HAL open science

LDApol: vers une méthodologie de contextualisation des discours politiques

Jeanne Vermeirsche, Eric Sanjuan, Tania Jiménez

To cite this version:

Jeanne Vermeirsche, Eric Sanjuan, Tania Jiménez. LDApol: vers une méthodologie de contextualisation des discours politiques. Traitement Automatique des Langues Naturelles, 2022, Avignon, France. pp.19-27. hal-03701469

HAL Id: hal-03701469

https://hal.science/hal-03701469v1

Submitted on 24 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LDApol: vers une méthodologie de contextualisation des discours politiques

Jeanne Vermeirsche [1] Eric SanJuan[2] Tania Jiménez[2]
(1) LBNC, 74 rue Louis Pasteur, 84 000 Avignon, France
(2) LIA, 339 chemin des Meinajaries BP 1228, 84 911 Avignon, France
jeanne.vermeirsche@univ-avignon.fr, eric.sanjuan@univ-avignon.fr, tania.jimenez@univ-avignon.fr

RÉSUMÉ

Nous comparons les distributions de mots dans les communiqués de presse politiques récents. Nous proposons une méthodologie pour objectiver des associations entre notions participant au débat politique. Nous montrons comment les modèles de langage probabilistes peuvent révéler les concepts sous-jacents en tant qu'associations fortes à plusieurs termes pour aider à clarifier le débat politique, notamment pour la surveillance des médias sociaux. Cette approche tente de modéliser les termes du débat comme des distributions de probabilités d'apparition des mots.

ABSTRACT

LDApol : towards a methodology of political speech contextualisation

We compare word distributions in recent political press releases. We propose a methodology to objectify associations between notions participating in the political debate. We show how probabilistic language models can reveal the concepts underlying as strong multi-term associations to help clarify political debate, especially for social media monitoring. This approach attempts to model the terms of the debate as probability distributions of words appearance.

MOTS-CLÉS : Contextualisation - Nationalisme - Polarisation - Discours politiques - LDA - corpus - modélisation.

KEYWORDS: Contextualization - Nationalism - Polarization - Political discourses - LDA - corpus - modelization.

1 Introduction

Un parti politique est un univers linguistique en lui-même. Il déploie son propre vocabulaire fait de marqueurs et de références spécifiques (Lefebvre, 2022). Ce lexique le fait exister et permet de le spécifier en regard des autres. Les partis essayent d'influencer l'agenda politique en le politisant autour de certaines questions et enjeux (Lefebvre, 2022), tentant de fait d'imposer des thèmes dont chaque parti serait le "propriétaire". Toutefois, les autres partis politiques ne l'entendent pas ainsi et cherchent à leur tour à s'approprier ces enjeux (immigration, sécurité...) (Lefebvre, 2022). Chacun s'oppose dans sa propre langue et son propre langage politique sur ces thèmes mis en avant et sur les prises de positions qui sont les leurs. Dans un débat politique ainsi polarisé, les termes ont souvent des significations différentes ou des références cachées qui ne peuvent être comprises que dans le contexte politique.

Le durcissement des discours et leur polarisation nous a conduit à expérimenter une modélisation multinomiale afin de pouvoir extraire automatiquement les associations de termes caractéristiques d'un parti ou d'une personnalité politique sur une période donnée. Pour calculer effectivement ces associations nous utilisons les multinomiales produites par l'algorithme d'allocation latente de Dirichlet (LDA) introduit en (Blei *et al.*, 2003). Nous utilisons les implémentations récentes disponibles dans R et les tests non paramétriques de corrélation pour vérifier la significativité des associations automatiquement extraites par le modèle ou formulées comme hypothèses par le chercheur. La détection du vocabulaire et des associations systématiques d'une communication politique est nécessaire à l'étude de l'éventuelle propagation des idées induites dans les médias, classiques (tel que la presse écrite en ligne), sociaux (tels que les applications de micro blog) ou encyclopédiques (tel que le WikiPedia).

Dans cet article nous présentons un premier travail qui permet de questionner notre méthodologie, que nous proposons pour objectiver des associations entre notions participant au débat politique. Nous montrons comment les modèles de langage probabilistes peuvent révéler les concepts sous-jacents en tant qu'associations fortes à plusieurs termes pour aider à clarifier le débat politique, notamment pour la surveillance des médias sociaux. Cette approche tente de modéliser les termes du débat comme des distributions de probabilités d'apparition des mots.

2 Contexte

« Les discours ne sont pas seulement des signes destinés à être compris, déchiffrés, ce sont aussi des signes de richesse destinés à être évalués, appréciés et des signes d'autorité destinés à être crus et obéis. » (Bourdieu, 1981). Les discours peuvent exprimer de diverses façons les relations de pouvoir. Le discours politique est d'abord une proposition de mise en ordre - en premier lieu des *« divisions du monde social »* (Bourdieu, 1981) - et donc de mise en forme du monde tel qu'il devrait être. C'est la langue de l'Etat, et donc le politique, qui impose une langue "officielle" et *«une « mise en forme » que suppose l'usage officiel » de la langue»* (Bourdieu & Thompson, 2001). Pour autant, le discours politique ne reflète que superficiellement ce qu'un locuteur politique souhaite exprimer (le vouloir dire) à un moment donné (Le Bart, 2003). Il ne reflète donc pas la vérité du monde social mais *«l'état du champ politique au moment où il est produit ainsi que la position occupée, dans ce champ, par celui qui parle»* (Le Bart, 2003). Le discours politique n'est pas le lieu de la vérité de la parole publique mais plutôt celui de sa véracité, de sa force de persuasion (Charaudeau, 2011).

La compétition partisane s'incarne dans un langage propre à chacun, décliné dans les thématiques portées par les partis politiques selon un agenda politique que chacun tente d'influencer : *«Parce qu'il cherche à administrer un sens commun, le parti cherche à promouvoir un lexique qui lui donne corps»* (Lefebvre, 2022). Ces lexiques, loin d'être figés, subissent des reprises, détournements et redéfinitions par l'ensemble de la classe politique. Actuellement, la compétition partisane est fortement polarisée autour de certains thèmes. De nombreuses idées sont ainsi partagées par l'extrême-droite et une partie de la classe politique, notamment de la droite traditionnelle, mais pas seulement puisque cette "confusion" s'étendrait de l'extrême-droite à l'extrême-gauche (Corcuff, 2021). Ainsi, les thématiques nationalistes fleurissent dans les discours et les programmes politiques, en temps de campagne mais aussi hors campagne. Porté et revendiqué depuis de nombreuses années par les formations d'extrême-droite (Greenfeld, 1992), celles-ci participent par là à la redéfinition et à la diffusion idéologique d'un nationalisme dit exclusif (Zimmer, 2003; Bonikowski *et al.*, 2019).

Nous proposons, à travers cette étude, de saisir et les éléments de langage empruntés à l'extrême-droite, et leur évolution, dans le discours politique des partis étudiés. Nous cherchons à mettre en évidence un répertoire linguistique, à travers un vocabulaire et des éléments de langage associés au nationalisme, et leur diffusion dans les discours des différents partis étudiés.

3 Méthodologie

Le schéma suivant précise notre méthodologie.

Il est constitué de quatre entités distinctes :

— le corpus de textes (représenté en orange) est composé des communiqués de presse produits par les partis politiques. (voir 3.1)

— calcul de modèles statistiques ; pour objectiver l'analyse de ces textes, nous calculons un modèle LDA à partir d'un sous-ensemble du corpus (par période et/ou parti) (voir 3.2)

— création d'un lexique nommé *"lex_asso"*(voir 3.3)

— extraction d'associations entre mots. De ces modèles il est possible d'objectiver un ensemble d'associations (représentées par le rectangle central) qui sont comparées par des concepts systématiques (voir 3.4)

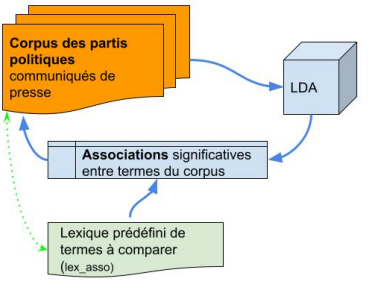


FIGURE 1 – Méthodologie

3.1 Constitution du corpus

L'une des premières tâches pour pouvoir analyser les discours politiques a été de collecter un corpus. Ce premier corpus se compose des communiqués de presse (CPs) de cinq partis politiques français (Rassemblement National, Les Républicains, La République en Marche, Parti Socialiste, et France Insoumise. L'analyse porte sur une période allant du 01/03/2016 (période de création du parti politique français LREM) au 01/08/2022 (deux mois après les élections présidentielles et législatives en France).

Les communiqués ont été organisés dans une base de données relationnelle accessible en ligne par date, nom du parti et auteur officiel du communiqué. La collecte a été effectuée dans un premier temps par scrapping depuis R, puis manuellement à partir de septembre 2021.

Nous avons aussi tenté d'automatiser entièrement la cueillette des discours politiques grâce à la wayback machine (http ://www.wayback.com/). Il s'est avéré que ces archives sont trop incomplètes. Toutefois, la wayback machine est utilisée de façon manuelle pour retrouver les CP qui n'ont pu être collectés dans la première phase de recueil automatique à cause des trop nombreux changements de scripts des sites internet des partis. En effet, la collecte de ces CP est un enjeu en lui-même. Ces documents de communication politique officielle ne sont pas archivés, notamment par la BNF. De fait, dès que le parti politique décide de modifier son site internet, ou tout simplement de les supprimer, les CP disparaissent. Recueillir et archiver les CP produits par les partis politiques dans une base de données est donc un élément central de ce travail.

Nous définissons des modèles linguistiques par parti ou leader politique. Nous n'appliquons aucune liste de mots vides. Ne pas utiliser de liste de mots vides semble être important lorsqu'il s'agit de textes politiques en français, car des pronoms comme "nous", "eux", "moi", "je" peuvent être utilisés pour renforcer un message politique. Les articles sont également importants en français. Par exemple « la femme » et « les femmes » peuvent faire référence à des concepts différents. Dans le cadre de ce travail pluridisciplinaire, nous pensons que le genre, féminin ou masculin, et le nombre, singulier ou pluriel, a son importance. Par exemple, lors de nos premières analyses, nous avons pu constater que :

— le Rassemblement national utilise majoritairement le terme "étrangers" plutôt que celui d'"étranger" dans ses discours. Les associations ne sont pas les mêmes selon que le mot soit au singulier ou au pluriel.
— Les Républicains utilisent "étrangers" et "étranger" de façon parfaitement semblable - les associations et le degré de probabilité associé sont à chaque fois les mêmes pour les deux termes. Cette situation est également celle de la France Insoumise.

Ce sont donc autant les associations entre termes que les emplois différenciés de ces termes (en genre et en nombre) selon les partis qui doivent retenir notre attention et être sujet de notre analyse.

La constitution de ce corpus est supervisée, ce n'est pas une simple collecte et sera mis à la disposition de la communauté.

3.2 Modélisation probabiliste du discours politique

Nous considérons les CP officiels des partis. Nous supposons que les termes à inclure dans un communiqué de presse sont choisis selon un agenda politique interne ciblant des groupes spécifiques d'électeurs. Notre approche pour révéler le public ciblé repose sur l'exploration d'ensembles de termes significativement corrélés sur la base d'un modèle de langage génératif et de tests statistiques.

Nous utilisons ainsi une représentation en sac de mots où nous avons choisi d'ignorer l'ordre des mots dans les CP politiques et de nous concentrer sur les co-occurrences et les fréquences des termes. En effet, les CP politiques diffèrent des textes argumentatifs ou narratifs. Ils sont courts et s'appuient souvent sur un nombre réduit d'affirmations. Nos recherches nous conduisent à poser l'hypothèse que les CP sont des instruments de communication officielle spécifiques des partis politiques. Nous supposons ainsi que les CP sont des mises en forme spécifique du discours politique, lui-même mise en ordre et mise en forme du monde tel qu'il devrait être (Bourdieu, 1981). Ils suivent un ordre établi rigoureux où chaque terme à son importance, où les sujets et les concepts ont été préalablement pensés et discutés avec des conseillers politiques.

Ces hypothèses nous permettent d'appliquer le modèle génératif d'allocation Dirichlet latente introduit par (Blei *et al.*, 2003) qui repose sur le concept d'échangeabilité mis en exergue par De Finetti (Cifarelli & Regazzini, 1996).

On considère un vocabulaire Ω de taille m. Un corpus $\mathcal{C}$ de textes est considéré comme une suite infinie de mots $(\omega_i)_{i\in\mathbf{N}} \in \Omega^{\mathbf{N}}$. Si l'on considère un vocabulaire déterminé et fixe, on ne suppose qu'il soit possible de disposer d'un ensemble exhaustif de textes. L'hypothèse d'une suite infinie correspond à l'idée que le corpus est incomplet par nature et qu'il est toujours possible de trouver et d'ajouter des textes à celui ci, ce qui correspond à la situation d'une constitution de corpus à partir d'une collecte sur les sites des partis politiques et dans des archives. Par contre il n'est pas possible d'enrichir le vocabulaire sans devoir recalculer l'ensemble du modèle. Cela nous contraint à travailler sur des temporalités où le vocabulaire est stable. L'hypothèse que l'ordonnancement des mot dans

un discours politique n'est pas informatif n'induit pas que ces mots sont choisis indépendamment les uns des autres. Il peut exister des associations fortes entre eux. L'hypothèse d'échangeabilité est moins forte que celle d'indépendance. Le théorème de De Finetti (Aldous, 1985) induit cependant que les probabilités conjointes d'un ensemble fini est un mélange de variables aléatoires indépendantes identiquement distribuées. Dans le cas des textes, chaque mot étant associé à une variable aléatoire de loi Binomiale (choisir ou pas ce mot), le théorème induit l'existence d'un mélange de variables multinomiales explicatives des dépendances entre mots.

Le modèle génératif LDA fait l'hypothèse simplificatrice qu'il existe une distribution de Dirichlet permettant d'exprimer ce mélange. Cette hypothèse induit que l'ensemble des multinomiales recherchées est fini et que le mélange se réduit à une moyenne pondérée de celles-ci. Plus formellement, le modèle génératif LDA suppose que les probabilités conjointes de mots peuvent s'écrire :

$$\mathbf{P_C}(X) = \sum_{i=1}^{i=k} \theta_i \prod_{\omega \in X} \beta_i(\omega) \tag{1}$$

avec :

- $X \subseteq \Omega$ un ensemble fini de mots.
- $\mathbf{P_C}(X)$ probabilité de trouver l'ensemble de mots dans X dans un même texte.
- $\theta = (\theta_1, ..., \theta_k) \in [0,1]^k$ est une suite fini de k pondérations tel que $\sum_{i=1}^{k} \theta_i = 1$. k est un entier qui est un paramètre du modèle génératif qu'il faut fixer à priori.
- $\beta = (\beta_1, ..., \beta_m) \in [0,1]^m$ est une distribution de probabilités sur les mots.

L'équation 1 induit une hypothèse forte sur la forme des mélanges alors que les preuves du théorème de De Finetti reposent sur la théorie des mesures de Borel Lebesgue, qui généralisent la notion de mélange. Par contre cette équation permet de mettre en oeuvre de multiples calculs approchés des multinomiales β. La qualité de cette approximation dépend principalement du paramètre k, qui ne peut pas être inféré, et du vocabulaire Ω sélectionné. Pour tenir compte de ce processus génératif aléatoire nous utilisons le test de corrélation de Pearson entre représentation des mots $v(\omega)$ définies par : $v(\omega) = (\beta_1(\omega), ..., \beta_k(\omega))$. On considère ainsi les k multinomiales β_i calculées par le modèle génératif LDA comme autant d'observations continues sur les mots et le théorème de De Finetti assure l'indépendance de ces observations dans l'hypothèse d'une convergence.

Pour s'assurer de cette convergence par la suite, on préfère utiliser l'algorithme d'espérance-maximisation par estimation variationnelle (VEM) proposé dans la publication d'origine (Blei *et al.*, 2003) que de procéder par échantillonnage de Gibbs. Bien que beaucoup plus rapide sur notre corpus, et donc mieux adapté à une analyse interactive, le test de normalité de Kolmogorov-Smirnov rejette l'hypothèse de normalité des représentions $v(\omega)$ des mots obtenues pour $k \geq 12$ par échantillonage de Gibbs, alors que cette hypothèse est plus rarement rejetée dans le cas d'utilisation de VEM. Il est aussi possible d'accélérer l'approche par VEM par l'utilisation de réseaux neurones, cependant celle-ci s'avère complexe dans le cas de distributions de Dirichlet (Tian *et al.*, 2020).

Le choix de fixer $k = 12$ permet d'obtenir un plongement du vocabulaire Ω en des vecteurs de distribution Gaussienne sur $[0,1]$ qui rendent possible l'application de tests statistiques paramétriques. Les métriques introduites en (Deveaud *et al.*, 2014) qui tiennent compte de la diversité des β_i induisent des valeurs de k plus élevées, mais les représentations qui en résultent ne satisfont pas les critères de normalité.

3.3 Création du lexique

Comme précisé en §3.2, la détermination du vocabulaire Ω étudié est un paramètre essentiel du modèle choisi. Nos lectures scientifiques, couplées à celles des CP et programmes politiques, ainsi

qu'à une veille politique et médiatique active, nous ont permis de définir une liste de 133 termes à ce jour - *"lex_asso"* - pour qualifier le nationalisme et le discours associé. Ce lexique est en partie extrait des CP politiques recueillis. À travers ce lexique, nous cherchons à révéler les glissements sémantiques, les modifications que peut subir le discours politique à propos du nationalisme. Nous travaillons à créer une liste de termes qui prennent en compte ces évolutions, notamment de sens.

Par exemple, les termes "nation" et "nationalisme" ont globalement une fréquence peu élevée au sein des communications officielles des partis politiques. Toutefois, comprendre et contextualiser leur(s) emploi(s) au sein du discours politique est central en regard de notre objet de recherche.

Nous cherchons à savoir plus spécifiquement comment ces termes spécifiques, même s'ils sont moins fréquemment employés dans le discours politique, sont associés avec quels autres termes, dans quels contextes et à quelles temporalités plus précises (temps de campagne ; temps hors-campagne ; pas de période spécifique ; etc). Notre corpus peut ainsi être interrogé à partir de ces termes du lexique et des associations définies. Nous nous intéressons donc dans le cadre de ce travail non pas à l'ensemble du discours politique mais à la partie du discours politique qui croise les termes choisis. Nous procédons à une analyse de discours supervisée par un lexique.

3.4 Exploration des associations

Nous cherchons à objectiver des éléments de langage et la récurrence des associations propres à chaque formation politique pour une ou plusieurs périodes données. Pour ce faire nous utilisons le lexique *"lex_asso"* introduit en §3.3.

Nous procédons à une analyse statistique des textes en comparant la distribution des mots dans les textes et les corrélations significatives par corpus. Un modèle LDA §3.2 est extrait par formation politique. Ces modèles sont comparés et interrogés sur les termes utilisés par chaque parti politique.

L'ensemble du dispositif a été mis en place en utilisant un serveur Rstudio couplé à une base de données PostGreSQL. La recherche des termes et des associations caractéristiques pour ce corpus se fait de manière interactive à l'aide d'un programme R (R Development Core Team, 2010) annoté (Rmarkdown) que nous mettons aussi à disposition de la communauté.

4 Résultats

Nous avons effectué une première phase d'analyse portant sur la période : du 1/03/2016 au 15/03/2022 sur 2 (FN/RN et PS) des 5 corpus de CP politiques présentés dans le tableau 1. A ces corpus nous ajoutons une extraction du WikiPedia que nous considérons ici comme un corpus de contrôle [1].

Les corpus sont de taille très variable. La proportion du vocabulaire pour lequel on trouve une représentation en 12 dimensions valide pour un test de Kolmogorov-Smirnov avec une p-valeur supérieure à 0.01 aussi. Pour tenir compte de cette grande variabilité, nous calculons l'indice r de corrélation de Pearson entre représentations $w(\omega)$ des termes. Nous ne retenons que les associations significatives (Student t-test avec une p.valeur inférieure à 1%).

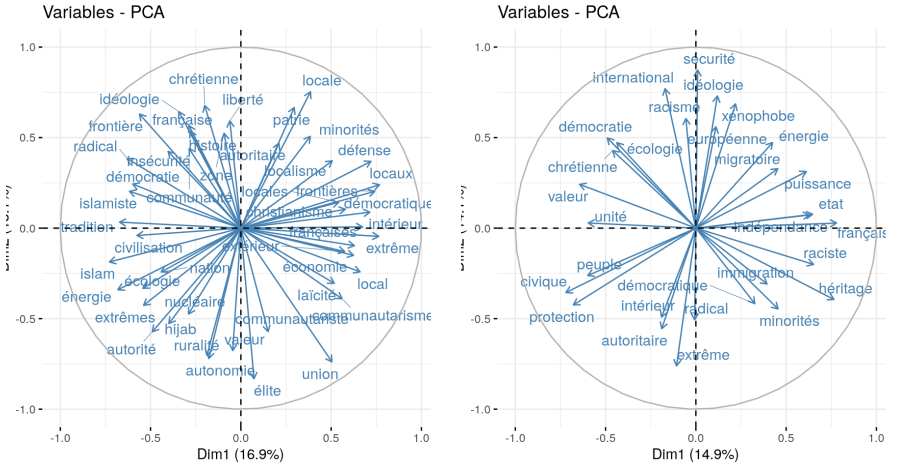
Nous nous focalisons sur l'ensemble des associations entre termes de notre lexique *"lex_asso"*. La

1. Nous avons extraits les 3000 résumés du WikiPedia français qui recouvrent le plus le vocabulaire lex_asso. De ces résumés, nous retenons ceux ayant été actualisés durant la même période que les communiqués politiques.

Corpus	Docs	Longueur	Ω	β	lex_asso
FN/RN	2846	364008	9814	9581	91
LR	398	24882	1587	194	4
LREM	269	33363	1293	379	4
PS	906	95624	4065	906	15
FI	511	49971	2678	545	7
WP	1933	100949	3995	3949	97

TABLE 1 – Caractéristiques des corpus : Nombre de documents (Docs) ; Longueur totale du corpus ; Nombre de mots de fréquence supérieure à 5 qui constituent l'ensemble Ω ; Nombre de mots ayant une représentation normale dans β ; Nombre de ces mots dans le vocabulaire d'étude lex_asso.

figure 2 montre les cercles de corrélation qui résultent de la projection des termes sur deux corpus différents. Le modèle génératif LDA est utilisé ici comme une méthode de lissage des corpus afin de pouvoir les soumettre à une analyse comparative.



FN/RN Parti Socialiste

FIGURE 2 – Projection du vocabulaire d'étude sur les corpus des CP du FN/RN et du PS.

Le choix de petites valeurs pour le nombre k 1 de dimensions conduit à des représentations qui vérifient les hypothèses pour l'application de tests paramétriques de significativité (p-valeur < 1%). Le calcul de l'indice r de Pearson permet de considérer à la fois les corrélations positives qui correspondent à des associations systématiques de termes dans les communiqués de presse, ainsi que négatives qui renvoient à des exclusions. Le tableau 2 présente les 10 corrélations positives les plus significatives dans le cas du PS.

Avant d'analyser ces corrélations, nous rappelons que nous travaillons sur un discours spécifique, celui lié au nationalisme. En testant d'autres lexiques pour d'autres discours, certaines corrélations que nous choisissons de ne pas traiter ici pourraient se révéler pertinentes selon l'objet d'étude.

Sur la base des résultats précédents, nous avons sélectionné 19 associations pour leur intérêt vis à vis de l'analyse politique dont on souhaité évaluer leur significativité sur les différents partis. Pour chacun

Terme 1	Terme 2	r	Terme 1	Terme 2	r
minorités	démocratique	0.77	européenne	sécurité	0.76
peuple	civique	0.76	souveraineté	france	0.75
immigration	frontière	0.74	protection	civique	0.74
européenne	gauche	0.74	indépendance	etat	0.73
idéologie	international	0.73	valeur	local	0.72

TABLE 2 – Corrélations les plus significatives entre termes du PS dans le vocabulaire d'étude lex_asso.

des corpus de CPs, nous avons effectué des tests de corrélation entre différentes représentations vectorielles calculées sur chacun des corpus :

— 3 modèles LDA pour $k \in \{12, 16, 20\}$.
— le modèle factoriel LSA (Peladeau & Davoodi, 2018).
— les plongements lexicaux Vord2Vec. (Mikolov *et al.*, 2013).

Le nombre de relations déterminées comme significatives par chacun de ces corpus et de ces modèles est reporté en table 3 [2].

Corpus	LDA 12	LDA 16	LDA 20	LSA	Word2Vec	total
FN/RN	**11**	10	7	5	8	41
LR			2		2	4
LREM		2		1	2	5
PS	2	1		**7**	2	12
FI	3	2	2		2	9
WP	1	1		4	5	11
total	17	16	11	17	**21**	82

TABLE 3 – Nombre d'associations classées comme significatives sur un total de 19 associations testées (Indice de corrélation > 0.3 et p-valeur < 0.1).

La méthodologie développée ici est bien celle qui détecte le plus d'associations significatives entre termes du lexique lex_asso sur le corpus FN/RN. Il nous reste à explorer le contexte pour chacune de ces associations pour valider ou pas une relation de causalité dans l'usage de ces termes. Par ailleurs, le nombre de relations classées comme significatives décroît avec l'augmentation du nombre de dimensions. Notre choix de travailler sur un petit nombre de dimensions facilite les rapprochements indirects entre termes. Il se confirme aussi que le modèle factoriel LSA diffère du LDA et semble s'avérer plus instable (Peladeau & Davoodi, 2018). Le modèle neuronal Word2Vec renvoie le plus grand nombre de tests positifs sur l'ensemble des corpus.

5 Conclusions et travaux futurs

Nous avons proposé une méthodologie pour révéler des associations indirectes entre termes employés dans une communication politique. Le modèle proposé se prête à l'application de tests paramétriques de corrélation et à une analyse comparative entre différents corpus. Par la suite l'idée est d'objectiver la proximité des thématiques sur des corpus de textes différents ou leur circulation au sein de l'espace public tel que énoncé dans (Velcin *et al.*, 2014).

2. Nous publions l'ensemble de ces résultats et le code qui en résulte sur : https://demo-lia.univ-avignon.fr/lda-pol

Références

ALDOUS D. J. (1985). Exchangeability and related topics. In D. J. ALDOUS, I. A. IBRAGIMOV, J. JACOD & P. L. HENNEQUIN, Éds., *École d'Été de Probabilités de Saint-Flour XIII — 1983*, Lecture Notes in Mathematics, p. 1–198, Berlin, Heidelberg : Springer. DOI : 10.1007/BFb0099421.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**(Jan), 993–1022.

BONIKOWSKI B., HALIKIOPOULOU D., KAUFMANN E. & ROODUIJN M. (2019). Populism and nationalism in a comparative perspective : a scholarly exchange. *Nations and Nationalism*, **25**(1), 58–81. DOI : https ://doi.org/10.1111/nana.12480.

BOURDIEU P. (1981). La représentation politique. *Actes de la Recherche en Sciences Sociales*, **36**(1), 3–24. DOI : 10.3406/arss.1981.2105.

BOURDIEU P. & THOMPSON J. B. (2001). *Langage et pouvoir symbolique*. Éditions Fayard. Google-Books-ID : GqJpQgAACAAJ.

CHARAUDEAU P. (2011). Réflexions pour l'analyse du discours populiste. *Mots. Les langages du politique*, (97), 101–116. DOI : 10.4000/mots.20534.

CIFARELLI D. M. & REGAZZINI E. (1996). De Finetti's Contribution to Probability and Statistics. *Statistical Science*, **11**(4), 253–282. Publisher : Institute of Mathematical Statistics.

CORCUFF P. (2021). *La Grande Confusion*. Textuel édition.

DEVEAUD R., SANJUAN E. & BELLOT P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, **17**(1), 61–84. DOI : 10.3166/dn.17.1.61-84.

GREENFELD L. (1992). *Nationalism : five roads to modernity*. Cambridge, Mass. : Harvard Univ. Press, 7. print édition.

LE BART C. (2003). L'analyse du discours politique : de la théorie des champs à la sociologie de la grandeur. *Mots. Les langages du politique*, (72), 97–110. DOI : 10.4000/mots.6323.

LEFEBVRE R. (2022). *Les mots des partis politiques*. Presses Universitaires du Midi.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26 : Curran Associates, Inc.

PELADEAU N. & DAVOODI E. (2018). Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction : A Lesson of History. DOI : 10.24251/HICSS.2018.078.

R DEVELOPMENT CORE TEAM (2010). *a language and environment for statistical computing : reference index*. Vienna : R Foundation for Statistical Computing. OCLC : 1120300286.

TIAN R., MAO Y. & ZHANG R. (2020). Learning VAE-LDA Models with Rounded Reparameterization Trick. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1315–1325, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.101.

VELCIN J., KIM Y. M., BRUN C., DORMAGEN J. Y., SANJUAN E., KHOUAS L., PERADOTTO A., BONNEVAY S., ROUX C., BOYADJIAN J. & OTHERS (2014). Investigating the Image of Entities in Social Media : Dataset Design and First Results. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

ZIMMER O. (2003). *Nationalism in Europe, 1890-1940*. Studies in European history. Basingstoke, Hampshire ; New York : Palgrave Macmillan.