Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Exploitation de modèles distributionnels pour l'étude de la nomination dans un corpus d'interviews politiques

Manon Cassier[1, 2]
(1) CY Cergy-Paris Université, AGORA, 33 Boulevard du port, 95011 Cergy, France
(2) Institut National des Langues et Civilisations Orientales, ERTIM, 2 Rue de Lille, 75007 Paris, France

manon.cassier@u-cergy.fr

RÉSUMÉ

En analyse de discours (AD), la nomination désigne la recatégorisation du référent par le locuteur à travers l'usage d'un nouveau nom ou d'un nom modifié. Parfois utilisé pour influencer l'autre sur sa vision de voir le monde, ce phénomène sert d'indice sur l'idéologie du locuteur voire, en contexte adéquat, sur son affiliation politique. L'AD ne dispose pas à ce jour d'outils en mesure d'appréhender efficacement ce qui relève ou non de l'idéologie ou d'une visée argumentative face à une simple réutilisation de mots dont le sens est déjà consensuel. Dans le cadre d'une thèse entre AD et TAL, nous nous intéressons à l'exploitation de modèles distributionnels pour repérer de manière automatique ces variations de sens en discours dans un corpus d'interviews politiques. Dans cet article, nous nous interrogeons sur l'impact de leurs paramètres d'entraînement pour de la désambiguïsation lexicale et explorons une méthode de représentation de la variation sémantique interdiscursive.

ABSTRACT

Speaker-specific semantic variation representations using vector space models.

The French Discourse Analysis (DA) concept of "Nomination" designates the cases of the speaker updating an entity's identity by using words in a particular way that involves its opinion. It sometimes constitutes a good way to gain access to the person's ideology and to predict its political stripe. Several tools are already used by discourse analysts, but they do not provide to properly represent the nomination semantic characteristics. In our work, we question word embeddings based models benefits to automatically detect these speaker-specific semantic changes. In this paper, we investigate the role of training features for a semantic oriented task and test some approaches to represent semantic variation.

MOTS-CLÉS : Nomination, analyse du discours (AD), désambiguïsation lexicale, modèles distributionnels.

KEYWORDS: Nomination, Political Discourse Analysis, Word Sense Induction (WSI), Word Sense Disambiguation (WSD), Word Embeddings Based Models.

1 Introduction

Le concept de nomination est étudié en analyse du discours (AD) comme la réassignation sémantique et référentielle en discours d'un mot, dans certains cas à des fins de valorisation idéologique et de persuasion de l'auditoire. Parfois considéré comme une pratique de l'argumentation, l'acte de nomination est réalisé dans des discussions autour de sujets propices à la controverse, dans des

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 71–84.
HAL : hal-02786188.

contextes socio-politiques qui portent au conflit, au sein desquels il sert à qualifier l'objet du désaccord (Koren, 2016). Le locuteur manipule alors le lexique de manière à négocier le sens du mot et imposer sa représentation du référent en influençant l'autre. Gauthier (2016) montre par exemple comment l'utilisation par le gouvernement canadien du mot « boycott plutôt que « grève » pour désigner la cessation d'activité des associations étudiantes contre la hausse des frais de scolarité sert une stratégie d'évitement du débat par une minimisation de l'impact de l'événement.

L'analyse de discours dispose déjà, à ce jour, de différents outils pour l'exploration de textes et la mesure des spécificités lexicales (i.e. le sur- ou sous-emploi d'une forme par un locuteur ou par un genre de discours particulier) (Tournier, 1981). Ces outils reposent tous sur des méthodes de calculs statistiques qui permettent par exemple d'accéder à la fréquence des cooccurents ou des segments répétés, mais qui ne permettent pas d'appréhender l'aspect sémantique indispensable pour l'étude de la nomination. En effet, s'il est possible d'observer des tendances à l'usage de certains mots choisis plutôt que d'autres (Marchand & Ratinaud (2012) montrent par exemple un sur-emploi des formes « écologique » et « équitable » dans le discours de Ségolène Royal lors des primaires socialistes de 2011), seule une analyse humaine permet à ce jour de discriminer les usages dits « neutres » de ceux qui relèveraient de la nomination (i.e. dont le sens serait spécifique à l'idéologie du locuteur).

Le projet ANR TALAD [1] se donne pour objectif de fournir à l'AD de nouveaux outils, adaptés des tâches du TAL, qui permettront d'assister plus efficacement la recherche de phénomènes tels que la nomination sur de gros volumes de textes.

Il s'agirait par exemple d'aider à l'identification, dans les exemples suivants [2], du candidat nomination « Europe ». Celui-ci est utilisé dans le premier cas pour référer à l'entité politique perçue comme bénéfique, mais réfère au contraire à une simple entité géographique dans le second pour contraster avec l'« Union européenne » qui endosse à sa place la référence à l'entité politique perçue comme négative.

Exemple 1. « Si la France veut continuer de se projeter dans le monde, elle doit avancer en **Europe**, rebâtir le projet européen avec détermination et là aussi ne rien céder à celles et ceux qui doutent. » (Discours d'Emmanuel Macron à Montpellier le 18/10/2016)

Exemple 2. « Donc l'Union européenne est une mauvaise chose, pas l'**Europe**. L'**Europe** c'est une entité, une réalité de civilisations, c'est une réalité géographique, historique. » (Interview de Marine Le Pen sur Europe1 le 01/05/2017)

Dans le cadre de notre thèse, nous nous focalisons en particulier sur l'exploration de méthodes distributionnelles pour mettre au jour les variations sémantiques inter-locuteurs au niveau du mot. En effet, de tels modèles ont déjà fait leurs preuves dans des travaux portés sur la sémantique des mots, notamment pour des tâches de désambiguïsation lexicale.

Néanmoins, notre problématique apporte un aspect qui, à notre connaissance, n'est pas encore abordé par la littérature existante, à savoir la mise au jour de sens nouveaux ou non consensuels (Pengam & Jackiewicz, 2019), et de variantes parfois si subtiles et peu fréquentes que même l'identification humaine en est rendue difficile. De même, notre objectif ne se résume pas à de la désambiguïsation lexicale, puisqu'il s'agit aussi de remarquer un déplacement du sens, parfois simplement d'un tour de parole à un autre (et donc au sein d'un même corpus). Enfin, notre intérêt final serait d'aboutir à une méthode qui puisse nous permettre de mesurer, sur le modèle des calculs de spécificités lexicales

1. https://anr.fr/Project-ANR-17-CE38-0012
2. Exemples tirés du corpus *Reticular* présenté en section 3.1

(Tournier, 1981), des spécificités sémantiques capables de caractériser chaque acteur en fonction du sens qu'il accorde à ses mots.

Nous commençons cet article par un état des lieux des outils déjà exploitables ou exploités en AD et présentons leurs lacunes pour des études focalisées sur la nomination en première partie de section 2. Nous présentons en seconde partie de cette même section un état de l'art des méthodes exploitées notamment pour la détection de la polysémie et la désambiguïsation lexicale desquelles nous nous inspirons pour notre travail. La section 3 détaille le dispositif expérimental employé dans cette étude, notamment les corpus et les méthodes de sélection des paramètres utilisés pour l'entraînement des modèles. La section 4 introduit une expérimentation centrée sur la prédiction d'un mot-pivot. Nous présentons ensuite la démarche que nous envisageons pour représenter la variation interdiscursive en section 5. La section 6 conclut et discute les choix méthodologiques et les poursuites envisagées.

2 État de l'art

2.1 Outiller l'analyse de discours

Les analystes du discours travaillent généralement autour de mots-pivots, sélectionnés car potentiellement vecteurs d'idéologie (Mazière, 2018). Ce type de méthode nécessite au préalable d'identifier le mot autour duquel il faudra construire un corpus qui réunira les contextes représentatifs du phénomène étudié. L'utilisation de logiciels de lexicométrie ou textométrie (e.g. Lexico, Iramuteq, Hyperbase, Alceste ou TXM pour ne citer que les plus utilisés en AD), en l'occurrence par le biais des concordanciers qu'ils proposent, peut faciliter la constitution de corpus, qui représente déjà en soi une étape assez chronophage du travail de l'analyste du discours. Néanmoins, l'utilisation de ces concordanciers seuls ne suffit pas pour repérer de nouvelles formes candidates susceptibles d'être concernées par la nomination lorsque le contexte nécessaire à leur reconnaissance est plus large que la fenêtre d'observation offerte par l'outil, ni lorsqu'il s'agit d'étudier des variantes lexicales pour une même nomination. Pengam & Jackiewicz (2019) montrent par exemple que la nomination autour de « musulman modéré » est susceptible d'entraîner des variantes telles que « islam modéré » ou « islamisme modéré » qu'il est nécessaire d'anticiper si elles n'apparaissent pas dans les mêmes contextes que la forme pivot pré-sélectionnée.

Aussi, le fait de travailler justement autour de formes pivots, généralement pré-sélectionnées pour leur caractère néologique (comme c'est le cas pour la collocation « musulman modéré ») et/ou polémique (comme c'est le cas pour « migrant » (Calabrese, 2018)) limite fortement l'étude de l'ensemble des nominations possibles et peut empêcher de détecter les cas les plus flagrants.

Ces logiciels, développés à l'origine pour des études en statistiques textuelles, offrent chacun des fonctionnalités qui s'avèrent utiles pour les études menées en analyse de discours (Longhi, 2017). Les calculs de cooccurrences et la recherche de segments répétés (Lebart & Salem, 1994) peuvent par exemple assister la détection de nouvelles collocations comme pour « musulman modéré », qui donnent un indice sur le changement de sens du mot tête (ici « musulman »). La classification de Reinert (Reinert, 1993) offre la possibilité de distinguer les différentes thématiques abordées par un corpus, qui peuvent par exemple aider à déterminer les cadres de la nomination (i.e. les thèmes les plus probables de susciter l'emploi de nominations). Le calcul de spécificités lexicales (Tournier, 1981) permet de déterminer, pour un corpus découpé en sous-parties (distinguées par exemple selon le genre textuel, l'année ou le locuteur), les formes en sur- ou en sous-emploi dans ces sous-parties (i.e. les formes qui en sont les plus ou les moins réprésentatives). Ce type de calcul offre la possibilité, dans le cadre de l'analyse du discours politique, de déterminer rapidement le vocabulaire qui est le

plus spécifique à chaque politicien. En revanche, puisque ces calculs ne fonctionnent que sur la base de statistiques sur la fréquence des mots dans chaque corpus, sans prendre en compte le contexte d'apparition des mots, il est compliqué d'en tirer des conclusions au niveau sémantique. Un tri manuel des contextes une fois les spécificités lexicales identifiées est toujours indispensable.

Sans vouloir dispenser de cette étape nécessaire d'étude manuelle de contextes, ce travail se place dans l'objectif de proposer un outil plus complet pour assister l'analyse de discours (et plus spécifiquement l'analyse de la nomination) en intégrant l'aspect sémantique à l'extraction de contextes candidats. L'étude présentée dans ce papier consiste en l'exploration de méthodes distributionnelles pour le repérage automatique de candidats nominations, dont nous présentons un état de l'art dans la section suivante.

2.2 Les modèles distributionnels au service de la désambiguïsation lexicale

Les représentations vectorielles de mots se fondent sur l'hypothèse distributionnelle introduite par Harris (1954) et Firth (1957) selon laquelle le sens d'un mot peut être déterminé grâce aux contextes dans lesquels il apparaît, et que les mots qui apparaissent dans des contextes similaires sont sémantiquement proches. Les pratiques de l'AD abordées dans la section précédente montrent également l'importance de ces contextes pour les analystes du discours qui fondent leur travail sur l'étude des cooccurrences et des segments répétés. Les différents modèles prédictifs qui ont pu voir le jour au cours de la dernière décennie suscitent toujours un grand engouement pour leurs performances sur de nombreuses tâches du TAL et leur capacité à fournir une représentation sémantique des mots en les situant les uns par rapport aux autres dans un espace vectoriel dense.

Les outils comme Word2Vec (Mikolov *et al.*, 2013) jouissent désormais d'une documentation suffisamment complète pour être éprouvés dans des études purement linguistiques, les mécanismes en jeu dans l'apprentissage des représentations faisant régulièrement l'objet de travaux qui les ont rendus beaucoup plus compréhensibles qu'à leurs débuts (e.g. Levy & Goldberg (2014); Levy *et al.* (2015); Patel & Bhattacharyya (2017) sur l'influence des paramètres d'apprentissage ; Pierrejean & Tanguy (2018, 2019) sur la variabilité des représentations et l'instabilité des modèles). En outre, les modèles prédictifs qu'ils proposent suscitent beaucoup d'intérêt pour la résolution de tâches focalisées sur des aspects sémantiques telles que l'analogie (i.e. reconstitution de paires de concepts qui entretiennent le même type de relation que les concepts d'une paire exemple, e.g. de type *Pays-Capitale* : Paris est à la France ce que Madrid est à l'Espagne). Enfin, les travaux de Baroni *et al.* (2014) et Levy *et al.* (2015) ont montré que ces modèles prédictifs, entraînés avec des paramètres adaptés à la tâche et au corpus d'étude, peuvent obtenir de meilleurs résultats que les approches à base de comptes (e.g. *Positive Pointwise Mutual Information* (PPMI), *Latent Semantic Analysis* (LSA)).

Néanmoins, ces modèles présentent le défaut de ne fournir qu'une seule représentation vectorielle par forme ou par mot, en encodant tous les sens possibles de cette forme ou de ce mot en un vecteur unique lorsqu'il serait préférable d'avoir un vecteur par contexte. Différentes méthodes ont déjà été explorées pour transformer ces vecteurs uniques en vecteurs de sens, notamment dans le cadre de travaux en induction de sens (*Word Sense Induction* ou WSI) et en désambiguïsation lexicale (*Word Sense Disambiguation* ou WSD)(Ruas *et al.*, 2019; Pelevina *et al.*, 2017).

La plupart des systèmes de WSI et WSD se basent sur un inventaire de sens, qui liste les différents sens possibles de chaque mot, pour calculer la probabilité d'un contexte de relever d'un des sens répertoriés pour un mot. De nombreux travaux exploitent par exemple des bases de données comme *WordNet* (Vial *et al.*, 2017) ou *BabelNet* (Dongsuk *et al.*, 2018) qui associent des groupes de mots

qui entretiennent des relations sémantiques (e.g. hypéronymie, méronymie) au sein de *synsets*. Les ressources de ce type sont généralement moins développées pour le français. En outre, dans le cas de la nomination, le sens du mot n'est pas forcément attesté dans un dictionnaire, ni partagé par l'ensemble de la communauté linguistique puisqu'il dépend de l'appréciation du locuteur du référent qu'il nomme. La ressource la plus exhaustive possible resterait donc inexploitable pour distinguer des variantes de sens ou des représentations du référent complètement nouvelles.

Pour contourner ce problème, Pelevina *et al.* (2017) proposent une méthode d'acquisition automatique d'inventaire de sens à partir de textes non annotés via la création d'*ego-networks* (i.e. identification de clusters de sens autour d'un mot pivot). Ces réseaux sont ensuite exploités pour calculer un vecteur de sens comme vecteur moyen des vecteurs de tous les mots appartenant au cluster. Les vecteurs de sens induits sont ensuite comparés aux vecteurs de contexte des mots à désambiguïser.

Nous notons, pour la suite de notre travail, que la plupart de ces approches sont évaluées sur des *benchmarks* pour des tâches précises comme la similarité ou l'analogie. Ces *benchmarks*, souvent constitués pour l'anglais, évaluent des relations sémantiques tirées du domaine général qui diffèrent beaucoup de celles que nous pouvons observer dans des corpus de spécialité comme les interviews politiques. Les particularités de notre tâche (i.e. évaluée sur du français, dans un corpus de spécialité, avec l'objectif supplémentaire de comparer les représentations interdiscursives) devront nous questionner sur l'adaptation de méthodes d'évaluation, sur le modèle par exemple de Bloem *et al.* (2019) qui proposent une mesure de cohérence des représentations apprises dans des corpus de spécialité.

Au regard de cet état de l'art, nous présentons dans cet article notre réflexion sur l'exploitabilité de modèles distributionnels pour représenter et détecter la variation sémantique interdiscursive au niveau du mot (plus précisément le nom) et leur évaluation sur un corpus d'interviews politiques françaises.

3 Dispositif expérimental

À ce stade de la thèse, nous avons choisi de concentrer nos expérimentations sur des modèles entraînés avec Word2Vec (Mikolov *et al.*, 2013). Ce choix s'explique notamment par l'accès facilité à de la documentation claire, la possibilité d'entraîner soi-même ses modèles pour un coût relativement faible face à des modèles de réseaux de neurones très gourmands en GPU, mais aussi par la multiplication de travaux qui ont déjà prouvé un fonctionnement suffisant sur des corpus limités (Bloem *et al.*, 2019). Nous envisageons, si les premiers résultats sont satisfaisants et que le temps nous le permet, de poursuivre nos explorations avec les modèles à base de réseaux de neurones plus élaborés qui font aujourd'hui état de l'art pour de nombreuses tâches du TAL (e.g. BERT, ELMo).

Comme nous l'expliquons dans la section 1, notre problématique nécessite de repérer à la fois les usages « inhabituels » de noms (i.e. dont le sens diffère de celui qui serait consensuel à l'ensemble des locuteurs de la langue et qui serait disponible dans une ressource dictionnairique ou inventaire de sens), mais aussi les variantes de sens d'un acteur politique à un autre.

Suivant ces deux contraintes, nos expérimentations sont basées sur deux corpus de genres différents. Le corpus *Reticular*, que nous présentons dans la sous-section suivante rassemble des interviews politiques et nous sert à construire les représentations subjectives à analyser (qui contiennent potentiellement des nominations). Pour comparer ces représentations à un usage plus « neutre », nous utilisons un corpus constitué des articles de la version française de Wikipédia de 2008 [3].

3. Corpus WikipediaFR2008 extrait par l'équipe CLLE-ERSS de l'Université de Toulouse Jean Jaurès, disponible à l'adresse suivante : http://redac.univ-tlse2.fr/corpus/wikipedia.html

3.1 Le corpus *Reticular*

Pour notre étude, nous nous basons sur un corpus d'interviews transcrites à la main. Ce corpus couvre 3166 interviews de 561 personnalités publiques données pour des émissions de radio dans le contexte des élections présidentielles françaises de 2017. Le corpus couvre la période de juin 2016 à décembre 2017 et compte environ 11 millions de mots.

Chacune des interviews du corpus est accompagnée de métadonnées indiquant la date et l'heure de l'interview et le nom de l'interviewé. Une interview est découpée en tours de parole, avec une ligne par tour de parole, alternant les questions et réactions du journaliste et de l'interviewé. La répartition des interviews n'est pas homogène, les émissions de radio accueillant à la fois des politiciens et des personnes extérieures à la sphère politique susceptibles d'éclairer un débat (e.g. économistes, dirigeants d'entreprise, etc.). La majorité des interviews sont néanmoins partagées par des têtes de listes et leur directeur de campagne.

Bien qu'il s'agisse d'un corpus transcrit de l'oral, le corpus *Reticular* ne contient que le texte des interviews, sans marques d'hésitations. De même, la plupart des répétitions, habituelles dans les corpus oraux, n'ont pas été transcrites. Aussi, les marques de ponctuation n'ont été utilisées par les transcripteurs que pour découper le texte en phrases cohérentes. Nous devons donc préciser qu'il nous est impossible d'utiliser ces marques à des fins d'analyse sémantique.

Pour toutes les expérimentations présentées dans les sections suivantes, nous utilisons une version lemmatisée avec Treetagger (Schmid, 2013) du corpus (en conservant les formes en lieu et place des lemmes non reconnus) à laquelle nous retirons toutes les marques de ponctuations. Nous ne considérons donc pas les phrases, mais les tours de parole marqués dans le corpus par des sauts de ligne.

Les mêmes prétraitements sont appliqués au corpus *Wikipédia*, en plus du retrait des caractères spéciaux qui n'apparaissent pas dans le corpus des interviews. Aussi, pour limiter l'impact des balises, hyperliens et phrases typiques de l'encyclopédie en ligne (e.g. « Ceci est un article qui concerne... »), nous faisons un tri grossier en retirant les lignes de moins de 10 mots.

3.2 Méthode de sélection des paramètres d'entraînement

Dans un premier temps, comme pour répondre à toute tâche de TAL, notre travail nécessite de déterminer les paramètres d'entraînement de nos modèles qui auront l'impact le plus positif sur les représentations distributionnelles qui en découleront. Certains travaux ont investigué l'incidence du choix des paramètres d'entraînement sur la performance des modèles. Patel & Bhattacharyya (2017) montrent par exemple que les représentations sont plus fiables pour des tâches de similarité, d'analogie et de classification lorsque le nombre de dimensions dépasse un seuil qui dépend du nombre de noeuds du plus large cluster disponible dans un graphe de similarités construit à partir de la matrice de cooccurrences du corpus. En dessous de ce seuil, le modèle n'est pas capable d'encoder tous les liens entre les mots du corpus et donne donc de moins bons résultats. En revanche, la performance du modèle a tendance à se stabiliser pour un nombre de dimensions supérieur à ce seuil. Les travaux de Levy & Goldberg (2014) et Levy *et al.* (2015) démontrent de leur côté qu'une taille de fenêtre petite (i.e. pour laquelle on réduit le contexte à prendre en compte autour du mot lors de l'apprentissage du modèle) permet de mieux représenter les caractéristiques syntaxiques du mot (e.g. quelle catégorie grammaticale a le plus de chance de se retrouver directement en contact avec le mot cible) alors qu'un contexte plus grand est capable d'encoder les aspects plus sémantiques (e.g. quel synonyme a le plus

de chance de se retrouver à la place de ce mot) puisqu'il a accès aux cooccurents plus éloignés du mot.

Puisque l'étude de la nomination demande de s'intéresser essentiellement à la catégorie nominale, nous souhaitons en premier lieu éprouver cette remarque en vérifiant si la taille de fenêtre sélectionnée lors de l'entraînement de nos modèles impacte particulièrement la prédiction des noms et des entités nommées (Nouvel *et al.*, 2015). Pour ce faire, nous exploitons simplement la fonction principale de l'architecture CBOW (que la littérature présentée dans la section 2 reconnaît comme la plus efficace pour des tâches de désambiguïsation lexicale), à savoir la prédiction d'un mot à partir d'un contexte.

À partir du corpus lemmatisé de *Wikipédia*, dans lequel nous ajoutons à chaque mot l'étiquette de sa catégorie grammaticale récupérée avec Treetagger, nous entraînons deux modèles avec les paramètres par défaut de Word2Vec (i.e. 100 dimensions avec un *negative sampling* et 5 itérations sur le corpus) en ne faisant varier que la taille de fenêtre (i.e. la taille du contexte considéré), fixée à 5 pour le premier modèle et à 15 pour le second. Nous fixons la fréquence minimale des mots à prendre en compte dans le calcul à 10 pour limiter l'impact des mots rares, susceptibles d'être répétés plusieurs fois dans les articles de Wikipédia.

Nous évaluons ensuite la capacité des deux modèles à prédire chaque catégorie grammaticale en fonction du contexte qui leur est fourni dans un système de texte à trou. Pour des raisons de temps de traitement, cette évaluation est menée sur un échantillon de 150 interviews du corpus *Reticular* (soit environ 5% du corpus total) sélectionnées de manière aléatoire pour une taille d'environ 450 000 mots. Nous ajoutons également son étiquette grammaticale à chaque mot.

Le modèle propose une liste triée par ordre de probabilités de prédictions du mot caché (qui occupe la position centrale du contexte). Par exemple, pour l'énoncé ci-dessous dans lequel le mot « référendum » [4] est masqué, le modèle entraîné avec une fenêtre de 5 considère le contexte restreint qui suit :

Énoncé : « Est-ce que le camp du maintien au sein de l'Union européenne n'est pas en train depuis déjà maintenant des semaines de reproduire les erreurs de 2005 lors du **référendum** constitutionnel, c'est-à-dire au fond en agitant la peur le catastrophisme en permanence ? »
Contexte considéré : ['lors_ADV', 'du_PRP', 'constitutionnel_ADJ', 'c'est-à-dire_ADV']

Avec l'architecture CBOW, le modèle récupère le vecteur du contexte donné (calculé comme le vecteur moyen de tous les mots du contexte) et donne une liste de prédictions de la taille du vocabulaire pour le mot manquant (dont nous restreignons l'affichage aux 5 premières propositions par souci de lisibilité des résultats), triée par ordre de probabilité (i.e. par ordre de similarité la plus élevée entre le vecteur du contexte et le vecteur de la prédiction).

3.3 Évaluation quantitative et qualitative des prédictions

Pour évaluer les prédictions de notre modèle, nous mesurons la moyenne et l'écart-type des distances (notées entre 0 et 1) entre le vecteur des mots cachés et ceux des prédictions faites par le modèle. Les résultats les plus proches de 0 correspondent donc aux cas où le mot à prédire se trouve parmi les 5 premières propositions du modèle.

Le tableau 1 donne les résultats de la prédiction pour chaque modèle avec une distance moyenne par catégorie grammaticale. La deuxième colonne précise la fréquence de chaque catégorie dans le

4. Dans tout l'article, nous utilisons le **gras** pour marquer les mots à prédire dans les énoncés.

corpus d'évaluation. La dernière colonne donne l'écart des moyennes des deux modèles. Les résultats sont affichés sur la base de la troisième colonne, par ordre de distance pour le modèle avec une taille de fenêtre fixée à 5.

Prédictions	Fréquence (en milliers)	Fenêtre de 5		Fenêtre de 15		Variation (en %)
		Distance	Écart-type	Distance	Écart-type	
ADV	41	0.3	0.11	0.31	0.17	3.4 %
PRO	81	0.32	0.08	0.35	0.12	**9.4 %**
KON	26	0.33	0.08	0.35	0.13	4.5 %
DET	42	0.35	0.06	0.35	0.07	**1.4 %**
VER	83	0.35	0.07	0.37	0.08	4.3 %
PRP	56	0.36	0.05	0.38	0.07	4.2 %
NOM	68	0.36	0.08	0.39	0.09	8.3 %
ADJ	21	0.37	0.07	0.4	0.08	**9.6 %**
NUM	5	0.4	0.09	0.46	0.08	**12.3 %**
NAM	20	0.42	0.09	0.45	0.09	7.2 %

TABLE 1 – Distance moyenne des prédictions par catégorie grammaticale

Au premier abord, nous remarquons que les distances moyennes sont toutes, sans exception, plus élevées pour le modèle entraîné avec une taille de fenêtre plus grande et que l'écart-type reste sensiblement le même pour les deux modèles, ce qui laisse entendre que toutes les prédictions de chaque catégorie sont affectées de la même manière par l'augmentation de la taille de la fenêtre.

Concernant la variation des distances moyennes en passant d'un modèle à l'autre, nous remarquons que les prédictions des pronoms, des numéraux, des adjectifs et des noms sont celles qui varient le plus. Ces résultats sont difficilement compréhensibles sans jeter un oeil au comportement des prédictions. Une analyse manuelle des prédictions faites par les deux modèles nous permet de faire les observations suivantes :

1. Concernant la prédiction des numéraux (NUM) :

   — Le modèle entraîné avec une taille de fenêtre plus petite est obligé d'accorder plus de poids aux quelques mots du contexte dont il dispose pour calculer le vecteur le plus proche. Il en résulte que pour la prédiction des numéraux, même s'il ne propose pas forcément le chiffre exact à prédire, il est plus simple pour lui de prédire les suites de chiffres (i.e. le modèle propose des chiffres lorsqu'il y en a déjà dans le contexte, e.g. « **000** » dans « déjà 300 **000** morts »), les dates (e.g. « **11** » dans « attentats du **11** septembre 2001 »), les siècles (e.g. « datent du **XIXème** siècle ») ou les noms de Républiques (e.g. « de la **Ve** République »). En revanche, le contexte lui pose problème lorsqu'il contient des mots polysémiques (e.g. pour le contexte « à peine **10** jours après », le modèle propose des mots comme « emprisonnement », « incarcération » et « perpétuité » à la place du nombre « **10** »).

   — À l'inverse, il est naturellement plus difficile pour le modèle entraîné avec une taille de fenêtre plus grande de prédire les chiffres qui sont souvent utilisés comme des déterminants et nécessitent d'avoir une vision très locale du contexte pour les prédire. La représentation du contexte en sacs de mots empêche toute prédiction lorsque le contexte

contient plus de catégories nominales et verbales. Ainsi, même soumis à évaluation humaine, il est difficile de dire qu'il manque un adjectif numéral dans un exemple comme « il arrive quand même systématiquement en **5ème** position » si on considère les mots indépendamment de leur ordre dans la phrase.

2. Concernant la prédiction des pronoms (PRO) :

   — De la même manière, une fenêtre réduite semble permettre au modèle de reconnaître plus facilement les pronoms personnels et impersonnels (ex.« **il** faut que ») pour lesquels le contexte contient généralement le verbe qu'ils accompagnent, mais l'empêche de prédire les pronoms lorsque le contexte contient les entités nommées auxquelles ils réfèrent. Le modèle montre alors une tendance à proposer plutôt les associations les plus probables pour cette entité (e.g. pour le contexte ['Patrick_NAM', 'Buisson_NAM', 'être_VER', 'quelqu'un_PRO'], le modèle ne propose que des noms de personnes, probablement souvent associées à l'entité nommée « Patrick Buisson » dans le corpus *Wikipédia* au lieu du pronom attendu « **ce** » pour « Patrick Buisson **c'**est quelqu'un de...»).

   — la taille de fenêtre plus grande semble au contraire permettre au modèle entraîné avec une fenêtre de 15 de prédire des mots qui ont un sens proche de celui à prédire, indépendamment de la catégorie grammaticale. Par exemple, pour une phrase « Je n'ai **aucune** espèce d'hésitation », le modèle ne propose que des mots du champ lexical de l'absence (i.e. « pas_ADV », « guère_ADV », « rien_ADV », « aucun_PRO » et « jamais_ADV »).

Nous notons que les catégories des noms communs (NOM), adjectifs (ADJ) et noms propres (NAM) présentent les distances moyennes les plus élevées. Ces résultats peuvent s'expliquer en partie par le fait qu'il s'agit de catégories productives et en liste ouverte, qui présentent parfois des formes très peu fréquentes dans le corpus d'entraînement. Puisque nous entraînons notre modèle sur un corpus très différent de notre corpus d'évaluation, nous pouvons également nous questionner sur l'impact des représentations apprises sur la prédictibilité des noms. En effet, il est peu probable que les catégories qui relèvent plutôt de la syntaxe (comme les conjonctions ou déterminants) présentent un usage très différent d'un corpus à l'autre. En revanche, les catégories ouvertes comme les noms et adjectifs ont plus de probabilité de voir leur sens modifié en passant d'un corpus encyclopédique à un corpus politique plus subjectif, rendant leur prédiction exacte plus compliquée.

En observant les étiquettes grammaticales des mots prédits par les modèles, nous pouvons observer néanmoins qu'elles sont également les catégories les plus faciles à reconnaître. Le tableau 2 donne le ratio moyen, par catégorie grammaticale, de prédictions d'étiquette grammaticale identiques à celle du mot à prédire parmi les 5 premières prédictions données par chaque modèle.

En effet, les résultats pour les noms communs et propres, les verbes, les adverbes et les adjectifs sont les plus élevés, ce qui signifie qu'il est plus facile pour les modèles de prédire si le mot manquant est un nom plutôt qu'un déterminant. Là encore, la variation quand on passe d'un modèle à l'autre est intéressante, puisque le modèle entraîné avec une plus grande fenêtre montre une tendance plus élevée à donner des prédictions d'étiquettes différentes du mot à prédire, particulièrement lorsqu'il s'agit de noms.

En réalité, en observant les données, nous pouvons nous rendre compte que le modèle, au lieu de ne proposer que des noms qui lui paraissent proches des mots qui apparaissent dans le contexte étudié, propose également les verbes et adjectifs dont le sens est proche du mot à prédire. Par exemple,

Prédictions	Fenêtre de 5 (en %)	Fenêtre de 15 (en %)
NOM	48	35
VER	41	40
ADV	32	27
NAM	23	12
PRO	18	18
ADJ	17	9
NUM	14	4
KON	6	7
PRP	3	1
DET	0	0

TABLE 2 – Ratio moyen de prédiction des catégories grammaticales

pour une phrase « il n'est pas en situation d'être candidat à l'**élection** présidentielle », le modèle va prédire des mots comme « présidentiel_ADJ » et « réélire_VER » en plus de « élection_NOM ». Ce phénomène est encore une fois dû à la prise en compte du contexte comme un sac de mots, qui est corrigé lorsque la fenêtre prise en compte est plus petite puisque le modèle fait un focus sur les cooccurrents directs du mots (ici l'article défini « le » et l'adjectif « présidentielle »). Néanmoins, dans certains contextes où les éléments nécessaires à la prédiction du mot sont plus éloignés dudit mot, cette vision plus élargie du contexte peut s'avérer utile. C'est le cas par exemple pour l'étude de la phrase « À l'origine, certaines personnalités chez vous militaient pour une large primaire à **gauche** qui engloberait tout le monde, des socialistes jusqu'à Jean-Luc Mélenchon. » pour laquelle le modèle donne des prédictions comme « gauche », « écologiste », « communiste » et « socialiste ».

De manière générale, il semble effectivement que réduire la taille du contexte pour l'apprentissage permet de mieux intégrer la syntaxe dans les prédictions du modèle, mais une taille de contexte plus grande est parfois nécessaire lorsque les éléments indispensables à la compréhension de l'extrait se trouvent plus loin ou que les éléments du contexte sont trop ambigus.

Si les objectifs du modèle sont de détecter la nomination, il est peu probable que tous les éléments nécessaires à la compréhension du nom soient disponibles dans le contexte restreint autour du mot-cible. Cette expérience nous conforte donc dans l'idée qu'un contexte plus large est préférable pour l'apprentissage de nos modèles.

En renouvelant l'expérience, cette fois avec un modèle entraîné avec une taille de fenêtre de 10, nous nous apercevons que les distances moyennes et écarts-types varient très peu par rapport à ceux mentionnés dans le tableau 1 (i.e. seulement 0.01 ou 0.02 points d'écart, toutes catégories confondues avec le modèle entraîné avec un contexte fixé à 15). Comme pour l'expérience menée par (Patel & Bhattacharyya, 2017), il semble donc qu'il existe un seuil de fenêtre à partir duquel les résultats ne varient plus. Cette observation nous pousse à fixer notre taille de contexte à 10 pour nos prochaines expérimentations.

4 Prédiction de mot-pivot

En utilisant le même principe que dans notre expérience sur les catégories grammaticales présentée dans les deux sections précédentes, nous souhaitons observer le comportement de notre modèle pour

la prédiction d'un mot pivot selon la méthode habituelle de l'analyse du discours. Nous supposons que si le modèle propose dans sa liste de prédictions des mots très éloignés du mot à prédire, ce résultat implique que le mot en question est employé de manière inhabituelle (i.e. apparaît dans un contexte peu probable) et peut relever de la nomination.

Pour cette étude, nous réunissons un sous-corpus autour du mot-pivot « Europe », déjà identifié par d'autres travaux (Gauthier, 2016) comme candidat nomination et très fréquent dans notre corpus d'interviews. Nous extrayons grâce au concordancier de l'outil TXM (Heiden *et al.*, 2010) 5855 contextes contenant la forme « Europe ».

Pour les raisons citées en fin de section précédente, nous travaillons cette fois avec un modèle entraîné avec une taille de fenêtre de 10. L'analyse manuelle des résultats nous permet déjà d'identifier plusieurs sens pour le mot « Europe », à savoir :

— un **territoire** : le modèle parvient à identifier les usages du mot au sens de lieu lorsque des mots qui s'en rapprochent ou qui indiquent une position font partie du contexte (e.g. « Monde », « France », « dans » , « en »).

— une **puissance économique** : on retrouve pour plusieurs contextes des prédictions du champ lexical de la puissance économique (e.g. « empire », « compétitivité », « prospérité », « capitalisme », « mondialisation », « libéralisation »). Des prédictions comme « utopie » ou « idéologie » permettent déjà de mettre en lumière cette vision colorée de l'entité dans un contexte comme « l'avenir progressiste de l'**Europe** et du monde ».

— un **peuple** : on trouve dans certains cas comme « l'humanité, notre pays, l'**Europe**, le monde » ou « l'**Europe** immobile, l'Europe du sceptique » des prédictions comme « civilisation » ou « civiliser » qui présentent l'Europe comme une entité dotée de vie et d'intelligence.

— une **victime** ou un **bourreau** : d'autres exemples permettent de déceler l'inquiétude du locuteur pour ou sur l'entité, par exemple dans « l'**Europe** cause de tout le mal », contexte pour lequel le modèle propose des mots comme « esclavage », « humiliation », « trahison » et « injustice », ou au contraire dans l'exemple « c'est une menace pour l'Europe » pour lequel il propose des mots comme « victime », « trahison » ou encore « esclave ».

Néanmoins, ces résultats restent difficiles à exploiter puisque la majorité des prédictions comprennent des résultats très éparses, souvent très éloignés sans que l'on puisse déterminer s'ils sont la conséquence d'une nouvelle utilisation du mot ou d'une mauvaise représentativité des données dans le corpus d'entraînement.

5 Discrimination de sens

Dans une seconde approche, nous envisageons de confronter directement des modèles appris sur différents corpus (chacun représentatif d'un locuteur différent) pour mesurer la variation sémantique interdiscursive (i.e. les changements de sens appliqués à chaque mot en fonction du locuteur).

Nous souhaitons pour cela comparer un modèle appris sur le corpus *Wikipédia* avec différents modèles appris sur des concaténations du corpus *Wikipédia* et les sous-corpus contenant à chaque fois les interviews d'un seul candidat. En suivant la méthode utilisée par Pierrejean & Tanguy (2018), nous comptons ainsi faire une comparaison par binômes de modèles en observant la variation des voisins distributionnels les plus proches (*Nearest Neighbors*) de chaque mot (i.e les mots dont le score de

similarité cosinus est le plus élevé par rapport au mot cible). Après avoir identifié le vocabulaire commun à chaque paire de corpus (i.e. en retirant de l'étude tous les mots du corpus *Wikipédia* qui n'apparaissent pas dans les interviews), nous utiliserons la distance de Jaccard pour mesurer le taux de variation entre chaque modèle.

Nous souhaitons également utiliser ces représentations pour regrouper des clusters de sens sur le modèle de Pelevina *et al.* (2017) pour tenter d'identifier de nouveaux aspects sémantiques, mais également rapprocher les usages de mots identiques d'un locuteur à l'autre.

Grâce à cette étude, nous souhaitons mettre au jour les représentations qui évoluent d'un modèle à l'autre, pour identifier les usages qui relèveraient soit d'un usage spécifique au locuteur, soit d'un usage spécifique au genre de l'interview politique.

Cette étude devra néanmoins prendre en compte l'instabilité inhérente aux méthodes d'apprentissage des modèles (Pierrejean & Tanguy, 2019) dans l'évaluation des représentations de variations sémantiques inter-locuteur pour discriminer les variations dues à l'entraînement des modèles des variations effectivement dues à des usages différents.

6 Conclusion et perspectives

Dans l'objectif de fournir une méthode capable de rendre compte de la variation sémantique de la nomination, qui viendrait compléter les fonctionnalités des outils déjà utilisés par l'analyse du discours, cet article propose une exploration de méthodes distributionnelles pour repérer de manière automatique des candidats nominations.

Notre problématique, qui nécessite de détecter les usages particuliers des noms (i.e. dont le sens s'écarte de l'habituel), se heurte à la représentation vectorielle unique de chaque forme calculée par les modèles prédictifs. Pour répondre à cette question, cet article propose un état de l'art des méthodes distributionnelles appliquées à la désambiguïsation lexicale et la représentation de la variété sémantique sous forme de vecteurs de sens duquel nous nous inspirons pour nos propres travaux.

Nos expérimentations, menées sur un corpus de transcriptions d'interviews politiques enregistrées dans des émissions de radio, questionnent la capacité de modèles prédictifs à représenter la variation sémantique d'un discours à l'autre. Une première expérience, menée sur la prédictibilité des catégories grammaticales, nous permet d'évaluer l'impact du contexte pris en compte dans l'entraînement des modèles sur la représentation des mots porteurs de sens. Sur le modèle de travaux en analyse de discours, nous utilisons le modèle qui résulte de cette expérimentation pour observer son utilisabilité pour la prédiction de nouveaux usages d'un mot-pivot. Les résultats, bien qu'encourageants, restent difficiles à exploiter en l'état et ne permettent pas de discerner facilement les résultats effectivement dus à un usage inhabituel ou à une lacune du modèle appris. Enfin, en nous inspirant d'approches concentrées sur l'étude des voisins distributionnels, notre article amorce une réflexion sur une méthode de détection de la variation sémantique interdiscursive basée sur une comparaison par binômes de modèles représentatifs respectivement d'un usage encyclopédique *vs.* en discours du lexique.

Nous espérons, à terme, que nos travaux nous permettrons de définir des critères de sélection automatique de candidats nominations à intégrer à des outils déjà existants pour l'analyse de discours.

Pour des raisons pratiques, nous avons décidé de ne pas nous pencher plus sur l'aspect diachronique de l'étude de la nomination, qui est normalement repérable en premier lieu sur le même modèle que la dénomination de Kleiber (1984), par son acte de baptême (Siblot, 2001) (i.e. lorsque le locuteur décide de nommer par tel nom précis l'entité de son choix). Ce phénomène n'étant repérable qu'à

l'introduction de la nomination, nous pouvons douter de la représentativité de notre corpus sur cet aspect, et nous cantonner pour le moment à une approche synchronique nous paraît plus sage. Néanmoins, nous ne tirons pas de trait définitif sur la perspective d'étendre ultérieurement l'approche à d'autres corpus pour inclure cette particularité à notre travail.

Aussi, nous n'abordons pas dans ce papier la question de modèles plus performants largement exploités dans le domaine du TAL, tels que BERT (Devlin *et al.*, 2018) et ELMo (Peters *et al.*, 2018). Nous avons choisi de débuter nos travaux avec des modèles dont la documentation nous paraît plus accessible et l'entraînement moins coûteux pour éprouver nos méthodes, mais une expérimentation de modèles neuronaux profonds est prévue dans la poursuite de notre étude.

Références

BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 238–247 : Association for Computational Linguistics. DOI : 10.3115/v1/P14-1023.

BLOEM J., FOKKENS A. & HERBELOT A. (2019). Evaluating the consistency of word embeddings from small data. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*.

CALABRESE L. (2018). Faut-il dire migrant ou réfugié ? débat lexico-sémantique autour d'un problème public. *Langages*, **210**, p.105–124. DOI : 10.3917/lang.210.0105.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DONGSUK O., SUNJAE K., KYUNGSUN K. & YOUNGJOONG K. (2018). Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph. In *Proceedings of the 27th International Conference on Computational Linguistics* : Association for Computational Linguistics.

FIRTH J. (1957). *A synopsis of linguistic theory*. Blackwell, Oxford.

GAUTHIER G. (2016). Le « printemps érable » au québec : « grève » ou « boycott » ? les enjeux stratégiques d'un conflit de nomination. *Argumentation et Analyse du Discours (AAD)*, **17**. DOI : 10.4000/aad.2248.

HARRIS Z. (1954). Distributional structure. *Word*, **10**, 146–162.

HEIDEN S., MAGUÉ J.-P. & PINCEMIN B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie – conception et développement. In *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010 (Volume 2)*, p. p.1021–1032, Roma, Italy.

KLEIBER G. (1984). « dénomination et relations dénominatives ». *Langages*, **76**, p.77–94. DOI : 10.3406/lgge.1984.1496.

KOREN R. (2016). La nomination et ses enjeux socio-politiques : Introduction. *Argumentation et Analyse du Discours (AAD)*, **17**. DOI : 10.4000/aad.2295.

LEBART L. & SALEM A. (1994). *Statistique textuelle*. Dunod.

LEVY O. & GOLDBERG Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 302–308 : Association for Computational Linguistics. DOI : 10.3115/v1/P14-2050.

LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving distributional similarity with lessons learned from word embeddings. In *Transactions of the Association of Computational Linguistics, Volume 3*, p. 211–225 : Association for Computational Linguistics. DOI : 10.1162/tacl_a_00134.

LONGHI J. (2017). Humanités, numérique : des corpus au sens, du sens aux corpus. *Questions de communication*, **31**, p.7–17.

MARCHAND P. & RATINAUD P. (2012). L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). In *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT*, p. 687–699.

MAZIÈRE F., Éd. (2018). *L'analyse du discours : Histoire et pratiques. "Que sais-je ?"*. Presses Universitaires de France.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint, arXiv : abs/1301.3781*.

NOUVEL D., EHRMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Group.

PATEL K. & BHATTACHARYYA P. (2017). Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*.

PELEVINA M., AREFYEV N., BIEMANN C. & PANCHENKO A. (2017). Making sense of word embeddings. *arXiv preprint, arXiv : abs/1708.03390v1*.

PENGAM M. & JACKIEWICZ A. (2019). Sens et emplois de l'expression « musulmans modérés » dans les discours médiatiques. *Open Library of Humanities*, **5**, p.45. DOI : 10.16995/olh.431.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365v2*.

PIERREJEAN B. & TANGUY L. (2018). Towards qualitative word embeddings evaluation : Measuring neighbors variation. In *Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 32–39, New-Orleans, United States. HAL : hal-01806468.

PIERREJEAN B. & TANGUY L. (2019). Investigating the stability of concrete nouns in word embeddings. In *13th International Conference on Computational Semantics*, p. 32–39, Gothenburg, Sweden. HAL : hal-02073705.

REINERT M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, **66**, p.5–39.

RUAS T., GROSKY W. & AIZAWA A. (2019). Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications*, **136**, 288–303. DOI : 10.1016/j.eswa.2019.06.026.

SCHMID H. (2013). Probabilistic part-of-speech tagging using decision trees. *New methods in language processing*, p. 154.

SIBLOT P. (2001). De la dénomination à la nomination. *Cahiers de praxématique*, **36**, p.189–214. DOI : 10.3406/lgge.1997.2124.

TOURNIER M. (1981). Spécificité politique et spécificité lexicale. *Mots. Les langages du politique*, **2**, 5–10.

VIAL L., LECOUTEUX B. & SCHWAB D. (2017). Sense embeddings in knowledge-based word sense disambiguation. In *12th International Conference on Computational Semantics*. HAL : hal-01599685.