La contribution des outils de Machine Learning et de Deep Learning à l'analyse de textes historiques : l'apport particulier de la sémantique quantitative

Yves Noblet

Année académique 2022/2023

Résumé

La sémantique quantitative constitue un domaine de recherche utile pour la recherche historique. Dans ce contexte, les outils de Machine Learning et de Deep Learning sont particulièrement intéressant et permettent d'entreprendre une grande variété d'études. Ainsi, après avoir définis en quoi consistent le Machine Learning et le Deep Learning, cet article vise à montrer ce que peuvent apporter de tels outils dans des études de sémantique quantitative en histoire.

1 Introduction

Ce travail a pour but de voir comment l'on peut saisir le sens des mots de manière quantitative, autrement dit faire de la sémantique quantitative, à l'aide d'outils informatiques. Le but est également de montrer quelques exemples du genre de travail que l'on peut réaliser dans ce domaine, non seulement sur des corpus de textes en langues modernes, mais aussi en langues anciennes. Comme nous allons le voir, les études de sémantique quantitative ne sont pas sans intérêt pour la recherche en histoire. Le sens des mots évoluant dans le temps, cela implique, par exemple, qu'un historien effectuant une recherche sur un sujet donné pourra ou non incorporer tel ou tel texte comportant les mots clés nécessaires à sa recherche selon que leur sémantique correspond bien à son sujet. La sémantique quantitative apporte donc un certain soutien au chercheur en histoire [9, p. 1]. Mais l'intérêt des études sémantiques ne s'arrête pas là, en connaissant le sens des mots, on peut effectuer des tâches de classification de textes [9, p. 1]. On peut également voir quels sont les termes associés à certains mots, ou encore saisir la similarité entre plusieurs termes. On peut voir à quel genre de textes certains mots sont le plus associés [12, p. 1]. Les possibilités d'études à l'aide d'outils de sémantique quantitative sont donc assez nombreuses et variées.

Afin d'effectuer de la sémantique quantitative, il faut faire appel à des systèmes de Machine Learning, ou encore de Deep Learning. Le Deep Learning n'étant rien de moins qu'un sous-ensemble du Machine Learning. Les systèmes de Machine Learning relèvent de l'intelligence artificielle. Avant de poursuivre, il convient de consacrer une section à la définition de ces notions d'intelligence artificielle, de Machine Learning et de Deep Learning. Une seconde section montre des exemples d'application de l'analyse sémantique dans le domaine de la recherche historique. En premier lieu, quelques exemples d'application sur des textes en langues anciennes et en second lieu sur des textes dans des langues plus modernes.

2 Définitions

2.1 L'"intelligence artificielle"

L'intelligence artificielle est une branche de l'informatique qui cherche à incorporer l'intelligence humaine dans des machines. On crée ainsi des systèmes permettant d'effectuer des tâches complexes pour lesquels on requière en temps normal cette intelligence humaine. Ils fonctionnent sur base d'algorithmes et de règles, avec un minimum d'intervention humaine [6, p. 686] et [5, pp. 131-132]. L'intelligence artificielle est donc une expression générique pour tout programme informatique ayant une forme d'intelligence humaine. Elle englobe à la fois le Machine Learning et le Deep Learning[5, p. 132].

2.2 Le Machine Learning

Deepak Jakhar et Ishmeet Kaur expliquent que le Machine Learning regroupe « toutes les approches permettant aux machines d'apprendre à partir de données sans être explicitement programmées » [5, p. 132]. Elles tournent sur base d'algorithmes et données et c'est via ces données et les informations traitées que les machines apprennent à prendre des décisions. Elles peuvent se modifier en étant exposées à plus de données. Par « learning » («apprentissage») on entend qu'elles font en sorte qu'il y ait le moins d'erreurs et que leurs prédictions soient le plus juste possible [5, p. 132]. Le Machine Learning cherche à effectuer des tâches cognitives grâce à des modèles analytiques construits automatiquement, le but étant, par exemple, de détecter des objets, traduire du langage. Les ordinateurs peuvent trouver des idées cachées et des modèles complexes, tout cela, il faut le rappeler, sans être explicitement programmés. Toutes ces tâches sont effectuées via des algorithmes qui, de manière itérative, font un apprentissage à partir de données d'entraînement [6, p. 686].

À partir d'un ensemble d'observation, un ensemble d'apprentissages, le Machine Learning a pour but d'obtenir une fonction de prédiction, la construction de cette fonction constitue l'apprentissage, ou l'entraînement du modèle [8, p. 115]. Selon Ted Dunning, figure de proue du Machine Learning, un algorithme de

Machine Learning doit avoir principalement les cinq qualités suivantes : il doit être facilement déployé, robuste, transparent (lorsque les performances d'une application dotée d'un Machine Learning se dégradent, l'algorithme doit le détecter le plus vite possible), adéquat aux compétences du personnel disponible, le temps et l'énergie investis doivent être proportionnels aux bénéfices, enfin il doit être performant [8, pp. 117-118]. Par ailleurs, on juge la qualité d'un algorithme de Machine Learning par rapport à sa capacité à généraliser les associations qu'il a apprises durant la phase d'entraînement à de nouvelles observations [8, p. 118].

Il existe trois types de Machine Learning : à apprentissage supervisé, non supervisé et par renforcement [6, p. 686]. Il y a plusieurs familles d'algorithmes, comportant de nombreuses variantes, on a les modèles de régression, les algorithmes basés sur les instances, les arbres de décisions, les méthodes bayésiennes et les réseaux de neurones artificiels (ANN) [6, p. 687]. Le choix de l'algorithme dépend d'abord du type de problème à résoudre [8, p. 122].

L'apprentissage supervisé est la forme de Machine Learning la plus courante [8, p. 123]. Dans ce type d'apprentissage, on dispose d'un ensemble de données d'apprentissage, avec en entrée des exemples (la variable x) et des réponses étiquetées (ou des valeurs de la variable cible y) en sortie. On entraîne le modèle, c'est-à-dire qu'une phase d'apprentissage est effectuée, durant laquelle on observe les associations entre les données d'entrée et de sortie. On peut ainsi calibrer les paramètres du modèle. Lorsque l'entraînement du modèle est un succès, on peut l'utiliser avec de nouvelles données d'entrée (variable x) pour prédire la variable cible y [6, p. 687] et [8, p. 123]. Cet apprentissage se subdivise encore en deux catégories. La régression, qui consiste à prédire une valeur numérique, la variable cible est une variable quantitative[8, p. 123], et la classification consistant à attribuer le résultat d'une prédiction à une classe, une catégorie, la variable cible est donc ici qualitative [6, p. 687] et [8, p. 123].

À l'inverse de l'apprentissage supervisé, dans l'apprentissage non supervisé, les données d'apprentissage ne sont pas étiquetées [6, p. 687] et [8, p. 123]. Le but est que le système retrouve des informations structurelles par lui-même afin qu'il regroupe les exemples fournis en entrée par catégorie, c'est ce qu'on appelle le clustering (partitionnement). Le but peut être aussi de faire une réduction de dimension c'est-à-dire de projeter des données d'un espace avec une haute dimension vers un espace avec une faible dimension [6, p. 687]. Ce système présuppose une notion de distance ou de similarité entre les observations [8, p. 123].

Enfin, il y a l'apprentissage par renforcement pour lequel on ne donne pas de paires d'entrée et de sortie mais on fournit au modèle une description de l'état du système en cours, un objectif et une liste d'actions autorisées et de contraintes imposées par l'environnement. Une fois le modèle de Machine Learning lancé, on

le laisse expérimenter par lui-même « en utilisant le principe d'essai et d'erreur pour maximiser une récompense » [6, p. 687].

2.3 Le Deep Learning

On retrouve au sein du Machine Learning ce que l'on appelle les réseaux de neurones artificiels (artificial neural networks ou ANN), il s'agit d'une structure flexible pouvant être modifiée selon divers contextes et être ainsi utilisée dans les trois types de Machine Learning [6, p. 687]. Le Deep Learning est justement une branche du Machine Learning comprenant des algorithmes imitant les réseaux de neurones d'un cerveau humain, on parle ainsi de réseau de neurones artificiels. À l'instar du cerveau, le réseau compare une nouvelle information à celles dont il dispose en stock afin de leur donner un sens, ainsi, il déchiffre, étiquette et assigne ces informations à la catégorie adaptée. Si on parle de « Deep » Learning, d'apprentissage « profond », c'est en raison du nombre de couches constituant le réseau. Le Deep Learning constitue donc une branche du Machine Learning. Il existe trois couches dans ce type de réseau. Il y a la couche d'entrée, recevant les données d'entrée, la couche de sortie, avec les résultats, et la couche cachée qui extrait les modèles depuis les données. Un réseau de neurones artificiels profond aura ainsi plus d'une couche cachée, une architecture profondément imbriquée, des neurones avancés, il peut donc effectuer des opérations avancées comme des convolutions [6, p. 687]. Le Deep Learning fonctionne bien en particulier sur de grands volumes de données non structurées. Il est également plus précis que le Machine Learning. Il est cependant plus coûteux à mettre en œuvre et nécessite par essence un énorme volume de données [5, p. 132]. Mais des modèles de Machine Learning peu profonds peuvent être parfois supérieurs au Deep Learning, notamment dans le cas où les données d'entraînement sont peu disponibles ou lorsque les entrées de données sont de faibles dimensions [p. 688]janiesch2021machine. Ainsi, grâce au Machine Learning et au Deep Learning, on peut notamment, entre autres applications, effectuer du traitement de langage naturel. Le traitement de langage naturel, ou Natural Language Processing (NLP), est un champ du Machine Learning qui a pour but d'analyser et extraire de précieuses informations depuis un texte [2, p.2].

3 Application de l'analyse sémantique dans la recherche historique

3.1 Outils et intérêt de l'analyse sémantique en histoire

Aujourd'hui une énorme part de sources et de documents textuels se trouvent sous format numérique [4, p. 135], derrière cette numérisation des textes sous format papier se trouve une volonté de conserver le patrimoine culturel et de faciliter son accès, non seulement à la communauté scientifique, mais aussi au grand public [4, p. 135] et [9, p. 1]. Cet état de fait donne un énorme corpus de travail pour des études requérant des outils de traitement de langage naturel.

Les outils de traitement du langage naturel peuvent améliorer qualitativement les recherches pour un historien, ils peuvent réduire la quantité d'information à consulter et rendre son travail plus efficace [4, p. 135]. Par une analyse des données des textes en langage naturel, c'est-à-dire du text mining, ou fouille de textes, on classifie et catégorise les textes, on recherche des informations (des informations documentaires non structurées du fait d'un besoin d'informations), on traite les changements dans les textes. Dans ce contexte, pour des besoins de classification et de recherche de l'information, on retrouve notamment la recherche d'informations sémantiques [4, p. 136].

Les études de sémantique ont un certain intérêt en histoire, elles apportent un soutien à la recherche historique. On peut, par exemple, faire une recherche afin de voir les évolutions des sens d'un ou plusieurs mots à travers l'histoire ou les régions. Avec une telle recherche, on comble ainsi « le fossé lexical entre les langues modernes et anciennes » [9, p. 1]. De plus, connaître les évolutions sémantiques en histoire permet de restreindre la recherche à certains sens en particulier [12, p. 1].

Ainsi, la mise à disposition d'un nombre de documents historiques sous format numérique toujours plus croissant a poussé à mettre en application les méthodes et outils de traitement du langage naturel [9, p. 1].

Le traitement du langage naturel est « une technique computationnelle qui permet l'analyse du langage » [2, p. 3]. Il consiste en deux étapes, d'abord un prétraitement du texte. Ce prétraitement a pour but de nettoyer le texte, c'est-à-dire de traiter la ponctuation, supprimer les mots vides, corriger les erreurs d'orthographe, etc. La deuxième étape est le word embedding, ou plongement lexical, qui consiste à transformer le texte en objet mathématique sur lesquels une opération peut être effectuée. Il s'agit d'une étape nécessaire à l'analyse computationnelle. Pour ce faire, un vecteur est attribué à chaque mot ou phrase et ainsi les termes sont projetés dans un espace vectoriel [2, p. 3]. À l'aide de ces vecteurs, on peut ensuite calculer la similarité entre les termes (avec la distance euclidienne ou encore la similarité cosinus) [2, p. 4]. Les mots ayant une sémantique similaire ont des vecteurs similaires. il s'agit d'une approche basée sur les réseaux neuronaux. Les modèles les plus populaires de word embedding sont ceux issus de Word2Vec, avec notamment les algorithmes Skip-gram et continuous bag-of-word. Il s'agit dans les deux cas de réseaux neuronaux à deux couches, peu profonds et qui reconstruisent les contextes linguistiques des mots [9, p. 7].

L'application d'outils de traitement du langage naturel aux textes historiques présente toutefois quelques difficultés. Ces textes comportent des propriétés linguistiques particulières, comme la grammaire, l'orthographe ou des abréviations qui ne sont pas standardisées. De plus, il y a accessoirement le problème de savoir la date à laquelle le texte a été écrit quand ce n'est pas précisé, ce qui peut

être gênant lorsque l'on veut réaliser une étude sur l'évolution sémantique d'un ou plusieurs mots [9, p. 1].

3.2 Le problème de la sémantique des langues anciennes

Un problème commun à toutes les langues quand on veut faire une étude diachronique sur un champ lexical c'est qu'une langue n'est jamais totalement « ancienne » ni « nouvelle » [9, p. 4], il y a à tout moment une coexistence des sens originaux avec des nouveaux [12, p. 1]. Il faut ajouter qu'il existe des tendances et des modes d'utilisation des langues, qui peuvent avoir un fort impact sur les champs lexicaux, y compris dans un espace de temps très restreint, parfois inférieur à cinquante ans [9, p. 4].

Jusqu'à récemment, la recherche sur les changements sémantiques s'était principalement concentrée sur les langues modernes [11, p. 1]. De ce fait, les outils de traitement du langage naturel ont pour la majorité été conçus pour les langues modernes, or les langues historiques sont différentes des langues modernes sur de nombreux aspects, ce qui rend problématique l'utilisation de tels outils pour ces langues [2, p. 2].

Les langues anciennes présentent effectivement quelques difficultés non négligeables. Ainsi, le Grec ancien, par exemple, est une langue où les mots peuvent avoir un certain nombre de sens différents. Les changements sémantiques qu'il est possible de repérer peuvent dès lors être fort liés à ce problème de polysémie, ce qui accentue la difficulté de trouver notamment un moment où un sens nouveau apparaît [12, pp. 2-4].

Une autre difficulté avec les textes anciens est le fait que les dates de publications sont souvent inconnues, ce qui complique la tâche lorsque l'on travaille sur une évolution du sens des mots dans le temps, ainsi que lors du classement des textes historiques par période d'écriture, comme c'est le cas dans l'étude de Chaya Liebeskind et Shmuel Liebeskind, *Deep Learning for Période Classification of Historical Hebrew Texts*. Il s'agit d'une étude ayant pour but de classifier un corpus [1] de texte en hébreu en quatre périodes (XI^{e} siècle-fin XV^{e} siècle, XVI^{e} siècle, XVII^{e} siècle-XIX^{e} siècle, XX^{e} siècle à aujourd'hui). Sur base de cette classification, ils ont pu effectuer une recherche de changements sémantiques dans le temps. Dans leur cas, l'hébreu moderne également soulève quelques problèmes, car il incorpore des mots de la Bible et de commentaires rabbiniques, l'usage de morphèmes d'hébreu biblique, de l'orthographe mishnique, de la prononciation séfarade, ainsi que d'expressions idiomatiques yiddish. Pour leur étude, ils ont fait usage de modèles de Deep Learning. Le Deep Learning s'avérait en effet être le plus efficace, car leur corpus comptant 1 406 208 mots, cela représentait un grand nombre de données, une situation dans laquelle excelle le Deep learning. Ils ont comparé trois modèles différents, à savoir les vecteurs de paragraphes

1. à savoir le corpus Responsa comptant 1 406 208 mots

(qui modélise l'espace thématique avec des vecteurs de paragraphes), le réseau convolutionel (convolutional neural network, ou CNN, un réseau neuronal à anticipation) et le réseau neuronal récurrent (recurrent neural network RNN, un réseau qui utilise « sa mémoire interne pour traiter des séquences arbitraires d'entrée »[9, p. 2]). Leurs modèles ont utilisé des word-embedding de 300 dimensions qui ont été produits par l'algorithme skip-gram de Word2Vec, avec une fenêtre de cinq mots . Ils ont tout d'abord testé une classification via trois méthodes de Machine Learning conventionnelles (un Naive Bayes, un modèle linéaire et un Multi-Layer Perceptron). Ces trois méthodes conventionnelles se sont finalement montrées inférieures aux trois méthodes d'apprentissage profond qu'ils ont choisies. En effet, les méthodes conventionnelles avaient une efficacité qui dépendait beaucoup des étapes préalables, lors de la mise en place de l'ingénierie. Alors que l'avantage de l'apprentissage profond est qu'à partir de données d'entrée brutes, peu importe le domaine d'application, on fixe une importante quantité de caractéristiques à découvrir automatiquement. Ils ont donc par la suite opté pour des algorithmes d'apprentissage profond, appartenant à la catégorie des Machines Learning supervisés. De leurs trois modèles conventionnels, le modèle Naive Bayes a été le plus performant, mais de tous les modèles testés, les plus performants ont été les modèles Deep Learning CNN et RNN. Une fois la classification du corpus en quatre périodes réalisée, ils ont procédé à une analyse de changements sémantiques. Pour ce faire, ils ont adopté un protocole d'entraînement continu, dans lequel les embeddings de chaque période initialisent le modèle de la suivante. Ils ont entraîné les vecteurs de mots avec le package open-source Gensim. Pour analyser les changements sémantiques, ils ont d'abord comparé la similarité cosinus des mots des première et quatrième périodes, en retirant les mots apparaissant moins de 500 fois dans l'ensemble du corpus et ceux apparaissant moins de 50 fois dans chaque période. Ils ont par après comparé les mots voisins de ces mots cibles. Ils en déduisent que, lorsque les mots avaient des « comportements » différents, cela pouvait « provenir de changement de sens ou de changement de formulation dans le contexte » [9, p. 15]. Ainsi, ils ont regardé l'évolution du sens des mots cibles par rapport à leurs mots voisins. Ils ont également calculé la similarité cosinus d'un même mot à travers les différentes périodes, en prenant la première comme période de référence, afin de retrouver des périodes de changement sans l'influence des mots voisins. Ils remarquent, alors, que la plupart des changements de sens ont eu lieu lors des deuxième (XVI^{e} siècle) et troisième périodes (XVII^{e}-XIX^{e}),selon eux, ces changements s'expliquent par le fait que l'Hébreu redevient une langue couramment parlée à la fin du XIX^{e} [9].

On voit déjà ce que de tels outils nous permettent de réaliser en termes de recherches de changements sémantiques. De telles expériences ont également été mises en cours sur des corpus de textes grecs et latins. C'est ainsi ce qu'ont entrepris Valerio Perrone *et al.* dans deux études, une sur un corpus de textes grecs [12] et une sur un corpus de textes latins en plus du grec [11]. Mais dans ces études, Valerio Perrone *et al.* y intègrent un paramètre particulier.

Outre les difficultés auxquelles sont confrontées les études sémantiques sur les langues historiques que nous avons présentées précédemment, la sémantique présente une autre problématique, c'est l'impact des genres de textes. Il s'agit d'un point particulièrement intéressant avec les langues anciennes. En effet, les genres de textes disponibles sont déséquilibrés dans les corpus de textes anciens [12, p. 1] et ce n'est pas parce qu'un sens est présent dans un certain texte à une période donnée qu'il est représentatif de cette période, car tous les genres ne sont pas disponibles dans les mêmes proportions. De plus, le genre peut parfois avoir plus d'impact que le temps lui-même sur les différents sens d'un même terme [12, p. 2].

Ainsi, en 2019, Valerio Perrone *et al.*, dans leur article *GASC : Genre-Aware Semantic Change for Ancient Greek*, présentait GASC (Genre-Aware Semantic Change), un nouveau modèle de mélange dynamique bayésien pour les changements sémantiques, appliqué à un corpus de textes grecs.

Ce type de modèle intégrant le genre permet de savoir, par exemple, quel est le genre dans lequel on retrouve le plus tel ou tel sens associé à un terme donné. Le cas du Grec ancien est ici particulièrement intéressant, car c'est une langue pour laquelle on dispose de données sur plusieurs siècles avec un grand nombre de genres et où les mots peuvent avoir un nombre important de sens. Le Grec ancien comporte ainsi une certaine difficulté pour repérer les moments où un mot change de sens, c'est que le changement sémantique y est fort lié à la polysémie[12, p. 2]. Pour leur expérience, ils ont choisi cinquante mots cibles identifiés, manuellement, comme polysémiques et issus du Corpus de Grec ancien Diorisis. Ils ont divisé le corpus, 80% du corpus a été destiné à l'entraînement et 20% à un test. Les extraits choisis consistaient en des fenêtres d'une taille de cinq mots à gauche et à droit du mot cible. Le but étant de déterminer le sens associé au mot cible selon le contexte donné et faire une description de l'évolution des proportions de sens au cours du temps. Pour vérifier la fiabilité de leur méthode, ils ont fait appel à deux experts en Grec ancien. Ces derniers ont annoté le corpus manuellement, en marquant le bon sens de chaque occurrence pour trois mots cibles sélectionnés par ces experts. Ils ont ainsi évalué la performance prédictive de trois modèles, en log-vraisemblances, des données retenues : SCAN (qui ne comporte pas d'informations sur le genre), GASC-all (le modèle GASC avec tous les genres disponibles), et GASC-narr (avec deux genres, les genres narratif et non narratif). En moyenne, le modèle GASC-narr surpasse SCAN. Les exploitations des informations sur le genre fournissent de meilleures prédictions, cependant l'exploitation de tous les genres ne donne pas de résultats satisfaisants, car certains ne sont pas assez représentés, un problème inhérent aux textes anciens. Les résultats montrent que pour les mots cibles les plus courants, les informations sur le genre permettent de mieux récupérer la vérité terrain [12].

En 2021, Valerio Perrone *et al.* publient une autre étude [11] dans laquelle ils font part d'une expérience similaire effectuée de nouveau avec le modèle de mélange dynamique Bayésien GASC.

Cette fois-ci ils se penchent non seulement sur le corpus de grec ancien annoté Diorisis (qui a été annoté par des experts), mais aussi sur le corpus de textes latins LatinISE. Ces corpus ont été lemmatisés et étiquetés en parties de discours. Dans cet article, les chercheurs expliquent qu'alors que leur étude compte intégrer le genre dans leurs variables observables, les opérations de traitement du langage naturel sont généralement effectuées sur des genres spécifiques. Mais, finalement ce n'est pas un véritable obstacle puisque la recherche sur l'identification des genres est assez avancée. Valerio Perrone *et al.* ont donc décidé d'ajouter le genre comme variable observable supplémentaire à leur modèle bayésien. Afin de tenir compte du genre, la structure de leur modèle bayésien a été modifiée. Pour mesurer les changements sémantiques entre deux vecteurs d'un même mot, entre deux périodes, ils ont utilisé la similarité cosinus entre deux vecteurs. Ils ont alors effectué une classification binaire, c'est-à-dire s'il y a eu un changement ou non. L'étude de Perrone *et al.*, comme ils l'expriment en conclusion de leur article, montre que les modèles de mélange dynamiques bayésiens peuvent détecter des changements binaires et peuvent représenter de manière assez complète les évolutions sémantiques, ce qui montre que ces modèles sont efficaces dans l'étude de changements sémantiques pour les langues anciennes [11].

3.3 Quelques exemples d'application sur des thématiques modernes

Les outils d'analyse sémantique peuvent également apporter un énorme soutien à des études en histoire, notamment en histoire sociale et histoire du genre comme le montre l'article de Nikhil Garg *et al.*, *Word embeddings quantify 100 years of gender and ethnic stereotypes.*

Dans cette étude, les auteurs font notamment usage du word embedding pour effectuer une quantification des stéréotypes de genre et ethniques aux États-Unis, aux XX^{e} et XXI^{e} siècles. Ils montrent que le word embedding a une dynamique temporelle qui saisit les changements des stéréotypes. Ils remarquent une forte corrélation entre les changements quantifiables aux États-Unis et les dynamiques des plongements lexicaux. Avec les fortes transitions repérées dans la géométrie des plongements, ils montrent que pendant les mouvements féministes des années 1960-1970 et la croissance démographique de la population d'origine asiatique dans les années 1960-1980, les descriptions des genres et des ethnies ont changé. Pour l'analyse contemporaine, ils ont pris les vecteurs standards du Google News Word2Vec, entrainés sur le Google News dataset. Quant aux analyses temporelles historiques, ils ont utilisé un ensemble de 9 plongements, chacun sur une décennie des années 1900, les plongements préentrainés GoogleBooks/Corpus of historical American English. Ils ont entraîné aussi des

plongements à partir du New York Time Annotated Corpus pour les années 1988 à 2005, avec l'algorithme GLoVe. Ils ont établi des listes de mots pour les genres, les ethnies (blancs, asiatiques, hispaniques) et de mots neutres (des adjectifs et des professions). Ils ont utilisé les données du recensement américain pour avoir le pourcentage de travailleurs par profession pour chaque sexe et groupe ethnique, afin de comparer avec les biais détectés dans les plongements. Ils peuvent ainsi établir une mesure de l'association (embedding bias) entre des mots neutres et un groupe avec les plongements et ces listes de mots. Ils remarquent que la géométrie des plongements des mots évolue au cours du temps et s'aligne sur les changements démographiques empiriques ayant eu lieu aux États-Unis. Avec les professions, ils peuvent comparer leurs résultats et valider leur méthode avec les taux enregistrés empiriquement dans les recensements américains. Les biais liés au sexe et à l'origine ethnique correspondent. De plus, ils expliquent que les associations d'adjectifs dans les plongements permettent d'apercevoir la façon dont la perception les groupes d'individus a évolué [3].

Les analyses sémantiques ne sont toutefois pas sans défaut, elles peuvent avoir des impacts sur la société. Ainsi, Sean Matthews *et al.* montrent que les analyses sémantiques peuvent propager des préjugés, contenir des biais qui peuvent se révéler être une nuisance pour de futures prédictions [10].

Les modèles prédictifs présentent, en effet, certains défauts. Ils peuvent notamment participer à la propagation de préjugés et stéréotypes, car dans les données qu'ils encodent, et via lesquels ils sont entraînés, ils reproduisent des préjugés. Par exemple, on retrouve un biais racial dans les modèles prédictifs de discours haineux pour les messages sur les réseaux sociaux. Dans les modèles prédictifs, on retrouve des biais historiques, mais aussi de représentation, de mesure, d'agrégation, d'évaluation et de déploiement. Pour leur étude, Sean Matthews *et al.* se sont concentrés sur les biais historiques et de représentations trouvés dans les plongements lexicaux exercés sur des avis judiciaires tirés de la jurisprudence américaine. Ils se sont également concentrés sur divers défis par rapport à la conception de modèles prédictifs dans un cadre juridique. Les approches de plongements lexicaux comme Word2Vec « représentent les mots dans un espace à n dimensions en encodant les statistiques de cooccurrences contextuelles pour les mots apparaissant dans de grands corpus de textes » [10, p. 1], de fait, des biais déjà présents dans ces textes vont impacter les représentations des mots si on ne prend pas des mesures préalables. Ainsi, il peut y avoir des biais qui sont peu importants ou dont l'impact est relativement neutre comme les fleurs qui sont associées à des mots agréables et les armes à des mots désagréables. Mais il peut y avoir des effets aussi plus préjudiciables, notamment avec l'encodage de discrimination basée sur des critères raciaux, de genre ou sociaux. Dans ce contexte, afin de repérer, visualiser et diminuer les effets de ces biais, différentes méthodes ont été créées, dont le Word Embedding Association Test (WEAT). Cet outil permet de détecter des préjugés qui ont été encodés dans des plongements lexicaux. Ce procédé est basé sur le Implicit Association

Test (IAT), un test d'association implicite. Ce test est utilisé en psychologie sociale afin d'étudier les préjugés implicites qu'ont les humains. Il consiste à mesurer le temps de réponse différentiel lorsque des individus catégorisent des groupes de mots cibles et des attributs, et à déterminer si la configuration dans laquelle ils sont associés est stéréotypée ou non. Pour ce faire, le modèle prend par exemple deux groupes de mots cibles, un groupe de noms féminins et un de noms masculins, il prend le premier groupe, calcule la similarité des plongements des mots cibles du premier groupe avec les plongements dans deux ensembles de mots attributs (comme les adjectifs agréables et désagréables), on calcule ainsi les forces d'associations puis on compare avec celle du second groupe de mots cibles, après avoir effectué la même opération. C'est avec cette méthode que l'on a découvert que des préjugés raciaux se glissaient dans les plongements lexicaux effectués sur des textes de loi, comme des avis de cours d'appel [10, p. 2]. Ils ont ainsi analysé les plongements créés à partir d'un important corpus d'avis juridiques américains comptant 12 millions d'avis venant de 1949 juridictions contemporaines et historiques. Leur corpus remonte jusqu'à 1650. Pour le plongement lexical, ils ont utilisé un modèle skip-gram Word2Vec. Pour leur modèle de WEAT, ils utilisent des listes de mots originales d'Aylin Caliskan, Joana Bryson et Arvind Narayanan [1]. Pour chaque test, avec des listes de mots cibles et d'attributs, ils calculent la taille de l'effet et l'erreur standard. Au préalable, ils ont dû adapter les tests au domaine juridique, notamment en rajoutant des listes d'attributs, à savoir les mentions légales positives vs négatives, des résultats juridiques (motions), une liste élargie de carrière par rapport à la famille et de cibles (avec des noms de famille par race, des termes masculins vs féminins, des noms de juge). Après avoir entraîné leur modèle, ils ont observé, par exemple, un biais positif pour les insulaires d'Asie-Pacifique pour le test « agréable vs désagréable » et un biais négatif pour le test « mention légale positive vs négative » [10, p. 3]. Les auteurs de l'étude mettent toutefois en garde sur le fait que leur corpus remontant jusque 1650, l'observation de préjugés racistes et sexistes lors des tests peut être dû à ce facteur, étant donné que ces préjugés étaient plus présents et plus fortement exprimés en ce temps-là qu'à notre époque. Pour voir l'effet de ces opinions historiques sur les biais encodés, ils ont formé des sous-ensembles de plongements lexicaux dont l'année de début varie afin de toujours intégrer des opinions anciennes, ces sous-ensembles comportent de fait toujours des opinions modernes. Avec cette méthode, pour le test de biais raciaux, ils ont détecté un biais racial négatif à toutes les périodes par rapport aux noms de famille afro-américains et hispaniques. Cependant, ce biais diminuait un peu quand ils mettaient moins de données historiques, mais cela n'avait rien de très flagrant. Dans le cas des scores des biais sur le genre, les WEATs ont montré que les biais de genre variaient positivement ou négativement au cours du temps. Ils ont cependant remarqué un biais de genre au niveau des carrières à toutes les périodes. Leur étude ayant pour cadre la justice, ils ont dû rajouter des expressions propres au contexte juridique et ont modifié la méthode de détection des biais WEAT pour l'adapter à ce langage juridique. Leur expérience montre qu'il est important d'adapter la méthode de détection des biais à la matière étudiés. Sans le savoir, les développeurs de système de

traitement du langage naturel peuvent reproduire des préjugés, même en ayant tenté de filtrer le plongement lexical. Pour diminuer les préjugés raciaux ou sexistes, utiliser une date limite ne s'est pas révélée être une manœuvre assez efficace, même si les mentalités ont changé. Leurs résultats montrent que les biais de carrière liés au genre sont très fort pour les prénoms, selon eux, cela laisse penser que les systèmes de traitement de langage naturel juridiques, qui effectuent des opérations à ce propos, peuvent faire des prédictions biaisées [10]. Ainsi, Sean Matthews et ses collègues, qui analysent donc dans cette étude les biais dans les représentations en elles-mêmes, expliquent que si ces représentations sont utilisées pour faire des prédictions accessibles aux utilisateurs, elles peuvent dès lors avoir un impact négatif sur la société.

Le word embedding est également un bon outil pour analyser les évolutions sémantiques sur un ensemble de documents historiques s'étendant sur une longue période, comme le montrent KyoHoon Jin *et al.* dans leur article intitulé *Korean Historical Documents Analysis with Improved Dynamic Word Embedding*. Pour leur étude, les chercheurs disposent d'une série de textes enregistrant des faits sur de longues périodes, par exemple le journal du Secrétariat royal, qui enregistre des faits allant de 1623 à 1910, ou encore le Ming Shilu, de 1368 à 1644. Pour des analyses de textes sur de si longues périodes, il convient de prendre en considération des changements de sémantique au cours du temps. Le sens et l'usage d'un mot variant au fil du temps. Ils prennent pour exemple le mot anglais « apple » qui veut dire à la base « pomme », mais qui peut maintenant se référer à la société d'électronique Apple ou à l'iPhone en particulier. Ils expliquent ainsi que « par conséquent, la connaissance de l'évolution du sens des mots est un facteur important pour déchiffrer les documents historiques écrits sur de longues périodes » [7, p. 1]. Parmi leurs corpus de sources, ils utilisent les Annales de la dynastie Joseon (AJD), un document historique coréen représentatif. Faisant partie du patrimoine mondial de l'UNESCO, c'est un vaste document historique s'étendant sur plus de 500 ans (avec pas moins de 27 rois) et comportant des informations diversifiées allant de la politique au climat en passant par l'économie, la culture et la société. Il compte 888 livres pour 1893 volumes et 50 millions de caractères, ce qui en fait donc un texte complet et détaillé. Diverses études ont été faites, mais, à chaque fois, sur l'ensemble de la période couverte c'est-à-dire 500 ans. Il est donc difficile de se rendre compte des caractéristiques ou des idéologies propres aux différents rois. Pour leur étude, les chercheurs ont utilisé le plongement lexical dynamique (dynamic word embedding) pour saisir les changements sémantiques, qui ont été quantifiés sur base des vecteurs de plongement. Ils ont ensuite fait usage des informations obtenues par le plongement lexical pour augmenter les performances de la reconnaissance des entités nommées (named entity recognition, NER) et la traduction automatique neuronale (neural machine translation, NMT) des documents historiques. Ils ont paramétré leur plongement de manière à identifier le moment où le sens d'un mot change dans les documents. En incorporant le plongement lexical dynamique à des opérations comme la NER, ils ont pu montrer qu'il était efficace.

Il leur permet de détecter les changements dans les mots utilisés pour tel ou tel roi et lorsque des informations pour les noms d'objets changent. Par ailleurs en appliquant au plongement lexical dynamique des paramètres obtenus à partir du modèle NER ils ont pu améliorer les traductions des documents. Ils ont donc utilisé un plongement lexical dynamique amélioré pour analyser les changements sémantiques puis ils ont effectué un classement des personnes et des organisations trouvées avec un modèle NER et ont utilisé les paramètres formés par ce modèle dans le NMT pour améliorer la traduction des documents. Pour disposer d'une meilleure réflexion sur les informations pour les différentes périodes, ils ont ajouté des informations supplémentaires sur les 27 rois dans le modèle. Leur modèle, relevant de l'apprentissage profond, a utilisé des données de l'AJD, ils ont « parcouru le texte original de l'AJD, les données de noms d'objets étiquetés par des experts et les données traduites par des experts » [7, p. 5], pour la NER, ils ont pris quatre types de noms d'objets : personne, organisation, livre et temps. Pour calculer la distance entre les vecteurs, ils ont utilisé la distance euclidienne. Leur méthode a présenté de bonnes performances, par exemple, les résultats montrent qu'à la fin de la dynastie Joseon, le mot « Japon » s'est rapproché d' « Angleterre » et « États-Unis », au moment où le japon s'est rapproché des cultures occidentales avec la restauration Meiji. Leur technique a mis en avant les changements sémantiques et ils ont démontré qu'ils pouvaient utiliser ces informations pour les opérations de NER et de NMT. Ainsi, ils exqpliquent qu'en combinant leur plongement lexical dynamique amélioré et des informations sur les rois avec une fonction bilinéaire, ils ont obtenu un meilleur score pour la NER [7].

3.4 Conclusion

On peut voir que les analyses sémantiques quantitatives peuvent avoir leur intérêt, particulièrement dans le domaine de l'histoire où elles peuvent apporter un soutien de poids à la recherche historique. Outre que la sémantique permet d'effectuer de meilleures recherches, car elle permet de faire des requêtes plus précises, les analyses sémantiques sont des sujets de recherches à part entière nous apportant des informations sur l'évolution des langues, des usages et du sens des mots. On peut ainsi réaliser des études de linguistique historique sur un mot et voir, par exemple, quel sens est associé à ce mot à telle ou telle époque. En d'autres termes, on peut ainsi saisir son champ sémantique selon les époques. En outre, comme l'a montré l'étude de Nikhil Garg *et al.*, ce type d'analyse permet de se rendre compte de la manière dont des groupes ou des individus peuvent être perçus par la société au cours du temps. Les possibilités de sujets de recherche avec ce type de méthode et de technologie sont donc assez vastes et variées. Bien que les outils disponibles soient plus développés pour des études réalisées sur base de corpus de textes dans des langues modernes, il est tout à fait possible de les adapter afin de réaliser des études similaires sur des corpus de textes anciens, comme le grec ancien, le latin ou encore l'hébreu de l'époque médiévale. Par ailleurs, les deux études de Valerio Perrone *et al.* ont montré que l'intégration du genre de texte dans le modèle permettait d'avoir

de meilleurs résultats pour repérer les changements de sémantiques. Peut-être qu'à l'avenir de nouveaux paramètres seront incorporés aux modèles d'analyse sémantique, permettant d'atteindre encore de meilleurs résultats.

Références

[1] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334) :183–186, 2017.

[2] Arianna Di Bernardo, Simone Poetto, Pietro Sillano, Beatrice Villata, Weronika Sójka, Zofia Piętka-Danilewicz, and Piotr Pranke. Latin writing styles analysis with machine learning : New approach to old questions. *CoRR*, abs/2109.00601, 2021.

[3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16) :E3635–E3644, 2018.

[4] Anna Glazkova, Valery Kruzhinov, and Zinaida Sokova. Automatic text processing for historical research. In *International Conference on Modern Information Technology and IT Education*, pages 135–146. Springer, 2017.

[5] Deepack Jakhar and Ishmeet Kaur. Artificial intelligence, machine learning and deep learning : definitions and differences. *Clinical and experimental dermatology*, 45(1) :131–132, 2020.

[6] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3) :685–695, 2021.

[7] KyoHoon Jin, JeongA Wi, KyeongPil Kang, and YoungBin Kim. Korean historical documents analysis with improved dynamic word embedding. *Applied Sciences*, 10(21) :7939, 2020.

[8] Pirmin Lemberger, Marc Batty, Médéric Morel, and Jean-Luc Raffaëlli. *Big Data et Machine Learning-3e éd. : Les concepts et les outils de la data science*. Dunod, 2019.

[9] Shmuel Liebeskind and Chaya Liebeskind. Deep learning for period classification of historical hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020, 2020.

[10] Sean Matthews, John Hudzina, and Dawn Sepehr. Gender and racial stereotype detection in legal opinion word embeddings. *arXiv preprint arXiv :2203.13369*, 2022.

[11] Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. Lexical semantic change for ancient greek and latin. *arXiv preprint arXiv :2101.09069*, 2021.

[12] Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. Gasc : Genre-aware semantic change for ancient greek. *CoRR*, abs/1903.05587 :56–66, 2019.