Exploration des relations terminologiques entre les termes multi-mots dans les modèles de sémantique distributionnelle

Yizhe Wang

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse - Jean Jaurès*

Présentée et soutenue le *04/10/2022* **par :**

Yizhe WANG

Exploration des relations terminologiques entre les termes multi-mots dans les modèles de sémantique distributionnelle

---

JURY

Béatrice DAILLE	Professeure, Université de Nantes	Directrice
Nabil HATHOUT	Directeur de recherche, CNRS	Directeur
Marie-Paule JACQUES	Maîtresse de conférences, Université Grenoble Alpes	Rapporteure
Olivier FERRET	Directeur de recherche, CEA	Rapporteur
Anne CONDAMINES	Directrice de recherche, CNRS	Examinatrice
Manel ZARROUK	Maîtresse de conférences, Université Paris 13	Examinatrice

---

École doctorale et spécialité :
CLESCO : Sciences du langage

Unité de Recherche :
Laboratoire Cognition, Langues, Langage, Ergonomie (UMR 5263)

Directeur(s) de Thèse :
Béatrice DAILLE et *Nabil HATHOUT*

Rapporteurs :
Marie-Paule JACQUES et *Olivier FERRET*

Résumé

Le terme est une unité lexicale qui a un sens spécialisé dans un domaine particulier. L'organisation des termes reflète la structure de la connaissance d'un domaine. Cette structure est basée sur les relations qui existent entre les concepts du domaine et par suite entre les termes simples (TS) ou multi-mots (TMM).

Les ressources terminologiques structurées, telles que les dictionnaires spécialisés, les banques et les bases de données terminologiques, sont conçues pour répondre aux besoins dans les domaines de recherche, de traduction, de rédaction technique, etc. Cependant, les relations entre TMM y sont souvent sous-représentées. Beaucoup de travaux portent en effet sur l'acquisition de relations entre TS et relativement peu sur l'acquisition de relations entre TMM. D'un autre côté, on observe depuis plusieurs années, l'utilisation massive et réussie des modèles sémantiques distributionnels (MSD) dans de nombreux travaux en sémantique. Notre problématique de recherche se place à l'articulation de ces deux thèmes. Elle se décline en deux questions : est-il possible de capturer des informations relationnelles entre TMM en utilisant des MSD ? Quels sont les relations entre TMM que ces modèles permettent d'identifier le mieux ? Pour répondre à ces deux questions, nous avons réalisé un ensemble d'expériences dans le domaine de l'environnement en français.

Notre hypothèse générale de travail est que les MSD statiques et contextuels permettent d'identifier des relations terminologiques entre TMM. Nous avons exploré les possibilités de ces modèles en utilisant deux méthodes. La première adoptant la substitution lexicale est fondée sur les prédictions d'un modèle de langue masqué (MLM). La seconde consiste à capter des relations sémantiques lexicales par analogie entre les représentations des termes générées par un modèle FastText. Ces méthodes sont testées sur deux jeux de données. Nous nous appuyons d'abord sur un jeu de données composé de TMM synonymes du domaine de l'environnement en français fournis par la banque IATE. Devant le manque de ressources intégrant des relations sémantiques variées entre TMM, nous avons construit un second jeu de données par projection sémantique à partir des termes simples et de leurs relations recensées dans le dictionnaire de termes de l'environnement, DiCoEnviro. Une annotation manuelle à l'aide des contextes des TMM est effectuée pour vérifier la préservation de la relation inférée entre TMM.

Les résultats expérimentaux que nous avons obtenus valident notre hypothèse concernant la possibilité de capter des relations terminologiques entre TMM par des MSD. L'analogie impliquant un modèle FastText s'avère plus performante que la substitution lexicale réalisée

avec un MLM pour capturer la synonymie, l'antonymie et l'hyponymie. Les résultats montrent aussi que la stratégie conditionnement permet d'avoir des prédictions du MLM reliées plus étroitement au mot masqué et que les performances de l'analogie sont améliorées lorsque les variantes de TMM sont traitées comme des occurrences des TMM. Un autre résultat notable est que la composition sémantique des TMM est modélisée par l'analogie et capturée dans une certaine mesure par les modèles de langage masqués. Ce résultat confirme l'avantage de combiner les approches distributionnelles et compositionnelles pour l'identification des relations sémantiques entre TMM. Les meilleurs résultats ont été obtenus avec l'analogie, avec un MRR de 0,793 pour la synonymie, de 0,720 pour l'antonymie, de 0,613 pour l'hyperonymie et 0,579 pour l'hyponymie.

Dans l'ensemble, cette thèse est l'une des premières tentatives pour identifier les relations lexicales entre TMM d'un domaine spécialisé, celui de l'environnement, en explorant les MSD. Nous avons construit et mis à disposition un jeu de données de TMM reliés par les relations lexicales variées. Ce travail fournit aussi un carnet de route pour l'application des MSD pour la tâche de structuration terminologique.

Abstract

A term is a lexical unit with specialized meaning in a particular domain. The organization of terms reflects the structure of domain knowledge, which is based on the relationships between domain concepts, i.e., between single terms (STs) or multi-word terms (MWTs).

Structured terminology resources, such as specialized dictionaries, terminology banks, and databases, are designed to meet the needs of research, translation, technical writing, etc. However, the relationships between MWTs are often underrepresented. On the other hand, we have seen the massive and successful use of distributional semantic models (DSMs) in many semantics works over the last few years. Our research problem is on the intersection of these two themes. It can be broken down into two questions : is it possible to capture relational information between MWTs using DSMs ? What are the relationships between MWTs that these models can best identify ? We conducted a set of experiments in the French environment domain to answer these two questions.

Our general working hypothesis is that static and contextual DSMs allow us to identify terminological relations between MWTs. We explored the possibilities of these models using two methods. The first one adopting lexical substitution is based on the predictions of a masked language model (MLM). The second one captures lexico-semantic relations by analogy between term representations generated by a FastText model. These methods are tested on two datasets. First, we rely on a dataset composed of French synonymous MWTs of the environment domain provided by the IATE database. Due to the lack of resources integrating various semantic relations between MWTs, we built a second dataset by semantic projection from the single terms and their relations listed in the dictionary of environmental terms, DiCoEnviro. A manual annotation using the contexts of the MWTs is performed to check the preservation of the inferred relationship between the MWTs.

The experimental results we obtained validate our hypothesis regarding the possibility of capturing terminological relations between MWTs by DSMs. Analogy involving a FastText model performs better than lexical substitution performed with an MLM in capturing synonymy, antonymy, and hyponymy. The results also show that the conditioning strategy results in MLM predictions that are more closely related to the masked word and that the performance of analogy is improved when variants of MWTs are treated as occurrences of MWTs. Another notable result is that the semantic composition of MWTs is modelled by analogy and captured partially by MLM. This result confirms the advantage of combining distributional and compositional

approaches for identifying semantic relations between MWTs. The best results were obtained with the analogy, with an MRR of 0.793 for synonymy, 0.720 for antonymy, 0.613 for hypernymy, and 0.579 for hyponymy.

Overall, this thesis is one of the first attempts to identify lexical relations between MWTs in a specialized domain, that of the environment, by exploring DSMs. We have constructed a dataset of MWTs linked by various lexical relationships, which also made available for further research purpose. Ideally, this work is hope to provide a roadmap for applying DSMs for the terminology structuring task.

Remerciements

Les années de thèse ont été marquées par de nouvelles rencontres mais aussi par le soutien de nombreuses personnes que je souhaite remercier ici. J'aimerais tout d'abord remercier sincèrement mes directeurs de thèse Béatrice Daille et Nabil Hathout, pour nos réunions régulières malgré leur agenda serré, leur pédagogie, leur bonne humeur et leur encadrement bienveillant. Difficile est la rédaction d'une thèse, et encore plus quand le français n'est pas ma langue maternelle. Leur sens du détail et leurs nombreuses relectures ont permis d'aboutir à ce manuscrit.

Je remercie chaleureusement les membres de mon jury, à savoir Marie-Paule Jacques et Olivier Ferret pour avoir accepté d'être les rapporteurs de ce travail, ainsi que Anne Condamines et Manel Zarrouk, qui ont bien voulu en être les examinatrices.

Cette thèse est réalisée dans le cadre du projet ADDICTE dont je tiens à remercier tous les membres : Emmanuel Morin, Cécile Fabre, Gaël de Chalendar, Hicham El Boukkouri, Ludovic Tanguy, Mérième bouhandi, Pierre Zweigenbaum, Thierry Hamon, Thomas Lavergne, ... Je les remercie pour leur retour sur mon travail et leur réponse rapide et détaillée à mes questions sur les études.

Je suis très chanceuse d'avoir passé mes quatre années de thèse au laboratoire CLLE-ERSS, dont je tiens à remercier les membres pour leur accueil et leur disponibilité. Je tiens plus particulièrement à remercier Josette Rebeyrolle qui a participé à mon comité de suivi de thèse et Lydia-Mai Ho-Dac qui m'a beaucoup aidé sur la préparation de la soutenance. Sans oublier le personnel administratif et l'équipe technique pour leur sympathie et leur disponibilité. La vie au laboratoire est aussi colorée grâce aux nombreux doctorants que j'ai connus avec un grand plaisir : Bénédicte, Camilla, Claire, Julie, Lena, Lison, Marine, Océane, Natalia, Nataly, ... J'ai toujours aimé notre discussion, notre entraide, et nos soirées après le travail. Je souhaite remercier particulièrement Lena et Daniele, mes collègues de bureau, qui m'ont fait goûter des gâteaux délicieux et découvrir les meilleures pizzas de Toulouse. Je voudrais également remercier Silvia et Filip qui m'ont beaucoup aidé au course de ces années.

Mes pensées vont aussi à mes chers professeurs de master à l'INALCO et à l'université Sorbonne Nouvelle qui m'ont guidé pour découvrir le TAL. Je remercie surtout Damien Nouvel qui était mon encadrant du mémoire de master.

Et pour finir, j'exprime toute la gratitude à mes parents et mes amis, ceux qui ont toujours été là, derrière moi, qui ont cru en moi. Merci pour leur amour inconditionnel et leur soutien sans faille.

Table des matières

Introduction

Problématique de recherche

En terminologie, on considère que l'organisation des termes reflète la structure des connaissances d'un domaine (Sager et Ndi-Kimbi, 1995 ; Zweigenbaum et Grabar, 2000 ; L'Homme, 2004). Cette structure est fondée sur les relations entre les concepts du domaine. L'organisation terminologique décrit les liens qui existent entre les termes simples (TS) et les termes multi-mots (TMM). Par exemple, il existe une relation d'antonymie entre *boisé* et *déboisé* et entre *humide* et *sec* ; une relation d'hyperonymie entre *combustible* et *charbon* et entre *énergie* et *électricité* ; de même, *air humide* est un antonyme de *air sec* et *énergie propre* est un hyperonyme de *électricité propre* ; etc.

Les relations entre termes peuvent être identifiées par des experts qui peuvent utiliser à cette fin des ressources existantes, notamment des corpus. De nombreuses recherches actuelles portent sur les relations entre TS (Grabar et Hamon, 2006 ; Zhang *et al.*, 2017 ; Zhu *et al.*, 2017). À l'inverse, peu de travaux s'intéressent à l'acquisition de relations entre TMM (Hazem et Daille, 2018). La plupart des travaux sur les relations entre TMM concernent l'exploitation de la structure interne des TMM en utilisant différents types d'informations linguistiques, notamment syntaxiques (Verspoor *et al.*, 2003) et sémantiques (Hazem et Daille, 2015).

Nous proposons dans cette thèse d'utiliser les représentations sémantiques distributionnelles pour explorer les relations sémantiques entre TMM. La sémantique distributionnelle (SD) (Lenci, 2008) est une méthode de représentation du sens fondée sur l'hypothèse distributionnelle (Harris, 1954) selon laquelle les mots qui ont des contextes linguistiques similaires tendent à avoir des sens similaires. La SD représente les unités lexicales par des vecteurs produits par des modèles sémantiques distributionnels (MSD). Ces vecteurs encodent la distribution des mots d'un corpus. Les vecteurs proches dans l'espace vectoriel distributionnel représentent des mots dont les distributions sont similaires. L'utilisation de l'analyse distributionnelle dans les domaines de spécialité est ancienne et remonte au moins aux années 1990 (Morlane-Hondère et Fabre, 2012). Cette utilisation est cependant limitée par la taille trop faible des corpus spécialisés. La tendance actuelle en TAL impliquant les MSD est de traiter de grands corpus. La SD a beaucoup été utilisée dans le domaine médical (Paullada *et al.*, 2020 ; Chen *et al.*, 2018 ; Bourigault, 2002). En revanche, il existe peu de travaux similaires sur le domaine environnemental (Hazem et Daille, 2018 ; Bernier-Colborne et Drouin, 2016).

Notre étude porte sur l'exploration des capacités des modèles sémantiques distributionnels à capter différentes relations sémantiques entre TMM en français dans le domaine de l'environnement. Nous nous intéressons aux TMM nominaux composés de deux mots lexicaux (bitermes). Notre étude est fondée sur deux hypothèses : (*i*) les relations lexicales entre TMM peuvent être identifiées par les MSD ; (*ii*) les MSD modélisent partiellement la nature compositionnelle des TMM. Pour ce faire, nous avons utilisé deux méthodes dans l'état de l'art. La première méthode est fondée sur la substitution lexicale qui vise à identifier les substituts possibles d'un mot cible dans une phrase. Ces substituts doivent être sémantiquement reliés au mot cible et doivent pouvoir s'intégrer dans la phrase contexte (McCarthy et Navigli, 2007). La substitution lexicale intervient dans diverses applications, telles que l'induction du sens des mots (Amrami et Goldberg, 2019), l'extraction de relations lexicales (Schick et Schütze, 2020), la génération de paraphrases (Guu *et al.*, 2018), la simplification des textes (Specia *et al.*, 2012), etc. La seconde méthode permet de capter les relations sémantiques entre TMM au moyen de l'analogie entre les représentations vectorielles des termes construites par un modèle statique FastText. L'analogie présente l'intérêt de retrouver des relations entre des concepts sans nécessiter d'entraînement sur de grandes quantités de données annotées. Elle est appliquée à de nombreuses tâches en TAL comme la désambiguïsation sémantique (Barbella et Forbus, 2013), les questions-réponses (Crouse *et al.*, 2018) et l'acquisition de relations (Chaudhri *et al.*, 2022).

Parmi les contributions les plus significatives de ce travail, notons la création et la mise à disposition d'un jeu de données composé de 180 couples de TMM reliés par trois types de relations lexicales [1] : (1) ANTI qui regroupe les relations contraires et contrastives ; (2) HYP composée de l'hyperonymie et l'hyponymie ; (3) QSYN composé de la synonymie, de la quasi-synonymie et de la co-hyponymie. Ce jeu de données a été créé par projection sémantique, une méthode souvent utilisée pour identifier des relations sémantiques entre TMM (Morin et Jacquemin, 1999 ; Hazem et Daille, 2015). Nous avons étendu les relations aux TMM entre TS de la ressource DiCoEnviro. Par exemple, la relation entre *sec* et *humide* peut être étendue aux TMM *air sec* et *air humide*.

Le projet ADDICTE

Notre thèse est réalisée dans le cadre du projet ADDICTE [2] (Analyse distributionnelle en domaine de spécialité), financée par l'Agence nationale de la recherche (ANR-17-CE23-0001)). Le projet ADDICTE réunit des membres de quatre équipes de recherche : CEA LIST [3], CLLE [4], LISN [5] et LS2N [6]. L'objectif général du projet est de proposer des solutions opérationnelles à

1. https://github.com/YizWang/List-of-semantically-linked-MWTs
2. https://anr-addicte.ls2n.fr/projet/
3. https://list.cea.fr/fr/
4. https://clle.univ-tlse2.fr
5. https://www.lisn.upsaclay.fr/
6. https://www.ls2n.fr

l'analyse sémantique distributionnelle en domaine de spécialité en exploitant des caractéristiques linguistiques et terminologiques du matériau textuel pour que l'analyse distributionnelle en domaine de spécialité puisse atteindre le même niveau de maturité que pour les grands corpus de langue générale. Trois aspects sont étudiés dans le cadre du projet : (*i*) l'amélioration endogène des contextes distributionnels en considérant des unités terminologiques qui sont les éléments essentiels d'une terminologie et véhiculent une part importante des connaissances d'un domaine de spécialité ; (*ii*) l'amélioration exogène des contextes distributionnels en enrichissant les contextes distributionnels par des ressources externes ; (*iii*) l'amélioration de la nature des contextes distributionnels en proposant une représentation distributionnelle pouvant tirer parti d'informations endogènes et exogènes. Notre étude se place dans le premier volet.

Le domaine de l'environnement

Nous nous intéressons dans ce travail aux TMM du domaine de l'environnement. L'environnement, au sens large, désigne l'ensemble des conditions naturelles (physiques, chimiques, biologiques) et culturelles (sociologiques) susceptibles d'agir sur les organismes vivants et sur les activités humaines (Robert et Rey, 2001). Il se distingue d'autres domaines de spécialité par sa nature multidisciplinaire : il fait appel à des connaissances relevant de domaines techniques, scientifiques, sociaux et économiques et couvre entre autres la climatologie, l'énergie, la biologie et l'agriculture. Les termes de l'environnement peuvent par exemple dénoter des ressources naturelles (*eau*, *carbone*, *forêt*), des impacts (*érosion*, *pollution*, *déforestation*), des acteurs (*agriculteur*, *écologiste*), des techniques (*fosse septique*, *station d'épuration*, *éolienne*), mais également des propriétés (*polluant*, *propre*) et des activités (*exploitation*, *déversement*, *recyclage*). Certaines connaissances du domaine ne sont pas encore stabilisées, notamment les concepts liés à des sujets récents comme le changement climatique (L'Homme, 2016).

Les textes dans ce domaine sont destinés à des publics hétérogènes, présentent différents niveaux de spécialisation et appartiennent à un éventail de genres. Rappelons que les termes peuvent se comporter différemment dans des textes de différents niveaux de spécialisation et que le sens des termes est souvent moins précis dans les textes de vulgarisation relativement à celui qu'ils ont dans les textes hautement spécialisés (Botta, 2013).

Plusieurs ressources terminologiques décrivent des termes du domaine de l'environnement. Certaines sont des terminologies générales qui contiennent des termes du domaine de l'environnement, telles que TermSciences [7] (portail terminologique multidisciplinaire du CNRS) et le *Grand dictionnaire terminologique* [8] (dictionnaire terminologique canadien). D'autres sont des terminologies spécialisées de termes d'un sous-domaine comme le *Glossaire écologique* [9] ou *Glossary of Water Resource Terms* [10] (glossaire anglais sur les ressources en eau). La plupart de

7. http://www.termsciences.fr
8. https://gdt.oqlf.gouv.qc.ca
9. https://www.conservation-nature.fr/ecologie/glossaire/
10. http://www.edwardsaquifer.net/glossary.html

ces ressources terminologiques fournissent uniquement les définitions des termes.

Structure du mémoire

Ce mémoire est organisé de la manière suivante.

- Les deux premiers chapitres sont consacrés aux notions de base en terminologie : les objets dans le chapitre 1 ; la structuration de terminologie et les relations principales entre termes dans le chapitre 2.
- Nous présentons dans le chapitre 3 les modèles sémantiques distributionnels, leur base théorique et l'hypothèse distributionnelle. Trois types de modèles sont présentés ainsi que leur adaptation aux domaines spécialisés.
- Nous présentons ensuite dans le chapitre 4 une revue de littérature sur l'acquisition des relations sémantiques et notamment sur la projection sémantique, la substitution lexicale et l'analogie qui sont trois méthodes utilisées dans notre étude.
- Notre étude s'appuie sur deux jeux de données. L'un est composé de TMM synonymes extraits de la ressource terminologique IATE et validés manuellement (chapitre 5). L'autre se compose de TMM reliés par des relations diverses, construit par projection sémantique (chapitre 6). Les différentes ressources pour effectuer la projection et la mise en œuvre de l'annotation manuelle sur les relations inférées sont décrites en détail dans le chapitre 6. Nous présentons notamment le corpus PANACEA utilisé pour créer les jeux de données.
- L'identification de la synonymie entre TMM fondée sur les données de IATE et l'acquisition de relations multiples entre TMM inférées par la projection sémantique sont respectivement présentées dans les chapitres 7 et 8. Nous présentons pour chacune les données de test et les expériences réalisées. Le cadre expérimental est détaillé au début du chapitre 7.

Publication liées à la thèse

1. Y. Wang, B. Daille et N. Hathout. Caractérisation des relations sémantiques entre termes multi-mots fondée sur l'analogie, In *Actes de la 28e conférence sur le traitement automatique des langues naturelles (TALN 2021)*, pages 115–124, Lille, 28 juin–2 juillet 2021.
2. Y. Wang, B. Daille et N. Hathout. A study of semantic projection from single word terms to multi-word terms in the environment domain, In *Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 95–100, Marseille, 11-16 mai 2020).

3. Y. Wang, B. Daille et N. Hathout. Exploring terminological relations between multi-word terms in semantic distributional models, accepté avec révisions mineures par la revue *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, édité par Kyo Kageura et Rita Temmerman, resoumis pour deuxième relecture.

Première partie

Contexte de travail

Chapitre 1

Notions fondamentales du terme

Avant d'aborder la question de l'identification de relations entre TMM, nous proposons une introduction sur les fondements linguistiques et terminologiques manipulés. La première notion que nous introduisons est celle de terme. Nous présentons quelques-unes de ses caractéristiques et propriétés. La compositionnalité des TMM sera aussi introduite dans cette partie.

1.1 Terme et concept

Le terme est une unité lexicale dont le sens est déterminé dans un domaine spécialisé (L'Homme, 2004). Il a une forme linguistique et une fonction dénotative. Le terme dénote une classe spécifique d'objets mentaux ou du monde réel. Il est en même temps lié à un domaine (Otman, 1995). La forme linguistique du terme peut dépendre du domaine ainsi que de l'application (Daille, 2005). Selon Bowker et Hawkins (2006), les termes doivent respecter les quatre principes suivants :

— **Monosémie** : un terme désigne un seul concept, et un concept est désigné par un seul terme ;
— **Précision linguistique** : un terme est conforme aux conventions morphologiques, syntaxiques, orthographiques et phonotactiques de la langue en question ;
— **Transparence** : un terme reflète les caractéristiques essentielles du concept qu'il désigne ;
— **Concision** : un terme est concis et conforme au principe de l'économie linguistique [1].

Le terme peut être simple ou complexe. Selon L'Homme (2004), les termes simples sont des unités lexicales composées d'une seule unité graphique (*froid*, *chaud*, *temps*) et les termes complexes sont constitués de plusieurs entités graphiques, séparées par des espaces ou des signes diacritiques tels que le trait d'union ou l'apostrophe (*air froid*, *changement du climat*, *poisson-clown*) [2].

1. L'économie linguistique vise à réduire l'effort des organes de parole, à économiser le temps et à faciliter le travail d'expression (Whitney, 1867).

2. Bien que cette définition soit opérationnelle dans des langues où les composés syntaxiques dominent, elles posent des problèmes dans les langues où les composés morphologiques prévalent, comme l'allemand (Cram et Daille, 2016).

Le sens d'un terme peut être décrit de différentes façons, et notamment dans une approche onomasiologique ou une approche sémasiologique. L'approche onomasiologique est conceptuelle. Elle est fondée sur les connaissances du domaine. Dans cette approche, le « concept » réfère aux éléments de connaissance dans un domaine spécialisé (L'Homme, 2020). Les concepts sont définis selon des critères sur lesquels les experts s'accordent. Une liste de conditions nécessaires et suffisantes détermine si un objet peut appartenir à une catégorie : les objets appartiennent à la catégorie s'ils remplissent toutes les conditions et ne remplissent que ces conditions. Par exemple, en zoologie, l'oiseau est défini comme un animal à sang chaud, qui a des plumes et qui pond des œufs. Il se distingue des mammifères par son mode de reproduction.

Les « concepts » définis par les experts peuvent coïncider ou ne pas coïncider avec la manière dont le sens est représenté dans l'esprit des non-experts. Reprenons le concept *oiseau*. Nous pouvons supposer que pour les non-experts, la capacité de voler est une caractéristique importante dans la définition des oiseaux. Cependant, cette caractéristique n'est pas associée aux oiseaux pour les experts. Les concepts dans un domaine spécialisé ne sont pas tous définis par les experts au moyen d'une liste de caractéristiques. C'est le problème principal que pose l'approche conceptuelle. Même en zoologie, où les espèces sont assez bien délimitées, établir une classification conceptuelle qui fait consensus entre les experts reste difficile.

L'approche sémasiologique est lexicale. Dans cette approche, un terme est une unité lexicale qui a un sens spécifique délimitée syntagmatiquement et paradigmatiquement. Le sens du terme est défini en contexte par son interaction avec les autres unités linguistiques. Plus précisément, cette approche ne dépend plus de la délimitation préalable des concepts et les sens des termes sont délimités en fonction des relations qu'ils partagent avec d'autres unités linguistiques. Par exemple, deux unités lexicales peuvent avoir des sens similaires (comme *conservation* et *protection*) ; deux unités lexicales peuvent avoir des sens contraires (comme *froid* et *chaud*) ; le sens d'une unité lexicale peut inclure celui d'une autre (comme *énergie* et *électricité*), etc.

La terminologie vise principalement à normaliser l'usage lexical dans les domaines spécialisés (L'Homme, 2007). Ainsi, les terminologues, avec l'aide des spécialistes du domaine, prennent des décisions visant à réduire la polysémie et la synonymie et établissent un certain nombre de « normes » linguistiques.

Dans l'approche onomasiologique et selon la Théorie générale de terminologie (TGT) (Wüster, 1974), un concept a uniquement une désignation et à l'inverse, une désignation ne correspond qu'à un concept. Les termes sont donc monosémiques dans un domaine de spécialité donné. À l'inverse, dans une approche sémasiologique, une désignation peut dénoter plusieurs concepts (L'Homme, 2020).

Les deux approches s'accordent sur le fait qu'un terme peut avoir des sens différents dans des domaines différents. Par exemple, *ligne* signifie 'connexion d'un ordinateur à un réseau de communication' dans le domaine Internet et télématique (1a), tandis qu'il est défini comme le 'trajet emprunté par un service régulier de transport en commun entre deux lieux' dans le domaine des transports en commun (1b). La difficulté réside alors à circonscrire les domaines et

à établir une distinction nette entre eux. Dans un domaine de spécialité comme l'environnement qui englobe de nombreux sous-domaines, les termes sont importés de différents sous-domaines comme l'énergie ou la chimie et peuvent se retrouver en conflit. Ainsi, *gaz* peut signifier 'corps gazeux, naturel ou manufacturé, utilisé comme combustible ou comme carburant' dans le domaine de l'énergie (2a) ou 'fluide indéfiniment expansible, c'est à dire occupant entièrement le récipient qui le contient, quel que soit le volume de celui-ci' dans le domaine de la « chimie » (2b). Gaz est donc un « terme polysémique ».

(1) a. *Le projet relatif à la formation en ligne revêt par nature une dimension mondiale.*

b. *Le métro parisien se compose d'un réseau de 16 lignes totalisant 172 kilomètres.*

(2) a. *Ce pourrait en effet être le gaz que vous alimentez votre voiture.*

b. *Les gaz qui circulent dans une cheminée intérieure refroidiront plus lentement, diminuant ainsi la formation de créosote*

Un sens peut aussi être lié à des termes différents d'un même domaine. Les synonymes sont considérés comme des variantes par la TGT permettant un processus de normalisation : une variante est choisie et devient le terme vedette. Cependant, dans la réalité, les termes synonymes ou quasi-synonymes en contexte sont fréquents.

L'approche sémasiologique est efficace à la fois pour distinguer les différents sens des termes (L'Homme, 2020) et pour identifier les termes synonymes. Pour distinguer les différents sens, elle examine les relations de chaque sens du terme polysémique en utilisant les sens des autres unités lexicales. Le Tableau 1 montre comment les 3 différents sens de *terre* sont liés aux autres unités lexicales. Pour identifier les termes synonymes, elle s'appuie sur les contextes qui partagent les mêmes unités lexicales. Par exemple, dans le domaine de la gestion environnementale, le terme *conservation* signifie 'mise en œuvre de mesures visant l'utilisation rationnelle, le maintien ou la remise en état des ressources naturelles'. Il peut être remplacé par son quasi-synonyme *préservation* dans dans le contexte (3) sans modification du sens.

(3) *Cependant, même dans les pays riches, l'éducation concernant la conservation des forêts ne fait pas toujours partie des programmes scolaires normaux.*

	Définition	**Lien lexical**
Terre$_1$	Planète du système solaire habitée par l'homme	système solaire, Mars
terre$_2$	Partie solide et émergée du globe, par opposition à la mer, aux eaux, à l'air	océan, mer
terre$_3$	Matière constituant la couche superficielle du globe où croissent les végétaux	sol, plante

TABLE 1 – Différents sens de *terre* et unités lexicales reliées

Les contextes peuvent également aider à résoudre les problèmes de polysémie, comme illustré dans le Tableau 2. Selon les contextes, le sens du terme *terre* est différent.

	Définition	**Contexte**
Terre$_1$	Planète du système solaire habitée par l'homme	Avec une population mondiale qui devrait atteindre 9 milliards d'individus en 2050, la Terre présente des symptômes inquiétants.
terre$_2$	Partie solide et émergée du globe, par opposition à la mer, aux eaux, à l'air	Le réchauffement des températures pourrait modifier la diversité des espèces présentes sur la terre et dans la mer.
terre$_3$	Matière constituant la couche superficielle du globe où croissent les végétaux & sol, plante	Une montée de un mètre du niveau de la mer mettrait en péril 4 600 hectares de terres agricoles très productives.

TABLE 2 – Différents contextes pour différents sens de *terre*

1.2 Structure linguistique du terme

Les termes comme unités lexicales peuvent avoir des catégories grammaticales différentes. Les termes complexes, qu'ils soient composés morphologiques ou syntaxiques [3], adoptent une structure morpho-syntaxique exprimée par des catégories grammaticales, c'est-à-dire des patrons (Daille, 2017).

Les termes simples appartiennent aux principales parties du discours. Les termes peuvent être des :

— noms (N) comme *température*, *climat*, *agriculture* ;
— adjectifs (A) comme *froid*, *rapide*, *climatique* ;
— adverbes (Adv) comme *globalement*, *dynamiquement* ;
— verbes (V) comme *reculer*, *accélérer*, *fondre*.

Les termes multi-mots peuvent être des composés morphologiques comme *céréaliculture* ou avoir une structure syntaxique comme *climat humide*. Les TMM instancient des patrons comme le patron NA pour le terme *climat humide*. Les TMM français sont principalement des composés syntagmatiques. Leur structure comporte le plus souvent un nom modifié par un adjectif (*climat chaud*), par un autre nom (*protection de la forêt*) ou par un syntagme prépositionnel (*température dans la masse*) (L'Homme, 2004). Le Tableau 3 décrit quelques patrons de TMM de longueur

3. Notre étude se concentre sur les TMM syntaxiquement construits. Nous avons écarté les termes morphologiquement construits peu représentés en français comparativement au termes syntagmatiquement construits.

2 du français. Daille (2017) distingue les $N\,A$, $N\,V_{:mp}$ et $N\,V_{:mg}$où $V_{:mp}$ et $V_{:mg}$ représentent respectivement un participe passé et un participe présent utilisés comme adjectifs.

Patron	**Exemple**
$N\,A$	réchauffement climatique
$N\,V_{:mp}$	aérogénérateur caréné
$N\,V_{:mg}$	contact glissant
$N\,N$	parc offshore
$N\,P\,N$	consommation en électricité
$N\,P\,D\,N$	protection de la flore
$A\,N$	petite antenne

TABLE 3 – Exemples de patrons de TMM nominaux de longueur 2 du français

La proportion des patrons varie d'une ressource terminologique à l'autre, mais les noms prévalent généralement. Seuls quelques adverbes (*globalement)* peuvent être considérés comme des termes et la plupart d'entre eux sont dérivés des adjectifs (*global*) (L'Homme, 2020). Le Tableau 4 présente la distribution des différents patrons pour les termes débutant par la lettre c dans les entrées du dictionnaire environnemental *Dictionary of environment & economy*. Certains termes peuvent avoir plus d'un patron comme *backscatter* 'rétrodiffusion' qui est soit une nom, soit un verbe. Le Tableau 4 montre que les noms (85,99 %) sont plus nombreux que toutes les autres catégories cumulées.

Nom	**Verbe**	**Adjectif**	**Adverbe**
362 (85,99 %)	22 (5,22 %)	36 (8,55 %)	1 (0,24 %)

TABLE 4 – Catégories grammaticales des entrées du *Dictionary of environment & economy* (5e édition) commençant par la lettre c

L'accent mis sur les noms dans la plupart des ressources terminologiques entraîne des incohérences. Par exemple, certains verbes et adjectifs comme *abiotique* (lié au domaine de la biologie) devraient donc être inclus dans les ressources terminologiques. Globalement, la décision de considérer les noms ou d'autres catégories grammaticales dépend de l'application pour laquelle les termes sont recueillis (L'Homme, 2020).

Un terme est une unité lexicale et il peut être réalisé en discours dans différentes formes. Ces formes différentes sont appelées « variantes ». La variation s'applique aux termes simples et complexes, et inclut toutes les parties du discours (Daille, 2017). Daille *et al.* (1996) définit la variante de terme de la manière suivante :

> *A variant of a term is an utterance which is semantically and conceptually related to an original term.*

Cette définition décrit la variante comme une forme rencontrée dans un corpus et liée à un terme original existant dans les ressources terminologiques.

Quatre catégories de variantes sont souvent rencontrées (Daille, 2017) :

— **Variantes dénominatives** : Les variantes dénominatives sont des formes lexicalisées, représentant le même concept que le terme visé, mais adoptant des formes lexicales différentes selon des contextes différents. Ce sont des quasi-synonymes des termes (Freixa, 2006) (*énergie éolienne* ↔ *courant éolien*).

— **Variantes conceptuelles** : Les variantes conceptuelles sont construites à partir de variantes dénominatives du terme. Ce sont des expansions ou réductions des dénominations des termes. L'expansion s'applique lorsque la dénomination du terme doit être élargie ou détaillée. Inversement, la réduction est utilisée lorsque le niveau de précision n'est pas jugé nécessaire dans le contexte (*énergie éolienne* ↔ *production d'énergie éolienne*).

— **Variantes linguistiques** : Lorsque le terme est invoqué au niveau du discours, une variante est choisie dans l'ensemble des variantes dénominatives. Cette variante dénominative est alors soumise aux règles orthographiques et grammaticales de la langue. Une variante linguistique est ainsi produite dont la forme diffère par rapport à la forme initiale du terme.

— **Variantes de registre** : Comme ce que nous venons de mentionner, les variantes linguistiques sont générées à partir d'une variante dénominative choisie parmi un ensemble de variantes dénominatives. Le choix d'une variante dénominative est soit arbitraire, soit soumis à des contraintes syntaxiques. Il peut également répondre à une situation de communication et à une interaction personnelle. Cela est considéré comme variante de registre (*hydrocarbure* est plus scientifique par rapport à *hydrocarbone*).

La variante peut être un synonyme du terme ou refléter une distance sémantique par rapport au terme. Elle peut aussi être elle-même un terme lié au terme original grâce à une relation conceptuelle ou sémantique. Les variantes dénominatives sont des termes potentiels des domaines spécialisés tandis que les variantes linguistiques ne sont que des formes synonymiques en contexte. Les variantes conceptuelles en tant que néologismes [4] potentiels peuvent avoir des variantes dénominatives et linguistiques (Daille, 2017).

Nous pouvons théoriquement distinguer les variantes et les termes par les définitions. Cependant, il est difficile en pratique de décider quels sont les termes et quelles sont les variantes sans tenir compte des contextes dans lesquels les termes et les variantes se présentent (et, dans certains cas, même les contextes ne permettent pas de les distinguer). Ce problème peut être expliqué par le fait que les processus qui produisent des néologismes et des variantes sont les mêmes. *Changement du climat* peut être une variante de *changement climatique* ou *changement climatique* peut être inversement une variante de *changement du climat*. À l'inverse, *resource* est

4. Un néologisme est un mot (nom commun, adjectif, expression) nouveau ou apparu récemment dans une langue, le phénomène de création de nouveaux mots communs étant appelé, de manière générale, « néologie » (Sablayrolles, 2002).

une variante linguistique du terme *ressource* (générée par la faute d'orthographe) mais il n'est pas un terme. La majorité des variantes sont donc potentiellement des termes (Daille, 2017).

1.3 Compositionnalité des TMM

Comme nous l'avons mentionné ci-dessus, un terme peut être simple ou multi-mots. Les dictionnaires spécialisés actuels accordent beaucoup d'importance aux termes multi-mots (L'Homme, 2007). La plupart des termes composant les entrées des dictionnaires spécialisés sont des TMM. Par exemple, dans le dictionnaire spécialisé du cyclisme (crée par l'Office québécoise de la langue français en 2018), sur 503 termes enregistrés, 77 % sont des TMM, comme *vélo électrique*. Par ailleurs, la plupart de ces termes sont compositionnels. C'est pourquoi nous nous intéressons à la compositionnalité sémantique des TMM. Nous introduisons dans un premier temps la notion de compositionnalité, puis nous nous intéressons à sa formulation pour les TMM.

1.3.1 Principe de compositionnalité

La compositionnalité sémantique est une propriété essentielle de la plupart des langues naturelles. Frege *et al.* (1892) est souvent considéré comme le premier à avoir explicitement formulé le principe de compositionnalité. Ce principe est d'abord posé comme une exigence de la sémantique des langages formels et des langues naturelles : un langage qui permet de construire un nombre indéfini d'expressions grammaticales porteuses de sens doit être gouverné par des principes généraux qui décrivent comment le sens d'une expression complexe dépend du sens de ses constituants (Amsili et Bras, 1998). À la suite de Frege *et al.* (1892), beaucoup de chercheurs ont donné leur propre interprétation de la compositionnalité. La version la plus citée est celle de Partee (1984) :

> *The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined.*

Dans cette définition, la fonction appliquée au sens des constituants d'une expression qui produit le sens de cette expression est le plus souvent formalisée par la syntaxe de langue. La compositionnalité sémantique établit donc que le sens d'une expression multi-mots est déterminé par la signification de ses composants et par la manière dont les composants se combinent (Pagin et Westerståhl, 2010). Ce principe permet aux locuteurs d'une langue de comprendre des expressions qu'ils n'ont jamais entendues. De nouvelles expressions peuvent ainsi être créées en combinant des unités lexicales dont le sens est connu au moyen de règles grammaticales (Pagin et Westerståhl, 2010).

1.3.2 Formalisation de la compositionnalité

Selon le principe de compositionnalité, pour une règle syntaxique α, il existe une fonction de composition [5] f telle que pour toute expression construite syntaxiquement, si l'expression est significative, son sens est obtenu en appliquant f aux sens de ses constituants et à la règle α.

Plus formellement, soit une règle syntaxique α, soient e_i et e_j deux expressions dans l'ensemble des expressions d'une langue E telles que $\mathbf{I}(e_1)$ soit le sens de l'expression e_1 et $\mathbf{I}(e_2)$ le sens de l'expression e_2 (plus généralement $\mathbf{I}(x)$ représente le sens de x), $\alpha(e_i, e_j)$ est une expression construite en appliquant la règle α aux expressions e_i et e_j. Si $\alpha(e_i, e_j)$ est porteuse de sens (c'est-à-dire que e_i et e_j permettent de construire une expression qui a du sens en fonction de la règle syntaxique α), il existe une fonction de composition f telle que $\mathbf{I}(\alpha(e_i, e_j)) = f(\alpha, \mathbf{I}(e_i), \mathbf{I}(e_j))$.

Par exemple, soit la règle syntaxique *na* exprimant la modification d'un nom par un adjectif. Pour toutes les expressions construites par la règle *na* porteuse de sens comme *culture biologique*, la fonction de composition f permet de dériver le sens de *culture biologique* à partir du sens de *culture* et de *biologique* de sorte que :

$$\mathbf{I}(\text{culture biologique}) = \mathbf{I}(na(\text{culture}, \text{biologique})) = f(na, \mathbf{I}(\text{culture}), \mathbf{I}(\text{biologique}))$$

1.3.3 Degrés de compositionnalité

Les TMM sont inclus dans l'ensemble des expressions multi-mots (EMM). Baldwin *et al.* (2003) classent les EMM selon le degré auquel le sens de l'expression peut être prédit en combinant les sens de ses composants. Les EMM sont classées en trois catégories : non-compositionelles, semi-compositionnelles et compositionnelles.

Les constituants des EMM compositionelles conservent leur sens. Ces constituants se combinent selon les règles syntaxiques pour produire une lecture compositionnelle (Sag *et al.*, 2002). Par exemple, la définition du TMM *réchauffement climatique* est : 'modification du climat de la Terre, caractérisée par un accroissement de la température moyenne à sa surface'. Son sens peut être directement obtenu à partir du sens de *réchauffement* ('action de réchauffer, de se réchauffer') et celui de *climatique* ('relatif au climat'), ainsi que de la règle syntaxique qui construit cette expression (un nom modifié par un adjectif).

Le sens des EMM semi-compositionnelles contient seulement une partie du sens de ses composants, c'est-à-dire que le sens des composants n'est pas totalement restitué dans le sens total de ces EMM même s'il est possible de prédire le sens total dans une certaine mesure. Les collocations appartiennent à ce type d'expressions (Manning *et al.*, 1999). Par exemple, *flore intestinal* qui s'interprète comme 'ensemble de micro-organismes présents au niveau des

5. Dans le but de simplifier la description (et à la suite de Pagin et Westerståhl (2010)), nous considérons une fonction de composition qui prend en entrée une paire de significations et une règle syntaxique et donne en sortie un sens.

intestins' est une collocation de type *N Adj* où la tête indiquant le collectif est *flore* et le collocatif est *intestinale* ('de l'intestin'). La sémantique de *flore* n'est pas ici liée au végétal ; Elle n'est ni prédictible, ni même transparente. Cependant, il est possible de prédire le sens total approximativement.

Les EMM non-compositionnelles, aussi appelées expressions figées, ont été largement étudiées en traitement automatique des langues (Tabossi *et al.*, 2008). Ce sont des expressions syntaxiques complexes qui sont interprétées de manière globale et dont aucune analyse décomposable n'est possible. Par exemple, l'expression *d'arrache-pied* qui signifie 'avec acharnement, persévérance' n'est lié à aucun des sens de ses constituants.

Les TMM ont souvent un sens compositionnel ou semi-compositionnel (L'Homme, 2004). Par exemple, le sens de *élevage biologique* peut être dérivé à partir du sens de *élevage* et de *biologique*, i.e. une méthode d'élevage conforme aux principes de l'agriculture biologique. Les occurrences de *biologique* dans un corpus spécialisé dans le domaine de l'environnement montrent qu'il modifie aussi d'autres noms : *agriculture biologique*, *activité biologique*, *matière biologique*, etc. De même, *élevage* apparaît dans d'autres expressions multi-mots : *élevage intensif*, *élevage laitier*, *élevage de montagne*, etc. Les sens de *biologique* et de *élevage* restent stables dans toutes ces expressions.

1.4 Synthèse

Dans ce chapitre, nous avons présenté les notions fondamentales du domaine de la terminologie. Nous avons rappelé la définition du terme, en distinguant termes simples et termes multi-mots. La compositionnalité de ces derniers jouant un rôle important dans notre étude, nous avons précisé et formalisé la notion de compositionalité sémantique. Les TMM ayant la plupart du temps un sens compositionnel, nous exploitons cette hypothèse à la fois pour la constitution d'une ressource terminologique structurée en relations sémantiques composée de TMM et pour la prédiction de ces relations. La ressource et les méthodes fondées sur la compositonalité des TMM seront présentées dans les parties II et III. Nous avons listé les principaux patrons de TMM en français. Nous ne retiendrons pour notre étude que les termes nominaux et les patrons de bitermes constitués de noms et d'adjectifs.

La terminologie d'un domaine rend compte de la structure du domaine grâce aux relations entre ses termes (Zweigenbaum et Grabar, 2000). Dans le chapitre suivant, nous abordons la structuration des terminologies par les relations sémantiques.

Chapitre 2

Structure des termes

La terminologie est une discipline dont l'objet est l'étude des termes et des relations entre termes. Elle considère que les termes et les relations entre eux reflètent la structure des connaissances d'un domaine (Sager et Ndi-Kimbi, 1995 ; Zweigenbaum et Grabar, 2000 ; L'Homme, 2004). Bien que les termes soient des mots ou des expressions d'une langue particulière, les terminologues ont conçu des principes et des méthodes pour les décrire d'une manière différente que ceux utilisés en lexicologie ou en sémantique lexicale (L'Homme, 2020). Les termes du domaine sont étudiés à travers leurs relations avec les autres termes, par exemple, la relation d'antonymie qui existe entre *boisé* et *déboisé* ou entre *charge* et *décharge* ou la relation d'hyperonymie qui existe entre *combustible* et *charbon* ou entre *énergie* et *électricité*.

Ce chapitre aborde les principales relations conceptuelles et terminologiques entre termes. Quatre ressources terminologiques sont présentées pour illustrer la variété de la description et de l'organisation des termes dans les ressources terminologiques. Trois de ces ressources sont utilisées dans notre travail.

2.1 Relations entre termes

Nous avons vu dans le chapitre 1.1 que le traitement du sens des termes peut s'effectuer selon deux approches. La première est l'approche conceptuelle dont l'objectif est de montrer comment les éléments de connaissance d'un domaine sont organisés en fonction de relations conceptuelles. La seconde est l'approche lexico-sémantique qui vise à expliquer le sens des termes et les relations terminologiques qui existent entre eux.

L'Homme (2020) a décrit les relations structurant les terminologies selon ces deux approches. Les relations conceptuelles dans l'approche conceptuelle comprennent principalement les taxonomies et les relations partitives. Les relations terminologiques dans l'approche lexicale sont classées en deux catégories : les relations paradigmatiques composées des relations comparables aux relations sémantiques lexicales classiques comme la synonymie et l'hyperonymie ; les relations syntagmatiques qui décrivent comment les termes sont combinés avec d'autres unités

lexicales.

Dans la suite de ce chapitre, ces relations entre termes sont présentées en détail relativement aux deux types d'approche.

2.1.1 Relations conceptuelles

Dans l'approche conceptuelle, un concept d'un domaine spécialisé est déterminé par sa position dans une structure conceptuelle obtenue par classification. La classification est effectuée relativement à un ensemble de caractéristiques partagées entre concepts. Certains concepts partagent la plupart de leurs caractéristiques tandis que d'autres n'en partagent que certaines. Par exemple, dans le domaine de la zoologie, *lion* a plus de caractéristiques communes avec *tigre* qu'avec *dauphin*.

Taxonomie

Les relations les plus importantes en terminologie dans l'approche conceptuelle sont taxonomiques. Elles regroupent les relations hiérarchiques et asymétriques. Elles s'établissent entre des concepts qui partagent une partie de leurs caractéristiques. L'exemple (4b) décrit quelques-unes des caractéristiques communes aux trois concepts (4a) du domaine de l'environnement.

(4) a. **Concepts :** gaz à effet de serre ; méthane ; dioxyde de carbone.

b. **Caractéristiques partagées :** substance ; sous forme gazeuse ; retenir la chaleur

Dans les taxonomies, certains concepts partagent toutes les caractéristiques d'un concept. Par exemple, les caractéristiques de *gaz à effet de serre* sont incluses dans celles de *méthane* et de *dioxyde de carbone* ; *gaz à effet de serre* est de ce fait considéré comme **concept générique** ; *méthane* et *dioxyde de carbone* sont définis comme **concept spécifique**. Les deux autres termes sont des **concepts spécifiques** qui se distinguent l'un de l'autre par au moins une caractéristique. La taxonomie est souvent représentée en utilisant les étiquettes <est-un> et <sorte de> (Otman, 1995). Elle peut également être représentée sous forme de graphe. La Figure 1 illustre une partie de la taxonomie du terme *ressources* extraite d'AGROVOC, une ressource terminologique qui est présentée dans la section 2.2.2. Dans cette figure, *ressources* est le concept le plus générique.

Relation partitive

Une autre relation conceptuelle est la relation partitive. Connue aussi sous le nom de <partie-de> (Otman, 1995), la relation partitive relie deux concepts si l'un (la *partie*) constitue une partie de l'autre (le *tout*). La relation partitive est hiérarchique. Différente de la taxonomie, les deux concepts dans cette relation ne partagent pas forcément de caractéristiques (L'Homme, 2020), comme ce qui est illustré par l'exemple 5. *Chaîne*, *siège* et *roue* sont des parties de *vélo*.

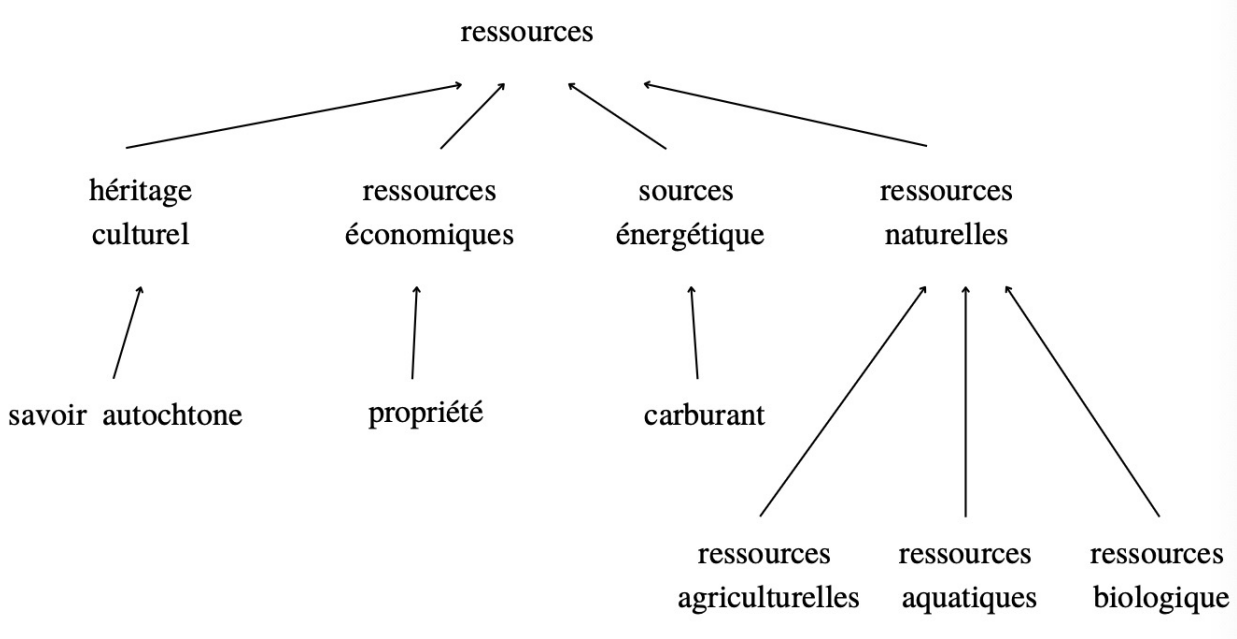


FIGURE 1 – Extrait de la taxonomie de *ressources*

Cependant, ils ont des caractéristiques (qui sont listées dans les parenthèses) différentes de celles de *vélo*.

(5) a. **Tout :** vélo (véhicule terrestre ; composé de deux roues alignées ; la force motrice est fournie par son conducteur)

b. **Parties :**
— chaîne à rouleaux (un ensemble de maillons)
— siège (un objet ; utilisé pour s'asseoir)
— roue (une pièce mécanique ; en forme circulaire ; tourner autour d'un axe passant par son centre)

De plus, les relations partitives sont plus complexes que les relations taxonomiques et peuvent être divisées en plusieurs sous-types (Pitar, 2006). Par exemple, dans certaines relations partitives, les parties sont différentes du tout. Si nous supprimons des parties, le fonctionnement du tout est affecté mais son existence ne l'est pas. Par exemple, un vélo est encore un vélo si le siège est enlevé. D'autres relations partitives sont différentes, comme celle entre *oxygène* et *dioxyde de carbone*. Si on enlève l'oxygène, il n'y a plus de dioxyde de carbone. Il existe aussi des cas où on peut supprimer des parties, sans que le fonctionnement et l'existence du tout ne soit affecté comme *éléphant* et *troupeau*.

Autres relations conceptuelles

Dans l'approche conceptuelle, d'autres relations sont utilisées en plus des relations taxonomiques et partitives. La Table 5 illustre certaines des relations conceptuelles proposées par

Sager (1990) dans différents domaines. La relation <cause-effet> a par exemple été étudiée dans le domaine médical au moyen des patrons de connaissances lexicaux par Marshman (2006).

Relation	Concept 1	Concept 2
<cause-effet>	*gaz à effet de serre*	*réchauffement climatique*
<matériel-propriété>	*aluminium*	*conductivité*
<matériel-état>	*fer*	*corrosion*
<phénomène-mesure>	*courant électrique*	*ampère*
<activité-endroit>	*culture*	*terre*

TABLE 5 – Autres relations conceptuelles

2.1.2 Relations terminologiques

Les relations terminologiques dans l'approche lexicale sont des relations entre termes considérés comme des unités lexicales et non plus comme des concepts. Elles sont souvent classées en deux catégories : les relations paradigmatiques et les relations syntagmatiques (L'Homme, 2020).

Relations paradigmatiques

Les relations paradigmatiques relient normalement des unités lexicales qui appartiennent aux mêmes parties du discours et qui peuvent remplir les mêmes fonctions syntaxiques. Cependant, elles comprennent aussi certaines relations moins prototypiques comme celles qui existent entre un verbe et un nom qui expriment son agent typique (*émettre-émetteur*). Les trois relations paradigmatiques fondamentales sont la synonymie, l'hyperonymie et l'antonymie.

Synonymie. La synonymie est une relation symétrique entre des termes qui ont des sens identiques ou similaires (Polguère, 2016). On distingue deux types de synonymie : la synonymie exacte et la quasi-synonymie.

Deux termes $term_1$ et $term_2$ sont reliés par la synonymie exacte s'ils remplissent les trois conditions suivantes (L'Homme, 2020) :

1. $term_1$ et $term_2$ ont les mêmes composantes sémantiques [1] (Fromkin *et al.*, 2018), c'est-à-dire que $term_1$ peut remplacer $term_2$ dans tous les contextes de $term_2$ sans modification du sens et vice-versa. Par exemple, *diesel* et *gazole* sont synonymes dans le domaine de l'environnement : *diesel* peut être utilisé dans les contextes (6a) et (6b) à la place de *gazole* sans modifier le sens général des deux phrases.

   (6) a. *Les véhicules à hydrogène sont tout aussi sûrs que les véhicules à essence ou au* ***gazole***.

1. Une composante sémantique est une composante du concept associé à une unité lexicale (e.g., HUMAIN + FÉMININ + JEUNE = FILLE). Plus généralement, il peut également s'agir d'une composante du concept associé à toute unité grammaticale, simple ou poly-lexicale (HUMAIN + FÉMININ + JEUNE = JEUNE FEMME ou FILLE).

b. *Nous pourrions les aider à se libérer de leur dépendance à l'égard du **gazole** et à mettre en œuvre des solutions durables.*

De même, *gazole* peut remplacer *diesel* dans les contextes (7a) et (7b).

(7) a. *L'essence et le **diesel** sont deux carburants d'origine fossile.*

b. *Si l'électrification permet en théorie une diminution des coûts d'approvisionnement en énergie, par la substitution du **diesel** par de l'électricité, à court et moyen terme, le passage à l'électrification comporte de nombreux surcoûts.*

2. $term_1$ et $term_2$ se trouvent dans les mêmes relations terminologiques. Par exemple, *diesel* et *gazole* sont tous les deux hyponymes de *carburant*. Par ailleurs, ils ont les mêmes co-localisations comme « *au* ~ » (*rouler au gazole*, *rouler au diesel*).
3. $term_1$ et $term_2$ peuvent être décrits par la même définition. Dans *Le Robert* (Robert et Rey, 2001), *diesel* et *gazole* sont tous définis comme 'Produit pétrolier utilisé comme combustible et comme carburant dans les moteurs Diesel'.

L'autre type de synonymie considéré en terminologie est la quasi-synonymie. Cette relation s'établit entre des termes qui partagent la plupart de leurs composantes sémantiques. Autrement dit, si $term_1$ et $term_2$ sont quasi-synonymes, $term_1$ peut remplacer $term_2$ dans certains des contextes de $term_2$ mais pas tous. Par exemple, *habitat* et *territoire* partagent une partie de leurs composantes sémantiques : ils indiquent tous les deux un type d'endroit spécifique pouvant être utilisé par différentes espèces animales ou végétales. Dans certains contextes, *habitat* et *territoire* peuvent être substitués l'un à l'autre sans que le sens des contextes ne soit changé (comme dans (8a) et (8b)) parce qu'ils représentent tous deux un territoire à l'échelle d'une espèce. Cependant, si nous remplaçons *territoire* par *habitat* dans le contexte (8c), la proposition devient étrange parce que *territoire* y signifie *zone qui entoure le nid d'un couple d'oiseaux*.

(8) a. *L'objectif du colloque est de faire le bilan de l'état des connaissances sur les incidences du bruit sur les animaux dans leur **habitat** naturel.*

b. *Avec la population humaine qui ne cesse de croître et l'expansion sans fin des villes qui grignotent la savane petit à petit, le **territoire** des éléphants s'est réduit à quelques îlots de végétation*

c. *les couvées restent généralement sur le **territoire** de nidification.*

Hyperonymie. L'hyperonymie (resp. hyponymie) est une relation d'inclusion entre un terme plus général (l'hyperonyme) et un terme plus spécifique (l'hyponyme). C'est un type de relation hiérarchique et asymétrique. Un hyperonyme et un hyponyme partagent une partie de leurs composantes sémantiques. Plus précisément, le sens de l'hyperonyme est inclus dans celui de l'hyponyme comme cela est illustré en Figure 2 : *combustible* a trois composantes sémantiques : MATIÈRE ; CAPABLE DE BRÛLER AU CONTACT DE L'OXYGÈNE ; PRODUISANT UNE QUANTITÉ DE CHALEUR UTILISABLE. *carburant* a pour sa part ces trois composantes plus une quatrième : QUI ALIMENTE UN MOTEUR À COMBUSTION INTERNE.

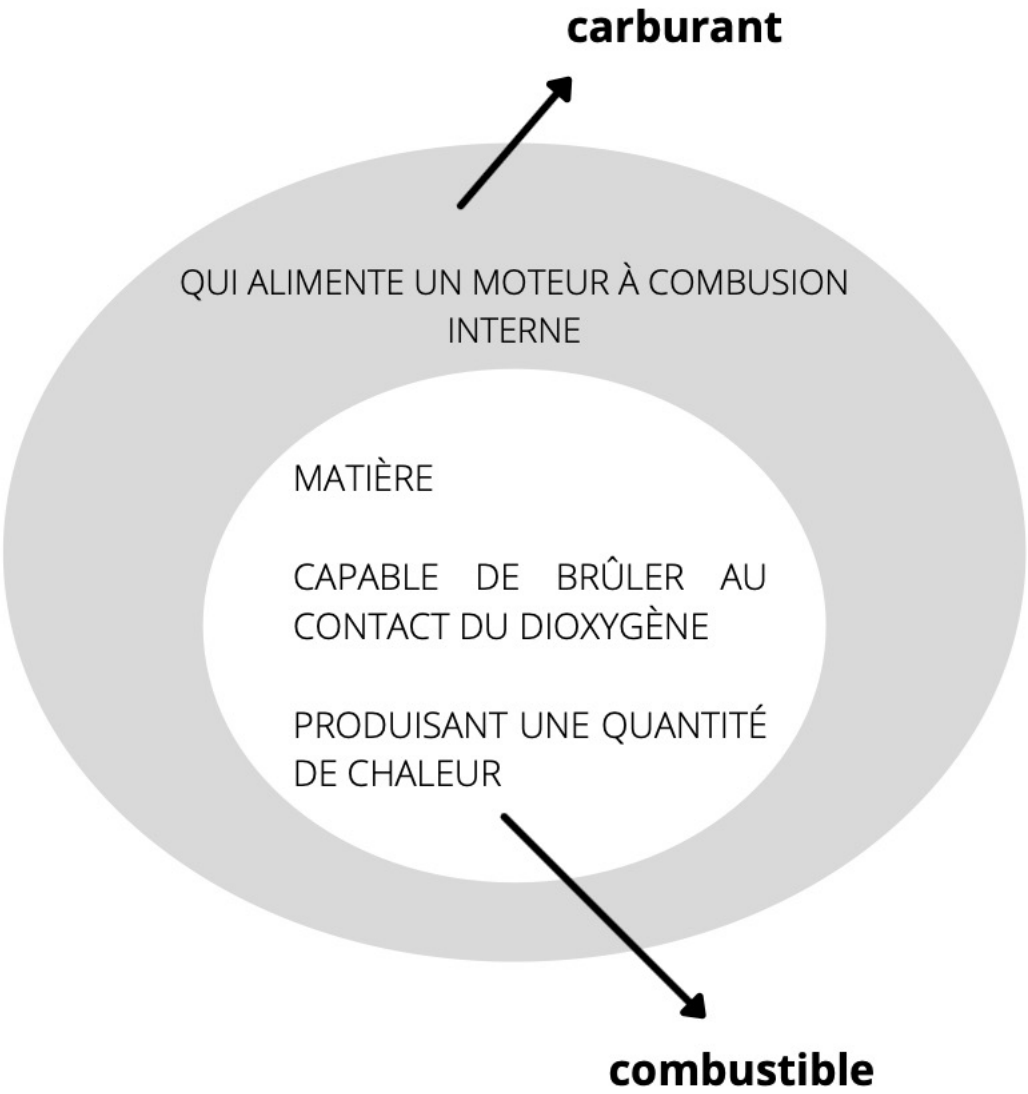


FIGURE 2 – Composantes sémantiques partagées par *combustible* et *carburant*

Pour valider la relation d'hyperonymie, Cruse (1986) proposent le test suivant :

X est une sorte de (ou un type de) *Y*

où *X* représente l'hyponyme et *Y* l'hyperonyme. Si la proposition obtenue en remplaçant *X* et *Y* par deux unités lexicales est vraie, nous pouvons considérer que ces deux unités sont reliées par une relation hiérarchique.

Les relations d'hyperonymie et d'hyponymie relient habituellement des termes nominaux. Un terme peut avoir plusieurs hyperonymes : *huile minérale*, *énergie fossile*, *combustible* sont trois hyperonymes de *pétrole*. De même, un terme peut avoir plusieurs hyponymes : *charbon*, *pétrole*, *méthane* sont trois hyponymes de *combustible*.

Opposition. L'antonymie est une relation d'opposition entre deux termes. Fondamentalement, l'antonymie relie deux termes qui partagent la plupart de leurs composantes sémantiques mais dont au moins une de leurs composantes sémantiques introduit une opposition (Murphy, 2003).

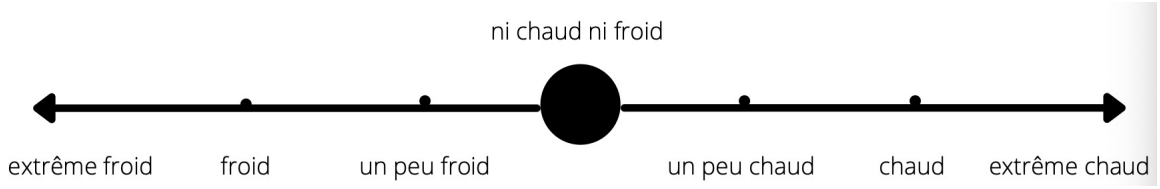


FIGURE 3 – Exemple de valeurs opposées sur une échelle

L'antonymie classique s'établit généralement entre les adjectifs comme *froid* et *chaud*. L'antonymie peut aussi s'exprimer syntagmatiquement comme entre *pollué* et *non pollué*. Les antonymies sont généralement classées en deux variétés : antonymie complémentaire et antonymie scalaire.

L'antonymie complémentaire est un type d'opposition qui existe entre deux termes (généralement des adjectifs) mutuellement exclusifs. Les termes reliés par ce type de relation sont appelés **antonymes complémentaires**, comme *organique* dans le contexte 9a et *inorganique* dans le contexte 9b.

(9) a. *Le biogaz est un gaz produit par fermentation de matière **organique** végétale ou animale dans un milieu anaérobie.*

b. *L'azote **inorganique**, le phosphore et les bactéries sont les principaux polluants découlant du fumier.*

L'antonymie scalaire est un type d'opposition qui concerne deux termes dénotant des valeurs situées aux deux extrêmes d'une échelle de valeurs comme *chaud* (dans le contexte 10a) et *froid* (dans le contexte 10b). La plage des températures est très large. Parmi les valeurs possibles, il y a celles qui peuvent être considérées comme élevées ou basses, comme l'illustre la Figure 3. Les valeurs élevées et basses sont opposées par rapport à un point central où la température n'est considérée ni comme élevée, ni comme basse. *Extrême chaud* est utilisé pour décrire les valeurs extrêmes élevées et *extrême froid* pour désigner les valeurs extrêmes basses.

(10) a. *La chaleur captée est transmise au point focal, source **chaude** d'une réaction thermodynamique.*

b. *Si le système est situé dans un environnement **froid**, les piles doivent être remplacées plus fréquemment.*

Il existe d'autres types d'opposition dans les ressources terminologiques (L'Homme, 2020) :

— **Directions opposées.** Certains termes désignent des activités (souvent des noms et des verbes) qui peuvent être opposés comme, *boisement* et *déboisement* : *déboisement* indique l'action d'enlever des arbres (défrichement, exploitation abusive, incendie, surpâturage) ; *boisement* désigne la plantation d'arbres forestiers. L'état initial de *déboisement* correspond à l'état final de *boisement*.

(11) a. *L'incidence de la mesure de **boisement** de terres agricoles a été faible.*

b. *la forêt de palétuviers qui protégeait la côte contre les puissants mouvements de marée résultant des cyclones avait presque totalement disparu par suite du* ***déboisement*** *irréfléchi*

— **Opposition selon la convention.** L'opposition contrastive (Polguère, 2016) est une relation qui implique deux termes désignant des entités qui expriment un contraste fondé sur des conventions plutôt que sur des composantes sémantiques précises. Par exemple, dans le domaine de l'environnement, *flore* et *faune* sont souvent opposées : *flore* indique l'ensemble des plantes (d'une région, d'un milieu) alors que *faune* représente l'ensemble des espèces animales vivant dans un même lieu.

(12) a. *Il aurait aussi une incidence sur l'environnement, non seulement pour les êtres humains, mais aussi pour la* ***faune****.*

b. *L'augmentation de l'utilisation des pesticides et herbicides en réponse aux nouvelles espèces parasitaires pourra endommager les communautés de* ***flore****, la qualité de l'eau et la santé humaine.*

Dérivation. Les relations paradigmatiques que nous avons présentées comme synonymie, antonymie et hyperonymie relient généralement deux termes ayant les mêmes parties du discours. Cependant, il existe également d'autres relations paradigmatiques qui relient deux termes ayant des parties de discours différentes, comme la relation de dérivation.

Deux termes reliés par la dérivation peuvent avoir le même sens mais des parties de discours différentes, comme des noms d'actions et des verbes. Par exemple, *conservation* 'action de maintenir quelque chose intact, de le maintenir dans le même état' en (13a) et *conserver* 'maintenir quelque chose intact, de le maintenir dans le même état' en (13b) ont le même sens même si l'un est un nom et l'autre est un verbe. La même situation peut également arriver entre les adjectifs de relation et les noms de propriété, comme *climatique* en (14a) et *climat* en (14b).

(13) a. À défaut d'un moratoire sur la coupe forestière, la **conservation** dans ces 15 pays représente tout de même un point de départ raisonnable pour la sauvegarde forestière.

b. Il peut donc y avoir perturbation de certains écosystèmes à haute teneur en carbone, ce qui libère du carbone dans l'atmosphère. De même, on doit **conserver** les écosystèmes à faible teneur en carbone.

(14) a. Le ralentissement voire l'arrêt du Gulf Stream est l'un des effets collatéraux du **réchauffement climatique** que l'on peut redouter.

b. Les catastrophes météorologiques, comme la vague de chaleur de l'été de 2003, devraient se faire plus fréquentes avec le **réchauffement du climat** .

Les termes reliés par la dérivation et ayant le même sens ont souvent la même structure actancielle, comme A climatique et A du climat ; conserver A et conservation de A.

La relation de dérivation peut aussi relier des termes ayant des sens différents mais étroitement liés, comme *polluer*, *pollué* et *polluant*. *Pollué* en (15a) signifie que quelque chose a déjà subi le processus exprimé par le verbe *polluer*, tandis que *polluant* en (15b) signifie que quelque chose pourrait subir le processus exprimé par le verbe *polluer*.

(15) a. De nombreux cours d'eau sont gravement pollués ; le réchauffement climatique pourrait détériorer davantage la qualité de l'eau, surtout si le ruissellement est réduit .

b. L'utilisation massive de charbon (le plus **polluant** des hydrocarbures en matière d'émission de CO2) pour produire de l'électricité ou synthétiser des substituts des carburants pétroliers accroîtrait fortement les émissions de CO2.

Les termes reliés par la dérivation et ayant les sens différents partagent certains de leurs arguments, comme A pollue B par C, B pollué et A polluant.

Relation syntagmatique

Les relations syntagmatiques décrivent la manière dont les termes se combinent avec d'autres unités lexicales dans la phrase (L'Homme, 2001). Les termes et les unités lexicales sont combinés relativement aux règles syntaxiques d'une langue. Ils remplissent ainsi des fonctions syntaxiques prédéfinies. Par exemple, un adjectif peut modifier un nom (16a) ou être utilisé comme un prédicat (16b).

(16) a. L'agriculture **écologique** peut être trouvée dans ce village.

b. L'agriculture est **écologique** dans ce village.

Les termes peuvent aussi se combiner avec des unités lexicales avec lesquelles ils partagent certaines composantes sémantiques. Par exemple, *propre* peut se combiner avec *hydrogène*, *charbon*, *carburant* désignant un type de combustible environnemental. D'autres combinaisons plus contraintes comme la collocation sont souvent rencontrées dans les domaines spécialisés. La **collocation** est une expression composée de deux unités lexicales : l'une est la base ; l'autre est le collocat sélectionné par la base (Polguère, 2016). Même s'il reste transparent, le sens des collocations ne peut pas être directement dérivé de celui de ses composants (cf. Section 1.3). Par exemple le terme *espèce menacée* peut être considérée comme une collocation : le sens du collocat *menacé* employé avec *espèce* indique une disparition de l'espèce dans un avenir proche.

2.2 Exemples de ressources terminologies structurées

Les ressources terminologiques sont de diverses formes : banques terminologiques, dictionnaires spécialisés, index, thésaurus, glossaires, etc. Ces ressources organisent les termes utilisés dans les domaines de spécialité à travers les relations entre eux.

Dans cette section, nous présentons quatre ressources terminologiques pour illustrer comment les termes sont décrits et structurés dans différents produits terminologiques. Trois d'entre elles sont utilisées dans les expériences que nous avons réalisées.

2.2.1 TERMIUM Plus

TERMIUM Plus [2] est l'une des plus grandes banques de données terminologiques et linguistiques du monde. Elle a été créée par le gouvernement du Canada. La banque enregistre des millions de termes en anglais, français, espagnol et portugais. Elle contient des acronymes, des abréviations, des synonymes, des noms de lieux géographiques, des appellations officielles d'organismes, de titres de lois et de programmes, des unités phraséologiques et des exemples d'utilisation.

Les termes dans TERMIUM Plus sont organisés sous forme de fiches terminologiques qui présentent de façon synthétique les informations caractéristiques d'un terme. Par exemple, la fiche terminologique du terme *préservation* (cf. Figure 4) précise son domaine, i.e. la conservation des aliments et conserverie, ses synonymes, ici *conservation*, sa définition et les unités phraséologiques dans lequel le terme *préservation* apparaît comme *conservation de la viande*.

FIGURE 4 – Fiche terminologique de *préservation* dans TERMIUM Plus

Les différents sens d'un terme dans les domaines différents sont décrits dans les fiches différentes. *Gaz* dans l'exemple de gauche de la Figure 5 décrit une source d'énergie sous forme gazeuse alors que dans l'exemple de droite, il désigne des agents volatiles qui peuvent être utilisés comme des psychotropes.

TERMIUM Plus distingue également les termes en fonction de leurs catégories grammaticales. La Figure 6 présente les fiches du terme *froid* comme nom et comme adjectif respectivement. Le domaine, la catégorie grammaticale et la définition du terme *froid* sont des informations répétées dans chacune de ces fiches.

2. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra

FIGURE 5 – Différentes fiches du terme *gaz* en fonction des sens différents dans TERMIUM Plus

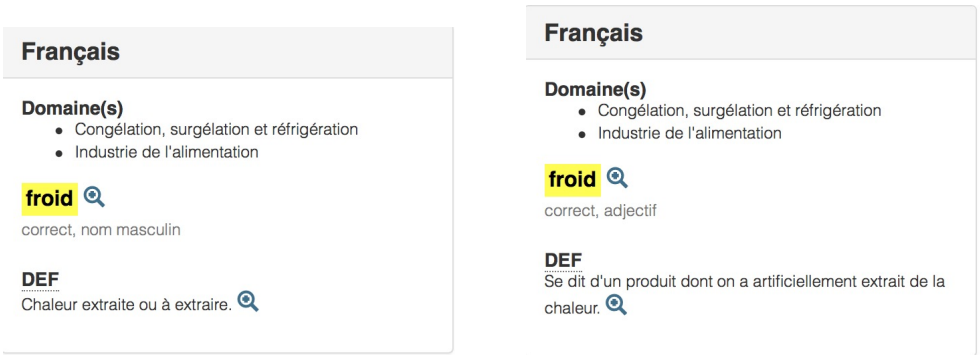


FIGURE 6 – Fiches du nom et de l'adjectif *froid* dans TERMIUM Plus

2.2.2 AGROVOC

AGROVOC est un thésaurus structuré multilingue qui couvre tous les domaines ayant trait à l'agriculture, à la forêt, à la pêche et à l'alimentation. Il contient plus de 36 000 termes et est disponible dans les six langues officielles de la FAO (Organisation des Nations unies pour l'alimentation et l'agriculture) : anglais, arabe, chinois, espagnol, français et russe. Il a aussi été traduit dans 30 autres langues.

AGROVOC se compose de termes simples et multi-mots. Ces termes sont organisés en thèmes et en sous-domaines (une macrostructuration), comme ce qui est illustré dans la Figure 7 : les termes tels que *corail*, *krill*, *insecte aquatique* sont des animaux aquatiques ; *animal aquatique* est un type de *animal*, tout comme *animal femelle*, *animal de combat*, *animal hyperprolifique*, etc. ; *animal* est un type d'*organisme*. La figure montre aussi que AGROVOC est composé de termes en plusieurs thèmes comme *mesure*, *méthode* et *objet*.

Pour chaque terme dans AGROVOC, un bloc mot décrit ses relations hiérarchiques et non hiérarchiques avec d'autres termes : termes génériques, termes spécifiques, termes apparentés et non-descripteurs, ainsi que les variantes. Des notes d'application, c'est-à-dire des notes sur l'utilisation et la couverture du descripteur, et des définitions sont également fournies dans AGROVOC afin de clarifier le sens et le contexte dans lesquels les termes s'utilisent. Les termes taxonomiques et géographiques sont marqués pour faciliter la recherche, le filtrage et le

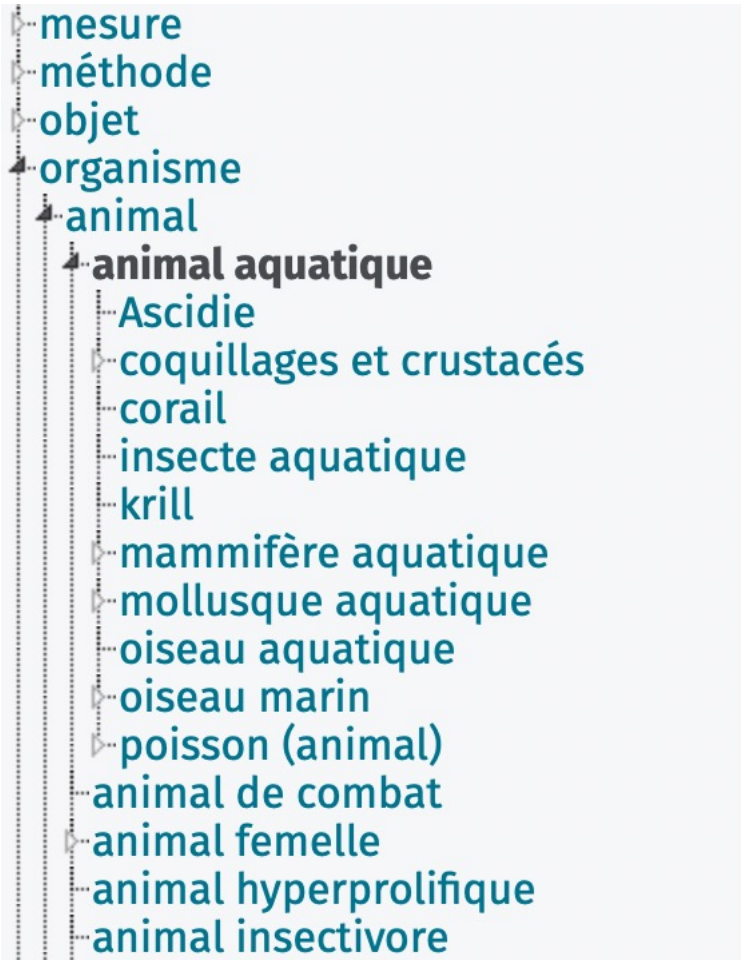


FIGURE 7 – Illustration de l'organisation hiérarchique des termes dans AGROVOC

téléchargement. Les traductions du terme en d'autres langues sont aussi fournies pour certains termes. La Figure 8 illustre un extrait du bloc de *mammifère aquatique*. Dans la figure, nous trouvons des concepts plus génériques ou plus spécifiques (p.ex., *baleine*) ; des concepts associés à *mammifère aquatique* (*cétologie* et *échouage*) ; le terme qui est relié à *mammifère aquatique* par la relation *a un membre* [3] (*Pnnipedia*) ; l'indication de classe du terme *mammifère aquatique* : il fait partie de la pêcherie ; des traductions en autres langues.

2.2.3 IATE

IATE (*Interactive Terminology for Europe*) est une ressource terminologique de traduction composée de plus de 8 millions termes dans les 26 langues de l'UE. Elle a été créée en 1999 (et est régulièrement mise à jour) dans le but de fournir une infrastructure en ligne pour toutes les ressources terminologiques de l'UE afin d'améliorer la disponibilité et la normalisation de l'information. IATE n'offre que des informations synonymiques entre les termes. Elle est cependant relativement riche et offre la possibilité d'en extraire facilement des jeux de données.

3. Y a un membre X. Un groupe social ou politique Y est constitué d'une ou plusieurs unités sociales ou politiques subsidiaires X.

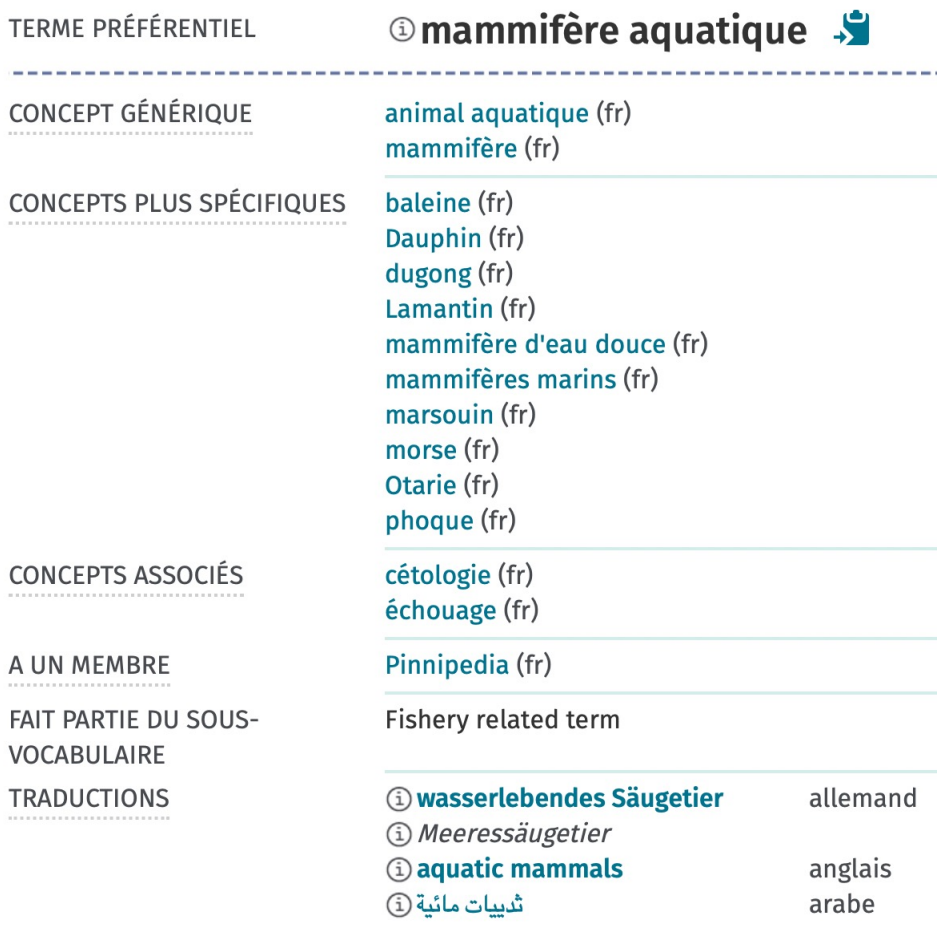


FIGURE 8 – Description du terme *mammifère auqatique* dans AGROVOC

Les termes simples et les termes multi-mots dans IATE sont d'abord organisés par domaine. Dans chaque domaine, les termes sont structurés par la relation de traduction (au niveau de la métastructuration) et par la relation synonymique (au niveau de la microstructuration). Les termes d'un domaine qui ont des sens similaires ont une identité commune, dans toutes les langues dans lesquels ils existent, comme cela est illustré dans la Figure 9 : *consommation des sols*, son synonyme en français *consommation des terres* et sa traduction en anglais *land consumption* ont une même identité *3527251*. Outre les informations relationnelles, des informations telles que la référence, la définition et les contextes (optionnels) sont également fournies.

2.2.4 DiCoEnviro

DiCoEnviro [4] est un dictionnaire multilingue de termes du domaine de l'environnement développé par l'Observatoire de linguistique Sens-Texte (OLST) [5]. DiCoEnviro décrit le sens et les propriétés linguistiques (notamment lexico-sémantiques) de 1 382 termes appartenant à divers sous-domaines du domaine de l'environnement, tels que les énergies, le changement climatique, les ressources et les transports. Les termes dans DiCoEnviro sont extraits au moyen

4. http://olst.ling.umontreal.ca/cgi-bin/DiCoEnviro/search_enviro.cgi
5. http://olst.ling.umontreal.ca

FIGURE 9 – Organisation et description des termes synonymiques d'un domaine dans IATE

de l'extracteur TermoStat [6] à partir d'un corpus spécialisé du domaine de l'environnement et leur statut de terme est validé par des spécialistes du domaine (L'Homme, 2004).

Les termes dans DiCoEnviro appartiennent aux quatre catégories majeures : substantifs (*pisciculture*), verbes (*piéger*), adjectifs (*humide*) et adverbes (*écologiquement*). Les entrées de DiCoEnviro sont majoritairement des termes simples (1 285). Le dictionnaire ne contient que 97 TMM comme *système climatique*.

Les entrées de DiCoEnviro sont décrites par des fiches terminologiques. Si un terme a plusieurs sens dans le domaine, ses sens sont traités comme des entrées séparées. La Figure 10 illustre les deux entrées du terme *gaz* : la première décrit une substance sous forme gazeuse et la seconde une source d'énergie. Chaque fiche fournit des informations grammaticales et sémantiques pour le terme vedette.

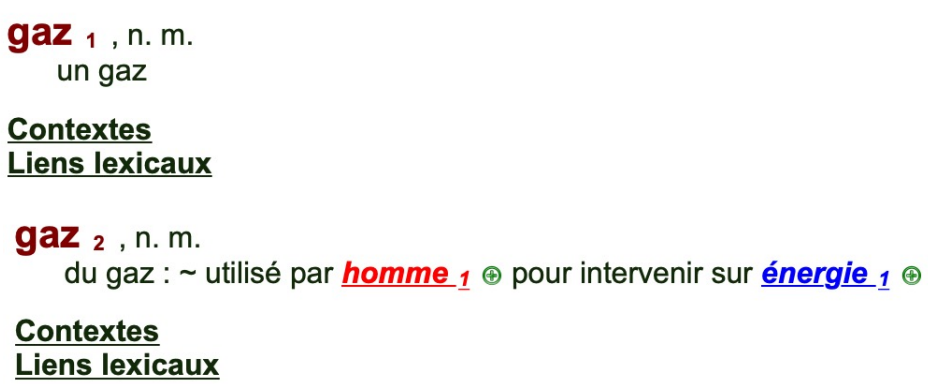


FIGURE 10 – Différentes fiches du terme *gaz* en fonction de ses sens distingués dans DiCoEnviro

La structure actancielle qui permet une meilleure compréhension du sens du terme est fournie dans la fiche. Elle se compose de deux parties : (a) les **rôles actanciels** énumèrent les actants sémantiques du terme ; (b) les **actants typiques** correspondent aux actants du terme dans des

6. http://termostat.ling.umontreal.ca

contextes linguistiques réels. Les actants sont représentés par des étiquettes qui permettent de décrire le rôle qu'ils jouent par rapport au terme telles que *AGENT*, *PATIENT* et *DESTINATION* (L'Homme et Lanneville, 2014). La Figure 11 illustre la structure actancielle du terme *flore* : *flore* a deux actants sémantiques *LIEU* et *PARTIE*. Dans les contextes réels, l'actant *LIEU* peut être réalisé par *mer* et l'actant *PARTIE* peut l'être par *plante*.

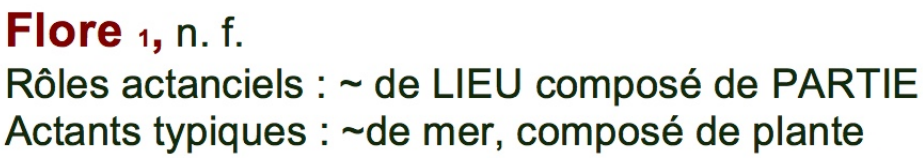


FIGURE 11 – Structure actancielle du terme *flore*

Les synonymes, les variantes graphiques et la flexion en genre d'une entrée sont listés à la suite de la description actancielle lorsqu'elle est présente (cf. Figure 12).

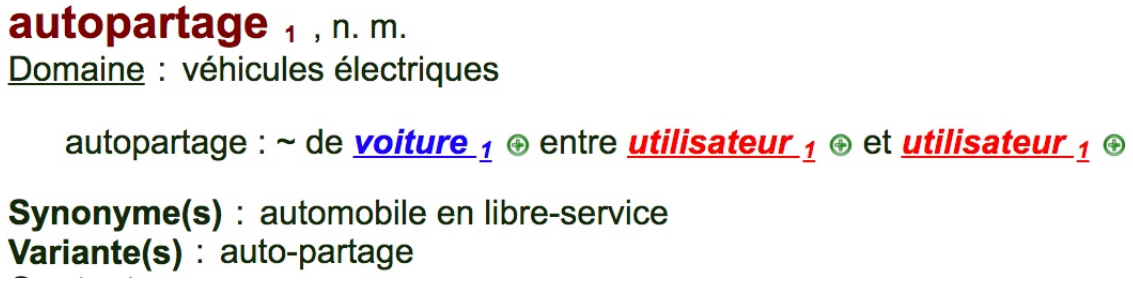
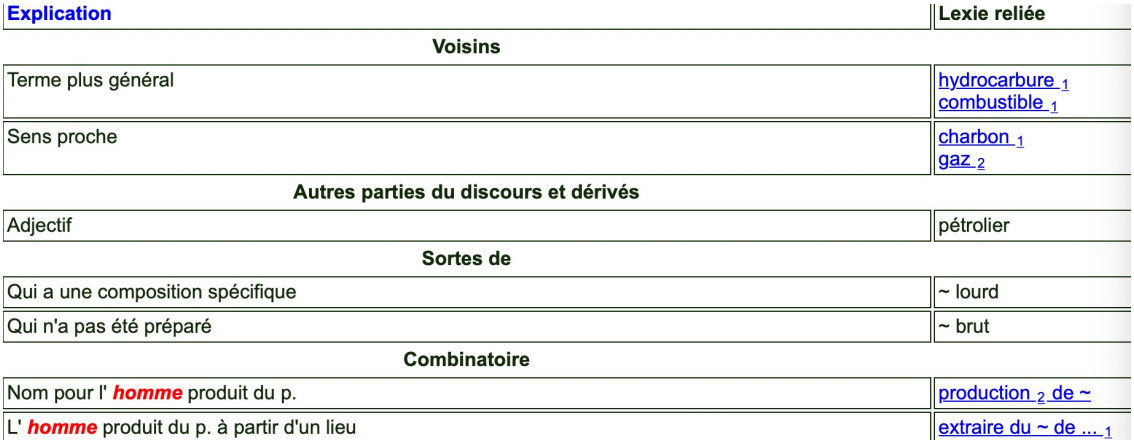


FIGURE 12 – Illustration de la description des synonymes et des variantes du terme *autopartage*

Les fiches DiCoEnviro contiennent également des relations paradigmatiques et syntaxiques listées dans un tableau illustré par la Figure 13 qui présente une partie des liens lexicaux de *pétrole*. La colonne de droite fournit les termes reliés à *pétrole* et la colonne de gauche caractérise ces liens : sens voisins (Voisins), dérivés adjectivaux (autres parties du discours et dérivés), combinaisons composées du terme et d'une modification (sortes de) et collocatifs verbaux (combinatoire) de *pétrole*.

La Table 6 récapitule les dix relations principales présentes dans DiCoEnviro, leurs dénominations et un exemple. Les fiches DiCoEnviro contiennent d'autres types de relations comme AUTRE qui sont des liens terminologiques moins réguliers comme la méronymie (*éolienne* → *nacelle*).



Explication	Lexie reliée
Voisins	
Terme plus général	hydrocarbure 1 combustible 1
Sens proche	charbon 1 gaz 2
Autres parties du discours et dérivés	
Adjectif	pétrolier
Sortes de	
Qui a une composition spécifique	~ lourd
Qui n'a pas été préparé	~ brut
Combinatoire	
Nom pour l' ***homme*** produit du p.	production 2 de ~
L' ***homme*** produit du p. à partir d'un lieu	extraire du ~ de ... 1

FIGURE 13 – Illustration des termes reliés à *pétrole*

Relation	**Exemple**
Synonymie	diesel — gazole
Variante	diesel — Diesel
Quasi-synonymie	augmentation — accroissement
Sens proche	routier — ferroviaire
Antonymie	accélérer — ralentir
Sens contrastif	flore — faune
Forme de spécifiques ou de combinaisons composées du terme et d'une modification	gaz → carbone
Hyperonymie	combustible → pétrole
Relation de dérivation	automobile — automobiliste
Collocatifs verbaux	carbone — libérer du carbone

TABLE 6 – Relations utilisées pour décrire les termes dans DiCoEnviro. Dénomination est la dénomination de la relation dans les fiches du dictionnaire.

DiCoEnviro distingue les synonymes exacts, les quasi-synonymes et un groupe très général qui contient les sens voisins et les sens proches. Les sens voisins ou proches peuvent être classés dans différentes catégories. Nous y retrouvons des co-hyponymes (ex. *ferroviaire* et *routier*), mais également d'autres termes évoqués par le terme comme *absorption* et *adsorption* ou *agricole* et *cultural*.

2.3 Synthèse

Nous avons présenté dans ce chapitre les principales relations entre les termes selon les deux approches conceptuelle et lexico-sémantique. Ces relations sont récapitulées dans le tableau 7.

L'approche conceptuelle			**L'approche lexico-sémantique**	
Relations	**Exemples**		**Relations**	**Exemples**
synonymie conceptuelle [7]	cardiopathy : cardiac disease		synonymie	diesel : gazole
taxonomie	combustible : charbon		hyperonymie et hyponymie	consommation : surconsommation
relation partitive	vélo : siège		relations opposées	sec : humide
			relations de dérivation	conservation : conserver
			combinaison syntagmatique	propre : charbon propre

TABLE 7 – Relations principales entre termes dans les approches conceptuelle et lexico-sémantique.

Nous avons d'abord présenté les deux relations les plus importantes de l'approche conceptuelle : les relations taxonomiques et les relations partitives. Nous avons ensuite décrit en détail les relations qui relèvent de l'approche lexico-sémantique, et notamment les relations paradigmatiques qui constituent l'objet de notre travail.

La synonymie que nous avons présentée comme un type de relation dans l'approche lexicale est en fait aussi considérée par l'approche conceptuelle. Cependant, l'approche conceptuelle adopte la synonymie exacte plutôt que la quasi-synonymie et la définit comme une relation entre désignations (L'Homme, 2020). La relation taxonomique est la relation la plus importante dans l'approche conceptuelle. Elle est un type spécifique de relations d'hyperonymie/hyponymie dans l'approche lexicale. Elle n'est valable qu'entre des paires de prédicats qui ont une relation de « type de » entre eux (Cruse, 1986). Par exemple, la relation entre l'hyperonyme *animal* et l'hyponyme *chien* est un exemple de taxonomie, car le chien est un type d'animal. En revanche, la relation entre l'hyperonyme *femme* et l'hyponyme *mère* n'instaure pas de taxonomie, car la mère n'est pas un type de femme.

Nous avons illustré la description et la structuration des terminologies par quatre ressources terminologiques structurées. La banque de données terminologiques et linguistiques Termium Plus, le thésaurus structuré AGROVOC, le dictionnaire de traduction IATE et le dictionnaire des termes de l'environnement DiCoEnviro. Nous avons décrit plus en détail DiCoEnviro sur lequel nous nous appuierons pour construire une nouvelle ressource, en particulier, son organisation des relations entre termes. Nous utiliserons aussi IATE et Termium Plus dans notre travail.

AGROVOC, s'il décrit également des termes en relation dans le domaine de l'environnement, comme les termes liés aux ressources naturelles, ne présente qu'une structuration en fonction des thèmes du grand domaine de l'alimentation et de l'agriculture. Il est donc difficile d'extraire les termes circonscrits au seul domaine de l'environnement. Par exemple, dans la classe « économie », *économie de l'environnement* est un terme du domaine de l'environnement mais d'autres termes comme *économie de la santé* ne le sont pas.

Toutes les quatre ressources présentées dans ce chapitre fournissent des fiches terminologiques qui contiennent la définition des termes et décrivent certaines des relations qu'ils entretiennent avec d'autres termes. Ces relations sont au coeur de la structuration des terminologies. Il est par conséquent essentiel d'élaborer des méthodes permettant d'automatiser leur identification.

Chapitre 3

Sémantique distributionnelle

Ces dernières années ont vu l'émergence de méthodes dédiées à l'acquisition de relations sémantiques pour la plupart adoptant des modèles sémantiques distributionnels (MSD).

La sémantique distributionnelle (SD) (Lenci, 2008) est une méthode de représentation du sens lexical en TAL, fondée sur l'hypothèse distributionnelle (Harris, 1954). Selon cette hypothèse, les mots qui ont les contextes linguistiques similaires ont tendance à avoir des significations similaires. La méthode représente les unités lexicales par des vecteurs construits à partir de leur distribution dans le corpus.

Pour mieux comprendre la sémantique distributionnelle moderne, nous revenons sur ses fondements et ses origines. Dans ce chapitre, nous présentons d'abord l'hypothèse distributionnelle et la notion de « contexte » qui est essentielle pour comprendre cette hypothèse. Puis, nous donnons un aperçu des modèles sémantiques distributionnels et nous les illustrons. Nous présenterons au final l'utilisation de MSD dans les domaines de spécialité.

3.1 Hypothèse distributionnelle

La sémantique distributionnelle repose sur l'hypothèse distributionnelle qui remonte au mouvement structuraliste américain. Elle a été formulée par Harris (1954) et Firth (1957). Elle stipule que les unités linguistiques qui ont des distributions similaires ont tendance à avoir des sens similaires. La distribution est définie de la façon suivante par Harris (1954) : « The distribution of an element will be understood as the sum of all its environments. », ce qui signifie que la distribution d'un élément est l'ensemble des contextes dans lesquels il apparaît. Nous pouvons formuler l'hypothèse distributionnelle pour différents éléments comme les phonèmes, les morphèmes et les mots. Par exemple, si nous nous intéressons au sens des mots, l'hypothèse est formulée comme : les mots qui ont des distributions similaires ont tendance à avoir des sens similaires.

Harris (1954) donne un exemple pour illustrer cette hypothèse :

> 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.

Selon cette citation, *oculist* 'oculiste' et *eye-doctor* 'médecin de yeux' sont des synonymes car ils partagent pratiquement les mêmes contextes. Cependant, des mots ayant des sens différents peuvent tout de même partager des contextes communs (comme *oculist* et *lawyer* 'avocat').

L'approche distributionnelle de Harris a une portée large et ne se limite pas à la sémantique. Par exemple, elle permet également d'identifier la tête d'un syntagme à l'aide de sa distribution et celle de ses composants : « if A occurs in environment X, and AB does too, but B does not, then A is the head of AB » (Harris, 1954). C'est-à-dire que étant donné un syntagme AB composé de deux mots A et B, si les distributions de AB et A sont similaires et que celles de AB et B sont différentes, alors A est la tête de AB.

Le travail de Firth (1957) est plus proche de la sémantique distributionnelle moderne (Clark, 2015). L'hypothèse de Firth s'appuie sur la notion de collocation. Elle est définie par cette citation : « You shall know a word by the company it keeps ! ». Selon Firth, les contextes distributionnels peuvent être utilisés pour expliquer le comportement des mots et le sens des mots peut être caractérisé par leurs collocations. La notion de *collocation* chez Firth repose sur les combinaisons fréquentes de mots, plutôt que sur la collocation lexicale. À la différence de l'approche de Harris, l'approche de Firth fournit des informations sur le degré de polysémie d'un mot, car en examinant les différents contextes d'un mot, il est possible d'obtenir des informations sur ses différents sens. De plus, l'approche de Harris met en contraste deux mots : les contextes que deux mots partagent donnent des informations sur leur degré de similarité, tandis que Firth considère séparément les contextes de chaque mot.

À la suite de Harris et Firth, beaucoup d'autres travaux sur l'hypothèse distributionnelle ont été proposés (Rubenstein et Goodenough, 1965 ; Harper, 1965 ; Cruse, 1986). En général, il y existe deux versions de l'hypothèse distributionnelle : une version faible et une version forte (Lenci, 2008). Dans la version faible, le comportement sémantique des mots est inféré à partir de leur distribution. L'hypothèse faible n'affirme pas que la distribution des mots peut prédire le sens des mots. L'idée est que le sens des mots a un effet spécifique sur leur distribution. En observant la distribution, nous pouvons déduire quelque chose sur leur sens. La distribution est alors considérée comme une variable latente du sens. La version forte de l'hypothèse, en revanche, affirme que la distribution des mots a une relation causale avec la façon dont le sens est déduit du texte. En d'autres termes, alors que la version faible affirme qu'il existe une relation asymétrique entre la distribution et le sens telle que $sens \longrightarrow distribution$, la version forte préconise que cette relation est symétrique. Les chercheurs en linguistique et TAL adoptent généralement la version faible de l'hypothèse distributionnelle.

3.2 Contextes

L'hypothèse distributionnelle stipule que les mots présents dans des contextes similaires ont tendance à avoir des sens similaires. La notion de « contexte » peut être définie de plusieurs manières. Traditionnellement, les contextes de type fenêtre sont distingués des contextes explorant les dépendances syntaxiques (Levy et Goldberg, 2014).

Les contextes de type fenêtre sont fondés sur les cooccurrences des mots dans une fenêtre glissante de taille fixe autour du mot cible. La fenêtre peut être orientée dans une direction (droite ou gauche) ou centrée sur le mot (droite et gauche) pour capturer les mots contexte à droite et/ou à gauche du mot cible.

La taille de fenêtre a un impact direct sur la représentation sémantique du mot cible, car elle détermine la qualité d'information extraite pour le mot cible. Considérons la phrase *des mesures adoptées en vue de lutter contre le réchauffement de la planète et le changement du climat peuvent uniquement être évaluées avec le recul*. Pour le mot cible *planète* dans une fenêtre de taille 3 orientée dans les deux directions, les mots contexte de *planète* sont *changement*, *de*, *et*, *la*, *le*, *réchauffement*. Si nous considérons une fenêtre de taille 5 orientée dans deux directions, les mots contexte sont *changement*, *climat*, *contre*, *de*, *du*, *et*, *la*, *le*, *réchauffement*.

Les mots contexte sont parfois filtrés pour supprimer les mots peu discriminants comme *le* et *la*. Par exemple, nous pouvons supprimer les mots ayant des fréquences hautes ou les mots fonctionnels dont la distribution est uniforme sur les textes de la collection.

Par ailleurs, la forme de la fenêtre peut être rectangulaire et triangulaire (Bernier-Colborne et Drouin, 2016). Une fenêtre rectangulaire est symétrique et attribue le même poids à tous les mots contexte, tandis qu'une fenêtre triangulaire donne des poids différents aux mots contexte. Les mots les plus proches reçoivent les poids les plus forts.

Les contextes de type fenêtre sont utilisés dans les MSD comme Word2Vec. C'est un type de contexte facile à implémenter qui a une complexité faible. Néanmoins, il ne prends pas en compte les informations syntaxiques entre les mots dans la fenêtre (Padó et Lapata, 2007).

Les contextes peuvent aussi être déterminés en s'appuyant sur des informations syntaxiques (i.e. les dépendances syntaxiques) : les contextes du mot dans une phrase sont représentés par ses relations syntaxiques avec d'autres mots dans la phrase. Les premières utilisations de ce type de contexte remontent aux années 1990s. Grefenstette (2012) propose le système SEXTANT dans le but de calculer la similarité entre les mots en utilisant des contextes syntaxiques. Cette méthode est aujourd'hui facile à utiliser grâce au développement de l'analyse syntaxique (Goldberg et Nivre, 2013). Prenez la phrase *Le principal facteur d'incertitude provient de l'environnement extérieur* comme exemple. La Figure 14 illustre l'analyse syntaxique de cette phrase par l'analyseur syntaxique FrMG. Par exemple, *principale* est un modifieur adjectif de *facteur* ; *facteur* est un sujet nominal de *provient* ; *incertitude* est un modifieur nominal de *facteur*. Les contextes dépendant des relations syntaxiques pour les mots *principal*, *facteur*, *extérieur* sont illustrés dans le Tableau 8. Les mots fonctionnels ne sont pas pris en compte. « −1n» représente

la relation inverse.

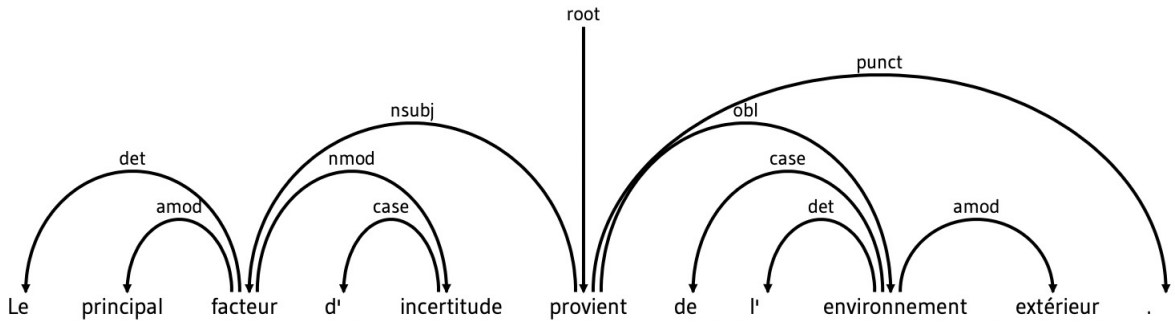


FIGURE 14 – Analyse syntaxique de la phrase *Le principal facteur d'incertitude provient de l'environnement extérieur* par l'analyseur syntaxique FrMG

Mot cible	**Contextes**
principale	facteur/$amod^{-1}$
facteur	principal/$amod$, provient/$nsubj^{-1}$, incertitude/$nmod^{-1}$
extérieur	environnement/$amod$

TABLE 8 – Contextes reposant sur des informations syntaxiques proposées par l'analyseur syntaxique FrMG pour les mots *principal*, *facteur*, *extérieur* dans la phrase *Le principal facteur d'incertitude provient de l'environnement extérieur*

Contrairement aux contextes de type fenêtre, dans les contextes de type dépendance syntaxique, les cooccurrences ne sont plus déterminées par une fenêtre de taille fixe. Par exemple, les mots contexte de *facteur* sont à des distances de 1, 2 et 3. De plus, les contextes de type dépendance syntaxique sont plus précis et l'ordre des mots peut être pris en compte. Il est démontré que les modèles incluant les informations syntaxiques tendent à capturer les mots qui sont taxonomiquement reliés, notamment les co-hyponymes parce que ces mots ont tendance à avoir les mêmes patrons syntaxiques et les mêmes relations de dépendance avec d'autres mots. Les modèles fondés sur la fenêtre ont tendance à capturer les relations associatives parce que les contextes extraits sont composés de tous les mots entourant le mot cible sans distinction.

3.3 Modèles sémantiques distributionnels

Un modèle de sémantique distributionnelle modélise le sens des mots par des vecteurs générés à partir d'un corpus. Les mots cible du modèle, i.e. les mots représentés dans le modèle peuvent être tous les mots du corpus ou une partie de ces mots ; ils peuvent aussi inclure des expressions multi-mots. Les représentations sémantiques [1] (nous parlerons parfois de représentations lexicales) des mots cible sont des vecteurs qui peuvent être construits de

1. La représentation sémantique est une caractérisation formelle de l'information transmise par les unités lexicales (Jackendoff, 1976).

différentes façons. Dans cette thèse, le terme « modèles sémantiques distributionnels » ainsi que son abréviation « MSD » désigne tous les types de modèles sémantiques distributionnels. Les modèles de type BERT peuvent également être considérés comme des MSD puisque leurs couches internes génèrent des plongements qui encodent plusieurs aspects du sens en fonction des propriétés de distribution des mots dans les textes (Mickus *et al.*, 2020).

Le Tableau 9 présente des exemples de MSD souvent utilisés en TAL. Ces modèles peuvent être classés en fonction de la méthode de construction : (*i*) les modèles fréquentiels ; (*ii*) les modèles prédictifs. Ils peuvent également être distingués selon des caractéristiques des représentations qu'ils produisent : (*i*) les modèles statiques ; (*ii*) les modèles contextuels. D'un point de vue chronologique, les modèles prédictifs statiques ont volé la vedette aux modèles fréquentiels dans un premier temps. Plus récemment, tous deux ont été remplacés dans de nombreuses tâches en TAL par les modèles prédictifs contextuels.

Modèle	**Référence**	**Classification des MSD selon la mode de construction**	**Classification des MSD selon les caractéristiques des représentations produites**
PPMI	(Landauer et Dumais, 1997)	Modèle fréquentiel	MSD statique
Analyse sémantique latente (LSA de l'anglais : *Latent semantic analysis*)	(Niwa et Nitta, 1994)		
Word2Vec (CBOW, Skip-Gram)	(Mikolov *et al.*, 2013b)	Modèle prédictif	
FastText	(Bojanowski *et al.*, 2017)		
Bidirectional Encoder Representations from Transformers (BERT)	(Devlin *et al.*, 2019)		MSD contextuel

TABLE 9 – MSD souvent utilisés en TAL

3.3.1 Modèles fréquentiels

La première génération de MSD remonte aux années 1990 et se caractérise par des modèles fondés sur les fréquences. Ces modèles construisent le vecteur de distribution d'une unité lexicale en utilisant les informations contextuelles souvent en forme de matrice de cooccurrence. C'est une méthode distributionnelle classique généralement appelée analyse distributionnelle (AD).

La première étape de l'AD consiste à calculer une matrice de cooccurrences en fonction de la fréquence de cooccurrence (généralement normalisée) des mots cible et de leurs cooccurrents (mots-contextes) dans un corpus (Lund et Burgess, 1996) [2]. Dans la matrice, chaque ligne est le vecteur de contexte d'un mot cible. Chaque dimension dans le vecteur de contexte représente une mesure de cooccurrence entre le mot-cible et un mot-contexte qui apparaît à proximité de ce mot cible dans un corpus. Le Tableau 10 donne un exemple. Le mot-cible *flore* a une cooccurrence de 9 dans la dimension étiquetée *plante* et une cooccurrence de 3 dans la dimension étiquetée *animal*.

Vocabulaire	plante	animal
flore	9	3
faune	2	8
végétation	8	1

TABLE 10 – Exemple de matrice de cooccurrences. Les lignes représentant les mots cible et les colonnes les mots contexte

Les modèles fréquentiels traitent chacune de ces lignes comme une représentation numérique de dimension D où D est le nombre de mots-contextes (avec $D = 2$ dans l'exemple du Tableau 10). L'hypothèse est que le sens d'un mot peut être déduit de ce vecteur. Par exemple, le mot *flore* serait représenté par le vecteur $V_{flore} = [9, 3]$ qui indique que les mots-contextes *plante* et *animal* apparaissent respectivement neuf et trois fois comme cooccurrent du mot-cible *flore* dans le corpus.

Deerwester *et al.* (1990) proposent pour la première fois d'utiliser cette matrice pour mesurer la similarité des mots. Il existe de nombreuses méthodes pour mesurer la similarité distributionnelle parmi lesquelles le cosinus de l'angle des vecteurs (Salton et Lesk, 1997) est la méthode la plus courante. Nous avons aussi utilisé cette mesure dans notre travail.

La Figure 15 illustre la façon dont le cosinus de l'angle des vecteurs permet de mesurer la similarité sémantique entre les mots en prenant les vecteurs du Tableau 10 comme exemples. Dans la figure 15, les vecteurs de *flore* et *végétation* forment un angle beaucoup plus petit que

2. Bien que ce soit la matrice de type contexte-mot qui est la plus utilisée, il existe également deux autres types de matrices (Turney et Pantel, 2010) : les matrices de type document-terme (Salton et McGill, 1986) dans lesquelles les lignes correspondent aux mots et les colonnes aux documents. Chaque cellule indique la fréquence d'un mot spécifique dans un document donné ; les matrices de type patron-paire (Lin et Pantel, 2001) dans laquelle les lignes correspondent à des paires de mots et les colonnes sont les règles d'inférence dans lesquelles les deux mots se présentent.

celui formé par les vecteurs de *faune* et *végétation*. Plus l'angle des vecteurs est petit, plus les sens des mots représentés par ces vecteurs sont similaires.

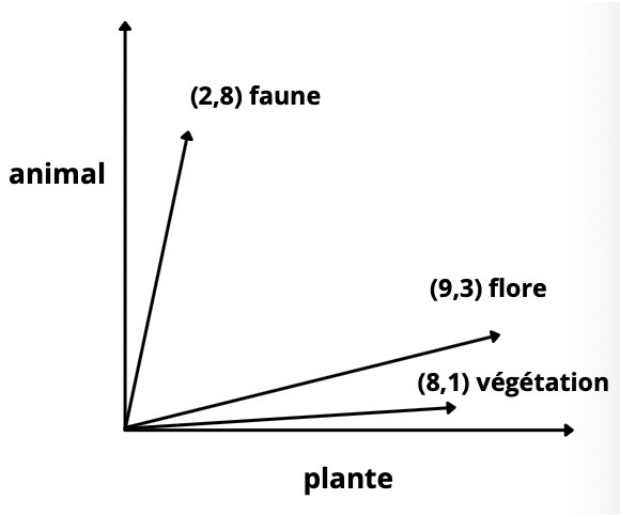


FIGURE 15 – Les vecteurs MSD des mots *flore*, *végétation* et *faune* en considérant *plante* et *animal* comme mots contexte

Plus formellement, la similarité cosinus est défini de la manière suivante :

$$cosinus(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2 \times \sum_{i=1}^{n}(B_i)^2}}$$

où A et B sont deux vecteurs de dimension n. La valeur du cosinus est dans l'intervalle [0,1]. La valeur de 0 indique une similarité minimale, tandis que la valeur de 1 indique une similarité maximale. Les valeurs intermédiaires permettent d'évaluer le degré de similarité. Reprenons l'exemple du Tableau 10. La similarité cosinus entre les vecteurs de *flore* et *végétation* est d'environ 0,98, une valeur beaucoup plus élevée par rapport à celle entre les vecteurs de *végétation* et *faune* qui est de 0,36.

Dans la pratique, les fréquences brutes ne fournissent pas toujours une mesure fiable pour modéliser le sens des mots. Les mots fonctionnels, tel que *le* co-occurrent fréquemment avec d'autres mots. Néanmoins, cette cooccurrence ne correspond pas nécessairement à une relation sémantique. De ce fait, il est préférable d'avoir une mesure qui puisse indiquer si une cooccurrence est informative ou pas, ce qui est souvent réalisé en pondérant le niveau de cooccurrence de toutes les paires de mots. Cette pondération peut être réalisée en comptant le nombre de cooccurrences (Baroni *et al.*, 2014), ou en calculant une mesure d'association entre le mot cible et le mot contexte, comme la PPMI (Church et Hanks, 1990 ; Morlane-Hondere *et al.*, 2014 ; Levy *et al.*, 2015b). Alternativement, certains travaux définissent la cooccurrence selon le fait que les deux mots partagent une relation de dépendance syntaxique (Padó et Lapata, 2007).

L'analyse sémantique latente (LSA) (Landauer *et al.*, 1998) est l'un des premiers modèles fréquentiels. Fondé sur l'hypothèse distributionnelle, le modèle suppose que :

- Le sens des phrases ou des documents est une somme du sens de tous les mots qui y apparaissent et le sens d'un mot donné est une moyenne de ses sens dans tous les documents où il apparaît.
- Les associations sémantiques entre les mots ne sont pas présentes de manière explicite, mais seulement de manière latente dans l'échantillon de langue.

La LSA est souvent utilisée dans les tâches de modélisation thématique et pour calculer la similarité des documents. Elle utilise une matrice document-terme [3] qui décrit les occurrences des termes dans les documents. Il s'agit d'une matrice creuse dont les lignes correspondent aux termes et les colonnes aux documents. Pour pondérer la matrice, la LSA utilise le TF-IDF des mots (term frequency-inverse document frequency) : le poids d'un élément de la matrice est proportionnel au nombre de fois où les termes apparaissent dans chaque document. L'importance relative des termes rares est ainsi plus forte. Pour réduire le nombre de lignes tout en préservant la similarité entre les colonnes, une décomposition en valeurs singulières (SVD) est appliquée sur la matrice générée. SVD est une technique d'algèbre linéaire qui permet de décomposer une matrice rectangulaire (une matrice dont le nombre de lignes et de colonnes est différent).

Bien que les modèles fréquentiels soient efficaces pour représenter le sens des textes pour des tâches de recherche d'informations, la représentation en sac de mots qu'ils produisent ignore la relation sémantique entre les mots et l'ordre des mots dans les phrases. Des modèles plus sophistiqués peuvent capturer ces informations, comme les modèles prédictifs que nous présentons ci-dessous.

3.3.2 Modèles prédictifs statiques

Avec l'émergence des méthodes d'apprentissage profond dans les années 2010 (Goodfellow *et al.*, 2016), un nouveau type de modèles prédictifs statiques est apparu en sémantique distributionnelle. Plutôt que de compter les cooccurrences comme les modèles fréquentiels, les modèles prédictifs statiques, comme Word2Vec (Mikolov *et al.*, 2013b) et FastText (Bojanowski *et al.*, 2017), utilisent des réseaux de neurones pour générer des vecteurs denses et de faible dimension dont l'objectif est d'améliorer leur capacité de prédiction. Ces vecteurs sont également appelés plongements de mots [4] (word embedding). Ils permettent de capter des informations sémantiques et syntaxiques entre les mots (Naseem *et al.*, 2021).

Les modèles prédictifs statiques sont fondés sur l'hypothèse distributionnelle . La technique la plus connue est celle adoptée par Word2Vec, qui entraîne un réseau de neurones sur une tâche de prédiction des relations entre les mots cible et les mots contexte. Deux approches ont été proposées : l'entraînement d'un réseau qui prédit un mot cible à partir d'autres mots dans la fenêtre de contextes (l'architecture CBOW) ; l'entraînement d'un réseau qui prédit les mots contexte probables d'un mot cible donné (l'architecture Skip-gram).

3. Le mot « terme » indique ici les mots d'intérêt dans le vocabulaire et non les termes d'un domaine spécialisé.
4. Les plongements de mots sont des vecteurs denses, de faible dimension et à valeur réelle, créés en utilisant des modèles de langue neuronaux (Turian *et al.*, 2009)

La Figure 16 présente l'architecture générale des deux approches, pour une fenêtre de 2 + 2 mots autour du mot cible. Dans les deux cas, chaque entrée est connectée à une couche cachée qui permet de réaliser la tâche. À l'issue de l'entraînement, les poids de la couche cachée sont utilisés comme plongements des mots.

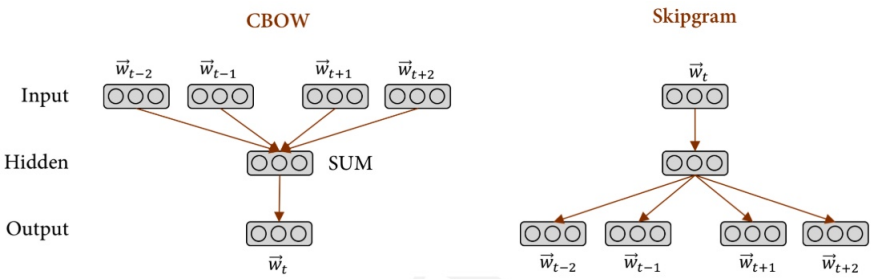


FIGURE 16 – Architecture d'apprentissage des modèles CBOW et Skip-gram

Source : (Mikolov *et al.*, 2013b)

Chacune des deux architectures a ses avantages et inconvénients. Selon Mikolov *et al.* (2013a), Skip-gram est adapté aux données de taille petite et représente bien les mots rares tandis que CBOW est plus rapide et offre de meilleures représentations pour les mots plus fréquents.

L'avènement des MSD neuronaux a apporté de nouvelles méthodologies et l'expression « plongements de mots » devient désormais le terme standard pour désigner les vecteurs distributionnels. L'apprentissage profond a radicalement modifié la portée et la diversité des applications de la sémantique distributionnelle. La première génération de MSD englobait essentiellement des méthodes informatiques visant à estimer la similarité ou la parenté sémantique entre les mots (Lenci *et al.*, 2021). Avec les modèles prédictifs statiques, la sémantique distributionnelle est devenue une approche générale qui fournit des informations sémantiques aux applications en TAL : les plongements de mots sont aujourd'hui couramment utilisés dans les architectures d'apprentissage profond pour initialiser les représentations des mots. Ils permettent aux réseaux neuronaux de capturer les similarités sémantiques entre les unités lexicales qui sont ensuite utilisées dans les tâches supervisées en aval.

La performance des modèles prédictifs statiques surpasse celle des modèles fréquentiels dans la plupart des tâches en sémantique (Baroni *et al.*, 2014). Cependant, les représentations sémantiques statiques souffrent d'une limite pénalisante (Scheepers *et al.*, 2018) : un mot aura toujours la même représentation quel que soit le contexte dans lequel ce mot se présente. Dans la pratique, une caractérisation plus fine du sens des mots est souvent nécessaire. Certaines tâches demandent une distinction de différents sens des mots. Les modèles contextuels comme ELMO (Peters *et al.*, 2018a) et BERT (Devlin *et al.*, 2019) apportent une solution à ce problème.

3.3.3 Modèles prédictifs contextuels

Une nouvelle génération de modèles sémantiques distributionnels a été récemment proposée. Elle est issue du développement des modèles de langue neuronaux profonds. Ces modèles prédictifs s'appuient généralement sur un réseau neuronal d'encodeur multi-couches et les vecteurs de mots sont appris en fonction de ses états internes. Dans ces modèles, un mot détermine les différents états d'activation dans différents contextes et est représenté par des vecteurs distincts dans chaque contexte. Par conséquent, les plongements produits par ces nouveaux modèles sont contextuels, par opposition aux plongements statiques générés par les MSD antérieurs. Par exemple, le terme *ligne* a des représentations différentes dans les deux phrases suivantes : *Ce robot se déplace uniquement en ligne droite* ; *Mon nouveau régime m'aidera à garder la ligne*. Ces modèles sont appelés modèles de langage contextuels. Une autre différence importante est que les modèles statiques sont entraînés sur les corpus choisis par l'utilisateur alors que les modèles contextuels sont pré-entraînés.

La génération de représentations lexicales n'est pas l'objectif final des modèles contextuels comme BERT ou GPT (Radford *et al.*, 2019). Ils sont principalement conçus comme des architectures générales pour développer des applications en TAL en les optimisant pour des tâches spécifiques. Leurs couches internes génèrent des plongements qui codent plusieurs aspects du sens en fonction des propriétés de distribution des mots dans les textes.

Un des modèles contextuels les plus utilisés dans la littérature est le modèle BERT. La principale innovation de BERT est qu'il est fondé sur l'architecture du modèle Transformer qui applique le mécanisme d'attention pour apprendre les relations contextuelles entre les mots (ou les sous-mots) d'un texte. Par exemple, étant donnée une phrase *la France a décidé d'accélérer le pas dans la voie des carburants verts*, pour déterminer que le mot « vert » fait référence au sens relié à l'écologie et non à la couleur verte, le Transformer peut apprendre à faire immédiatement attention au mot « carburants » et prendre cette décision en une seule étape.

Le modèle BERT est pré-entraîné sur deux objectifs de modélisation de langue : la prédiction de mots masqués (MLM) et la prédiction de la phrase suivante (NSP). MLM masque aléatoirement certains des tokens de la séquence d'entrée (15 % par exemple), en les remplaçant par un token spécial « [MASQUE] ». Par exemple, si nous masquons *eau* dans la séquence *la capacité de stockage en eau des sols*, le modèle doit prédire le token « [MASQUE] » dans le contexte *la capacité de stockage en [MASQUE] des sols* en fonction des informations disponibles à partir des tokens non masqués de la séquence. NSP est une tâche qui vise à identifier si une phrase donnée peut être considérée comme la phrase qui suit à la phrase actuelle ou non. Cet objectif est motivé par le fait que pour réussir certaines tâches, le modèle doit encoder les relations entre phrases ou recourir à des informations qui se situent au-delà de la limite de la phrase.

BERT s'appuie sur un Transformer. Un Transformer de base consiste en un encodeur pour lire le texte en entrée et un décodeur pour produire une prédiction pour la tâche. Contrairement aux modèles directionnels, qui lisent le texte d'entrée de manière séquentielle (de gauche à

droite ou de droite à gauche), les Transformers lisent la séquence entière en une seule fois. Par exemple, étant donnée une phrase *la France a décidé d'accélérer le pas dans la voie des carburants verts*, un modèle contextuel unidirectionnel représenterait *carburants* en fonction d'une partie de la phrase : *la France a décidé d'accélérer le pas dans la voie des* (si le modèle lit la phrase de gauche à droite) ou *verts* (si le modèle lit la phrase de droite à gauche). Cependant, BERT représente le mot *carburants* en utilisant à la fois son contexte précédent et son contexte suivant.

BERT ne se sert que de la partie encodeur du Transformer. L'entrée de l'encodeur de BERT est une séquence de tokens. Les tokens sont définis par le segmenteur de tokens WordPiece tokenizer. WordPiece tokenizer est une méthode de tokenisation en sous-mots, proposée par Schuster et Nakajima (2012). Il fonctionne en représentant les mots en formes complètes, soit en les divisant en morceaux de mots (lorsque les mots ne sont pas fréquents dans le corpus d'entraînement), comme « piscicole » qui va être divisé en « pis », « s », et « cicole ». Le tokenizer permet de résoudre les problèmes engendrés par la tokenisation en mots (grande taille du vocabulaire ; grand nombre de mots OOV (out-of-vocabulary) ; signification différente de mots très similaires) et la tokenisation en caractères (mauvaises représentations des séquences très longues et des tokens individuels moins significatifs).

Les tokens d'entrée sont d'abord convertis en vecteurs, puis traités dans le réseau neuronal. Comme illustré en Figure 17, les vecteurs d'entrée sont la somme des plongements de tokens, des plongements de segmentation et des plongements de position :

- **Plongements de tokens** : un token spécial [CLS] est utilisé par BERT pour marquer le début de la première phrase. Un token [SEP] est inséré à la fin de chaque phrase.
- **Plongements de segmentation** : un marqueur indiquant la phrase A ou la phrase B est ajouté à chaque token, ce qui permet à l'encodeur de distinguer les phrases.
- **Plongement de position** : un plongement de position est ajouté à chaque token pour indiquer sa position dans la phrase.

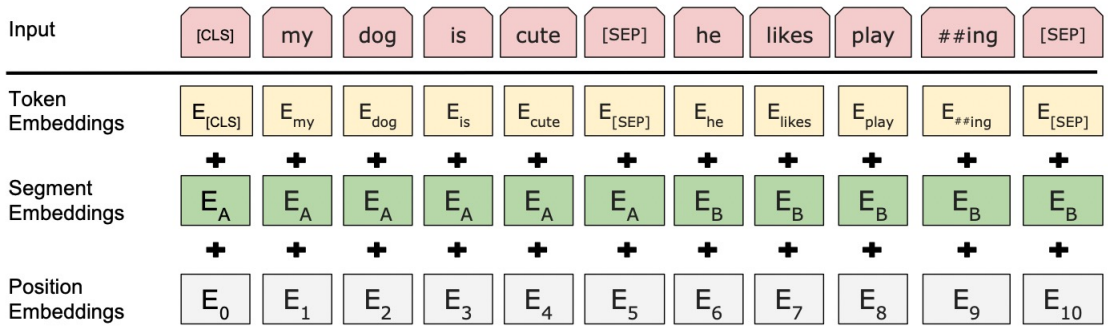


FIGURE 17 – Représentation d'entrée du modèle BERT

(Devlin *et al.*, 2019)

Essentiellement, le transformer empile une couche qui fait correspondre les séquences aux

séquences, de sorte que la sortie du modèle est également une séquence de vecteurs avec une correspondance entre les tokens d'entrée et de sortie de même position, comme ce qui est illustré dans la Figure 18. La représentation du token T_i peut être utilisée pour prédire si le token était masqué ou non. Pour prédire si la deuxième phrase est la phrase suivante de la première ou non, la représentation C qui correspond au token d'entrée « [CLS] » est utilisée. Ce dernier est transformé en un vecteur à l'aide d'une couche de classification simple. L'étiquette (IsNext) est attribuée à l'aide de la fonction softmax [5].

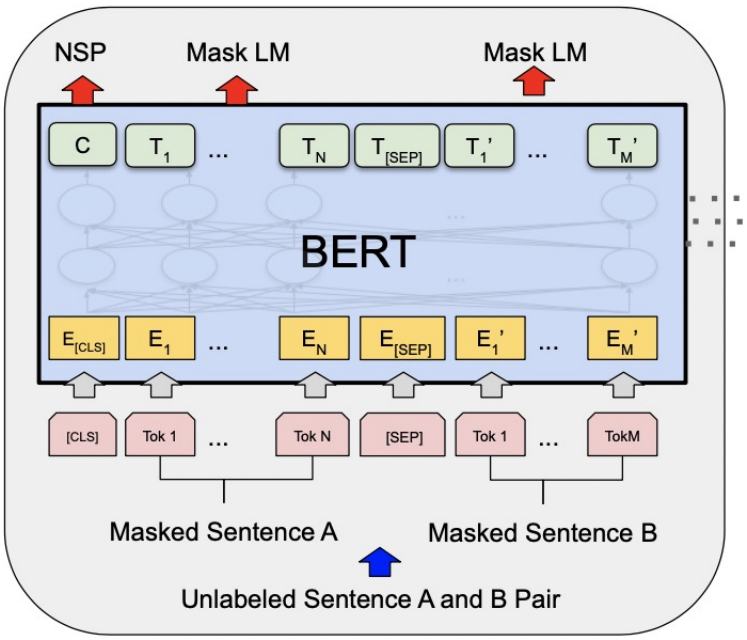


FIGURE 18 – Architecture du modèle BERT

(Devlin *et al.*, 2019)

Les améliorations apportées par les modèles contextuels la plupart des tâches du TAL ont conféré une grande popularité aux plongements contextuels qui ont rapidement remplacé les plongements statiques, en particulier dans les tâches en aval. La raison de ce succès est attribuée à la capacité de ces représentations à capturer de nombreuses caractéristiques linguistiques (Tenney *et al.*, 2019). En particulier, le sens des mots dépend du contexte (Wiedemann *et al.*, 2019), ce qui permet de surmonter une limite importante des plongements statiques qui fusionnent les différents sens d'un mot dans un seul vecteur.

3.4 MSD dans les domaines spécialisés

Les modèles de réseaux neuronaux modernes en TAL sont des systèmes d'apprentissage puissants qui s'adaptent facilement à de nouvelles tâches et à de nouveaux jeux de données

5. La fonction softmax est une fonction qui transforme un vecteur de K valeurs réelles en un vecteur de K valeurs réelles dont la somme est égale à 1.

lorsqu'une supervision suffisante est fournie. Cependant, ils sont également très fragiles lorsqu'ils sont utilisés sur des données de domaines spécialisés car la distribution des informations et les objectifs des tâches peuvent être très différents (Ma *et al.*, 2019).

Plus précisément, il a été montré que les plongements de mots varient d'un domaine à l'autre en raison de variations lexicales et sémantiques (Hamilton *et al.*, 2016). De ce fait, la plupart des modèles de représentation sémantique entraînés avec des corpus généraux ne modélisent pas de façon adéquate les termes qui ont des fréquences basses ou qui ont des sens spécifiques au domaine. Dans la section suivante, nous aborderons l'utilisation des MSD dans les domaines spécialisés.

3.4.1 Modèles spécifiques

La solution la plus simple pour obtenir des représentations de termes adaptées consiste à créer des modèles neuronaux spécifiques à un domaine. Ces modèles sont créés en utilisant soit des corpus spécialisés soit à la fois des corpus généraux et spécialisés. Le vocabulaire du modèle et les plongements de mots générés sont alors mieux adaptés au domaine spécialisé. En plus des modèles statiques qui peuvent être entraînés par les utilisateurs sur leurs propres corpus spécialisés, il existe aussi des modèles pré-entraînés tels que BioBERT (Lee *et al.*, 2020) et SciBERT (Beltagy *et al.*, 2019).

Pour créer BioBERT, les auteurs ont complété le corpus utilisé pour entraîner le modèle BERT original avec PubMed [6] et PMC [7]. PubMed est une ressource de citations et de résumés biomédicaux tandis que PMC est une archive électronique d'articles des revues biomédicales et de sciences de la vie. Le modèle est construit en l'initialisant avec les poids du modèle pré-entraîné $BERT_{BASE}$. Ce modèle pré-entraîné qui inclut à la fois des corpus de domaine général et des corpus biomédicaux permet d'obtenir des représentations des termes plus spécialisées. Il surpasse largement $BERT_{BASE}$ dans les tâches de fouille de textes biomédicaux.

SciBERT est un modèle BERT entraîné sur une grande quantité de textes scientifiques. Il est totalement entraîné en utilisant la même configuration et la même taille de données que $BERT_{BASE}$, sans initialisation avec les poids d'un modèle BERT. Les données d'entraînement sont un corpus de 3,17 milliards de tokens extraits de Semantic Scholar Corpus [8] comprenant des articles biomédicaux (82 %) et informatiques (18 %). Le vocabulaire de sous-mots SCIVOCAB a aussi été généré à partir de ces corpus. Le modèle SciBERT a été évalué sur douze tâches, dans le domaine informatique, le domaine biomédical et sur des tâches multi-domaines. Il a obtenu les meilleurs résultats dans les huit tâches.

Bien que la création des modèles neuronaux spécifiques au domaine soit une méthode efficace pour générer des représentations de termes, elle nécessite de grands corpus spécialisés qui font défaut dans de nombreux domaines. De plus, l'entraînement des modèles est long et coûteux.

6. https://catalog.data.gov/dataset/pubmed
7. https://www.ncbi.nlm.nih.gov/pmc/
8. https://www.semanticscholar.org/

De ce fait, beaucoup de recherches adaptent les modèles de langue générale aux domaines spécialisés.

3.4.2 Adaptation de domaine

L'adaptation de domaine (Csurka, 2017) vise à entraîner un réseau de neurones sur un ensemble des données source dans le but d'obtenir une bonne précision sur des données cible qui sont différentes de données source.

L'optimisation des modèles neuronaux est une approche supervisée souvent utilisée pour adapter les modèles à des domaines spécifiques. Un réseau de base est entraîné avec les données source dans un premier temps. Puis, les *n* premières couches du réseau de base sont copiées. Les données étiquetées du domaine cible sont ensuite utilisées pour optimiser les dernières couches du réseau. L'adaptation peut optimiser les couches copiées ou les laisser inchangées (Chu *et al.*, 2016). Comme la méthode d'optimisation nécessite souvent une quantité importante de données annotées du domaine cible, des méthodes non-supervisées ont été proposées. Les deux méthodes non-supervisées qui permettent d'aligner les domaines source et cible (Wilson et Cook, 2020) les plus courantes sont la minimisation de la divergence et la méthode adverse.

La minimisation de la divergence entre la distribution source et la distribution cible vise à générer des descripteurs invariants relativement au domaine, i.e. les descripteurs ayant la même distribution sur les données d'entrée du domaine source et du domaine cible. Pour mesurer la distance entre les distributions de la source et de la cible, plusieurs approches sont proposées, telles que la divergence moyenne maximale (Gretton *et al.*, 2006), l'alignement par corrélation (CORAL) (Sun *et al.*, 2016) et la divergence de domaine contrastive (Kang *et al.*, 2019). Les divergences sont généralement mesurées par des formules mathématiques non paramétriques et définies par les humains. Cette approche n'est pas spécifique à des données particulières ou à des tâches spécifiques.

Les méthodes adverses impliquent souvent les réseaux discriminants [9] (Ryu et Lee, 2020 ; Nishida *et al.*, 2020 ; Karouzos *et al.*, 2021). Ce type de méthodes permet à la fois de rapprocher les deux domaines et d'apprendre des descripteurs discriminants et invariants par rapport au domaine. Ces méthodes fonctionnent également pour l'adaptation à un domaine multi-sources (Li *et al.*, 2018). Cependant, ce type de modèles est généralement difficile à entraîner car le problème d'optimisation implique une minimisation par rapport à certains paramètres et une maximisation par rapport aux autres.

9. Le discriminateur est simplement un classificateur qui distingue les données réelles des données créées par un générateur.

3.5 Synthèse

Dans cette section, nous avons introduit l'hypothèse distributionnelle à la base des MSD. Puis, nous avons présenté trois types de MSD. À partir de Word2Vec, les modèles entraînés sur de grands corpus ont démontré une puissance significative dans la plupart des tâches en TAL. Plus récemment, les modèles contextuels permettent de capter les sens différents des mots dans des contextes différents. Ces modèles peuvent être directement optimisés pour des tâches spécifiques.

Notre travail portant sur les termes, nous avons également rappelé comment les MSD sont utilisés dans les domaines spécialisés. Les représentations des termes peuvent être directement générées par les modèles entraînés sur des corpus spécialisés. Des méthodes qui permettent d'adapter les MSD du domaine général à un domaine spécialisé ont aussi été présentées.

Les MSD vont être explorés dans notre étude pour l'identification des relations lexicales entre TMM. Nous avons entraîné des modèles FastText sur un corpus spécialisé du domaine de l'environnement. De même, le modèle CamemBERT que nous utilisons est entraîné sur un corpus du domaine général. L'adaptation de domaine d'un modèle pré-entraîné dépasse le cadre de notre travail.

Chapitre 4

Acquisition de relations sémantiques

Dans une terminologie, les termes sont connectés par différents types de relations. Ces relations peuvent être identifiées par des experts, extraites de ressources existantes ou acquises à partir de corpus. La plupart des techniques proposées dans la littérature pour la caractérisation des relations sémantiques sont indépendantes du domaine (Lafourcade et Ramadier, 2016). Les relations terminologiques de base étant très proches des relations sémantiques classiques, il est possible de réutiliser les techniques conçues pour l'acquisition des secondes, dans le domaine général, pour identifier les relations entre termes dans les domaines spécialisés (Lee *et al.*, 2004 ; Abacha et Zweigenbaum, 2011 ; Chaudhri *et al.*, 2022). Dans cette section, nous présentons une brève revue de la littérature sur l'acquisition des relations sémantiques. Nous examinons successivement les patrons lexico-syntaxiques, les modèles sémantiques distributionnels (MSD), et les méthodes que nous avons mis en œuvre : la projection sémantique, la substitution lexicale et l'analogie.

4.1 Patrons lexico-syntaxiques

Les relations sémantiques peuvent être identifiées dans les textes à l'aide de patrons lexico-syntaxiques, c'est-à-dire d'expressions régulières sur des séquences de mots. Les patrons lexico-syntaxiques peuvent exploiter diverses annotations dont les mots peuvent être munis comme des catégories grammaticales ou des classes sémantiques (Kamel et Aussenac-Gilles, 2009). Les patrons sont utilisés pour identifier les séquences textuelles qui en sont des instances.

L'acquisition de relations sémantiques au moyen de patrons consiste à définir des configurations dans lesquelles une relation visée s'établit entre deux ou plusieurs termes. Par exemple, le patron `SN est un type de SN` permet de prédire une hyperonymie entre deux syntagmes nominaux (`SN`) comme entre *le carburant* et *combustible* à partir de la séquence *le carburant est un type de combustible*.

Les méthodes fondées sur les patrons comportent généralement quatre étapes (Auger et Barrière, 2008) : (*i*) définition de la relation sémantique cible à acquérir ; (*ii*) définition ou

découverte des patrons qui expriment explicitement cette relation dans les corpus (*iii*) rechercher des instances de la relation en utilisant les patrons ; (*iv*) ajouter les nouvelles instances à une terminologie nouvelle ou existante après validation par des experts ou des terminologues. Par exemple, un patron comme `SN, c'est un autre mot pour dire SN` permet d'identifier des termes en relation de synonymie. Si on trouve dans un corpus une instance de ce patron comme *voiture, c'est un autre mot pour dire automobile*, on peut prédire que *voiture* et *automobile* sont synonymes. La définition et la découverte de patrons (2e étape) ont fait l'objet de nombreux travaux dont nous discutons dans la suite de cette section.

4.1.1 Définition manuelle de patrons

La définition manuelle de patrons est coûteuse en temps (Hakenberg *et al.*, 2010). Elle implique parfois la participation d'experts pour caractériser les relations spécifiques aux domaines. Certains patrons comportent des abstractions exprimées au moyen d'étiquettes qui représentent des classes de mots et des constituants syntaxiques. Par exemple, Hearst (1992) définit des patrons qui contiennent des syntagmes nominaux, comme `SN, surtout {SN}* {ou|et} SN` pour extraire des couples en relations d'hyperonymie comme dans la séquence textuelle *[la plupart des pays européens]$_{SN}$, surtout [la France]$_{SN}$, [l'Angleterre]$_{SN}$ et [l'Espagne]$_{SN}$....* Cette instance du patron permet de prédire trois relations : *France* est un hyponyme de *pays européen* ; *Angleterre* est un hyponyme de *pays européen* ; *Espagne* est un hyponyme de *pays européen*.

Dans la même lignée, Liu *et al.* (2011b) proposent des patrons permettant de rechercher diverses relations sémantiques du domaine biomédical comme la synonymie, l'hyponymie ou la relation partitive. Les patrons de ces auteurs incluent ceux de Hearst (1992) et de Berland et Charniak (1999). Ils sont complétés par des patrons définis manuellement à partir des documents cliniques. L'étude a permis de montrer que les patrons génériques de langue générale sont également efficaces pour la découverte de relations entre des termes du domaine biomédical. Pour leur part, Ng et Wong (1999) ont conçu des patrons décrits au moyen d'étiquettes sémantiques qu'ils utilisent pour identifier des relations entre les protéines, comme $A * R * B$ dans laquelle A et B représentent des noms de protéines et où R un verbe qui décrit une relation d'interaction entre deux protéines comme *activer*, *inhiber* ou *lier*.

Ramadier et Lafourcade (2016) utilisent des patrons pour identifier des relations sémantiques dans des comptes-rendus radiologiques en français, comme `x au niveau de y` pour la relation de localisation. Certains patrons sont ambigus, comme `x du y` qui peut signaler une relation de localisation ou d'holonymie. Ainsi, l'occurrence *lobe caudé du foie* peut s'interpréter comme une relation de localisation (r_{lieu}), i.e., le lobe codé se trouve dans le foie ou comme une relation d'holonymie (r_{holo}) : le *lobe caudé* est une partie du *foie*. Pour résoudre ce problème, Ramadier et Lafourcade (2016) proposent d'ajouter des contraintes sémantiques aux patrons. Par exemple, les contraintes en (17) permettent de distinguer les relations identifiés par le patron `x du y` :

(17) a. $x\ r_{\text{isa}}$ lieu_anatomique $\wedge$ $y\ r_{\text{isa}}$ lieu_anatomique $\Rightarrow$ $x\ r_{\text{holo}}\ y$.

b. $x\ r_{\text{isa}}$ maladie et $\wedge\ y\ r_{\text{isa}}$ lieu_anatomique $\Rightarrow x\ r_{\text{lieu}}\ y$.

Lefeuvre (2017) souhaite évaluer l'influence de différents paramètres extralinguistiques et situationnels sur la capacité des patrons lexico-syntaxiques à identifier l'hyperonymie, la méronymie, et la relation causale entre des termes des domaines de la volcanologie et du cancer du sein chez la femme. Elle a élaboré une liste de 340 patrons à partir de travaux existants, comme (Hearst, 1992), puis extrait et annoté 10 000 occurrences de ces patrons en indiquant si les termes sont bien dans la relation correspondante. Par exemple, *le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules* est une occurrence du patron *X est un Y+ caractéristiques-différentielles*) dans laquelle il existe une relation d'hyperonymie entre *cancer* et *maladie*. Lefeuvre (2017) calcule en même temps la fréquence d'apparition et la productivité des patrons dans différents sous-corpus scientifiques et de vulgarisation. Elle montre dans son analyse que le domaine a un effet sur l'apparition de certaines sous-relations et que la distribution des patrons de cause varie de manière significative en fonction du domaine. Elle note que le genre textuel (i.e., les différents registres de discours) influence la présence et la précision de certains patrons. Par exemple, dans le genre scientifique, la précision des patrons d'hyperonymie est de 63 % tandis qu'elle s'élève à 65 % dans le genre vulgarisé.

Les patrons définis manuellement ont généralement une précision élevée. En contrepartie, le rappel est relativement faible. Par ailleurs, ces approches sont difficiles à étendre et s'adaptent peu à de nouveaux domaines (Zhou *et al.*, 2008).

4.1.2 Génération automatique de patrons

Des méthodes permettant de générer automatiquement des patrons sont proposées pour augmenter le rappel des systèmes. Ces patrons peuvent être générés par des méthodes par amorçage (*bootstrap*) (Wang *et al.*, 2011) ou directement à partir de corpus (Liu *et al.*, 2011a). Les méthodes par amorçage sont semi-supervisées. Elles utilisent un petit ensemble de patrons initiaux pour identifier des couples de termes qui se trouvent dans une relation particulière. Ces termes sont ensuite repérés dans un corpus afin de collecter les configurations dans lesquelles ils apparaissent. De nouveaux patrons peuvent ainsi être abstraits pour cette relation. À leur tour, ces derniers permettent de créer de nouveaux patrons et ainsi de suite.

Agichtein et Gravano (2000) utilisent une méthode par amorçage pour extraire la relation organisation-siège social dans une grande collection de textes. À la différence des patrons lexico-syntaxiques classiques, ces patrons sont composés d'entités nommées tels que des noms de lieux (représentés par l'identifiant <LOCATION>), des noms d'organisations (<ORGANISATION>) et des noms de personnes (<PERSON>). Par exemple, un patron comme `<LOCATION>-based <ORGANISATION>` identifie les couples lieu-organisation reliés par le marqueur *-based* tels que <LOCATION> (resp. <ORGANISATION>) correspond aux entités nommées de type lieu (resp. organisation). Cette méthode permet d'obtenir des scores de précision et rappel supérieurs à ceux d'une méthode de référence fondée sur la fréquence de cooccurrence des organisations et des

lieux.

Deepika et Geetha (2021) proposent un système plus élaboré d'acquisition par amorçage de relations biomédicales. Le système distingue trois niveaux de structuration : entités, dépendances et ancres lexicales. Il apprend de nouveaux patrons en relâchant les contraintes qui portent sur les objets de l'un des niveaux. Considérons par exemple le patron amorce `Ropivacaine nsubj (inhibit)* dobj CTP2D6` où *Ropivacaine* et *CTP2D6* sont deux entités biomédicales et où `nsubj` et `dobj` sont des étiquettes de relations syntaxiques qui indiquent respectivement que la première entité est un sujet nominal et que la seconde est un objet direct. `(inhibit)*` représente un séquence textuelle incluant l'ancre lexicale *inhibit*.

Au niveau des entités, un nouveau patron peut être formé en réalisant une abstraction des entités biomédicales contenues dans un patron amorce et en les remplaçant par leurs types. Par exemple, dans le patron amorce `Ropivacaine nsubj (inhibit)* dobj CTP2D6`, *Ropivacaine* et *CTP2D6* peuvent être remplacés par les types *Drug* 'médicament' et *Gene* 'gène' pour permettre d'identifier de nouveaux couples d'entités qui apparaissent dans une structure similaire à celle du patron amorce, comme *alogliptin* 'alogliptine' et *DPP-4* dans la phrase *Alogliptin potently inhibited human DPP-4 in vitro* 'L'alogliptine inhibe fortement la DPP-4 humaine in vitro'.

Le deuxième niveau est celui des relations de dépendance. Par exemple, en relâchant la contrainte que *Drug* est un `nsubj` dans le patron amorce `Drug nsubj (inhibit)* dobj Gene`, il est possible d'extraire des patrons qui comportent d'autres relations de dépendance comme *acl :recl*. Par exemple, le patron `Ropivacaine acl:recl (inhibit)* dobj CTP2D6` peut être extrait à partir d'une phrase comme *..., aminoguanidine, which inhibits iNOS, protected against ...* '..., l'aminoguanidine, qui inhibe la iNOS, a protégé contre ...'. Il est également possible d'extraire de nouveaux patrons par abstraction de l'ancre lexicale *inhibit* 'inhiber' afin d'identifier d'autres relations. Par exemple, en relâchant la contrainte que l'ancre *inhibit* dans le patron amorce `Drug nsubj (inhibit)* dobj Gene`, une nouvelle ancre *activate* peut être repérée dans la phrase *CDDO-Me activates the epidermal_growth_factor_receptor (EGFR) related ONA repair responses.* 'Le CDDO-Me active les réponses de réparation de l'ONA liées au récepteur de croissance épidermique (EGFR).' Le système de Deepika et Geetha (2021) a permis d'extraire 37 450 patrons qui représentent différentes relations biomédicales dont 64 % correspondent à des relations présentes dans la base de données CTD [1].

Dans d'autres travaux, les patrons sont générés directement à partir des textes sans utiliser de patrons amorces. C'est notamment le cas de la méthode de Monnin et Hamon (2018) qui se servent de couples d'entités nommées pour créer des patrons à partir de phrases extraites du Web. Par exemple, pour créer des patrons capables d'identifier la relation entre le nom d'une personne et sa date de naissance, ils partent d'un couple d'entités comme *Victor Hugo* et *26 février 1802*. Les phrases contenant ces deux entités sont d'abord collectées en ligne puis filtrées

1. `https://www.cdtdb.neuroinf.jp/CDT/Top.jsp`

pour ne conserver que celles qui sont composées de moins de 20 tokens. Puis, une analyse en dépendances est effectuée sur ces phrases pour générer des patrons comportant des parties du discours, des étiquettes de dépendances et des ancres lexicales. 34 patrons lexico-syntaxiques sont ainsi générés pour la relation nom-date de naissance. Les patrons sont ensuite appliqués sur le corpus pour extraire de nouveaux couples d'entités reliées. Cette méthode a permis d'extraire 1 380 couples d'entités dont 47 % sont en relation nom-date de naissance.

En comparaison avec les systèmes fondés sur des patrons construits manuellement, le rappel des systèmes qui utilisent des patrons construits automatiquement est meilleur, mais la précision est plus faible. Le nombre de patrons générés est élevé mais un nombre important parmi eux ne sont valides que pour un faible nombre d'occurrences. La plupart des patrons créés sont trop spécifiques et ne permettent pas d'identifier les relations visées. À l'inverse, d'autres sont trop généraux et ramènent des couples qui ne sont pas dans la relation recherchée.

4.2 Modèles de sémantique distributionnelle

Les modèles sémantiques distributionnels (MSD) sont des modèles du sens des mots composés des représentations vectorielles de grande dimension (Evert, 2010). Ils reposent sur l'hypothèse que le sens d'un mot peut, dans une certaine mesure, être déduit de sa distribution dans un corpus. De nombreux travaux proposent des méthodes permettant de découvrir des relations sémantiques à partir des représentations contenues dans les MSD et d'optimiser les MSD à cette fin.

La méthode la plus simple consiste à utiliser directement la proximité des représentations dans les MSD. Elle est fondée sur l'hypothèse distributionnelle que les mots représentés par des vecteurs proches dans l'espace sémantique sont sémantiquement reliés. Bernier-Colborne et Drouin (2016) étudient différentes relations lexicales entre des termes simples du domaine de l'environnement (par exemple, la relation d'antonymie entre *humide* et *sec*) en utilisant la similarité distributionnelle entre les représentations des mots. Ils comparent des modèles fréquentiels à Word2Vec en faisant varier les hyperparamètres des modèles. Dans cette étude, les modèles fréquentiels obtiennent de meilleurs résultats que Word2Vec (les scores de MAP sont respectivement de 0,411 et de 0,382). Ils captent mieux les relations lexicales sémantiques (synonymie, hyperonymie, antonymie) tandis que Word2Vec s'avère plus efficace pour les relations de dérivation. Les auteurs soulignent que la performance des modèles varie en fonction des hyperparamètres. Par exemple, les modèles construits avec une fenêtre plus large captent mieux la relation de dérivation ; à l'inverse, les représentations des modèles construits avec une fenêtre plus petites sont mieux adaptées à l'identification de relations sémantiques comme la synonymie.

D'autres méthodes plus élaborées ont été proposées pour capter des relations sémantiques. Mikolov *et al.* (2013c) montrent que les modèles de type Word2Vec sont relativement performants dans les tâche de complétion d'analogie dans lesquelles les relations sont conçues

comme des différences entre des représentations de mots. Dans la lignée de ce travail, d'autres manières de représenter ces relations ont été proposées, notamment la compositions de plongements lexicaux et la concaténation des modèles (Levy *et al.*, 2015b). Ces méthodes reposent sur l'hypothèse qu'il existe des dimensions, ou des directions des plongements dans les espaces vectoriels qui représentent des relations particulières qui peuvent être identifiées et exploitées pour le partitionnement (*clustering*) et la classification (Vylomova *et al.*, 2016).

Plus précisément, Vylomova *et al.* (2016) testent l'hypothèse que les différences de vecteurs captent les relations lexicales du domaine général. Ils construisent à cet effet un jeu de données à partir de sources diverses. Le jeu de données est composé de relations sémantiques lexicales comme l'hyponymie et des relations flexionnelles (p. ex. le nombre et le temps). L'évaluation est effectuée dans deux contextes d'apprentissage. Les auteurs réalisent d'abord un partitionnement spectral (*spectral clustering*) pour vérifier la possibilité d'utiliser le modèle pour identifier des ensembles de couples de mots ayant une forte similarité relationnelle. Vylomova *et al.* (2016) réalisent ensuite une classification supervisée pour vérifier s'il est possible de caractériser avec précision les relations sémantiques au moyen des différences de vecteurs. Les résultats montrent que la classification supervisée à partir de différence de vecteurs permet d'acquérir une large palette de relations sémantiques lexicales. Les auteurs ont également mis en évidence le fait que le partitionnement est mieux adapté à l'acquisition de relations flexionnelles qu'à celle de relations sémantiques, avec une F-mesure de 0,98 pour les relations flexionnelles et de 0,93 pour l'hyponymie.

Dans une étude similaire, Levy *et al.* (2015b) comparent les performances de la concaténation et la différence pour classer différentes relations entre mots. Ils utilisent pour ce faire 5 jeux de données dont deux ne sont constitués que de couples en relation d'hyperonymie. Les trois autres couvrent un ensemble de relations plus variées. Leurs résultats dont ils font état présentent des différences avec ceux de Vylomova *et al.* (2016). En utilisant la même méthode, ils obtiennent un score de F-mesure de 0,665 pour les hyperonymes de BLESS [2] (Baroni et Lenci, 2011), ce qui est largement inférieur à celui rapporté par Vylomova *et al.* (2016) qui est de 0,93. La relation d'intérêt et la méthode étant les mêmes, les scores donnés par Vylomova *et al.* (2016) pourraient être spécifiques aux données qu'ils ont utilisées. Selon Levy *et al.* (2015b), les méthodes exploitant des plongements de mots isolés ne sont pas capables d'apprendre à inférer des relations lexicales mais seulement les propriétés particulières des mots.

Cette observation découle directement du fait que l'objectif premier des plongements de mots est de capter la similarité qui existe entre les mots, ce qui limite la capacité des modèles à capter certains types de relations (Bouraoui *et al.*, 2018). Il est par ailleurs difficile d'encoder à la fois (*i*) qu'un mot m_1 se trouve dans une relation particulière avec un autre mot m_2 et (*ii)* que les mots qui ont des vecteurs similaires à m_1 et m_2 ne se trouvent pas dans la même relation. Par exemple, Bouraoui *et al.* (2018) ont constaté que si les modèles Word2Vec et GloVe décrivent que *Berlin* et *Allemagne*, *Moscou* et *Russie* sont dans la même relation capitale_de, ils tendent

2. `http://clic.cimec.unitn.it/distsem`

à sur-générer des relations incorrectes comme entre *Berlin* et *Russie* car les deux capitales ont des représentations similaires.

La moyenne des vecteurs des mots peut aussi être utilisée pour créer des vecteurs de relations entre des couples de mots sémantiquement reliés (Espinosa-Anke et Schockaert, 2018). Cette méthode calcule ces vecteurs à partir des couples en relation qui se trouvent dans les mêmes phrases. Pour apprendre le vecteur relationnel entre deux mots m_1 et m_2, Camacho-Collados *et al.* (2019) font la moyenne des plongements de mots w_i qui se trouvent entre m_1 et m_2 dans un ensemble de contextes extraits du corpus (en l'occurrence, le corpus de Wikipédia en anglais). Ils pondèrent les plongements des w_i par le nombre de fois où ils apparaissent entre m_1 et m_2 dans tous les contextes où m_1 et m_2 cooccurrent dans la même phrase. L'ordre d'apparition de m_1 et m_2 n'est pas pris en compte ; leurs vecteurs de relation calculés sont de ce fait symétriques. Leur méthode a été évaluée sur le jeu de données BLESS (Baroni et Lenci, 2011) et a permis d'obtenir une F-mesure de 0,943.

Pour dépasser les limitations des méthodes fondées sur les plongements de mots isolés, des méthodes qui consistent à apprendre directement des vecteurs de relation à partir de leurs distributions sont proposées. Dans ces modèles, les relations ne sont pas décrites par des compositions de plongements de mots sémantiquement reliés. Elles sont directement représentées par des vecteurs appris à partir des contextes dans lesquels apparaissent les mots en relation.

L'idée d'apprendre un vecteur de relation remonte au moins à la méthode d'analyse relationnelle latente (LRA) de Turney (2005). La LRA est fondée sur l'hypothèse que les paires de mots qui cooccurrent dans des patrons lexico-syntaxiques similaires tendent à avoir des relations sémantiques similaires. Les vecteurs de relations peuvent ainsi être appris en factorisant une matrice dont les lignes correspondent à des couples de mots et dont les colonnes correspondent à des patrons construits à partir de phrases contenant ces couples de mots (Turney, 2005 ; Riedel *et al.*, 2013). Turney (2005) a évalué les vecteurs de relations obtenus par la LRA sur des analogies de tests SAT (Turney *et al.*, 2003) et établit qu'ils permettent de répondre correctement à 56,4 % de ces analogies.

Les récents modèles de langue neuronaux pré-entraînés qui permettent d'obtenir des plongements de mots contextuels tels que ELMo (Peters *et al.*, 2018a), XLNet (Yang *et al.*, 2019) et BERT (Devlin *et al.*, 2019) peuvent aussi être utilisés et optimisés (*finetuning*) pour déterminer les relations sémantiques qui peuvent exister entre les mots. Les tâches d'acquisition des relations sont reformulées comme une tâche de classification de textes (Hou *et al.*, 2020 ; Yao *et al.*, 2019 ; Bouraoui *et al.*, 2018). Par exemple, ces modèles peuvent prendre en entrée des phrases contenant deux mots et proposer en sortie une étiquette de relation ou l'existence d'une relation particulière entre eux.

Peng *et al.* (2019) optimisent le modèle BERT pour extraire les relations spécifiques au domaine médical comme *TeRP* qui relie un test et un problème médical révélé par le test. Le problème de l'extraction des relations est ainsi reformulé comme une tâche de classification des phrases dans lesquelles les deux termes d'intérêt sont remplacés par des étiquettes prédéfinies

(comme @Gene$, @DRUG$). Par exemple, pour connaître la relation qui existe entre *citalopram* et *SERT*, ils utilisent une phrase comme *Citalopram protected against the RTI-76-induced inhibition of SERT binding* dans laquelle ils remplacent les deux termes par les étiquettes @CHEMICAL$ et @GENE$. Après substitution, la phrase *@CHEMICAL$ protected against the RTI-76-induced inhibition of @GENE$ binding* est fournie comme entrée au modèle BERT. Le système a été évalué sur plusieurs jeux de données du domaine médical. La F-mesure la plus élevée est de 79,9 %.

Au lieu d'utiliser des phrases sélectionnées et annotées manuellement qui expriment la relation entre deux mots comme dans (Peng *et al.*, 2019), Bouraoui *et al.* (2020) proposent de sélectionner des contextes qui expriment la relation considérée en général. Par exemple, *Paris is the capital of France* est un contexte qui décrit la relation *capital of* d'une manière explicite et générale, tandis que le contexte *The Eiffel Tower is in Paris, France* ne le fait que façon implicite. Par ailleurs, nous ne pouvons pas utiliser dans cette phrase d'autres couples de mots reliés par *capital de* à la place de *Paris* et *France* comme c'est le cas dans la première phrase. Afin de sélectionner les contextes pour un couple de mots m_1 et m_2 reliés par une relation R, les auteurs utilisent le modèle BERT. Ils masquent respectivement m_1 et m_2 dans chaque contexte où m_1 et m_2 co-occurrent puis ils filtrent ces contextes en fonction de la position de m_1 (lorsque m_2 est masqué) et de m_2 (lorsque m_1 est masqué) dans les prédictions du modèle. Une fois les contextes sélectionnés, il est possible de tester l'existence de la relation R entre couple de mots M_3 et M_4 en remplaçant m_1 et m_2 par m_3 et m_4 dans ces contextes puis en réalisant une tâche de classification binaire fondée sur BERT pour prédire si la phrase résultante est une affirmation correcte ou incorrecte. Si c'est le cas, m_3 et m_4 se trouvent dans la même relation que celle entre m_1 et m_2. La méthode a été évaluée sur deux jeux de données contenant plusieurs types de relations, lexicales et morphologiques. Elle a été comparée à deux modèles de base respectivement fondée sur une classification par SVM et sur un apprentissage de la distribution gaussienne de différences de vecteurs. Le modèle proposé par Bouraoui *et al.* (2020) obtient relativement à ces derniers de meilleurs résultats pour les relations causales et encyclopédiques.

Les travaux de Peng *et al.* (2019) et de Bouraoui *et al.* (2020) permettent l'acquisition de relations entre des mots présents dans la même phrase. D'autres travaux portent sur la classification des relations entre des mots qui se trouvent dans des phrases différentes (Yu *et al.*, 2020 ; Lai et Lu, 2020 ; Christou et Tsoumakas, 2021). C'est également le cas de la méthode de Yang *et al.* (2021) qui permet de caractériser les relations entre des concepts du domaine médical en utilisant trois modèles de type Transformer : BERT, RoBERTa (Zhuang *et al.*, 2021) et XLNet (Yang *et al.*, 2019). Les auteurs définissent d'abord la distance CSD comme le nombre de frontières de phase qui se trouvent entre les deux concepts. Par exemple, si les deux concepts reliés se trouvent dans la même phrase, CSD vaut 0 ; CSD vaut 1 si les deux concepts sont dans des phrases consécutives. Yang *et al.* (2021) considèrent les couples de concepts dont le CSD est compris entre 0 et 4. Dans un second temps, ils optimisent des modèles BERT pour les couples de concepts qui ont le même CSD. Les entrées des modèles optimisés sont les

deux phrases qui contiennent les deux concepts reliés séparées par le token spécial *[SEP]*. Le début et la fin des séquences qui correspondent aux concepts sont signalés par deux couples de marqueurs [S1], [E1] et [S2], [E2]. Si les deux concepts se trouvent dans la même phrase, les deux phrases d'entrée sont identiques mais avec des marqueurs différents. L'évaluation de la méthode est effectuée sur deux jeux de données. La F-mesure de la méthode est de 0,961. Cependant, cette méthode employant une classification binaire, elle ne peut pas être généralisée à des tâches plus complexes dans lesquelles plusieurs types de relations doivent être acquises simultanément.

4.3 Projection sémantique

La projection sémantique est une méthode qui permet d'inférer des relations entre termes et notamment entre TMM (Morin et Jacquemin, 1999 ; Hamon et Nazarenko, 2001). Dans ce cas, la projection sémantique s'appuie sur l'hypothèse que le sens des TMM est compositionnel. Concrètement, elle tire profit de la présence d'un invariant syntaxique entre les TMM des couples en relation.

Les premiers travaux dans lesquels la projection sémantique est utilisée pour détecter la synonymie entre TMM candidats sont probablement ceux de Hamon *et al.* (1998). Les synonymes potentiels sont extraits d'un corpus spécialisé du domaine des centrales électriques. Chaque TMM candidat (p. ex., *matériel électrique*) est analysé en une tête (*matériel*) et une expansion (*électrique*). Le sens des TMM candidats étant supposé compositionnel, le remplacement de la tête, de l'expansion ou des deux par leurs synonymes fournis par un dictionnaire général préservera la synonymie. Pour le TMM *matériel électrique*, la substitution de la tête *matériel* par son synonyme *équipement* produit le TMM candidat *équipement électrique*. 396 liens ont ainsi été inférés et annotés par un expert. 37 % d'entre eux ont été validés comme des liens de synonymie corrects. La plupart des erreurs sont dues au manque d'informations contextuelles pour les mots polysémiques et aux données bruitées présentes dans le dictionnaire général (certaines relations ou sens d'un mot présents dans le dictionnaire général n'existent pas dans le domaine spécialisé). Hamon et Nazarenko (2001) appliquent la même méthode à trois corpus spécialisés issus des domaines biomédicaux et énergétiques pour extraire des TMM candidats toujours en relation de synonymie. Ils montrent dans cette seconde étude que les relations fournies par le dictionnaire de langue générale sont complémentaires de celles issues de ressources lexicales spécialisés construites manuellement. Les premières permettent d'obtenir un meilleur rappel, et les secondes une meilleure précision.

Dans une étude similaire, Morin et Jacquemin (1999) appliquent des règles d'inférence pour identifier la relation d'hyperonymie entre TMM du domaine agroalimentaire. Ils s'intéressent eux aussi à l'influence des différentes sources de relations d'amorce entre mots ou termes simples sur les résultats de la projection. L'étude de Morin et Jacquemin (1999) compare les amorces constituées de termes simples extraits du corpus AGRO en utilisant différents patrons

lexico-syntaxiques appris sur ce même corpus à celles fournies par le thésaurus AGROVOC [3]. Les règles d'inférence de Morin et Jacquemin (1999) sont différentes que celles de Hamon et Nazarenko (2001). Les mots lexicaux non sémantiquement reliés dans les deux candidats TMM ne sont pas forcément identiques. Ils peuvent être morphologiquement reliés, comme *isotope* et *isotopique* dans le couple *contenu en isotope* et *teneur isotopique*. Les résultats montrent que la projection effectuée avec les relations d'amorce extraites du corpus permettent d'obtenir une précision de 58,4 % légèrement meilleure que celle obtenue en utilisant les relations d'amorces fournies par le thésaurus (57,8 %).

Les relations d'amorces peuvent aussi être inférées à partir de MSD : deux mots sont reliés sémantiquement s'ils partagent les mêmes contextes lexicaux (Hazem et Daille, 2015 ; 2018). Hazem et Daille (2015) proposent une méthode semi-compositionnelle non supervisée qui croise analyse compositionnelle et analyse distributionnelle pour identifier des synonymes de TMM dans les domaines de l'éolien et du cancer du sein. Cette méthode se distingue de celle de Hamon et Nazarenko (2001) sur deux points. Les synonymes ne sont pas déterminés en utilisant des ressources existantes. Hazem et Daille (2015) utilisent une méthode distributionnelle pour identifier les mots sémantiquement reliés aux constituants du TMM. Par exemple, pour trouver des synonymes de *énergie renouvelable*, les auteurs sélectionnent les mots dont les vecteurs sont proches de ceux de *énergie* et de *renouvelable*. Dans un deuxième temps, ils construisent des expressions par projection sémantique puis les filtrent au moyen d'un corpus monolingue spécialisé. Notons que Hazem et Daille (2015) étendent les règles d'inférence de Hamon et Nazarenko (2001) à des TMM de longueur quelconque. L'étude a été menée sur des corpus multilingues. Leurs données de test sont composées de 54 couples de termes en français, 36 couples de termes en anglais et 26 couples de termes en espagnol. Le meilleur résultat a une MAP de 40,4 %. Il est obtenu en adoptant une approche semi-compositionnelle sur le corpus en français du cancer du sein. Il est par ailleurs très supérieur à celui de la méthode de base fondée sur les relations entre TS fournies par un dictionnaire de langue générale (4,92 %).

Hazem et Daille (2018) reprennent et adaptent la méthode semi-compositonnelle en utilisant des plongements de mots à la place des représentations vectorielles fréquentielles. Ils proposent deux méthodes supplémentaires qui exploitent les plongements de mots : l'une utilise la propriété additive des plongements pour calculer des représentations vectorielles des TMM ; l'autre considère un TMM comme un token unique et génère sa représentation de la même manière pour les mots simples. Hazem et Daille (2018) calculent une similarité entre les vecteurs qu'ils utilisent pour prédire la relation entre TMM. Les nouvelles méthodes impliquant les plongements de mots sont comparées à la méthode semi-compositionnelle (Hazem et Daille, 2015) et celle de Hamon et Nazarenko (2001). Les données de test sont composées de 54 TMM en français et 36 TMM en anglais des domaines de l'éolien et du cancer du sein. Le meilleur résultat est obtenu en utilisant la méthode semi-compositionnelle impliquant les plongements calculés à partir du corpus français du domaine de l'éolien avec un score de MAP de 34,9 %.

3. https://agrovoc.fao.org/browse/agrovoc/en/

Il est supérieur aux résultats obtenus en adoptant la méthode semi-compositionnelle (Hazem et Daille, 2015) (31,4 %). En comparaison, la méthode de Hamon et Nazarenko (2001) obtient une MAP de 0,25 %.

La projection sémantique a été utilisée pour créer le jeu de données DataProj que nous présentons dans le chapitre 6. Elle nous a permis de prédire les relations sémantiques entre TMM à partir de relations entre TS fournis par un dictionnaire spécialisé. Par ailleurs, notre étude ayant pour objet les relations sémantiques entre TMM, nous n'avons pas considéré les relations morphologiques ni flexionnelles ni dérivationnelles comme l'ont notamment fait Morin et Jacquemin (1999) (cf. Section 7.1). De plus, les règles d'inférence que nous avons utilisées ne comportent pas de contraintes, ni sur l'ordre des constituants, ni sur les patrons syntaxiques des TMM candidats. Une autre spécificité du jeu de données DataProj et l'utilisation de contextes lors de la validation des relations inférées afin de prendre en compte l'éventuelle ambiguïté des TMM et la façon dont les contextes déterminent leurs sens. Parmi les études présentées dans cette section, seule celle de Hamon et Nazarenko (2001) s'appuie sur des contextes pour la validation des relations.

4.4 Substitution lexicale

Une autre méthode permettant l'acquisition de relations sémantiques est la substitution lexicale qui est souvent mise en œuvre au moyen de MSD. L'objectif de la substitution lexicale est de proposer des substituts d'un mot cible dans un contexte donné. Par exemple, dans une phrase comme *selon les cas, la voiture est présentée comme propre, bio ou encore vert*, *voiture* peut être remplacé par son quasi-synonyme *véhicule* sans que cela ne change le sens de la phrase. D'autres candidats, notamment des hyponymes comme *berline*, des co-hyponymes comme *camion* ou des hyperonymes comme *transport*, peuvent aussi être substitués à *voiture*. La substitution entraîne dans ce cas une modification légère du sens à la phrase. La tâche de substitution lexicale peut être réalisée au moyen de modèles de langue masqués (MLM). Ces modèles sont entraînés pour prédire les tokens qui peuvent être substitués à un token spécial <MASK> dans une phrase donnée. Ils considèrent pour cela les contextes gauche et droit du token masqué. Les substituts proposés par les MLM sont généralement sémantiquement proches de la cible, i.e., du token remplacé par <MASK>.

Schick et Schütze (2020) utilisent la substitution lexicale au moyen du modèle BERT pour évaluer l'influence de la fréquence des mots-clés sur la capacité du modèle à capter différentes propriétés sémantiques sans optimisation spécifique à la tâche. L'étude porte sur l'antonymie, l'hyperonymie et la co-hyponymie dans le domaine général. Pour chaque relation, les auteurs extraient des triplets de WordNet de la forme $< k, r, T >$ où k est un mot-clé, c'est-à-dire une entrée de WordNet ; où r est une relation ; où T est un ensemble de mots cibles reliés à k par la relation r et appartenant au vocabulaire du modèle. Schick et Schütze (2020) se servent également de patrons définis manuellement qui peuvent exprimer la relation entre le mot-clé k (<W>) et

un mot cible (_), comme <W> `n'est pas _` pour l'antonymie. Par ailleurs, ils ont adapté la méthode *Attentive Mimicking* (Schick et Schütze, 2019) pour construire des plongements de mots pour les mots rares. Les conclusions de leur étude sont que BERT est capable de capter les relations sémantiques et que la précision de la méthode varie en fonction de la fréquence des mots-clés : elle est moins performante pour les mots-clés de fréquence faible [4]. Avec la méthode *Attentive Mimicking* et le modèle BERT-large, les scores MRR sont de 0,529 pour l'antonymie, de 0,299 pour l'hyperonymie et de 0,227 pour la co-hyponymie.

La substitution lexicale fondée sur les MLM est également utilisée par Arefyev *et al.* (2020) pour évaluer sa capacité à capturer des relations sémantiques. Ces auteurs proposent deux méthodes qui permettent de prédire les substituts en prenant en compte le sens du mot masqué. L'une ajoute un estimateur de la proximité entre le mot masqué et les substituts. L'autre utilise des patrons dynamiques tels que `T and then <mask>` 'T et ensuite <mask>' où T représente le mot masqué. Par exemple, pour retrouver les substituts de *climate* 'climat' dans le contexte *the effects of climate change on humans* 'les effets du changement climatique sur l'homme', ils génèrent un contexte *the effects of climate and then <mask> change on humans* qu'ils fournissent en entrée à BERT. Arefyev *et al.* (2020) observent que la plupart des substituts ramenés sont des synonymes et des co-hyponymes lorsque le mot masqué est un nom.

Une étude connexe proposée par Ferret (2021) porte sur les propriétés sémantiques des MLM. L'étude est réalisée en utilisant la substitution lexicale inversée qui estime l'adéquation des candidats substituts à un contexte du mot cible. Plus précisément, soit m_1 un mot dont on recherche les substituts ; soit m_2 un substitut candidat ; soit $Cont_{m_1}$ un contexte de m_1 ; soit V_{m_1} la représentation vectorielle de m_1 dans $Cont_{m_1}$ et V_{m_2} celle de m_2 dans le même contexte. Les représentation peuvent être construites par un modèle BERT ou ELMo. La substitution lexicale inversée estime la possibilité de substituer m_2 à m_1 dans le contexte $Cont_{m_1}$ par la similarité de V_{m_1} et V_{m_2}. La méthode a été testée dans le domaine général pour l'anglais. Par exemple, étant donné un terme *disaster* 'désastre' et un de ses contextes *Since the 1946 disaster there have been 15 tsunamis in the Pacific...* 'Depuis le désastre de 1946, il y a eu 15 tsunamis dans le Pacifique...'la substitution lexicale inversée permet de savoir dans quelle mesure le modèle est davantage orienté vers la synonymie ou vers l'hyperonymie en remplaçant *disaster* par son synonyme *catastrophe* 'catastrophe' et par son hyperonyme *misfortune* 'malheur'. Une fois générées les représentations de *disaster*, *catastrophe* et *misfortune* en contexte ; il est possible de comparer la similarité des représentations de *disaster* et de *catastrophe* et celle des représentations de *disaster* et de *misfortune* pour connaître la préférence du modèle. Les expériences menées montrent que les modèles contextuels de type BERT favorisent l'acquisition de la synonymie et de la co-hyponymie.

Nous utilisons la substitution lexicale dans les expériences que nous présentons dans la

4. Une adaptation de la méthode Schick et Schütze (2020) à l'acquisition de relations entre TMM en français du domaine de l'environnement serait pénalisée par le même problème. En effet, le modèle CamemBERT étant entraîné avec des textes du domaine général, les termes des domaines spécialisés y sont sous-représentés.

suite de la thèse. Cependant, nos objectifs diffèrent dans ceux des travaux que nous venons de présenter. (*i*) Nous cherchons à identifier des relations lexicales entre TMM en français dans le domaine de l'environnement, alors que ces travaux portent sur des relations entre mots simples en anglais dans le domaine général. (*ii*) Nous utilisons des contextes extraits de corpus plutôt que des patrons pouvant exprimer les relations sémantiques lexicales comme le font (Schick et Schütze, 2020). (*iii*) Nous utilisons une stratégie de conditionnement permettant d'apporter au modèle des informations supplémentaires sur le mot masqué, mais elle est différente de celle de Arefyev *et al.* (2020) (cf. Section 7.1.1). (*iv*) Notre étude porte sur toutes les relations lexicales classiques, à l'inverse de celle de Schick et Schütze (2020) qui ignore la synonymie et de celle de Arefyev *et al.* (2020) qui ignore l'antonymie.

4.5 Analogie

La recherche actuelle associant l'analogie et les plongements de mots se concentre sur « l'analogie proportionnelle » du type $a : b :: c : d$, c'est-à-dire telle que « a est à b comme c est à d ». Le point de départ de cette recherche est l'étude de Mikolov *et al.* (2013c) dans laquelle ils montrent que les modèles Word2Vec captent une variété de relations syntaxiques et sémantiques entre les mots comme entre les formes au singulier et au pluriel des noms communs ; la relation cause-effet ; la taxonomie ; etc. et que l'identification de ces relations peut être estimée par la différence entre leurs vecteurs sémantiques : $V_a - V_b \approx V_c - V_d$. Dans la lignée de Mikolov *et al.* (2013c), de nombreuses études ont porté sur la capacité de l'analogie à capter diverses relations lexicales, encyclopédiques ou relevant de de domaines spécialisés (Pennington *et al.*, 2014 ; Köper *et al.*, 2015 ; Wohlgenannt *et al.*, 2019).

Gladkova *et al.* (2016) évaluent la possibilité de détecter au moyen de l'analogie différents types de relations linguistiques (flexionnelles, dérivationnelles, lexicales sémantiques, encyclopédiques) entre des mots simples. L'évaluation est faite sur les jeux de données BATS (Gladkova *et al.*, 2016) et *Google analogy* (Mikolov *et al.*, 2013a) en calculant la précision moyenne. Leurs résultats montrent que, parmi les quatre types de relations, les relations lexicales sont les plus difficiles à capter avec une précision inférieure à 10 %. Ce résultat est compatible avec les conclusions de Köper *et al.* (2015).

Chen *et al.* (2018) extraient des relations sémantiques entre termes médicaux au moyen de l'analogie. Leur jeu de données se compose de 33 000 questions analogiques et de six relations médicales, telles que *peut traiter* qui peut relier un médicament comme *atropine* et une maladie comme *uvéite*. Leurs résultats montrent que les performances de l'analogie varient en fonction de la relation et du modèle. La précision qu'ils ont obtenue à TOP 5 est inférieure à 20 %.

Alors que la plupart des études se concentrent sur l'analogie entre mots ou termes simples, Chaudhri *et al.* (2022) s'intéressent à la résolution d'équations analogiques entre termes simples et multi-mots dans le domaine biologique en anglais. Leur jeu de données est composé de quadruplets de termes tels que *Carbon 14 atom :Radioactivity : :C4 plant :C4 photosynthesis*.

Les relations qu'ils considèrent sont spécifiques au domaine biologique comme *a type of* 'un type de'. Chaudhri *et al.* (2022) utilisent un modèle Seq2seq [5] (Sutskever *et al.*, 2014) pour résoudre des équations analogiques de type $a : b :: c :?$. Le modèle Seq2Seq comporte trois encodeurs, qui modélisent respectivement les entrées a, b et c. Les auteurs ont également utilisé des modèles Seq2Vec [6]. Ils ont constaté que le modèle Seq2Vec ELMo fournit la meilleure précision à Top 1 (0,507).

L'étude de Paullada *et al.* (2020) porte elle aussi sur l'analogie entre termes simples et multi-mots dans le domaine biomédical. L'objectif est l'acquisition de relations spécifiques au domaine, telle que gène-maladie. Afin de générer des plongements des termes multi-mots, ces derniers sont indexés comme des unités lexicales uniques construites en concaténant leurs composants. Par exemple, *forêt primaire* est indexé comme forêt_primaire. Les auteurs ont créé des plongements à partir d'un corpus de phrases extraites de la littérature biomédicale et annotées en dépendances syntaxiques. Pour évaluer les méthodes, les auteurs définissent un score 1 − rang normalisé pour tous les termes inconnus dans les équations d'analogie. Comparés à un modèle skip-gram (SGNS), les plongements créés par les auteurs à partir du corpus syntaxiquement analysé améliorent effectivement la résolution des équations analogiques biomédicales sur la plupart des jeux de données. Par exemple, sur les données PGKB (Whirl-Carrillo *et al.*, 2012), leur méthode obtient un score de 0,969 tandis que le modèle SGNS n'obtient que 0,705.

Notre étude diffère de celles que nous venons de présenter sur plusieurs points. Comme nous l'avons déjà souligné, nous travaillons pour l'identification de relations terminologiques entre TMM en français dans le domaine de l'environnement. Comme Paullada *et al.* (2020), nous utilisons le décalage vectoriel au lieu de modèles Seq2Vec et Seq2Seq comme le font Chaudhri *et al.* (2022). Cependant, le modèle que nous utilisons est différent de celui de Paullada *et al.* (2020) qui encodent les TMM sous forme de termes simples. Nous utilisons un modèle FastText (Bojanowski *et al.*, 2017) où les TMM et ses composants sont représentés dans le même espace vectoriel. De plus, nous nous intéressons à des relations sémantiques classiques entre TMM et non à des relations spécifiques à un domaine particulier.

4.6 Synthèse

Nous avons présenté dans ce chapitre un état de l'art sur les méthodes d'extraction de relations sémantiques dans le domaine général et dans différents domaines spécialisés.

Nous avons d'abord décrit des méthodes fondées sur des patrons lexico-syntaxiques qui caractérisent explicitement les relations visées. Les patrons peuvent être définis manuellement en analysant des phrases contenant la relation d'intérêt, ce qui permet d'obtenir une précision élevée mais un rappel faible. Ils peuvent aussi être générés automatiquement par amorçage à

5. Un modèle Seq2Seq est un modèle qui prend en entrée une séquence (de mots, de lettres, etc.) et qui produit une autre séquence en sortie.

6. Des modèles dont l'entrée est une séquence et la sortie est un vecteur unique.

partir d'un corpus, ce qui augmente le rappel par rapport aux patrons définis manuellement. En contrepartie, la précision baisse. Par ailleurs, Lefeuvre (2017) a mis en évidence que le domaine et le genre textuel ont un effet sur les performances de l'acquisition des relations sémantiques.

Nous avons ensuite présenté des méthodes qui utilisent les MSD, la projection sémantique, la substitution lexicale et l'analogie. Ces dernières peuvent être calculées à partir de plongements de mots, éventuellement optimisés à cette fin.

Dans le cadre de la thèse, nous explorons la capacité des modèles distributionnels statiques et contextuels à capter les relations lexico-sémantiques entre TMM en utilisant deux méthodes. La première, inspirée de (Schick et Schütze, 2020), est une méthode par substitution lexicale fondée sur un modèle MLM. La seconde inspirée de (Paullada *et al.*, 2020) capte les relations lexico-sémantiques entre TMM au moyen de l'analogie. Nous avons par ailleurs construit au moyen de la projection sémantique un jeu de données utilisé dans le chapitre 8.

Deuxième partie

Jeu de données de relations sémantiques entre TMM du domaine de l'environnement

Chapitre 5

Création d'un jeu de donnés à partir des TMM de IATE

Notre étude s'appuie sur des données composées de couples de TMM sémantiquement reliés du domaine de l'environnement. Cependant, peu de ressources terminologiques fournissent des relations entre termes, et moins encore entre TMM. C'est le cas pour la synonymie, la relation qui est la plus représentée dans les produits terminologiques. L'une des ressources terminologiques en ligne qui contiennent des termes du domaine de l'environnement est IATE (que nous avons déjà présenté en Section 2.2.3). Elle fournit un nombre important de couples de TMM synonymes pouvant servir de jeu de données pour nos expériences.

Dans ce chapitre, nous présentons les ressources et le processus utilisés pour créer ce jeu de données. Rappelons que notre travail porte sur les bitermes nominaux comme *changement du climat*.

5.1 Le corpus PANACEA

Pour construire le jeu de données extrait de IATE nous avons utilisé le corpus PANACEA (code catalogue ELRA-W0065) [1] construit dans le cadre du projet PANACEA [2]. Il sera désigné comme « corpus PANACEA » dans les parties suivantes. Il s'agit d'un corpus monolingue français du domaine de l'environnement composé d'environ 50 millions de mots. Il est distribué gratuitement pour la recherche par l'Agence ELDA.

Ce corpus contient 23 514 documents détectés automatiquement comme étant en langue française et comme relevant du domaine de l'environnement par le Focused Monolingual Crawler (FMC) qui a été développé dans le cadre du projet. Ces documents sont collectés en ligne à partir d'encyclopédies, de blogs, de sites gouvernementaux et de sites d'organisations non gouvernementales. Les documents ont des niveaux de spécialisation variables et appartiennent à une variété de genres. Le corpus est plus hétérogène qu'un corpus spécialisé typique (composé

1. http://catalog.elra.info/en-us/repository/browse/ELRA-W0065/
2. http://www.panacea-lr.eu/en/info-for-researchers/data-sets/monolingual-corpora

uniquement de textes spécialisés), ce qui pourrait s'expliquer par la nature hétérogène du domaine de l'environnement (Bernier-Colborne, 2016).

Nous avons appliqué des opérations de prétraitement au corpus avant de les utiliser. Ces opérations courantes comprennent l'extraction de texte à partir des documents XML, la normalisation des caractères et la lemmatisation. Toutes ces opérations sont réalisées sur les documents comprenant plus de 50 mots au moyen d'un programme Python adapté de celui mis à disposition par Bernier-Colborne (2016).

Le corpus est composé de documents au format TEI [3]. Nous avons extrait le contenu textuel des documents XML. Le programme localise d'abord les éléments XML qui contiennent un contenu textuel dans chaque document. Chaque paragraphe qui se trouve entre des balises <p> est caractérisé par des attributs qui indiquent : (a) si le paragraphe est dans une langue autre que français ; (b) si le paragraphe est considéré comme trop court ou s'il s'agit d'un paragraphe « passe-partout » (boilerplate) comme *Un article de Wikipédia, l'encyclopédie libre*. Seuls les paragraphes en français qui ne sont ni trop courts, ni passe-partout sont extraits. Le titre du document est également conservé lorsqu'il est présent. Le corpus étant construit automatiquement, tous les paragraphes ne sont pas constitués de phrases complètes. Nous y trouvons par exemple une phrase comme *sont utilisés des barrages flottants , des dispersants et des moyens humains de ramassage : pelleteuses , pelles et seaux ..*

Après avoir extrait le contenu textuel de fichiers en format XML, nous avons normalisé les caractères. Le corpus PANACEA est encodé en UTF-8, un codage qui utilise entre 1 et 4 octets pour représenter tous les caractères. L'encodage UTF-8 offre une couverture beaucoup plus complète que Latin-1. Cependant, certains caractères peuvent être représentés en UTF-8 de plusieurs façons. Par exemple, la lettre *é* peut être représentée en utilisant soit le code U+00E9, soit la combinaison des codes de la lettre *e* (U+0065) et celui de l'accent aigu (U+0301). Cette variabilité pose problème lorsque nous construisons des modèles distributionnels car certaines occurrences d'une forme linguistique sont ignorées car elles sont codées de manière différente. Le processus de normalisation des caractères comprend les étapes suivantes :

1. Tous les mots du corpus sont transformés en minuscules.
2. Les caractères *œ* (majuscules et minuscules) sont remplacées par les chaînes de caractères « oe » ;
3. Les caractères accentués seront représentés en utilisant le code unique. Nous avons utilisé la bibliothèque Unicodedata3 [4] comme référence ;
4. Les caractères absents du codage Latin-1 sont supprimés, p. ex., emoji.

Le corpus résultant est en UTF-8.

La dernière étape est la lemmatisation. Pour des raisons grammaticales, les documents peuvent utiliser différentes formes d'un même terme, comme *protection des forêts* et *protection*

3. La Text Encoding Initiative (TEI) est un format de balisage du domaine de humanités numériques qui propose des guides pour la création et la gestion de tout type de données sous forme numérique (Ide et Véronis, 1995).

4. `https://docs.python.org/2/library/unicodedata.html`

de la forêt. L'objectif de la lemmatisation est de ramener les différentes formes flexionnelles d'un mot ou d'une lemme à une forme conventionnelle. Par exemple, elle remplace *protection des forêts* et *protection de la forêt* par *protection de le forêt*. Nous avons réalisé cette opération parce que nous ne cherchons pas à détecter des relations entre les formes fléchies des TMM. De plus, certains TMM sont rares dans le corpus. La lemmatisation permet d'avoir une seule représentation vectorielle dans les modèles FastText pour une unité lexicale quelle que soit sa flexion.

La lemmatisation du corpus est réalisée en utilisant TreeTagger (Schmid, 2013). Nous conservons la forme lemmatisée des mots lorsque TreeTagger en propose une ou à défaut la forme originale du mot.

Nous avons également utilisé Treetagger pour segmenter les paragraphes en phrases. Treetagger sépare les phrases par l'étiquette *SENT*. Signalons que certaines phrases sont répétées en double dans les documents. Elles ne sont pas filtrées pendant le pré-traitement.

5.2 Extraction des couples de TMM synonymiques de IATE

Nous avons d'abord extrait à partir de IATE 20 154 TMM nominaux composés de deux mots lexicaux du domaine de l'environnement (comme *séparateur magnétique*). Nous avons utilisé Treetagger pour identifier le patron syntaxique des TMM. Le tableau 11 montre un extrait des bitermes nominaux extraits de IATE. Pour chaque biterme, son code et son domaine sont aussi extraits. Comme le domaine de l'environnement est hétérogène, des termes de sous-domaines comme l'énergie, le changement climatique, le transport, sont aussi pris en compte.

Puis, nous avons constitué des couples de TMM en utilisant leur code. Dans IATE, les termes synonymes ont le même code. Par exemple, *déferreur magnétique* et *séparateur magnétique* ont le même code *1427234*. Ensuite, nous avons filtré les couples de TMM pour ne conserver que ceux qui sont présents dans le corpus PANACEA. Seuls les couples contenant deux TMM dont les formes lemmatisées sont présentes dans le corpus PANACEA lemmatisé sont conservés. À l'issue du filtrage, 1 064 paires de TMM sont conservées, comme *reconstitution des forêts* et *restauration des forêts*. Le tableau 12 présente des exemples de couples de TMM extraits de IATE.

Nous avons noté que certains TMM qui ont le même code ne sont pas vraiment synonymes. Certains couples sont composés de TMM dont l'un est la forme plurielle de l'autre (*pollution des eaux* et *pollution d'eau*). D'autres couples contiennent des TMM dont l'un est un dérivé synonymique (*réchauffement planétaire* et *réchauffement de la planète*) ou une variante de l'autre (*dynamique de la population* et *dynamique des populations*). Par ailleurs, certains couples contiennent des TMM qui ne sont pas des bitermes, comme *déplacement domicile-travail*.

Afin de conserver uniquement les TMM synonymes, nous avons validé manuellement les relations entre les couples de TMM. Nous avons aussi remarqué que certains bitermes sont plutôt reliés par une relation hiérarchique, même si dans certains cas, un hyperonyme peut aussi

Code	Domaine	TMM
35145	LAW ; FINANCE ; ENVIRONMENT	vérification des comptes
35145	LAW ; FINANCE ; ENVIRONMENT	contrôle des comptes
35666	ENVIRONMENT	méthode d'analyse
35666	ENVIRONMENT	méthode analytique
43594	pharmaceutical industry ; ENVIRONMENT	conditionnement primaire
43624	ENVIRONMENT	délégation de pouvoir
43711	ENVIRONMENT	rougeur de la face
43766	ENVIRONMENT	acide muriatique
43766	ENVIRONMENT	acide chlorhydrique
43766	ENVIRONMENT	acide chlorydrique
43774	ENVIRONMENT	amélioration de l'image
43774	ENVIRONMENT	accentuation d'image
43774	ENVIRONMENT	amélioration d'image
44445	ENVIRONMENT	déchets agrochimiques
44467	ENVIRONMENT	sciure de bois
44485	ENVIRONMENT	résidus de pelanage
44490	ENVIRONMENT	boues sans chrome
44509	ENVIRONMENT	boues de dessalage
44514	ENVIRONMENT	goudrons acides
44557	ENVIRONMENT	oxydes métalliques
44569	ENVIRONMENT	scories phosphoriques
44616	ENVIRONMENT	encre séchée
44636	ENVIRONMENT	bains de fixation
44662	ENVIRONMENT	autres boues
44668	ENVIRONMENT	poussières d'alumine
44670	ENVIRONMENT	vieilles brasques
44680	ENVIRONMENT	arséniate de calcium
44695	ENVIRONMENT	verre usagé
44695	ENVIRONMENT	déchet de verre
44695	ENVIRONMENT	déchets de verre
44698	ENVIRONMENT	moules déclassés
44712	ENVIRONMENT	boues de phosphatation
44720	ENVIRONMENT	autres déchets
44734	ENVIRONMENT	boues d'usinage
44736	ENVIRONMENT	déchets de soudure

TABLE 11 – Exemples de bitermes extraits de IATE

être considéré comme un synonyme (Polguère, 2016). Nous avons donc gardé ces couples dans le jeu de données. Le résultat de la validation est décrit dans le tableau 13. Parmi les 1 064 couples de TMM, 928 sont composés de bitermes synonymes. Cet ensemble de couples est appelé DataIATE dans la suite de la thèse.

Nous avons observé que plus de 90 % des couples dans DataIATE partagent un mot lexical comme *analyse du risque* et *étude de risque*. Cependant, les couples dans DataIATE n'ont pas les

TMM1	TMM2
action pilote	projet pilote
activité humaine	activité anthropique
activité terrestre	occupation terrestre
activité verte	emploi vert
affectation des sols	utilisation des sols
affectation des sols	utilisation des terres
affectation des terres	affectation des sols
affectation des terres	utilisation des sols
affectation des terres	utilisation des terres
aide financière	aide au financement
ail des ours	ail des bois
aire de répartition	aire de distribution
altération de l'environnement	changement écologique
amélioration des terres	aménagement foncier
amélioration foncière	amélioration des terres
aménagement de l'espace	aménagement du territoire
aménagement du territoire	amélioration des terres
aménagement du territoire	aménagement foncier
aménagement foncier	développement du territoire
aménagement hydraulique	gestion de l'eau
aménagement hydraulique	gestion des eaux
aménagement spatial	aménagement de l'espace
aménagement spatial	aménagement du territoire
ami de l'environnement	hygiéniste du milieu
amplitude de la température	amplitude thermique
analyse de sûreté	analyse de sécurité
analyse des déchets	étude déchets
analyse du risque	étude de risques
analyses biologiques	test biologiques
animal d'élevage	animal de ferme
arbre à pain	fruit à pain

TABLE 12 – Exemples de couples de bitermes nominaux présents dans le corpus PANACEA extraits de IATE

	Synonymique	Pluriel	Dérivation synonymique	Variante	Ternaire	Total
Nbr	928	10	83	37	6	1064

TABLE 13 – Résultats de la validation sur les relations entre les couples de TMM extraits de IATE

mêmes caractéristiques linguistiques et statistiques. Par exemple, les TMM d'un couple peuvent ne pas avoir la même structure syntaxique comme *réchauffement climatique* et *réchauffement de la planète*. Pour sélectionner les données utilisées pour les expériences, nous avons effectué des analyses supplémentaires sur les couples de TMM de DataIATE. Elles sont résumées dans le tableau 14. DataIATE comprend 563 paires de termes binaires, qui partagent un mot lexical

et qui ont la même structure syntaxique, comme *environnement de travail* et *milieu de travail*. Désignons ce sous-ensemble comme DataIATE_MLM. 599 couples de TMM dans DataIATE sont composées de deux termes binaires ayant une fréquence supérieure à 5 dans le corpus PANACEA. Nous désignons ce sous-ensemble comme DataIATE_FastText.

Caractéristiques	**Dénomination**	**Nbr**
Couples de TMM	DataIATE	928
Couples de TMM qui ont le même patron et qui partagent un élément lexical	DataIATE_MLM	563
Couples de TMM de fréquence supérieure à 5 dans PANACEA	DataIATE_FastText	599

TABLE 14 – Caractéristiques des couples de TMM synonymes extraits de IATE

5.3 Synthèse

Dans ce chapitre, nous avons présenté la construction du jeu de données composé de TMM synonymes du domaine de l'environnement extraits de IATE, que nous appelons DataIATA. Pour le constituer, nous avons extrait des 23 514 documents constituant le corpus monolingue français PANACEA, 20 154 termes binaires nominaux présents dans IATE. Nous avons pu construire à partir de ces 20 154 termes, 928 couples de termes binaires nominaux reliés par la synonymie qui constituent le jeu de données DataIATE. Les occurrences de termes de DataIATE dans PANACEA permettent de mettre en œuvre la méthode de substitution lexicale qui exploite les contextes des TMM. De même, ces contextes servent aussi à construire les plongements des TMM pour la méthode par analogie.

Chapitre 6

Création par projection sémantique d'un jeu de données de TMM sémantiquement reliés

Les relations lexicales entre TMM ne se limitent pas à la synonymie. Elles incluent également l'antonymie (comme *climat sec* : *climat humide*), l'hyperonymie (comme *stockage de combustible* : *stockage de gaz*) et les relations contrastives (comme *préservation de la flore* : *préservation de la faune*) qui sont toutes des relations terminologiques essentielles (L'Homme, 2020). Cependant, aucune ressource terminologique existante ne fournit ces relations entre les TMM du domaine de l'environnement. Nous avons donc créé par projection sémantique un jeu de données pour ces relations. La projection sémantique est une méthode fondée sur la compositionnalité qui permet d'étendre aux TMM des relations sémantiques entre TS. Rappelons que nous avons observé que plus de 90 % des couples de TMM synonymes extraits de IATE partagent un mot lexical. La projection sémantique semble donc bien adaptée à l'identification des autres relations lexicales si l'on admet que cette caractéristique est aussi partagée par les couples de TMM antonymes et par ceux qui sont dans des relations hiérarchiques.

Nous présentons dans ce chapitre les ressources que nous avons utilisées et les étapes de la construction de ce jeu de données, composé de couples de TMM reliés par différentes relations lexicales.

6.1 Extraction des candidats termes

La première étape de la création du jeu de données consiste à extraire des candidats bitermes à partir de corpus PANACEA. Le candidat terme est une notion utilisée en extraction terminologique. Elle désigne les unités lexicales identifiées par un système automatiquement (Drouin et Langlais, 2006). Elle est liée à la notion appelée *termhood* en anglais, qui désigne un ensemble de critères pouvant être mis en œuvre pour filtrer une liste de mots et de syntagmes candidats.

Ces critères peuvent être fondés sur des mesures statistiques, sur des indices morphosyntaxiques ou contextuels ou une combinaison de ces critères.

Dans la suite de la section, nous présentons dans un premier temps le fonctionnement et l'application de l'extracteur TermSuite. Nous décrivons ensuite les candidats extraits par TermSuite à partir de corpus PANACEA.

6.1.1 TermSuite, un outil d'extraction terminologique

Pour extraire les candidats TMM, nous avons utilisé TermSuite (Cram et Daille, 2016). TermSuite est une boîte à outils développée dans le laboratoire LS2N [1]. Il traite 7 langues : anglais, français, allemand, espagnol, letton, chinois, russe. La présentation que nous en faisons se limite à sa fonction de reconnaissance de termes.

Comme les autres outils d'extraction terminologique, TermSuite procède classiquement en deux temps (Cram et Daille, 2016). Il collecte dans un premier temps des candidats termes à partir du corpus au moyen des opérations suivantes : tokenisation, étiquetage morphosyntaxique, lemmatisation, racinisation, segmentation en phrases. Puis, il filtre les candidats qui pourraient ne pas être des termes et les classe selon une mesure de *termhood*. TermSuite possède en outre un composant supplémentaire qui permet d'identifier les variantes de termes afin de détecter davantage de termes et de rendre le classement des candidats termes plus précis (Daille et Blancafort, 2013).

Afin de reconnaître les candidats termes simples, la spécificité du terme dans le corpus par rapport à un corpus de langue générale (des journaux compilés fournis par CLEF 2004 (Peters, 2005)) est calculée. L'identification des candidats TMM est effectuée à l'aide de spécifications linguistiques exprimées dans UIMA Tokens Regex (Ferrucci et Lally, 2004). UIMA Tokens Regex est un outil permettant de définir des patrons syntaxiques sous forme d'expressions régulières et d'identifier des TMM et leurs variantes en corpus. La plupart des patrons syntaxiques de TMM encodés dans TermSuite sont ceux de syntagmes nominaux (L'Homme, 2004).

TermSuite fournit un ensemble de caractéristiques des candidats termes et des relations qui existent entre eux. Elles sont décrites dans le guide d'utilisation [2]. En voici quelques exemples :

— Propriétés des candidats termes :

pilot : forme la plus courante du terme.

pattern : patron du terme (concaténation des étiquettes morphosyntaxiques de ses composants).

spec : spécificité du terme. C'est une mesure statistique qui indique si les occurrences d'un terme paraissent en surnombre dans un corpus spécialisé par rapport à un corpus général.

freq : nombre d'occurrences du terme dans le corpus.

1. https://www.ls2n.fr
2. http://termsuite.github.io/documentation/gui/

— Propriétés des relations entre candidats termes :

DerivationType : type de dérivation de la variation, lorsque la relation est une variation.

VariationRank : rang de la variation parmi toutes les variations d'un même candidat terme, lorsque la relation est une variation.

IsSyntagmatic : la relation est-elle une variation syntagmatique ?

L'entrée de TermSuite est un corpus fourni sous la forme d'une collection de fichiers ***.txt** dans un répertoire. La Figure 19 montre la hiérarchie (non exhaustive) du répertoire. Les fichiers à partir desquels TermSuite extrait des candidats termes sont placés dans le répertoire dont le nom indique la langue des textes dans les fichiers.

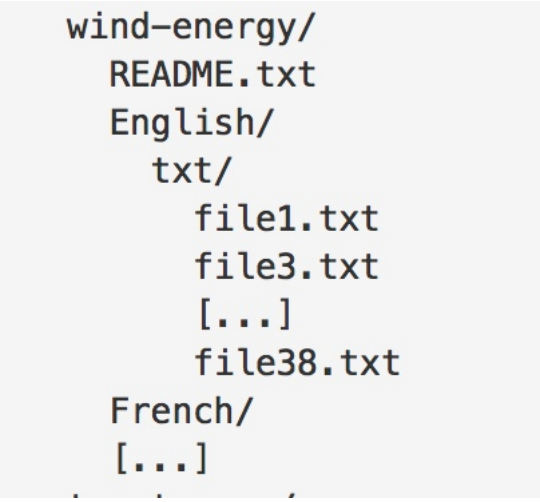


FIGURE 19 – Répertoire qui contient le corpus donné en entrée à TermSuite

Source: `http://termsuite.github.io/getting-started/prepare-corpus`

La sortie de TermSuite est un fichier au format TSV contenant des informations linguistiques et statistiques sur les candidats termes. La Figure 20 présente un extrait des sorties. La première colonne représente le rang des termes. Les variantes d'un terme ont le même rang que le terme. La deuxième colonne indique que le candidat extrait est un terme (*T*) ou que c'est une variante du terme (*V*). La lettre entre crochets après *V* indique le type de variante (syntagmatique, morphologique, graphique, sémantique, etc.). La colonne *dFreq* représente le nombre de documents du corpus dans lesquels le candidat terme est présent. Par exemple, *wind turbine rotor* 'rotor d'éolienne' est une variante syntagmatique du terme *wind turbine* 'éolienne' classé au rang 2. Il apparaît 31 fois dans le corpus et a un score de spécificité de 3,38. Il est présent dans 12 documents du corpus.

6.1.2 Lancement de TermSuite

Les candidats termes utilisés pour la projection sémantique sont identifiés par leur partie du discours, leur spécificité et leur fréquence. La Figure 21 présente la ligne de commande

#	type	pattern	pilot	spec	freq	dFreq		
1	T	N	rotor	4,82	848	30		
2	T	N N	wind turbine	4,56	1855	37		
2	V[s]	N N N	wind turbine rotor		3,38	31		12
2	V[s]	A N N	offshore wind turbine		3,26	47		7
2	V[s]	N N N	wind turbine noise		3,53	43		3
2	V[s]+	N N N	wind turbine technology		3,34	28		10
2	V[s]	N N N	wind turbine system		3,40	32		7
2	V[s]	A N N	modern wind turbines		2,82	17		7
2	V[s]	N N N	wind turbine tower		3,07	15		9
2	V[s]	A N N	large wind turbines		3,12	17		10
2	V[s]	N N N	wind turbine power		2,89	10		6
3	T	N N	wind energy	4,51	414	32		
3	V[s]	N N N	wind energy potential		3,07	15		5
3	V[s]	A N N	offshore wind energy		3,56	47		7
3	V[s]	N N N	wind energy development		3,29	25		5
4	T	N N	wind power	4,34	278	26		
4	V[s]	N N N	wind turbine power		2,89	10		6
4	V[s]	N N N	Wind Power Plant		3,76	74		9
4	V[s]	A N N	offshore wind power		3,01	13		4
5	T	N	airfoil 4,26	236	8			

FIGURE 20 – Exemple de sortie de TermSuite

Source: http://termsuite.github.io/documentation/terminology-tsv-output/

utilisée pour extraire les candidats termes et leurs propriétés. Les paramètres **-t** et **-c** introduisent respectivement l'outil d'étiquetage morphosyntaxique (en l'occurrence, le TreeTagger) et le répertoire du corpus. Les paramètres **-l** et **–tsv** spécifient respectivement la langue du corpus d'entrée et le format du fichier de sortie. Le paramètre **–tsv-properties** indique les propriétés des termes à inclure dans les sorties.

```
java -Xmx25g -cp '/home/yizhewang/Bureau/termsuite/termsuite-core-3.0.10.jar'
fr.univnantes.termsuite.tools.TerminologyExtractorCLI
-t '/home/yizhewang/Bureau/termsuite/TreeTagger'
-c '/home/yizhewang/Bureau/termsuite/panacea'
-l fr
--tsv panacea.tsv
--tsv-properties "pattern,pilot,freq,spec"
--info
```

FIGURE 21 – Ligne de commande pour l'extraction de termes

Le Tableau 15 présente un sous-ensemble des candidats termes extraits par TermSuite à partir du corpus PANACEA. Par exemple, *zone intertropicale humide* est une variante syntagmatique du candidat terme *zones humides* classé au rang 6. Il apparaît 6 fois dans le corpus et a un score de spécificité de 0,97.

#	type	patron	pilote	freq	spec
6	T	N A	zones humides	6 320	3,95
6	V[s]	N A A	zone intertropicale humide	6	0,97
6	V[s]	N P N A	conservation des zones humides	42	1,78
9	T	N A	parc national	10 196	3,85
16	T	N	polluants	7 883	3,74
18	T	N A	eaux souterraines	3 802	3,73
18	V[v]	N A C A	eaux superficielles et souterraines	87	2,09
226	T	N P N A	adaptation au changement climatique	1 914	3,13
70	T	N N	mission agrobiosciences	1 904	3,42

TABLE 15 – Extrait de sortie de TermSuite

6.1.3 Analyse des candidats termes extraits

Nous avons extrait 624 482 candidats du corpus PANACEA, parmi lesquels 519 749 candidats termes et 104 733 variantes. Par exemple, le candidat terme *pollution des fleuves* a deux variantes : *pollution des fleuves et des nappes* et *pollution des fleuves côtiers*.

Les candidats TMM extraits par TermSuite sont tous des syntagmes nominaux (Cram et Daille, 2016). Parmi les 519 749 candidats termes, 280 183 sont composés de deux mots lexicaux, comme *fluides froides*. Le Tableau 16 présente les patrons syntaxiques des candidats bitermes extraits.

Patron syntaxique	**Nombre**
N A	100 179
N P N	158 926
N N	21 078

TABLE 16 – Patrons syntaxiques des candidats termes binaires extraits par TermSuite

La notion de candidat met en évidence la possibilité d'erreurs dans les résultats des extracteurs automatiques et la nécessité de valider ces résultats avant de les utiliser. Par exemple, certains candidats termes extraits ne sont pas syntaxiquement complets, comme *azote issu*. Dans le contexte (18), l'expression complète est en réalité *azote issu de l'elorn*.

(18) même la rénovation de la step de la zip , qui a considérablement réduit les rejets azotes urbains à partir de 2006 et a donc augmenté la part de l' azote issu de l' elorn dans les ulves ...

D'autres candidats ne sont pas des termes du domaine de l'environnement, comme *ministère autrichien* qui serait plutôt une entité nommée du domaine général. Ces problèmes sont inhérents aux extracteurs terminologiques qui malgré la sophistication de leurs traitements incorporent toujours des candidats termes mal formés ou hors domaine.

6.2 Relations de référence

Pour construire le jeu de données de couples de TMM par projection sémantique, il est nécessaire de disposer d'une référence, en l'occurrence les relations entre TS RefDiCoEnviro. RefDiCoEnviro est constitué de 830 couples de termes simples (noms et adjectifs) reliés par des relations sémantiques lexicales. Ces couples sont extraits de DiCoEnviro (L'Homme et Lanneville, 2014), un dictionnaire spécialisé du domaine de l'environnement que nous avons présenté en section 2.2.4.

Les relations entre TS dans RefDiCoEnviro sont classées dans les trois catégories suivantes (Drouin et Langlais, 2006) :

1. **QSYN**
   - synonymie (*diesel* & *gazole*)
   - quasi-synonymie (*conserver* & *protéger*)
   - sens proche (*boisement* & *plantation*)
   - variante (*autopartage* → *auto-partage*)
2. **HYP**
   - hyponymie (*autoroute* → *route*)
   - hyperonymie (*combustible* → *pétrole*)
3. **ANTI**
   - antonymie (*accélérer* & *ralentir*)
   - sens contrastif (*flore* & *faune*)

Le Tableau 17 présente un extrait des couples de TS des RefDiCoEnviro : la colonne **Entrée** donne le terme d'entrée dans DiCoEnviro ; la colonne **Termes reliés** contient les termes sémantiquement reliés au terme d'entrée ; les colonnes **POS1** et **POS2** fournissent respectivement la catégorie grammaticale du terme d'entrée et celle de son terme sémantiquement relié ; la colonne **Relation** indique la relation qui relie les deux termes.

Entrée	POS1	Termes reliés	POS2	Relation
neige	NN	glace	NN	QSYN
mondial	JJ	planétaire	JJ	QSYN
infrastructure	NN	arrêt	NN	HYP
rail	NN	infrastructure	NN	HYP
abiotique	JJ	biotique	JJ	ANTI
boisement	NN	déboisement	NN	ANTI

TABLE 17 – Extrait des enregistrements des RefDiCoEnviro

La distribution des trois types de relations dans RefDiCoEnviro est déséquilibrée, comme nous pouvons le voir dans le Tableau 18 : RefDiCoEnviro contient 116 couples d'ANTI composés de 107 termes ; 191 couples d'HYP composés de 122 termes ; 523 couples de QSYN composés de 415 termes.

	ANTI	HYP	QSYN	Total
Nombre de couples	116	191	523	830
Nombre de termes	107	122	415	644

TABLE 18 – Distribution des relations sémantiques dans RefDiCoEnviro

Les chiffres du Tableau18 ne tiennent pas compte des verbes et des types de relations présentes dans DiCoEnviro car notre étude porte uniquement sur les TMM nominaux. Par ailleurs, 63 % des couples de termes simples dans RefDiCoEnviro ont été utilisés pour générer des couples de candidats TMM.

6.3 Mise en œuvre de la projection sémantique

Afin d'étendre les relations entre TS aux candidats TMM, nous avons appliqué les règles de projection sémantique aux couples de TS dans RefDiCoEnviro et aux candidats bitermes nominaux extraits de corpus PANACEA par TermSuite. Cette méthode est basée sur l'hypothèse que le sens des TMM est compositionnel, ce qui signifie que si nous remplaçons une partie d'un TMM par un mot ou une séquence de mots ayant le même sens, le sens du TMM ne change pas (Pagin et Westerståhl, 2010). Par exemple, étant donné un TMM *croissance de la population*, si nous remplaçons *croissance* par son quasi-synonyme *accroissement*, l'expression *accroissement de la population* devrait avoir le même sens que *croissance de la population*.

Dans cette section, nous décrivons en détail les différentes étapes réalisées lors de la projection, puis nous présentons les résultats obtenus et analysons les caractéristiques des couples de candidats produits.

6.3.1 Prédire des relations entre TMM

La Figure 22 illustre les différentes étapes réalisées pour construire le jeu de données.

TermSuite a extrait 280 183 candidats bitermes tels que *climat humide* et *réchauffement du climat*. Pour réaliser la projection sémantique, nous n'avons gardé que les candidats présents au moins 5 fois dans le corpus PANACEA (selon la fréquence calculée par TermSuite), soit 129 383 candidats. Ce filtrage a été réalisé pour permettre l'utilisation de ces candidats dans les expériences fondées sur des modèles distributionnels. Les modèles distributionnels sont en effet sensibles à la fréquence des unités indexées.

Nous avons ensuite appliqué les règles d'inférence de Hamon et Nazarenko (2001) sur les couples de TS de RefDiCoEnviro sans aucune restriction sur l'ordre des constituants ni sur les patrons syntaxiques des candidats TMM.

Plus formellement, soit t_1 et t_2 deux TMM tels que $\mathbf{voc}(t_1) = \{u_1, v_1\}$ et $\mathbf{voc}(t_2) = \{u_2, v_2\}$ où $\mathbf{voc}(x)$ représente l'ensemble des mots lexicaux qui composent x. Si u_1 et u_2 sont deux TS reliés par une relation R et si $v_1 = v_2$, alors t_1 et t_2 sont aussi reliés par R. En d'autres termes,

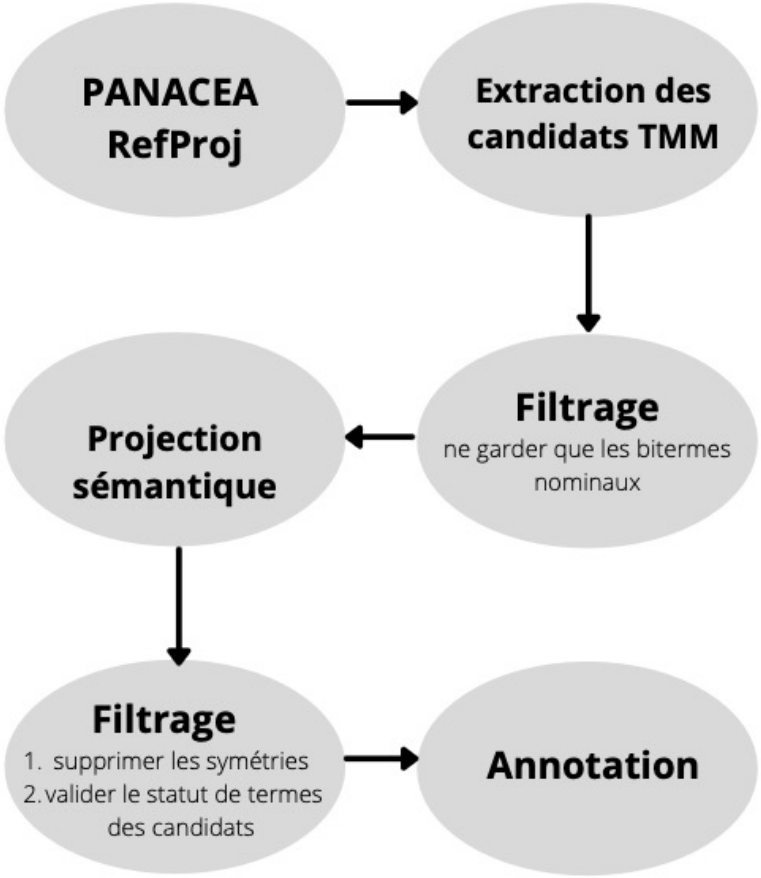


FIGURE 22 – Étapes de la construction du jeu de données de TMM reliés sémantiquement

si $\mathcal{M}$ est l'ensemble des TMM d'un domaine et $\mathcal{S}$ est l'ensemble des TS, la règle de projection peut être formulé comme suit :

$$\begin{aligned} &\forall t_1 \in \mathcal{M}, \forall t_2 \in \mathcal{M} \\ \mathbf{voc}(t_1) = \{u_1, v_1\} \wedge \mathbf{voc}(t_2) &= \{u_2, v_2\} \\ \wedge\, u_1 \in \mathcal{S} \wedge u_2 \in \mathcal{S} \wedge v_1 &= v_2, \\ [\exists R, R(u_1, u_2) &\Rightarrow R(t_1, t_2)] \end{aligned}$$

Considérons par exemple les deux candidats TMM *climat sec* et *climat humide* ; *climat sec* est composé de deux mots lexicaux *climat* et *sec* ; *climat humide* est composé de *climat* et *humide*. Puisque *sec* et *humide* sont des antonymes et que l'autre élément lexical dans les deux TMM est identique (i.e. *climat*), nous pouvons considérer *climat sec* et *climat humide* comme étant eux aussi des antonymes.

6.3.2 Résultats de la projection sémantique

7 874 couples de candidats bitermes (composés de 5 986 candidats bitermes) ont été générés, parmi lesquels 1 503 couples reliés par ANTI ; 1 059 couples sont reliés par HYP ; 5 312 couples sont reliés par QSYN. Le Tableau 19 présente le nombre de couples de candidats bitermes générés par la projection initiale pour chaque catégorie de relation.

	ANTI	HYP	QSYN
Nombre de couples	1 503	1 059	5 312

TABLE 19 – Résultats de la projection initiale

Dans RefDiCoEnviro, un couple de termes simples est composé d'une entrée de DiCoEnviro et d'un terme qui lui est sémantiquement relié. Le terme sémantiquement relié pouvant aussi être une entrée dans le dictionnaire, certaines relations sont symétriques, comme le couple *régional & local* et le couple *local & régional*. Plusieurs couples de candidats bitermes symétriques sont ainsi générés (comme *climat régional & climat local* et *climat local & climat régional*). Dans tous les cas de symétrie, nous avons supprimé l'un des deux couples. Nous avons également supprimé l'un des couples dans les cas où deux relations hiérarchiques sont symétriques. Par ailleurs, la direction des relations hiérarchiques n'a pas été prise en compte lors de la création du jeu de données.

Le Tableau 20 présente le nombre de couples de candidats bitermes conservés après la suppression des relations symétriques : les 58 couples de TS ANTI génèrent 912 couples de candidats bitermes ; les 63 couples de TS HYP produisent 720 couples de candidats bitermes ; les 234 couples de TS QSYN engendrent 3 259 couples de candidats bitermes. Le nombre de couples de TMM QSYN est largement supérieur à celui des autres deux types de relations car le nombre de couples de TS QSYN est plus élevé.

	ANTI	HYP	QSYN
Nombre de couples de TMM candidats	912	720	3 259
Nombre de couples de TS	58	63	234

TABLE 20 – Résultats de la projection sémantique après filtrage des relations symétriques

Le Tableau 21 présente la distribution des patrons syntaxiques des couples de candidats bitermes générés. Plus de 90 % des couples générés sont composés de deux candidats ayant les mêmes patrons : le plus souvent deux *NA* (comme *activité biologique* et *agriculture biologique*) ou deux *NPN* (comme *teneur en azote* et *teneur en gaz*).

	NA-NA	NA-NN	NA-NPN	NN-NN	NN-NPN	NPN-NPN
Nombre	1073	15	61	45	217	3480
Pourcentage	21,94 %	0,30 %	1,25 %	0,92 %	4,44 %	66,63 %

TABLE 21 – Distribution des patrons des couples de candidats TMM

Bien que tous les candidats TMM extraits soient nominaux, nous observons que certains patrons syntaxiques sont relativement plus fréquents pour certaines relations. Nous savons en effet que les antonymes sont principalement des adjectifs et des verbes, tandis que les hyponymes et les hyperonymes sont principalement des noms (Polguère, 2016). La Figure 23 présente la distribution des couples de candidats TMM relativement à leurs patrons syntaxiques et à leur type de relation. La proportion correspond au nombre de couples de TMM ayant ces patrons divisé par le nombre total de couples dans la relation. La figure 23 montre que pour les trois types de relations, plus de la moitié des couples de candidats TMM ont des patrons **NPN-NPN**. Cependant, la proportion de couples dont les patrons sont **NPN-NPN** est relativement faible pour ANTI en comparaison avec celles des relations HYP et QSYN. À l'inverse, les couples de candidats qui ont des patrons **NA-NA** sont plus nombreux pour ANTI, ce qui suggère que dans ces couples, l'antonymie s'établit principalement entre les adjectifs.

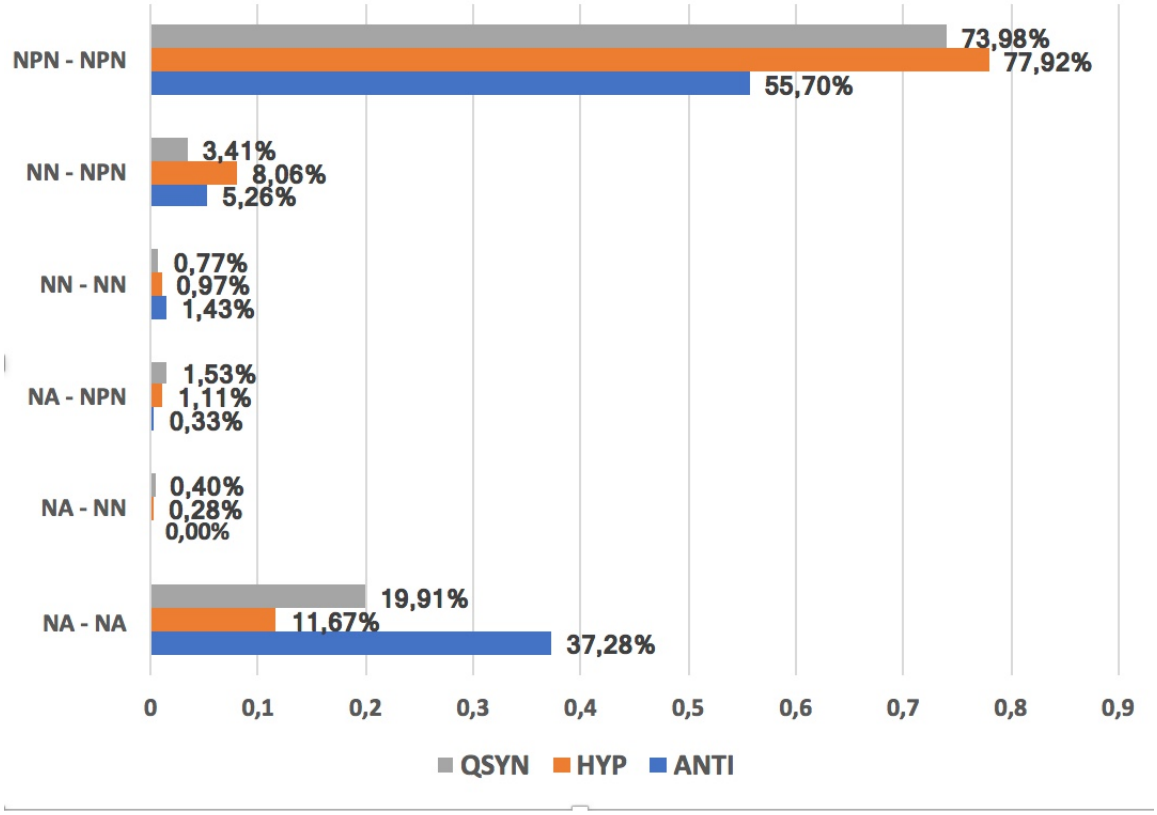


FIGURE 23 – Distribution des couples de candidats TMM relativement à leurs patrons syntaxiques et à leur type de relation

6.4 Annotation du statut de terme des candidats TMM et de la relation inférée

Une partie des candidats TMM présentés en Section 6.1.3 ne sont pas des termes du domaine de l'environnement. Par exemple, *lutte contre le changement* n'est pas un terme car il est syntaxiquement incomplet ; en contexte, le terme complet est *lutte contre le changement climatique*. Il existe aussi des candidats comme *cadre régional* qui n'appartiennent pas au domaine de l'environnement. De plus, certaines relations inférées sont incorrectes. Par exemple, la relation HYP qui existe entre *terre* et *planète* est projetée sur le couple de TMM *climat de la terre* et *climat de la planète*. Cependant, nous observons qu'en contexte ces TMM dans une relation QSYN. Une validation du statut de terme des candidats et de la relation inférée est donc nécessaire.

6.4.1 Validation du statut de terme des candidats TMM

Afin d'éviter les éventuels désaccords entre annotateurs nous avons utilisé trois ressources terminologiques externes pour vérifier le statut de terme des candidats : TERMIUM Plus [3], Le Grand Dictionnaire (GDT) [4] et IATE (Interactive Terminology for Europe) [5]. Ces trois banques sont construites manuellement par des terminologues. De plus, les choix terminologiques sont encadrés et guidés par des politiques éditoriales documentées. Ces bases de données sont donc assez fiables.

Nous considérons de ce fait tout candidat présent dans l'une de ces ressources comme un terme du domaine de l'environnement puisqu'il a été extrait d'un corpus spécialisé de ce domaine. La mise en correspondance des données d'IATE avec les candidats a été traitée automatiquement par programme. À l'issue de la validation fondée sur IATE, 100 couples QSYN composés de candidats valides ont été conservés, mais seulement 10 couples ANTI et 8 couples HYP. Afin d'augmenter le nombre de couples de candidats valides pour HYP et ANTI, nous avons sélectionné aléatoirement 500 couples HYP et 500 couples ANTI et vérifié la présence des TMM dans TERMIUM Plus et GDT en réalisant des requêtes manuelles, ces deux ressources terminologiques ne permettant pas le téléchargement de leurs fichiers de données.

De nombreux candidats extraits sont très spécifiques, tels que *conservation des papillons*. De ce fait, seule une fraction des couples constitués deux candidats TMM sont présents dans au moins une des ressources. Le Tableau 22 présente le nombre de couples composés de candidats valides pour chaque catégorie de relation. Les données pour chaque type de relation sont déséquilibrées avec 80 couples ANTI, 51 couples HYP et 100 couples de QSYN. De plus, certains couples de TS sont plus productifs que les autres. Par exemple, trois couples de TMM sont engendrés par *agricole* et *piscicole*, quatre par *augmentation* et *diminution*, mais seulement un par *décharge* et *recharge*.

3. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra
4. http://www.granddictionnaire.com/
5. https://iate.europa.eu/

ANTI	HYP	QSYN	total
80	51	100	231

TABLE 22 – Données dont les TMM ont été validés en utilisant les banques terminologiques

6.4.2 Évaluation des relations inférées

Une fois établi le statut de terme des TMM des couples présentés en Section 6.4.1, une annotation manuelle de la préservation de la relation inférée a été effectuée par trois annotateurs.

Critères d'annotation

L'annotation de la préservation de la relation sémantique est binaire. Nous considérons que la relation est préservée lorsque la relation qui s'établit entre les deux TS existe aussi entre les deux TMM qui contiennent les deux TS. Pour réaliser cette tâche, les annotateurs se sont appuyés sur leurs connaissances et sur le sens contextuel des TMM.

Nous avons donc extrait aléatoirement du corpus PANACEA 5 contextes pour chaque TMM. Si la relation inférée existe entre les sens d'au moins une occurrence de chacun des TMM du couple, nous considérons que la relation inférée est préservée. Par exemple, *froid* et *chaud* sont dans une relation ANTI. Pour savoir si *temps froid* et *temps chaud* qui contiennent *froid* et *chaud* préservent la relation ANTI, nous examinons les sens de *temps froid* et *temps chaud* dans les contextes extraits. Les sens des *temps froid* et *temps chaud* dans les contextes (19a) et (19b) étant opposés, nous considérons que la relation ANTI est préservée. À l'inverse, la relation HYP entre *diversité* et *biodiversité* n'est pas préservée pour les couples *gestion de la diversité* et *gestion de la biodiversité* parce que les deux TMM ont un sens similaire dans tous les contextes extraits comme illustré en (20).

(19) a. par **temps froid**, cette technique consiste à ne pas laisser tourner son moteur au ralenti plus de 30 secondes .

b. par **temps chaud**, le compromis entre confort et pratique est difficile à trouver .

(20) a. les variétés paysannes, issues de millénaires de **gestion de la diversité** par les agriculteurs sont trop vivantes pour se plier aux critères d'inscription .

b. elle même distincte de l'utilisation (par les agriculteurs) des semences, la **gestion de la biodiversité** cultivée réunit dans un processus continu ...

Les contextes jouent un rôle déterminant dans notre annotation. Ils permettent de mieux comprendre le sens des TMM, certains TMM étant plus techniques que d'autres ou ayant des sens particuliers lorsqu'ils sont utilisés dans un domaine particulier. Par exemple, *zone de recharge* fait référence à un aquifère libre caractérisé par une infiltration rapide où l'eau s'accumule. Comme ce terme n'est pas utilisé couramment dans la vie quotidienne, il est difficile de l'associer aux aquifères sans connaissances préalables ou sans l'aide du contexte.

Les contextes servent ainsi à mettre en évidence le sens fin des TMM, comme pour *changement du climat*. Ce TMM signifie 'variabilité climatique' dans le contexte (21b) et 'réchauffement climatique' dans le contexte (21a). Dans certains contextes, un TMM peut se retrouver dans une conjonction avec une expression qui exprime l'un de ses sens soit parce que l'énoncé est redondant, l'information étant exprimée plusieurs fois, soit parce que le sens du TMM est différent. Par exemple, dans le contexte (21c), *réchauffement de la planète* et *changement du climat* sont réunis par une conjonction. Nous considérons dans ce cas que les sens des deux TMM sont différents : *changement du climat* signifie ici 'variabilité du climat ou refroidissement du climat' plutôt que 'réchauffement du climat'.

(21) a. il a établi que le **changement du climat** était « sans équivoque » et que les émissions de gaz à effet de serre provenant des activités humaines étaient responsables (avec 90 % de certitude) de l'augmentation des températures depuis cent ans .

b. à quelle vitesse la réduction des concentrations atmosphériques de GES de courte durée entraînerait un **changement du climat** .

c. par exemple , des mesures adoptées en vue de lutter contre le réchauffement de la planète et le **changement du climat** – peuvent uniquement être évaluées avec le recul.

Les contextes peuvent également aider à identifier la relation entre les TMM. Par exemple, sans contextes, nous pouvons considérer que *environnement global* est un hyperonyme d'*environnement régional*. Cependant, dans les contextes (22a) et (22b) leurs sens sont plus contrastifs que hiérarchiques. Le premier exemple décrit un phénomène naturel local qui n'est pas mondial, tandis que le second concerne l'environnement mondial.

(22) a. sel, de sable et de fines poussières qui causent un énorme préjudice à l'**environnement régional** et à la santé des populations ...

b. manière suffisamment forte et audible un sentiment d'urgence face à l'évolution de notre **environnement global** - partagé par la plupart des scientifiques - , tout en gardant un discours ...

Problèmes rencontrés lors de l'annotation

Les annotateurs ont rencontré deux problèmes. L'un est lié au fait que les couples TS utilisés pour générer les couples de candidats TMM ne sont pas suffisamment diversifiés, en particulier pour la catégorie HYP. Le Tableau 23 donne le nombre de couples de TS contenus dans les couples de TMM pour chaque type de relation. Ce problème est susceptible d'avoir un impact sur les résultats de l'annotation et de rendre les données moins informatives. Par exemple, 5 couples de TMM ont été générés à partir du couple de TS *terre* et *planète*. Dans DiCoEnviro, *terre* est un hyponyme de *planète*. Cependant, dans les contextes des TMM contenant *planète*, ce terme a pour référence la Terre. De ce fait, aucun des cinq couples de TMM ne préserve la relation HYP.

	ANTI	**HYP**	**QSYN**	**total**
Nombre de couples de TS	34	19	62	115
Nombre de couples de TMM	80	51	100	231

TABLE 23 – Nombre de couples de TS contenus dans les couples de TMM pour chaque type de relation

Le second problème concerne les cas fortuits : certains candidats TMM ne sont syntaxiquement complets dans aucun de leurs contextes. Ils y ont toujours un complément. Ce complément qualifie souvent une partie du TMM plutôt que le TMM dans son entier et change ainsi le sens du TMM. Par exemple, *séquestration des gaz* est un TMM qui existe dans les ressources terminologiques en ligne. Dans ce TMM, le sens de *gaz* est 'corps gazeux utilisé comme combustible ou comme carburant'. Cependant, dans tous les contextes extraits, *séquestration des gaz* est suivi du complément *à effet de serre*. Ce complément qualifie le mot *gaz* et lui donne un sens particulier : celui de corps qui se trouve à l'état gazeux à température et pression ordinaires. Ainsi, le sens en contexte de *séquestration des gaz* varie. Par ailleurs, le candidat n'est pas complet. Les cas fortuits sont susceptibles d'affecter la qualité des modèles distributionnels créés pour les expériences présentées dans la troisième partie de la thèse. Selon l'hypothèse distributionnelle, un TMM est caractérisé par les mots avec lesquels il cooccurre. Néanmoins, si un candidat TMM n'est pas syntaxiquement complet, certains mots de son contexte ne peuvent pas être utilisés pour le caractériser distributionnellement.

Pour décider si un candidat TMM est complet ou ne l'est pas, nous vérifions s'il existe dans le corpus un contexte dans lequel le candidat est syntaxiquement complet. Si aucun n'est trouvé, la relation est annotée comme non préservée. Dans le cas contraire, l'annotation est réalisée comme nous l'avons décrit ci-dessus. Par exemple, bien que *puits de gaz* apparaisse avec un complément dans les cinq contextes extraits, il existe un contexte (23) dans le corpus dans lequel le candidat TMM n'a pas de complément. Nous continuons alors à vérifier la préservation de la relation entre *puits de carbone* et *puits de gaz*.

(23) nous avons trouvé une corrélation assez forte entre la proximité d' un **puits de gaz** et la concentration de méthane dans l' eau potable » , a déclaré robert (vrai gaz naturel) .

La situation est similaire pour le candidat *conservation de la nature* (24) qui apparaît comme une partie d'une entité nommée dans les cinq contextes extraits. Dans ce cas, toutes les occurrences du candidat dans le corpus sont vérifiées. Si toutes sont une partie de l'entité nommée, l'annotation de la préservation de la relation est négative.

(24) primates les plus menacés dans la liste rouge de l'union internationale pour la conservation de la nature ...

6.4.3 Résultats d'inter-annotation

À l'issue de l'annotation, l'accord inter-annotateur a été calculé en utilisant le kappa de Fleiss (1971). Le kappa K varie de 0 à 1. Landis et Koch (1977) proposent le Tableau 24 pour l'interprétation de ses valeurs.

K	**Interprétation**
< 0	Pauvre concordance
0,01 - 0,20	Faible concordance
0,21 - 0,40	Légère concordance
0,41 - 0,60	Concordance moyenne
0,61 - 0,80	Concordance importante
0,81 - 1,00	Concordance presque parfaite

TABLE 24 – Interprétation de valeurs du kappa de Fleiss

Le Tableau 25 présente la valeur K de l'annotation du jeu de données pour chaque type de relation. L'accord est plus fort pour la relation ANTI que pour les deux autres types de relation. L'accord inter-annotateurs sur l'ensemble de données est assez fort puisqu'il est de 0,69. Une phase d'adjudication a ensuite été réalisée pour créer le jeu de données. Les tableaux 26 et 27

	ANTI	**HYP**	**QSYN**	**total**
K	0,77	0,68	0,61	0,69

TABLE 25 – Score de kappa de Fleiss

présentent des exemples des couples valides et non-valides.

Terme	**Terme relié**	**Relation**	**Annotation**
conservation des espèces	protection des espèces	QSYN	1
contamination bactériologique	pollution bactériologique	QSYN	1
enfouissement des déchets	élimination des déchets	QSYN	1
réservoir de carbone	puits de carbone	HYP	1
gaz combustible	gaz de pétrole	HYP	1
écologie terrestre	écologie aquatique	ANTI	1
environnement global	environnement régional	ANTI	1
courant alternatif	courant continu	ANTI	1

TABLE 26 – Exemples des couples valides

Les 231 couples de TMM dont le statut de terme a été validé (cf. Section 6.4.1) permettent de construire 231 quadruplets formés d'un couple de TS et d'un couple de TMM qui les contiennent. La relation sémantique entre TS est préservée entre TMM dans 180 de ces quadruplets. Dans ce qui suit, nous appelons DataProj le sous-ensemble de quadruplets où la relation est préservée. DataProj se compose de 69 couples ANTI, 26 couples HYP et 85 couples QSYN. Plus de trois quart des couples de TMM préservent la relation inférée, ce qui suggère que la plupart des TMM

Terme	**Terme relié**	**Relation**	**Annotation**
dégradation écologique	catastrophe écologique	QSYN	0
recyclage des matériaux	matériaux de récupération	QSYN	0
habitat rural	milieu rural	QSYN	0
agriculture itinérante	culture itinérante	HYP	0
climat de la planète	climat de la terre	HYP	0
combustible de remplacement	remplacement du pétrole	HYP	0
eau atmosphérique	air atmosphérique	ANTI	0
eau libre	air libre	ANTI	0
vitesse de croissance	diminution de la vitesse	ANTI	0

TABLE 27 – Exemples des couples non-valides

ont un sens compositionnel (ou un sens faiblement compositionnel). Cette observation confirme l'affirmation de L'Homme (2004). Bien qu'aucune restriction sur les patrons syntaxiques n'ait été imposée lors de la génération des couples de TMM, nous observons que tous les couples valides sont composés de TMM ayant le même patron syntaxique et que les TS qu'ils contiennent apparaissent tous dans les mêmes positions.

51 couples de TMM ne préservent pas la relation inférée. Ils se répartissent en trois groupes. (*i*) Les couples de TMM contenant deux TS sémantiquement reliés qui ne se trouvent pas à la même position dans les TMM, comme *eau de surface* et *surface de la terre*. Ici, *eau* et *terre* sont reliés par ANTI mais les TMM ne le sont pas, car *eau* est à la 1re position dans *eau de surface* tandis que *terre* est situé à la dernière position dans *surface de la terre*. (*ii*) Le sens du TS n'est pas préservé dans le TMM qui le contient comme *route* et *route maritime*. Dans cet exemple, *route* est un hyperonyme d'*autoroute* quand il signifie 'voie de communication terrestre', mais dans *route maritime*, *route* réfère à l'itinéraire et *route maritime* n'est donc pas l'hyperonyme d'*autoroute maritime*. (*iii*) Le changement de sens peut également provenir du mot lexical partagé par deux TMM comme dans *air libre* et *eau libre* où *libre* dans *air libre* signifie 'extérieur' et a un sens différent de son sens dans *eau libre*.

Le jeu de données contient également quatre cas de transfert de la relation où les deux TMM sont bien reliés par une relation sémantique différente de celle qui existe entre les TS.

— **ANTI** entre TS → **HYP** entre TMM. Par exemple, *terre* et *eau* ont des sens contrastifs selon DiCoEnviro. Cependant *ressources de la terre* et *ressources en eau* sont reliés plutôt par HYP dans les contextes (25a) et (25b), car *ressources de la terre* signifie 'ressources de la planète' dans le contexte (25b).

— **HYP** entre TS → **QSYN** entre TMM. Par exemple, selon DiCoEnviro, *planète* est un hyperonyme de *terre*. Néanmoins, dans les contextes (26a) et (26b), les sens fins de *réchauffement de la terre* et *réchauffement de la planète* sont similaires.

— **HYP** entre TS → **ANTI** entre TMM. Par exemple, *agriculture* et *élevage* sont reliés par HYP selon DiCoEnviro. Mais le sens fin d'*élevage biologique* dans (27a) et celui d'*agriculture biologique* dans (27b) sont contrastifs, car le premier concerne les animaux

et la seconde concerne les plantes.

— **QSYN** entre TS → **Cause** entre TMM. Par exemple, selon DiCoEnviro, *catastrophe* et *dégradation* ont des sens similaires dans certains contextes. Cependant, *catastrophe écologique* est plutôt la cause de *dégradation écologique* au vu de leurs sens dans les contextuels (28a) et (28b).

— **QSYN** entre TS → **Action** entre TMM. Par exemple, *recyclage* et *récupération* ont des sens similaires dans DiCoEnviro. Néanmoins, *recyclage des matériaux* et *matériaux de récupération* sont dans une relation Action dans les contextes (29a) et (29b).

(25) a. de réflexions sur la manière de gérer de façon plus efficace et plus durable les **ressources en eau** du moyen-orient .

b. nous ne voyons que le bénéfice à titre d' exemple , si la chine et l' inde décidaient de consommer autant de papier qu' un occidental , tous les arbres de la planète disparaîtraient en moins d' un an . nous sommes aveugles . nous ne voyons que le bénéfice immédiat à exploiter toutes les **ressources de la terre** " ?

(26) a. ces combustibles polluent l'air et envoient dans l'atmosphère des gaz qui provoquent le **réchauffement de la terre** .

b. la prise de conscience de la dégradation de la nature (pesticides , pollution, **réchauffement de la planète**, diminution de la couche d'ozone, effet de serre, contamination nucléaire

(27) a. par exemple, il vaut mieux manger du saumon issu de l' **élevage biologique** car cette espèce est menacée .

b. l'**agriculture biologique** peut atténuer les effets de nouveaux problèmes, comme les changements climatique, grâce à des mesures comme la fixation améliorée du carbone du sol.

(28) a. sombré ou sont endommagés, des marins sont morts ou portés disparus tandis qu' une **catastrophe écologique** sans précédent se profile dans la région .

b. les avocats de la décroissance, qui prétendent que la prospérité économique s' accompagne de la **dégradation écologique**, sous-entendu que les chinois auraient mieux fait d'en rester au stade de ...

(29) a. ... mettant en avant les meilleures conditions économiques et écologiques pour le traitement et le **recyclage des matériaux** des véhicules en fin de ...

b. ... les habitants des bidonvilles dans les zones inondables , dont les abris construits en **matériaux de récupération** sont extrêmement vulnérables .

6.5 Synthèse

Dans ce chapitre, nous avons présenté la construction par la projection sémantique d'un jeu de données composé de couples de TMM du domaine de l'environnement sémantiquement reliés.

Nous avons d'abord présenté le fonctionnement et l'utilisation de l'extracteur TermSuite pour extraire les candidats termes. Nous avons observé que plus de la moitié des candidats bitermes nominaux extraits avaient un patron syntaxique **N P N**. Nous avons aussi remarqué que certains candidats ne sont pas de vrais termes du domaine car ils ne sont pas syntaxiquement complets ou appartiennent à d'autres domaines que celui de l'environnement.

Nous avons ensuite présenté les TS nominaux ou adjectivaux reliés par les types de relation QSYN, ANTI et HYP extraits de DiCoEnviro. Ces couples de TS sont utilisés comme référence pour identifier les relations qui existent entre les TMM. Nous avons présenté le processus de projection. À l'issue de la projection, les trois types de relations lexicales entre termes simples ont été étendus aux 7 874 couples de candidats bitermes.

Le statut de terme des candidats a été validé en croisant les candidats avec les termes du domaine de l'environnement de trois ressources terminologiques : IATE, Termium Plus et GDT. Nous avons au final 231 couples composés de deux candidats qui sont de vrais termes du domaine. La relation inférée entre ces couples de TMM ont été examinés par trois annotateurs. Elle est préservée dans 180 couples de TMM (DataProj).

Il est intéressant de noter que DataProj et DataIATE sont complémentaires. En effet, l'intersection des couples QSYN de DataProj et DataIATE ne contient que 9 couples de TMM. Cette complémentarité est probablement due à la nature de IATE et de DiCoEnviro. Les synonymes de IATE sont destinés à la traduction et à la rédaction tandis que DiCoEnviro a été créé pour décrire et structurer les termes du domaine de l'environnement. En outre, il y a une légère différence dans les relations elles-mêmes. Certaines relations de DataIATE sont en effet hyperonymiques [6] comme *vie sauvage* : *animaux sauvages* ; de même, QSYN contient des relations de co-hyponymie comme *réseau ferroviaire* : *réseau routier*.

6. Dans certains cas, l'hyperonymie peut aussi être décrite comme une synonymie (Polguère, 2016)

Troisième partie

Exploration des MSD pour l'acquisition de relations sémantiques entre TMM

Chapitre 7

Acquisition de synonymie entre TMM

Ce chapitre est consacré à l'acquisition de synonymie entre TMM au moyen de la substitution lexicale et de l'analogie. Nous disposons en effet pour cette relation d'un jeu de données de test extrait de la banque IATE présenté en chapitre 5. Le jeu de données de test contient environ 300 couples de TMM, soit 6 fois plus que celui de Hazem et Daille (2018). Dans ce qui suit, nous présentons la substitution lexicale et l'analogie, les modèles de langue neuronaux et les mesures d'évaluation que nous avons utilisés. Pour chaque méthode, nous décrivons de façon détaillée les données de test, la mise en œuvre des méthodes et les résultats.

7.1 Cadre expérimental

Dans cette section, nous présentons d'abord la modélisation de la substitution lexicale et de l'analogie pour la tâche d'identification de relations lexicales entre TMM. Puis, nous décrivons les MSD sur lesquelles les deux méthodes ont été mises en œuvre. Nous décrivons ensuite leur évaluation.

7.1.1 Modélisation des méthodes

Substitution lexicale

Le sens d'un couple d'expressions et par suite la relation qui s'établit entre elles dépend des contextes dans lesquels elles apparaissent (Depraetere, 2019). Par exemple, *changement du climat* a un sens proche de celui de *réchauffement du climat*, dans le contexte (21a), mais signifie 'variabilité du climat' dans le contexte (21b). Dans le deuxième contexte, *changement du climat* n'est pas synonyme de *réchauffement du climat*. Pour distinguer ces deux acceptions de *changement du climat*, il est nécessaire d'utiliser une méthode qui prend en compte le sens des termes en contexte. Plus précisément, nous proposons d'identifier les relations lexicales entre TMM en utilisant une méthode par substitution lexicale fondée sur un modèle de langue masqué (MLM). Les MLM sont particulièrement adaptés à cette tâche car ils sont conçus pour

prédire les mots qui peuvent apparaître dans un contexte donné dans la position signalée par le token spécial <mask>. Les mots prédits sont les unités du vocabulaire du modèle qui sont les plus à même de replacer <mask> dans le contexte.

Dans notre étude, nous comparons deux stratégies d'interrogation : l'une au moyen de requêtes MLM de base et l'autre de requêtes MLM conditionnées.

La stratégie qui utilise des requêtes MLM de base pour identifier les relations lexicales peut être décrite formellement comme suit :

— soient TMM_1 et TMM_2 deux TMM ayant la même structure syntaxique tels que TMM_1 contient les mots lexicaux M_1 et M_3 et TMM_2 contient M_2 et M_3 ;
— soient S_1 un contexte de TMM_1 et S_2 un contexte de TMM_2 ;
— soit k_1 le rang de M_1 parmi les prédictions du MLM pour la requête obtenue en masquant M_2 dans S_2 ;
— soit k_2 le rang de M_2 parmi les prédictions du MLM pour la requête obtenue en masquant M_1 dans S_1 ;
— nous supposons que TMM_1 et TMM_2 ont un sens compositionnel. De ce fait, M_3 contribue de manière identique au sens de TMM_1 et TMM_2 ;
— soit N le nombre des voisins que l'on considère comme proches ;
— si $k_1 < N$ ou si $k_2 < N$, nous prédisons que (*i*) TMM_1 et TMM_2 sont sémantiquement reliés et (*ii*) la relation qui s'établit entre TMM_1 et TMM_2 est probablement la même que celle qui existe entre M_1 et M_2.

La méthode peut être illustrée par l'exemple suivant. L'existence d'un contexte S_1 permettant de construire une requête Q_1 dont les $N = 10$ premières réponses contiennent l'autre mot (*protection*) suffit à prédire que la relation (synonymie) entre les deux TS existe aussi entre les deux TMM.

Couple de TMM : préservation des forêts ; protection des forêts

M_1 : préservation

M_2 : protection

Relation d'intérêt : synonymie

Contexte S_1 masqué : l' aide financière à apporter pour la <mask> des forêts sera l' un des grands sujets abordés lors de la conférence .

N = 10

10 premières des prédictions pour la requête Q_1 : préserver, protection, conservation, restauration, reproduction, régénération, dégradation, durabilité, disponibilité, production

Observation : $k_2 = 2$, *protection* apparaît au rang 2 dans la liste des prédictions de la requête Q_1 ;

Conclusion : TMM_1 et TMM_2 sont synonymes.

La seconde stratégie trouve son origine dans l'observation de Zhou *et al.* (2019) que les MLM produisent des candidats qui peuvent être sémantiquement très différents du mot masqué tout en étant parfaitement adaptés au contexte. Qiang *et al.* (2019) proposent une solution à ce problème qui consiste à ajouter la phrase originale (dans laquelle aucun mot n'est masqué) comme un conditionnement supplémentaire sur la prédiction des tokens masqués. Qiang *et al.* (2019) appliquent cette méthode au modèle BERT pour réaliser une tâche de simplification lexicale. Espinosa Anke *et al.* (2021) utilisent la même méthode de conditionnement pour étudier les collocations. Plus précisément, la méthode de conditionnement consiste à utiliser comme entrée du MLM la concaténation de la phrase originale (où le mot cible n'est pas masqué) et de la phrase masquée (où le mot cible est masqué) afin d'améliorer la prédiction des voisins sémantiques du mot masqué. Les requêtes d'entrée fournies au MLM sans et avec conditionnement sont illustrées ci-dessous :

Requête sans conditionnement : l'aide financière à apporter pour la <mask> des forêts sera l'un des grands sujets abordés lors de la conférence.

Requête avec conditionnement : l'aide financière à apporter pour la préservation des forêts sera l'un des grands sujets abordés lors de la conférence [SEP] l'aide financière à apporter pour la <mask> des forêts sera l'un des grands sujets abordés lors de la conférence.

La méthode par substitution lexicale ne fournit que des relations entre des couples de TMM qui partagent un élément lexical et qui ont la même structure syntaxique. Les couples de TMM ignorés du fait de la première condition sont peu nombreux. Nous avons en effet constaté que 90 % des couples de TMM en relation de synonymie dans IATE partagent un élément lexical (cf. chapitre 5). La seconde condition a en revanche une incidence plus forte : seuls 68 % des couples de TMM en relation de synonymie dans IATE partagent la même structure syntaxique. Conjointement, les deux conditions sont vérifiées par 61 % des couples de DataIATA.

La seconde méthode que nous avons utilisée permet de relâcher la seconde contrainte. L'analogie permet en effet de généraliser l'exploration aux couples de TMM ayant des structures syntaxiques différentes. En revanche, elle ne peut être utilisée que pour des couples de TMM qui partagent un élément lexical.

Analogie

L'analogie est une relation proportionnelle entre des couples d'objets qui se trouvent dans la même relation (Lepage et Ando, 1996 ; Lepage, 1998 ; Skousen, 2002 ; Claveau et L'Homme, 2005 ; Turney, 2008 ; Langlais *et al.*, 2009). Elle peut notamment être utilisée pour identifier des relations entre les entrées d'un MSD. Mikolov *et al.* (2013a) sont les premiers à avoir utilisé l'analogie proportionnelle dans les plongements de mots pour découvrir des relations sémantiques à partir des contextes des mots. Ils ont montré que certaines relations entre les mots peuvent être captées dans une large mesure par les décalages qui existent entre leurs plongements vectoriels, comme permet de l'illustrer la Figure 24. Dans cette figure la relation sémantique

entre deux entrées est représentée par le décalage entre leurs vecteurs. En l'occurrence, la relation qui s'établit entre les trois couples de mots est la même et par suite le décalage entre les vecteurs des trois couples de mots sont identiques.

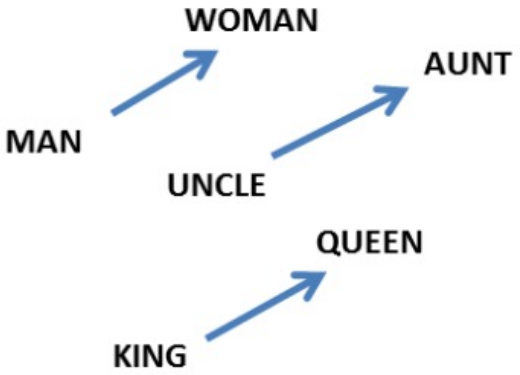


FIGURE 24 – Décalages entre les représentations vectorielles de trois couples de mots reliés par la relation de genre (Mikolov *et al.*, 2013c)

La détection des relations entre TMM par analogie découle de l'observation que si $M_1 : M_2 :: TMM_1 : TMM_2$ est une analogie proportionelle alors la relation entre M_1 et M_2 est identique que celle entre TMM_1 et TMM_2. Par exemple, s'il existe une analogie entre *sec*:*humide*::*climat sec*:*climat humide*, alors *climat sec* et *climat humide* sont reliés par la même relation que celle qui relie *sec* et *humide*.

La découverte des relations entre TMM par analogie consiste à résoudre des équations analogiques sur les représentations vectorielles. Les solutions d'une équation $a : b :: c :?$ sont recherchées parmi les mots dont la représentation vectorielle V_d est similaire au vecteur $V_c - V_a + V_b$. La solution de l'équation peut ainsi être obtenue comme :

$$\underset{d \in Voc}{\operatorname{argmax}}(\text{similarity}(V_d, V_c - V_a + V_b))$$

En d'autres termes, on cherche dans le vocabulaire *Voc* les termes pour lesquels le vecteur V_d est le plus proche du vecteur $V_c - V_a + V_b$ relativement à la mesure *similarity*. Cette méthode est connue sous le nom de 3cosADD lorsque la mesure de similarité est cosinus.

Dans notre cas, nous connaissons la relation entre M_1 et M_2 et nous cherchons à savoir si cette même relation existe entre TMM_1 et TMM_2. Soit V_{M_1}, V_{M_2}, V_{TMM_1} et V_{TMM_2} les représentations vectorielles de M_1, M_2, TMM_1 et TMM_2. L'information que nous cherchons à découvrir étant la relation entre les deux TMM, nous utilisons donc un de ces derniers comme inconnu dans l'équation analogique. Plus précisément, le test analogique est réalisé deux fois pour les relations symétriques (synonymie et antonymie), en prenant l'un puis l'autre des deux TMM comme inconnu. Ainsi, chaque quadruplet donne lieu à deux équations analogiques : $M_1 : M_2 :: TMM_1 :?$ ou $M_1 : M_2 ::? : TMM_2$. Si nous choisissons par exemple TMM_2, comme inconnue, nous chercherons alors à estimer la distance entre la représentation de TMM_2

et le vecteur attendu $V_{attendu} = V_{TMM_1} - V_{M_1} + V_{M_2}$. Le résultat final est la moyenne des rangs des TMM inconnus dans les prédictions pour les deux équations. En revanche, les relations hiérarchiques (hyperonymie et hyponymie) étant orientées, seule une équation analogique est utilisée : $M_1 : M_2 :: TMM_1 :?$. L'exemple ci-dessous illustre la méthode par analogie :

Quadruplet : empreinte :impact : :empreinte environnemental :impact environnemental

Relation connue : la synonymie entre *empreinte* et *impact*

Équations d'analogie : $équation_1$: *empreinte :impact : :empreinte environnemental : ?* ; $équation_2$: *empreinte :impact : : ? :impact environnemental*

TMM inconnu : *impact environnemental* pour $équation_1$; *empreinte environnemental* pour $équation_2$

Nombre des voisins considérés comme proche : 5

5 premières prédictions pour $équation_1$ **:** *empreinte écologique, empreinte carbone, empreinte environnemental, bilan écologique, bilan carbone*

5 premières prédictions pour $équation_2$ **:** *impact environnemental, impact écologique, impact positif, effet environnemental, coût de dépollution*

Observation : *empreinte environnemental* apparaît au rang 3 dans la liste de prédiction pour $équation_1$; *impact environnemental* apparaît au rang 1 dans la liste de prédiction pour $équation_2$

Résultat final : la moyenne la moyenne des rangs des TMM inconnus égale à 2

Conclusion : *empreinte environnemental* et *impact environnemental* sont aussi reliés par la synonymie

7.1.2 Modèles

Nous avons utilisé deux types de modèles de sémantique distributionnel : le MSD contextuel CamemBERT (Martin *et al.*, 2020) pour la méthode par substitution lexicale et le MSD statique FastText (Bojanowski *et al.*, 2017) pour la méthode par analogie.

FastText

FastText est essentiellement une extension de Word2Vec dont les entrées sont les mots présents dans le corpus d'entraînement et les *n*-grammes avec la longueur de 3 à 6 de ces mots. Considérons par exemple le mot *climat*. Des chevrons sont ajoutés pour indiquer le début et la fin de chaque occurrence du mot *climat* : climat ⟶ <climat>. Puis, les *n*-grammes de caractères sont générés en faisant glisser une fenêtre de *n* caractères. Le Tableau 28 présente les *n*-grammes de caractères de *climat* pour chaque valeur de *n*.

Cette extension permet de calculer de meilleures représentations pour les mots rares en tirant profit du fait qu'ils partagent des *n*-grammes de caractères avec d'autres mots. Dans FastText, La

n	n-grammes de caractères
3	<cl, cli,lim, ima, mat, at>
4	<cli, clim, lima, imat, mat>
5	<clim, clima, limat, imat>
6	<clima, climat, limat>

TABLE 28 – *n*-grammes de caractères de climat de longueurs de 3 à 6

représentation d'un mot est en effet la moyenne des représentations du mot et de ces *n*-grammes. De même, FastText fournit des représentations aux mots OOV (hors du vocabulaire). Dans ce cas, le plongement est obtenu en faisant la moyenne des représentations des *n*-grammes de caractères consécutifs contenus dans le vocabulaire du modèle. Pour l'entraînement des modèles, FastText dispose de 10 hyperparamètres possibles [1] qu'il est possible d'optimiser. Nous en présentons certains en Section 7.2.2.

CamemBERT

CamemBERT est un MSD contextuel du français qui adopte l'architecture RoBERTa (Liu *et al.*, 2019). RoBERTa est un modèle BERT qui utilise une approche d'entraînement différente. Le modèle n'est pas entraîné à prédire les phrases suivantes. L'entraînement comporte par ailleurs un masque dynamique. Dans BERT, l'entrée n'est masquée qu'une seule fois de sorte qu'elle a les mêmes mots masqués à toutes les époques tandis que dans RoBERTa, les mots masqués changent d'une époque à l'autre. CamemBERT se distingue aussi par le fait qu'il est entraîné sur un corpus en français, i.e., le sous-corpus français du corpus multilingue OSCAR [2].

CamemBERT utilise SentencePiece (Kudo et Richardson, 2018) comme tokenizer. La taille du vocabulaire est ainsi prédéterminée et l'entraînement du modèle est réalisé sur le corpus brut (aucun prétraitement n'est réalisé). Le vocabulaire du modèle contient des mots entiers et de sous-mots (*word pieces*). Lorsqu'un mot n'existe pas dans le vocabulaire du modèle, il est tokenisé en sous-mots. Par exemple, *piscicole* est divisé en 3 sous-mots : *pis*, *s*, *cicole*.

7.1.3 Mesures d'évaluation

Nous évaluons les deux méthodes à l'aide du score MRR (*mean reciprocal rank*) (Radev *et al.*, 2002 ; Chowdhury, 2010) et de la précision à TOP1, TOP5 et TOP10, i.e. la proportion de réponses correctes parmi les premiers candidats, parmi les 5 premiers candidats de chaque requête et parmi 10 premiers candidats. Le score MRR est utilisé pour estimer la performance des méthodes qui prennent en entrée une requête et fournissent en sortie une liste ordonnée de réponses. Ce score ne prend en compte que le rang de la première réponse correcte dans cette liste. Il est définit comme suit :

1. https://FastText.cc/docs/en/options.html
2. https://oscar-corpus.com

$$MRR = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rang_i}$$

où $|W|$ est le nombre de requêtes et où $Rang_i$ est le rang de la première réponse correcte pour la i-ième requête. Plus le score MRR est proche de 1, plus le modèle est performant. Le Tableau 29 illustre le calcul du score MRR pour 3 requêtes. Le terme cible est présenté en colonne 1, les candidats en colonne 2, le rang de la réponse correcte en colonne 3 et son inverse en colonne 4. Pour chaque requête, le modèle propose une liste de trois candidats. Le score MRR calculé pour les requêtes est :

$$\frac{1}{3} \times (1 + \frac{1}{2} + \frac{1}{3}) = \frac{11}{18} \approx 0,61$$

TMM inconnu	**Candidats ordonnées**	**Rang**	**1/Rang**
air chaud	air chaud, air humide, taux d'évaporation	1	1
capital humain	espace humain, capital humain, capacité de charger	2	1/2
climat local	réchauffement local, climat, climat local	3	1/3

TABLE 29 – Illustration du calcul du score MRR

La formule de la précision est la suivante :

$$\text{Précision} = \frac{n}{|W|}$$

où n est le nombre de requêtes qui produisent un résultat correct parmi les réponses à TOP1, TOP5 ou TOP10 et où $|W|$ est le nombre total de requêtes. Par exemple, pour une expérience qui comporte 10 requêtes, si 7 parmi elles fournissent le terme cible parmi les 5 premières réponses du modèle, la précision à TOP5 est de $7/10$ soit $0,7$.

7.2 Expériences avec DataIATE

Les premières expériences que nous présentons permettent d'explorer la capacité des MSD à identifier la synonymie entre TMM en utilisant les couples de TMM synonymes de DataIATE comme jeu de test. Nous présentons dans un premier temps les expériences par substitution lexicale, puis celle par analogie.

7.2.1 Substitution lexicale

Données de test

La méthode par substitution lexicale présentée en Section 7.1.1 requiert des couples de bitermes ayant le même patron syntaxique et partageant un élément lexical, comme *prestation*

de service et *fourniture de service*. Par ailleurs, les deux unités lexicales différentes doivent être contenues dans les bitermes et doivent faire partie du vocabulaire du modèle. Nous avons donc constitué un jeu de données DataIATE_MLM obtenu en retirant de DataIATE les couples dont ces deux unités ne sont pas incluses dans le vocabulaire du modèle. En effet, comme indiqué en Section 7.1.2 les mots hors vocabulaire sont divisés en sous-mots ce qui les exclut de fait des réponses des requêtes qui comportent un seul token <mask>.Le jeu de données DataIATE_MLM contient 396 couples de TMM synonymes. Il sera utilisé pour tester les méthodes par substitution lexicale.

Les requêtes MLM sont construites à partir des contextes des TMM. Pour chaque TMM qui figure dans les 396 couples de DataIATE_MLM, nous sélectionnons à partir du corpus PANACEA 100 contextes riches en informations (dans la mesure du possible). Meyer (2001) considère les contextes contenant des informations précieuses comme des contextes riches en connaissances (KRC ; *knowledge rich contexts*). Les KRC contiennent à la fois un terme d'intérêt d'un domaine particulier et un patron de connaissance (KP ; *knowledge pattern*) qui indique comment ce terme est relié à d'autres termes du domaine. Barrière (2004) élargit la définition de KP pour y inclure des informations sémantiques. Hmida *et al.* (2015) proposent de compléter les KP par des collocations. Les contextes contenant des KP et des collocations sont effectivement souvent riches en informations, cependant leur nombre est limité. De ce fait, nous avons utilisé certains des critères de qualité définis par Kilgarriff *et al.* (2008) pour sélectionner les bons contextes. Parmi ces critères, nous retenons : (*i*) la longueur du contexte qui doit être comprise entre 10 et 100 mots ; (*ii*) le fait que les contextes doivent être des phrases ; (*iii*) le fait que les contextes doivent contenir au moins un autre terme spécifique au domaine de l'environnement en plus du TMM considéré. Pour la vérification de la troisième condition, nous avons utilisé l'ensemble des termes du domaine de l'environnement dans IATE.

Pour certains TMM, le nombre de contextes correspondant à ces critères est inférieur à 100. Par ailleurs, nous avons supprimé les couples composés de TMM dont aucun contexte ne satisfait les critères, comme *taxe d'émission*. À l'issue de cette sélection, les données de test se composent de 317 couples de TMM et de 24 265 contextes. Les Tableaux 30 et 31 présentent un extrait des couples de TMM et des contextes extraits.

Expérimentation

Nous avons utilisé le modèle CamemBERT-large pour les expériences visant à acquérir la synonymie entre TMM du domaine de l'environnement par la méthode de substitution lexicale. Dans ces expériences, nous considérons uniquement les réponses qui sont des termes simples du domaine. Les TMM étant extraits de IATE, une solution serait de ne considérer que les termes simples qui apparaissent dans cette même banque. Cependant, le nombre de termes simples dans IATE s'est avéré trop faible. Nous avons donc utilisé un vocabulaire plus étendu constitué des 784 unités lexicales qui apparaissent dans les TMM de DataIATE et qui font partie du vocabulaire du modèle. Par exemple, ce vocabulaire contient *restauration* et *forêt* qui proviennent du TMM

TMM1	**TMM2**
évolution de le population	dynamique de le population
densité maximal	densité potentiel
équipement sportif	infrastructure sportif
finance vert	financement vert
bassin de carbone	réserve de carbone
réservoir aquifère	formation aquifère
moyenne glissant	moyenne mobile
accord volontaire	accord environnemental
espèce caractéristique	espèce indicateur
camomille sauvage	camomille allemand
année de référence	année de base
couvert naturel	végétation naturel
énergie doux	technologie doux

TABLE 30 – Exemples de couples de TMM du jeu de test DataIATE_MLM utilisé pour la substitution lexicale

restauration des forêts et qui font tous deux partie du vocabulaire de CamemBERT-large.

La prédiction est réalisée en utilisant les contextes non lemmatisés, car le modèle CamenBERT-large est pré-entraîné sur un corpus non lemmatisé. Par ailleurs, la vérification de la présence des termes dans les contextes est réalisée en utilisant les formes lemmatisées des TMM. Pour ce faire, les contextes lemmatisés et non lemmatisés ont été alignés. On peut ainsi déterminer la position du mot à masquer dans le contexte lemmatisé puis masquer la forme correspondante dans le contexte non lemmatisé. Par exemple, pour trouver la position du mot à masquer *sol* dans un contexte non lemmatisé du terme *horizon du sol* comme (30a), nous recherchons d'abord sa position dans le contexte lemmatisé (30b) : *sol* se trouve en 13^{e} position. Nous pouvons alors masquer le mot qui se trouve dans la même position dans le contexte non lemmatisé (30c).

(30) a. il s' agit d' éviter le compactage des horizons du sol et la mortalité des racines superficielles

b. il s ' agir d ' éviter le compactage du horizon du sol et le mortalité du racine superficiel

c. il s ' agit d ' éviter le compactage des horizons du <mask> et la mortalité des racines superficielles

Résultats et analyses

Résultats et analyses généraux. Le Tableau 32 donne les résultats des requêtes sans et avec conditionnement. Les résultats sont améliorés avec les requêtes avec conditionnement. Les scores MRR passent de 0,302 à 0,374 ce qui signifie qu'en moyenne, la bonne réponse se trouve en 3^{e} ou 4^{e} position dans le premier cas, et au 2^{e} ou 3^{e} dans le second. Les scores de précision sont également améliorés.

TMM	**Contexte**	**Contexte lemmatisé**
centre de compostage	en ce qui concerne les déchets organiques , des efforts devront être faits sur les normes assurant la qualité du compost et des contrôles devront être effectués à l ' entrée des centres de compostage afin de s ' assurer de la qualité du compost produit .	en ce qui concerner le déchet organique , du effort devoir être faire sur le norme assurer le qualité du compost et du contrôle devoir être effectuer à l ' entrée du centre de compostage afin de s ' assurer de le qualité du compost produire .
installation de compostage	il est également nécessaire de prendre en compte les contraintes d ' environnement de l ' installation de compostage (disponibilité foncière , proximité d ' habitations , risques d ' odeurs) et d ' écoulement du compost , dans la conception technique de l ' opération (stockage du compost produit dont l ' écoulement est saisonnier) .	il être également nécessaire de prendre en compte le contrainte d ' environnement de l ' installation de compostage (disponibilité foncier , proximité d ' habitation , risque d ' odeur) et d ' écoulement du compost , dans le conception technique de l ' opération (stockage du compost produire dont l ' écoulement être saisonnier) .
sac à déchet	faites attention aux substances que vous versez dans les tuyaux d ' écoulement ou que vous mettez dans des sacs à déchets ordinaires .	faire attention au substance que vous verser dans le tuyau d ' écoulement ou que vous mettre dans du sac à déchet ordinaire .
sac poubelle	par ailleurs pour le consommateur , l ' élimination par le sac poubelle ne serait pas meilleur marché .	par ailleurs pour le consommateur , l ' élimination par le sac poubelle ne être pas meilleur marché .
installation compact	une faible emprise au sol des unités de traitement et l ' existence d ' une l ' offre d ' installations compactes ,	un faible emprise au sol du unité de traitement et l ' existence d ' un l ' offrir d ' installation compact ,
plante à fleur	les plantes à fleurs et les abeilles entretiennent une relation d ' interdépendance : l ' une ne peut pas exister sans l' autre .	le plante à fleur et le abeille entretenir un relation d ' interdépendance : l ' un ne pouvoir pas exister sans l ' autre .
horizon du sol	il s ' agit d ' éviter le compactage des horizons du sol et la mortalité des racines superficielles .	il s ' agir d ' éviter le compactage du horizon du sol et le mortalité du racine superficiel .

TABLE 31 – Exemples de contextes extraits du corpus PANACEA pour les TMM du jeu de données DataIATE_MLM

	MRR	**Précision**		
		TOP1	TOP5	TOP10
MLM sans conditionnement	0,302	0,189	0,416	0,532
MLM avec conditionnement	0,374	0,253	0,502	0,613

TABLE 32 – Résultats des prédictions pour les requêtes MLM sans et avec conditionnement sur le jeu de données DataIATE_MLM

Une analyse qualitative des prédictions permet de compléter l'analyse quantitative de ces résultats. Nous avons analysé les 10 premières prédictions de 100 requêtes avec conditionnement sélectionnées aléatoirement. Nous avons observé que la plupart des prédictions sont sémantiquement proches du mot masqué. Il s'agit le plus souvent de synonymes et de variantes notamment dérivationnelles. Ces résultats renforcent l'observation faite par Ferret (2021) et Arefyev *et al.* (2020). Pour certaines requêtes, les variantes ou synonymes du terme cible apparaissent parmi les premiers candidats. Par exemple, lorsque *habitation* est masquée dans un contexte *habitation individuelle*, le cible *maison* n'apparaît qu'au rang 71, mais son synonyme *logement* apparaît lui au rang 2.

Influence de la relation entre les éléments lexicaux des TMM sur la prédiction. Nous remarquons également que les meilleurs résultats ont été obtenus pour les couples TMM dont M_1 et M_2 sont des synonymes. Pour confirmer cette observation qualitative, nous avons refait l'expérience sur 25 couples TMM dont les éléments lexicaux M_1 et M_2 sont des synonymes dans IATE. 2 432 requêtes ont ainsi été testées. Le score MRR obtenu avec les requêtes avec conditionnement est de 0,561, soit une amélioration de 50 %. Ce résultat confirme l'intérêt, pour l'identification des TMM synonymes, de combiner l'approche distributionnelle avec une méthode compositionnelle qui génère des TMM synonymes en substituant un synonyme à l'un des composants d'un TMM de départ (Daille, 2017).

7.2.2 Méthode par analogie

Données de test

Les données de test pour les expériences par analogie sont construites à partir des couples de TMM de DataIATE_FastText (cf. Section 5.2). Nous générons un quadruplet $M_1 : M_2 :: TMM_1 : TMM_2$ pour chaque couple de TMM tel que (*i*) TMM_1 et TMM_2 partagent un constituant commun M_3 ; (*ii*) TMM_1 contient un constituant M_1 et TMM_2 un constituant M_2 et M_1 et M_2 sont des synonymes. La synonymie de M_1 et M_2 a été vérifiée en utilisant RefDicoenviro_qsyn, RefIATE et RefDicosyn. RefDicoEnviro_qsyn est un jeu de données constitué de 523 couples de termes simples QSYN dans RefDicoEnviro cf. chapitre 6) comme *conservation : préservation*. RefIATE est un jeu de données composé de 551 couples de termes simples (noms et adjectifs) synonymes appartenant au domaine de l'environnement extraits d'IATE, comme *rejet : déversement*. RefDS est composé d'un jeu de données de 833 891 couples de noms et d'adjectifs synonymes extraites de Dicosyn, comme *empreinte : impact*. Dicosyn est une compilation de sept dictionnaires des synonymes du français qui relève de la langue générale. Le Tableau 33 présente des exemples de quadruplets construits en utilisant DataIATE_FastText. Par ailleurs, trois ensembles de quadruplets ont été construits à partir de RefIATE, RefDicosyn et RefDicoEnviro_qsyn. Il s'agit de Quadruplet_RefIATE qui comprend 20 quadruplets, Quadruplet_RefDicoEnviro_qsyn qui comprend 63 quadruplets et Quadruplet_RefDicosyn qui

comprend 156 quadruplets.

Constituant 1	**Constituant 2**	**TMM1**	**TMM2**
feu	incendie	feu de forêt	incendie de forêt
conservation	préservation	conservation de le biodiversité	préservation de le biodiversité
environnemental	écologique	sécurité environnemental	sécurité écologique
vert	écologique	label vert	label écologique
pénurie	raréfaction	pénurie du ressource	raréfaction du ressource
dégradation	détérioration	dégradation environnemental	détérioration environnemental
couvert	couverture	couvert forestier	couverture forestier
vert	écologique	économie vert	économie écologique
environnemental	écologique	préoccupation environnemental	préoccupation écologique
phénomène	événement	phénomène naturel	événement naturel
dégradation	détérioration	dégradation du sol	détérioration du sol
région	zone	région côtier	zone côtier
vert	écologique	voiture vert	voiture écologique
protection	préservation	protection du paysage	préservation du paysage

TABLE 33 – Extrait des données de test construites à partir des TMM synonymes extraits de IATE pour l'évaluation de la méthode par analogie

Expériences

La tâche portant sur l'acquisition de relations entre TMM, les rangs des solutions candidates à l'équation analogique sont calculés relativement à un jeu de données composé de 5 002 TMM extraits de IATE. Ce jeu de données est composé de bitermes nominaux qui apparaissent au moins 5 fois dans le corpus PANACEA. Nous utilisons un modèle FastText pour l'acquisition de synonymes par analogie car il permet de représenter dans le même espace, de manière indépendante les TMM et leurs constituants. Les représentations des premiers ne doivent en effet pas être calculées par la composition à partir des représentations de leurs constituants car l'équation d'analogie serait alors toujours trivialement vraie. Pour calculer des représentations pour les TMM et leurs composants dans le même espace vectoriel, nous avons annoté le corpus de sorte que les uns et les autres soient indexés séparément. Par exemple, un TMM tel que *air froid* produit les trois tokens : `air`, `froid` et `air_froid`. Nous avons par ailleurs imposé au modèle à ne pas découper les mots en *n*-grammes de caractères en définissant le paramètre `maxn` à 0.

La performance du modèle FastText peut être impactée significativement par ses hyper-paramètres (Levy *et al.*, 2015a). Par ailleurs, la précision de la résolution des analogies varie énormément pour les relations différentes. Cependant, l'optimisation de la précision moyenne sur un ensemble de relations variées ne garantit pas nécessairement de meilleures performances sur une relation particulière (Gladkova *et al.*, 2016). Nous n'avons donc pas réalisé ces optimisations. Les hyper-paramètres que nous avons utilisés sont :

min_count = 3 : le nombre minimal d'occurrences du mot est 3 ;

maxn = 0 : la longueur maximale d'un *n*-gramme de caractères est 0 ;

ws = 5 : la taille de la fenêtre est 5 ;

model = skipgram : l'architecture du modèle est skipgram ;

lr = 0,05 : le taux d'apprentissage est 0,05.

Tous les autres paramètres sont fixés à leurs valeurs par défaut.

Résultats et analyses

Résultats. Le Tableau 34 présente les résultats produits par la méthode analogie. Les scores MRR sont élevés et tous supérieurs à 0,6. Il témoignent du fait que l'analogie est une méthode efficace pour l'acquisition de synonymie entre TMM. Les quadruplets de quadruplet_RefIATE donnent les meilleurs résultats avec un score MRR de 0,744. Ce bon chiffre peut s'expliquer par le fait que les relations de synonymie entre les TMM et les termes simples qui composent les quadruplets de quadruplet_RefIATE proviennent de la même base de données, i.e., IATE. Un autre résultat intéressant est que les résultats obtenus avec les quadruplets construits à partir des bases de données spécialisées et à partir de synonymes du domaine général sont proches. Cela pourrait être dû au fait que les termes simples du domaine de l'environnement sont moins spécialisés que les TMM. Par exemple, des termes simples de DiCoEnviro tels que *mondial*, *impact* et *effet* sont (aussi) des mots courants.

	MRR	**Précision**		
		TOP1	TOP5	TOP10
Quadruplet_RefIATE	0,744	0,650	0,875	0,900
Quadruplet_RefDicoEnviro_qsyn	0,624	0,548	0,723	0,746
Quadruplet _RefDicosyn	0,612	0,522	0,728	0,766

TABLE 34 – Acquisition de TMM synonymes par analogie en utilisant les représentations FastText et les données extraites de IATE

Nous avons effectué une analyse qualitative similaire à celle que nous avons réalisée pour la substitution lexicale. Nous avons examiné les prédictions à TOP5 de toutes les requêtes dont l'inconnue est TMM_2. Nous avons ainsi observé que TMM_2 est fourni parmi les cinq premiers candidats pour 73,6 % des quadruplets. Lorsque TMM_2 n'apparaît pas au TOP5, nous avons trouvé dans la plupart de cas un de ses synonymes parmi ces candidats. Par exemple, pour le quadruplet *effet :incidence : :effet sur l'environnement :incidence sur l'environnement*, le terme inconnu *incidence sur l'environnement* se trouve au rang 3$\tilde{6}$98, mais la première prédiction est une variante dérivationnelle, i.e., *incidence environnementale*.

Influence du vocabulaire. Afin d'estimer la généralité de notre méthode, nous avons réalisé une seconde expérience en utilisant Quadruplet_RefIATE comme données de test et le vocabulaire du modèle comme vocabulaire dans lequel les solutions de l'équation analogique sont

recherchées. Le vocabulaire est composé de 141 253 expressions. Sa taille est très supérieure à nombre de bitermes nominaux extraits d'IATE (5 002 TMM). Pour cette nouvelle configuration, nous avons obtenu un score de MRR de 0,574, inférieur à celui obtenu avec le vocabulaire que nous avons construit et une précision de 0,775 et 0,825 à TOP5 et TOP10, supérieures à celles obtenues avec Quadruplet_RefDicoEnviro_qsyn et Quadruplet_RefDicosyn. Nous pouvons donc considérer que la méthode par analogie présente une certaine généralité.

7.3 Synthèse

Nous avons présenté dans ce chapitre le cadre expérimental et les expériences réalisées pour tester la capacité des MSD à capter la synonymie sur un jeu de données composé de TMM synonymes extraits de IATE. Notre jeu de données de test est constitué d'environ 300 couples de TMM. Nous avons présenté les méthodes par substitution lexicale et par analogie. Deux types de requêtes ont été utilisées pour la méthode par substitution lexicale. La première utilise la capacité principale des MLM de prédire les mots qui correspondent à un contexte particulier. La seconde utilise une stratégie de conditionnement qui permet de fournir au modèle davantage d'informations sur le mot masqué afin d'améliorer les substituts prédits.

Les modèles de langue utilisés pour les expériences sont le modèle contextuel CamemBERT-large pour la substitution lexicale et un modèle statique FastText pour l'analogie. CamemBERT-large est un modèle pré-entraîné sur un corpus en français sur une tâche de prédiction de token masqué avec un masquage dynamique. Ce modèle est mieux adapté aux expériences de prédiction de mots masqués entiers relativement à d'autres modèles comme FlauBERT (Le *et al.*, 2020) pré-entraîné pour la prédiction de sous-mots masqués. Pour la méthode par analogie, nous avons utilisé un modèle FastText car contrairement à d'autres modèles statiques comme Word2Vec, il permet de générer indépendamment des vecteurs pour les mots simples et les TMM dans le même espace vectoriel. Nous avons présenté les scores MRR et de précision que nous avons utilisés pour évaluer les méthodes proposées.

Nous avons ensuite décrit les expériences qui ont été réalisées. Globalement, les modèles que nous avons testés captent relativement bien la synonymie entre TMM. La substitution lexicale au moyen de requêtes avec conditionnement permet d'obtenir un MRR 0,374. Le score MRR obtenu par la méthode par analogie est quant à lui presque 2 fois de plus élevé ce qui suggère une meilleure prise en compte par l'analogie de la compositionnalité des TMM. Cette compositionnalité semble également prise en compte par les modèles contextuels comme l'indique l'amélioration du score MRR de 50 % pour les quaduplets dont les composants M_1 et M_2 sont synonymes. Une analyse contrastive des deux méthodes et une comparaison de leurs résultats avec ceux de travaux existants sont présentées dans le chapitre suivant.

Chapitre 8

Acquisition de relations ANTI, HYP et QSYN

Les méthodes que nous venons de présenter peuvent aussi être utilisées pour l'acquisition d'autres types de relations lexicales, notamment les relations d'opposition et les relations hiérarchiques qui sont considérées comme faisant partie des relations terminologiques essentielles (L'Homme, 2020). Le jeu de données utilisé pour les expériences que nous avons menées est DataProj. Il est composé de trois types de relations lexicales : ANTI, HYP et QSYN. QSYN contient des couples de synonymes et de quasi-synonymes mais aussi des co-hyponymes comme *réseau ferroviaire* et *réseau routier*. On retrouve ainsi un recouvrement partiel entre les relations de DataProj et celles de DataIATE qui contient des synonymes, des quasi-synonymes. Par ailleurs, DataIATE contient des hyperonymes mais pas de co-hyponymes. Les deux jeux de données sont néanmoins relativement différents puisque seuls 9 couples de TMM sont communs à DataIATE et à la partie QSYN de DataProj. Ceci explique notre choix de présenter séparément les expériences sur DataIATE et sur la partie QSYN de DataProj.

Les expériences présentées dans ce chapitre sont similaires à celles présentées en Section 7.2. La seule différence est qu'elles sont réalisées en utilisant DataProj. Pour chaque méthode, les données de test, la mise en œuvre et les résultats sont présentés de façon détaillée. La dernière section est consacrée à une discussion des méthodes et les résultats obtenus.

8.1 Substitution lexicale

8.1.1 Données de test

Les expériences qui utilisent la méthode par substitution lexicale ont été réalisées sur des données de test composées des quadruplets de DataProj dans lesquels les deux TS appartiennent au vocabulaire du modèle. Ce sous-ensemble contient 61 couples ANTI, 16 couples HYP et 71 couples QSYN. Les Tableaux 35 et 36 présentent un extrait des couples de TMM et des contextes que nous avons utilisés pour ces expériences.

TMM1	**TMM2**	**TS1**	**TS2**	**Rel**
air chaud	air froid	chaud	froid	ANTI
aménagement urbain	aménagement rural	urbain	rural	ANTI
assainissement collectif	assainissement individuel	collectif	individuel	ANTI
augmentation de le biodiversité	diminution de le biodiversité	augmentation	diminution	ANTI
augmentation du effectif	diminution du effectif	augmentation	diminution	ANTI
augmentation du tarif	baisse du tarif	augmentation	baisse	ANTI
climat de le planète	climat de le terre	planète	terre	HYP
gestion du combustible	gestion de le biomasse	combustible	biomasse	HYP
hydrate de gaz	hydrate de carbone	gaz	carbone	HYP
planète vivant	terre vivant	planète	terre	HYP
environnement rural	milieu rural	environnement	milieu	QSYN
habitat urbain	milieu urbain	habitat	milieu	QSYN
agriculture organique	agriculture biologique	organique	biologique	QSYN
air frais	air froid	frais	froid	QSYN
approvisionnement en électricité	approvisionnement en énergie	électricité	énergie	QSYN
augmentation de le population	croissance de le population	augmentation	croissance	QSYN
augmentation du tarif	hausse du tarif	augmentation	hausse	QSYN
bilan environnemental	bilan écologique	environnemental	écologique	QSYN

TABLE 35 – Extrait de couples de TMM utilisés pour tester la capacité des MSD à identifier des relations lexicales entre TMM par substitution lexicale

TMM	**Contexte**
patrimoine animal	pourquoi nos choix de consommation déterminent les activités de la chaîne de production alimentaire et le devenir des savoir-faire comme du patrimoine animal , végétal et culturel ?
patrimoine végétal	d' ici peu de temps il sera accessible à un large public national et international rendant ainsi les collections utilisables pour une meilleure valorisation des travaux de recherche et du patrimoine végétal national .
règne animal	c' est la théorie de malthus appliquée à tout le règne animal et tout le règne végétal .
règne végétal	dans le sable puis sous une tente de chercheur , ils pourront s' amuser à identifier l' origine et le nom de leurs trouvailles et retracer ainsi les grandes étapes de l' évolution du règne végétal .
augmentation du tarif	l' augmentation des tarifs de l' électricité aurait pour origine , nous dit -on aussi , la seule énergie photovoltaïque .

TABLE 36 – Exemples de contextes utilisés pour créer des requêtes pour l'acquisition de relations lexicales entre TMM par substitution lexicale

8.1.2 Expérimentation

Nous avons réalisé une expérience par substitution lexicale similaire à celle présentée en Section 7.2.1 en utilisant 100 contextes (ou moins) extraits du corpus PANACEA pour chaque TMM qui apparaît dans l'un des quadruples des données de test. Comme précédemment, le modèle utilisé est CamemBERT-large. De même, le rang des candidats relatifs à un vocabulaire composé de tous les termes simples du DiCoEnviro qui apparaissent dans le corpus PANACEA et dans le vocabulaire du modèle (soit 796 termes simples).

8.1.3 Résultats et analyses

Résultats et analyses généraux

Les Tableaux 37 et 38 présentent les résultats obtenus pour les requêtes sans et avec conditionnement. Nous pouvons voir dans le Tableau 37 que les requêtes sans conditionnement capturent QSYN avec un MRR de 0,54, ce qui signifie qu'en moyenne la réponse correcte est soit la première soit deuxième. En revanche, ANTI et HYP sont plus difficiles à identifier. Le Tableau 38 permet de voir que la méthode avec conditionnement améliore le MRR et la précision de manière significative, en particulier pour l'hyperonymie avec une augmentation de 69 %. Comme dans l'expérience d'acquisition de synonymie, nous avons analysé manuellement les dix premières prédictions de 100 requêtes avec conditionnement sélectionnées aléatoirement. Les conclusions sont les mêmes que précédemment : on retrouve pour de nombreuses requêtes, des variantes flexionnelles et dérivationnelles du terme cible ainsi que des synonymes parmi les premiers candidats.

MLM sans conditionnement	**MRR**	**Précision**		
		TOP1	TOP5	TOP10
ANTI	0,316	0,160	0,498	0,687
Hyperonyme	0,360	0,229	0,513	0,635
Hyponyme	0,398	0,289	0,506	0,645
QSYN	**0,541**	**0,385**	**0,739**	**0,839**

TABLE 37 – Résultats de la prédiction MLM sans conditionnement pour chaque type de relation lexicale

MLM avec conditionnement	**MRR**	**Précision**		
		TOP1	TOP5	TOP10
ANTI	0,358	0,195	0,567	0,747
Hyperonyme	**0,614**	**0,448**	0,814	**0,938**
Hyponyme	0,449	0,286	0,652	0,874
QSYN	0,604	0,421	**0,843**	0,928

TABLE 38 – Résultats de la prédiction MLM avec conditionnement pour chaque type de relation lexicale

Influence de la longueur des contextes sur la qualité de la prédiction

Nous avons par ailleurs observé que la longueur des contextes [1] influe sur les prédictions. Afin de vérifier cette observation, nous avons regroupé les contextes en fonction de leur longueur L puis nous avons calculé la moyenne P des rangs des mots sémantiquement reliés aux mots masqués dans les prédictions. Plus la moyenne P est petite, plus les prédictions des modèles sont pertinentes. Les chiffres présentés dans le Tableau 39 montrent que plus les contextes sont longs, meilleure est la prédiction. Cela pourrait être expliqué par le fait que les contextes longs sont plus informatifs. Rappelons que la longueur maximum des contextes extraits est de 100 mots. Par ailleurs, un contexte trop long peut éventuellement contenir un grand nombre de mots sans lien avec le mot cible et dégrader la prédiction.

	$L \leq 20$	$20 < L \leq 60$	$L > 60$
P	33 614	32 263	30 318

TABLE 39 – Position moyenne des mots sémantiquement reliés parmi les candidats pour différentes longueurs de contextes

Influence de la position du mot cible dans le contexte sur la prédiction

La position du mot cible dans le contexte est un autre facteur dont nous avons souhaité étudier l'effet. Nous avons regroupé les contextes selon la position PT du mot masqué dans les contextes et calculé la moyenne P des rangs des mots sémantiquement reliés au mot masqué dans les prédictions. Les chiffres du Tableau 40 suggèrent que les prédictions sont plus pertinentes lorsque le mot masqué se trouve à la fin du contexte. Cette observation rejoint l'un des critères de Kilgarriff *et al.* (2008) qui stipule que dans les bons contextes, le mot cible apparaît à la fin car son sens est ainsi est mieux spécifié.

	$PT \leq \frac{1}{3}$	$\frac{1}{3} < PT \leq \frac{2}{3}$	$PT < \frac{2}{3}$
P	34 611	35 679	32 111

TABLE 40 – Position moyenne des mots sémantiquement reliés parmi les candidats pour différentes positions du mot masqué

8.2 Méthode par analogie

8.2.1 Données de test

Le jeu de données de test de l'expérience est DataProj qui pour rappel contient 180 quadruplets. Le format de données est illustré dans le Tableau 35.

1. La longueur d'un contexte est estimée par le nombre de mots qu'il contient.

8.2.2 Expériences

Nous avons réalisé une expérience similaire à celle que nous avons présentée en Section 7.2.2. Pour les relations symétriques ANTI et QSYN, chaque quadruplet génère deux équations analogiques ; les résultats sont la moyenne des rangs des TMM inconnus dans les prédictions pour les deux équations. Pour les relations orientées (hyperonymie et hyponymie), seule une équation est produite ; les résultats sont fournis séparément pour chacune des deux relations.

Vocabulaire

Pour cette expérience, le rang de la solution est calculé relativement à un ensemble de 1 809 TMM nominaux composé des 386 bitermes de DataProj et des 1 423 bitermes nominaux de DicoEnviro qui apparaissent dans le corpus PANACEA [2], comme par exemple *biomasse forestière*.

Annotation du corpus

Comme précédemment, nous avons construit un modèle FastText qui contient les représentations des TS et celles des TMM. Pour ce faire, nous avons annoté toutes les occurrences des bitermes dans le corpus (cf. Section 7.2.2). Nous avons par ailleurs estimé la contribution des variantes extraites par TermSuite sur la qualité des prédictions. Pour cette annotation, trois stratégies ont été testées :

1. pas de prise en compte des variantes ;
2. les variantes sont indexées comme des occurrences du terme ;
3. les variantes qui sont des bitermes sont indexées comme telles ; les variantes qui ne sont pas des bitermes sont indexées comme des occurrences du terme.

TermSuite a extrait des variantes pour 529 des 1 809 bitermes du vocabulaire ; 1 815 variantes ont ainsi été extraites.

L'annotation des variantes a été adaptée pour tenir compte de trois configurations particulières. La première est celle où deux bitermes apparaissent toujours comme des parties de la même expression comme *perte de le banquise* et *banquise arctique* que l'on ne trouve dans le corpus que dans l'expression *perte de le banquise arctique*. Dans ce cas, l'expression est considérée comme une occurrence de chacun des deux bitermes. Pour l'exemple précédent, toutes les occurrences de *perte de le banquise arctique* sont annotées comme < `perte de le banquise arctique perte_de_le_banquise banquise_arctique` >. Il arrive également qu'une variante soit une conjonction. Par exemple, *protection de le faune et de le flore* est une variante de *protection de le faune* mais aussi de *protection de le flore*. Dans ce cas, la conjonction est considérée comme une occurrence des deux bitermes. Par exemple, *acidification du sol et du eau* est annoté comme < `acidification sol eau acidification_du_sol`

2. Les bitermes présents aussi dans les 386 bitermes de DataProj ne sont pas pris en compte.

acidification_du_eau >. Dans la troisième stratégie, les variantes qui sont des bitermes sont indexées comme telles. Pour vérifier si une variante est un biterme ou ne l'est pas, nous avons extrait les variantes composées de deux mots lexicaux en nous appuyant sur les patrons syntaxiques fournis par TermSuite puis nous les avons recherchées manuellement dans les banques terminologiques TERMIUM Plus, GDT et IATE. Les variantes présentes dans l'une de ces banques sont considérées comme des bitermes. 60 variantes ont ainsi été sélectionnées.

8.2.3 Résultats et analyses

Résultats globaux

Le Tableau 41 présentent les résultats obtenus lorsque les variantes ne sont pas prises en compte. Les scores MRR pour des relations ANTI que QSYN sont très satisfaisants. Ces résultats rejoignent l'observation de Ferret (2021). Si les relations hiérarchiques sont légèrement plus difficiles à capter, la précision à TOP5 et TOP10 témoigne de l'intérêt de la méthode par analogie pour l'acquisition de relations lexicales entre TMM.

	MRR	**Précision**		
		TOP1	**TOP5**	**TOP10**
ANTI	0,720	0,590	0,926	0,967
Hyperonymie	0,613	0,423	0,808	0,923
Hyponymie	0,579	0,346	**0,962**	**1,000**
QSYN	**0,793**	**0,697**	0,937	0,958

TABLE 41 – Résultats de la méthode par analogie sans prise en compte des variantes

Influence des variantes

Pour estimer l'effet de la prise en compte des variantes des TMM, nous avons comparé les résultats de la méthode par analogie dans trois modèles : Le premier (Modèle 1) entraîné sans prise en compte des variantes ; le deuxième (Modèle 2) entraîné en considérant les variantes comme des occurrences des termes ; le troisième (Modèle 3) entraîné en prenant en compte les variantes qui sont des bitermes et en annotant les autres comme des termes de référence. Le Tableau 42 présente les scores de MRR obtenus pour les trois modèles.

	Modèle 1	**Modèle 2**	**Modèle 3**
MRR	0,726	0,746	0,735

TABLE 42 – Résultats de la méthode par analogie pour les différentes stratégies de prise en comptes des variantes des TMM

La prise en compte des variantes comme occurrences des termes permet d'améliorer légèrement les scores de MRR de la méthode par analogie (Modèle 2). Une explication possible serait

que les représentations vectorielles du Modèle 2 sont de meilleure qualité. On voit aussi que l'amélioration est plus faible pour le Modèle 3 car seule une partie de variantes sont considérées comme occurrences des termes.

8.3 Discussion

Nous comparons dans cette section les méthodes et les résultats que nous avons obtenus pour l'acquisition de synonymie entre les TMM de DataIATE et celles des autres relations lexicales entre les TMM de DataProj. La comparaison est ensuite étendue aux résultats de travaux existants.

8.3.1 Comparaison des méthodes et de leurs résultats

La principale différence entre les méthodes par substitution lexicale et par analogie est que les prédictions du MLM sont fortement dépendantes du contexte contrairement aux prédictions analogiques qui sont indépendantes du contexte car les modèles FastText ne sont pas contextuels. Ces derniers sont de ce fait mieux adapté à l'identification de relations lexicales. Cette observations rejoint celles de Peters *et al.* (2018b) qui montrent que les plongements de mots générés par des modèles de langue contextuels sous-performent par rapport aux modèles statiques sur les tâches d'identification de relations sémantiques par analogie. Ces résultats suggèrent par ailleurs que la composition sémantique des TMM est mieux captée par l'analogie qu'elle ne l'est pas les MLM. Ils conforment ainsi aux conclusions de Hupkes *et al.* (2020) que les modèles de type Transformer ont un faible niveau de généralisation compositionnelle.

La supériorité de la méthode par analogie s'observe notamment pour la synonymie. La méthode par analogie obtient de meilleurs résultats que la méthode par substitution lexicale sur les jeux de données DataIATE et sur la partie QSYN de DataProj. Une explication possible est que les couples QSYN ont été créés par projection sémantique et sont donc plus fortement contraints : les TMM partagent la même structure syntaxique et les deux TS qu'ils contiennent sont sémantiquement reliés. De plus, l'analogie permet d'obtenir de meilleurs résultats.

Afin de savoir s'il existe une complémentarité entre les méthodes par analogie et par substitution lexicale nous avons analysé qualitativement leurs résultats à TOP5 pour les requêtes générées par les 147 quadruplets dans DataProj dans lesquels les deux TS appartiennent au vocabulaire du modèle (Tableau 43). Nous avons ainsi pu observer que les deux méthodes sont relativement complémentaires. Sur les 10 quadruplets pour lesquels la solution fournie par la méthode par analogie se situe au-delà de la 5^{e} position, 6 ont une solution qui fait partie du Top5 pour la méthode par substitution lexicale. De même, sur les 66 quadruplets pour lesquels la solution fournie par la méthode par substitution lexicale se situe au-delà de la 5^{e} position, 62 ont une solution qui fait partie du Top5 pour la méthode par analogie. Cependant, la taille de notre jeu de données n'est pas suffisante pour nous permettre de caractériser effectivement les

quadruplets sur lesquels les modèles échouent.

	ANA > 5	**ANA** ≤ 5	**MLM** > 5	**MLM** ≤ 5
ANA > 5	10	0	4	6
ANA ≤ 5	0	137	62	75
MLM > 5	4	62	66	0
MLM ≤ 5	6	75	0	81

TABLE 43 – Nombres de requêtes pour lesquelles le terme cible apparaît à un rang inférieur ou égal à 5 ou supérieur à 5 parmi les résultats fournis par les méthodes par analogie (ANA) et par substitution lexicale (MLM)

8.3.2 Comparaison de nos résultats avec ceux des autres

Les résultats que nous avons obtenus avec la méthode par substitution lexicale peuvent être comparés à ceux de Schick et Schütze (2020). Les méthodes utilisées sont proches et les relations considérées sont similaires. La principale différence avec leur travail est que notre étude est menée dans un domaine spécialisé. Les scores MRR de Schick et Schütze (2020) sont meilleurs que les nôtres pour l'antonymie mais plus faibles pour l'hyperonymie : 0,570 et 0,462 respectivement alors que les nôtres sont de 0,358 et 0,614. L'antonymie est en effet difficile à capturer par la méthode par substitution lexicale. Les bons résultats obtenus par Schick et Schütze (2020) pour l'antonymie peuvent probablement s'expliquer par le fait que les relations lexicales sont exprimées à l'aide de patrons prédéfinis tandis que nous utilisons des requêtes construites à partir de contextes extraits du corpus. Ces patrons guident la prédiction des antonymes de manière plus précise que les contextes car ils contiennent des expressions négatives explicites. Inversement, les contextes extraits un corpus fournissent plus d'informations sur le mot masqué pour la prédiction de la synonymie et de l'hyperonymie, en particulier parce qu'ils sont plus longs.

Les résultats de la méthode par analogie peuvent quant à eux être comparés à ceux de Chaudhri *et al.* (2022) qui utilisent un modèle Seq2Seq et un modèle Seq2Vec avec une couche linéaire pour résoudre les équations analogiques. Ces méthodes sont relativement complexes en comparaison de la nôtre qui utilise une simple différence vectorielle. De plus, les travaux de Chaudhri *et al.* (2022) se concentrent sur des relations spécifiques au domaine de la biologie, qui sont différentes des relations sémantiques lexicales qui nous intéressent. Au-delà de ces différences, nous pouvons néanmoins comparer globalement leurs méthodes aux nôtres. Le meilleur score de précision à Top 1 obtenu par Chaudhri *et al.* (2022) pour les équations analogique contenant des TMM est de 0,51 (avec le modèle Seq2Vec ELMo) ce qui est nettement inférieur à celui que nous avons obtenu qui est de 0,697. L'utilisation d'une méthode similaire à celle que nous proposons pourrait donc fortement améliorer leurs résultats.

Nos résultats peuvent également être comparés à ceux de Hazem et Daille (2018) pour la synonymie. Hazem et Daille (2018) proposent différentes méthodes pour capturer la synonymie

entre 54 couples de TMM dans les domaines de l'éolien et du cancer du sein. Ils ont obtenu une MAP de 0,349 pour les TMM français du domaine de l'éolien en utilisant une méthode par projection sémantique et un modèle word2vec afin de tirer profit de leurs représentations distributionnelles. Cette valeur est proche de celle que nous avons obtenue sur les données de test extraites de IATE en utilisant la méthode par substitution lexicale (le score MRR est de 0,374). Cependant, ce score est bien inférieur à celui obtenu par la méthode par analogie dont le score MRR est de 0,744. D'autre part, nous avons comme eux utilisé la projection sémantique pour créer des couples de TMM en relation QSYN. Rappelons que QSYN est principalement composée de synonymes. La précision que nous avons obtenue pour les couples de QSYN est de 0,60. Elle est nettement plus élevée que le score MAP de 0,25 de Hazem et Daille (2018). Une raison qui peut expliquer ces différences est le fait que les TMM que nous considérons sont dans les mêmes relations que les termes simples qu'ils contiennent, alors que Hazem et Daille (2018) ont utilisé des relations autres que la synonymie pour inférer la synonymie entre TMM.

8.3.3 Observations

Nous avons vu en Section 8.1 que les résultats de la méthode par substitution qui utilise des requêtes avec conditionnement sont meilleurs pour l'hyperonymie que pour la synonymie. Sachant que les requêtes avec conditionnement favorisent normalement les réponses sémantiquement proches du mot masqué, on aurait pu s'attendre à ce que le conditionnement favorise davantage la synonymie. Or, c'est le score MRR pour l'hyperonymie qui augmente le plus. Cela est dû en partie au fait que les contextes des couples en relation d'hyperonymie sont moins informatifs que ceux des couples de synonymes. Pour le vérifier, nous avons analysé les 641 contextes utilisés pour prédire les 16 couples en relation d'hyperonymie de DataProj qui font partie du vocabulaire du modèle. Nous avons remarqué que ces contextes étaient généralement relativement pauvres : ils sont courts et contiennent peu de connaissances linguistiques. Ces requêtes sont donc celles qui bénéficient le plus de la stratégie conditionnement.

Nous avons également observé que les résultats de la méthode par analogie sont meilleurs pour les relations symétriques que pour les relations asymétriques. Deux raisons peuvent expliquer cette différence. La première est l'utilisation d'une distance symétrique (cosinus) qui est probablement moins adaptée aux relations asymétriques. Une solution possible pour améliorer les prédictions pour les relations orientées serait d'utiliser une mesure de similarité orientée comme WeedsPrec. Lenci et Benotto (2012) montrent en effet que les mesures de similarité orientées (*directional similarity measures*) sont plus performantes que le cosinus pour l'identification des hyperonymes. La deuxième raison est que la différence entre deux plongements de mots ne représente pas précisément la relation qui existe entre les TMM. La relation n'est identifiée que de façon approximative. Cette approximation désavantage davantage les relations asymétriques que les relations symétriques.

8.4 Synthèse

Nous avons présenté dans ce chapitre un ensemble d'expériences dans lesquelles nous avons appliqué nos méthodes aux relations sémantiques entre TMM de DataProj. Nous avons d'abord évalué la capacité de modèle MLM à identifier ces relations lexicales. Les résultats montrent que la méthode par substitution lexicale est capable de capter des propriétés relationnelles, principalement QSYN avec un score de MRR de 0,604. Nous avons observé que plus la taille des contextes est grande, meilleurs sont les résultats et que la position du mot cible dans le contexte a un impact sur la qualité des prédictions. Nous avons ensuite appliqué la méthode par analogie aux mêmes données. Nous avons à nouveau remarqué que QSYN est la relation la plus facile à identifier avec un score de MRR de 0,720. Nous avons aussi étudié l'impact des variantes lors de la création des plongements de termes et vu que la qualité des plongements est améliorée légèrement lorsqu'elles sont prises en compte : le score MRR augmente de 0,726 à 0,746.

La comparaison des méthodes et des résultats nous a permis de mettre en lumière le fait que l'analogie fondée sur des modèles indépendants du contexte permet d'obtenir des résultats meilleurs que ceux de la méthode par substitution lexicale dont les prédictions sont fondées sur des modèles contextuels. Cette différence pourrait également être expliquée par le fait que la composition sémantique serait mieux captée par l'analogie que par les modèles MLM. Par ailleurs, les résultats obtenus pour la partie QSYN de DataProj sont meilleurs que ceux obtenus avec les synonymes de DataIATE du fait de la meilleure compositionnalité des TMM de DataProj. Nous avons proposé une explication au fait que les résultats obtenus en utilisant la stratégie avec conditionnement qui favorise l'hyperonymie et au fait que l'analogie capte mieux les relations symétriques. Nous avons aussi comparé nos résultats avec ceux obtenus dans les travaux proches et vu que nos résultats sont globalement meilleurs.

Conclusion et perspectives

Conclusion

Cette thèse a pour objet l'étude de la capacité des modèles sémantiques distributionnels (MSD) à capter différentes relations sémantiques entre termes multi-mots (TMM) au moyen de deux méthodes, la substitution lexicale et l'analogie. Cette étude a porté sur les relations entre bitermes nominaux du domaine de l'environnement. Ce travail est réalisé dans le cadre du projet ADDICTE financé par l'ANR (ANR-17-CE23-0001). Deux hypothèses ont guidé notre travail : (*i*) les relations lexicales entre TMM peuvent être identifiées à partir des représentations des MSD ; (*ii*) les MSD modélisent partiellement la composition sémantique des TMM.

La première partie de la thèse est consacrée à l'état de l'art. Le chapitre 1 présente quelques-unes des notions fondamentales de terminologie : termes, variantes, structure syntaxique des TMM. Nous y abordons d'autre part la notion de compositionnalité qui est centrale dans la thèse notamment parce que la très grande majorité des TMM ont un sens compositionnel. Cette observation est à l'origine de notre seconde hypothèse qui fonde d'une part la construction par projection sémantique du jeu de données DataProj et les méthodes d'acquisition des relations sémantiques que nous avons proposées. La structuration des terminologies par les relations sémantiques, et en particulier les relations paradigmatiques, est discutée dans le chapitre 2. Le chapitre présente d'autre part quatre ressources terminologiques structurées. Trois d'entre elles sont utilisées dans notre travail : Termium Plus, DiCoEnviro et IATE. Le chapitre 3 aborde les modèles distributionnels sémantiques. Nous y rappelons l'hypothèse distributionnelle qui fonde ces modèles puis nous décrivons les trois types principaux de MSD : les modèles fréquentiels ; les modèles neuronaux statiques ; les modèles neuronaux contextuels. Nous discutons également dans ce chapitre de l'adaptation de ces modèles aux domaines de spécialité. Les méthodes d'acquisition automatique de relations sémantiques sont l'objet du chapitre 4. Nous y abordons notamment les travaux qui utilisent la projection sémantique, la substitution lexicale et l'analogie qui sont trois méthodes principales utilisées dans notre travail.

Les expériences présentées dans la troisième partie de la thèse s'appuient sur deux jeux de données, DataIATE et DataProj, respectivement présentés dans les chapitres 5 et 6. Nous utilisons également pour réaliser ces expériences le corpus monolingue français PANACEA qui sert à la création de DataProj, à la construction de modèles distributionnels statiques et à la sélection de contextes pour l'interrogation des modèles de langue masqués. Le jeu de données

DataIATE a été extrait de la banque terminologique IATE. Il se compose de 928 couples de bitermes nominaux synonymes du domaine de l'environnement. L'une des contributions de la thèse est le jeu de données DataProj que nous avons construit par la projection sémantique à partir des couples de termes simples de DiCoEnviro. DataProj contient 180 quadruplets de type $TS_1 : TS_2 : TMM_1 : TMM_2$ où TS_1 et TS_2 sont deux TS qui sont des composants de deux TMM TMM_1 et TMM_2 respectivement et tels que la relation entre TS_1 et TS_2 et celle entre TMM_1 et TMM_2 sont identiques. Pour construire DataProj, nous avons utilisé TermSuite pour extraire des candidats termes du domaine de l'environnement à partir du corpus PANACEA. Trois types de relations lexicales, QSYN, ANTI et HYP, entre les termes simples de DicoEnviro ont été étendus aux candidats bitermes extraits. Une annotation du statut de terme des TMM candidats et de la relation sémantique inférée a été réalisée. L'analyse de la préservation de ces relations fait apparaître que la plupart des TMM ont un sens compositionnel, confirmant ainsi notre seconde hypothèse et l'observation de L'Homme (2004).

Les expériences que nous avons réalisées sont présentées dans les chapitres 7 et 8. Le chapitre 7 présente les deux méthodes que nous avons utilisées pour identifier les relations lexicales entre TMM : la substitution lexicale et l'analogie. Il décrit également les modèles sémantiques distributionnels utilisés et les méthodes d'évaluation des résultats des expériences. Les expériences présentées dans le chapitre 7 portent sur l'acquisition de la synonymie entre TMM et utilisent le jeu de données DataIATE. Les expériences du chapitre 8 reprennent celles du chapitre 7 en utilisant DataProj au lieu de DataIATE.

Les résultats de ces expériences font apparaître que la substitution lexicale fondée sur les MLM et l'analogie dans des modèles FastText sont globalement performantes sur les deux jeux de données. Cela confirme nos hypothèses. Par ailleurs, les méthodes par analogie sont plus performantes que celles par substitution lexicale pour la synonymie, l'antonymie et l'hyponymie. Les meilleurs résultats obtenus avec la méthode par analogie ont un score MRR de 0,793 pour la synonymie, de 0,720 pour l'antonymie, 0,613 pour l'hyperonymie et 0,579 pour l'hyponymie. Nous avons aussi mis en évidence que les modèles dans lesquels les variantes sont considérées comme des occurrences des TMM améliorent les performances de la méthode par analogie.

L'une des conclusions de la thèse est que la combinaison de la projection sémantique fondée sur l'hypothèse de compositionnalité des TMM et et les approches distributionnelles est adapté à l'identification des relations sémantiques entre ces dernier. Par ailleurs, nos résultats suggèrent que la composition sémantique présente dans les plongements statiques est bien captée par l'analogie ; à l'inverse, le niveau de généralisation compositionnelle des modèles de type Transformer semblent plus faible.

Perspectives

Nous envisageons de poursuivre l'étude à laquelle cette thèse est consacrée dans plusieurs directions. La première sera l'amélioration de la qualité des contextes afin d'augmenter les

performance de la méthode par substitution lexicale dont les résultats sont très inférieurs aux ceux obtenus par la méthode d'analogie. Nous avons en effet pu observer le fort impact des contextes sur les prédictions des MLM. L'utilisation de corpus de plus grande taille permettra la sélection des contextes qui satisfont davantage de critères de Kilgarriff *et al.* (2008) comme le fait que les bons contextes ne doivent pas contenir des pronoms et des anaphores. Une autre piste pour l'amélioration des candidats fournis pas les MLM consiste à sélectionner des contextes dans lesquels les TMM ont un sens fin qui correspond à celui de la relation que l'on cherche à acquérir. Pour ce faire, il sera nécessaire d'utiliser des méthodes de désambiguïsation en contexte du sens des expressions poly-lexicales, similaires à celle de Hey *et al.* (2021).

Nous prévoyons également d'utiliser des modèles génératifs comme GPT-3 (Brown *et al.*, 2020) au lieu des MLM pour réaliser la tâche de substitution lexicale. Les modèles génératifs sont des modèles de langue qui permettent de générer du texte. La substitution lexicale deviendrait ainsi une tâche de génération de textes de langue naturelle. Cette voie a déjà été explorée par Lee *et al.* (2021) qui ont comparé les modèles de type BERT et les modèles génératifs sur une tâche de substitution lexicale en anglais. Dans cette étude, les scores de F-mesure à TOP 10 des modèles génératifs sont plus élevés que ceux des modèles de type BERT lorsque les candidats considérés ne sont pas limités à ceux de la liste de référence.

Une troisième piste consiste à améliorer les capacités de généralisation compositionnelle des modèles Transformer étant donné la nature compositionnelle du sens des TMM. Nous pourrons nous inspirer du travail de Ontanon *et al.* (2022) qui montre que des gains significatifs dans la généralisation compositionnelle peuvent être obtenus en adaptant la configuration architecturale des modèles et notamment leur taille et le type du décodeur.

Nous envisageons d'étendre ce travail aux TMM composés de plus de 2 mots lexicaux. Les TMM en relation sémantique ne sont en effet pas toujours de même longueur et peuvent avoir des structures syntaxiques différentes comme *audit* et *contrôle des comptes*. Cette extension pourrait être réalisée en utilisant des modèles de graphes qui incluent des informations relationnelles comme dans le travail de Xu *et al.* (2021).

D'autre part, la compositionnalité des TMM qui est confirmée par les travaux précédents (L'Homme, 2004 ; Hazem et Daille, 2018) est une hypothèse importante dans notre travail à la fois pour la construction du jeu de données et notre méthode de substitution lexicale. Cependant, nous avons observé dans les données de IATE que le niveau de compositionnalité des TMM n'est pas identique. Par exemple, *accord volontaire* peut être utilisé comme un synonyme de *accord environnemental*. Dans ce cas, son sens ne peut pas être calculé directement à partir de ceux de *accord* et de *volontaire*. Nous souhaiterions étudier la capacité des modèles comme BERT et XLNet à capter ces sens non-compositionnels. Plus généralement, l'étude sera élargie à deux nouvelles questions : (*i*) Quelles sont les caractéristiques linguistiques des TMM qui ont un impact sur leur niveau de compositionnalité ? Nous envisageons notamment d'étudier la relation de sens entre les constituants des TMM. (*ii*) Les MSD sont-ils capables de capter les différents niveaux de compositionnalité des TMM ?

Table des figures

Liste des tableaux

Bibliographie

Abacha, A. B. et Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of Biomedical Semantics*, 2 :S4 – S4.

Agichtein, E. et Gravano, L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA. Association for Computing Machinery.

Amrami, A. et Goldberg, Y. (2019). Towards better substitution-based word sense induction.

Amsili, P. et Bras, M. (1998). Drt et compositionnalité. *Traitement Automatique des Langues*, 39(1) :131–160.

Arefyev, N., Sheludko, B., Podolskiy, A., et Panchenko, A. (2020). A comparative study of lexical substitution approaches based on neural language models. *arXiv preprint arXiv :2006.00031*.

Auger, A. et Barrière, C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, 14(1) :1.

Baldwin, T., Bannard, C., Tanaka, T., et Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions : analysis, acquisition and treatment-Volume 18*, MWE '03, pages 89–96, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Barbella, D. et Forbus, K. (2013). Analogical word sense disambiguation. *Advances in Cognitive Systems*, 2(1) :297–315.

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

Baroni, M. et Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK.

Barrière, C. (2004). Knowledge-rich contexts discovery. In *Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI*, pages 187–201, London, Ontario, Canada. Springer.

Beltagy, I., Lo, K., et Cohan, A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Berland, M. et Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, page 57–64, USA. Association for Computational Linguistics.

Bernier-Colborne, G. (2016). *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*. PhD thesis, Université de Montréal, Montréal, Canada.

Bernier-Colborne, G. et Drouin, P. (2016). Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique (evaluation of distributional semantic models : The case of syntactic derivation). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs)*, pages 125–138, Paris, France. AFCP - ATALA.

Bojanowski, P., Grave, E., Joulin, A., et Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5 :135–146.

Botta, M. (2013). La terminologie de l'environnement en vulgarisation scientifique : la famille lexicale de la régénération des forêts en portugais. *Equivalences*, 40(1) :277–298.

Bouraoui, Z., Camacho-Collados, J., et Schockaert, S. (2020). Inducing relational knowledge from bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463, New York, USA. AAAI Press.

Bouraoui, Z., Jameel, S., et Schockaert, S. (2018). Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bourigault, D. (2002). Upery : Un outil d'analyse distributionnelle Étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 75–84, Nancy, France. ATALA.

Bowker, L. et Hawkins, S. (2006). Variation in the organization of medical terms : Exploring some motivations for term choice. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(1) :79–110.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., et Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., et Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Camacho-Collados, J., Espinosa Anke, L., et Schockaert, S. (2019). Relational word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3286–3296, Florence, Italy. Association for Computational Linguistics.

Chaudhri, V. K., Xu, J., Aung, H. L., et Weerawardhena, S. (2022). *A Corpus of Biology Analogy Questions as a Challenge for Explainable AI*, pages 327–337. Springer International Publishing, Cham.

Chen, Z., He, Z., Liu, X., et Bian, J. (2018). Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC medical informatics and decision making*, 18(2) :53–68.

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing, London,England.

Christou, D. et Tsoumakas, G. (2021). Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9 :62574–62582.

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., et Darrell, T. (2016). Best practices for fine-tuning visual classifiers to new domains. In *ECCV Workshops*, pages 435–442, Amsterdam, The Netherlands. University of Amsterdam.

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

Clark, S. (2015). *Vector space models of lexical meaning*, pages 493–522. Wiley Online Library, Hoboken, New Jersey, USA.

Claveau, V. et L'Homme, M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie. Utilisation comparée de ressources endogènes et exogènes. In *Actes de la conférence terminologie et intelligence artificielle (TIA-2005)*, Rouen.

Cram, D. et Daille, B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Crouse, M., McFate, C., et Forbus, K. (2018). Learning from unannotated qa pairs to analogically disambiguate and answer questions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, volume 32 of *AAAI'18/IAAI'18/EAAI'18*, pages 654–662. AAAI Press.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, United Kingdom.

Csurka, G. (2017). Domain adaptation for visual applications : A comprehensive survey. *arXiv preprint arXiv :1702.05374*, abs/1702.05374.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 11(1) :181–197.

Daille, B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Daille, B. et Blancafort, H. (2013). Knowledge-poor and knowledge-rich approaches for multilingual terminology extraction. *Res. Comput. Sci.*, 70 :173–186.

Daille, B., Habert, B., Jacquemin, C., et Royauté, J. (1996). Empirical observation of term variations and principles for their description. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2) :197–257.

Deepika, S. et Geetha, T. (2021). Pattern-based bootstrapping framework for biomedical relation extraction. *Engineering Applications of Artificial Intelligence*, 99 :104–130.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6) :391–407.

Depraetere, I. (2019). Meaning in context and contextual meaning : A perspective on the semantics-pragmatics interface applied to modal verbs. *Anglophonia. French Journal of English Linguistics*, 28.

Devlin, J., Chang, M.-W., Lee, K., et Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Drouin, P. et Langlais, P. (2006). Évaluation du potentiel terminologique de candidats termes. *Actes des 8e Journées internationales d'analyse statistique des données textuelles (JADT-2006)*, pages 379–388.

Espinosa Anke, L., Codina-Filba, J., et Wanner, L. (2021). Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 1406–1417, Online. Association for Computational Linguistics.

Espinosa-Anke, L. et Schockaert, S. (2018). SeVeN : Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Evert, S. (2010). Distributional semantic models. In *NAACL HLT 2010 Tutorial Abstracts*, pages 15–18, Los Angeles, California. Association for Computational Linguistics.

Ferret, O. (2021). Exploration des relations sémantiques sous-jacentes aux plongements contextuels de mots (exploring semantic relations underlying contextual word embeddings). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 26–36, Lille, France. ATALA.

Ferrucci, D. et Lally, A. (2004). Uima : an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4) :327–348.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-55.*, volume 1952-59, pages 1–32. The Philological Society, Oxford.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378.

Frege, G. *et al.* (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1) :25–50.

Freixa, J. (2006). Causes of denominative variation in terminology : A typology proposal. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(1) :51–77.

Fromkin, V., Rodman, R., et Hyams, N. (2018). *An introduction to language*. Cengage Learning, Boston, Massachusetts, USA.

Gladkova, A., Drozd, A., et Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings : What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California.

Goldberg, Y. et Nivre, J. (2013). Training deterministic parsers with non-deterministic oracles. *Transactions of the Association for Computational Linguistics*, 1 :403–414.

Goodfellow, I., Bengio, Y., et Courville, A. (2016). *Deep learning*. MIT press, Cambridge, MA.

Grabar, N. et Hamon, T. (2006). Terminology structuring through the derivational morphology. In *International Conference on Natural Language Processing (in Finland)*, pages 652–663, Berlin Heidelberg. Springer.

Grefenstette, G. (2012). *Explorations in automatic thesaurus discovery*, volume 278. Springer Science & Business Media, Berlin.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., et Smola, A. J. (2006). A kernel method for the two-sample-problem. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, page 513–520, Cambridge, MA, USA. MIT Press.

Guu, K., Hashimoto, T. B., Oren, Y., et Liang, P. (2018). Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6 :437–450.

Hakenberg, J., Leaman, R., Vo, N., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C., et Gonzalez, G. (2010). Efficient extraction of protein-protein interactions from full-text articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7 :481–494.

Hamilton, W. L., Clark, K., Leskovec, J., et Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 595, Austin, Texas, USA. NIH Public Access.

Hamon, T. et Nazarenko, A. (2001). Detection of synonymy links between terms : experiment and results. *Recent advances in computational terminology*, 2 :185–208.

Hamon, T., Nazarenko, A., et Gros, C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 498–504, Montreal, Quebec, Canada. Association for Computational Linguistics.

Harper, K. E. (1965). Measurement of similarity between nouns. In *COLING 1965*, New York, NY, USA. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3) :146–162.

Hazem, A. et Daille, B. (2015). Méthode semi-compositionnelle pour l'extraction de synonymes des termes complexes. *Traitement Automatique des Langues*, 56(2) :51–76.

Hazem, A. et Daille, B. (2018). Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 297–303, Miyazaki, Japan. European Language Resources Association (ELRA).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 2 of *COLING '92*, page 539–545, USA. Association for Computational Linguistics.

Hey, T., Keim, J., et Tichy, W. F. (2021). Knowledge-based sense disambiguation of multiword expressions in requirements documents. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 70–76, Online. IEEE.

Hmida, F., Morin, E., et Daille, B. (2015). Extraction de contextes riches en connaissances en corpus spécialisés. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 109–115, Caen, France. ATALA.

Hou, J., Li, X., Yao, H., Sun, H., Mai, T., et Zhu, R. (2020). Bert-based chinese relation extraction for public security. *IEEE Access*, 8 :132367–132375.

Hupkes, D., Dankers, V., Mul, M., et Bruni, E. (2020). Compositionality decomposed : How do neural networks generalise ? *Journal of Artificial Intelligence Research*, 67 :757–795.

Ide, N. et Véronis, J. (1995). *Text encoding initiative : Background and contexts*, volume 29. Springer Science & Business Media, USA.

Jackendoff, R. (1976). Toward an explanatory semantic representation. *Linguistic inquiry*, 7(1) :89–150.

Kamel, M. et Aussenac-Gilles, N. (2009). Construction automatique d'ontologies à partir de spécifications de bases de données. In *IC 2009 : 20es Journées Francophones d'Ingénierie des Connaissances «Connaissance et communautés en ligne»*, pages 85–96, Hammamet, Tunisia. Irit.

Kang, G., Jiang, L., Yang, Y., et Hauptmann, A. G. (2019). Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, Long Beach, CA, USA. IEEE.

Karouzos, C., Paraskevopoulos, G., et Potamianos, A. (2021). UDALM : Unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2579–2590, Online. Association for Computational Linguistics.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., et Rychlý, P. (2008). Gdex : Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, pages 425–432, Barcelona, Spain. Documenta Universitaria.

Köper, M., Scheible, C., et Schulte im Walde, S. (2015). Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th international conference on computational semantics*, pages 40–45, London, UK. Association for Computational Linguistics.

Kudo, T. et Richardson, J. (2018). Sentencepiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lafourcade, M. et Ramadier, L. (2016). Semantic relation extraction with semantic patterns experiment on radiology reports. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4578–4582, Portorož, Slovenia. European Language Resources Association (ELRA).

Lai, P.-T. et Lu, Z. (2020). Bert-gt : cross-sentence n-ary relation extraction with bert and graph transformer. *Bioinformatics*, 36(24) :5678–5685.

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

Landauer, T. K., Foltz, P. W., et Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3) :259–284.

Landis, J. R. et Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Langlais, P., Yvon, F., et Zweigenbaum, P. (2009). Improvements in analogical learning : Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 487–495, Athens, Greece. Association for Computational Linguistics.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., et Schwab, D. (2020). Flaubert : des modèles de langue contextualisés préentraînés pour le français. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278. ATALA ; AFCP.

Lee, C.-H., Khoo, C., et Na, J.-C. (2004). Automatic identification of treatment relations for medical ontology learning : An exploratory study. *Advances in knowledge organization*, 9(2004) :245.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et Kang, J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240.

Lee, M., Donahue, C., Jia, R., Iyabor, A., et Liang, P. (2021). Swords : A benchmark for lexical substitution with improved data coverage and quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4362–4379, Online. Association for Computational Linguistics.

Lefeuvre, L. (2017). *Analyse des marqueurs de relations conceptuelles en corpus spécialisé : recensement, évaluation et caractérisation en fonction du domaine et du genre textuel*. PhD thesis, Toulouse 2, Toulouse, France.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1) :1–31.

Lenci, A. et Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 75–79, Montréal, Canada. Association for Computational Linguistics.

Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., et Miliani, M. (2021). A comprehensive comparative evaluation and analysis of distributional semantic models.

Lepage, Y. (1998). Solving analogies on words : An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, volume 2, pages 728–735, Montréal.

Lepage, Y. et Ando, S. (1996). Saussurian analogy : a theoretical account and its application. In *COLING 1996 Volume 2 : The 16th International Conference on Computational Linguistics*.

Levy, O. et Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Levy, O., Goldberg, Y., et Dagan, I. (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3 :211–225.

Levy, O., Remus, S., Biemann, C., et Dagan, I. (2015b). Do supervised distributional methods really learn lexical inference relations ? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Pum, Montréal, Canada.

L'Homme, M.-C. (2020). *Lexical Semantics for Terminology : An Introduction*, volume 20. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Li, Y., Baldwin, T., et Cohn, T. (2018). What's in a domain ? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana. Association for Computational Linguistics.

Lin, D. et Pantel, P. (2001). Dirt - discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, page 323–328, New York, USA. Association for Computing Machinery.

Liu, H., Komandur, R., et Verspoor, K. (2011a). From graphs to events : A subgraph matching approach for information extraction from biomedical text. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 164–172, Portland, Oregon, USA. Association for Computational Linguistics.

Liu, K., Chapman, W., Savova, G., Chute, C., Sioutos, N., et Crowley, R. S. (2011b). Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Methods of information in medicine*, 50(05) :397–407.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., et Stoyanov, V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. *CoRR*.

Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2) :203–208.

L'Homme, M.-C. (2001). Combinaisons lexicales spécialisées. regroupement des mots clés par classes conceptuelles. *Journée d'étude de l'ATALA. La collocation. Rapport de recherche, Nantes, Institut de recherche en informatique de Nantes*, pages 19–22.

L'Homme, M.-C. (2007). *Using explanatory and combinatorial lexicology to describe terms*, pages 163–202. John Benjamins, Amsterdam / Philadelphia.

L'Homme, M.-C. (2016). Terminologie de l'environnement et sémantique des cadres. In *SHS Web of Conferences*, volume 27, page 05010, Tours, France. EDP Sciences.

L'Homme, M.-C. et Lanneville, M.-E. (2014). Dicoenviro. dictionnaire fondamental de l'environnement. *Consulté à l'adresse http ://olst. ling. umontreal. ca/cgibin/dicoenviro/search. cgi*.

Ma, X., Xu, P., Wang, Z., Nallapati, R., et Xiang, B. (2019). Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Manning, C. D., Manning, C. D., et Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press, Cambridge, MA.

Marshman, E. (2006). *Lexical knowledge patterns for semi-automatic extraction of cause-effect and association relations from medical texts : a comparative study of English and French*. PhD thesis, Université de Montréal, Montréal, Canada.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, E., Seddah, D., et Sagot, B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

McCarthy, D. et Navigli, R. (2007). SemEval-2007 task 10 : English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2 :279–302.

Mickus, T., Paperno, D., Constant, M., et van Deemter, K. (2020). What do you mean, BERT ? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., et Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., et Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Mikolov, T., Yih, W.-t., et Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics : Human language technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Monnin, C. et Hamon, O. (2018). Construction de patrons lexico-syntaxiques d'extraction pour l'acquisition de connaissances à partir du web (relation pattern extraction and information extraction from the web). In *Actes de la Conférence TALN. Volume 2 - Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT*, pages 3–16, Rennes, France. ATALA.

Morin, E. et Jacquemin, C. (1999). Projecting corpus-based semantic links on a thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, page 389–396, USA. Association for Computational Linguistics.

Morlane-Hondere, F., Fabre, C., Hathout, N., et Tanguy, L. (2014). Disambiguating distributional neighbors using a lexical substitution dataset. *Natural Language Processing and Cognitive Science*, pages 27–38.

Morlane-Hondère, F. et Fabre, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. In *3e CMLF*, volume 1, pages 1001–1015, France. EDP Sciences.

Murphy, M. L. (2003). *Semantic relations and the lexicon : Antonymy, synonymy and other paradigms*. Cambridge University Press, Cambridge, United Kingdom.

Naseem, U., Razzak, I., Khan, S. K., et Prasad, M. (2021). A comprehensive survey on word representation models : From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5) :1–35.

Ng, S.-K. et Wong, M. (1999). Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10 :104–112.

Nishida, K., Nishida, K., Saito, I., Asano, H., et Tomita, J. (2020). Unsupervised domain adaptation of language models for reading comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5392–5399, Marseille, France. European Language Resources Association.

Niwa, Y. et Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING 1994 : The 15th International Conference on Computational Linguistics*, volume 1, Kyoto, Japan. Association for Computational Linguistics.

Ontanon, S., Ainslie, J., Fisher, Z., et Cvicek, V. (2022). Making transformers solve compositional tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3591–3607, Dublin, Ireland. Association for Computational Linguistics.

Otman, G. (1995). *Les représentations sémantiques en terminologie : la modélisation des unités terminologiques sous la forme de réseaux sémantico-terminologiques*. PhD thesis, Paris 4, Paris, France.

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

Pagin, P. et Westerståhl, D. (2010). Compositionality i : Definitions and variants. *Philosophy Compass*, 5(3) :250–264.

Partee, B. (1984). Compositionality. *Varieties of formal semantics*, 3 :281–311.

Paullada, A., Percha, B., et Cohen, T. (2020). Improving biomedical analogical retrieval with embedding of structural dependencies. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 38–48, Online. Association for Computational Linguistics.

Peng, Y., Yan, S., et Lu, Z. (2019). Transfer learning in biomedical natural language processing : An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Pennington, J., Socher, R., et Manning, C. (2014). GloVe : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, C. (2005). *Multilingual Information Access for Text, Speech and Images : 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491. Springer Science & Business Media, Berlin, Germany.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Zettlemoyer, L., et Yih, W.-t. (2018b). Dissecting contextual word embeddings : Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Pitar, M. (2006). Les relations entre les concepts en terminologie. *JOURNÉES DE LA FRANCOPHONIE 2006–2007*, pages 89–98.

Polguère, A. (2016). *Lexicologie et sémantique lexicale : Notions fondamentales*. Presses de l'Université de Montréal, Montréal.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., et Wu, X. (2019). A simple bert-based approach for lexical simplification. *ArXiv*, abs/1907.06226.

Radev, D. R., Qi, H., Wu, H., et Fan, W. (2002). Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., et Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8) :9.

Ramadier, L. et Lafourcade, M. (2016). Patrons sémantiques pour l'extraction de relations entre termes - application aux comptes rendus radiologiques (here the title in English). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, pages 514–521, Paris, France. AFCP - ATALA.

Riedel, S., Yao, L., McCallum, A., et Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics.

Robert, P. et Rey, A. (2001). *Le Grand Robert de la langue française*, volume 5. Dictionnaires Le Robert, Paris, France.

Rubenstein, H. et Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10) :627–633.

Ryu, M. et Lee, K. (2020). Knowledge distillation for bert unsupervised domain adaptation. *arXiv preprint arXiv :2010.11478*.

Sablayrolles, J.-F. (2002). Fondements théoriques des difficultés pratiques du traitement des néologismes. *Revue française de linguistique appliquée*, 7(1) :97–111.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., et Flickinger, D. (2002). Multiword expressions : A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002*, pages 1–15, Mexico City, Mexico. Springer.

Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing, Amsterdam / Philadelphia.

Sager, J. C. et Ndi-Kimbi, A. (1995). The conceptual structure of terminological definitions and their linguistic realisations : A report on research in progress. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 2(1) :61–85.

Salton, G. et Lesk, M. E. (1997). *Computer Evaluation of Indexing and Text Processing*, page 60–84. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Salton, G. et McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

Scheepers, T., Kanoulas, E., et Gavves, E. (2018). Improving word embedding compositionality using lexicographic definitions. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1083–1093, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Schick, T. et Schütze, H. (2019). Attentive mimicking : Better word embeddings by attending to informative contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Schick, T. et Schütze, H. (2020). Rare words : A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8766–8774, New York, USA. AAAI Press.

Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Schuster, M. et Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152, Wuhan, China. IEEE.

Skousen, R., editor (2002). *Analogical Modeling. An exemplar-based approach to language*. Number 10 in Human Cognitive Processing. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Specia, L., Jauhar, S. K., et Mihalcea, R. (2012). SemEval-2012 task 1 : English lexical simplification. In **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Sun, B., Feng, J., et Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065. AAAI Press.

Sutskever, I., Vinyals, O., et Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 3104–3112, Montreal, Canada. MIT Press.

Tabossi, P., Fanari, R., et Wolf, K. (2008). Processing idiomatic expressions : Effects of semantic compositionality. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 34(2) :313–327.

Tenney, I., Das, D., et Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Turian, J., Ratinov, L., Bengio, Y., et Roth, D. (2009). A preliminary evaluation of word representations for named-entity recognition. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, pages 1–8, Whistler, Canada. MIT press.

Turney, P. D. (2005). Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 1136–1141, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of The 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 905–912, Manchester.

Turney, P. D., Littman, M. L., Bigham, J., et Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, Borovets, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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

Verspoor, C. M., Joslyn, C., et Papcun, G. J. (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *SIGIR workshop on Text Analysis and Search for Bioinformatics*, pages 51–56, Toronto, Canada.

Vylomova, E., Rimell, L., Cohn, T., et Baldwin, T. (2016). Take and took, gaggle and goose, book and read : Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Wang, H.-C., Chen, Y.-H., Kao, H.-Y., et Tsai, S.-J. (2011). Inference of transcriptional regulatory network by bootstrapping patterns. *Bioinformatics*, 27(10) :1422–1428.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R. B., et Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4) :414–417.

Whitney, W. D. (1867). *Language and the study of language : Twelve lectures on the principles of linguistic science*. C. Scribner, New York.

Wiedemann, G., Remus, S., Chawla, A., et Biemann, C. (2019). Does bert make any sense ? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, pages 161–170, Erlangen, Germany. Chair of Computational Corpus Linguistics.

Wilson, G. et Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5) :1–46.

Wohlgenannt, G., Chernyak, E., Ilvovsky, D., Barinova, A., et Mouromtsev, D. (2019). Relation extraction datasets in the digital humanities domain and their evaluation with word embeddings. *arXiv preprint arXiv :1903.01284*.

Wüster, E. (1974). Die allgemeine terminologielehre–ein grenzgebiet zwischen sprachwissenschaft, logik, ontologie, informatik und den sachwissenschaften. *Linguistics*, 12 :61–106.

Xu, J., Chen, Y., Qin, Y., Huang, R., et Zheng, Q. (2021). A feature combination-based graph convolutional neural network model for relation extraction. *Symmetry*, 13(8) :1458.

Yang, X., Yu, Z., Guo, Y., Bian, J., et Wu, Y. (2021). Clinical relation extraction using transformer-based models. *ArXiv*, abs/2107.08957.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., et Le, Q. V. (2019). Xlnet : Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv :1906.08237*, abs/1906.08237.

Yao, L., Mao, C., et Luo, Y. (2019). Kg-bert : Bert for knowledge graph completion. *arXiv preprint arXiv :1909.03193*.

Yu, D., Sun, K., Cardie, C., et Yu, D. (2020). Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Zhang, L., Li, J., et Wang, C. (2017). Automatic synonym extraction using word2vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*, pages 5629–5632, Dalian, China. IEEE.

Zhou, D., He, Y., et Kwoh, C. K. (2008). From biomedical literature to knowledge : mining protein-protein interactions. In Smolinski, T. G., Milanova, M. G., et Hassanien, A.-E., editors, *Computational Intelligence in Biomedicine and Bioinformatics*, pages 397–421. Springer, New York.

Zhou, W., Ge, T., Xu, K., Wei, F., et Zhou, M. (2019). Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

Zhu, Y., Yan, E., et Wang, F. (2017). Semantic relatedness and similarity of biomedical terms : examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC medical informatics and decision making*, 17(1) :1–8.

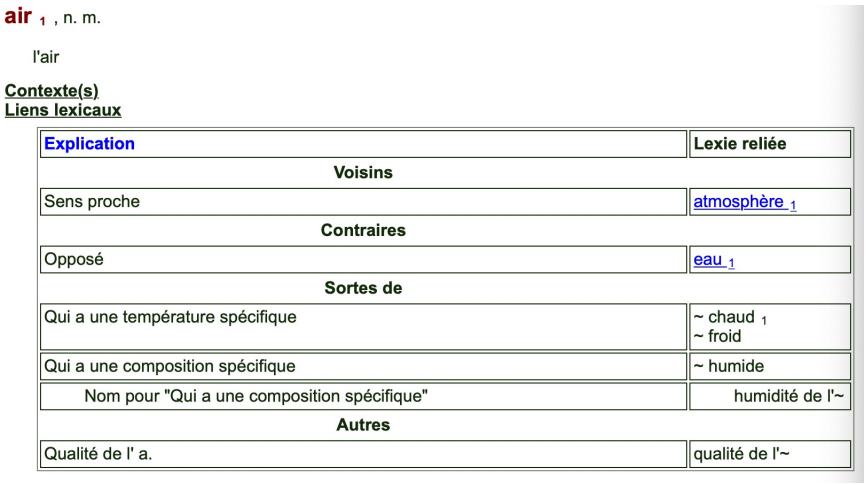
Zhuang, L., Wayne, L., Ya, S., et Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Zweigenbaum, P. et Grabar, N. (2000). Liens morphologiques et structuration de terminologie. In *IC 2000 : Ingénierie des connaissances*, pages 325–334, Toulouse, France. Irit.

Annexe A

Bitermes nominaux extraits de DiCoEnviro

Nous avons extrait des bitermes nominaux de DiCoEnviro pour construire le vocabulaire où nous cherchons la solution d'analogie et pour indexer le corpus pour les expériences. Les entrées de DiCoEnviro sont principalement des termes simples. Il n'existe que 73 bitermes nominaux qui sont les entrées de DiCoEnviro présents dans le corpus PANACEA. De ce fait, nous avons également extrait 1 354 bitermes nominaux sémantiquement reliés aux termes simples qui composent les entrées du dictionnaire. Par exemple, la Figure 25 illustre les termes sémantiquement reliés du terme *air*. Les bitermes extraits sont *air chaud*, *air foid*, *air humide*, *humidité de l'air* et *qualité de l'air*.



air 1 , n. m.

l'air

Contexte(s)
Liens lexicaux

Explication	Lexie reliée
Voisins	
Sens proche	atmosphère 1
Contraires	
Opposé	eau 1
Sortes de	
Qui a une température spécifique	~ chaud 1 ~ froid
Qui a une composition spécifique	~ humide
Nom pour "Qui a une composition spécifique"	humidité de l'~
Autres	
Qualité de l' a.	qualité de l'~

FIGURE 25 – Fiche terminologique du terme *air* dans DiCoEnviro

Annexe B

Extrait de bitermes nominaux de DiCoEnviro

matière animal
bas troposphère
augmentation du rendement
transformation de le biomasse
biome terrestre
bac brun
produit périssable
perte de diversité
faible autonomie
variation du climat
faune riche
moteur diesel
zone continental
survie de espèce
teneur en carbone
migration du population
acide faible
moteur à explosion
véhicule vert
patrimoine arboricole
bac vert
bois tendre
oiseau migrateur
marais salé
au niveau local

Table 44 continued from previous page

émission de hydrocarbure
forçage anthropique
période de reproduction
élévation de le teneur
grand cétacé
mode électrique
disparition de un espèce
variation de le pluviosité
statut précaire
fumier liquide
évolution du climat
parc éolien
combustion de combustible
teneur en ion
rétroaction positif
couverture terrestre
petit fruit
changement climatique
destruction du couvert
acide sulfurique
résidu solide
plaque océanique
abaissement du niveau
rétroaction négatif
eau ménager
constructeur automobile
biomasse organique
diminution de le pluviosité
oiseau de mer
produit jetable
insecte ravageur
combustible carboné
prolifération du algue
diminution de le vulnérabilité
faible solubilité
agriculture biologique
aménagement durable
énergie solaire

Table 44 continued from previous page

biomasse forestier
dégradation de le forêt
défrichement de un forêt
chaudière à vapeur
oiseau de proie
changement du climat
environnement aquatique
réchauffement du globe
consommation de énergie
croissance forestier
grave sécheresse
égout séparatif
au niveau mondial
espèce terrestre
îlot de végétation
culture de le luzerne
défrichement illégal
énergie géothermique
conservation de environnement
plantation agricole
particule de sulfate
plante aquatique
solaire passif
variabilité du rendement
cycle de azote
diminution du manteau
substance nocif
animal domestique
fort précipitation
déchet combustible
teneur en azote
vie sauvage
voiture diesel
électricité solaire
population citadin
petit étang
écosystème marin

Annexe C

Ensembles de variantes qui sont des conjonctions

Conjonction	**TMM1**	**TMM2**
acidification du sol et du eau	acidification du sol	acidification du eau
agriculture durable et compétitif	agriculture durable	agriculture compétitif
agriculture naturel ou biologique	agriculture naturel	agriculture biologique
agriculture paysan et biologique	agriculture paysan	agriculture biologique
alimentation humain et animal	alimentation humain	alimentation animal
alimentation humain ou animal	alimentation humain	alimentation animal
aménagement agricole et rural	aménagement agricole	aménagement rural
bilan écologique et éthique	bilan écologique	bilan éthique
capital naturel et humain	capital naturel	capital humain
catastrophe écologique et humain	catastrophe écologique	catastrophe humain
catastrophe écologique et social	catastrophe écologique	catastrophe social
changement climatique et politique	changement climatique	changement politique
changement environnemental et climatique	changement environnemental	changement climatique
climat chaud et humide	climat chaud	climat humide

Table 45 continued from previous page

Conjonction	TMM1	TMM2
conservation de le faune et de le flore	conservation de le faune	conservation de le flore
conservation de le flore et de le faune	conservation de le flore	conservation de le faune
conservation de le nature et de le biodiversité	conservation de le nature	conservation de le biodiversité
conservation du espèce et du écosystème	conservation du espèce	conservation du écosystème
conservation du espèce et du habitat	conservation du espèce	conservation du habitat
conservation du site et du monument	conservation du site	conservation du monument
conservation du sol et du eau	conservation du sol	conservation du eau
consommation de viande et de poisson	consommation de viande	consommation de poisson
consommation humain ou animal	consommation humain	consommation animal
contamination chimique ou bactérien	contamination chimique	contamination bactérien
couverture de neige et de glace	couverture de neige	couverture de glace
croissance démographique et économique	croissance démographique	croissance économique
croissance économique et démographique	croissance économique	croissance démographique
développement agricole et rural	développement agricole	développement rural
développement économique et social	développement économique	développement social
développement économique et touristique	développement économique	développement touristique
développement local et économique	développement local	développement économique
développement pastoral et économique	développement pastoral	développement économique
développement social et économique	développement social	développement économique

Table 45 continued from previous page

Conjonction	TMM1	TMM2
eau libre ou circulant	eau libre	eau circulant
élimination du déchet et du détritus	élimination du déchet	élimination du détritus
énergie éolien et solaire	énergie éolien	énergie solaire
énergie fossile ou nucléaire	énergie fossile	énergie nucléaire
énergie renouvelable et nucléaire	énergie renouvelable	énergie nucléaire
énergie solaire et éolien	énergie solaire	énergie éolien
énergie vert et renouvelable	énergie vert	énergie renouvelable
environnement littoral et marin	environnement littoral	environnement marin
environnement marin et côtier	environnement marin	environnement côtier
environnement naturel et humain	environnement naturel	environnement humain
érosion du sol et du delta	érosion du sol	érosion du delta
exploitation forestier et minier	exploitation forestier	exploitation minier
exploitation minier et pétrolier	exploitation minier	exploitation pétrolier
extraction de pétrole et de gaz	extraction de pétrole	extraction de gaz
flore marin et terrestre	flore marin	flore terrestre
gestion agricole et forestier	gestion agricole	gestion forestier
impact environnemental et sanitaire	impact environnemental	impact sanitaire
impact environnemental et social	impact environnemental	impact social
impact environnemental et socio-économique	impact environnemental	impact socio-économique
impact sanitaire et environnemental	impact sanitaire	impact environnemental
impact sanitaire ou environnemental	impact sanitaire	impact environnemental
impact social et environnemental	impact social	impact environnemental
industrie du gaz et du eau	industrie du gaz	industrie du eau
lutte contre le désertification et pour le atténuation	lutte contre le désertification	lutte pour le atténuation
milieu côtier et marin	milieu côtier	milieu marin

Table 45 continued from previous page

Conjonction	TMM1	TMM2
milieu littoral et marin	milieu littoral	milieu marin
milieu marin et littoral	milieu marin	milieu littoral
milieu marin et souterrain	milieu marin	milieu souterrain
milieu marin ou lacustre	milieu marin	milieu lacustre
milieu naturel ou urbain	milieu naturel	milieu urbain
milieu physique et biologique	milieu physique	milieu biologique
milieu physique et humain	milieu physique	milieu humain
milieu terrestre et aquatique	milieu terrestre	milieu aquatique
paysage naturel et urbain	paysage naturel	paysage périurbain
paysage naturel ou urbain	paysage naturel	paysage urbain
paysage urbain et rural	paysage urbain	paysage rural
perte de biodiversité et de résilience	perte de biodiversité	perte de résilience
pollution du sol et du eau	pollution du sol	pollution du eau
population rural et urbain	population rural	population urbain
population urbain et rural	population urbain	population rural
préservation de le faune et de le flore	préservation de le faune	préservation de le flore
préservation de le nature et de le environnement	préservation de le nature	préservation de le environnement
processus de croissance et de décomposition	processus de croissance	processus de décomposition
production agricole et alimentaire	production agricole	production alimentaire
production agricole et forestier	production agricole	production forestier
protection de le faune et de le flore	protection de le faune	protection de le flore
protection de le flore et de le faune	protection de le flore	protection de le faune
protection de le nature et de le agriculture	protection de le nature	protection de le agriculture
protection de le nature et de le biodiversité	protection de le nature	protection de le biodiversité
protection de le nature et de le environnement	protection de le nature	protection de le environnement

Table 45 continued from previous page

Conjonction	TMM1	TMM2
protection du espèce et du écosystème	protection du espèce	protection du écosystème
protection du espèce et du habitat	protection du espèce	protection du habitat
protection du espèce ou du écosystème	protection du espèce	protection du écosystème
protection du sol et du eau	protection du sol	protection du eau
protection du sol et du transport	protection du sol	protection du transport
puits de pétrole et de gaz	puits de pétrole	puits de gaz
ressource en eau et du écosystème	ressource en eau	ressource du écosystème
ressource en eau et du milieu	ressource en eau	ressource du milieu
ressource en eau et en sol	ressource en eau	ressource en sol
risque de sécheresse et de inondation	risque de sécheresse	risque de inondation
science de le terre et de le vie	science de le terre	science de le vie
surface de terre et de eau	surface de terre	surface de eau
surface de terre ou de mer	surface de terre	surface de mer
taux de récupération et de valorisation	taux de récupération	taux de valorisation
temps chaud et sec	temps chaud	temps sec
tourisme écologique et équitable	tourisme écologique	tourisme équitable
transport routier et aérien	transport routier	transport aérien
transport routier et ferroviaire	transport routier	transport ferroviaire
valorisation du déchet et du boue	valorisation du déchet	valorisation du boue
destruction du couvert forestier	destruction du couvert	couvert forestier
perte de le banquise arctique	perte de le banquise	banquise arctique
expansion du couvert forestier	expansion du couvert	couvert forestier
élimination du couvert forestier	élimination du couvert	couvert forestier
dégradation de le matière organique	dégradation de le matière	matière organique

Table 45 continued from previous page

Conjonction	TMM1	TMM2
augmentation de le dégradation du sol	augmentation de le dégradation	dégradation du sol
couche de ozone arctique	couche de ozone	ozone arctique
conservation de un ressource naturel	conservation de un ressource	ressource naturel
grand delta fluvial	grand delta	delta fluvial
aggravation du réchauffement climatique	aggravation du réchauffement	réchauffement climatique
aggravation du réchauffement planétaire	aggravation du réchauffement	réchauffement planétaire
fermeture de le saison de pêche	fermeture de le saison	saison de pêche
fermeture de le saison de chasse	fermeture de le saison	saison de chasse
rendement de un moteur à combustion	rendement de un moteur	moteur à combustion
rendement de un moteur électrique	rendement de un moteur	moteur électrique
rendement de un moteur à essence	rendement de un moteur	moteur à essence

Annexe D

Ensembles de variantes qui sont des bitermes

réseau autoroutier
sommet montagneux
environnement périurbain
industrie agroalimentaire
réchauffement climatique
production hydroélectrique
transport interurbain
région subarctique
production électrique
niveau marin
production alimentaire
production gazier
variation climatique
production pétrolier
ressource en terre
eau en surface
protection naturel
couche de ozone
exploitation pétrolier
puits de carbone
couverture neigeux
émission gazeuses
production fourrager
activité microbiologique
émission en gaz

Table 46 continued from previous page

production céréalier
protection solaire
composé inorganique
changement climatique
écosystème planétaire
consommation électrique
eau de surface
zone périurbain
espace périurbain
modification climatique
puits à carbone
ressource marin
pénurie en eau
tortue marin
produit agrochimique
carbone inorganique
climat subtropical
évolution climatique
culture céréalier
réchauffement planétaire
valorisation du matière
exploitation forestier
eau alimentaire
milieu périurbain
substance inorganique
climat planétaire
sommet du montagne
industrie pétrolier
risque de inondation
manteau neigeux
environnement biophysique
produit biochimique
oiseau marin
réseau électrique
biodiversité ultramarin