A Linguistic Contribution to an Automatic Classification of Communities and their Analysis

Dalia Saigh, Boris Borzic, Abdulhafiz Alkhouli, Julien Longhi

questions de communication

Questions de communication

31 | 2017
Humanités numériques, corpus et sens

Contribution linguistique à une classification automatique des communautés de sens et à leur analyse

La controverse sur le statut des intermittents du spectacle

A Linguistic Contribution to an Automatic Classification of Communities and their Analysis

Dalia Saigh, Boris Borzic, Abdulhafiz Alkhouli et Julien Longhi

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/questionsdecommunication/11097
DOI : 10.4000/questionsdecommunication.11097
ISSN : 2259-8901

Éditeur
Presses universitaires de Lorraine

Édition imprimée
Date de publication : 1 septembre 2017
Pagination : 161-182
ISBN : 9782814303256
ISSN : 1633-5961

Référence électronique
Dalia Saigh, Boris Borzic, Abdulhafiz Alkhouli et Julien Longhi, « Contribution linguistique à une classification automatique des communautés de sens et à leur analyse », *Questions de communication* [En ligne], 31 | 2017, mis en ligne le 01 septembre 2019, consulté le 15 avril 2024. URL : http://journals.openedition.org/questionsdecommunication/11097 ; DOI : https://doi.org/10.4000/questionsdecommunication.11097

questions de communication, 2017, 31, 161-182

> DOSSIER

DALIA SAIGH
Agora
Université de Cergy-Pontoise
F-95011
daliasaigh1988@gmail.com

BORIS BORZIC
Équipes Traitement de l'information
et systèmes
Université de Cergy-Pontoise
Centre national de la recherche scientifique
École nationale supérieure
de l'électronique et de ses applications
F-95000
boris.borzic@ensea.fr

ABDULHAFIZ ALKHOULI
Équipes Traitement de l'information et
systèmes
Université de Cergy-Pontoise
Centre national de la recherche scientifique
École nationale supérieure
de l'électronique et de ses applications
F-95000
abdulhafiz.alkhouli@ensea.fr

JULIEN LONGHI
Agora
Université de Cergy-Pontoise
F-95011
julien.longhi@u-cergy.fr

CONTRIBUTION LINGUISTIQUE À UNE CLASSIFICATION AUTOMATIQUE DES COMMUNAUTÉS DE SENS ET À LEUR ANALYSE

LA CONTROVERSE SUR LE STATUT DES INTERMITTENTS DU SPECTACLE

Résumé. — Le 22 mars 2014, un accord sur l'indemnisation du chômage des intermittents du spectacle a été signé par des partenaires sociaux. Ce texte a suscité des inquiétudes et oppositions parmi les intermittents, des mouvements d'occupation de plusieurs places ou théâtres, à Paris et en région, qui ont duré plusieurs jours. Ces réactions ont rapidement envahi les réseaux socionumériques, Twitter en particulier. Des millions de *tweets* ont été échangés à partir de la diffusion des premières informations. L'objectif de ce travail est d'attester, par le croisement de perspectives linguistiques et informatiques, de la pertinence d'une recherche fondée sur l'analyse des *tweets* pour rendre compte d'enjeux sociaux et politiques. Il s'agira alors de croiser des méthodes issues de sciences humaines, sociales, techniques, de la modélisation, etc., afin de contribuer à l'appréhension de ce sujet. Ainsi montrerons-nous la nécessaire complémentarité de tous ces niveaux dans la prise en compte des humanités numériques, dont le traitement se veut résolument complexe.

Mots clés. — controverse, *tweets*, textométrie, classification automatique, communautés

Le 22 mars 2014, un nouvel accord sur l'indemnisation du chômage des intermittents du spectacle a été signé par des partenaires sociaux Ce texte a suscité des inquiétudes et oppositions parmi les intermittents, des mouvements d'occupation de plusieurs places ou théâtres, à Paris et en région, qui ont duré plusieurs jours. Ces réactions ont rapidement envahi les réseaux socionumériques, Twitter en particulier. Des millions de *tweets* ont été échangés à partir de la diffusion des premières informations. L'objectif de notre travail est de montrer, par la complémentarité d'analyses linguistiques et informatiques, la pertinence d'une recherche fondée sur l'analyse des *tweets* pour rendre compte d'enjeux sociaux et politiques. Ce travail exploratoire est présenté ici sous la forme d'un « *data paper* »[1] dont l'objectif est de socialiser les premiers résultats d'une collaboration pluridisciplinaire qui combine plusieurs méthodologies et préoccupations et essaie de les intégrer dans une même analyse, sans se contenter de les ajouter ou de les superposer.

Constitution du corpus et méthodologie d'analyse

La constitution du corpus « #intermittent » vise à permettre de travailler sur ce genre de discours (des *tweets* relatifs à un sujet controversé), de le caractériser et de l'appréhender sous différentes formes afin de prolonger des travaux antérieurs à propos des intermittents du spectacle sur la période 2003-2004 (par exemple, voir Longhi, 2006). La méthodologie employée est importante car, comme le souligne Martin Grandjean (2016 : 2), « quand il est clair que le contenu des *tweets* d'un utilisateur n'est pas toujours révélateur de son domaine de spécialisation – en raison du bruit produit par les nombreux messages personnels, les blagues, la politique, etc. – nous devons nous tourner vers un réseau dont la structure semble plus facilement analysable en termes de "communauté" »[2]. Aussi, la récupération des *tweets* a suivi le processus suivant :

– en 2014, récupération de 13 074 *tweets* contenant le mot-dièse *#intermittent(s)* publiés par 4 617 comptes Twitter ;

– en 2015, adoption d'un seuil d'au moins 10 *tweets* avec le *hashtag #intermittent(s)* : identification de 215 comptes ayant produit au moins 10 *tweets* explicitement référencés comme appartenant à la thématique ;

– en récupérant tous les *tweets* de ces 215 personnes, nous avons obtenu 586 239 *tweets* dont 10 876 contenant le mot-dièse *#intermittent(s)* : le corpus #intermittent est composé de ces 10 876 *tweets*[3].

[1] Il s'agit d'une publication présentant le jeux de données (corpus), l'organisation et les premières analyses croisées du point de vue linguistique et informatique.

[2] Sauf mention contraire, nous traduisons les citations extraites d'œuvres non francophones : « *When it becomes clear that the content of a user's tweets is not always indicative of his field of specialisation—due to the noise produced by the many personal messages, jokes, politics, etc.—we need to turn to a network whose structure seems more readily analysable in terms of "community"* ».

[3] La finalisation du corpus #intermittent a été possible grâce au soutien financier de l'équipement d'excellence Outils et ressources pour un traitement optimisé de la langue (Ortolang).

Cette méthodologie s'inspire de travaux d'historiens utilisant les réseaux sociaux, notamment Twitter, pour analyser la représentation d'événements historiques, comme Frédéric Clavert (2016) l'a fait à propos du *hashtag #ww1* dans le cadre du Centenaire de la Grande Guerre sur Twitter.

Frédéric Clavert, Benoît Majerus et Nicolas Beaupré (2015) montrent aussi que l'utilisation de l'interface de programmation applicative (API) de Twitter, bien qu'elle soit d'une aide précieuse pour accéder aux représentations d'un événement sur ce réseau social, n'est pas sans poser question :

> « Le fait que l'API de Twitter, bien que parfois très instable, soit très pratique à utiliser, est l'un des critères de ce choix. Mais est-ce vraiment pertinent en termes de recherche ? Ne devrions-nous pas avoir des sources plus larges ? Comment extrapoler les résultats du projet à d'autres réseaux sociaux en ligne ? Enfin, la difficulté d'anticiper les *hashtags* à collecter pourrait introduire des biais dans notre recherche »[4].

Pour cette raison, cette collecte n'a pas seulement été entreposée dans une base de données, mais a été suivie d'un processus de constitution de corpus : nous avons opéré une sélection des données et métadonnées. À cette fin, l'unité de recherche Équipes Traitement de l'information et systèmes (Etis) a effectué un processus de sélection en trois étapes :

– utilisation de l'API de Twitter : appel d'une dizaine de ses fonctions concernant le descriptif détaillé des *tweets*, des *twittos*, des listes thématiques, des relations entre *twittos* (*followers*, *following*) ;

– double enrichissement d'une base de données locale en natif (JSON) et avec un design de base propre (une dizaine de tables et une cinquantaine de champs spécifiés pour préparer nos calculs de façon optimale) ;

– export sur mesure, avec les informations stockées, dans les format de données suivants : JSON, TEI, Iramuteq, etc.

Ensuite, l'ensemble des *tweets* a été mis en forme selon le schéma de description de données recommandé par la *Text Encoding Initiative* et la base de données est structurée dans un fichier XML pour devenir un corpus, afin de répondre aux enjeux institutionnels du projet CoMeRe[5] et de permettre d'en réaliser l'analyse du discours outillée. Ce processus de mise en forme permet de valoriser la dimension sémiotique et symbolique des données, et privilégie le corpus plutôt que la base de données, par le choix des balises TEI qui indiquent que tel élément linguistique est une mention, une adresse, un *hashtag*, etc.

---

[4] « *The fact that Twitter API, though sometimes very unstable, is very convenient to use is one of the criteria of this choice. But is it really pertinent in terms of research? Shouldn't we have broader sources? How to extrapolate the project's results to other online social networks? Last but not least, the difficulty to anticipate the hashtags to be collected might introduce biases in our research* ».

[5] Le projet « Communication médiée par les réseaux » (CoMeRe) a reçu l'appui financier et scientifique du consortium national Corpus-écrits et de l'équipement d'excellence Ortolang.

L'analyse textométrique du corpus #intermittent : première approche lexicale

Une première manière d'aborder ce corpus est de recourir aux méthodes statistiques développées par l'analyse du discours et la linguistique de corpus. La textométrie « (ou statistique textuelle, lexicométrie, logométrie) propose une approche instrumentée des corpus, articulant synthèses quantitatives et analyses à même le texte » (Lebart, Salem, 1994, cité par Pincemin, 2012 : en ligne). Celle-ci met en œuvre des principes différentiels, ce qui permet de mettre en évidence les similitudes et différences observées dans le corpus. En plus de fournir des procédures de tri et de calculs statistiques pour l'étude de corpus numériques, elle « établit une modélisation contextuelle et contrastive : le texte est caractérisé par ses mots par rapport à leur usage dans le corpus, le mot est caractérisé par ses cooccurrents, etc. » (Pincemin, 2012 : en ligne). La textométrie se présente comme particulièrement pertinente pour l'exploitation des corpus en sciences humaines et sociales (SHS). Elle permet une observation à la fois fine et globale des textes, tout en restant proche de la matérialité de ces derniers par le recours à des instruments d'accès aux données, et en mettant en valeur la réalité langagière qui est un terrain d'observation important pour les SHS[6].

Pour commencer cette analyse, nous procédons avec le logiciel *Iramuteq*[7] qui offre une multitude de possibilités de traitements pour la description et l'analyse de corpus textuels. L'une de ses principales méthodes est Alceste, qui permet de segmenter un corpus en « unités de contexte », d'effectuer des comparaisons et groupements au sein du corpus segmenté selon les lexèmes qu'il contient, puis de rechercher « des distributions stables » (Reinert, 1998). En plus de la méthode Alceste, *Iramuteq* fournit d'autres types d'analyse comme l'analyse prototypique, l'analyse des similitudes et celle des nuages de mots. En conséquence, cet outil est très utile « lorsque l'on souhaite cartographier la dynamique du discours des différents sujets engagés dans une interaction » (Reinert, 1999 : en ligne).

Pour la mise en forme de notre corpus, nous avons choisi un formatage avec trois variables représentatives (c'est-à-dire trois éléments indiqués en métadonnées) : la première, « intermittent », renvoie à notre thème et constitue donc le mot clé de ce corpus ; la deuxième est relative au nom d'utilisateur qui changera donc d'un *tweet* à un autre ; la troisième permet de comptabiliser le nombre de *tweets* et *retweets* relatifs à la publication de ce dernier.

La capture d'écran ci-dessous illustre le formatage du corpus #intermittent.

---

[6] Voir le site du projet *Textométrie*. Accès : http://textometrie.ens-lyon.fr/spip.php?article69.

[7] Développé par Pierre Ratinaud, le logiciel *Iramuteq* a notamment été popularisé grâce aux travaux de P. Marchand, seul ou en collaboration avec le développeur du logiciel. Voir aussi le site de ce dernier. Accès : http://www.iramuteq.org/.

Figure 1. Formatage du corpus #intermittent

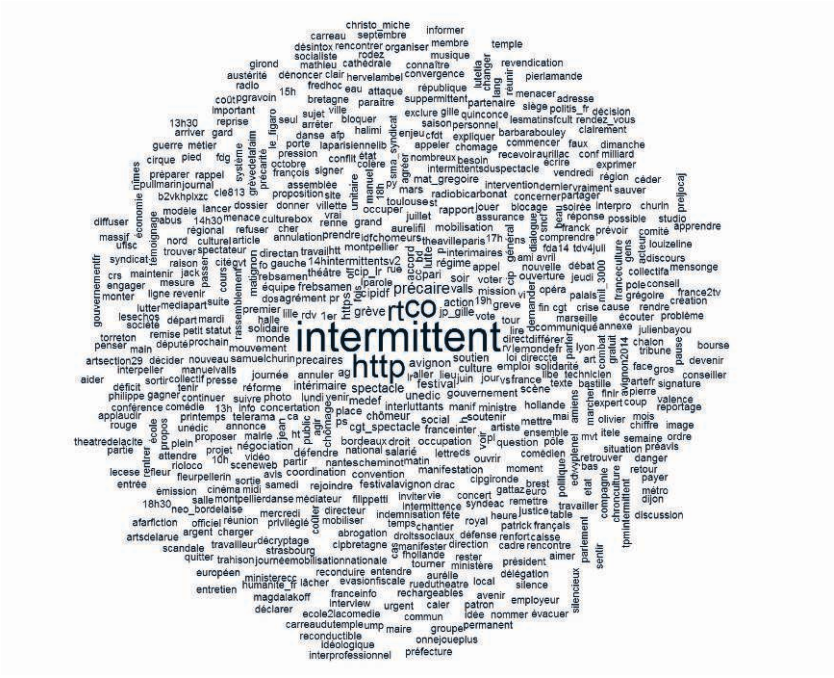
```
**** *#DirectAN *tweet_1
RT @jp_gille : #intermittents : Revue de presse du rapport intermittence du spectacle du 07 janvier #Tours #DirectAN @socialistesAN http://t…
**** *#DirectAN *tweet_2
RT @jp_gille : Demain sortie de "soumission" de M. Houellebecq et du rapport sur les #intermittents qui n'est pas le fruit d une "sous-miss…
**** *AFAR *tweet_1
#Intermittents : @aurelifil Filippetti obtient la révision de l'accord d'assurance-chômage http://t.co/uLE1IrZe3cvia @Le_Figaro
**** *AFAR *tweet_2
#Intermittents : "Ma mission n'est pas de renégocier" http://t.co/L0ybaOr8RPvia @LeNouvelObs @jp_gille
**** *AFAR *tweet_3
#Intermittents : Hollande promet de « défendre toujours la culture » http://t.co/JcUnzTLmoVvia @lemondefr
**** *AFAR *tweet_4
#Intermittents : Le gouvernement joue la montre http://t.co/HW04WOVXIYvia @humanite_fr
```

Une fois ce formatage réalisé, plusieurs analyses peuvent être menées.

Nuage de mots et analyse des similitudes

Le logiciel *lramuteq* contient une fonction réalisant une visualisation d'un document du point de vue lexical : les formes linguistiques sont représentées avec des tailles variables selon leur importance. Concrètement, plus une forme est présente dans un corpus, plus elle apparaîtra en gros dans le nuage de mots. Ceci permet de mieux mettre en évidence les mots clés utilisés par les *twittos*[8].

Figure 2. Nuage de mots du corpus #intermittent



[8] Le terme *twitto* (au pluriel, *twittos*) désigne un utilisateur de Twitter.

Ce nuage de mots met en évidence les occurrences les plus fréquentes dans les *tweets* qui se positionnent au centre du nuage. Il y a donc l'occurrence *intermittent* qui est le mot clé de notre corpus, suivi de marqueurs spécifiques tel *co* et *http* qui renvoient à des liens partagés sur Twitter, ces derniers étant automatiquement abrégés en *http://co* afin de permettre le partage de longues URL dans un *tweet* sans dépasser le nombre de caractères maximum autorisé. On trouve également le signe *rt* qui signifie « *retweet* » ; celui-ci consiste à reposter le *tweet* d'un compte, permettant ainsi aux usagers de le partager rapidement avec tous leurs abonnés.

Autour de ces formes, s'en ajoutent d'autres qui ont plus au moins la même fréquence d'où l'égalité de taille que l'on observe : parmi elles, certaines renvoient au champs lexical de la république et du gouvernement français tels *Manuel Valls*, *république*, *député*, *français*, *F.Hollande*, *Filippetti*, *ministre*, etc. ; d'autres évoquent des mouvements ou actions tels *accord*, *grève*, *mobilisation*, *manifestation*, *convention*, *combat* ; enfin, des noms ou adjectifs renvoient aux intermittents et décrivent leur situation comme *chômeurs*, *précaires*, *interluttants*, *comédiens*…

Cette description reste très générale. Pour préciser l'analyse, nous avons recours à une autre fonction que propose *lramuteq* qui permet une autre représentation graphique des formes présentes dans un corpus : l'analyse de similitude, qui conserve l'idée de taille proportionnelle à la fréquence, mais introduit les relations de cooccurrences entre les formes.

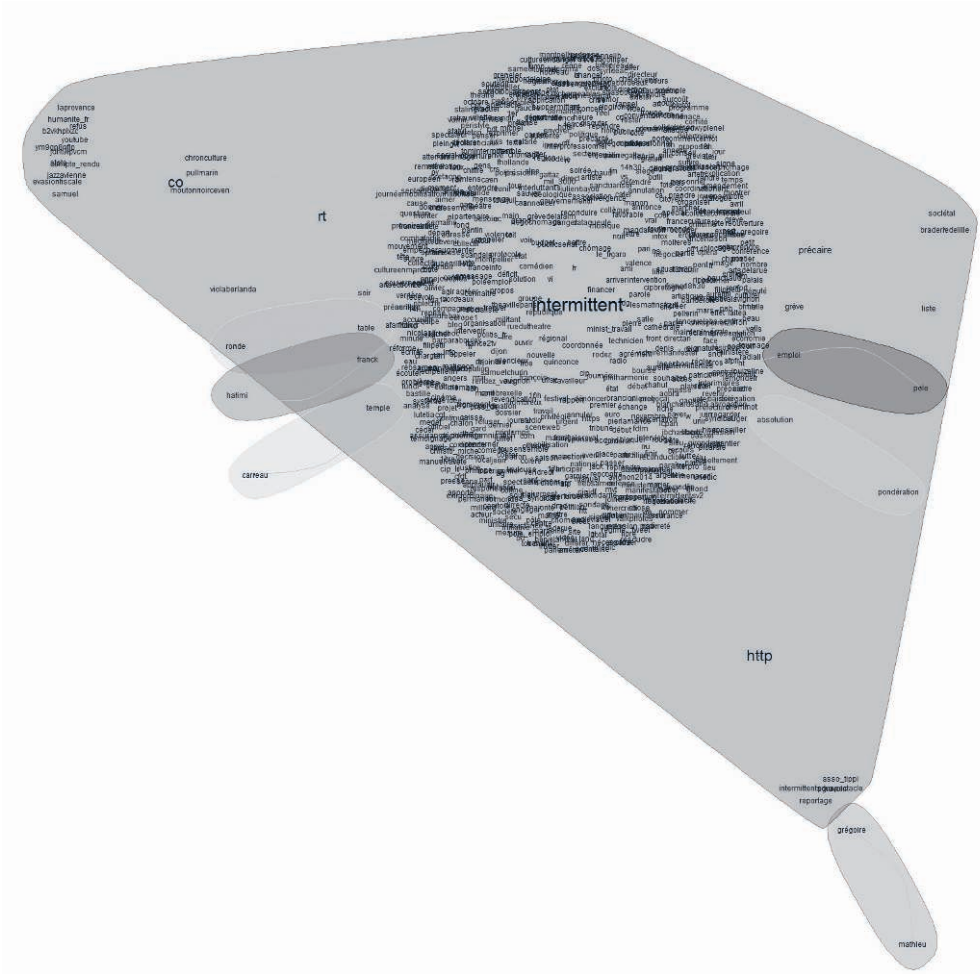
L'analyse des similitudes est une technique fondée sur la théorie des graphes (Flament, 1962, 1981). Elle présente graphiquement la structure d'un corpus en distinguant les parties communes et les spécificités des variables codées, ce qui permet de mettre en avant la relation entre les différentes formes dans les segments de texte (Marchand, Ratinaud, 2012). La représentation obtenue avec notre corpus. est reproduite ci-contre (figure 3).

Le discours est très homogène avec une seule forme centrale autour de laquelle gravite la plus grande partie du lexique du corpus. Cette figure montre un seul *cluster* principal, avec quelques autres, de tailles très réduites et qui sont peu significatifs. Ce *cluster* est constitué d'un nuage de mots qui renferme en son centre le mot clé *intermittent*, autour duquel se regroupe un lexique très dense et très lié (ce qui n'est pas étonnant puisque le corpus a été constitué autour de cette forme précédée du #).

Il y a tout de même quelques petits groupes présents dans le *cluster* principal, liés directement (avec des arêtes) à celui qui est le plus important et ayant comme terme principal *intermittent*. Parmi eux : le *cluster htpp* dans lequel nous retrouvons le terme *intermittentdespectacle* et, un peu plus loin, un tout petit *cluster* contenant le nom *GrégoireMathieu* qui est un sociologue ayant écrit le livre *Les Intermittents du spectacle. Enjeux d'un siècle de luttes* (2013). Certains *tweets* renvoient donc les usagers à des pages web où l'on cite le nom du sociologue. On voit aussi le sigle *rt* qui regroupe les termes *chronculture*, *pullmarin*, *dinamopress*, *angelin*, etc., qui renvoient donc aux comptes qui ont le plus *retweeté*. Puis, il y a *co* qui est la forme abrégée des liens sur Twitter.

Les résultats affichés sur cette figure ont permis de constater que le corpus #intermittent contient énormément de liens, de *retweets* relatifs aux intermittents du spectacle, ce qui décrit leurs différentes actions et leur situation (exemple du *cluster précaire*). Cela dit, du fait de la densité du lexique qui se rattache au mot clé *intermittent*, la fonction « analyse de similitude » nous a seulement aidé à décrire la nature et l'objet des *tweets* (*tweets* avec des liens, *retweets*). Pour préciser davantage la structure du corpus, nous utilisons la fonction « classification hiérarchique descendante » (méthode Alceste développée par Max Reinert).

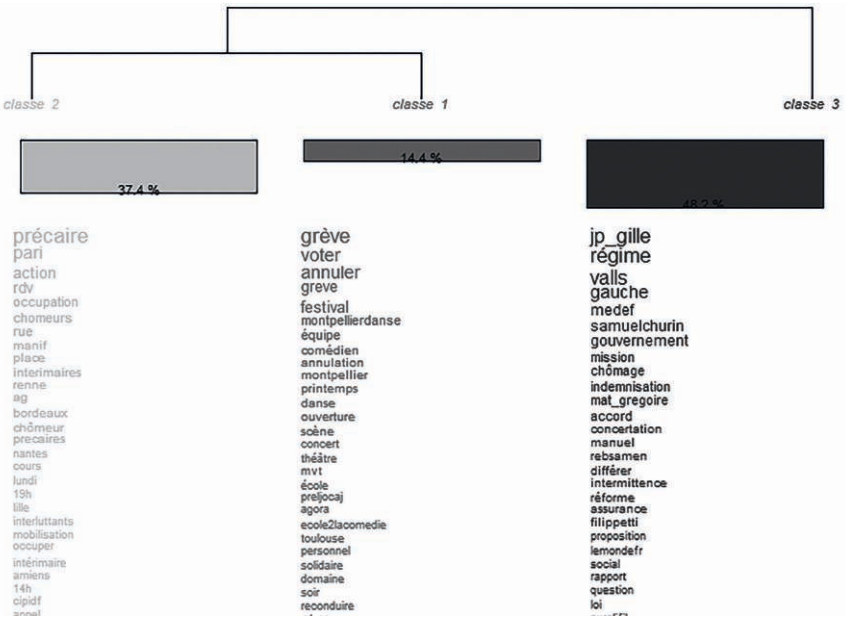
Figure 3. L'analyse des similitudes



Extraire des classes thématiques

Cette méthode permet d'extraire des classes de sens, constituées par les mots et phrases les plus significatifs : les classes obtenues permettent de dégager les thèmes dominants du corpus, avec une présentation sous forme d'un dendrogramme.

Figure 4. Résultats de l'analyse avec la méthode Alceste



Deux ensembles se distinguent dans cette figure, le premier a deux classes associées (classe 1 et classe 2), et le dernier où l'on trouve une seule classe (classe 3).

La classe 1 regroupe des formes associées aux différents mouvements de mobilisation des intermittents comme l'occupation des rues, des théâtres ou d'autres places, des manifestations. Voici un extrait de segments caractéristiques (avec un score élevé[9]) qui contiennent les mots les plus fréquents associés à la classe 1 comme *manif*, *cipdfjournée*, *action* (ces mots sont mis en valeur par l'italique).

Figure 5. Segments caractéristiques de la classe 1

**** *intermittent *CQFjournal *tweet10
score :1458.88
rt *cipidf journée* d *action paris 10h république 14h manif ministère* du *travail* 127 *rue* de *grenelle intermittents précaires*

**** *intermittent *CIP_IDF *tweet782
score : 1431.31
Rvd paris journée d *actions coordonées 11h* devant *bourse* du *travail* 3 *rue* du *château* d *eau* intermittents *précaires*

[9] Pour calculer le score absolu, les segments de texte sont classés en fonction de l'association statistique forte d'un segment à une classe lexicale (test statistique du chi2).

La classe 2 fait plutôt référence aux grèves et annulations de spectacles et concerts par les intermittents. Elle contient des mots comme *grève*, *festival*, *annulé*… Les exemples suivants montrent les segments caractéristiques de cette classe.

Figure 6. Segments caractéristiques de la classe 2

****** *intermittent *CIP_LR* tweet155**
score :2058.76
intermittents rencontres photos arles la *grève* a été *votée* pour *lundi* 7 *juillet jour* de l *ouverture* du *festival* le vernissage *annulé*

****** *intermittent *cie813* tweet48**
score :1877.02
second *soir* de *grève* et d *annulations* au *printemps* des *comédiens* à *montpellier* opéra occupé *représentation traviata annulée* intermittents

La classe 3 est relative aux *tweets* renvoyant au régime d'assurance chômage des intermittents et aux entités politiques impliquées dans la question Voici les segments caractéristiques de cette classe.

Figure 7. Segments caractéristiques de la classe 3

****** *intermittent *AFARfiction *tweet42**
score : 973.15
rt *jp_gille* intermittents je viens de *remettre* mon *rapport* à *manuel valls premier ministre* avec *aurelifil* et *frebsamen*

**** *intermittent *laparisiennelib *tweet39
score : 732.63
intermittents donc *valls* vient de *confier* une *mission* de *propositions* à *jp_gille* qui *tweetait* il y a 2 jours

Cette analyse permet de voir comment les *twittos* ont réagi lors de l'annonce du nouveau régime concernant l'assurance chômage des intermittents. Grâce à l'analyse des similitudes, nous avons constaté un grand nombre de liens pointant vers ce sujet avec des références telles celles au sociologue Mathieu Grégoire ou aux publications de journaux comme *Le Monde*. La méthode de Max Reinert indique que le discours est subdivisé en deux grands thèmes : d'une part, des *tweets* décrivent la précarité des intermittents et leur différents mouvement de contestation contre ce nouveau régime ; d'autre part, des *tweets* dénoncent l'accord, avec des liens renseignant sur cet acte et des citations de différentes personnalités politiques impliquées dans la polémique. Il y a donc une valeur argumentative importante dans ce corpus, liée aux positions énonciatives des acteurs du débat.

Classification automatique par institutions de sens[10]

Dans cette partie, nous commencerons par présenter les résultats d'un projet de typologie de la *twittosphère* francophone, puis expliciterons comment nous avons caractérisé chaque *twitto* (journaliste, artiste, politicien, citoyen…) au sein de notre corpus #intermittent, en les projetant sur notre référentiel « domaine d'activités » issu de différentes sphères d'influence, puis interpréterons ces résultats en concluant sur la relation avec les classes *Iramuteq* préalablement calculées. Ce travail permet une première approche d'une identification et d'une analyse d'institutions et de communautés de sens sur Twitter. En effet, l'un des enjeux IHM (Interactions homme-machine) autour du filtrage automatique de l'information sur les réseaux sociaux est d'éviter l'enferment des usagers dans un filtre de bulles. Développé par Eli Pariser, le concept *filter bubble* désigne l'état dans lequel se trouve un usager lorsque les informations auxquelles il accède sur l'internet sont le résultat d'une personnalisation mise en place à son insu. Une piste pourrait être d'enrichir le réseau affinitaire d'un utilisateur par des informations pertinentes provenant de *twittos* non abonnés mais sélectionnés par une proximité sémantique significative.

Pour répondre à cette double exigence, nous avons créé un dispositif expérimental d'analyse de métriques autour de la circulation de l'information et l'interaction des usagers sur Twitter afin de délimiter la frontière entre différentes institutions de sens comme la politique, l'espace médiatique, la filière artistique et le mouvement citoyen autour de thématiques polémiques, tout en prenant en considération l'intersection avec des communautés de sens.

Issue de la théorie des graphes, l'analyse de liens implicites repose sur des indices de centralité et de densité calculés à partir des relations affinitaires et interactionnelles des utilisateurs afin de détecter les sous-communautés dynamiques associées à chaque individu appartenant à une institution de sens donnée, tout en l'assignant à sa ou ses communauté(s) de sens calculées puis validées suivant un référentiel donné. L'originalité de notre travail réside dans le fait de compléter ce calcul en ajoutant la définition de communautés « statiques » par une classification prédictive issue des techniques d'apprentissage artificiel réalisée à partir d'une analyse sémantique et numérique de chaque *twitto*. Après avoir dénombré les comptes présents dans les listes thématiques de Twitter, notre approche met en œuvre des outils issus du domaine de l'apprentissage de type SVM (machines à

[10] Dans les travaux menés par G.-E. Sarfati (2008), une formation de sens commun se comprend comme l'ensemble des manières de dire et de signifier des membres d'une même communauté de sens. L'auteur prend pour exemple le domaine médical : il constitue une institution de sens, mais dans l'expérience des sujets-acteurs, ce domaine n'est accessible qu'à partir de ses subdivisions disciplinaires (médecine générale, médecine spécialisée), lesquelles définissent les communautés de sens afférentes. Notre projet de recherche entend aboutir à une caractérisation des sujets-acteurs sur la base de leurs productions discursives, en fonction des institutions et communautés discursivement repérables.

vecteurs de support issues du *machine learning*) qui s'effectue en deux étapes. La première consiste à constituer un échantillon de profils représentatifs de chaque communauté de façon automatique. Il suffit de dépasser un seuil de présence dans les listes thématiques représentatives de chaque communauté. Nous obtenons des données d'entraînement pour lesquelles on connaît la meilleure prédiction. La seconde étape repose sur l'exécution de notre algorithme d'apprentissage SVM à partir des profils validés sur les autres profils ne dépassant pas le seuil pour les classer dans la bonne communauté.

Typologie de *twittos* à partir de six sphères d'influence

Afin de créer des frontières entre *twittos* pour réaliser des analyses croisées du type institutions et communauté de sens, nous proposons une méthodologie originale pour tendre vers une typologie des usagers de Twitter. En effet, le même message n'a pas le même impact s'il provient d'un journaliste renommé, d'un artiste engagé, d'un politicien en campagne, d'un industriel concerné ou d'un citoyen lambda éclairé ; et l'analyse qui en découle peut être améliorée en amont en dégageant les grandes tendances des profils d'usagers[11]. Nous calculons un score local et global d'influence pour chaque utilisateur à partir de la densité et la centralité de son nœud à partir du graphe social étudié. Puis nous calculons pour chaque *hashtag*, chaque nom propre et commun (après étiquetage morphosyntaxique dans le texte intégral) un équivalent à la méthode de pondération *Term Frequency-Inverse Document Frequency* (TF/IDF), qui accorde une pertinence lexicale à un terme au sein d'une biographie ou d'un *tweet* pour chaque *twitto* par rapport à l'ensemble des *tweets* et *twittos* du corpus. Le TF-IDF applique une relation entre un *tweet* et un ensemble de *tweets* partageant des similarités en matière de *hashtag*/nom propre/nom commun. Il s'agit d'atteindre une relation de quantité/qualité lexicale à travers un ensemble de *tweets*. Nous mesurons également les relations sociales entre chaque utilisateur ciblé, les actions sociales globales de chaque *tweet* récupéré (*retweet*, mention, *reply*) et les actions sociales locales relatives à un utilisateur.

[11] Nous nous appuyons sur des travaux réalisés dans le cadre du Concours mondial d'innovation pour lequel l'unité de recherche Etis été lauréate (partenariat laboratoire Équipes Traitement de l'information et systèmes/Qwant – moteur de recherche européen orienté médias sociaux ; projet innovation 2030 COLLECTEURQWANTV2 QWANT SAS) en 2014 de la phase d'amorçage pour la valorisation des données massives (*big data*) dont l'objectif était de créer une typologie des profils d'utilisateurs de la twittosphère francophone comprenant les six sphères d'influence suivantes : politique, médiatique (journalistique), citoyenne, artistique (issue du divertissement), industrielle et scientifique. Nous distinguons des profils d'influenceurs (médiatique, politique, industriel) et de non-influenceurs (citoyens, suiveurs) par un score d'influence personnel par opposition à la solution industrielle klout (Alkhouli, Vodislav, Borzic, 2014, 2015).

Méthodologie d'attribution d'une catégorie à chaque *twitto* du corpus

Nous formalisons nos besoins par la définition d'une problématique de méthodes d'apprentissage supervisées qui consiste à apprendre à classer dans la bonne catégorie les *twittos* émetteurs de messages de notre corpus selon leurs champs description et leur présence dans les listes thématiques de Twitter. Rappelons qu'une telle liste permet pour chaque *twitto* d'organiser ses contacts Twitter par thématiques, qu'il soit ou non abonné à ces comptes. Chaque liste est décrite par plusieurs champs de description, ce qui permettra d'effectuer une analyse sémantique pour préparer l'appariement avec les catégories désirées. Notre approche utilise la distribution des *twittos* au sein de ces listes pour, dans un premier temps, qualifier ces listes et, dans un second temps, inverser le processus pour attribuer le profil idoine à chaque *twitto*.

La validation du projet de la typologie des *twittos* a permis de créer des jeux d'essai d'entraînement (*training dataset*), sortes de collections d'exemples instanciés automatiquement, contrairement aux méthodes traditionnelles qui imposent une validation humaine. Ces jeux d'essai se présentent sous formes de tableaux de données dont chaque ligne est une instance représentant une observation, elle-même décrite par un vecteur. Chaque colonne représente une variable (*attribute*) qui peut être quantitative (*numeric*), qualitative (*nominal*) ou textuelle (*string*). À part la classe à prédire qui est une variable discrète (choix entre les six sphères d'influence), toutes les autres variables sont numériques, continues et correspondent aux fréquences de mots reliés à des branches de thésaurus (scénarios, voir le logiciel *Tropes*[12]), dont le vocabulaire fermé dépend des six domaines prédéfinis.

En cas d'incertitude dans le choix de l'attribution de la classe, de nouvelles valeurs numériques viennent enrichir l'observation du *twitto*, comme son nombre de *followers*, de *following*, de statuts, de favoris, qui peuvent caractériser un domaine.

Rappelons que les catégories médiatiques et politiques ont des indicateurs d'influence plus importants en moyenne que les catégories scientifiques, citoyennes ou artistiques. La catégorie industrielle est plus complexe à cerner car, par exemple dans le domaine des médias sociaux, les indicateurs ont des valeurs importantes.

Résultats sur les *twittos* les plus représentatifs

Au regard des résultats de la classification après validation (taux d'erreur de moins de 7 %), nous remarquons que les 100 *twittos* les plus influents de notre corpus sont présents dans les six grands domaines d'activité avec, en deuxième observation, une homogénéité de la description de *twittos* au sein d'une même sphère d'influence. Ainsi, dans la catégorie artistique, obtenons-nous de façon explicite, dans les champs de

[12] Accès : http://tropes.fr/.

description, deux sortes de profil : soit des regroupements, soit des individus, liés à des d'associations, de fédérations syndicales, de troupes de théâtre, de coordinations des intermittents, d'écoles d'art dramatique, de fédération nationale d'art dramatique. Individuellement, les profils les plus récurrents sont des assistants, des comédiens, des réalisateurs, des écrivains, des truquistes-graphistes, des metteurs en scène ou simplement des intermittents sans précision de branche d'activité. Au niveau de la catégorie médiatique, se dégagent des chroniqueurs ou des rédacteurs en chef. Au niveau de la catégorie politique, le courant représenté est plutôt à « gauche » avec des élus aux étiquettes aussi variées que Front de gauche, socialiste et écologiste. Au niveau de la catégorie scientifique, il est question d'enseignement et de physique alors que, dans la catégorie industrielle, nous avons affaire à du consulting, du conseil, de la direction au sein d'agence et à de la diffusion de spectacle.

Alors que les six premières catégories ont été discriminées par des descriptions « métiers », la catégorie citoyenne est caractérisée par des énoncés parfois génériques, des citations, dont voici quelques exemples :

> « Les plus grandes trahisons sont commises au nom même des idéaux qu'elles prétendent défendre » (accès : https://twitter.com/cleromancie).

> « À l'heure où la machine tend de plus en plus à remplacer la main de l'homme, gifler autrui reste une des dernières joies tangibles de ce siècle inhumain » (accès : https://twitter.com/david_sillet).

Ces éléments sont des extraits de la biographie de *twittos*, et non des *tweets* : ceci illustre le fait que la détection d'un citoyen se réalise par l'analyse de sa propre biographie (en plus de l'analyse de la description des listes auxquelles il est lié). Une telle caractéristique diffère des autres catégories, pour lesquelles on trouve des indications telles « je suis journaliste », « je suis député », etc. Dans le cas de la catégorie « citoyen », on trouve rarement « je suis citoyen », mais plutôt des citations qui évoquent ce qu'il pense en tant que citoyen.

Exploitation des catégories sur les *tweets*

Analysons la distribution, la volumétrie de la publication et les indices d'influencabilité de nos différentes populations. Nous avons retenu quatre critères pour démontrer le caractère innovant de ce type de travaux : la représentativité de chaque catégorie du corpus, la répartition des *tweets* au sein de ces catégories (un *twitto* peut être plus prolifique que dix *twittos* sur un même sujet), la moyenne des scores d'influence des *twittos* d'une même catégorie et, enfin, la moyenne des *retweets* de *tweets* originels (qui ne sont eux même pas des *retweets*). Ainsi avons-nous de nouvelles métriques pour mesurer l'impact, le comportement et la stratégie de publication des acteurs sociaux en fonction de leur appartenance à certaines communautés d'usagers. La métrique la plus problématique concerne le score d'influence des *twittos* car la notion d'« influence » fait évidemment débat. Nous proposons un calcul original, détaillé dans le cadre d'un projet de veille de *tweets*

par centre d'intérêts par l'optimisation d'un algorithme innovant « Continuous top-k queries in social networks » (Alkhouli, Vodislav, Borzic, 2014)[13] :

$$score(m, u) = \alpha\ sim\ (m, p(u)) + (1 - \alpha)\ social(m, u)$$
$$social(m, u) = \beta\ global(m) + (1 - \beta)\ f(u, u^m)$$
$$global(m) = \gamma UI(u^m) + (1 - \gamma)\ AI(m)$$

Ce calcul permet de démystifier l'indice Klout – concept d'influence très critiqué, mais très utilisé dans le monde industriel – qui prétend mesurer l'activité d'un utilisateur (*Kloutscore*) sur les différents réseaux sociaux – Facebook, Twitter, LinkedIn, Google + – en leur attribuant un score de 1 à 100 projeté sur une matrice d'influence très détaillée (Quercia *et al.*, 2011).

Twitter attribue des indices d'influence de type *Kloutscore*, mais ces calculs peuvent poser problème, notamment parce que les explications de calculs fournies ne sont pas très précises. Ainsi Twitter propose-t-il deux listes de résultats au sein de son interface de recherche « À la une » (mélange de *Kloutscore* et de *tweetscore*), par opposition à son tri chronologique sous la dénomination « Récemment. »

Nous postulons que la notion d'« influence » ne peut s'exercer domaine par domaine, communauté par communauté, tout en étant projetée dans un classement unique. Nous avons réglé cette incohérence par notre indice d'« influence » contextuel et singulier pour chaque individu, au sein du jeu d'essai par la position des individus analysés dans le graphe social, en calculant le nombre d'interactions par un individu sur la thématique de l'intermittence.

Tableau 1. Représentativité des différentes catégories de *twittos*

	Artiste	Citoyen	Journaliste	Politicien	Industriel	Scientifique
Twittos	48 %	26 %	11 %	10 %	2 %	3 %
Tweets	80 %	9 %	5 %	4,5 %	0,5 %	1 %
Score d'influence	41	38	52	59	50	54
Moyenne de retweets	3	1,1	3,4	3,2	0	1,9

[13] *score(m, u)* représente l'importance d'un message *m* pour un utilisateur *u*, en combinant les facteurs de contenu et de réseau social ;
sim(m,p (u)) mesure la similarité entre les mots clés du message *m* et ceux du profil utilisateur *p(u)* ;
social(m, u) désigne l'importance du message *m* au sein du réseau social ;
global(m) est l'importance du message dans le réseau social indépendamment de l'utilisateur à la question;
$f(u, u^m)$ représente l'importance relative entre l'émetteur *u* du message et le récepteur u^m ;
$UI(u^m)$ désigne l'importance globale de l'émetteur u^m du message dans le réseau social ;
AI(m) mesure l'importance du message par rapport aux actions provoquées par d'autre utilisateurs.

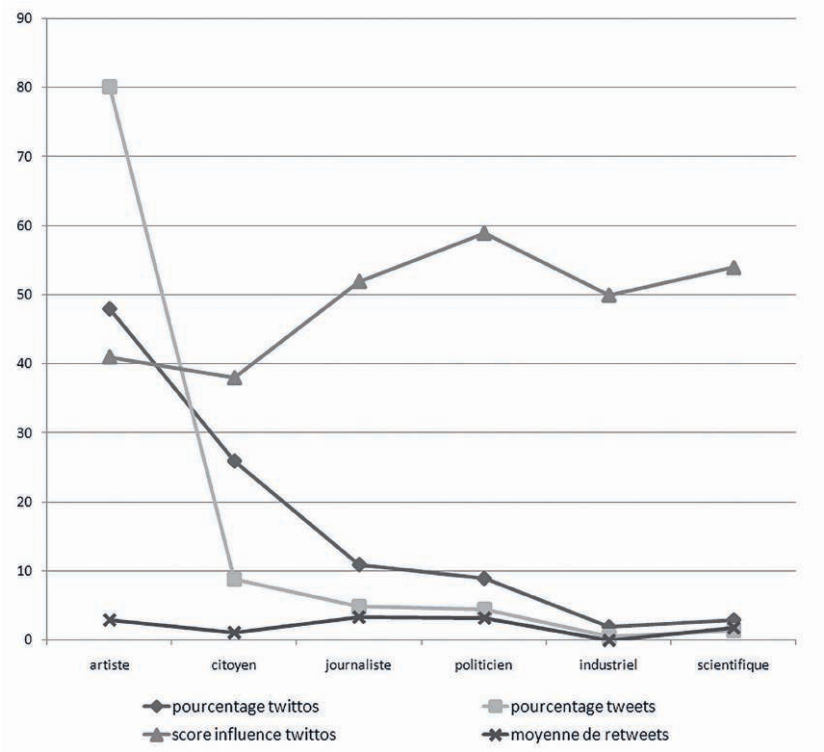
Il est intéressant de souligner que, dans notre exemple, toutes les catégories sont représentées, avec une forte dominante citoyenne/artiste, un deuxième couple plus modeste politique/journaliste et, de façon plus marginale, la paire science/industrie souvent liée aux technologies.

Les *twittos* artistes représentent près de la moitié de l'échantillon, ceux citoyens un quart, les journalistes et politiques sont autour de 10 % chacun, et les *twittos* scientifiques et industriels avoisinent les 3 %.

Dans une perspective émetteur-contenu, nous comptabilisons la provenance des *tweets*. Ceux-ci émanent à 80 % de la sphère artistique, à 9 % de la sphère citoyenne, à 4-5 % des sphères médiatico-politique chacune, et à un peu plus de 1 % de la sphère scientifique et 0,5 % de la sphère industrielle.

On voit que, au niveau des métriques d'influence des *twittos* et des calculs de buzzabilité des *tweets*, la tendance s'inverse : les *twittos* politiques ont un score d'influence de 59 sur 100, les journalistes de 52, suivi des industriels et scientifiques autour de 50, quand les artistes et citoyens ferment la marche avec un score proche de 40. La moyenne du nombre de *retweets* par *tweet* est favorable au couple médiatico-politique, suivi de près par la communauté artiste et, enfin, les communautés scientifique, citoyenne et industrielle ont des valeurs presque négligeables.

Figure 8. Répartition des *twittos* et de leurs caractéristiques



Liens avec les trois classes *Iramuteq*

La dernière partie concerne la projection des trois classes issues de l'analyse des similitudes par *Iramutek* sur nos six sphères d'influence. Si les classes sont définies comme des esquisses de communautés de sens, nous voyons fort bien se profiler la cohérence d'un chevauchement institution/communauté de sens. En effet, la première classe est très fortement représentée dans les sphères politique et artistique, la seconde impacte les sphères citoyenne et artistique et la dernière caractérise plutôt les sphères médiatique, citoyenne et artistique.

Ainsi le *tweet* originel le plus partagé dans la sphère politique provient-il du compte « Jp-Gille », député membre du Parti socialiste et auteur du rapport sur les métiers artistiques : « Dossier de l'intermittence : un conflit complexe à 3 personnages bipolaires #directAN #intermittents http://t.co/V0eEdEcroO ». Ce *tweet* est caractéristique de la troisième classe *Iramuteq*, car il traite bien des protagonistes impliqués dans la réflexion du régime d'assurance chômage des intermittents.

Le *tweet* originel le plus partagé dans la sphère médiatique provient du compte « laparisiennelib » qui est chroniqueuse à *Médiapart* : « Comité d'accueil des #intermittents et #précaires au ministère du travail, en ce moment. Photo de @JoachimSalinger ». Ce qui correspond à la définition informative de la seconde classe *Iramuteq* regroupant des formes associées aux différents mouvements de mobilisations comme l'occupation ou des manifestations à Paris par les intermittents.

Comme prévu, car peu représentatives, les deux sphères non impactées par les classes *Iramuteq* sont caractérisées par d'autres agrégats de mots clés. La sphère industrielle se définit par les concepts : spectacle, artiste, argent et finance, théâtre, intervention, négociation, crise et conflits alors que la sphère scientifique se résume aux expressions *spectacle*, *théâtre*, *courrier*, *réglementation*, *culture*, *conversation*, *accord*, *chef du gouvernement*, *festival*, *mission*, *soutien*...

L'analyse textométrique du corpus #intermittent : explorations argumentatives et textuelles

Cette classification établie, nous cherchons à caractériser plus finement les spécificités du corpus, notamment au regard de la visée pragmatique et argumentative des *tweets*. Un moyen stratégique est l'utilisation du *hashtag*, la combinaison d'un terme et d'un marqueur qui permet aux utilisateurs de Twitter de donner du contexte supplémentaire à leur message pour compenser le nombre restreint de caractères (140 maximum). Cette forme (#*mot*) a un grand pouvoir communicatif : il permet de transformer le message ou la discussion en un hyperlien, cliquable, où sont centralisés les autres messages contenant le même *hashtag*. C'est donc un moyen rapide de créer des espaces communautaires autour d'un sujet de conversation.

Les *hashtags* et leur cooccurrence

Grâce à la fonction « segments répétés » de *Lexico3*[14] (qui donne comme résultat des suites de formes dont la fréquence est supérieure à 2 dans le corpus), nous avons analysé la manière dont les *twittos* utilisent les *hashtags*. Les tableaux suivants montrent les différents segments répétés ainsi que leur fréquence dans le corpus #intermittents.

Tableau 2. Intermittents et précaires

Lg	Segment	Frq
2	de la	1313
2	des #intermittents	1227
2	les #intermittents	965
2	| RT	929
2	#intermittents http	739
2	de l	597
2	#intermittents #précaires	524
2	à la	466
2	les #intermittents	455
2	RT @CIPIDF	449
2	RT @CIP	396
2	du spectacle	392
2	à l	385
2	RT @interluttants	349
2	#intermittents du	328
2	#intermittents et	321
2	RT @intermittentsV2	316
2	#précaires #intermittents	310
2	RT @cgt	302
2	sur les	268
2	#intermittents RT	262
2	aux #intermittents	251
2	#précaires http	237
2	la grève	233
2	sur le	231
2	#intermittents du spectacle	214
2	pour les	210
2	#intermittents #précaires	206

[14] *Lexico3* est un logiciel d'analyse des données textuelles qui établit des statistiques textuelles et propose des fonctionnalités linguistiques (segments répétés, concordanciers, etc.).

Tableau 3. Intermittents et intérimaires

Lg	Segment	Frq
2	et #précaires	179
2	ce soir	173
2	dans la	172
2	c| RT	166
2	avec les	161
3	sur les #intermittents	158
2	soutien aux	156
2	sur la	155
2	par les	149
2	RT @ip	145
2	dans le	144
2	en grève	144
3	#intermittents et #précaires	141
2	#intermittents à	139
2	co|RT	139
2	#intermittents #intérimaires	138
2	régime des	137
3	#intermittents #précaires http	131
2	est pas	130
2	t|RT	128
2	http|RT	126
2	RT @RadiBiCarbonat	123
2	le régime	123
2	soutien aux #intermittents	122
2	RT @franceinter	122
2	ce matin	121
3	régime des #intermittents	121
3	des #intermittents et	120
2	la culture	119
2	pas de	118

Tableau 4. Intermittents et chômeurs

Lg	Segment	Frq
2	#intermittents #interimaires	96
2	pour la	96
2	#Intermittents du	95
2	tous les	94
2	contre l	93
2	du festival	93
2	en lutte	93
2	voté la	92
2	#précaires #chomeurs	92
3	| RT @CIPIDF	91
2	tout l	91
2	la #GREVE	90
2	#précaires #chômeurs	90
3	avec les #intermittents	90
2	que les	90
2	RT @sceneweb	89
2	des #Intermittents	88
2	la Vilette	87
4	des #intermittents de spectacle	86
2	la lutte	84
2	RT @TheaVilleParis	83
2	ht|RT	82
2	et la	82
2	et #intermittents	81
2	#intermittents #chômeurs	79
2	du travail	78
2	#precaires http	77
2	#|RT	77
2	en cours	77
2	la mission	77
2	#chômeurs #précaires	76

Ces figures montrent que la combinaison de *hashtags* est utilisée dans de nombreux *tweets* et qu'elle se présente généralement sous deux formes : soit un lexème employé seul et renvoyant le plus souvent à des adjectifs (#*précaires*, #*intérimaires*) ou à des noms (#*grève*, #*intermittents*), soit des cooccurrences composées de deux ou plusieurs *hashtags* classés les uns à côté des autres (#*intermittents* #*précaires*, #*intermittents* #*chômeurs*, #*précaires* #*intermittents* #*intérimaires*). Ces différents types de *hastags* sont placés au début, au milieu ou à la fin d'un *tweet*.

Afin de savoir comment les *hashtags* s'intègrent dans la phrase, et préciser leur rapport syntaxique avec les composants d'un *tweet*, nous nous sommes intéressés aux cooccurrences. Le constat est que le degré d'intégration du *hashtag* varie, allant d'une indépendance totale, qui donne lieu à des cooccurrences ayant des composants de nature figée, à l'intégration syntaxique complète où les constituants solidaires du polylexème se défigent. Le degré de figement des unités des cooccurrences de *hashtags* augmente à partir de trois *hashtags* consécutifs. *A contrario*, on remarque que les coocurrences ayant deux *hashtags* s'intègrent plus fréquemment dans la structure syntaxique de la phrase. Les exemples qui suivent illustrent ce raisonnement :

> « Grand rassemblement avec les #intermittents #précaires et amis #solidaires aujourd'hui à #Bordeaux. Sur la brèche ! » (accès : https://twitter.com/bbotte33/status/478618788850839552).

> « Valls à 3 temps ? Plutôt un slam de rue pour les #intermittents #précaires surtout à #Nantes #cultureenmarche http://fb.me/6CJxVIIWE »
> (accès : https://twitter.com/ECOLIGHTnews/status/481668055895142400).

Ces exemples contiennent des cooccurrences formées de deux *hashtags*, #*intermittents* et #*précaires*. Ils illustrent tous le cas d'intégration syntaxique totale où le *hashtag* remplit une fonction syntaxique. Dans le premier extrait, le binôme #*intermittents* #*précaires* s'intègre parfaitement d'un point de vue syntaxique au reste de la phrase, la préposition *avec* introduit un complément d'objet indirect constitué d'un déterminant *les*, d'un nom *intermittents* et d'un adjectif *précaires*.

Défigements et problèmes syntaxiques

Les exemples ci-dessous illustrent les cas où l'on commence à apercevoir la structure syntaxique canonique, avec un certain degré de défigement, mais avec tout de même quelques anomalies qui causent une certaine incongruence syntaxique, ceci étant peut-être dû à l'économie dans la rédaction des *tweets* :

> « Comme je serai au débat #intermittents & #précaires, je vais louper #Notre_Dame_des_Landes de Fer par #GirO : ARG ! ! ! » (accès : https://twitter.com/ValKphotos/status/484713083839844352).

> « 7-9 FranceInter perturbé par #intermittents #chômeurs ce matin pour appel grève spectacle-intérim 01/10 Effacé des podcasts de la chaîne ! :) » (accès : https://twitter.com/CieJolieMome/status/515476648452689920).

> « Les muscles du gouvernement sont en train de déloger #intermittents #précaires de la Drac de #Lille Besoin de renforts #intermittents » (accès : https://twitter.com/interluttants/status/479874550717751296).

Enfin, les cas suivants montrent des cooccurrences qui ne remplissent pas les fonctions syntaxiques. Ce sont des procédés morphologiques qui se concurrencent produisant plusieurs signifiants pour un référent renvoyant à une même thématique :

> « Lundi à #RENNES AG à 9H30 aux Ateliers du Vent puis RDV à 14h30, pl. de la Mairie #intermittents #interimaires #précaires NON À L'AGRÉMENT ! » (accès : https://twitter.com/cgt_spectacle/status/478217232955478016).

> « Quimper Blocage de l'entreprise BTP Le Bris (patron pdt Medef Finistère) #intermittents #chômeurs #précaires #appel » (accès : https://twitter.com/CIPIDF/status/479153377105149952).

Il s'agit ici d'un objectif communicationnel, de référencement, pour renforcer la visibilité d'un thème (dans notre cas le thème étant « les intermittents de spectacle qui luttent contre le nouveau régime d'assurance chômage »).

Dans le premier exemple, les termes *intermittents*, *intérimaires*, et *précaires* sont des noms et ne peuvent être d'une autre nature comme des adjectifs car ils représentent des thèmes, ils deviennent de ce fait complètement figés et indépendants syntaxiquement. C'est une forme de polylexicalité qui se constitue par la répétition d'une séquence réussie et qui se reprend parce qu'elle correspond originellement à un but dénominatif.

Grâce à cette analyse des cooccurrences de *hashtags*, nous avons pu remarquer que ces marqueurs parviennent à construire des associations lexicales et thématiques. Dans le corpus, *intermittent* est donc associé à la précarité, au chômage, de manière très efficace par le référencement qui est créé entre les *hashtags* construits.

Conclusion

Nous avons essayé de mettre en avant l'originalité de notre approche par son caractère hybride qui articule des méthodologies souvent cloisonnées. Notre méthodologie est issue des méthodes de la linguistique de corpus, de la sémantique textuelle, mais aussi des techniques d'apprentissage artificiel. Elle s'appuie sur la description, puis la classification prédictive après une analyse de la description textuelle et numérique de chaque *twitto*, après une analyse linguistique du corpus, une étude des spécificités argumentatives et discursives, et un dénombrement des comptes présents dans des listes Twitter thématiques correspondant à des domaines d'activité bien délimités. À l'issue de la partie descriptive qui permet d'informer sur la nature du corpus, l'apprentissage s'effectue après avoir entraîné le classifieur sur des *twittos* bien identifiés et validés grâce à un procédé itératif de sauts de seuils (classement par nombre d'usagers dans les listes, nombre de listes pour chaque usager). Il s'agit d'une analyse à la fois descriptive et prédictive, qui s'appuie sur les formes explicites et sur les liens implicites. L'algorithme de prédiction analyse les valeurs des comptes de *twittos* déjà présents dans la base Twitter francophone validée, et prend une décision pour attribuer à chaque nouveau *twitto* sa catégorie correspondante. L'attribution effectuée est corrélée et confirmée par l'analyse textométrique et linguistique préalablement

identifiée, puis précisée qualitativement par des analyses linguistiques de cooccurrences de *hashtags*. Cette collaboration entre linguistes et informaticiens semble nécessaire pour appréhender de manière originale et optimale les grands volumes de données sociales, auxquels la constitution en corpus structurés confère le statut d'humanités numériques, analysables et interprétables avec des outils facilitant l'accès au sens.

Références

Alkhouli A., Vodislav D., Borzic B., 2014, « Continuous Top-k Processing of Social Network Information Streams: A Vision », pp. 35-48, *in*: Kotzinos D., Wei Choong Y., Spyratos N., Tanaka Y., eds, *Information Search, Integration, and Personalization. 9th International Workshop, ISIP 2014, Kuala Lumpur, Malaysia, October 9-10, 2014. Revised Selected Papers*, Cham, Springer.

Alkhouli A., Vodislav D., Borzic B., 2015, « Algorithms for continuous top-k processing in social networks », *in*: *International Symposium on Web AlGorithms*, Deauville, juin.

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C.R., Hriba L., Longhi J., Seddah D., 2014, « The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres », *Journal of Language Technology and Computational* Linguistics, pp. 1-30.

Clavert F., 2016, « Échos du centenaire sur le web et sur Twitter : partage de liens et Grande Guerre », *in* : *Revisiter la commémoration. Pratiques, usages et appropriation du Centenaire de la Grande Guerre*, Nanterre, mars. Accès : https://halshs.archives-ouvertes.fr/halshs-01295550.

Clavert F., Majerus B., Beaupré N., 2015, « #ww1. Twitter, the Centenary of the First World War and the Historian », *Twitter for Research 2015*, Lyon, avr. Accès : https://halshs.archives-ouvertes.fr/halshs-01148548.

Ducrot,O, 1972, *Dire et ne pas dire*, Paris, Hermann.

Ducrot O., 1973, *Qu'est-ce que le structuralisme ?*, Paris, Éd. Le Seuil.

Flament C., 1962, « L'analyse de similitude », *Cahiers du Centre de recherche opérationnelle*, 4, pp. 63-97.

Flament C., 1981, « L'analyse de similitude : une technique pour les recherches sur les représentations sociales », *Cahier de psychologie cognitive*, 1, pp. 375-395.

Grandjean M., 2016, « A Social Network Analysis of Twitter: Mapping the Digital Humanities Community », *Cogent Arts & Humanities*, 3. Accès : https://doi.org/10.1080/23311983.2016.1171458.

Jackiewicz A., Vidak M., 2014, « Étude sur les mots-dièse », *SHS Web of Conferences*, 8. Accès : https://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01198.pdf.

Lebart L., Salem A., 1994, *Statistique textuelle*, Paris, Dunod.

Longhi J., 2006, « De *intermittent du spectacle* à *intermittent* : de la représentation à la nomination d'un objet du discours », *Corela. Cognition, représentation, langage*, 4 (2). Accès : http://corela.revues.org/457.

Longhi J., 2008, « Sens communs et dynamiques sémantiques : l'objet discursif intermittent », *Langages*, 170, pp. 109-124.

Marchand P., Ratinaud P., 2012, « L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011) », pp. 687-699, *in* : Dister A., Longrée D., Purnelle G., éds, *JADT 2012. 11e Journées internationales d'analyse statistique des données textuelles. Actes - Proceedings*, Liège/Bruxelles, Université de Liège/Facultés universitaires Saint-Louis-Bruxelles. Accès : http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Marchand,%20Pascal%20et%20al.%20-%20L'analyse%20de%20similitude%20appliquee%20aux%20corpus%20textuels.pdf.

Mathieu G., 2013, *Les Intermittents du spectacle. Enjeux d'un siècle de luttes*, Paris, Éd. La Dispute.

Pariser E., 2011, *The Filter Bubble: What the Internet is hiding from You*, New York, Penguin Press.

Pincemin B, 2011, « Sémantique interprétative et textométrie – Version abrégée », *Corpus*, 10, pp. 259-269. Accès : http://corpus.revues.org/2121.

Quercia D., Ellis J., Capra L., Crowcroft J., 2011, « In the Mood for Being Influential on Twitter », pp. 307-314, *in*: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (PASSAT-SOCIALCOM)*, Washington, Conference Publishing Services.

Reinert M., 1998, « Quel *objet* pour une *analyse statistique du discours* ? Quelques réflexions à propos de la réponse *Alceste* », pp. 557-690, *in* : *Actes JADT'1998 en ligne*. Accès : http://lexicometrica.univ-paris3.fr/jadt/jadt1998/reinert.htm.

Reinert M., 1999, « Quelques interrogations à propos de l''objet'' d'une analyse de discours de type statistique et de la réponse ''Alceste'' », *Langage & société*, 90 (1), pp. 57-70. Accès : www.persee.fr/doc/lsoc_0181-4095_1999_num_90_1_2897.

Sarfati G. E., 2008, « Pragmatique linguistique et normativité : remarques sur les modalités discursives du sens commun », *Langages*, 170, pp. 92-108. Accès : http://www.cairn.info/revue-langages-2008-2-page-92.htm.

Corpus CoMeRe : https://corpuscomere.wordpress.com
Iramuteq : www.iramuteq.org
Ortolang : https://www.ortolang.fr/market/home
TROPES : http://tropes.fr/