limits. Dans *Proceedings of Computerm 2002, the 2nd International Workshop on Computational Terminology*, pages 1–7. ACL, 2002.

[81] Rebecca Green, Bonnie J. Dorr et Philip Resnik. Inducing frame semantic verb classes from WordNet and LDOCE. Dans *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL '04)*, pages 375–382, Barcelona, 2004.

[82] Gregory Grefenstette. SEXTANT : Exploring unexplored contexts for semantic extraction from syntactic analysis. Dans *Proceedings of the 30th Meeting of the Association for Computational Linguistics (ACL '92)*, pages 324–326. ACL, 1992.

[83] Gregory Grefenstette. Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. Dans *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 143–153, Columbus, 1993.

[84] Gregory Grefenstette. Corpus-derived first, second and third-order word affinities. Dans *Proceedings of Euralex*, pages 279–290, 1994.

[85] Amaru Cuba Gyllensten et Magnus Sahlgren. Navigating the semantic horizon using relative neighborhood graphs. Dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2451–2460. ACL, 2015.

[86] Benoît Habert et Pierre Zweigenbaum. Contextual acquisition of information categories : What has been done and what can be done automatically ? Dans Bruce E. Nevin et Stephen B. Johnson, éditeurs, *The Legacy of Zellig Harris : Language and information into the 21st century*, volume 2, pages 203–231. John Benjamins, 2002.

[87] Aric A. Hagberg, Daniel A. Schult et Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. Dans *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, 2008.

[88] Thierry Hamon et Adeline Nazarenko. Detection of synonymy links between terms : experiment and results. Dans *Recent Advances in Computational Terminology*, volume 2, pages 185–208. John Benjamins, 2001.

[89] Thierry Hamon, Adeline Nazarenko et Cécile Gros. A step towards the detection of semantic variants of terms in technical documents. Dans *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, volume 1, pages 498–504, Montréal, 1998. ACL.

[90] Kenneth E. Harper. Measurement of similarity between nouns. Dans *Proceedings of the 1965 Conference on Computational Linguistics (COLING)*, pages 1–23, Bonn, 1965. ACL. URL `http://dx.doi.org/10.3115/990314.990321`.

[91] Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

[92] Amir Hazem et Béatrice Daille. Méthode semi-compositionnelle pour l'extraction de synonymes des termes complexes. *TAL*, 56(2):51–76, 2015.

[93] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. Dans *Proceedings of the 14th Conference on Computational Linguistics (COLING '92)*, volume 2, pages 539–545, Nantes, 1992. ACL.

[94] Kris Heylen et Ann Bertels. Sémantique distributionnelle en linguistique de corpus. *Langages*, 201:51–64, 2016.

[95] Kris Heylen, Yves Peirsman et Dirk Geeraerts. Automatic synonymy extraction : A comparison of syntactic context models. Dans *Proceedings of the 18th Meeting of Computational Linguistics in the Netherlands*, pages 101–116, 2008.

[96] Felix Hill, Roi Reichart et Anna Korhonen. SimLex-999 : Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456, 2014. URL `http://arxiv.org/abs/1408.3456`.

[97] David Hull. Using statistical testing in the evaluation of retrieval experiments. Dans *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, Pittsburgh, 1993. ACM. URL `http://doi.acm.org/10.1145/160688.160758`.

[98] Fidelia Ibekwe-SanJuan. Inclusion lexicale et proximité sémantique entre termes. Dans *Proceedings of Terminology and Artificial Intelligence (TIA)*, Rouen, 2005.

[99] Christian Jacquemin. Guessing morphology from terms and corpora. Dans *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 156–167, Philadelphia, 1997.

[100] Michael N. Jones et Douglas J.K. Mewhort. Representing word meaning and

order information in a composite holographic lexicon. *Psychological Review*, 114 (1):1–37, 2007.

[101] Colette Joubarne et Diana Inkpen. Comparison of semantic similarity for different languages using the Google N-gram Corpus and second-order co-occurrence measures. Dans *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, pages 216–221, 2011. URL `http://dl.acm.org/citation.cfm?id=2018192.2018218`.

[102] Douwe Kiela et Stephen Clark. A systematic study of semantic vector space model parameters. Dans *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, pages 21–30. ACL, 2014.

[103] Mathieu Lafourcade et Alain Joubert. JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. Dans *Actes des Journées internationales d'analyse statistique des données textuelles*, pages 657–666, 2008.

[104] Thomas K. Landauer et Susan T. Dumais. A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[105] Gabriella Lapesa et Stefan Evert. Evaluating neighbor rank and distance measures as predictors of semantic priming. Dans *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 66–74, Sofia, 2013. ACL. URL `http://aclweb.org/anthology/W13-2608`.

[106] Gabriella Lapesa, Stefan Evert et Sabine Schulte im Walde. Contrasting syntagmatic and paradigmatic relations : Insights from distributional semantic models. Dans *Proceedings of the Third Joint Conference on Lexical and Computa-*

*tional Semantics (*SEM 2014)*, pages 160–170, Dublin, 2014. ACL/DCU. URL http://www.aclweb.org/anthology/S14-1020.

[107] Alberto Lavelli, Fabrizio Sebastiani et Roberto Zanoli. Distributional term representations : An experimental comparison. Dans *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 615–624. ACM, 2004.

[108] Omer Levy et Yoav Goldberg. Dependency-based word embeddings. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 302–308, 2014.

[109] Omer Levy et Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. Dans *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 171–180, Ann Arbor, 2014. ACL. URL http://www.aclweb.org/anthology/W14-1618.

[110] Omer Levy, Yoav Goldberg et Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[111] Marie-Claude L'Homme. *La terminologie : Principes et techniques*. Presses de l'Université de Montréal, Montréal, 2004.

[112] Marie-Claude L'Homme. Sur la notion de « terme ». *Meta*, 50(4):1112–1132, 2005.

[113] Marie-Claude L'Homme. Le verbe terminologique : un portrait de travaux récents. Dans *Actes du 3e Congrès mondial de linguistique française*, pages 93–107, Lyon, 2012.

[114] Marie-Claude L'Homme. Découverte de cadres sémantiques dans le domaine de l'environnement : le cas de l'influence objective. *Terminàlia*, 12:29–40, 2015.

[115] Marie-Claude L'Homme et Marie-Eve Laneville. DiCoEnviro : Le dictionnaire fondamental de l'environnement. http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf, 2009. Consulté : 2015-09-04.

[116] Marie-Claude L'Homme et Benoît Robichaud. Frames and terminology : Representing predicative terms in the field of the environment. Dans *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 186–197, Dublin, 2014. ACL/DCU. URL http://www.aclweb.org/anthology/W14-4723.

[117] Marie-Claude L'Homme, Benoît Robichaud et Carlos Subirats Rüggeberg. Discovering frames in specialized domains. Dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1364–1371, Reykjavík, 2014. ELRA.

[118] Ping Li, Curt Burgess et Kevin Lund. The acquisition of word meaning through global lexical co-occurrences. Dans *Proceedings of the 30th annual Child Language Research Forum*, pages 166–178, 2000.

[119] Dekang Lin. Automatic retrieval and clustering of similar words. Dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*, volume 2, pages 768–774. ACL, 1998.

[120] Dekang Lin et Patrick Pantel. Induction of semantic classes from natural language text. Dans *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 317–322. ACM, 2001.

[121] Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka et Shanfeng Zhu. MeSHLabeler : improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–

i347, 2015. URL http://bioinformatics.oxfordjournals.org/content/31/12/i339.abstract.

[122] William Léchelle et Philippe Langlais. Utilisation de représentations de mots pour l'étiquetage de rôles sémantiques suivant FrameNet. Dans *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 36–45, 2014.

[123] Markus Maier, Matthias Hein et Ulrike Von Luxburg. Cluster identification in nearest-neighbor graphs. Dans *Algorithmic Learning Theory*, pages 196–210. Springer, 2007.

[124] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze. Introduction to information retrieval (HTML edition). http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html, 2008. Consulté : 2015-08-12.

[125] Mounira Manser. État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. Dans *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 3, pages 163–175, Grenoble, 2012.

[126] Elizabeth Marshman et Marie-Claude L'Homme. Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés. Dans François Maniez, Pascaline Dury, Nathalie Arlin et Claire Rougemont, éditeurs, *Corpus et dictionnaires de langues de spécialité*, pages 87–110. Presses Universitaires de Grenoble, 2008.

[127] Alastair McKinnon. From co-occurrences to concepts. *Computers and the Humanities*, 11(3):147–155, 1977.

[128] Igor' Aleksandrovič Mel'čuk, Nadia Arbatchewsky-Jumarie, Léo Elnitsky, Lidija Iordanskaja, Adèle Lessard et André Clas. *Dictionnaire explicatif et combinatoire*

du français contemporain : Recherches lexico-sémantiques I. Presses de l'Université de Montréal, Montréal, 1984.

[129] Igor' Aleksandrovič Mel'čuk, André Clas et Alain Polguère. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve, 1995.

[130] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. Efficient estimation of word representations in vector space. Dans *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*, 2013.

[131] Tomas Mikolov, Quoc V. Le et Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013. URL `http://arxiv.org/abs/1309.4168`.

[132] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado et Jeffrey Dean. Distributed representations of words and phrases and their compositionality. Dans C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K.Q. Weinberger, éditeurs, *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3111–3119. Curran Associates, Inc., 2013.

[133] Tomas Mikolov, Wen-tau Yih et Geoffrey Zweig. Linguistic regularities in continuous space word representations. Dans *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751, Atlanta, 2013. ACL. URL `http://www.aclweb.org/anthology/N13-1090`.

[134] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký et Sanjeev Khudanpur. Recurrent neural network based language model. Dans *Proceedings of INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048, Makuhari, 2010.

[135] Jeff Mitchell et Mirella Lapata. Vector-based models of semantic composition. Dans *Proceedings of ACL-08 : HLT*, pages 236–244, 2008.

[136] Andriy Mnih et Geoffrey Hinton. Three new graphical models for statistical language modelling. Dans *Proceedings of the 24th International Conference on Machine Learning*, pages 641–648. ACM, 2007.

[137] Mikaël Morardo et Éric Villemonte de La Clergerie. Vers un environnement de production et de validation de ressources lexicales sémantiques. Dans *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, pages 167–180, Les Sables d'Olonne, 2013.

[138] François Morlane-Hondère et Cécile Fabre. Le test de substituabilité à l'épreuve des corpus : Utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. Dans *Actes du Congrès mondial de linguistique française (CMLF)*, pages 1001–1015, 2012.

[139] François Morlane-Hondère et Cécile Fabre. L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? Dans *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN)*, Montréal, 2010.

[140] Wolf Moskowich et Ruth Caplan. Distributive-statistical text analysis : A new tool for semantic and stylistic research. Dans G. Altmann, éditeur, *Glottometrika*, pages 107–153. Studienverlag Dr. N. Brockmeyer, Bochum, 1978.

[141] Claire Mouton. Induction de sens de mots à partir de multiples espaces sémantiques. Dans *Actes des 11es Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*. ATALA, 2009.

[142] Adeline Nazarenko, Pierre Zweigenbaum, Jacques Bouaud et Benoît Habert.

Corpus-based identification and refinement of semantic classes. Dans *Proceedings of the AMIA Annual Fall Symposium*, pages 585–589, 1997.

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

[144] Patrick Pantel et Dekang Lin. Discovering word senses from text. Dans *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619. ACM, 2002.

[145] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[146] Yves Peirsman et Dirk Geeraerts. Predicting strong associations on the basis of corpus data. Dans *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 648–656. ACL, 2009.

[147] Jeffrey Pennington, Richard Socher et Christopher D. Manning. GloVe : Global vectors for word representation. Dans *Proceedings of Empiricial Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[148] Amandine Périnet et Thierry Hamon. Hybrid acquisition of semantic relations based on context normalization in distributional analysis. Dans *Proceedings of Terminology and Artificial Intelligence (TIA)*, pages 113–120, 2013.

[149] Amandine Périnet et Thierry Hamon. Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité. Dans *Actes des 12es Journées d'analyse statistique des données textuelles (JADT)*, pages 507–518, 2014.

[150] Amandine Périnet et Thierry Hamon. Reducing VSM data sparseness by generalizing contexts : application to health text mining. Dans *Proceedings of the 5th*

International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 90–95, 2014.

[151] Amandine Périnet et Thierry Hamon. Réduction de la dispersion des données par généralisation des contextes distributionnels : Application aux textes de spécialité. Dans *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 232–243, Marseille, 2014.

[152] Amandine Périnet et Thierry Hamon. Analyse distributionnelle appliquée aux textes de spécialité : Réduction de la dispersion des données par abstraction des contextes. *TAL*, 56(2):77–102, 2015.

[153] Martin Phillips. *Aspects of Text Structure : An Investigation of the Lexical Organization of Text*. Elsevier, Amsterdam, 1985.

[154] Janine Pimentel. Methodological bases for assigning terminological equivalents : A contribution. *Terminology*, 19(2):237–257, 2013.

[155] Alain Polguère. *Lexicologie et sémantique lexicale : notions fondamentales*. Presses de l'Université de Montréal, 2003.

[156] Prokopis Prokopidis, Vassilis Papavassiliou, Antonio Toral, Marc Poch Riera, Francesca Frontini, Francesco Rubino et Gregor Thurmair. Final report on the corpus acquisition & annotation subsystem and its components. Rapport technique WP-4.5, PANACEA Project, 2012.

[157] Behrang QasemiZadeh. *Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora*. Thèse de doctorat, National University of Ireland, Galway, 2015.

[158] Reinhard Rapp. Identifying word translations in non-parallel texts. Dans *Proceedings of the 33rd annual meeting of the Association for Computational Linguistics*, pages 320–322. ACL, 1995.

[159] Reinhard Rapp. The computation of word associations : Comparing syntagmatic and paradigmatic approaches. Dans *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, volume 1, pages 1–7, Taipei, 2002. ACL. URL http://dx.doi.org/10.3115/1072228.1072235.

[160] Joseph Reisinger et Raymond Mooney. Cross-cutting models of lexical semantics. Dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1405–1415. ACL, 2011.

[161] Joseph Reisinger et Raymond J. Mooney. Multi-prototype vector-space models of word meaning. Dans *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. ACL, 2010.

[162] Herbert Rubenstein et John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. URL http://doi.acm.org/10.1145/365628.365657.

[163] Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson et Jan Scheffczyk. FrameNet II : Extended theory and practice. http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf, 2010. Consulté : 2016-07-26.

[164] Magnus Sahlgren. Towards pertinent evaluation methodologies for word-space models. Dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 2006.

[165] Magnus Sahlgren. *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, 2006.

[166] Magnus Sahlgren, Anders Holst et Pentti Kanerva. Permutations as a means to encode order in word space. Dans *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1300–1305, 2008.

[167] Gerard Salton et Michael E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

[168] Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Lausanne/Paris, 1916. Éd. Charles Bailly et Albert Sechehaye avec la collaboration d'Albert Riedlinger.

[169] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. Dans *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[170] Thomas Schmidt. The Kicktionary : Combining corpus linguistics and lexical semantics for a multilingual football dictionary. Dans Wolfgang Stadler, Andrew Skinner, Gerhard Pisek et Eva Lavric, éditeurs, *The Linguistics of Football*, pages 11–21. Narr, Tübingen, 2008.

[171] Hinrich Schütze. Dimensions of meaning. Dans *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing (Supercomputing'92)*, pages 787–796, 1992.

[172] Hinrich Schütze. Part-of-speech induction from scratch. Dans *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 251–258. ACL, 1993.

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

[174] Roy Schwartz, Roi Reichart et Ari Rappoport. Symmetric pattern based word embeddings for improved word similarity prediction. Dans *Proceedings of the*

19th Conference on Computational Language Learning (CoNLL), pages 258–267, 2015.

[175] Holger Schwenk. Continuous space language models. *Computer Speech & Language*, 21(3):492–518, 2007.

[176] Karen Spärck Jones. Some points in a time. *Computational Linguistics*, 31(1): 1–14, 2005.

[177] Mark Steyvers et Joshua B. Tenenbaum. The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.

[178] Ludovic Tanguy, Franck Sajous et Nabil Hathout. Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *TAL*, 56(2):103–127, 2015.

[179] Rita Temmerman. *Towards New Ways of Terminology Description : The Sociocognitive Approach*. John Benjamins, Amsterdam/Philadelphia, 2000.

[180] Aristomenis Thanopoulos, Nikos Fakotakis et George Kokkinakis. Automatic extraction of semantic relations from specialized corpora. Dans *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, volume 2, pages 836–842. ACL, 2000.

[181] Aristomenis Thanopoulos, Nikos Fakotakis et George Kokkinakis. Automatic extraction of semantic similarity of words from raw technical texts. Dans *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, Athènes, 2000.

[182] Peter D. Turney et Patrick Pantel. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[183] Tim Van de Cruys. A comparison of bag of words and syntax-based approaches for word categorization. Dans *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 47–54, 2008.

[184] Tim Van de Cruys. A non-negative tensor factorization model for selectional preference induction. Dans *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 83–90. ACL, 2009.

[185] Giulia Venturi, Alessandro Lenci, Simonetta Montemagni, Eva Maria Vecchi, Maria Teresa Sagri et Daniela Tiscornia. Towards a FrameNet resource for the legal domain. Dans *Proceedings of the Third Workshop on Legal Ontologies and Artificial Intelligence Techniques*, Barcelona, 2009.

[186] Jean Véronis. Cartographie lexicale pour la recherche d'information. Dans *Actes de la 10e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 265–274, Batz-sur-mer, 2003.

[187] Ornella Wandji, Marie-Claude L'Homme et Natalia Grabar. Discovering semantic frames for a contrastive study of verbs in medical corpora. Dans *Proceedings of Terminology and Artificial Intelligence (TIA)*, Paris, 2013.

[188] Warren Weaver. Translation. Dans William Nash Locke et Andrew Donald Booth, éditeurs, *Machine Translation of Languages : Fourteen Essays*, pages 15–23. MIT Press, Cambridge, 1955.

[189] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir et Bill Keller. Learning to distinguish hypernyms and co-hyponyms. Dans *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, pages 2249–2259, Dublin, 2014. ACL/DCU. URL `http://www.aclweb.org/anthology/C14-1212`.

[190] Julie Weeds, James Dowdall, Gerold Schneider, Bill Keller et David Weir. Using distributional similarity to organise biomedical terminology. *Terminology*, 11(1): 107–141, 2005.

[191] Julie Weeds, David Weir et Diana McCarthy. Characterising measures of lexical distributional similarity. Dans *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. ACL, 2004.

[192] Julie Weeds, David Weir et Jeremy Reffin. Distributional composition using higher-order dependency vectors. Dans *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL*, pages 11–20, 2014.

[193] Eugen Wüster. L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie et les sciences des choses. Dans V. I. Siforov, éditeur, *Fondements théoriques de la terminologie*, pages 55–114. Groupe interdisciplinaire de recherche scientifique et appliquée en terminologie, Québec, 1981.

[194] Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond et Asuka Sumida. Hypernym discovery based on distributional similarity and hierarchical structures. Dans *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 2, pages 929–937. ACL, 2009.

[195] George Kingsley Zipf. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256, 1945.

[196] Pierre Zweigenbaum et Natalia Grabar. Liens morphologiques et structuration de terminologie. Dans Pierre Tchounikine, éditeur, *Actes de IC'00*, Toulouse, 2000.

[197] Pierre Zweigenbaum, Fadila Hadouche et Natalia Grabar. Apprentissage de relations morphologiques en corpus. Dans *Actes de la 10e conférence sur le traitement automatique des langues naturelles (TALN)*, 2003.

Annexe I

Méthode de normalisation de caractères

Le programme[1] que nous avons développé pour le prétraitement du corpus PANACEA fait appel à la bibliothèque Unidecode[2] pour réaliser la normalisation des caractères en anglais. Cette bibliothèque offre une fonction qui remplace les codes de caractères qui sont absents du codage ASCII par un code ASCII équivalent lorsque possible. Par exemple, les codes de caractères correspondant à la lettre accentuée *é* sont remplacés par le code ASCII du *e* non accentué.

Pour le français, on doit adopter une stratégie différente, puisque les lettres accentuées doivent être conservées, du moins certaines lettres accentuées. Le programme de prétraitement traite ce problème de la façon suivante :

1. Les séquences lettre+accent sont remplacées par la lettre accentuée équivalente ; le programme exploite à cette fin la bibliothèque Unicodedata[3].

2. Le programme remplace les lettres *œ* et *Œ* par les chaînes *oe* et *OE* respectivement.

3. Les codes correspondant à des guillemets qui sont absents du codage Latin-1 (codes hexadécimaux 2018, 2019, 201C et 201D) sont remplacés par des codes Latin-1 équivalents (codes décimaux 34 et 39).

4. Le programme encode le corpus en Latin-1, en supprimant tout caractère absent de ce codage. Le corpus résultant est ensuite ré-encodé en UTF-8, mais ne contient que des codes présents dans le codage Latin-1.

---

[1] Voir `https://github.com/gbcolborne/exp_phd/blob/master/preprocess_PANACEA.py`.

[2] `https://pypi.python.org/pypi/Unidecode` (page consultée le 6 octobre 2014).

[3] `https://docs.python.org/2/library/unicodedata.html` (page consultée le 30 juillet 2015).

Annexe II

Pondérations

Afin de pondérer les fréquences de cooccurrence, nous utilisons les 6 mesures d'association simples décrites par Evert [59, ch. 4] :

$$\text{MI} = \log_2 \left(\frac{O}{E} \right) \tag{II.1}$$

$$\text{MI}^k = \log_2 \left(\frac{O^k}{E} \right) \tag{II.2}$$

$$\text{local-MI} = O \times \log_2 \left(\frac{O}{E} \right) \tag{II.3}$$

$$\text{simple-LL} = 2 \times \left(O \times \ln \left(\frac{O}{E} \right) - (O - E) \right) \tag{II.4}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \tag{II.5}$$

$$\text{t-score} = \frac{O - E}{\sqrt{O}} \tag{II.6}$$

Dans toutes ces équations, O désigne la fréquence de cooccurrence observée, que l'on pourrait aussi noter $f(w_i, w_j)$, c'est-à-dire la fréquence à laquelle le mot-contexte w_j apparait dans la fenêtre de contexte centrée sur toutes les occurrences du mot-cible w_i dans le corpus. Quant à E, il s'agit de l'espérance de la fréquence de cooccurrence $\mathbb{E}[f(w_i, w_j)]$. Evert [59] l'estime de la façon suivante :

$$E = \mathbb{E}[f(w_i, w_j)] = \frac{k \times f(w_i) \times f(w_j)}{N} \tag{II.7}$$

où k est le nombre de positions dans la fenêtre de contexte, $f(w)$ est la fréquence d'un mot w et N est le nombre de mots dans le corpus.

Nous définissons E d'une façon légèrement différente :

$$E = \mathbb{E}[f(w_i, w_j)] = \frac{\sum_{k=1}^{n} f(w_i, w_k) \times \sum_{k=1}^{n} f(w_k, w_j)}{\sum_{k=1}^{n} \sum_{l=1}^{n} f(w_k, w_l)} \tag{II.8}$$

où n est le nombre de mots-cibles (et de mots-contextes, puisque nous avons choisi d'utiliser le même ensemble pour les mots-cibles et les mots-contextes). Cette formule est équivalente à celle d'Evert [59] si l'ensemble des mots-cibles et l'ensemble des mots-contextes contiennent tous les mots dans le corpus. Nous préférons cette définition principalement parce qu'elle est basée seulement sur le contenu de la matrice de cooccurrence, ce qui simplifie un peu les choses du point de vue de l'implémentation de l'AD et de ses différents paramètres, à notre avis.

Dans le cas de MI^k, nous testons deux valeurs de k, à savoir 2 et 3.

Pour toutes les pondérations, nous ne conservons que les valeurs non négatives suivant Lapesa et al. [106] ; cette contrainte de non-négativité est souvent utilisée en sémantique distributionnelle. Dans le cas de la mesure simple-LL, qui ne distingue pas entre les cas où $O \ll E$ et ceux où $O \gg E$, le résultat étant positif dans les deux cas [59, p. 21], nous fixons la fréquence pondérée à 0 si $O < E$.

Pour certaines pondérations (simple-LL, local-MI, z-score et t-score), nous appliquons une transformation aux fréquences de cooccurrence pondérées, parce que cela produit parfois de meilleurs résultats [105, 106]. Deux transformations possibles sont le logarithme et la racine carrée :

$$\log(x) = \ln(x+1) \tag{II.9}$$

$$\text{sqrt}(x) = \sqrt{x} \tag{II.10}$$

Nous appliquons la transformation log à la pondération simple-LL et sqrt à la pondération z-score, suivant Lapesa et al. [106]. Nous appliquons également une transformation aux pondérations local-MI et t-score (log et sqrt respectivement).

Nous utilisons également une de ces transformations (log) comme pondération, c'est-à-dire en appliquant simplement cette transformation aux fréquences de coocurrence, sans les pondérer au moyen d'une mesure d'association au préalable.

Annexe III

Méthode d'extraction des couples de référence

Le fonctionnement du programme que nous avons développé pour extraire les couples de référence se résume essentiellement aux étapes suivantes :

1. Extraire de l'index du dictionnaire l'URL de chaque article.
2. Extraire les relations ciblées de tous les articles ; stocker pour chaque relation les deux termes (l'entrée et le terme relié, en minuscules) ainsi que leur partie du discours (PDD) et la nature de la relation.
3. Exclure les couples de termes ayant la même forme (p. ex. $animal_{NN} \rightarrow animal_{JJ}$).
4. Filtrer les couples en fonction de la PDD. Exclure les couples suivants : les couples contenant un adverbe ; les dérivés ayant la même PDD, ceux-ci représentant des erreurs (p. ex. identification erronée de la PDD du terme relié par le programme) ; et les couples ayant des PDD différentes, à l'exception des dérivés.
5. Éliminer les doublons, c'est-à-dire les couples qui ont été extraits plus d'une fois. Ces doublons peuvent être présents parce qu'une même paire de formes correspond à plus d'une relation, ou parce qu'elles appartiennent à plus d'une PDD.
6. Éliminer les couples contenant un terme qui ne fait pas partie des mots-cibles.
7. Écrire les couples.

Lors de l'extraction des relations (étape 2), le programme commence par lire l'article (en format XML). Pour chaque lexie (chaque acception de l'entrée), il vérifie le statut de rédaction de la description de cette lexie. Si ce statut est égal ou inférieur à 2, ce qui indique que la rédaction est suffisamment avancée pour que la description puisse être

consultée par le public, il extrait la PDD de la lexie ainsi que tous les termes reliés à la lexie par une des relations énumérées ci-dessus. Le programme exclut les termes reliés qui sont une collocation ou un modificateur de la lexie (p. ex. *moteur→~ électrique*). Ensuite, pour chaque terme relié, le programme tente d'obtenir sa PDD en consultant son article. S'il ne la trouve pas, il procède de la façon suivante :

- dans le cas des noms dérivés, verbes dérivés et adjectifs dérivés, il présume qu'il s'agit d'un nom, d'un verbe ou d'un adjectif respectivement, ce qui devrait être le cas ;
- pour toutes les autres relations, il présume que la PDD du terme relié est la même que celle de la lexie, ce qui devrait être vrai pour ces relations.

Annexe IV

Ensembles de référence pour le français

1. accumulation, accumuler, concentration
2. adaptation, adapter
3. boisement, boiser, reboisement, reboiser
4. agriculture, cultiver, culture
5. froid, température
6. stabilisation, stabiliser
7. charger, recharger
8. sensible, vulnérable
9. transport, transporter
10. capture, capturer, piégeage, piéger
11. stabilisation, stabiliser, équilibrer
12. développer, industrialisation, urbanisation
13. compostage, composter, transformer
14. appauvrir, appauvrissement
15. accélérer, aggraver, amplification, amplifier, atténuation, atténuer, intensifier, limiter, ralentir, renforcer
16. abaisser, augmentation, augmenter, diminuer, réduction, réduire

17. altération, altérer, bouleversement, changer, modifier, perturbation, perturber
18. propulser, propulsion
19. refroidir, refroidissement, réchauffement, réchauffer
20. disparaître, disparition, extinction, perte, pénurie, raréfaction, raréfier, épuisement, éteindre
21. recul, reculer, éroder, érosion
22. accélération, accélérer, aggraver, intensification, intensifier, ralentir, ralentissement
23. dégel, fondre, fonte, gel, geler
24. refroidir, refroidissement, réchauffement, réchauffer
25. augmentation, augmenter, baisse, baisser, croissance, diminuer, diminution, élever, élévation
26. climat, microclimat
27. collecte, ramassage, ramasser
28. acidification, contamination, contaminer, eutrophisation, polluer
29. conversion, convertir, transformation, transformer
30. dégradation, dégrader, détérioration, détériorer, endommager
31. dégrader, détériorer
32. injecter, libération, libérer, émettre, émission
33. décharge, décharger
34. estimation, estimer

35. croissance, développement
36. charge, charger, plein, recharge, recharger
37. inondation, inonder
38. enfouir, enfouissement
39. environnemental, propre, vert, écologique
40. extrême, intense, intensité, violent
41. extraction, extraire
42. affecter, effet, impact, incidence, influence, influencer, influer, répercussion
43. conduire, conduite
44. plantation, planter
45. prédire, prévision, prévoir
46. recyclage, recycler
47. conservation, conserver, préservation, préserver
48. conditionnement, traitement, traiter
49. protection, protéger
50. récupération, récupérer, valorisation, valoriser
51. élimination, éliminer
52. déboisement, déboiser, défrichement, défricher
53. lutte, lutter
54. modèle, scénario

55. réemploi, réutilisation, réutiliser
56. déplacement, déplacer, trajet, voyage
57. menace, nocif
58. décomposer, dégrader
59. tri, trier
60. modélisation, modéliser, reproduire, simulation, simuler
61. absorber, absorption
62. stockage, stocker
63. cyclone, orage, ouragan, tempête, tornade, tsunami
64. durable, rationnel
65. emprisonner, piéger
66. changement, changer, fluctuation, modification, modifier, variation, varier, évoluer
67. consommation, consommer, exploitation, exploiter, surexploitation, surexploiter
68. évaporation, évaporer, évapotranspiration
69. phénomène, événement

Annexe V

Ensembles de référence pour l'anglais

1. accumulate, accumulation, concentration
2. adapt, adaptation
3. afforestation, reforest, reforestation
4. agriculture, cultivate
5. cold, cool, temperature, warm
6. assess, assessment
7. stabilization, stabilize
8. sensitivity, threatened, vulnerability, vulnerable
9. contaminated, polluted
10. carry, transportation
11. calculate, calculation
12. capture, sequester, sequestration
13. development, industrialization, urbanization
14. deplete, depletion
15. abate, abatement, accelerate, amplify, intensify, mitigate, mitigation
16. decrease, increase, reduce, reduction
17. alter, alteration, change, perturbation

18. cool, warm
19. disappear, disappearance, extinction, loss, shortage
20. erosion, retreat
21. accelerate, acceleration, intensification, intensify
22. freeze, frost, melt, melting, thaw
23. cool, cooling, warm, warming
24. decline, decrease, grow, growth, increase, rise
25. acidification, acidify, contaminate, pollute
26. conversion, convert, hybridization, transform
27. damage, degradation, degrade
28. emission, emit, inject, radiate, release
29. endanger, threaten
30. estimate, estimation
31. grow, growth
32. charge, recharge
33. clean, environmental, green
34. extreme, intense, intensity, severe, severity
35. manage, management
36. drive, travel
37. affect, effect, impact, influence

38. drive, ride
39. precipitation, rain, rainfall
40. forecast, predict, project, projection
41. recycle, recycling
42. conservation, conserve, preserve
43. process, treat
44. protect, protection
45. discard, disposal, dispose
46. clear, clearing, deforest, deforestation
47. model, scenario
48. commute, transit, travel
49. harmful, hazard, hazardous, threat
50. separate, separation, sort
51. model, modelling, simulate
52. absorb, absorption, uptake
53. cyclone, hurricane, storm, thunderstorm, tornado, tsunami
54. sustainability, sustainable, unsustainable
55. change, fluctuate, fluctuation, shift, variation, vary
56. consume, consumption
57. evaporate, evaporation, evapotranspiration

Annexe VI

Remarques sur la F-mesure

Il existe plusieurs façons de calculer la mesure F_β [121, sec. 4.3.2]. Une façon consiste à calculer la mesure F_β pour chaque requête, puis prendre la moyenne sur toutes les requêtes, que nous noterons $\bar{F}_\beta$:

$$\bar{F}_\beta(Q) = \frac{1}{n}\sum_{i=1}^{n}\left((1+\beta^2)\times\frac{p(q_i)\times r(q_i)}{(\beta^2\times p(q_i))+r(q_i)}\right)$$

où $Q = \{q_1,\ldots,q_n\}$ est l'ensemble des requêtes, $p(q_i)$ est la précision du voisinage de la requête q_i et $r(q_i)$, le rappel. On appelle cette mesure la *mesure F_β moyenne*.

Une autre façon consiste à calculer d'abord la précision moyenne et le rappel moyen sur toutes les requêtes, puis à calculer la mesure F_β à partir de la précision moyenne et du rappel moyen :

$$F_\beta(Q) = (1+\beta^2)\times\frac{P(Q)\times R(Q)}{(\beta^2\times P(Q))+R(Q)}$$

où $P(Q)$ et $R(Q)$ sont la précision moyenne et le rappel moyen sur toutes les requêtes. On appelle cette mesure la *macromesure F_β*.

Nous avons opté pour la macromesure parce qu'elle nous permet de calculer *a posteriori* la mesure F_β (notamment avec différentes valeurs de β) à partir de la précision moyenne et du rappel moyen, qui sont les seules mesures que nous stockons pour chaque graphe (et pour chaque jeu de données).

À titre indicatif, nous avons vérifié dans quelle mesure ces deux façons de calculer la mesure F_β produisent des résultats différents. La Figure VI.1 montre la dispersion de la mesure F_1 moyenne et de la macromesure F_1, calculées sur les ensembles de référence ; les résultats comprennent tous les modèles évalués (AD et `word2vec`) dans les deux langues. La figure montre que la macromesure a une dispersion plus élevée et un

maximum plus élevé.

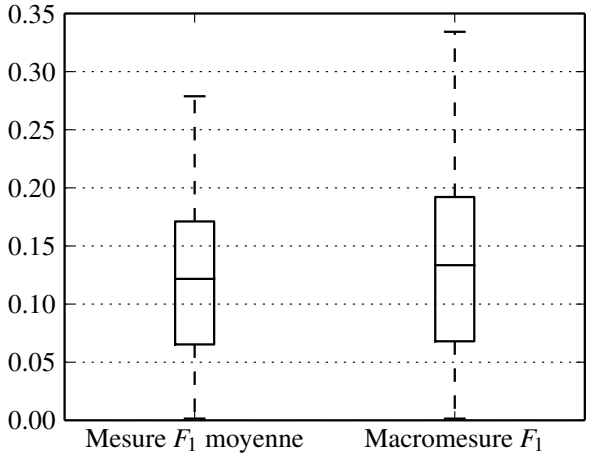


Figure VI.1 – Dispersion de la mesure F_1 moyenne et de la macromesure F_1, calculées sur les ensembles de référence.

La Figure VI.2 montre la répartition de la différence, pour chaque graphe évalué, entre la macromesure F_1 et la mesure F_1 moyenne $(F_1(Q) - \bar{F}_1(Q))$, calculées sur les ensembles de référence. Cette figure montre que la macromesure est presque toujours plus élevée que la mesure F_1 moyenne, sauf dans 22 des 34 560 cas, et que la différence se situe entre 0 et 0.01 dans la majorité (17 552) des cas. Ainsi, la macromesure F_1 tend à être légèrement plus élevée que la mesure F_1 moyenne.

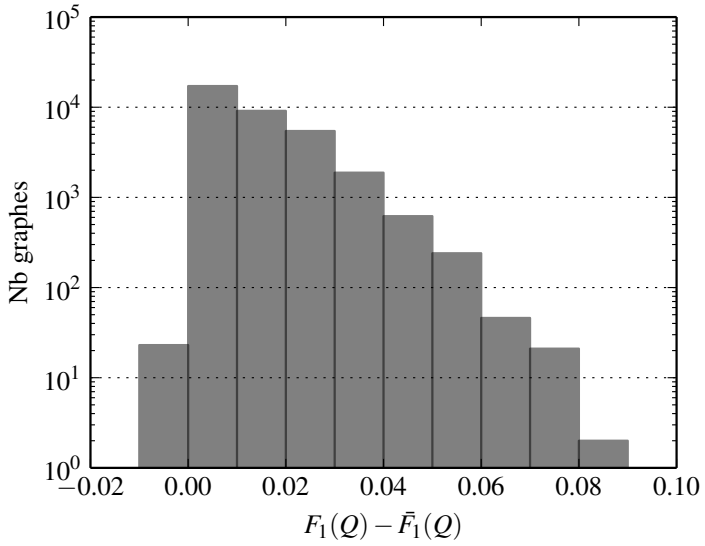


Figure VI.2 – Répartition de la différence entre la macromesure F_1 et la mesure F_1 moyenne (ordonnée en échelle logarithmique).