UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LES CHANGEMENTS CLIMATIQUES DANS LES PUBLICATIONS DE MÉDIAS QUÉBÉCOIS SELON LA SÉMANTIQUE VECTORIELLE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN LINGUISTIQUE

PAR

MYLÈNE VÉZINA-BOUCHER

AOÛT 2022

REMERCIEMENTS

Je veux d'abord remercier mon directeur de recherche Grégoire Winterstein, pour son infinie patience et sa bienveillance. Ce fut un plaisir de discuter avec toi. Cette expérience m'a éprouvée, mais grandie aussi.

Un énorme merci à mes parents pour le soutien et l'amour. Merci de me laisser m'envoler. Vous me manquez.

Merci, Clément, pour toute l'énergie, la patience et les chocolats. On va enfin pouvoir consacrer notre énergie à d'autres projets.

À Ariane, Laurie, Laurence et Isabelle, merci pour votre présence et votre soutien. Je suis heureuse d'avoir vécu ces dernières années avec vous. Vous êtes précieuses.

À Érika, Amélie, Kim, Édith, Stéphanie et Alexandra, merci de me changer les idées quand c'est possible. Merci de rester les mêmes malgré tous ces changements, malgré le temps et la distance.

Audréanne, on a réussi.

Merci à Thomas Leu et à John Lumsden, pour leur curiosité contagieuse et leur passion dévorante pour la morphosyntaxe et la sémantique. Je n'aurais peut-être pas entrepris le projet de mémoire s'ils ne m'avaient pas recommandé autant de lectures intéressantes.

DÉDICACE

Les faits appartiennent tous au problème à résoudre,

non pas à sa solution.

- Ludvig Wittgenstein (1922)

AVANT-PROPOS

Ce mémoire est le fruit de longues réflexions sur la pertinence de relever les termes employés relativement aux changements climatiques dans les articles de journaux au Québec. Les questions qui me taraudaient le plus étaient : en quoi est-ce utile? Quelle différence cela fait-il dans notre capacité à nous adapter aux changements climatiques? Certes, cela nous permet de mieux comprendre le paysage informatif auquel nous sommes confrontés tous les jours. Cela nous permet également de mieux nous outiller quant aux stratégies à utiliser pour informer la population, stratégies qui doivent demeurer les plus neutres possible pour préserver l'absence de partisanerie dans les médias d'information. Transmettre un message ou ne pas le transmettre constitue déjà une prise de position en soi. Malheureusement, il est très difficile de mesurer l'absence d'information transmise. Ce mémoire se concentre donc sur l'information transmise dans les médias. Et comme les changements climatiques sont en cours, et que la population mondiale demeure apathique face à ces changements, ce mémoire se veut une piste de réflexion pour l'amélioration des interventions en vue d'une prise de conscience collective.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

La couverture des enjeux sociétaux par les médias traditionnels (les journaux, les nouvelles télévisuelles, les annonces radiophoniques, etc.) a un impact important sur la perception des Québécois face à leur rôle de citoyens responsables (Bérubé, 2010). Ces médias prétendent à une couverture factuelle de ces enjeux. Alors qu'on pourrait penser que certains articles ne contiennent que des éléments factuels dénués d'opinion, des études (p. ex. Caliskan *et al.* (2017)) ont démontré qu'un large corpus de textes n'ayant pas d'orientation connue à priori, tirés d'Internet, contenait certains biais (ou stéréotypes) liés à l'histoire et à la culture humaine, telle qu'une plus grande facilité à associer les prénoms masculins aux termes liés au métier de médecin et les prénoms féminins aux termes liés au métier d'infirmière, que l'inverse. Il est donc possible de dégager du sens implicite des textes traités par des modèles d'apprentissage automatique non supervisé.

L'objet principal de cette recherche est de dégager et d'analyser les associations implicites, les analogies et les plus proches voisins relatifs aux changements climatiques contenus dans les nouvelles produites par les entreprises médiatiques au Québec. Plus précisément, il s'agit d'utiliser le cadre de la sémantique vectorielle pour dégager du sens implicite d'un corpus sur les changements climatiques. Il s'agit également de déterminer à quels champs sémantiques sont associées les thématiques liées aux changements climatiques.

Mots clés : sémantique vectorielle, changements climatiques, médias québécois, linguistique informatique

INTRODUCTION

Cette recherche aborde le traitement des changements climatiques par les médias québécois du point de vue de la sémantique vectorielle. Si les ressources en sémantique vectorielle sont devenues des outils standards en traitement automatisé du langage, peu d'études en comparaison se sont penchées sur l'interprétation de ces ressources et de l'information qu'elles encodent. Caliskan *et al.* (2017) démontrent que des modèles sémantiques construits sur la base de larges corpus de textes, constitués sans orientation spécifique à priori encodent des biais d'association similaires aux biais humains. Nous empruntons alors le cadre de la sémantique vectorielle afin d'y étudier des éléments d'un corpus sur les changements climatiques. Il s'agit donc de dégager et d'analyser les analogies, les plus proches voisins et les champs lexicaux liés aux changements climatiques. Pour ce faire, un corpus des articles publiés par TVA Nouvelles, La Presse et Radio-Canada de mai 2016 à mai 2019 a été collecté. Il a ensuite été traité à l'aide d'outils de sémantique vectorielle, tel que l'algorithme word2vec (Mikolov *et al.*, 2013b).

Ce mémoire est divisé en cinq chapitres. Le premier discute de la problématique des changements climatiques et de son traitement dans les médias québécois. Le second met en contexte le cadre théorique de la sémantique vectorielle. Le troisième aborde la méthode utilisée pour atteindre les objectifs mentionnés ci-dessus. Le quatrième présente les résultats de cette recherche. Le cinquième analyse et discute des résultats obtenus.

CHAPITRE 1
LA PROBLÉMATIQUE DES CHANGEMENTS CLIMATIQUES

Dans ce chapitre, les changements climatiques sont discutés en tant qu'enjeu dont les impacts ne sont pas concrètement observés localement. La population québécoise obtient essentiellement l'information des médias dits traditionnels ainsi que des médias sociaux. Sachant que certains préjugés et stéréotypes sont transmis de manière implicite même au sein de textes se voulant neutres, les médias pourraient influencer l'opinion de la population par la reproduction, involontaire, de ces biais.

1.1 Bref portrait des changements climatiques

En 2018, le GIEC a publié un rapport concernant les dernières recherches scientifiques sur le réchauffement climatique (IPCC - Intergovernmental Panel on Climate Change, 2018). Ce rapport souligne que les changements climatiques sont les conséquences du réchauffement planétaire, qui est un phénomène naturel amplifié par l'activité humaine. Selon Cook *et al.* (2013), 97 % des publications scientifiques sur le réchauffement planétaire aux États-Unis appuient le fait que les activités humaines sont la principale cause des changements climatiques.

1.1.1 Les impacts des changements climatiques

Au Québec, la température moyenne a augmenté de 1 ^{o}C à 3 ^{o}C de 1990 à 2011. Cette augmentation de température a un impact sur les activités économiques et sociales québécoises. Publié en 2015, le rapport *Évaluation des impacts des changements climatiques et de leurs couts pour le Québec et pour l'État québécois* fait état des projections des conséquences des changements climatiques sur la population québécoise, notamment l'augmentation des décès liés à l'exposition à la température élevée, l'augmentation des risques de contracter des maladies comme la maladie de Lyme ou le virus du Nil, et l'augmentation du risque d'inondations dans certaines régions (Bureau de projet des changements climatiques, 2015). Les changements climatiques affectent également directement le secteur touristique, puisque la période estivale est plus propice au tourisme. Une plus grande variation de température affecte directement la possibilité de tenir des activités touristiques. L'état des routes est aussi affecté par les plus grands épisodes de variation de température, puisque la succession de gels et de dégels endommage les routes. Les épisodes de smog sont également plus fréquents en raison des hausses de température

(Gouvernement du Québec, 2015). Les changements climatiques ont donc une importante emprise sur l'évolution des activités économiques et sociales au Québec.

1.1.2 Les prises de position politiques et scientifiques

Sachant les impacts importants des changements climatiques sur la population canadienne et québécoise, les gouvernements québécois et canadiens ont tous deux établi des plans d'action à adopter pour limiter les effets des changements climatiques. Ceux-ci sont disponibles sur leurs sites web respectifs. La plateforme canadienne aborde essentiellement les impacts des changements climatiques sur la population, et la résilience canadienne. Il n'y a aucune mention d'un lien causal entre des activités humaines et les changements climatiques. Il est question d'adaptation aux changements climatiques et d'atténuation des effets sur l'économie (Ressources naturelles Canada, 2013). Le plan d'action québécois, quant à lui, prend position quant à la causalité des actions humaines sur les changements climatiques. Il est écrit que « certaines de nos habitudes de vie engendrent, directement ou indirectement, des émissions de gaz à effet de serre (GES). En modifiant certains de nos comportements, nous contribuons à la lutte contre les changements climatiques […] » (Gouvernement du Québec, 2015). Le gouvernement québécois incite donc davantage la population à prendre conscience de l'importance des gestes posés et à contribuer individuellement à la lutte aux changements climatiques.

1.2 La perception de la population

Malgré tous les faits rapportés par les rapports scientifiques et les prises de position politiques, 25 % des Québécois sont climatosceptiques, c'est-à-dire qu'ils rejettent l'idée que les changements climatiques sont prouvés scientifiquement ou le fait qu'ils sont dus à l'activité humaine (de Marcellis-Warin *et al.*, 2015). Et même lorsqu'ils sont convaincus que les changements climatiques existent et sont le fruit des activités humaines, les citoyens québécois se considèrent comme très écologiques. Ils sont cependant parmi les plus grands pollueurs au monde (Bérubé, 2010).

Mildenberger *et al.* (2016) ont modélisé des données de sondages probabilistes réalisés par téléphone au Canada sur les préférences et perceptions des Canadiens quant aux changements climatiques et à l'adoption de politiques publiques pour s'adapter à leurs répercussions dans le cadre des projets Sondage canadien sur l'énergie et l'environnement (*Canadian Surveys on Energy and the Environment* ou CSEE) et Cartes de l'opinion publique canadienne sur le climat (COPCC). La figure 1.1 ci-dessous présente une partie des résultats obtenus au Québec.

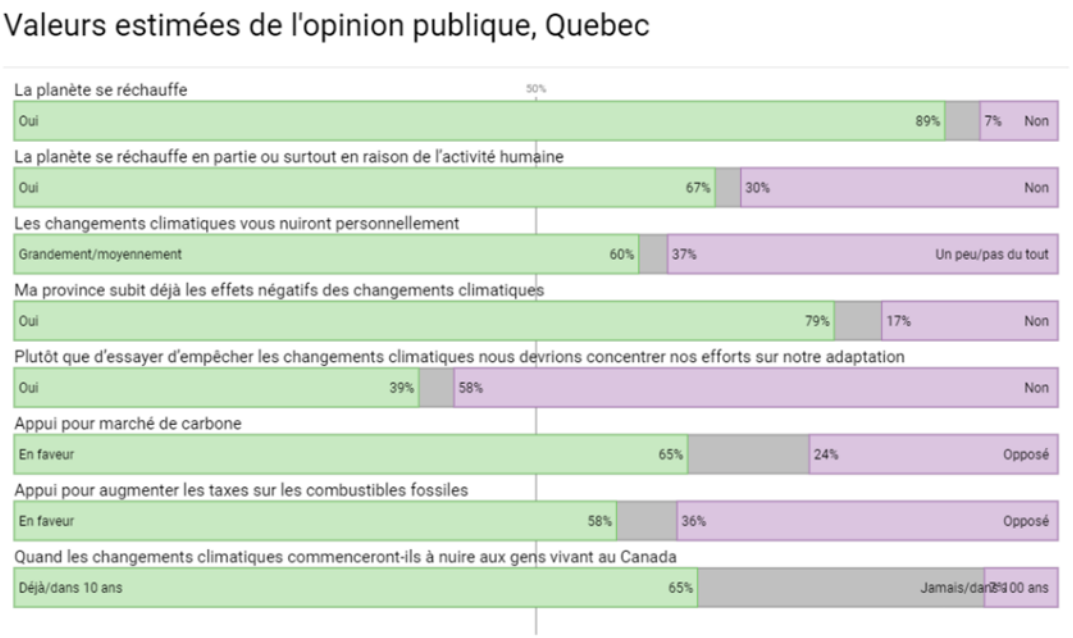

Figure 1.1 L'opinion publique au Québec[1]

La carte du projet CSEE sur l'opinion publique des changements climatiques souligne qu'alors que 89 % de la population québécoise croit que la planète se réchauffe et que 79 % croient que la province de Québec subit déjà des effets négatifs des changements climatiques, 37 % des Québécois ne croient pas ou peu que les changements climatiques leur nuiront personnellement (Mildenberger *et al.*, 2016). Il y a donc un écart important entre les impressions à l'échelle individuelle et à l'échelle provinciale.

Good (2008) a fait une étude comparative d'analyse de contenu dans les médias sur le traitement des changements climatiques à l'aide du modèle de propagande de Herman et Chomsky (2002) en comparant les médias américains, canadiens et internationaux. Elle considère que les perceptions et les réalités au sujet des changements climatiques sont éloignées en raison de la nature de leurs causes et de leurs conséquences : les émissions de gaz à effet de serre, qui causent les changements climatiques, sont principalement d'origine humaine et mesurables (donc concrètes), alors que les conséquences sont principalement naturelles, et n'ont pas nécessairement un impact direct et mesurable sur les activités

[1] Le CSEE et le COPCC n'endossent ni n'acceptent aucune responsabilité quant aux analyses ou interprétations des données présentées ici.

humaines. Par exemple, les causes des changements climatiques recensées sont observables, puisqu'il est possible de mesurer sa propre quantité d'émissions de gaz carbonique sur une période, ou la quantité de plastique consommé qui n'est pas nécessairement recyclé. En retraçant l'origine de la production d'un produit, il est possible de mesurer l'énergie dépensée par sa mise en marché. Cependant, les impacts des conséquences des changements climatiques tels que la plus grande fréquence des tornades dans certaines régions dans le monde ou l'augmentation du niveau des océans ne sont pas nécessairement observables à l'échelle locale, à moins d'habiter dans une région propice aux tornades ou près de l'océan. De plus, il est difficile d'établir scientifiquement et hors de tout doute le lien causal entre l'augmentation de la fréquence des tornades sur plusieurs années et l'augmentation de l'émission des gaz à effet de serre due au transport, notamment.

De son côté, Baron (2006) a conduit une étude en psychologie sur les biais implicites que les gens présentent face aux changements climatiques. Plus spécifiquement, il a testé la plus grande facilité des Américains à répondre en ligne à certaines questions qu'à d'autres, par exemple à leur capacité à payer pour améliorer les effets du réchauffement climatique causés par l'humain ou par la nature. Ses résultats démontrent que les Américains sont plus enclins à payer pour réparer leurs propres erreurs que pour réparer celles des autres, et pour compenser les effets de l'activité humaine sur le réchauffement climatique que pour inverser le réchauffement planétaire de causes naturelles, ce qui rejoint l'hypothèse de Good (2008) en ce qui concerne la perception « concrète » des conséquences des changements climatiques.

Comme il est moins aisé d'observer concrètement de chez soi les effets des changements climatiques à l'échelle de la planète, les citoyens québécois doivent se fier aux médias (spécialisés ou non, traditionnels ou non) pour connaitre l'état actuel des changements climatiques. Ces médias peuvent rapporter plusieurs types d'information, que ce soit des données factuelles, des reportages sur le terrain, des appels aux experts, des témoignages ou des images. De plus, la proportion d'information sur les questions environnementales dans les médias dépend vraisemblablement de l'intérêt du public sur ces questions (Létourneau, 2014).

1.3 Les opinions dans les médias

Les médias ont donc une énorme influence sur la perception de la population (Bérubé, 2010) concernant les enjeux courants de la société. Afin de préserver la qualité de l'information transmise à la population

dans les médias, le guide de déontologie journalistique (Conseil de presse du Québec, 2015) prévoit qu'elle soit exacte, rigoureuse, impartiale, équilibrée et complète. Il distingue également deux genres journalistiques : le journalisme factuel et le journalisme d'opinion. Ces deux genres doivent être aisément identifiables afin de ne pas tromper le public. Alors que les médias traitant des nouvelles globales tendent à une couverture factuelle des enjeux, les proportions d'expositions aux différents points de vue des sujets d'actualité ne sont pas réglementées et varient en fonction de ce que les médias jugent être des sujets importants dans l'instant présent. L'augmentation de la couverture médiatique sur les changements climatiques dans les dernières années témoigne d'un certain intérêt de la population (Létourneau, 2014).

Dans les pays européens, les médias donnent très peu de visibilité aux points de vue climatosceptiques, alors que les médias américains mettent plutôt l'accent sur le doute que soulève les acteurs politiques sur l'existence des changements climatiques (Comby, 2012) (Wetts, 2020). De plus, une forte corrélation a été observée entre le traitement des questions de climat dans les médias et l'attitude du public quant aux changements climatiques (Martin, 2020). Les textes dits objectifs des médias pourraient-ils être biaisés? Est-il possible de séparer l'opinion des faits dans les médias, et dans les textes en général? Caliskan *et al.* (2017) ont démontré que des ressources lexicales construites à partir d'un large corpus de textes issus d'Internet, à priori sans cohérence idéologique, encodent des stéréotypes susceptibles d'être néfastes dans la transmission de la culture humaine. À l'ère du traitement des données de masse, des applications utilisant l'intelligence artificielle et reposant sur de telles ressources peuvent donc perpétuer, voire intégrer ces stéréotypes dans leur fonctionnement : ce qui est écrit peut alors avoir un impact considérable sur les préjugés et les stéréotypes transmis.

La sémantique vectorielle, dont il sera question dans le prochain chapitre, permet de mieux saisir la teneur de ces biais transmis en observant l'utilisation de termes (fréquences, cooccurrences, relations sémantiques) dans un corpus de textes journalistiques sur les changements climatiques. Dans son analyse du discours médiatique sur les enjeux environnementaux rapportés par les experts en 2009 au Québec, Létourneau (2014) fait le constat que, bien que les enjeux environnementaux soient présents en faible nombre en proportion des autres sujets, les angles d'analyse sont fidèles aux données présentées et demeurent loin du scepticisme dont font preuve certains médias américains. Il sera intéressant d'opposer ces constats aux résultats obtenus sur le corpus d'articles de presse écrite de 2016 à 2019, et d'observer si, comme Caliskan *et al.* (2017), il est possible d'extraire des tendances d'un corpus à l'aide de la sémantique vectorielle.

CHAPITRE 2
CADRE THÉORIQUE

Dans ce chapitre, l'hypothèse distributionnelle et son lien avec la représentation du sens des expressions lexicales sont présentés. Ensuite, sont abordés les concepts et les applications de la sémantique vectorielle, qui constituent un héritage des principes de l'hypothèse distributionnelle mis en pratique sur des corpus de grande taille. La question de recherche se trouve à la fin de ce chapitre.

2.1 Le sens déterminé par sa distribution

2.1.1 L'hypothèse distributionnelle et la représentation du sens

Cette recherche s'inscrit dans le cadre de l'hypothèse distributionnelle en linguistique, défendue notamment par Harris (1954), qui soutient que le sens d'une expression est déterminé par son usage. Ainsi, cette hypothèse propose que les termes qui apparaissent dans les mêmes cotextes (donc entourés des mêmes termes, des mêmes constructions syntaxiques) possèdent des éléments de sens similaires. Cela s'explique par le fait que les locuteurs ont tendance à interchanger naturellement dans leur production langagière des termes similaires dans des situations similaires (Gastaldi, 2020). Il existe donc une certaine corrélation entre les termes qui apparaissent dans les mêmes contextes et la similarité de sens de ces termes. Voici des exemples d'inférences, rendues possibles grâce à cette corrélation, avec le terme imaginaire « chimiga » dans ses contextes de production fictifs (exemples 1-2-3), et d'autres termes utilisés dans les mêmes contextes (exemples 4-5-6) :

1. Ce **chimiga** aime jouer dans la cour de Mme Sophie.
2. Un **chimiga** se nourrit de carottes.
3. Le **chimiga** de son voisin est poilu et blanc.

4. ...ce **chien** est sorti jouer dans la cour...
5. ...les **lapins** sont appâtés par les carottes...
6. ...les **chèvres** ont un pelage blanc…

Les syntagmes ci-dessus illustrent que le terme « chimiga » des exemples 1, 2 et 3 désigne un animal de compagnie ou un mammifère en constatant qu'il est produit dans les mêmes cotextes (soit *jouer*, *cour*, *carottes*, *blanc*, *poilu*, *pelage*) que « chien », « lapins » ou « chèvres » en 4, 5 et 6. Cette association entre production dans les mêmes contextes et similarité de sens se révèle aussi en discriminant les cotextes auprès desquels « chimiga », « chien », « lapins » ou « chèvres » n'apparaissent pas, comme en 7a-b et en 8a-b :

7.
    a. ...ces **élèves** apprennent le français…
    b. *…ces **chèvres** apprennent le français…
8.
    a. …**Mme Sophie** discute de politique avec sa sœur…
    b. *…le **chimiga** discute de politique avec sa sœur…

L'astérisque devant les exemples 7b et 8b désigne l'absence d'attestation d'occurrences de ces syntagmes en production spontanée d'un locuteur du français. En effet, le terme « chimiga » a peu de chance d'être produit dans un contexte où un effort cognitif de l'entité « chimiga », comme l'apprentissage d'une langue ou une discussion politique, est requis. L'hypothèse distributionnelle repose donc sur la discrimination des cotextes pour déterminer la similarité de groupes de termes (Gastaldi, 2020). Gastaldi explique cette constatation par la conjecture suivante : si « x » et « y » apparaissent fréquemment dans les cotextes *a*, et si « y » et « z » apparaissent fréquemment dans les cotextes *b*, la conjonction des cotextes *a* et *b* donne une intersection similaire à « y », qui exclura « x » et « z ». Selon Gastaldi, cette conjonction correspond à la multiplication de leurs probabilités respectives, ou à l'addition des vecteurs de terme en relation avec ces probabilités. L'addition d'*a* et *b* résulte donc en un vecteur près de « y ».

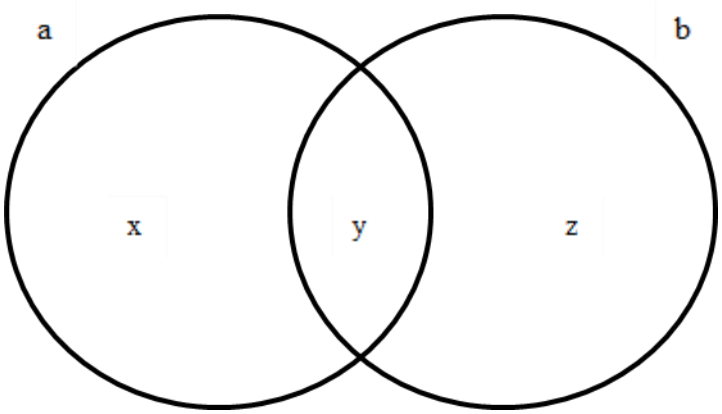

Figure 2.1 Illustration de la logique des interactions des cotextes de termes

Gastaldi présente *a* comme les cotextes de « x » et « y », et *b* comme les cotextes de « y » et « z », où, par exemple, x = « quebecois », y = « fleuve Saint-Laurent » et z = « fleuve ». Si « fleuve Saint-Laurent » apparait fréquemment dans les mêmes phrases que « québécois » et « fleuve », la somme des vecteurs de termes « x » et « z » donnera un vecteur de caractéristiques près de « fleuve Saint-Laurent ». Suivant cette logique et en revenant à l'exemple de « chimiga », de « chien » et de « chèvre » : si « chèvre » et « chimiga » apparaissent souvent dans les mêmes cotextes, et idem pour « chimiga » et « chien », il est possible de déduire ce qui est semblable à « chimiga » en additionnant les vecteurs de termes de « chèvre » et de « chien ».

Ainsi, plus les termes apparaissent dans plusieurs cotextes similaires, plus on leur trouve des éléments communs de sens. Les synonymes ont donc des significations similaires dans l'usage, puisqu'ils sont presque interchangeables dans leurs cotextes, en modifiant très peu le sens de la proposition. Par exemple, les termes « prophète » et « messager » pourraient être synonymes dans les phrases suivantes :

9. Jésus est le **prophète** de la bonne nouvelle.
10. Jésus est le **messager** de la bonne nouvelle.

Il existe cependant des cotextes de « messager » dans lesquels « prophète » n'apparaitra pas, lorsqu'il est question d'un facteur par exemple. Ainsi, bien qu'il soit possible de trouver des éléments communs de sens aux termes apparaissant dans plusieurs cotextes similaires, ils ne seront jamais parfaitement identiques.

2.1.2 Approches traditionnelles en sémantique lexicale

En sémantique lexicale, le principe de contraste (Clark, 1987) postule que deux termes sémantiquement similaires mais n'ayant pas la même forme linguistique ont nécessairement des différences dans leur signification. Il n'existe donc pas de réels synonymes et il devrait donc être possible d'extraire la signification de deux termes par la différence entre leurs cotextes.

11. Mon frère est un **prophète** pessimiste, il est persuadé qu'une récession frappera notre société en raison des ravages de la COVID-19.
12. ?Mon frère est un **messager** pessimiste, il est persuadé qu'une récession frappera notre société en raison des ravages de la COVID-19.

Par exemple, si « prophète » apparait dans le contexte 11, l'utilisation de « messager » dans le contexte 12 est improbable, puisque cet usage fait appel à la signification de « prévisions, intuitions » de prophète, alors que « messager » ne possède pas ce sens dans l'usage.

De leur côté, les antonymes, dont la signification est souvent en opposition binaire ou sur un spectre (Hill *et al.*, 2015), sont fortement liés dans leur distribution, puisqu'ils qualifient généralement les mêmes types de termes, comme en 13 et 14. Leurs cotextes sont donc semblables.

13. C'est un grand magasin.
14. C'est un petit magasin.

En comparant les exemples 13 et 14, il est possible de déduire que les antonymes « grand » et « petit » apparaissent généralement dans les mêmes cotextes. Selon les exemples 15 et 16, les antonymes « grand » et « petit » sont complémentaires et se comportent différemment de synonymes, puisqu'ils ne peuvent qualifier le même élément simultanément (Cruse, 1986). Les exemples en 15 et en 16 démontrent que des antonymes ne peuvent qualifier un même élément simultanément, alors que des synonymes le peuvent.

15. ?L'éléphant est grand et petit.
16. L'éléphant est gigantesque et immense.

La signification commune de ces deux antonymes pourrait donc être le spectre de la taille (comme traité par Kennedy et McNally, 2005). En traitement automatique, les antonymes sont difficiles à départager des synonymes (Jurafsky et Martin, 2019). Cependant, leur différence de sens pourrait potentiellement être capturée par des tâches d'analogie (voir Turney, 2008 pour des pistes en ce sens). Celles-ci sont détaillées dans la sous-section 2.3.2.

Un autre lien de sens pouvant être déduit de l'hypothèse distributionnelle est les relations de parenté (appelées configurations sémantiques par Cruse, 1986). Les relations taxonomiques telles que l'hyponymie (boxer < chien < animal) et l'hyperonymie (animal > chien > boxer) font aussi parties des relations de parenté pouvant être déduites de l'hypothèse distributionnelle. L'hyperonymie et l'hyponymie sont des relations hiérarchiques où un terme est l'extension du second terme plus général ou plus spécifique, c'est-à-dire que le boxer est une race de chien, le chien est une catégorie animale et, par conséquent, le boxer est un animal. Ces relations peuvent être capturées par des outils de sémantique vectorielle (dont il sera question dans la prochaine section), tel qu'appliqué par Nayak (2015) en entrainant un modèle de plongement (word embedding) à l'aide de la base de données Wordnet pour sélectionner les paires types de termes hyperonymes dans une tâche de classification, afin de permettre au modèle de prédire d'autres paires de termes en relation d'hyperonymie.

Les cadres sémantiques, au sens de Lehrer (2012), peuvent aussi être déduits des regroupements de termes effectués à l'aide des modèles de plongement. Les termes sont alors présents conjointement dans des cotextes similaires, puisqu'ils sont liés par un événement commun, un domaine commun. Cet événement commun peut ressembler à ce contexte, dans lequel l'outil souvent utilisé par une cuisinière est le fouet :

17. La cuisinière fit monter les blancs d'œuf en neige avec son fouet.

Si certains termes apparaissent régulièrement dans un contexte de recettes de cuisine, comme « fouet », « four », « cuisinière », « couper », leur cadre sémantique sera intuitivement le domaine de la cuisine. Les grappes (clusters) sont entre autres utilisées pour cerner les domaines sémantiques d'un corpus.

Enfin, une dernière relation de sens intéressante à observer dans les corpus est la connotation. Les connotations sont liées aux émotions, aux évaluations, aux opinions, positives ou négatives, de l'auteur, comme démontré par Osgood et al. (1957) (*dans* Jurafsky et Martin, 2019), qui ont proposé l'idée que ces

sentiments pouvaient être représentés par un vecteur à trois dimensions, chaque dimension se rapportant à une échelle présentant les facteurs suivants : la valence (*evaluation*, comme bon-mauvais, beau-laid, propre-sale), l'activation (*activity*, comme rapide-lent, actif-passif, chaud-froid) et la dominance (*potency*, comme grand-petit, fort-faible, lourd-léger). Les stéréotypes sont une forme de connotation (Wodak, 1989), qui peuvent émerger automatiquement des corpus de texte à l'aide des associations effectuées par le modèle (voir la section 2.3 pour une élaboration à ce propos).

2.2 La sémantique vectorielle

Avec l'avènement de l'apprentissage profond (deep learning) en traitement automatique des langues naturelles (Bengio *et al.*, 2012), dû notamment à une meilleure accessibilité aux masses de données du Web (big data), de nouvelles perspectives de recherche s'ouvrent en linguistique informatique pour traiter quantitativement les contextes des termes. En effet, plus la taille d'un corpus est grande, plus l'entrainement d'un modèle est performant et plus les résultats obtenus permettront des analyses fines (Lai *et al.*, 2016). La sémantique vectorielle est un cadre d'analyse utile en linguistique informatique pour améliorer les performances des algorithmes de traitement de données linguistiques, permettant au passage d'observer des régularités dans les fréquences d'association de certains termes. Comme ces algorithmes sont entrainés à l'aide de corpus de données écrites, ils ne peuvent apprendre réellement la signification des termes de la réalité physique. Néanmoins, ils permettent d'en apprendre davantage sur les régularités dans l'usage de la langue, notamment quant aux catégories lexicales ou syntactiques, permettant de mieux saisir des éléments de sens des termes par leurs contextes (Bender et Koller, 2020). Ainsi, la sémantique vectorielle repose sur l'analyse d'énormes corpus à travers les modèles de plongement, qui condensent les données, et qui permettent de les visualiser plus facilement en dégageant des analogies, des similarités entre les distributions des termes et des associations implicites.

2.2.1 Le vecteur et la matrice terme-terme

Avant de poursuivre, une définition du vecteur associé à un terme et de la nature de ses dimensions s'imposent. Un vecteur est représenté par un ensemble de coordonnées dans un espace donné. La nature de ces coordonnées varie selon la nature du modèle vectoriel projeté. Néanmoins, dans la grande majorité des cas, ces modèles étudient la fréquence des cooccurrences. Dans cette perspective, un vecteur de terme contient l'information de la nature et la fréquence des termes apparaissant dans l'environnement d'un terme ciblé. Autrement dit, chaque dimension d'un vecteur de terme capture de l'information liée à

l'occurrence des termes dans ses contextes. Un contexte correspond au nombre de termes délimités avant et après le terme cible.

Prenons la phrase suivante :

18. La linguistique est un domaine vaste, mais très intéressant.

On considère ici comme contexte une fenêtre de cotexte de quatre termes de chaque côté du terme cible. Pour le terme « domaine », les éléments du contexte seront donc les termes « la », « linguistique », « est », « un », « vaste », « mais », « très », « intéressant ». Ces vecteurs de termes sont compilés dans une matrice terme-terme (word-word matrix). Dans une matrice terme-terme, une cellule correspond au nombre de fois qu'un terme apparait dans le contexte d'un autre.

Tableau 2.1 Exemple de matrice terme-terme, inspirée de celle de Manning et Schütze (1999, p. 297)

	La	Linguistique	Est	Un	Domaine	Vaste	Mais	très	intéressant
La	1	1	1	1	1	0	0	0	0
Linguistique	1	1	1	1	1	1	0	0	0
Est	1	1	1	1	1	1	1	0	0
Un	1	1	1	1	1	1	1	1	0
Domaine	1	1	1	1	1	1	1	1	1
Vaste	0	1	1	1	1	1	1	1	1
Mais	0	0	1	1	1	1	1	1	1
Très	0	0	0	1	1	1	1	1	1
Intéressant	0	0	0	0	1	1	1	1	1

Le terme « intéressant » n'apparait donc jamais en présence du terme « linguistique ». Cependant, « vaste » et « intéressant » ont plus de chances d'apparaitre dans leur contexte mutuel.

Le choix de la taille de la fenêtre cotextuelle semble avoir un effet sur le type d'information encodée par les modèles de plongement. Plus la fenêtre de contexte est petite, mieux le modèle performera dans des tâches d'analyses syntaxiques, alors que plus la fenêtre de contexte est grande, mieux le modèle

performera dans les tâches d'analyses sémantiques (Pennington *et al.*, 2014). Cela s'explique notamment par le fait que les configurations des particules syntaxiques telles que les déterminants, articles et adjectifs sont souvent rapprochées des groupes nominaux, une fenêtre contextuelle plus petite capturant ainsi les interactions entre les noms et leurs modifieurs, notamment. Une fenêtre de contextes plus grande capturera davantage l'interaction entre les différents prédicats d'un syntagme. Gastaldi (2020) avance qu'une telle constatation renforce l'hypothèse que la syntaxe et la sémantique participent toutes deux d'une même contribution à l'élaboration du sens, puisque le sens d'une phrase (ou d'un corpus) ne peut être entièrement saisi sans l'interaction entre la sémantique et la syntaxe des éléments. Ainsi, bien que la grandeur de la fenêtre cotextuelle semble avoir une incidence sur la nature de l'information encodée par les modèles, elle n'a pas autant d'impact que supposé sur l'interaction entre les vecteurs de termes similaires, puisque l'information encodée est complémentaire à leur signification.

2.2.2 Les modèles vectoriels

Les modèles vectoriels permettent d'effectuer des opérations sur les vecteurs de termes. Les matrices terme-terme comme celle du tableau 2.1 présentent un défaut : elles contiennent majoritairement des « zéros », c'est-à-dire des contextes où des termes n'apparaissent pas ensemble, surtout lorsque la fenêtre cotextuelle est petite. Ces zéros rendent la factorisation d'éléments de sens moins aisée, puisque plus la matrice à traiter est de grande dimension, plus complexes seront les opérations de manipulation de vecteurs. Afin d'améliorer les traitements computationnels sur les vecteurs, il faut donc compresser les contextes de termes pour éliminer les « vides » dans la matrice terme-terme. La compression de contextes d'un terme dans un vecteur est désignée par un modèle de plongement. Pour construire des modèles vectoriels, une matrice terme-terme dense peut être construite de deux façons différentes : en partant d'une matrice de cooccurrences et en compressant les vecteurs obtenus, ou en produisant directement un modèle dense à l'aide d'une tâche de prédiction (Baroni *et al.*, 2014). Ainsi, pour obtenir des modèles de plongement, soit on compresse des matrices issues de l'algorithme tf-idf (Spärck Jones, 2004), soit on utilise des algorithmes comme word2vec ou GloVe (Mikolov *et al.*, 2013c ; Pennington *et al.*, 2014). Ces deux méthodes sont présentées ci-dessous.

2.2.2.1 L'algorithme tf-idf

L'algorithme Term Frequency – Inverse Document Frequency (tf-idf) est souvent utilisé dans le domaine de la recherche d'information, notamment pour comparer des vecteurs de documents (Spärck Jones, 2004). Il effectue deux opérations afin de créer des vecteurs de termes. D'abord, il divise la fréquence de

chaque terme d'un document par le nombre de termes total dans un document (ou un corpus), ce qui correspond à « Term Frequency ». Ensuite, « Inverse Document Frequency » mesure l'informativité d'un terme à l'aide de la mesure N/df, qui correspond au nombre total de documents dans un corpus, divisé par la fréquence du terme dans l'ensemble du corpus (Scott, 2019). Cette mesure assignera plus de poids aux termes spécifiques au voisinage de certains termes, et moins d'importance aux vecteurs de termes comme « et » ou « de », qui sont très fréquents en français et apparaissent dans les contextes de la quasi-totalité des termes du lexique du français (Tirtha, 2020). L'algorithme tf-idf crée une matrice de tous les contextes possibles d'un terme incluant les contextes « zéros », c'est-à-dire les contextes où un terme n'apparait pas. Il ne résout donc pas le problème de dimensionnalité mentionné ci-dessus; son utilité principale est de fournir une meilleure approximation du lien entre termes. Pour produire des matrices denses, on utilise des techniques de réduction de dimensionnalité dans les modèles suivants afin d'augmenter la performance de traitement des vecteurs. Ce type d'approche forme le cœur des travaux en sémantique latente, qui utilisent la décomposition en valeur singulière pour produire des modèles denses (Deerwester *et al.*, 1990).

2.2.2.2 Le modèle word2vec

Word2vec crée des modèles de plongement denses à l'aide de l'algorithme skip-gram ou de celui de continuous bag of words (CBOW). Ce type de modèle élude le problème des « zéros » dans la matrice (Mikolov *et al.*, 2013d). Au lieu de réduire une matrice de cooccurrences, l'algorithme cherche à prédire la probabilité d'apparition d'un terme dans le cotexte d'un autre terme. Pour ce faire, on entraine un réseau d'apprentissage neuronal profond, la supervision de l'entrainement se faisant sur la base de la matrice de cooccurrence. Les vecteurs associés à chacun des termes du lexique sont obtenus en extrayant les paramètres de la couche interne du réseau, communément appelée la boite noire du réseau neuronal. L'algorithme skip-gram utilise le terme donné pour prédire ses cotextes, alors que CBOW prédit le terme donné à l'aide des cotextes avant et après celui-ci. (voir figure 2.2 ci-dessous pour une illustration de la différence entre les deux algorithmes).

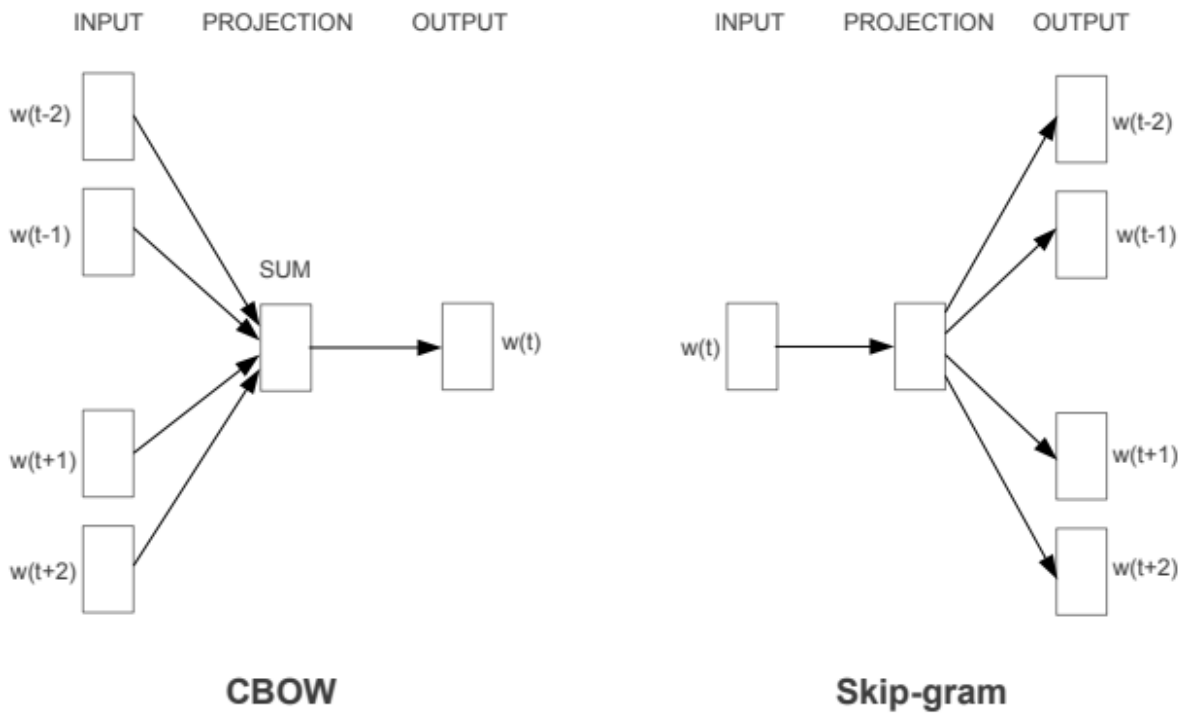

Figure 2.2 Représentation du traitement des données des modèles CBOW et Skip-gram (Mikolov *et al.*, 2013a)

La phrase « Le pauvre chat noir miaule furieusement. » servira d'exemple. Il est demandé au modèle de travailler avec une fenêtre de 2, c'est-à-dire deux termes avant et deux termes après le terme cible. Avec l'algorithme skip-gram, le terme en intrant w(t) serait « chat ». L'algorithme tente alors de trouver les termes cotextuels situés en positions w(t-2), w(t-1), w(t+1) et w(t+2), comme « pauvre » en w(t-1), « noir » en w(t+1) et « miaule » en w(t+2), à l'aide du vocabulaire et de la matrice de poids présents dans le modèle. L'algorithme ne tient pas compte des termes communs comme « le » ou « de ». Le modèle CBOW, quant à lui, combine tous les cotextes d'un terme. Dans la phrase en exemple, les cotextes combinés seraient « pauvre », « noir » et « miaule », et l'algorithme calcule alors les probabilités que le terme « chat » se trouve en w(t).

En utilisant un large corpus de textes pour entrainer le modèle et obtenir les vecteurs de termes de ce corpus, Mikolov *et al.* (2013) démontrent des gains de performance en utilisant les représentations obtenues (voir la sous-section 2.3.2 ci-dessous pour une illustration).

2.2.2.3 Le modèle GloVe

GloVe est un autre algorithme pour obtenir des modèles de plongement, en créant une matrice de cooccurrence qui ajoute un poids aux « non-zéros ». Pennington *et al.* (2014) ont compilé un ensemble de vecteurs à l'aide d'un corpus de 6 milliards de termes. Pour produire un modèle de plongement, GloVe exploite également une matrice de cooccurrences, sur la base de laquelle des ratios de probabilité de cooccurrences sont calculés. Le but de l'algorithme est alors d'associer aux termes du lexique des vecteurs pour que le produit scalaire de deux vecteurs approxime ces ratios.

2.2.3 Les avantages de la sémantique vectorielle

C'est donc dans la sémantique lexicale et l'hypothèse distributionnelle que la sémantique vectorielle prend racine et permet de découvrir des facettes jusqu'ici insoupçonnées de grands corpus de textes. L'avantage de la sémantique vectorielle réside notamment dans le fait qu'elle ne fait intervenir aucun critère de sélection dans la constitution et l'organisation des données. Les données sont brutes, peu annotées et sont utilisées telles quelles à l'aide de matrices de probabilités contextuelles. Ainsi, pour la constitution du corpus à l'étude dans cette recherche, tous les articles obtenus grâce à l'entrée du vocable « changement climatique » dans la barre de recherche des entreprises médiatiques québécoises sont retenus sans égard à leur contenu ou à leur format. Un annotateur a souvent pour principale tâche d'étiqueter des informations métalinguistiques, sémantiques ou syntaxiques quant aux parties du discours du corpus (Bender et Friedman, 2018). En supprimant cette étape d'annotation, on évite donc un possible biais de l'annotateur dans la composition du corpus, ce qui permet la démonstration de biais préexistants dans le contenu du corpus (voir 2.3.3).

La sémantique vectorielle permet également d'éliminer ce que certains, comme Baker (2012), critiquent de l'analyse de discours et des techniques d'apprentissage supervisé : les exemples donnés en entrées et sorties du système ne sont pas prédéterminés ou sélectionnés, ce qui permet d'écarter (du moins partiellement) le biais du chercheur (ou de la chercheuse dans ce contexte). L'interprétation des résultats obtenus par l'apprentissage non supervisé est plutôt exploratoire, dans la mesure où on ne peut prédire le résultat de la recherche. La sémantique vectorielle laisse parler le corpus de textes, ce qui n'empêche pas que le modèle puisse être biaisé en raison des données d'apprentissage (Friedman et Nissenbaum, 1996).

2.3 Les applications de la sémantique vectorielle

Lorsque le corpus est acquis et le modèle de plongement est construit, il ne reste plus qu'à interroger les données. Ci-dessous, il sera question des informations sémantiques qu'il est possible de déduire des vecteurs de termes à partir de tâches spécifiques, tels que l'analyse de la proximité sémantique et des analogies.

2.3.1 Les plus proches voisins

Les plus proches voisins des vecteurs de termes sont les vecteurs les plus proches d'un vecteur cible dans l'espace vectoriel. Il est possible de calculer la similarité de deux vecteurs en évaluant leur distance à l'aide de la mesure du cosinus de leur angle. S. Perone (2013) souligne que bien que des exemples en 2D soient utilisés pour illustrer le calcul de la similarité de deux vecteurs, il est mathématiquement possible de calculer les angles et la similarité de vecteurs dans un espace vectoriel de plus de deux dimensions.

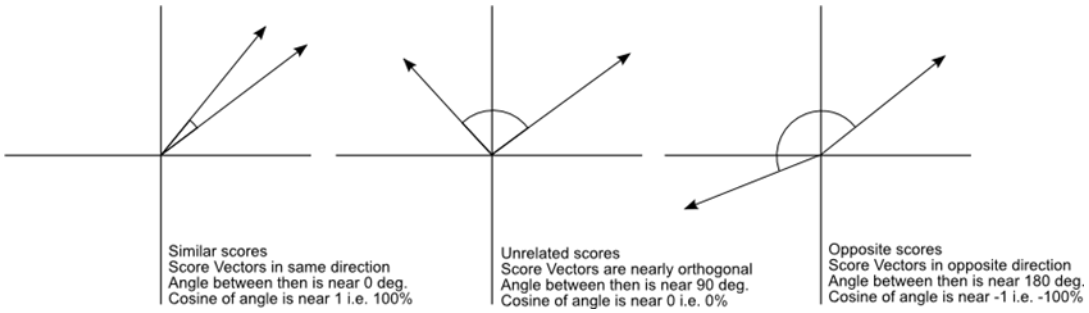

Figure 2.3 Représentation de la notion de distance entre les vecteurs (S. Perone, 2013)[2]

Deux vecteurs sont maximalement similaires lorsqu'ils forment un angle nul et qu'alors la mesure du cosinus de l'angle est égale à 1. Ils sont orthogonaux lorsque leur angle est de 90 degrés et que la mesure du cosinus de leur angle est égale à 0. Le score des vecteurs est opposé lorsqu'ils forment un angle de 180 degrés et que la mesure du cosinus de leur angle est égale à -1.

[2] « Scores similaires : Les vecteurs de score pointent dans la même direction, l'angle entre eux est près de 0 degré, le cosinus de l'angle est près de 1, c'est-à-dire 100 %. Scores non liés : Les vecteurs de score sont presque orthogonaux, l'angle entre eux est près de 90 degrés, le cosinus de l'angle est près de 0, c'est-à-dire 0 %. Scores opposés : Les vecteurs de score pointent dans des directions opposées, l'angle entre eux est près de 180 degrés, le cosinus de l'angle est près de -1, c'est-à-dire -100 %. » [Notre traduction]

Plus la mesure est proche de la valeur 1, plus les vecteurs sont similaires. Il est possible d'affirmer qu'une mesure de similarité du cosinus de l'angle de 0,676545 entre deux vecteurs signifie que ces deux vecteurs sont similaires à 67 %. En effet, comme la mesure du cosinus de l'angle est nécessairement entre 0 et 1 et que ces vecteurs partent tous de la même origine, une mesure de cosinus de l'angle de 0 voudra dire qu'il n'y a aucune dimension en commun entre les vecteurs testés, soit 0 % des dimensions sont similaires. Inversement, un vecteur dont la mesure du cosinus est de 1 signifiera que ces vecteurs ont toutes leurs dimensions en commun, donc 100 % des dimensions. La recherche des plus proches voisins de certains termes peut faire ressortir des régularités sur la signification de ceux-ci, notamment sur le domaine thématique du terme ou sur sa catégorie lexicale (revoir 2.1.2), puisque l'algorithme compare les vecteurs de contextes de termes.

2.3.2 Les tâches d'analogie

Les tâches d'analogie permettent de dégager des relations similaires entre deux paires de vecteurs testées en postulant qu'*a* est à *b*, ce que *c* est à *d*, par exemple, une pomme est à un pommier ce qu'une poire est à un poirier (Jurafsky et Martin, 2019). Une opération sur les vecteurs d'*a*, *b* et *c* est ensuite effectuée : $y = x_b - x_a + x_c$, *y* étant la représentation de l'espace dans lequel se trouve la meilleure réponse. La mesure de similarité du cosinus de l'angle est alors utilisée pour trouver le vecteur le plus proche de *y,* qui est *d*.

Mikolov *et al.* (2013b) se sont basés sur les données de la seconde tâche de SemEval-2012 (Jurgens *et al.*, 2012) pour trouver *d*. Ils ont pu produire un système capable de résoudre les tâches d'analogie proposées. Par exemple, Mikolov *et al.* (2013) démontrent que lorsqu'on soustrait le vecteur *man* du vecteur *king*, et qu'on y additionne *woman*, on obtient un vecteur proche du vecteur *queen*, résultat suggérant que le modèle encode la relation linguistique entre certaines paires de termes. Ces relations peuvent être de différentes natures : sémantiques (hyperonymie/hyponymie, antonymie, singulier/pluriel), mais aussi syntactiques (nom/adjectif/verbe). Il est possible d'observer d'autres régularités dans la figure 2.4 ci-dessous, comme le fait que les vecteurs de termes évoquant la royauté ou la famille soient plus près les uns des autres.

Les modèles de plongement permettent d'effectuer des tâches d'analogie et de trouver les plus proches voisins des vecteurs, entre autres. Ils permettent aussi de visualiser les données dans des espaces vectoriels (voir le code de Schmidt (2015) pour des exemples de visualisation possibles). Voici un exemple de visualisation de relation entre des vecteurs tiré de GloVe :

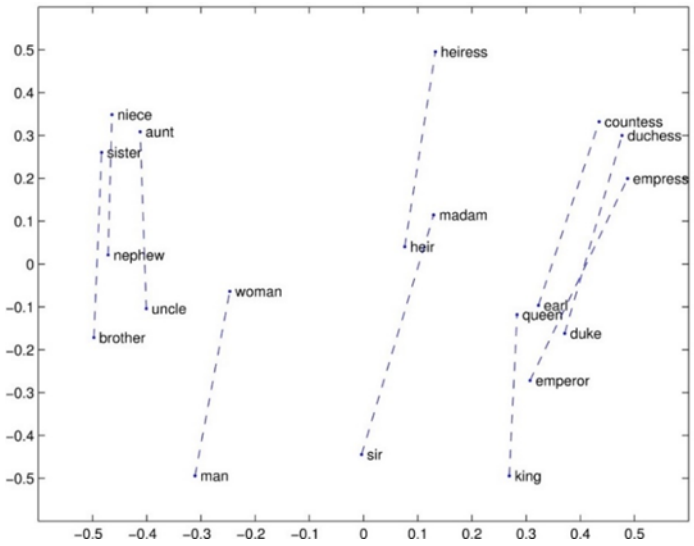

Figure 2.4 Représentation des analogies de vecteurs de terme (Pennington *et al.*, 2014)

La figure 2.4 représente des tâches d'analogie effectuées sur des vecteurs de termes. Les valeurs des axes du graphique sont modulées en fonction de la représentation des coordonnées des vecteurs de termes. Les lignes pointillées représentent les relations entre ces vecteurs de termes que l'outil de visualisation a fait ressortir. En fonction des termes testés, les relations représentées dans le graphique peuvent être représentées à l'horizontal; elles ne sont pas exclusivement verticales. Il faut cependant garder en tête que les tâches d'analogie présentées ci-dessus doivent demeurer un outil d'exploration de corpus, plutôt qu'un outil d'évaluation du contenu du corpus, puisque la méthode comporte certaines limites que signalent entre autres Linzen (2016) et Gladkova *et al.* (2016). En effet, la tâche d'analogie telle que présentée ici ne permet pas d'interpréter le contenu sémantique d'une relation entre deux paires de vecteurs, elle permet plutôt d'inférer leurs éléments communs, puisque les relations entre les paires de termes testés sont avant tout identifiées (par la personne qui interprète les données).

Ce qui est tout de même notable, c'est que le modèle reproduit le résultat escompté avec une certaine précision, qui varie selon la catégorie de la relation testée. Linzen souligne entre autres que les tâches d'analogie testant des relations d'adjectifs à superlatifs ne performent pas bien, alors que celles testant les termes au pluriel et au singulier obtiennent de très bons résultats. Gladkova *et al.* (2016), de leur côté, observent que les tâches d'analogie sur les relations de morphologie flexionnelle performent mieux que les relations de morphologie dérivationnelle, ce qui rejoint les observations de Linzen quant à la

performance de la catégorie des relations entre les termes au singulier et au pluriel. Par exemple, un modèle trouvera plus facilement le résultat de la tâche d'analogie 19 (morphologie flexionnelle) que celui de la 20 (morphologie dérivationnelle :

19. Chats – chat + chien = chiens
20. Retardement – retard + avance = avancement

Les modèles vectoriels sont tout de même reconnus pour leurs bonnes performances dans la résolution de tâches d'analogie impliquant des relations d'hyperonymie/hyponymie.

2.3.3 La sémantique vectorielle et corpus annoté spécialisé sur le climat

Luo *et al.* (2020) ont étudié la possibilité de détecter automatiquement les opinions sur le débat des changements climatiques dans les médias en montant un corpus de 2 000 phrases annotées à partir de 56 000 articles de presse sur ce sujet de 2000 à 2020 de 63 sources aux États-Unis nommé DeSMOG. Ces phrases ont été retenues si elles contenaient les mots-clés suivants : climate change, global warming, fossil fuels, carbon dioxide, methane, co2. Ils ont annoté ces 2 000 phrases selon un cadre argumentatif (SOURCE, PRÉDICAT, OPINION) afin de dégager le lexique de l'opinion. Comme les opinions détectées n'étaient pas nécessairement sur les changements climatiques, ils ont retenu uniquement les éléments d'opinion dont les phrases contenaient un des 73 mots-clés liés aux changements climatiques sélectionnés manuellement :

> *climat, climact, global, warm, carbon, fossil, oil, energi, environ, co2, green, ice, glacier, glacial, melt, sea, temperatur, heat, hot, methan, greenhous, arctic, antarct, celsiu, fahrenheit, ecosystem, pole, environ, coal, natur, human, economi, electr, futur, health, scienc, econom, air, pollut, fire, wildfir, ipcc, epa, market, scientist, earth, planet, wind, solar, record, fuel, ocean, nuclear, scientif, pipelin, emit, emiss, concensu, renew, accord, forest, pruitt, drought, hurrican, atmospher, activist, coast, agricultur, water, plant, weather, polar.*

Une sélection de ces mots-clés sert de termes ciblés sur les changements climatiques dans la présente étude (voir la sous-section 3.4.2.2).

Ils ont ensuite entrainé un classifieur BERT (Devlin *et al.*, 2019) à identifier ces opinions et à les classer « pour », « contre » ou « neutre ». Ils ont aussi développé un lexique de prédicats d'affirmation et de doute (savoir, affirmer, prétendre) et de modificateurs de SOURCE (évalué par les pairs, trompeur, controversé).

Ils ont comparé les résultats obtenus aux résultats attendus selon l'orientation politique des entreprises médiatiques (de droite ou de gauche).

Luo *et al.* (2020) ont démontré que le classificateur BERT était aussi performant que des annotateurs humains pour étiqueter des opinions comme « pour » ou « contre » l'existence d'éléments sur les changements climatiques. Selon leurs résultats, les entreprises médiatiques « pro » et « sceptiques » utilisent des stratégies argumentatives semblables pour véhiculer des opinions, avec une légère propension des entreprises « sceptiques » à démontrer des doutes face aux propos des « opposants ».

Il est donc possible d'extraire des opinions sur les changements climatiques d'un corpus d'articles de presse annoté grâce à un cadre argumentatif et aidé d'un lexique. Mais qu'en est-il d'un corpus peu annoté?

2.3.4 La sémantique vectorielle et corpus peu annoté

Caliskan *et al.* (2017) ont étudié la possibilité que les biais implicites étaient contenus dans le corpus de textes d'une langue. Les biais implicites sont ici associés à la définition de « stéréotypes », autrement dit une association entre deux termes qui est utilisée couramment dans l'usage et qui est qualitative. La notion de « stéréotype » semble renvoyer à une définition « péjorative » ou à un biais, un stéréotype référant dans le cadre de cette étude à une différence d'association qui peut être socialement discriminatoire. Caliskan *et al.* (2017) ont reproduit des résultats obtenus lors de l'Implicit Association Test (IAT), durant lequel les participants devaient associer, par exemple, des termes plaisants et non plaisants à des entités neutres. Voici un exemple de passation de l'IAT (Greenwald *et al.*, 1998) :

Une personne participante découvre quatre lexiques : « famille », « carrière », « mâle » et « femelle ». Des termes sont associés à chacun des lexiques (par exemple, « enfant » entre dans la catégorie « famille »). La personne participante doit ensuite cliquer à droite ou à gauche de l'écran en fonction du terme central qu'elle doit associer à une ou l'autre des catégories de chaque côté de l'écran. Plus précisément, dans la figure ci-dessous, la personne participante doit sélectionner « E » ou « I » en fonction de catégories et d'ensemble de termes prévisualisés. Ici, « children » entre dans la catégorie « family », « E » doit donc être sélectionné.

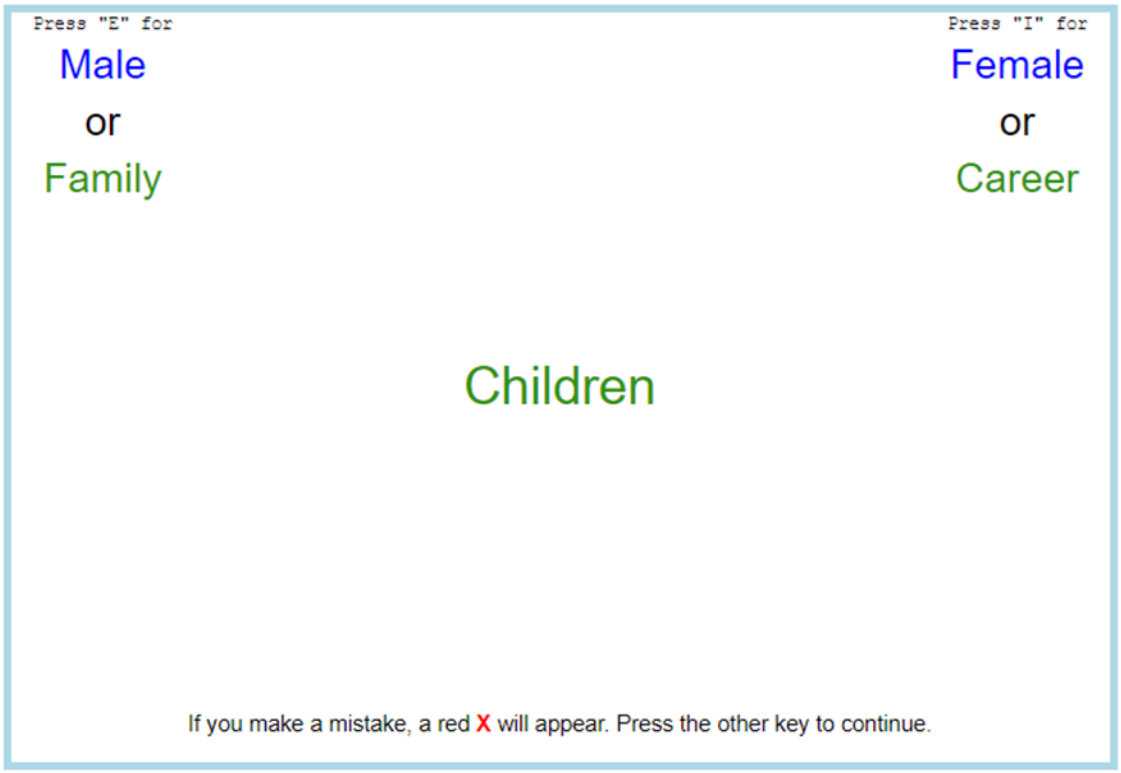

Figure 2.5 Interface de l'IAT.

Le temps de réaction est mesuré, et comparé en fonction des associations faites aux catégories. Greenwald *et al.* (1998) ont donc testé le temps de réaction des personnes participantes pour conclure que certaines catégories étaient plus faciles à associer entre elles que l'inverse (le temps de réponse était plus rapide). Ainsi, la catégorie « fleurs » s'associait plus aisément avec les termes plaisants et la catégorie « insectes », avec des termes moins plaisants, que l'inverse (Bryson, 2017).

Caliskan *et al.* (2017) ont reproduit un test équivalent en reprenant ces paires de termes pour les tester à l'aide des modèles vectoriels GloVe et word2vec présentés plus haut, afin de vérifier la proximité des vecteurs à 300 dimensions de ces termes. Leurs résultats quant à la proximité des vecteurs correspondent aux résultats des associations implicites. Par exemple, Caliskan *et al.* (2017) ont reproduit les résultats démontrant que le groupe de prénoms de femmes était plus facilement associé au groupe de termes du lexique de la famille et le groupe de prénoms masculins aux groupes de termes liés à la carrière. Ces résultats démontrent que les corpus utilisés pour produire les modèles vectoriels, bien que ne contenant

à priori aucune orientation explicite, encodent en fait des éléments culturels et que des biais implicites humains se trouvent dans les corpus de textes non annotés (Bender *et al.*, 2021).

Le fait que ces éléments aient des liens forts avec des faits historiques laisse croire que les textes sont imprégnés des acquis de leurs auteurs et qu'ils se révèlent en acquérant seulement les régularités quant à la fréquence d'apparition de certains termes dans des contextes semblables, qui permettent de déduire des éléments de sens (Bryson, 2017). Ainsi, nul besoin de comprendre la logique et les arguments d'une phrase, il suffit de *comprendre* la relation sémantique entre les termes et leur agencement pour saisir ce qu'on tente d'exprimer.

Cependant, un modèle de plongement ne *comprend* pas le sens des mots au même titre qu'un être humain : il apprend les statistiques d'apparition dans différents contextes, il fait ressortir certaines catégories lexicales, mais ne peut pas comprendre le monde comme un être humain le ferait (Bender et Koller, 2020). Néanmoins, un modèle de plongement peut démontrer des récurrences de l'association de certains termes pour décrire une réalité, ce qui renforcerait certains biais dans l'interprétation de l'information présentée dans le texte.

Comme le soulève Bryson, il s'agit ici d'une brève incursion dans une modélisation de la manière dont le langage est utilisé pour transmettre des idées de façon implicite. Il n'est pas dit que c'est exactement ce qui se passe dans la réalité, mais le fait que la sémantique vectorielle ait pu faire cette association augmente les probabilités que ce soit possible. Le choix des termes utilisés dans les articles de presse est donc crucial pour véhiculer l'information nécessaire à la compréhension des impacts des changements climatiques sur la planète. La sémantique vectorielle permettrait ainsi de déceler les termes privilégiés pour qualifier les événements rapportés dans les médias, liés aux changements climatiques dans le cas présenté en 2.3.3 et dans le cadre de la présente étude.

2.4 Question de recherche

En supposant que les constats de Caliskan *et al.* (2017), c'est-à-dire qu'il soit possible de saisir des régularités implicites à propos de la réalité uniquement en étant exposé au langage, peuvent aussi être observables dans les médias francophones au Québec, la question principale de cette recherche est la suivante :

Que nous apprend la sémantique vectorielle dans le cadre de l'approche d'un corpus d'articles québécois sur les changements climatiques?

Bien que la démarche de cette étude soit plutôt exploratoire, plusieurs hypothèses sont posées au moment de l'élaboration de cette question de recherche.

La première hypothèse est que le modèle de plongement créé à partir d'un petit corpus spécialisé en français sera en mesure de résoudre des tâches d'analogie posées à l'aide de termes non spécialisés sur les changements climatiques, ce qui serait compatible avec les hypothèses soulevées par Caliskan *et al.* (2017) et Bender et Koller (2020).

La seconde hypothèse est que les changements climatiques sont associés au lexique scientifique (scientifique, mesure, température, giec, soutient, prouve, etc.). Cette hypothèse vient du fait que les enjeux environnementaux sont souvent abordés dans les médias à l'aide de l'appel aux experts, procédé par lequel des scientifiques s'expriment sur leur sujet d'étude (Létourneau, 2014). Pour vérifier cette hypothèse, l'étude de Luo *et al.* (2020) est utilisée comme référence afin de comparer l'usage de prédicats d'opinion dans le corpus.

La troisième hypothèse est que les changements climatiques sont davantage présentés d'un point de vue proclimatique plutôt que climatosceptique, comme Létourneau (2014) a pu l'observer dans son étude de la presse écrite de 2009. Des références aux activistes proclimatiques et climatosceptiques seront vérifiées.

CHAPITRE 3
MÉTHODE

Dans ce chapitre, la méthode utilisée est présentée, entre autres en ce qui concerne la constitution et le traitement du corpus.

3.1 Choix du corpus

Trois grands médias constituent le corpus : TVA Nouvelles, La Presse et Radio-Canada. Ces médias ont été choisies, puisqu'elles rejoignent une plus grande part d'audience en ligne à travers la province du Québec. En effet, selon une étude de Vividata rapportée dans Robert (2018), les trois quotidiens ayant la plus grande part d'audience en ligne sont La Presse au premier rang, Le Journal de Montréal et le Journal de Québec au second et troisième rang. TVA Nouvelles est une propriété de Groupe TVA Inc., filiale de Québecor Média, qui est une entreprise québécoise qui existe depuis 1974 (Groupe TVA, 2018). Puisque Le Journal de Montréal et Le Journal de Québec sont des quotidiens régionaux, reconnus pour verser dans le journalisme d'opinion, mais qu'ils appartiennent tous deux à la même entreprise, soit Groupe TVA, leur contrepartie télévisuelle qui parait pratiquer le journalisme factuel à l'échelle provinciale, soit TVA Nouvelles, a été sélectionnée. La Presse, une fiducie de Power Corporation du Canada, est un quotidien québécois depuis 1884, dont la dernière édition papier était en décembre 2017 (La Presse, 2018). La Société Radio-Canada est un radiodiffuseur public national francophone du Canada depuis 1920, numérique depuis 2000 (Radio-Canada, 2018).

Pour faciliter la constitution du corpus, seules les nouvelles publiées en ligne ont été retenues. Les chroniques et les éditoriaux n'ont pas été sélectionnés, puisque l'objectif est d'obtenir un corpus ayant une vision d'ensemble de l'information diffusée par les médias qui disent pratiquer le journalisme factuel au Québec. Selon Ho-Dac et Küppers (2011), ce choix d'articles en ligne pour le corpus peut avoir des répercussions sur la qualité de l'information transmise :

> Pour la version en ligne, le standard est autre : par rapport aux "classiques du métier" (e.g. évaluation et validation de l'information, temps de rédaction jusqu'au moment du bouclage, etc.), le journaliste web subit davantage l'immédiateté de la mise à disposition de l'information, cherchant à mettre la nouvelle en ligne le plus vite possible, à être le premier à publier l'information pour avoir l'exclusivité, ce qui peut l'amener à simplement copier-coller et adapter à la ligne éditoriale une dépêche d'agence de presse (p. 4).

Néanmoins, le corpus est constitué d'articles papier et en ligne, puisque les parutions papier ont été archivées en ligne sans distinction des publications en ligne. La période choisie pour la constitution du corpus est de 2016 à 2019, afin d'obtenir au moins 500 000 termes-occurrences (*tokens*).

3.2 Collecte de données

La collecte de données est faite à l'aide du moteur de recherche de chaque média. Le mot clé « changements climatiques » est entré dans le champ de recherche et les résultats obtenus sont tous retenus pour constituer le corpus, sans exception. Les doublons sont éliminés au fur et à mesure en vérifiant la date de publication et le titre de l'article ajouté. Les articles sont insérés en format XML dans un document texte. Les balises suivantes sont utilisées afin de bien délimiter les données pertinentes.

```
<texte>
<titre></titre>
<corps></corps>
<source>
                    <auteur></auteur>
                    <date></date>
                    <URL></URL>
                    <media></media>
</source>
</texte>
```

Figure 3.1 Balises XML pour délimiter les données

Seul le texte inséré dans les balises « titre » et « corps » constitue le corpus. Les infographies et leur légende ne sont pas retenues, afin de faciliter la prise de données. Les légendes de ces infographies sont par ailleurs souvent des reprises de phrases de l'article. Le corpus DataClimat contient un total de 972 articles entrés à la main dans un fichier texte, et un total de 552 194 termes-occurrences (*tokens*).

La relative petite taille du corpus soulève un enjeu de performance non négligeable. Cependant, selon les observations de Lai *et al.* (2016), le domaine du corpus a un plus grand impact sur la performance des modèles de plongement que sa taille, c'est-à-dire que, dans une tâche spécifique, un corpus d'un seul domaine performera mieux qu'un corpus mixte de différents domaines. Néanmoins, l'augmentation de la taille d'un corpus d'un même domaine améliorera sa performance. Dans le cas de la présente recherche, les ressources de temps et d'énergie étant limitées, la spécialisation du corpus contribue certainement à

sa performance dans les tâches données, mais il serait intéressant d'observer la différence de performance et de précision dans les résultats en augmentant la période de temps durant laquelle des articles journalistiques sont retenus.

3.3 Traitement du corpus

Le corpus est d'abord indexé, uniformisé (suppression de la ponctuation, mise en minuscule de l'ensemble du corpus, suppression des accents) et segmenté en termes. Ce traitement a été effectué pour faciliter l'encodage du corpus en python et le traitement par les modèles, car les bibliothèques utilisées pour créer les modèles de plongement n'était pas compatibles avec les caractères autres qu'anglais, et l'encodage des caractères accentués [ASCII ou utf-8] pouvait causer des problèmes.

Certaines expressions multimots, comme « changements climatiques », sont alors associées, afin qu'elles soient traitées comme une seule expression (voir annexe A pour la liste des expressions multimots). Cette opération est effectuée afin d'éviter, entre autres, qu'un ministère intitulé « ministère de l'Environnement et de la Lutte contre les changements climatiques » influence l'interprétation du corpus quant à la proximité des termes « changements climatiques » et « ministere ». L'espace entre ces termes a été remplacé par ZZZ dans le corpus, pour éviter la segmentation de ces termes par l'algorithme de plongement. À des fins de lisibilité, ZZZ sera remplacé par un trait d'union dans cette étude.

Le corpus n'a pas été lemmatisé, puisque cette opération peut ajouter du « bruit » aux résultats. De plus, pour répondre aux objectifs de cette étude, les données utilisées doivent demeurer les plus brutes possible. Ainsi, les distinctions grammaticales, comme singulier-pluriel, sont maintenues et permettent de tester la validité du modèle de plongement pour des tâches d'analogie.

Ensuite, les modèles de plongement sont créés à l'aide de l'algorithme word2vec, à partir du code de Schmidt (2015). Word2vec est utilisé dans RStudio (R) (voir l'annexe B pour les exemples de manipulations effectuées dans RStudio). Les paramètres du modèle de plongement ont été établis à 200 dimensions, avec une fenêtre de 10 termes, puisque l'objectif est davantage de capturer des éléments sémantiques que syntaxiques (revoir la sous-section 2.2.1 pour l'influence de la grandeur de la fenêtre sur l'information capturée par le modèle).

L'algorithme CBOW a été préféré à skip-gram dans la création des modèles de plongement, puisqu'il effectuait la tâche plus rapidement que skip-gram et que plusieurs études ont démontré une meilleure performance dans la résolution de tâches sur les termes les plus fréquents d'un corpus (Mikolov *et al.*, 2013a).

3.4 Analyse

Deux méthodes sont testées pour dégager des associations implicites du corpus d'articles, soit la recherche des plus proches voisins et la résolution de tâches d'analogie.

3.4.1 Les plus proches voisins

La première avenue consiste en l'analyse du voisinage sémantique en dégageant les vecteurs les plus proches de certains termes liés aux changements climatiques, tel que démontré, entre autres, dans les possibilités de GloVe (Pennington *et al.*, 2014).

Les plus proches voisins des vecteurs de terme « changements-climatiques », « environnement », « emissions », « rechauffement » et « climat » sont trouvés à l'aide de la mesure du cosinus de l'angle de deux vecteurs de termes. Ces cinq termes ont été choisis à partir du vocabulaire extrait du corpus. Les termes les plus fréquents en français étant les particules « de » ou « a », un tri a été effectué pour retenir uniquement les termes qui, intuitivement, sont en lien avec la thématique du corpus. La sélection a ensuite été effectuée en fonction de leur plus grande fréquence dans le corpus d'articles et de leur lien possible avec la thématique des changements climatiques, en sélectionnant les termes apparaissant également dans la liste des mots-clés thématiques de (Luo *et al.*, 2020).

3.4.2 Les tâches d'analogie

La seconde avenue consiste en la résolution d'une tâche d'analogie à l'aide de la juxtaposition de deux paires de vecteurs de termes afin d'en extraire la différence (Pennington *et al.*, 2014). Cette tâche d'analogie permet de faire ressortir certaines caractéristiques spécifiques de la relation sémantique entre ces deux paires, en plaçant les vecteurs en opposition dans un espace vectoriel. Les termes testés sont choisis intuitivement en fonction de leur lien possible avec la thématique des changements climatiques, de leur fréquence d'apparition dans le corpus à l'étude et de leur compatibilité aux mots-clés thématiques issus de l'étude de Luo *et al.* (2020).

La tâche se déroule en deux étapes. D'abord, pour évaluer la performance du modèle, des analogies sont générées à partir de paires de termes cernant des catégories lexicales ou syntaxiques. En guise de rappel, la formule utilisée est la suivante : $y = x_b - x_a + x_c$, y étant la représentation de l'espace dans lequel se trouve la meilleure réponse. La formule peut être verbalisée ainsi : « x_a est à x_b ce que x_c est à y ». Si le modèle trouve le terme recherché (ou un terme similaire sémantiquement) dans les dix premières occurrences des résultats, il est possible d'affirmer que les résultats sont cohérents et que le modèle de plongement peut encoder des régularités sur des éléments lexicaux ou syntactiques.

3.4.2.1 Analogies des relations attendues testées

Le modèle de plongement sera donc testé à l'aide d'analogies dont le résultat attendu est tiré des connaissances communes de la population québécoise francophone (à laquelle appartient la personne qui réalise cette étude) afin d'évaluer qualitativement la performance du modèle de plongement. Les groupes de termes retenus sont présentés ci-dessous. Les résultats attendus de ces analogies se trouvent dans la section 4.2.

Le premier groupe de termes permet de déterminer le niveau de performance du modèle pour résoudre les relations métonymiques :

- citoyens – villes + entreprises
- montreal – quebec + ontario
- trudeau – canada + quebec

Le second groupe de termes cible les relations syntactiques pronominales :

- nous – je + tu
- il – le + la
- moi – me + te

Le troisième groupe de termes cible les relations de morphologie flexionnelle :

- vendre – vend + achete
- initiatives – initiative + solution
- citoyen – citoyenne + rectrice

Le dernier groupe cible les relations de morphologie dérivationnelle :

- vendre – vente + achat
- decideurs – acteurs + decision
- urgent – urgence + prudence

3.4.2.2 Analogies des changements climatiques

Ensuite, le modèle de plongement est interrogé plus spécifiquement sur les enjeux des changements climatiques. Afin de saisir la notion de « changement » dans les relations d'analogie, une des paires de termes testées servira de base pour l'ensemble des tâches d'analogie, soit la métonymie « rechauffement – temperature ». L'hypothèse soulevée est que cette paire de termes fera ressortir les changements subis par les éléments affectés par les changements climatiques, comme l'atmosphere, le climat, les glaciers, mais aussi les conséquences de ces changements, comme les incendies et la pollution.

Une liste des termes retenus pour effectuer des observations sur les associations implicites aux changements climatiques dans le corpus d'articles de journaux a été élaborée. Ces termes sont tirés des 73 mots-clés permettant de classer des phrases comme « liées aux changements climatiques » de l'étude de Luo *et al.* (2020), introduite dans la sous-section 2.3.3, qui ont été traduits. Afin de restreindre les possibilités, seuls les mots-clés ayant une fréquence d'apparition de 100 ou plus dans le corpus Dataclimat ont été retenus.

Les termes retenus sont les suivants :

climat, carbone, fossile, energie, co2, glace, glaciers, fonds, oceans, chaleur, serre, ges, arctique, antarctique, environnement, charbon, nature, economique, electricite, sante, scientifiques, pollution, feux, giec, marche, planete, gaz, emissions, accord, forets, atmosphere, agriculture, eau, meteo.

Les résultats des tâches de recherche des plus proches voisins et des tâches d'analogie sont présentés dans le prochain chapitre.

CHAPITRE 4

PRÉSENTATION DES RÉSULTATS

Dans ce chapitre, les résultats recueillis à l'aide de word2vec sont présentés et analysés.

4.1 Les plus proches voisins

Les plus proches voisins des vecteurs de termes « changementsclimatiques », « environnement », « emissions », « rechauffement » et « climat » se présentent dans des cotextes semblables à ceux présentés dans le tableau 4.1 dans le corpus Dataclimat. Par souci de concision, seules les dix premières occurrences de vecteurs de termes sont retenus.

Tableau 4.1 Le vecteur de terme « changements-climatiques »

Vecteurs de terme	**Mesure de similarité du cosinus**
changement-climatique	0,7608384
vaccins	0,6594123
négatifs	0,6575138
consommateurs	0,6435922
climatosceptiques	0,6421913
attaques	0,6228403
contribuables	0,6200939
gouvernements	0,6146607
débats	0,6097029
tribunaux	0,6081878

Les neuf des dix plus proches voisins du vecteur de terme « changements-climatiques » sont similaires de 60 % à 65 % à celui-ci. Sans surprise, le vecteur de terme le plus près de « changements-climatiques » est « changement-climatique » à 76 %. Il faut noter néanmoins que quelques-uns des autres plus proches voisins sont associés à un lexique à connotation négative, soit « négatifs », « climatosceptiques », « attaques », « débats » et « tribunaux », ce qui pourrait être interprété comme un indice que le sujet des changements climatiques est préoccupant pour les différents acteurs présentés dans les médias.

Le tableau suivant présente les plus proches voisins du vecteur de terme « environnement ».

Tableau 4.2 Le vecteur de terme « environnement »

Vecteurs de terme	**Mesure de similarité du cosinus**
catherine-mckenna	0,7647613
harjit-sajjan	0,7422043
mckenna	0,7357775
education	0,7294248
egalite	0,6801143
finances	0,6710224
premier	0,6675548
morneau	0,6624278
infrastructure	0,6607434
universite-d-ottawa	0,6558796

Dans les plus proches voisins du vecteur de terme « environnement », c'est la présence des vecteurs de terme « mckenna » et « catherine-mckenna », similaires à 76 % et 73 %, qui ne surprend pas, puisque la locution « la ministre de l'environnement catherine mckenna » apparait fréquemment dans les articles de presse du corpus. Il est donc concluant de les retrouver dans des cotextes semblables. Il faut également noter une dimension politique aux vecteurs de termes associés à « environnement », puisque « morneau » et « harjit-sajjan », faisant possiblement référence à Bill Morneau et à Harjit Sajjan sont des figures politiques au niveau fédéral au même titre que Catherine Mckenna.

Le tableau 4.3 présente les plus proches voisins du vecteur de terme « emissions ».

Tableau 4.3 Le vecteur de terme « emissions »

Vecteurs de terme	**Mesure de similarité du cosinus**
concentrations	0,7579442
objectifs	0,7342156
protoxyde	0,7258067
taxes	0,7219601
cibles	0,7216665
azote	0,7203202
factures	0,7108933
dioxyde	0,7095012
plastiques	0,7075305
ges	0,7062900

Les plus proches voisins du vecteur de terme « emissions » sont similaires de 70 % à 75 % aux vecteurs de terme du lexique des gaz, soit « concentrations », « protoxyde », « axote », « dioxyde » et « ges ». Il faut également noter la présence de « cibles » et « objectifs », qui marquent une volonté de contrôler les émissions à venir. Il est cependant surprenant de ne pas retrouver de vecteurs de termes liés à la « réduction » des émissions.

Le tableau 4.4 présente les plus proches voisins du vecteur « rechauffement ».

Tableau 4.4 Le vecteur de terme « rechauffement »

Vecteurs de terme	**Mesure de similarité du cosinus**
dereglement	0,8810694
crise	0,7666087
bouleversement	0,7354149
cataclysme	0,7158563
horloge	0,7153583
lobby	0,7073729
phenomene	0,7037679
deficit	0,7023640
atlas	0,7019280
changement-climatique	0,6999236

Avec 88 % de similarité, le vecteur de terme « dereglement » est le plus proche voisin de « rechauffement ». Avec « dereglement », les autres plus proches voisins du vecteur de terme « rechauffement », soit « crise », « bouleversement », « cataclysme », « deficit » et « changement-climatique », semblent marquer l'instabilité que le réchauffement climatique amène.

Le prochain tableau présente les plus proches voisins du vecteur de terme « climat ».

Tableau 4.5 Le vecteur de terme « climat »

Vecteurs de terme	**Mesure de similarité du cosinus**
changement-climatique	0,7110562
planete	0,6872609
pese	0,6770585
climatiques	0,6562573
climatique	0,6358401
nigeria	0,6356122
mars	0,6344928
geneve	0,6339611
pologne	0,6330813
futur	0,6267096

Les plus proches vecteurs de terme de « climat » sont similaires de 62 % à 71 %. Une partie de ces vecteurs de termes proviennent du lexique de « climat », soit « changement-climatique », « climatiques », « climatique ». Il faut noter la présence de vecteurs de termes dénotant des emplacements géographiques, soit « planete », « nigeria », « geneve » et « pologne ». Le lien entre « climat » et « Pologne » vient probablement du sommet pour le climat (COP24) qui y a eu lieu. Le sommet pour le climat COP2 a eu lieu à Genève, et pourrait expliquer la proximité des vecteurs de termes « climat » et « geneve ». De son côté, le Nigéria est très touché par les changements climatiques et a adopté une loi sur le climat créant un fonds pour le changement climatique (Trésor, 2021), ce qui peut expliquer la proximité entre les vecteurs de termes « climat » et « nigeria »

En résumé, les plus proches voisins de « changements-climatiques » et de « rechauffement » semblent associés à un lexique chargé d'une connotation négative importante. Ceux d'« environnement », « emissions » et « climat » semblent associés à un lexique plus neutre.

4.2 Les analogies observées : relations testées

Dans cette section, les résultats des analogies observées sont présentés. Les résultats attendus sont du type : A – a + b = B, c'est-à-dire que la relation vectorielle entre les deux premiers termes et la relation vectorielle entre les deux derniers termes sont similaires dans l'espace vectoriel. Par souci de concision, seuls les dix vecteurs de termes les plus similaires sont retenus pour analyse.

Les douze premiers tableaux en annexe C présentent les vecteurs de termes testés pour éprouver la performance du modèle sur différents types de relations syntactiques ou sémantiques.

4.2.1 Les relations métonymiques

Dans le tableau C.1, pour que « citoyens » et « villes » aient une relation similaire à « entreprises » et « y », la réponse attendue de cette première tâche d'analogie était intuitivement un actant, comme « employés » ou « consommateurs ». Les vecteurs de termes « canadiens », « politiciens », « consommateurs », « élus » et « gouvernements » répondent à cette attente avec une mesure de similarité de 89 % à 93 %.

La réponse attendue pour l'analogie du tableau C.2 était « toronto » ou « ottawa », qui ne se retrouvent pas dans les cinq vecteurs de termes avec les mesures de similarité du cosinus de l'angle les plus élevées. Il est possible que la nature spécialisée du corpus soit en cause dans ce cas-ci.

La tâche d'analogie du tableau C.3 répond presque parfaitement aux attentes, puisque Philippe Couillard et François Legault ont tous deux été premiers ministres du Québec de 2016 à 2019 (années de la collecte du corpus d'articles). Ainsi, le fait que la mesure de similarité du cosinus des paires de vecteurs de termes « trudeau – canada » et « quebec – couillard » soit à 97 % démontre que le modèle fonctionne pour les tâches d'analogie concernant les relations métonymiques.

4.2.2 Les relations syntactiques pronominales

La réponse recherchée dans la tâche d'analogie du tableau C.4 était « vous », ce qui correspond au premier résultat obtenu, avec 83 % de mesure de similarité du cosinus.

En C.5, le modèle devait pouvoir trouver « elle », qui se trouve en 3^{e} position dans le tableau. Les trois premiers résultats obtenus correspondent à des pronoms, soit « la », « il » et « elle ».

Dans la tâche d'analogie du tableau C.6, la réponse attendue était « toi ». Cependant, après vérifications, « toi » apparait seulement à cinq reprises dans le corpus, ce qui peut avoir un impact sur les résultats attendus. C'est également cohérent avec le style de discours habituellement utilisé dans les articles de journaux; on privilégie habituellement la troisième personne pronominale.

Le modèle semble à même d'identifier des pronoms dans le corpus, bien qu'ils ne soient pas toujours du bon genre ou du bon nombre.

4.2.3 Les relations de morphologie flexionnelle

Le vecteur de terme recherché dans l'analogie du tableau C.7 était « acheter », ce que le modèle n'a pas trouvé. Il faut noter néanmoins que le modèle semble avoir capturé la notion de « verbe à l'infinitif » pour d'autres vecteurs de termes similaires, comme « respecter », « ramener », « fonctionner » et « dévoiler ».

Pour la tâche d'analogie du tableau C.8, la réponse attendue était « solutions », ce que le modèle a fait émerger dans la deuxième proposition. Également, le modèle semble avoir capturé la notion de pluriel pour toutes les solutions trouvées. La mesure de similarité du cosinus est particulièrement près de 100 % pour cette tâche.

Bien que la tâche d'analogie du tableau C.9 donne l'impression que le modèle ait tout faux, puisque le vecteur du terme recherché « recteur » n'y est pas, il est possible de constater que le modèle a tout de même saisi l'élément « masculin ». Il semble également avoir capturé l'élément de sens « fonction », avec les vecteurs de terme « militant », « cofondateur » et « adjoint ». Enfin, Damon Matthews est un professeur en sciences du climat à l'université Concordia; « damon-matthews » serait donc pour « rectrice » ce que « citoyen » est à « citoyenne » à 99 % dans ce corpus.

4.2.4 Les relations de morphologie dérivationnelle

Dans le tableau C.10, le modèle a trouvé le résultat recherché : « acheter », ce qu'il a fait émerger en 10^{e} position. Il semble également avoir saisi l'élément de sens « verbe à l'infinitif ».

Dans la tâche d'analogie du tableau C.11, des vecteurs de termes sémantiquement proches de la réponse attendue « action » ont été trouvés, soit « proposition », « promesse », « motion », « resolution ».

La tâche d'analogie du tableau C.12 a été moins bien réussie, si on considère que la réponse attendue était « prudent ». Néanmoins, « lache » ou « effrayant » illustrent une retenue qui peut être similaire à de la prudence.

La première séquence de tâches d'analogie a démontré que le modèle de plongement entrainé sur un corpus spécialisé d'articles de journaux sur les changements climatiques de seulement 500 000 termes-occurrences répond aux attentes quant à sa performance de détection de régularités, bien que le corpus avec lequel il a été entrainé soit de petite taille. Le modèle semble donner des résultats encore plus cohérents lorsque des relations flexionnelles de termes sont impliquées, comme l'ont observé Gladkova *et al.* (2016) et Linzen (2016) (revoir la sous-section 2.3.2).

L'intérêt de ces manipulations repose sur le fait que les résultats obtenus n'ont pas été préannotées dans l'ensemble des textes. Le modèle a donc en quelque sorte *fait émerger* des régularités à partir d'un ensemble de textes bruts. Considérant que ces modèles n'ont jamais été testés sur un corpus québécois d'articles de journaux, le fait de retrouver des associations connues suggère que toute autre association testée dans ces modèles serait légitime.

4.3 Les analogies observées : les changements climatiques

Les tableaux présentant les résultats des tâches d'analogie effectuées sur les 34 termes tirés du lexique des changements climatiques se trouvent en annexe C de cette étude. Voici les observations sur les tâches d'analogie effectuées en lien avec les changements climatiques dans le corpus Dataclimat.

Ces tâches d'analogie comparent la relation entre les vecteurs de termes « rechauffement » et « temperature » à la relation entre les vecteurs de termes ciblés et les termes trouvés par le modèle.

Comme constaté durant la lecture des résultats de ces tâches, une seule analogie fait ressortir des éléments du lexique du climatoscepticisme. En effet, le tableau C.23 (rechauffement – temperature + serre = y) montre que le vecteur de terme « serre » est en relation avec « transcanada », « climatosceptique » et « lobby ». Durant les années ciblées du corpus, TransCanada faisait partie des acteurs qui agissaient « contre » les plans d'adaptation aux changements climatiques en investissant auprès des lobbys climatosceptiques aux États-Unis (CROTEAU, 2016). Il faut noter que « serre » apparait 507 fois dans le corpus, alors que « climatosceptique » apparait seulement huit (8) fois dans le corpus (« climatosceptiques », neuf [9] fois), « transcanada » apparait neuf (9) fois, et « lobby », six (6) fois.

Dans le tableau C.27 (rechauffement – temperature + environnement), le terme « environnement » est associé au lexique des acteurs pro-environnement, comme « harjit-sajjan », « dominic-leblanc »,

« activistes » et « greenpeace ». En effet, Harjit Sajjan, ancien ministre fédéral de la Défense et aujourd'hui ministre du Développement international, est reconnu pour ses sorties fréquentes sur les changements climatiques. Dominic LeBlanc, quant à lui ministre fédéral des Affaires intergouvernementales, a aussi quelques sorties à son actif au sujet du plan du gouvernement libéral pour les changements climatiques.

Le tableau C.33 (rechauffement – temperature + scientifiques) associe « scientifiques » à « experts », « rapports », « auteurs », « conclusions », ce qui semble faire émerger la légitimité de ces acteurs dans le discours de ce corpus.

Les tableaux C.34 et C.35 (rechauffement – temperature + pollution/feux = y) illustrent les effets des changements climatiques, avec des vecteurs de termes comme « cout », « fleau », « incendies », « degats, « violents ».

L'analogie du tableau C.38 (rechauffement – temperature + planete = y) marque le changement subi par la planète avec des termes comme « cataclysme », « mouvement », « dereglement » et « changement-climatique ».

Le tableau C.41 (rechauffement – temperature + accord) présente des termes en relation avec « accord » qui ont une connotation négative, comme « deficit », « echec » et « urgence », ce qui pourrait faire émerger l'insuffisance des accords établis entre les pays pour prendre des mesures afin de réduire les impacts des changements climatiques.

Les analogies des tableaux C.45 et C.46 (rechauffement – temperature + eau/meteo = y) font référence à des vecteurs de termes indiquant un lexique d'incertitude, d'instabilité, soit les vecteurs de terme « inquietude », « urgence », « incapacite », « dereglement », « crise » et « bouleversement ».

Il ressort de ces constats à l'aide des tâches d'analogie que, globalement, les termes référents à des éléments affectés par les changements climatiques, comme l'eau, la météo, la planète, les feux et la pollution sont associés à des termes qui traduisent l'urgence de la situation dans le corpus, soit « inquietude », « urgence », « dereglement », « bouleversement », « fleau », « degats » et « crise » et sont peu associés à des effets concrets locaux, qui auraient pu être illustrés, notamment, par des noms de lieux ou de personnalités victimes de ces effets.

CHAPITRE 5

INTERPRÉTATION DES RÉSULTATS

Dans ce chapitre, nous évaluons comment les résultats obtenus fournissent des réponses à la question de recherche. Nous validons également les hypothèses soulevées.

5.1 Retour sur les résultats

Ainsi, que nous apprennent les outils de sémantique vectorielle dans le cadre de l'approche d'un corpus d'articles québécois sur les changements climatiques?

5.1.1 Les plus proches voisins

Les résultats des plus proches voisins permettent de cerner les grands thèmes associés aux changements climatiques. En effet, il a été possible d'observer une récurrence du thème politique sous plusieurs enjeux, mais essentiellement près du vecteur de terme « environnement », ce qui était attendu, étant donné que plusieurs instances gouvernementales ont le terme « environnement » dans leur nom officiel. De plus, la proximité du vecteur de terme « changements-climatiques » à des vecteurs de termes à connotation négative, dont « climatosceptiques », « negatifs », « attaques », « tribunaux », soulève l'idée que l'information véhiculée sur les changements climatiques est présentée dans un contexte controversé dans les articles de presse. Cette hypothèse est renforcée par plusieurs résultats de tâches d'analogie, qui démontrent une association des enjeux des changements climatiques à l'incertitude et à l'instabilité, comme « inquietude », « urgence », « dereglement », « incapacite », « crise », « bouleversement », « cataclysme ».

Le modèle de plongement a donc cerné une tendance à connotation négative dans les plus proches voisins du vecteur de terme « changements-climatiques ». Il y a lieu de s'interroger sur les éventuelles intentions de communication des médias dans ce domaine. Considérant que le guide de déontologie journalistique du Conseil de presse commande une information exacte, rigoureuse et impartiale, entre autres, nous pouvons supposer que les médias ont pour objectif de transmettre une information la plus objective et exacte possible au public, l'objectivité consistant à décrire le plus fidèlement possible un objet ou à rapporter des faits, sans jugement et sans connotation. Dans le cas où cette objectivité serait considérée comme maintenue par les trois entreprises médiatiques durant les années couvertes par le corpus DataClimat, il est possible de soulever que le traitement de l'information sur les changements climatiques

relate des inquiétudes et une insécurité générale. Ainsi, les associations entre les vecteurs de termes « changements-climatiques » et « climatosceptiques », « negatifs », « attaques » et « tribunaux » ainsi qu'entre plusieurs enjeux des changements climatiques et « bouleversement », « cataclysme », « crise », « dereglement » et « incapacite » illustrent possiblement un biais inconscient des médias quant à la description de cette situation planétaire instable, puisque ces termes sont alarmistes et chargés de connotation négative. Le fait que le modèle les aient fait émerger automatiquement, sans annotation, à partir d'un corpus composé de trois sources d'information différentes suppose un biais culturel médiatique proclimatique qui rapporte l'urgence de la situation.

5.1.2 Les tâches d'analogie

De leur côté, les résultats des tâches d'analogie permettent de conclure, entre autres, que le modèle de plongement entrainé sur le corpus Dataclimat identifie des tendances sémantiques logiques dans le texte, puisque les résultats obtenus reflètent les connaissances générales et communes d'une citoyenne québécoise (lire ici, l'autrice de cette recherche), par exemple, sur l'association entre « feu » et « incendie ».

Cependant, même si une tâche d'analogie a fait émerger des relations implicites dans ce corpus, une critique quant au traitement de ces données soulèverait que certains termes ne sont pas suffisamment fréquents dans le texte pour influencer l'interprétation des enjeux sur les changements climatiques par les lecteurs des journaux, puisqu'il y aurait une faible probabilité de lire un terme apparaissant neuf fois dans les 972 articles de presse. Comme l'une des intuitions derrière la sémantique distributionnelle est qu'il est possible d'obtenir des éléments de sens d'un terme à l'aide de peu d'observations de contextes de ce terme, il est possible de supposer que les relations implicites capturées par le modèle de plongement représentent une fraction des possibilités de sens de ces termes et que d'autres éléments de sens sont à identifier. Le fait que le modèle associe ces termes indépendamment de leur fréquence signifie qu'il fait émerger des régularités sémantiques sous-jacentes qui peuvent également être capturées par un lectorat non initié et influencer son interprétation des enjeux sur les changements climatiques.

5.2 Retour sur les hypothèses

5.2.1 Première hypothèse

Les observations effectuées dans le cadre de la première séquence de tâches d'analogie sur différents types de relations sémantiques et syntactiques permettent de conclure que le modèle de plongement

peut faire émerger des relations extérieures à sa spécialisation. Ainsi, bien qu'il soit entrainé spécifiquement à l'aide d'articles de presse sur les changements climatiques, le fait qu'il puisse capturer d'autres informations sur la réalité à partir d'un petit corpus spécialisé corrobore les constats de Caliskan *et al.* (2017) et de Bender et Koller (2020), qui ont observé que les modèles de plongement font émerger des biais implicites dans l'apprentissage des statistiques de cotextes de termes (revoir la sous-section 2.3.4). Cependant, comme le soulignent Bender et Koller (2020), le modèle de plongement est limité dans la compréhension des intentions de communication, puisqu'il n'a pas accès aux contextes sensoriels de ces situations de communication. Dans la compréhension d'articles de presse, il faut également avoir accès au contexte historique, par exemple, quant aux actions posées par certains activistes ou aux politiques adoptées par certains gouvernements.

5.2.2 Deuxième hypothèse

Les observations générales des tâches d'analogie n'ont pas permis de conclure que les changements climatiques étaient spécifiquement associés au lexique scientifique. Il y a certes une présence marginale de termes référant aux rapports rédigés par les experts scientifiques. Cependant, une tendance des associations implicites de l'environnement au lexique du politique (acteurs politiques, institutions gouvernementales, organismes) indique que les décisions quant aux actions prises pour l'adaptation aux changements climatiques sont débattues dans la sphère publique, c'est-à-dire dans les médias.

Bien qu'il aurait été intéressant d'observer des vecteurs de termes référant aux éléments du cadre thématique de l'argumentation au sens de Luo *et al.* (2020) dans les tâches d'analogie, celles-ci n'ont pas permis d'observer des prédicats significatifs faisant la démonstration de la présence d'opinions marquées dans le discours du corpus. Néanmoins, la récurrence de l'association implicite de l'environnement à des actants politiques, comme des ministres ou des ministères, soulève une forme de point de vue politique suggérant que le politique pose des actions pour l'adaptation aux changements climatiques et véhicule une opinion proclimatique.

5.2.3 Troisième hypothèse

Il est possible d'inférer que la récurrence de certaines associations implicites alarmistes et la grande fréquence des références aux personnes et groupes proclimatiques démontrent que le point de vue proclimatique est dominant dans ce corpus et que celui-ci est présenté de manière à saisir l'urgence de la situation climatique. En effet, le fait que les éléments affectés par les changements climatiques soient

associés à des vecteurs de termes tels que « fleau », « degats », « bouleversement », « crise », « urgence », « cataclysme », « incapacite » et « dereglement » illustre les effets négatifs des changements climatiques et renforce la position proclimatique. De plus, les références à des actants politiques comme des ministres fédéraux et des ministères qui posent des actions pour l'adaptation aux changements climatiques renforce également cette position. Enfin, la référence marginale à des groupes ayant des intérêts climatosceptiques comme TransCanada démontre que ces groupes ne sont pas absents de la sphère médiatique, mais qu'ils n'y sont pas dominants. Ainsi, comme Létourneau (2014) l'a observé dans sa propre étude, le point de vue climatosceptique n'est que très peu présenté dans le corpus d'articles Dataclimat.

Il est rassurant de constater que les trois plus importantes entreprises médiatiques au Québec ne semblent pas adopter une ligne éditoriale proche de celle de leurs collègues aux États-Unis, comme The New York Times,The Wall Street Journal, et le USA Today de 1985 à 2014, qui sont reconnus pour la mise en avant des interventions climatosceptiques controversées de l'ancien président Trump et de certaines grandes entreprises exprimant des doutes sur l'existence des changements climatique (Wetts, 2020). Néanmoins, le fait que les plus proches voisins du terme « changements-climatiques » soient proches du lexique climatosceptique suggère que la couverture médiatique n'est pas homogène et reflète le débat actuel sur l'adaptation aux changements climatiques dans la société québécoise.

5.2.4 Apports et limites

Les conclusions inférées des analyses de cette étude ne reflètent pas nécessairement ce qui se passe dans la réalité, mais le fait que le modèle de plongement ait pu faire ces associations à partir d'un petit corpus spécialisé sur les changements climatiques augmente les probabilités que ce soit possible (Bryson, 2017). Puisque le modèle de plongement a une fenêtre d'apprentissage limitée sur les associations possibles aux changements climatiques, comme le corpus de textes est uniquement axé sur les changements climatiques et qu'il n'y a pas d'autres intrants sur la réalité (par exemple, des images, des odeurs, des sons), il fait émerger des associations de termes dont il est possible d'interpréter une relation spécifiquement sur ce sujet. Le modèle de plongement fonctionne également pour des tâches d'analogie extérieures aux changements climatiques, avec certaines limitations de portée (puisque le vocabulaire du corpus est limité).

De plus, l'utilisation d'une seule paire de termes de référence pour tester les analogies des termes en lien avec les changements climatiques limite la profondeur de l'analyse possible. Cette limite se reflète dans la

précision des termes obtenus, c'est-à-dire dans la similarité des dimensions des paires de vecteurs de termes. Les relations testées sur les changements climatiques étant toutes de même nature, l'angle d'analyse des résultats est le même pour toutes ces analogies, alors que ces termes auraient pu révéler de plus profondes significations en testant, par exemple, une plus grande variété de relations sémantiques et syntactiques.

Enfin, le fait que le corpus soit très peu annoté limite le biais de l'annotateur, mais les observations effectuées sur les résultats des tâches des plus proches voisins et d'analogies constituent des inférences effectuées à partir des connaissances acquises dans la réalité par la personne effectuant ces observations. Il faudrait effectuer une tâche d'associations implicites à partir de ces données auprès d'un groupe témoin afin de comparer les résultats obtenus aux résultats générés par le modèle de plongement et, ainsi, valider quantitativement sa performance.

CONCLUSION

Ce mémoire constitue un aperçu des possibilités de traitement de la sémantique vectorielle. Il démontre notamment qu'il est possible d'utiliser des modèles vectoriels comme word2vec sur un petit corpus de langue française orienté thématiquement. Le traitement de l'information contenue dans les médias québécois sur les changements climatiques rend compte des associations sémantiques auxquelles nous sommes exposés, et permet de constater comment les faits exposés dans les journaux ne nous appellent pas à réagir face aux changements climatiques. En effet, puisque l'information véhiculée sur les changements climatiques est, généralement, soit en lien avec l'écologie, soit en lien avec la prise de décision politique, les solutions proposées aux citoyens pour mettre la main à la pâte semblent marginales de leur point de vue : il semble que l'élection de partis politiques qui prônent la mise en place d'actions climatiques concrètes soit leur voie principale. Et, selon le baromètre de l'action climatique 2021, la confiance de la population québécoise envers les gouvernements quant à leur volonté d'agir est au plus bas (Champagne, 2021).

Des réflexions quant à la manière dont l'information est véhiculée s'imposent donc. Les articles de presse semblent mettre en perspective les réflexions entourant l'adaptation aux changements climatiques que soulèvent les acteurs politiques au Québec et au Canada. Ils présentent surtout les effets des changements climatiques à l'échelle régionale ou provinciale, dont les impacts sont observables sur la faune et la flore, alors que des anecdotes plus spécifiques de citoyens pourraient rejoindre les individus qui ne se sentent pas personnellement atteints par les catastrophes naturelles qui ont lieu dans une autre région ou une autre province.

L'échantillon des articles de presse de cette étude étant restreint, il serait intéressant de comparer ces résultats à ceux d'un corpus constitué des articles de 2000 à 2020, comme celui de Luo *et al.* (2020). Il serait également intéressant de pousser la recherche en comparant la production de contenu des trois entreprises médiatiques québécoises présentées précédemment, afin de dégager des différences significatives dans leur traitement de l'information. Enfin, une autre avenue de recherche serait de comparer la performance des modèles de plongement créés à partir de différents algorithmes sur un corpus en français, puisque l'état des connaissances en traitement automatique des langues se concentre essentiellement sur les corpus en anglais.

L’analyse du corpus de textes spécialisé sur les changements climatiques a permis de cerner que les termes associés aux enjeux des changements climatiques inspirent généralement un sentiment d’urgence, mais ne véhiculent pas énormément de concepts concrets, ce qui peut expliquer pourquoi les Québécois joignent peu les gestes à la parole lorsqu’il s’agit de prendre des actions concrètes pour limiter son empreinte carbone. Le choix des termes utilisés dans les articles de presse est donc crucial pour véhiculer l’information nécessaire à la compréhension des impacts des changements climatiques sur la planète. La sémantique vectorielle permet ainsi de déceler les termes privilégiés pour qualifier les événements liés aux changements climatiques rapportés dans les médias.

ANNEXE A

LISTE DES EXPRESSIONS MULTIMOTS

changements-climatiques 1429
changement-climatique 670
etats-unis 362
accord-de-paris 287
aujourd-hui 264
donald-trump 202
peut-etre 168
francois-legault 129
justin-trudeau 127
quebec-solidaire 124
taxe-carbone 121
colombie-britannique 106
nations-unies 94
1-5-degc 93
saint-laurent 89
new-york 82
catherine-mckenna 80
nouveau-brunswick 77
hydro-quebec 76
1-5 72
doug-ford 69
m-legault 64
apres-midi 58
jean-francois 57
ministere-de-l-environnement 55
manon-masse 53
mariechantal-chasse 50
rendez-vous 49
1-5degc 48
parti-liberal 47
bernie-sanders 46
barack-obama 46
centre-ville 44
valerie-plante 44
royaume-uni 44
radio-canada 40
activites-humaines 39
parti-vert 38
parti-quebecois 38
nicolas-hulot 37
antonio-guterres 35
parti-conservateur 35
nouvelle-ecosse 34
trans-mountain 33
coalition-avenir-quebec 33
patrick-bonin 33
troisieme-lien 32
philippe-couillard 32
saint-jean 32
san-francisco 31
pays-bas 31
maxime-bernier 30
andrew-scheer 28
mont-royal 27
scott-moe 27
greta-thunberg 27
activite-humaine 27
presse-canadienne 26
3-degc 26
mme-masse 26
m-hansen 25
gabriel-nadeau-dubois 24
emmanuel-macron 24
sainte-flavie 23
sylvain-gaudreault 22
universite-laval 22
el-nino 21
capitale-nationale 20
dominic-champagne 20
union-europeenne 20
mme-mckenna 20
arabie-saoudite 19
trois-rivieres 19
gouvernement-caquiste 19
climate-change 18
m-bonin 18
hillary-clinton 18
steven-guilbeault 17
jerry-brown 17
m-penner 17
michael-bloomberg 16
nouvelle-zelande 16
universite-mcgill 16
revue-nature 16
m-lisee 16
carlos-leitao 16

grande-bretagne 16
terre-neuve 16
lac-ontario 15
mme-chasse 15
m-gaudreault 15
jason-kenney 14
dominique-paquin 14
5-degc 14
catherine-potvin 14
dustin-duncan 14
ministere-des-transports 14
abitibi-temiscamingue 14
simon-tremblay-pepin 14
nicolas-marceau 14
m-martel 14
elizabeth-warren 14
amy-klobuchar 13
rouyn-noranda 13
prince-edouard 13
statu-quo 13
maxime-pedneaud-jobin 13
brise-glaces 13
ministre-wilson-raybould 13
croix-rouge 12
lac-winnipeg 12
gerald-butts 12
pete-buttigieg 12
philippe-gachon 12
gilles-brien 12
nouveau-parti-democratique 12
pierre-arcand 12
ruba-ghazal 12
marie-montpetit 12
genevieve-guilbault 12
amerique-latine 12
jair-bolsonaro 12
whistler-blackcomb 12
kamala-harris 12
sommet-sur-le-climat 11
rene-levesque 11
national-geographic 11
benoit-charette 11
burnaby-sud 11
brise-glace 11
universite-du-quebec-a-rimouski 11
kristalina-georgieva 11
jagmeet-singh 11
dominic-leblanc 11
andre-belisle 11
xxe-siecle 11
jean-charest 11
mario-dumont 11
ministre-brison 11
vincent-marissal 10
christine-hallquist 10
john-horgan 10
sept-iles 10
m-brien 10
m-guilbeault 10
xxie-siecle 10
jennifer-morgan 10
jean-jouzel 10
organisation-mondiale-de-la-sante 10
saint-pierre 10
declaration-citoyenne-universelle-d-urgence-climatique 10
monsef-derraji 10
julie-gelfand 10
mme-gelfand 10
wall-street 10
sainte-marie 10
climato-sceptique 10
mike-schreiner 10
joe-biden 10
saint-anicet 10
courant-jet 10
ban-ki 9
rocher-perce 9
brian-gallant 9
diane-dufresne 9
catherine-dorion 9
brian-pallister 9
adam-mckay 9
saint-hyacinthe 9
jody-wilson-raybould 9
simon-legault 9
chaudiere-appalaches 9
buenos-aires 9
los-angeles 9
universite-d-ottawa 9
marie-eve 9
yan-boulanger 9
rod-phillips 9
angela-merkel 9
francois-philippe 9
dick-cheney 9

marquis-bissonnette 9
julian-castro 9
alexandria-ocasio-cortez 9
louis-couillard 9
jean-christophe 9
communaute-metropolitaine-de-montreal 8
jean-philippe 8
alain-rayes 8
gulf-stream 8
caisse-de-depot-et-placement-du-quebec 8
etienne-leblanc 8
fort-mcmurray 8
alain-webster 8
john-roome 8
karel-mayrand 8
youri-chassin 8
daniel-green 8
agence-france-presse 8
mme-garneau 8
annie-chaloux 8
rodrigue-turgeon 8
eric-martel 8
hindou-oumarou 8
nouvelle-angleterre 8
kim-jong 8
lac-erie 8
tulsi-gabbard 8
cory-booker 8
south-bend 8
sara-montpetit 8
etienne-kapikian 8
ministere-des-finances 8
saint-barthelemy 8
bill-gates 7
sud-africaine 7
neo-brunswickois 7
chauve-souris 7
ouragan-michael 7
emilise-lessard-therrien 7
victoria-beckham 7
fashion-week 7
afrique-australe 7
marie-claude 7
universite-du-quebec-a-montreal 7
sylvia-kreutzer 7
saint-michel 7
christian-dube 7
mont-bleu 7
saint-jacques 7
grands-parents 7
kathleen-wynne 7
amerique-centrale 7
nick-sloane 7
petteri-taalas 7
universite-concordia 7
damon-matthews 7
adam-sprott 7
jacob-lebel 7
natalie-hasell 7
jay-inslee 7
affaire-snc-lavalin 7
leger-boyer 7
mme-wilson-raybould 7
ministre-philpott 7
alain-royer 7
zero-dechet 6
rachel-notley 6
saint-julien 6
arnold-schwarzenegger 6
britanno-colombiens 6
eric-girard 6
romeo-saganash 6
sainte-marthe 6
sol-zanetti 6
dame-nature 6
isabelle-melancon 6
baie-james 6
universite-catholique 6
in-fine 6
patricia-espinosa 6
tara-buakamsri 6
stephen-harper 6
new-delhi 6
ian-mauro 6
paul-romer 6
agence-qmi 6
agence-internationale-de-l-energie 6
gaetan-barrette 6
st-pierre 6
dominique-anglade 6
francois-bonnardel 6
don-iveson 6
van-oldenborgh 6
cols-bleus 6
bassin-mediterraneen 6
science-fiction 6

sidney-ribaux 6
parry-sound 6
harjit-sajjan 6
blaine-higgs 6
serge-bourgeois 6
roger-cooke 6
hec-montreal 6
sainte-anne 6
vladimir-poutine 6
3e-lien 6
boyer-villemaire 6
seyni-nafo 6
david-phillips 6
shirley-mainprize 6
dianne-saxe 6
miguel-anxo 6
georges-beaudoin 6
daniel-boyer 6
san-antonio 6
robert-gravel 6
da-silva 6
mike-layton 6
peter-demarsh 6
petrin-desrosiers 6
snc-lavalin 6
jorge-muller 6
me-hunter 6
christophe-cloutier-roy 6
universite-simon-fraser 5
1-5deg 5
conseil-de-l-industrie-forestiere-du-quebec 5
nadeau-dubois 5
institut-national-de-sante-publique-du-quebec 4
direction-de-la-sante-publique 4
ministre-des-affaires-intergouvernementales 4
association-quebecoise-de-lutte-contre-la-pollution-atmospherique 4
ministre-du-developpement-durable 3
centre-climatique-des-prairies 3
autorite-regionale-de-transport-metropolitain 3
pedneaud-jobin 3
15-degc 3
new-yorkais 3
cloutier-roy 3
commission-scolaire-de-montreal 2
ministre-de-l-environnement-de-l-ontario 2
neo-brunswickoises 2
31-5 2
33-degc 2
prince-edouardiennes 1
international-pour-l-analyse-des-systemes-appliques 1
communaute-economique-des-etats-d-afrique-de-l-ouest 1
changement-climatiques 1
loi-sur-la-gestion-et-la-reduction-des-ga-Za-effet-de-serre 1
tremblay-pepin 1
new-yorker 1
35-degc 1
sud-africaines 1
51-5 1
1-53 1
new-yorkaise 1
1-500 1
25-degc 1
neo-brunswickoise 1
wilson-raybould 1
lessard-therrien 1
ocasio-cortez 1

ANNEXE B

MANIPULATIONS DANS RSTUDIO

```
library(word2vec)

# Definition du dossier de travail

setwd("C:/Users/Public/Documents/VectorSemantics-Exp/MyScripts")

# Creation d'un nouveau modele sur la base du contenu du fichier vmDataClimat

model <- word2vec(x = "vmDataClimat.txt", type = "cbow", dim = 200, iter = 10)

# Sauvegarde du modele dans un fichier

write.word2vec(model, "vmDataClimat.bin")

# Chargement d'un modele depuis un fichier

model <- read.word2vec("vmDataClimat.bin")

# Les plus proches voisins d'un ensemble de termes

lookslike <- predict(model, c("climat", "environnement"), type = "nearest", top_n = 10)

lookslike

# Tâche d'analogies

wv <- predict(model, newdata = c("climat", "rechauffement", "planete"), type = "embedding")

wv <- wv["climat", ] - wv["rechauffement", ] + wv["planete", ]

predict(model, newdata = wv, type = "nearest", top_n = 10)
```

ANNEXE C

TABLEAUX D'ANALOGIES SUR LES RELATIONS TESTÉES ET SUR LES CHANGEMENTS CLIMATIQUES

Tableau C.1 Les vecteurs de terme : citoyens - villes + entreprises = y

Vecteurs de terme	**Mesure de similarité du cosinus**
entreprises	0,9953823
canadiens	0,9349531
politiciens	0,9296143
options	0,9092650
demandes	0,9065172
consommateurs	0,9020304
conseils	0,9003477
initiatives	0,8989437
elus	0,8967922
gouvernements	0,8957154

Tableau C.2 Les vecteurs de terme : montreal - quebec + ontario = y

Vecteurs de terme	**Mesure de similarité du cosinus**
ontario	0,9693330
montreal	0,8842540
alberta	0,8435338
hotel	0,8010201
economie	0,7915492
saskatchewan	0,7900424
assise	0,7831272
toronto	0,7829541
armee	0,7736651
ottawa	0,7708693

Tableau C.3 Les vecteurs de terme : trudeau - canada + quebec = y

Vecteurs de terme	Mesure de similarité du cosinus
couillard	0,9704525
liberal	0,9551017
ford	0,9504787
legault	0,9378312
quebec	0,9200084
federal	0,9003798
offert	0,8637405
sortant	0,8511202
parti-quebecois	0,8471630
albertain	0,8445598

Tableau C.4 Les vecteurs de terme : nous - je + tu = y

Vecteurs de terme	Mesure de similarité du cosinus
vous	0,8385008
investies	0,8293469
tu	0,8091211
mme-wilson-raybould	0,7860416
agi	0,7795014
accompli	0,7567810
vaut	0,7520407
te	0,7504119
placee	0,7496921
moi	0,7365866

Tableau C.5 Les vecteurs de terme : il - le + la = y

Vecteurs de terme	Mesure de similarité du cosinus
la	0,9545819
il	0,9360896
elle	0,8189526
intelligente	0,7750613
lac-winnipeg	0,7483875
telle	0,7398089
evidente	0,7331821
slogan	0,7280744
civile	0,7260361
logement	0,7235857

Tableau C.6 Les vecteurs de terme : moi - me + te = y

Vecteurs de terme	Mesure de similarité du cosinus
te	0,9652900
moi	0,8863127
urgent	0,8767092
injuste	0,8371704
agi	0,8283274
ministre-brison	0,8147601
eue	0,8009323
fatalite	0,7918536
ideologique	0,7892742
difficile	0,7821628

Tableau C.7 Les vecteurs de terme : vendre – vend + achete = y

Vecteurs de terme	Mesure de similarité du cosinus
achete	0,9066963
avions	0,8557397
respecter	0,8417833
prenons	0,8408071
ramener	0,8390045
fonctionner	0,8348119
devoiler	0,8304691
entiere	0,8278037
realiser	0,8239516
suivre	0,8235935

Tableau C.8 Les vecteurs de terme : initiatives - initiative + solution = y

Vecteurs de terme	Mesure de similarité du cosinus
initiatives	0,9830231
solutions	0,9739804
habitudes	0,9593960
facons	0,9389681
delais	0,9206732
moyens	0,9193389
solution	0,9180979
mineurs	0,9176983
taxes	0,9160677
ordures	0,9133504

Tableau C.9 Les vecteurs de terme : citoyen - citoyenne + rectrice = y

Vecteurs de terme	Mesure de similarité du cosinus
damon-matthews	0,9962849
rebecca	0,9867713
guy	0,9860172
harrison	0,9808043
militant	0,9799221
cofondateur	0,9772682
barnes	0,9723589
adjoint	0,9717188
shawn	0,9714425
philippe	0,9691300

Tableau C.10 Les vecteurs de terme : vendre - vente + achat = y

Vecteurs de terme	Mesure de similarité du cosinus
vendre	0,9658133
achat	0,9178868
etranger	0,8531471
elaboration	0,8423592
electricite	0,8384490
payer	0,8361267
incitatifs	0,8334022
utiliser	0,8176484
argent	0,8172807
acheter	0,8137206

Tableau C.11 Les vecteurs de terme : decideurs - acteurs + decision = y

Vecteurs de terme	Mesure de similarité du cosinus
proposition	0,9957581
promesse	0,9875244
session	0,9833546
lancee	0,9804779
motion	0,9734414
resolution	0,9729390
tribune	0,9575054
carriere	0,9535725
entendue	0,9486470
tenue	0,9423397

Tableau C.12 Les vecteurs de terme : urgent - urgence + prudence = y

Vecteurs de terme	Mesure de similarité du cosinus
prudence	0,9900379
lache	0,9614322
urgent	0,9359829
pollen	0,9020900
invoque	0,8970833
contient	0,8840512
effrayant	0,8729374
couvrir	0,8692090
amene	0,8662255
ambitieuse	0,8659008

Tableau C.13 Les vecteurs de termes : rechauffement - temperature + climat = y

Vecteurs de terme	Mesure de similarité du cosinus
changementZZZclimatique	0,8759751
mouvement	0,8642596
rechauffement	0,8189713
consensus	0,8105184
pacte	0,8074388
rassemblement	0,8021005
patrimoine	0,7861190
sommet	0,7819625
message	0,7778505
groupe	0,7746851

Tableau C.14 Les vecteurs de termes : rechauffement – temperature + carbone = y

Vecteurs de terme	Mesure de similarité du cosinus
recours	0,8359355
pipeline	0,8106746
rechauffement	0,8105464
fardeau	0,8082898
cout	0,8064814
plan	0,7992349
allie	0,7911131
système	0,7909755
code	0,7899674
autocollant	0,7818742

Tableau C.15 Les vecteurs de termes : rechauffement – temperature + fossile = y

Vecteurs de terme	Mesure de similarité du cosinus
deficit	0,9458306
collectif	0,9043016
instigateur	0,8835917
eolien	0,8811041
anti	0,8733441
outil	0,8713215
financement	0,8599728
one	0,8592991
pipeline	0,8588142
completer	0,8518162

Tableau C.16 Les vecteurs de termes : rechauffement – temperature + energie = y

Vecteurs de terme	Mesure de similarité du cosinus
energie	0,9859743
outil	0,8687105
collectif	0,8613461
investissement	0,8524832
exporter	0,8434327
atlas	0,8347761
deficit	0,8043363
pipeline	0,7975984
exploitation	0,7960253
infrastructure	0,7953026

Tableau C.17 Les vecteurs de termes : rechauffement – temperature + co2 = y

Vecteurs de terme	Mesure de similarité du cosinus
co2	0,9196218
rechauffement	0,9129369
changement	0,8183014
gaz	0,7826429
cout	0,7809001
methane	0,7634210
carbone	0,7622710
système	0,7406502
seuil	0,7390013
aspect	0,7324806

Tableau C.18 Les vecteurs de termes : rechauffement – temperature + glace = y

Vecteurs de terme	Mesure de similarité du cosinus
rechauffement	0,8437069
glace	0,8039216
wilkes	0,7575947
fond	0,7253675
recouvertes	0,7230160
sol	0,7162269
continent	0,7129677
blanc	0,7092164
froid	0,7020354
couvert	0,7004744

Tableau C.19 Les vecteurs de termes : rechauffement – temperature + glaciers = y

Vecteurs de terme	Mesure de similarité du cosinus
rechauffement	0,9059276
glaciers	0,9021899
pergelisol	0,7781147
feux	0,7770150
saumons	0,7762617
insectes	0,7737835
dereglement	0,7693146
predateurs	0,7538246
incendies	0,7483587
ecosysteme	0,7368394

Tableau C.20 Les vecteurs de termes : rechauffement – temperature + fonds = y

Vecteurs de terme	Mesure de similarité du cosinus
financement	0,9958919
programme	0,9717232
budget	0,9158342
soutien	0,9009686
pipeline	0,8832558
completer	0,8792636
programmes	0,8726779
deficit	0,8707485
vert	0,8486674
rechauffement	0,8459783

Tableau C.21 Les vecteurs de termes : rechauffement – temperature + oceans = y

Vecteurs de terme	Mesure de similarité du cosinus
oceans	0,8691870
rechauffement	0,8020908
peches	0,7968996
patrimoine	0,7582237
douter	0,7391065
maldives	0,7323740
dereglement	0,7289661
pergelisol	0,7265625
dangers	0,7260602
atlas	0,7203771

Tableau C.22 Les vecteurs de termes : rechauffement – temperature + chaleur = y

Vecteurs de terme	Mesure de similarité du cosinus
chaleur	0,9652190
rechauffement	0,9440477
dereglement	0,8763437
crise	0,8527004
feux	0,8048381
phenomene	0,7975026
incendies	0,7659866
extreme	0,7539474
inondations	0,7486267
changement	0,7421832

Tableau C.23 Les vecteurs de termes : rechauffement – temperature + serre = y

Vecteurs de terme	Mesure de similarité du cosinus
serre	0,9859563
rechauffement	0,8326391
domino	0,8144056
minimal	0,7573437
deficit	0,7371917
gaz	0,7069784
transcanada	0,7066121
liquefie	0,7001954
climatosceptique	0,6990115
lobby	0,6907240

Tableau C.24 Les vecteurs de termes : rechauffement – temperature + ges = y

Vecteurs de terme	Mesure de similarité du cosinus
ges	0,9522771
rechauffement	0,8335836
cibles	0,8145530
gaz	0,8115319
deficit	0,8067668
partizzzquebecois	0,7664381
plan	0,7628519
pq	0,7592442
cout	0,7527078
eliminer	0,7526699

Tableau C.25 Les vecteurs de termes : rechauffement – temperature + arctique = y

Vecteurs de terme	Mesure de similarité du cosinus
arctique	0,9136785
horloge	0,8342263
ecosysteme	0,8196276
rechauffement	0,7972629
atlas	0,7941623
hemisphere	0,7928714
dereglement	0,7825447
embleme	0,7822846
continent	0,7699782
pacifique	0,7697948

Tableau C.26 Les vecteurs de termes : rechauffement – temperature + antarctique = y

Vecteurs de terme	Mesure de similarité du cosinus
antarctique	0,8748351
rechauffement	0,8658162
ecosysteme	0,8518405
continent	0,8326745
horloge	0,8083361
wilkes	0,8007272
himalaya	0,7913609
atlas	0,7869480
aspect	0,7839478
cercle	0,7786871

Tableau C.27 Les vecteurs de termes : rechauffement – temperature + environnement = y

Vecteurs de terme	Mesure de similarité du cosinus
environnement	0,9631104
atlas	0,8217234
infrastructure	0,8062977
harjitzzzsajjan	0,8059306
education	0,7830722
egalite	0,7677876
dominiczzzleblanc	0,7648205
apogee	0,7471917
activistes	0,7431836
greenpeace	0,7416950

Tableau C.28 Les vecteurs de termes : rechauffement – temperature + charbon = y

Vecteurs de terme	Mesure de similarité du cosinus
charbon	0,9897351
petrole	0,8893552
batiment	0,8699588
pipeline	0,8632236
rechauffement	0,8576999
nom	0,8458229
bout	0,8375742
recyclage	0,8327684
deficit	0,8230218
financement	0,8221400

Tableau C.29 Les vecteurs de termes : rechauffement – temperature + nature = y

Vecteurs de terme	Mesure de similarité du cosinus
nature	0,9432575
rechauffement	0,7955876
dereglement	0,7861788
bouleversement	0,7838266
cataclysme	0,7835886
wwf	0,7793288
science	0,7670752
lobby	0,7649409
crise	0,7630923
douter	0,7370358

Tableau C.30 Les vecteurs de termes : rechauffement – temperature + economique = y

Vecteurs de terme	**Mesure de similarité du cosinus**
forum	0,8815668
environnemental	0,8621771
ontarien	0,8540826
instaurer	0,8522288
pacte	0,8507944
collectif	0,8490881
socio	0,8413746
patrimoine	0,8410239
deficit	0,8400069
commercial	0,8387534

Tableau C.31 Les vecteurs de termes : rechauffement – temperature + electricite = y

Vecteurs de terme	**Mesure de similarité du cosinus**
electricite	0,8981899
eliminer	0,8267787
pipeline	0,8131119
usage	0,8127108
exploitation	0,8061821
investissement	0,8021775
achat	0,7969126
incitatifs	0,7966514
energie	0,7939980
interdire	0,7889534

Tableau C.32 Les vecteurs de termes : rechauffement – temperature + sante = y

Vecteurs de terme	**Mesure de similarité du cosinus**
sante	0,9349015
direct	0,8668083
sensibiliser	0,8390656
alerter	0,8226991
resilience	0,8211818
dereglement	0,8073688
lobby	0,7980821
religieux	0,7926375
socio	0,7925506
expertise	0,7918607

Tableau C.33 Les vecteurs de termes : rechauffement – temperature + scientifiques = y

Vecteurs de terme	Mesure de similarité du cosinus
scientifiques	0,9588935
experts	0,9049238
rapports	0,8832440
rechauffement	0,8784183
alarme	0,8581539
tsunamis	0,8469436
auteurs	0,8455456
migrations	0,8283267
conclusions	0,8258854
dereglement	0,8253973

Tableau C.34 Les vecteurs de termes : rechauffement – temperature + pollution = y

Vecteurs de terme	Mesure de similarité du cosinus
pollution	0,9187546
rechauffement	0,8886423
cout	0,8150091
deficit	0,8100230
regime	0,8008260
dereglement	0,7993311
système	0,7680250
vert	0,7605866
fleau	0,7532179
bouleversement	0,7459240

Tableau C.35 Les vecteurs de termes : rechauffement – temperature + feux = y

Vecteurs de terme	Mesure de similarité du cosinus
incendies	0,9549279
rechauffement	0,8826839
boreale	0,8675279
feu	0,8357770
dereglement	0,8303149
crise	0,7727180
degats	0,7712488
verre	0,7629069
incendie	0,7554567
violents	0,7543737

Tableau C.36 Les vecteurs de termes : rechauffement – temperature + giec = y

Vecteurs de terme	Mesure de similarité du cosinus
rechauffement	0,9172747
appellent	0,9071994
consensus	0,9059008
recent	0,9047388
dereglement	0,8990670
minimal	0,8768389
lobby	0,8682846
groupe	0,8608575
cataclysme	0,8600235
mouvement	0,8569091

Tableau C.37 Les vecteurs de termes : rechauffement – temperature + marche = y

Vecteurs de terme	Mesure de similarité du cosinus
mouvement	0,9114959
rassemblement	0,9074042
dossier	0,8900143
deficit	0,8848814
lobby	0,8763220
pacte	0,8642131
ontarien	0,8633977
nom	0,8583806
sommet	0,8508633
rechauffement	0,8473521

Tableau C.38 Les vecteurs de termes : rechauffement – temperature + planete = y

Vecteurs de terme	Mesure de similarité du cosinus
planete	0,7518640
cataclysme	0,7374101
rechauffement	0,7174968
mouvement	0,7160303
dereglement	0,7135607
changement-climatique	0,6973752
patrimoine	0,6955723
climat	0,6916475
consensus	0,6867291
alarme	0,6819122

Tableau C.39 Les vecteurs de termes : rechauffement – temperature + gaz = y

Vecteurs de terme	Mesure de similarité du cosinus
rechauffement	0,8942825
deficit	0,8063808
domino	0,7848002
ramener	0,7642383
outil	0,7618791
changement	0,7574379
ges	0,7446054
donner	0,7330379
adoptees	0,7307883
pipeline	0,7281364

Tableau C.40 Les vecteurs de termes : rechauffement – temperature + emissions = y

Vecteurs de terme	Mesure de similarité du cosinus
emissions	0,9287351
rechauffement	0,8685114
objectifs	0,7786261
deficit	0,7756656
cibles	0,7496645
efforts	0,7356661
eliminer	0,7217869
minimal	0,7126138
plan	0,7124218
factures	0,7082435

Tableau C.41 Les vecteurs de termes : rechauffement – temperature + accord = y

Vecteurs de terme	Mesure de similarité du cosinus
action	0,9496175
deficit	0,9320658
minimal	0,9295279
aspect	0,9153943
investissement	0,9084327
echec	0,8873239
horloge	0,8759647
rechauffement	0,8754583
œil	0,8746275
urgence	0,8731332

Tableau C.42 Les vecteurs de termes : rechauffement – temperature + forets = y

Vecteurs de terme	Mesure de similarité du cosinus
forets	0,9291949
insectes	0,8027015
principalement	0,7902309
parcs	0,7895726
maladies	0,7847496
ecosystemes	0,7833319
boreale	0,7821259
zones	0,7769358
rechauffement	0,7739993
terres	0,7708734

Tableau C.43 Les vecteurs de termes : rechauffement – temperature + atmosphere = y

Vecteurs de terme	Mesure de similarité du cosinus
atmosphere	0,9127930
rechauffement	0,8203665
immediat	0,8047813
changement	0,7961333
atlas	0,7530248
incapacite	0,7515857
ecosysteme	0,7460533
capter	0,7456804
auto	0,7437480
max	0,7431774

Tableau C.44 Les vecteurs de termes : rechauffement – temperature + agriculture = y

Vecteurs de terme	Mesure de similarité du cosinus
agriculture	0,9817982
implantation	0,8651520
industrie	0,8553855
assistance	0,8542352
ecologie	0,8197597
accessibilite	0,8175422
exploitation	0,8145549
technologique	0,8027503
biologique	0,7970988
urbanisme	0,7957860

Tableau C.45 Les vecteurs de termes : rechauffement – temperature + eau = y

Vecteurs de terme	**Mesure de similarité du cosinus**
eau	0,9102408
rechauffement	0,7956253
inquietude	0,7497285
oxygene	0,7480062
atlas	0,7393367
urgence	0,7377059
incapacite	0,7215405
interet	0,7191010
iceberg	0,7120664
aspect	0,7112084

Tableau C.46 Les vecteurs de termes : rechauffement – temperature + meteo = y

Vecteurs de terme	**Mesure de similarité du cosinus**
dereglement	0,9858649
meteo	0,9793733
rechauffement	0,9439958
crise	0,8993381
horloge	0,8609118
vivante	0,8541189
atlas	0,8351483
survenue	0,8221085
bouleversement	0,8144285
simulation	0,8047851

RÉFÉRENCES

Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, *9*(3), 247-256. https://doi.org/10.1080/17405904.2012.688297

Baron, J. (2006). Thinking About Global Warming. *Climatic Change*, *77*(1-2), 137-150. https://doi.org/10.1007/s10584-006-9049-y

Baroni, M., Dinu, G. et Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 238-247). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1023

Bender, E. M. et Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A. et Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (p. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Bender, E. M. et Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 5185-5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bengio, Y., Courville, A. et Vincent, P. (2012). Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]*. http://arxiv.org/abs/1206.5538

Bérubé, C. (2010). *Changements climatiques et distorsion de la perception des Québécois: de la communication à l'action* [essai, Université de Sherbrooke].

Bryson, J. J. (2017, 13 avril). We Didn't Prove Prejudice Is True (A Role for Consciousness). *We Didn't Prove Prejudice Is True (A Role for Consciousness)*. https://joanna-bryson.blogspot.com/2017/04/we-didnt-prove-prejudice-is-true-role.html

Bureau de projet des changements climatiques. (2015). *Évaluation des impacts des changements climatiques et de leurs coûts pour le Québec et l'État québécois*, 97.

Caliskan, A., Bryson, J. J. et Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Champagne, É.-P. (2021, 8 décembre). Baromètre de l'action climatique 2021 | Les Québécois jugent sévèrement les gouvernements. *La Presse*. https://www.lapresse.ca/actualites/environnement/2021-12-08/barometre-de-l-action-climatique-2021/les-quebecois-jugent-severement-les-gouvernements.php

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. Dans *Mechanism of Language Acquisition*.

Comby, J.-B. (2012). Chapitre 7 - Les médias face aux controverses climatiques en Europe: un consensus fragilisé mais toujours structurant. Dans E. Zaccaï, F. Gemenne, J.-M. Decroly et V. Masson-Delmotte (dir.), *Controverses climatiques, sciences et politique*. Presses de la Fondation nationale des sciences politiques.

Conseil de presse du Québec. (2015). *Guide de déontologie journalistique du Conseil de presse du Québec.* http://epe.lac-bac.gc.ca/100/200/300/conseil_presse_qc/guide_deontologie/index.html

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P. et Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2), 024024. https://doi.org/10.1088/1748-9326/8/2/024024

CROTEAU, M. (2016, 14 mars). TransCanada torpille indirectement le plan climatique d'Obama. *La Presse*, section Environnement. https://www.lapresse.ca/environnement/201603/14/01-4960816-transcanada-torpille-indirectement-le-plan-climatique-dobama.php

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

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. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

de Marcellis-Warin, N., Peignier, I., Hoang Bui, M., F. Anjos, M., A. Gabriel, S. et Guerra, C. (2015, mai). *L'énergie et les changements climatiques: perceptions québécoises*. CIRANO et IET. https://www.cirano.qc.ca/files/publications/2015RP-08.pdf

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. http://arxiv.org/abs/1810.04805

Friedman, B. et Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330-347. https://doi.org/10.1145/230538.230561

Gastaldi, J. L. (2020). Why Can Computers Understand Natural Language? *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00393-9

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. Dans *Proceedings of the NAACL Student Research Workshop* (p. 8-15). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2002

Good, J. E. (2008). The Framing of Climate Change in Canadian, American, and International Newspapers: A Media Propaganda Model Analysis. *Canadian Journal of Communication*, *33*(2), 233-255.

Gouvernement du Québec. (2015). *Comment agir*. Faisons-le pour eux. https://www.faisonslepoureux.gouv.qc.ca/fr/comment-agir

Greenwald, A. G., McGhee, D. E. et Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* (University of Washington), section 6, 1464-1480.

Groupe TVA. (2018). *Profil de société | Groupe TVA*. http://www.groupetva.ca/legroupe/profil-societe

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Herman, E. S. et Chomsky, N. (2002). *Manufacturing consent: the political economy of the mass media*. Pantheon Books.

Hill, F., Cho, K., Jean, S., Devin, C. et Bengio, Y. (2015). Embedding Word Similarity with Neural Machine Translation. *arXiv:1412.6448 [cs]*. http://arxiv.org/abs/1412.6448

Ho-Dac, L.-M. et Küppers, A. (2011). La subjectivité à travers les médias: étude comparée de les médias participatifs et de la presse traditionnelle. *Corpus*, *10*, 179-199.

IPCC - Intergovernmental Panel on Climate Change. (2018). *Activités de l'organisation*. http://www.ipcc.ch/home_languages_main_french.shtml

Jurafsky, D. et Martin, J. H. (2019). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Third Edition draft). University of Standford. https://web.stanford.edu/~jurafsky/slp3/

Jurgens, D., Mohammad, S., Turney, P. et Holyoak, K. (2012). SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. Dans **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)* (p. 356-364). Association for Computational Linguistics. https://aclanthology.org/S12-1047

Kennedy, C. et McNally, L. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, *81*(2), 345-381. https://doi.org/10.1353/lan.2005.0071

La Presse. (2018). *À propos de nous*. La Presse. https://www.lapresse.ca/a-propos-de-nous/la-presse/

Lai, S., Liu, K., He, S. et Zhao, J. (2016). How to Generate a Good Word Embedding. *IEEE Intelligent Systems*, *31*(6), 5-14. https://doi.org/10.1109/MIS.2016.45

Lehrer, A. (2012). A theory of meaning. *Philosophical Studies*, *161*(1), 97-107. https://doi.org/10.1007/s11098-012-9934-3

Létourneau, A. (2014). Figures et importance de l'« expertise environnementale » dans la presse écrite. *VertigO - la revue électronique en sciences de l'environnement*, (Volume 14 Numéro 1). https://doi.org/10.4000/vertigo.14702

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. Dans *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (p. 13-18). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2503

Luo, Y., Card, D. et Jurafsky, D. (2020). DeSMOG: Detecting Stance in Media On Global Warming. Dans *Findings of the Association for Computational Linguistics: EMNLP 2020* (p. 3296-3315). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.296

Manning, C. D. et Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martin, S. (2020). L'opinion publique sur le climat en France. *Futuribles*, *N° 435*(2), 35-55.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Le, Q. V. et Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation (version 1). *arXiv:1309.4168 [cs]*. http://arxiv.org/abs/1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013c). Distributed Representations of Words and Phrases and their Compositionality. Dans C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K. Q. Weinberger (dir.), *Advances in Neural Information Processing Systems 26* (p. 3111-3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Mikolov, T., Yih, W. et Zweig, G. (2013d). Linguistic Regularities in Continuous Space Word Representations. Dans *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 746-751). Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1090

Mildenberger, M., Howe, P., Lachapelle, E., Stokes, L., Marlon, J. et Gravelle, T. (2016). The Distribution of Climate Change Public Opinion in Canada. *PLOS ONE*, *11*(8), e0159774. https://doi.org/10.1371/journal.pone.0159774

Nayak, N. (2015). *Learning Hypernymy over Word Embeddings*. Stanford. https://cs224d.stanford.edu/reports/NayakNeha.pdf

Pennington, J., Socher, R. et Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. https://nlp.stanford.edu/projects/glove/

Radio-Canada. (2018). *Découvrez CBC/Radio-Canada*. Radio-Canada. http://www.cbc.radio-canada.ca/fr/decouvrez/

Ressources naturelles Canada. (2013, 7 octobre). *Changements climatiques*. Gouvernement du Canada. https://www.rncan.gc.ca/environnement

Robert, P.-Y. (2018, 20 avril). *Voici l'état du lectorat des quotidiens et des magazines au Québec en 2017*. Infopresse. https://www.infopresse.com/article/2018/4/20/vividata

S. Perone, C. (2013, 12 septembre). Machine Learning :: Cosine Similarity for Vector Space Models (Part III). *Terra Incognita*. http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Schmidt, B. (2015, 25 octobre). Vector Space Models for the digital humanities. http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html

Scott, W. (2019, 21 mai). *TF-IDF for Document Ranking from scratch in python on real world dataset.* Medium. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

Spärck Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *60*(5), 493-502. https://doi.org/10.1108/00220410410560573

Tirtha. (2020, 24 avril). *Word Embeddings vs TF-IDF: Answering COVID-19 Questions*. Medium. https://towardsdatascience.com/word-embeddings-vs-tf-idf-answering-covid-19-questions-703e3d99f783

Trésor, D. générale du. (2021, 21 juillet). *La politique du Nigéria face au changement climatique*. Direction générale du Trésor. https://www.tresor.economie.gouv.fr/Articles/2021/07/21/la-politique-du-nigeria-face-au-changement-climatique

Turney, P. D. (2008). A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. *arXiv:0809.0124 [cs]*. http://arxiv.org/abs/0809.0124

Wetts, R. (2020). In climate news, statements from large businesses and opponents of climate action receive heightened visibility. *Proceedings of the National Academy of Sciences*, *117*(32), 19054-19060. https://doi.org/10.1073/pnas.1921526117

Wodak, R. (dir.). (1989). *Language, power and ideology: studies in political discourse*. Benjamins.