

Croiser ADT et NLP pour caractériser les commentaires en ligne et détecter les tendances complotistes : le cas des vaccins

Pascal Marchand¹, Pierre Ratinaud²

¹Université de Toulouse – pascal.marchand@iut-tlse3.fr

²Université de Toulouse – pierre.ratinaud@univ-tlse2.fr

Abstract

We analyze the comments on articles and reports published on the *France Info* website about vaccination from 07/27/2020 to 09/30/2022. The cleaned, lemmatized, recognized and segmented corpus is subject to a descending hierarchical classification (CDH; Iramuteq software) which makes it possible to define 15 interpretable classes according to a cognitive processing model. We then apply, to the characteristic segments of each of the classes resulting from the CDH, “NLP” models, as well as lexicometric indices (richness, temperature, etc.) or morphosyntactic indexing. The results make it possible to identify hypotheses for the detection of signs of “abnormality” in large textual corpora.

Keywords: Lexical classification, Natural Language Processing, morphosyntax, weak signals.

Résumé

On analyse les commentaires des articles et reportages publiés sur le site de France Info à propos de la vaccination du 27/07/2020 au 30/09/2022. Le corpus nettoyé, lemmatisé, reconnu et segmenté fait l’objet d’une classification hiérarchique descendante (CDH ; logiciel Iramuteq) qui permet de définir 15 classes interprétables selon un modèle de traitement cognitif. On applique ensuite, aux segments caractéristiques de chacune des classes issues de la CDH, des modèles « NLP », ainsi que des indices lexicométriques (richesse, température ...) ou des indexations morphosyntaxiques. Les résultats permettent de dégager des pistes pour la détection de signes « d’anormalité » dans les grands corpus textuels.

Mots clés : Classification lexicale, *Natural Language Processing*, morphosyntaxe, signaux faibles.

1. Introduction

La détection des expressions de haine, de complotisme, de désinformation, de dérives sectaires... circulant sur *Internet* est devenue une préoccupation majeure, notamment en contexte de cybersécurité. Issus de l’intelligence économique (Ansoff, 1975), de tels « signaux faibles » ne sont pas toujours très clairement définis, du moins opérationnellement. On pense à des fréquences faibles, des effets d’implicite, des variations fines de cotexte, de faisceaux d’indices, des changements de configurations dans des systèmes complexes... inabordables par approche statistique et renvoyant parfois à des démarches plus ou moins interprétatives.

L’intelligence artificielle (IA) n’est pas exempte de promesses pour détecter de tels « signaux faibles » en traitant des données massives pour modéliser des évolutions globales et détecter des inflexions accidentelles. Opérationnellement, il s’agit, le plus souvent, de se donner les moyens d’identifier des signaux « d’anormalité » textuelle dans les discours.

Le plus souvent, on recherche des modes opératoires (marqueurs ou patterns codes) identifiés comme porteurs d’un risque potentiel. Il peut s’agir de repérer des occurrences appartenant à un lexique préalablement défini comme malveillant. On peut rechercher des formes ou

tournures inhabituelles. Mais, le plus souvent, il s'agit, à partir d'un corpus d'apprentissage, de constituer des lexiques de mots-clés renvoyant à des thématiques malveillantes pour les « projeter » sur des corpus analyser. On trouvera, par exemple, un développement des intérêts et limites de cette approche dans le projet *Hatometer* (Laurent, 2020).

Plus marginalement, on peut chercher à identifier des « signaux faibles » de menace qui dépassent les attentes lexicales que l'on peut projeter sur les données textuelles en ligne. Le rapprochement entre les traitements par textométrie et modèles de langages n'est pas nouveau et on se référera avec profit aux résultats de Vanni, Mayaffre et Longrée (2018), qui montrent que le *Deep Learning* peut identifier des structures phraséologiques ou motifs linguistiques sensibles aux occurrences et à leur organisation syntagmatique mais que l'ADT « classique » ne faisait pas émerger.

C'est dans cette voie que nous initions une analyse comparée entre une analyse textométrique par classification hiérarchique et l'ajout de ressources issues du traitement automatique du langage.

2. Corpus, traitements et résultats textométriques

On s'intéresse ici aux **commentaires** des articles et reportages extraits sur le site de France Info à propos de la vaccination du 27/07/2020 au 30/09/2022¹. Le corpus représente 135 620 textes, 5 481 450 occurrences et 53 835 formes.

2.1. Classification lexicale

Après lemmatisation et reconnaissance des locutions, le corpus fait l'objet d'une classification hiérarchique descendante (méthode de Reinert, 1983, logiciel *Iramuteq*) qui permet de définir 15 classes (figure 1).

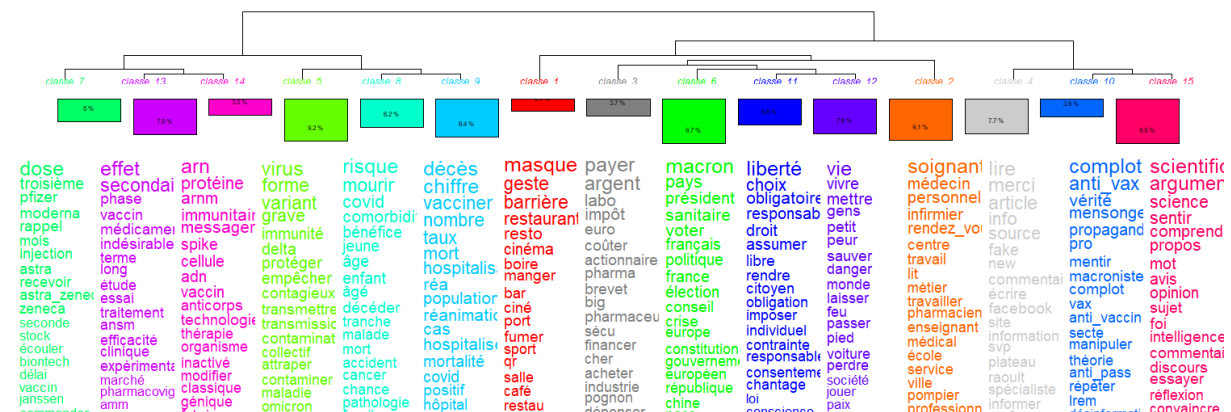


Figure 1 : Classification hiérarchique descendante du corpus "pro/anti-vax" (Reinert, 1983, Iramuteq).

Cette CDH peut être interprétée selon un modèle cognitif inspiré des travaux de Beauvois et Dubois (1991) ou de Kahneman (2012) et permettant de caractériser la valeur informationnelle *versus* sociale des commentaires. En effet, le « bloc » de gauche se caractérise par les formes *vaccin*, *virus*, *variant*, *grave*, *covid*, *forme*, *effet*, *risque*, *immunité*, *dose*, *efficace*, *maladie*, *secondaire*, *vacciner*, *delta*, *arn*, *efficacité*, *grippe*, *protéger*, *cas*, *décès*, *immunitaire*, *contamination*, *omicron*, *anticorps*, *attraper*, *contagieux*, *mort*, *arnm*..., qui renvoient à une

¹ <https://www.francetvinfo.fr/sante/maladie/coronavirus/vaccin/>.

objectivation ou une dimension descriptive, factuelle, informationnelle (autorisant le doute et la controverse). Par contraste, les formes du « bloc » de droite sont *macron*, *politique*, *gouvernement*, *sanitaire*, *pass*, *liberté*, *droit*, *français*, *commentaire*, *président*, *messenger*, *public*, *soignant*, *loi*, *médecin*, *social*, *payer*, *anti_vax*, *citoyen*, *complotisme*, *conseil*, *merci*, *penser*, *obligation*, *voter*, *demandeur*, *personnel*... et renvoient à une subjectivation ou une dimension évaluative, symbolique, sociale (impliquant la polémique et la suspicion). On ne fera ici que mentionner le fait que l'attitude *pro* vs *anti*-vaccins ne semble pas structurer cette opposition.

Plus précisément, chacune des classes peut être interprétée selon des modes de traitement ou des « biais cognitifs » mis en œuvre dans certains discours. Cela conduit à repérer deux classes potentiellement candidates à l'hébergement de discours complotistes : la classe 11, autour de la notion de *liberté*, et surtout la classe 12, autour de la *mise en danger de la vie d'autrui* (retournement de stigmatisation). Il s'agira ici de tester si des outils développant des mesures sémantiques peuvent attirer l'attention sur ces deux classes, particulièrement la classe 12.

3. Caractérisation des classes

3.1. Formes, locutions et segments spécifiques

Certaines approches considèrent le complotisme comme un type de discours et cherchent à en identifier les caractéristiques linguistiques. Ainsi, Paola Catenaccio (2022) identifie trois traits discursifs prédominants dans le discours conspirationniste : (1) une occurrence élevée de références métadiscursives à des termes tels que *rapport*, *version* et *histoire*, associés à des verbes évaluatifs (par exemple, *affirmer*) et l'adjectif *officiel* ; (2) le recours fréquent à des termes tels que *preuves* (dont l'absence caractérise les récits officiels) et *faits* (généralement posés comme étant au cœur des versions conspiratrices).

L'extraction des formes et segments caractéristiques des classes 11 et 12 permet de repérer des marqueurs possiblement candidats pour le complotisme.

Formes spécifiques :

- *Conscience*, *crime*, *criminel*, *liberticide*, *déguiser*, *(se) (faire) piquer*, *sacrifier*, *foutre*, *dictature*, *Corée (du nord)*, *politiciens*, *politicards*, *dirigeants*, *cupides*, *totalitaire*, *soumettre*, *propagande*, *scandale*, *affaire*, *naïf*, *amen*, *alléluia*...

Segments spécifiques :

- *comme par hasard*, *on vous dit*, *ne vous inquiétez pas*, *mise (mettre) en danger*, *roulette russe*, *à leur(s) fin(s)*, *sur le / a bon dos*, *gens comme vous*, *prêt à*, *mettre en place*, *(il faut) arrêter*, *jouer (à/avec)*...

A titre d'exemple, on projette les formes *heureusement*, *tant pis*, *sciemment* et *personnellement*, qui sont fortement reliées aux classes cibles (figure 2).

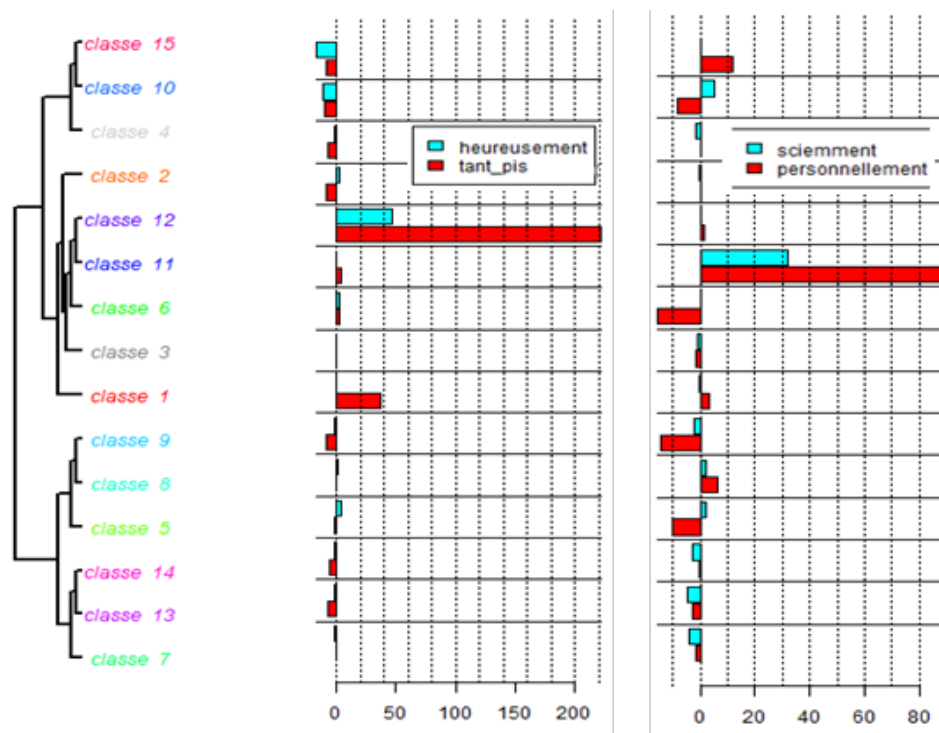


Figure 2 : Projection de formes spécifiques sur les classes de la CDH.

Un dictionnaire pourrait donc être imaginé, constitué de formes candidates au marquage complotiste sans être réductible à la thématique même du complot.

Pour identifier d'autres signes « d'anormalité » dans le corpus, on extrait les segments caractéristiques de chacune des classes issues de la CDH (qui constituent donc une métadonnée dans ce sous-corpus). On applique ensuite, à ce sous-corpus des modèles de « NLP », ainsi que des indices lexicométriques (richesse, température ...) ou des indexations morphosyntaxiques.

3.2. TAL, NLP et autres « Bert »

L'analyse des sentiments peut être utilisée pour décrire les discours complotistes, notamment sur les réseaux socionumériques. Cosgrove et Bahr (2022) ont montré, à l'aide du LIWC, que les *discussions complotistes (sur Twitter et Redit) montraient des taux significativement plus élevés de colère, d'anxiété, de pouvoir et de mort, mais également de tristesse, de famille et de religion que les discussions scientifique et générales sur les mêmes plate-formes*. Dans cet esprit, nous avons testé quatre dispositifs proposant de l'analyse de sentiment pour des textes en français. Ils ont en commun d'être disponible sous des licences libres.

3.2.1. Textblog-fr

TEXTBLOB (Loria, 2020) est une bibliothèque python d'ADT qui propose, entre autres, un module d'analyse de sentiment. Une version française de cette bibliothèque est également disponible (<https://github.com/sloria/textblob-fr>). L'analyse de sentiment repose sur un dictionnaire et sur un ensemble de règles appliquées sur ce dictionnaire. A chacun des mots de ce dictionnaire est associé un score de polarité qui varie de -1 pour les mots estimés les plus négatifs et 1 pour les mots les plus positifs. Les résultats obtenus sur chacun des segments des classes se présentent sous la forme d'un score également compris entre -1 et 1, en conservant la même logique. Nous avons calculé la moyenne de polarités des segments de chacune des classes (score *Moyenne polarity*). Cette mesure a été complétée par une estimation du

pourcentage de segments évalués plutôt négativement dans chacune des classes (score *Pol2*). La figure 3 rend compte de ces résultats.

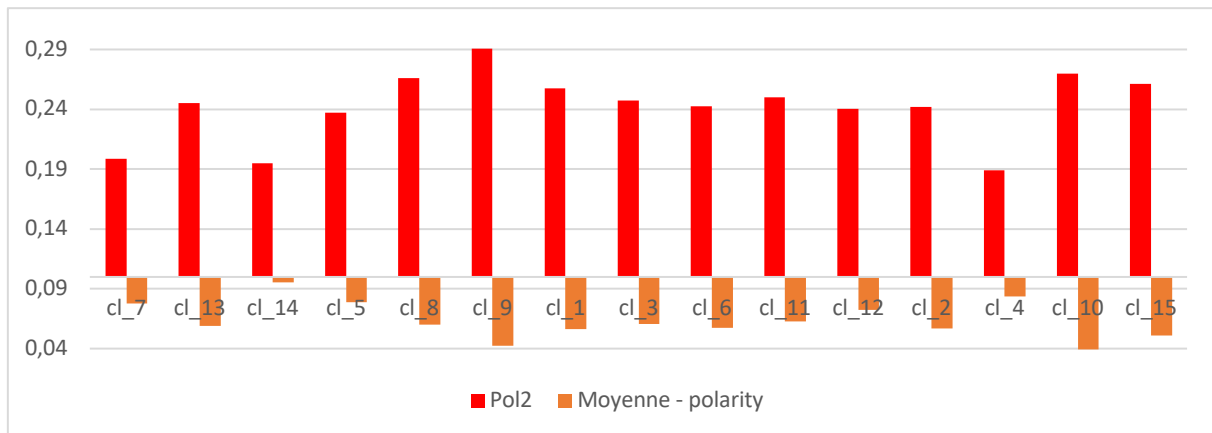


Figure 3 : TextBlob (Loria & coll., 2020).

3.2.2. VADER

Le *Valence Aware Dictionary and sEntiment Reasoner* (VADER) (Hutto & Gilbert, 2014) fonctionne sur le même principe que l'outil précédent. Un dictionnaire de mots accompagné d'un score de polarité et un ensemble de règles permettent d'estimer la valence des textes. L'auteur signale que cet outil est particulièrement adapté aux émotions exprimées sur les réseaux sociaux où la textualité est très proche de notre corpus. L'outil renvoie, pour chacun des segments, trois scores compris entre 0 et 1 : positivité, négativité et neutralité. Il fournit également un score nommé « compound » qui représente la somme des scores des mots du segment, en tenant compte des règles, et qui est normalisé entre -1 pour le très négatif et 1 pour le très positif. La figure suivante représente la moyenne de ce score pour chacune des classes. Nous utilisons ici la version française de VADER (https://github.com/thomas7lieues/vader_FR).

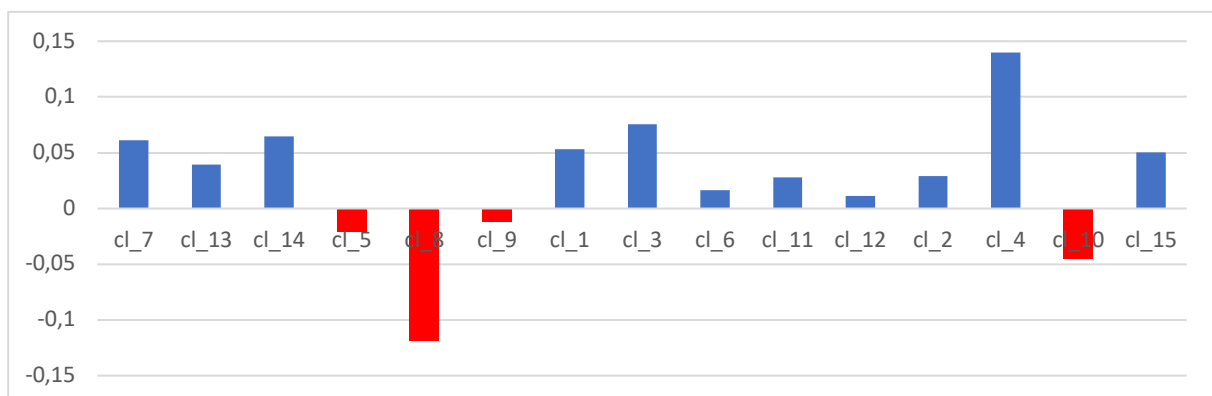


Figure 4 : Vader-Sentiment-Analysis-FR (Hutto & Gilbert, 2014).

3.2.3. BertAllo

BertAllo (Blard, 2020) est un modèle de langage (*language model*) basé sur le modèle *CamemBERT* (Martin et al. 2020) lui-même dérivé de *BERT* (Devlin et al., 2019). Tous ces modèles relèvent de l'apprentissage machine (*machine learning*). *BertAllo* a été paramétré finement (« fine tuned ») à partir d'un corpus de 200 000 évaluations de films sur le site *allocine.fr* (100 000 positives et 100 000 négatives). Un corpus d'entraînement assurément

éloigné du corpus analysé ici. Pour chaque segment, le modèle renvoie une évaluation binaire (positif ou négatif) ainsi qu'un score compris entre 0 et 1 qui reflète le niveau de confiance dans l'évaluation binaire. Nous avons utilisé ce score pour affiner la prédiction binaire. Les segments avec un score inférieur ou égal à 0,7 seront considérés comme *indéterminés* du point de vue de leur valence. Les segments avec un score supérieur, qu'ils soient jugés positivement ou négativement, seront comptabilisés. La figure 5 représente le pourcentage de segments positifs, négatifs et indéterminés pour chacune des classes de l'analyse.

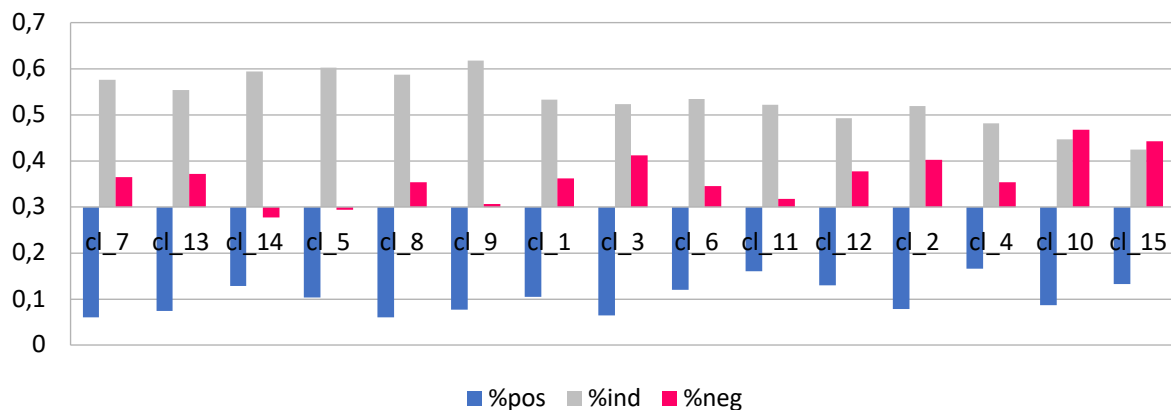


Figure 5 : Bert "allo" (Blard, 2020).

3.2.4. DistilCamemBERT-Sentiment

Comme son nom l'indique, *DistilCamemBERT-Sentiment* (Delestre et Amar, 2022) est un dérivé d'une version allégée de *CamemBERT*. Il a été paramétré à partir de la même collection que le modèle précédent, complété par un corpus de commentaires français postés sur le site *Amazon.fr*. Le modèle renvoie cinq scores compris entre 0 et 1 pour chaque segment. Ils correspondent à la probabilité que le segment soit noté 1, 2, 3, 4 ou 5 étoiles, 1 correspondant à une note négative et 5 à une note positive. La somme des scores est de 1. Dans la figure 6, nous avons regroupé les scores 1 et 2 et les scores 4 et 5.

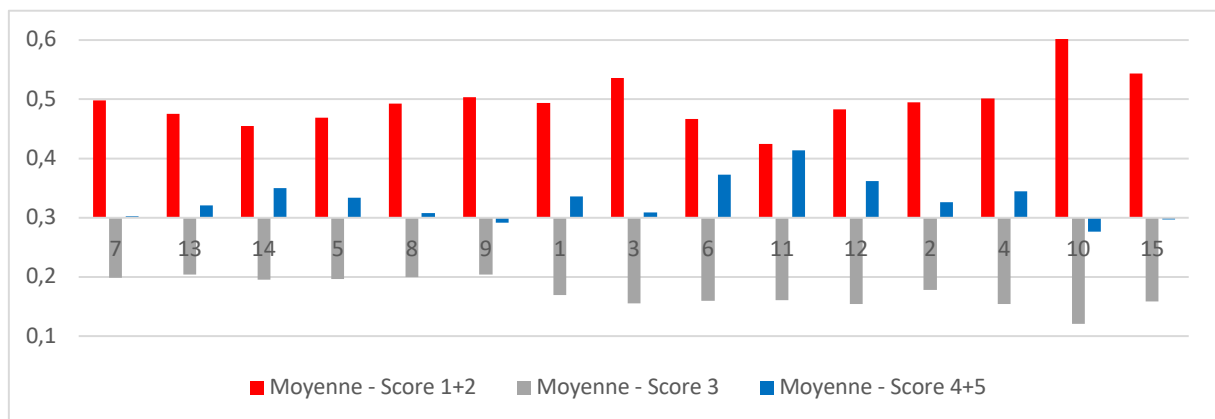


Figure 6 : DistilCamemBERT-Sentiment (Delestre & Amar, 2022).

Dans l'ensemble, les ressources mobilisées ne permettent pas d'identifier les classes (11 et surtout 12) candidates au complotisme. La plupart du temps, les segments organisés autour des termes « merci » et « liberté » attirent spontanément des valences positives, tandis que les segments organisés autour des termes « mort » et « complot » attirent des valences négatives. En revanche, les modèles « Bert » semblent différencier les classes « informationnelles » et les classes « évaluatives », ce qui n'est pas sans intérêt. Pour être adapté, un affinage des modèles

serait nécessaire pour s'adapter au contexte de l'étude. Notons également que nous appliquons ces modèles sur des textes ayant subi le prétraitement d'*Iramuteq* (suppression des majuscules, de la ponctuation, des apostrophes, des smileys, etc...) alors qu'ils ont potentiellement été entraînés sur des textes plus proche de l'original. Mais plus généralement, le recours à de tels modèles doit prendre en compte l'adéquation au contexte et notamment les stratégies de contournement, connues en modération pour éviter le *shadow ban*, depuis les raccourcis phonétiques et les modifications graphiques (*algotalk*, *algospeak* ou *algoridiomes*)² jusqu'au « langage ésopien » et aux implicites de savoirs partagés (« appel du pied » ou *dog whistle* : Henderson et McCready, 2018)³. Les dictionnaires devraient donc s'adapter en permanence aux évolutions constantes du lexique, et il faudrait accepter d'avoir systématiquement un temps de retard. Nous cherchons donc, quant à nous, à détecter des signes moins contrôlés cognitivement.

3.3. Attitudes à l'égard du langage (P_A/S_A)

Le modèle des attitudes à l'égard du langage, proposé par Beauvois et Ghiglione (1981) permet de distinguer les attitudes « paradigmatic » (centrée sur la référence, le contenu, l'objet, notée P_A) et « syntagmatic » (centrée sur la relation, la régulation, les sujets et notée S_A), essentiellement sur la base des « mots-outils », modalisateurs, connecteurs... Reprenant la méthode des types généralisés (*Tgen* : Lamalle & Salem, 2002), nous avons constitué un dictionnaire des formes habituellement corrélées avec ces deux attitudes et nous les projetons sur les classes (figure 7).

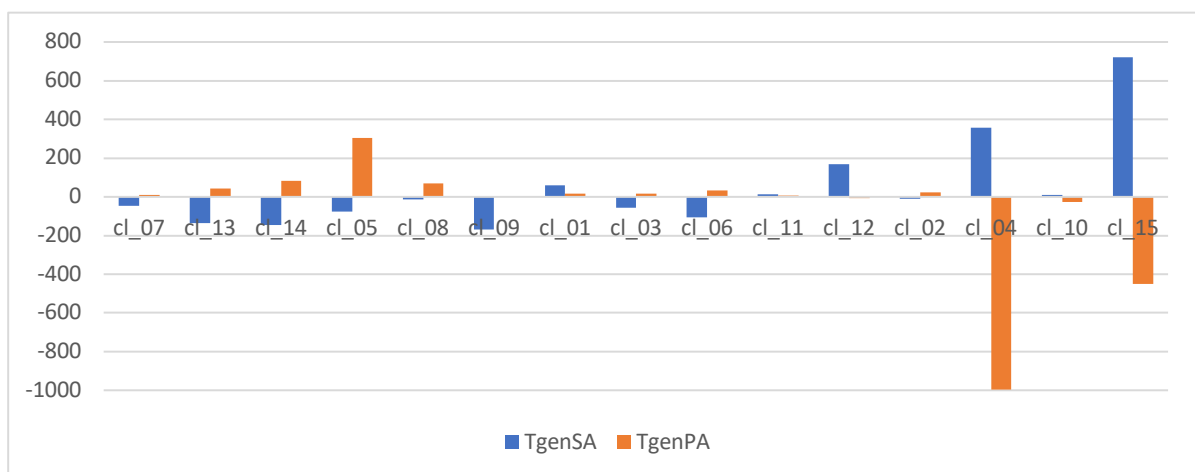


Figure 7 : Projection des $Tgen P_A$ et S_A sur les classes de la CDH.

Si les *Tgens* semblent, ici encore, différencier assez nettement les classes « informationnelles » et les classes « évaluatives », les classes ciblées ne sont pas discriminées par cet indice.

3.4. Richesse (*TTR*, *Hapax...*) et température

Bessi, Coletto, Davidescu, Scala & Caldarelli (2015) montrent, sur un échantillon de 1,2 millions d'individus sur Facebook, que les consommateurs habituels de théorie du complot sont plus concentrés et autonomes sur leur contenu spécifique. Mais Miani, Hills et Bangerter (2022) présentent des résultats plus complexes : les textes complotistes sont plus interconnectés, plus thématiquement hétérogènes, et plus similaires les uns aux autres, révélant une plus faible

² Exemples : *eske*, *koï 2 9 ? le\$bean*, *seggs*, *Ir@n*, *v10lence*...

³ A propos de vaccins, sur les réseaux sociaux anglophones, « panoramic » peut remplacer « pandemic » (ou « *Backstreet Boys Reunion Tour* ») et les groupes antivaccins peuvent être appelés « *dance party* » ou « *dinner party* », tandis que les personnes vaccinées seront des « *swimmers* » ou « *joining swim club* » (cf. Laugée, 2022).

cohésion au sein des textes mais une plus grande cohésion entre les textes et fournissant un solide soutien empirique à une vision globale du complot.

De tels indices lexicométriques ont pour intérêt d'être, du moins le supposera-t-on, moins contrôlables cognitivement. Si les noms, verbes, adjectifs... sont choisis pour leur valeur sémantique et pragmatique, l'organisation quantitative du discours est moins facile à contrôler. A titre de première exploration, nous avons repris la mesure de la richesse lexicale, relativement classique en textométrie.

Le « type/token ratio » ou le rapport hapax / formes, même pondérés par la taille des classes, ne donnent pas de résultats très probants (figure 8).

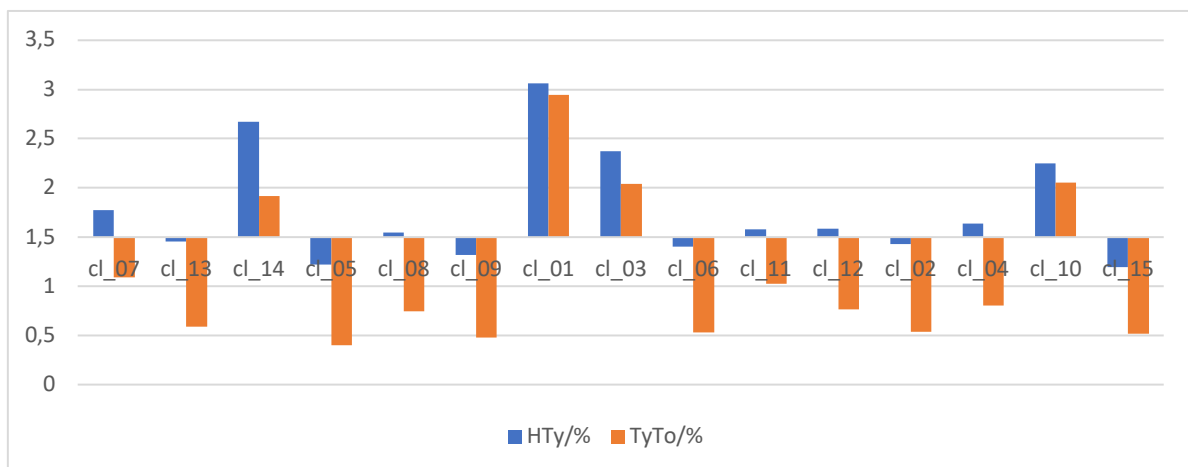


Figure 8 : Richesse (Type / token / hapax pondérés par le % classes).

Mais Jean-Pierre Deconchy (1971, 1980) avait proposé un indice de « température » considérant le rapport du log des formes et du log des occurrences. Il montrait ainsi que le discours « orthodoxe » était plus pauvre que le discours général. Ici, cet indice semble encore différencier les classes « informationnelles » et les classes « évaluatives » (figure 9). En revanche, elle n'identifie pas nos classes cibles. Cette direction devrait être approfondie.

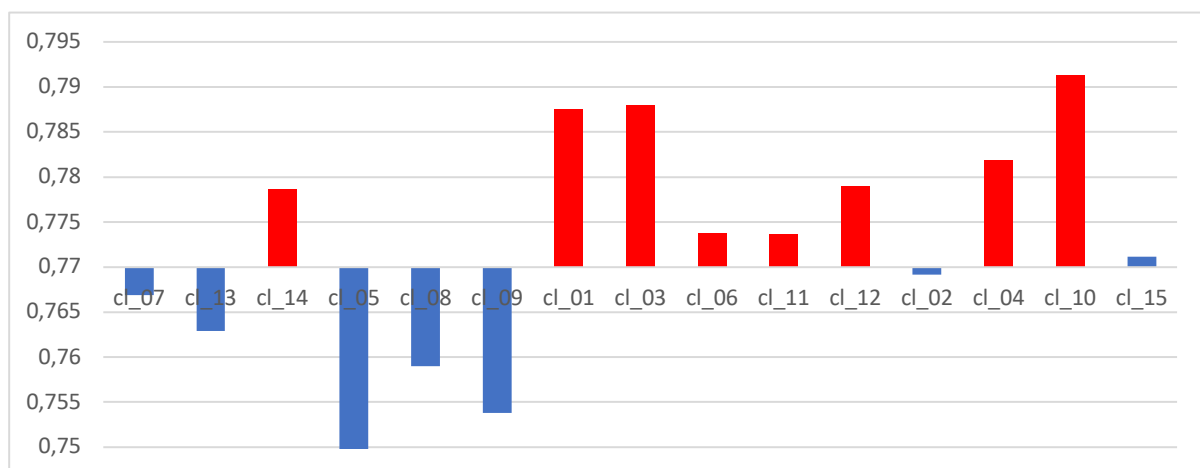


Figure 9 : « Température » (LogDico/LogTaille, Deconchy, 1991).

3.5. Types morphosyntaxiques

Catenaccio (O.C.) souligne l'importance des structures hypothétiques, signalées par l'utilisation récurrente de la modalisation « serait », qui « remettent en question les récits officiels en utilisant une forme d'argumentation réfutationnelle *modus tollens* ».

L'indexation des « types grammaticaux » lors de l'étape de lemmatisation permet d'analyser indépendamment les « mots pleins » (généralement substantifs, verbes, adjectifs...) et les marqueurs syntaxiques (modalisateurs, connecteurs, « mots-outils »...).

La projection des marqueurs syntaxiques sur les classes (non-représentée ici) montre des spécificités qui confirment nos précédentes analyses : le discours « informationnel » est plus nominal et descriptif, tandis que le discours « évaluatif » est plus verbal et modalisé. Une analyse des correspondances du tableau « marqueurs syntaxiques » X classes est plus intéressante (figure 10).

4. Perspectives

Ces premiers résultats permettent de dégager des pistes pour la détection de signes « d'anormalité » dans les grands corpus textuels. On montre notamment que les modèles « NLP » (supervisés), basés sur des mots-clés, sont contextuels et peu généralisables. La convergence distributionnelle entre les formes et les types morphosyntaxiques incite à dépasser les formes contextuelles (qui peuvent être reprises) et à privilégier les formes et locutions génériques ou patterns de types (moins contrôlés). A un niveau moindre encore de contrôle cognitif, les indices lexicométriques plus basiques (et non-contrôlés), comme la richesse et la « température », doivent être approfondis.

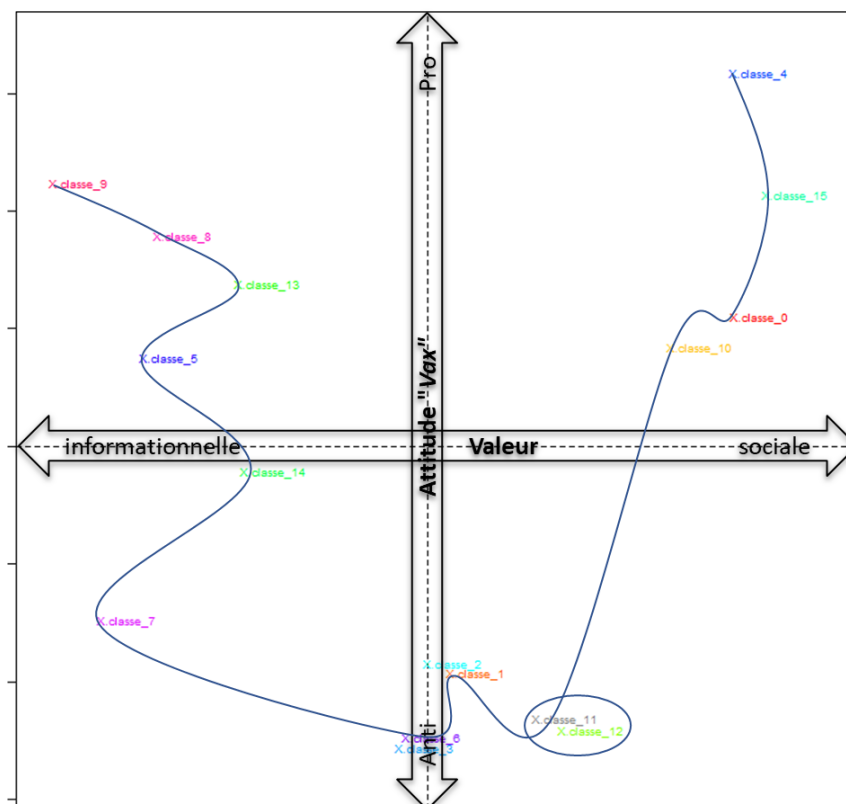


Figure 10 : AFC des types grammaticaux.

Bibliographie

- Ansoff H. I. (1975). Managing strategic surprise by response to weak signals. *California management review*, 18 (2), 21-33.
- Beauvois J.-L. et Dubois N. (1991). Internal/external orientations and psychological information processing. *CPC: European Bulletin of Cognitive Psychology*, 11, 193-212.

- Beauvois J.-L. et Ghiglione R. (1981), *L'homme et son langage - Attitudes et enjeux sociaux*. Paris : PUF.
- Bessi A., Coletto M., Davidescu G. A., Scala A., Caldarelli G. et Quattrocioni W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PLoS ONE*, 10 (2), e0118093. doi:10.1371/journal.pone.0118093.
- Blard T. (2020). French sentiment analysis with BERT. GitHub repository : <https://github.com/TheophileBlard/french-sentiment-analysis-with-bert>.
- Catenaccio P. (2022). A corpus-driven exploration of conspiracy theorising as a discourse type. Lexical indicators of argumentative patterning. In Demata M., Zorzi V. et Zottola A. (Éds.), *Conspiracy Theory Discourses*, Discourse Approaches to Politics, Society and Culture, 98. John Benjamins, 25-48.
- Cosgrove T. et Bahr M. (2022). Psycholinguistics of Conspiracy Theories: Negative Emotions and Themes Facilitate Diffusion Online. *SSRN Electronic Journal*. <http://dx.doi.org/10.2139/ssrn.4158239>.
- Deconchy J.-P. (1971). *L'orthodoxie religieuse. Essai de logique psycho-sociale*. Paris : Éditions ouvrières.
- Deconchy J.-P. (1980). *Orthodoxie religieuse et sciences humaines*. Paris-La Haye : Mouton.
- Delestre C. et Abibatou A. (2022). DistilCamemBERT : une distillation du modèle français CamemBERT, *Conférence sur l'apprentissage automatique (Cap)*. Vannes. hal-03674695.
- Devlin J., Chang M., Lee K. et Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein J., Doran C. et Solorio T. (Éds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1. Minneapolis : Association for Computational Linguistics, 4171-7186.
- Henderson R. et McCready E. (2018). *How Dogwhistles Work*. In Mineshima K. et al. (Éds.), *New Frontiers in Artificial Intelligence*, Lecture Notes in Computer Science, 10838. Springer, 231-240.
- Hutto C. J. et Gilbert E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 8 (1) . Michigan, 216-225.
- Kahneman D. (2012). *Système 1 / Système 2 : Les deux vitesses de pensée*. Paris : Flammarion.
- Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In Morin A. et Sébillot P. (Éds.), *Actes des 6e Journées d'Analyse des Données Textuelles (JADT 2002) : volume 2*. St Malo : Irisa et Inria, 529-538.
- Laugée F. (2022). Alogospeak. *La Revue Européenne des Médias et du Numérique*. En ligne : <https://la-rem.eu/2022/07/alospeak/>.
- Laurent M. (2020). Project Hatemeter: helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media. *Procedia Computer Science*, 176, 2143-2153.
- Loria S. (2018). textblob Documentation. *Release 0.15*, 2. <https://textblob.readthedocs.io/en/dev>.
- Martin L., Muller B., Suarez P.J.O. et al. (2019) CamemBERT: A Tasty French Language Model. In Jurafsky D., Chai J., Schluter N. et Tetreault J. (Éds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7203-7219. *arXiv:1911.034894* / <https://doi.org/10.18653/v1/2020.acl-main.645>.
- Miani A., Hills T. et Bangerter A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, 8 (43), en ligne. <https://www.science.org/doi/10.1126/sciadv.abq3668>.
- Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8 (2), 187-198.
- Reinert M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologique*, 26, 24-54.
- Vanni L., Mayaffre D. et Longree D. (2018). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. In *Proceedings of the 14th International Conference on Statistical Analysis of textual data (JADT 2018)*. Rome, 459-466. En ligne : <http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf>.