Tweets politiques : corrélation entre forme linguistique et information véhiculée Julien Longhi

Julien Longhi

Chapitre 9

Tweets politiques : corrélation entre forme linguistique et information véhiculée

Julien Longhi

Bien que Twitter soit de plus en plus étudié dans différents champs de recherche (fouille de données, information-communication, sociologie du numérique, etc.), la complexité énonciative des productions sur Twitter nécessite une prise en compte minutieuse de ses différents aspects. En particulier, notre travail s'intéresse à la corrélation entre le statut énonciatif ou discursif de certains tweets, et le sens qu'ils véhiculent. En effet, bien qu'il puisse exister plusieurs typologies de tweets, nous nous attachons dans la recherche que nous menons à établir des sous-ensembles de tweets qui partagent un certain nombre de propriétés linguistiques.

Notre étude est menée à partir du corpus *Polititweets* (Longhi *et al.* 2014), soit 34 273 tweets issus de 205 comptes influents, au sein duquel nous avons déjà mené des recherches, notamment pour extraire les tweets idéologiques, les tweets négatifs, ou les tweets efficaces (ces trois caractéristiques pouvant éventuellement se recouper).

Dans ce chapitre, nous proposons de partir de ces trois sous-ensembles de tweets déjà identifiés dans d'autres travaux, et de comparer les résultats obtenus, afin de proposer une corrélation entre la typologie linguistique des tweets et la nature de l'information qu'ils véhiculent.

Méthodologie d'étude des tweets

Pour mener cette étude, nous nous appuyons sur trois analyses distinctes précédemment menées : Djemili *et al.* (2014) sur l'idéologie, Longhi *et al.* (2016) sur la négation, et Longhi (2017) qui aborde les tweets efficaces.

Identification des tweets idéologiques

Les tweets idéologiques sont identifiés à partir des critères de définition de l'idéologie exposés dans Sarfati (2014), convertis en règles permettant de chercher informatiquement notamment l'absence de marques de *deixis* spatiotemporelle, l'absence de sujets de l'interlocution et la présence du « non-sujet », l'absence de noms propres indiquant des lieux, des personnes, des données factuelles trop précises, un gommage de l'argumentation, une moindre présence des temps du passé, pas/peu de marques de discours autre, etc. Sur 20 040 tweets soumis à l'analyse, 321 ont été identifiés comme idéologiques par l'outil d'analyse, parmi lesquels 172 ont été confirmés par un « expert » (moi-même) et deux étudiants de master journalisme chargés d'étiqueter les tweets comme idéologiques ou non idéologiques. Ces 172 tweets correspondent au sous-corpus des tweets idéologiques.

Les tweets négatifs sont les tweets qui ont été détectés comme portant une forme de négation, qu'elle soit syntaxique, sémantique, ou pragmatique ; 3 190 tweets ont été identifiés comme négatifs.

Les tweets efficaces sont les 600 tweets du corpus les plus retweetés et/ ou favorisés (c'est-à-dire qu'ils ont eu le plus de « succès »).

Ces trois sous-ensembles sont relativement hétérogènes du point de vue de leur taille. Nous pensons néanmoins qu'une analyse appuyée sur des critères statistiques pourra nous permettre de donner quelques indications sur leurs spécificités. Cette analyse doit en outre articuler des aspects quantitatifs et qualitatifs, afin de mettre à profit les niveaux d'analyse linguistique pertinents.

Une méthode mixte d'analyse

Nous adoptons donc pour appréhender ces corpus une méthode mixte, qui s'intègre dans les préconisations de François Rastier. Concernant notre objet d'étude, le corpus *Polititweets* correspond bien à la définition qu'il en donne :

> Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. […] De fait, tout regroupement de textes ne mérite pas le nom de corpus. (Rastier 2004)

Pour l'étude des tweets, cette approche est intéressante car elle diffère des conceptions computationnelles : pour nous un corpus n'est pas « un sac

de mots » (*ibid.*), et le geste d'informatisation des corpus est foncièrement lié aux aspects philologiques et herméneutiques. Notre démarche vise donc également à suivre les préconisations de F. Rastier :

> La linguistique ne peut être véritablement appliquée que si elle est également impliquée. Elle se doit d'intervenir à diverses étapes : création des logiciels, constitution des corpus, balisages, expérimentations avec outils sur corpus (balisés), interprétation et discussion des résultats. À toutes ces étapes de la chaîne de traitement, des connaissances linguistiques, et plus largement sémiotiques, se révèlent indispensables. (Rastier 2008)

***Polititweets*, un « terrain » d'analyse et d'observation**

En s'inspirant de la démarche de l'enquête de terrain du sociologue ou de l'anthropologue (voir Longhi 2015, à propos de Beaud & Weber 2012), on peut chercher à aborder le corpus comme un « terrain » d'analyse. Il convient alors de :

— se poser une question de départ ;
— passer du thème à la question d'enquête ;
— transformer « une question "abstraite" en une série décomposée de pra-tiques sociales et d'événements » (Beaud & Weber 2012 : 36).

La question de départ est celle mentionnée précédemment : peut-on établir des corrélations entre le statut énonciatif des tweets et l'information qu'ils véhiculent ? Pour cela, nous préconisons un « mix méthodologique » entre des moyens relevant tantôt de ce qui est qualifié d'analyse qualitative, tantôt d'analyse quantitative. En effet, nous pensons que l'opposition « quantité »/« qualité » radicalise des positionnements qui pourraient être articulés. Par exemple, d'un côté le « bon » corpus pourrait être qualifié ainsi parce qu'il est un « grand » corpus (et parce que la méthode se fonde sur la quantité des données, par le biais d'outils informatisés), alors que d'un autre côté la « qualité » du corpus et des méthodes d'analyse (plus « manuelles ») pourraient prévaloir, quitte à travailler sur des corpus de moindre envergure. Notre objectif est donc d'apporter aux corpus des traitements qualitatifs à partir d'indicateurs linguistiques minutieux, appliqués à des quantités de données importantes, et d'opérer un va-et-vient entre les résultats quantifiés et leur spécification dans les observables.

Concernant le corpus, il articule les deux soucis mentionnés précédemment : le choix des comptes Twitter a été guidé par le critère de l'influence sur ce réseau, car ce critère pouvait être exploité par la suite à propos de questions d'efficacité. Pour cela, les étapes suivantes ont été appliquées :

1. partir de 7 personnalités de 6 groupes politiques ;
2. récupérer toutes les listes créées par des utilisateurs où ces personnalités étaient citées : cela a conduit à 7 087 listes ;
3. sélectionner parmi ces listes celles qui avaient au moins 6 twittos (utilisateurs de Twitter) et qui contenaient la chaîne de caractères *politic* dans le nom ou descriptif de la liste : 120 listes ;
4. sur ces 120 listes récupérer les 2 934 twittos ;
5. travailler par seuil, en ne retenant que les comptes présents dans plus de 12 listes : nous arrivons ainsi à 205 twittos politiciens ;
6. sur ces 205 comptes nous avons récupéré les 200 derniers tweets de chacun au 27 mars 2014, soit 34 273 tweets.

Cette procédure nous permet d'avoir un ensemble de messages provenant de comptes « politiques », en garantissant que cette étiquette n'est pas projetée sur les données par le chercheur, mais qu'elle est choisie par les usagers (à l'étape 3) par la dénomination de leur liste avec la chaîne *politic*.

Une fois ces procédures établies, nous avons, en collaboration avec les participants au projet du domaine informatique (Boris Borzic et Abdulhafiz Alkhouli), opéré une sélection des données et des métadonnées. Pour cela, ils ont développé une application sur mesure en trois étapes :

1. appel d'une dizaine de fonctions de l'API de Twitter selon nos besoins, et récupération de toutes les informations sous format JSON que nous convertissons ;
2. import de ces informations dans une base de données ayant un design propre (une dizaine de tables et une cinquantaine de champs) ; enrichissement de champs supplémentaires par des indices calculés automatiquement par nos programmes ;
3. puis export sur mesure, avec les informations stockées, dans n'importe quel format de données.

La figure 1 donne un exemple de la base de données avec quelques champs sélectionnés :

tweet_text	screen_name	favoritecount	retweetcount
Toute ma gratitude va à lextraordinaire personnel ...	valtrier	2286	6337
Merci du fond du cœur à tous ceux qui ont envoyé d...	valtrier	1830	3429
"Soyons dignes, soyons patriotes, soyons Français ...	NicolasSarkozy	1632	5711
Avant dentrer à la présidence de la République, je ...	fhollande	1401	6806
Que chacun dentre vous sache combien je suis reco...	NicolasSarkozy	1350	2617
"Vous pourrez compter sur moi à chaque fois quil y ...	NicolasSarkozy	614	1832
"Jamais je ne pourrai vous rendre tout ce que vous...	NicolasSarkozy	499	2499
@nicodomenach je pourrais aussi entrer en arrière ...	ramayade	464	2979
Joie daccueillir le père Georges, rayonnant de gén...	fhollande	446	1129
#FF @Elysee pour suivre toutes les informations su ...	fhollande	427	995

Figure 1. Capture d'écran de la base de données développée au laboratoire ETIS

L'enjeu pour une approche linguistique est donc d'utiliser ce matériau pour élaborer le corpus *Polititweets*. Les différentes questions philologiques et herméneutiques ont été prises en charge par le balisage TEI ou *Text Encoding Initiative* (initiative pour l'encodage du texte), qui « est une communauté académique internationale dans le champ des humanités numériques visant à définir des recommandations pour l'encodage de documents textuels. Depuis 1987, le modèle théorique s'est adapté à différentes technologies, d'abord sous la forme d'une DTD SGML, puis XML. Dans sa version P5 (2007), le schéma TEI est représenté dans plusieurs langages, et notamment, Relax-NG [1] ».

Le corpus *Polititweets* permet donc de tenir compte, par le biais d'un codage mobilisant des balises TEI spécifiques, des différentes ressources sémiotiques des tweets politiques (voir Longhi 2017 pour la description exhaustive de ces enjeux).

L'analyse des trois sous-corpus

Dans cette partie, nous analyserons successivement les trois sous-corpus.

Le sous-corpus de tweets idéologiques

Ce corpus est relativement restreint, puisqu'il se compose de 172 tweets. Nous avons procédé à son traitement dans Iramuteq (voir Marchand & Ratinaud 2012). Le tableau 1 regroupe, par ordre de fréquence, les mots les plus employés :

1. <https://fr.wikipedia.org/wiki/Text_Encoding_Initiative>.

Tableau 1. Fréquence des mots les plus employés dans le sous-corpus idéologique

Mot	Fréquence	Catégorie	Mot	Fréquence	Catégorie
politique	16	nom	seul	8	adjectif
européen	14	adjectif	question	8	nom
société	12	nom	gouvernement	8	nom
public	12	nom	français	8	adjectif
gauche	10	nom	Français	8	nom propre
droite	10	nom	devenir	8	verbe

Nous pouvons en donner une représentation visuelle grâce au nuage de mots, où la taille des mots est proportionnelle à leur fréquence (figure 2).

Figure 2. Nuage de mots du sous-corpus idéologique

Ces deux éléments nous permettent de voir que ce sont des noms assez généraux qui sont employés dans ce type de tweets : « politique », « société », « public », ainsi que le clivage « gauche » et « droite ». L'adjectif « européen » est également employé.

Pour aller plus loin, on peut utiliser l'analyse des similitudes, qui présente les cooccurrences sous forme graphique. Cela permet de visualiser les réseaux qui constituent les textes, et les associations fréquentes (figure 3).

Cette représentation est plus dynamique, puisqu'elle fait ressortir cer-tains usages, comme « extrême droite », « parlement européen », et pointe l'attention sur certaines préoccupations telles que le « devenir », la « jeunesse » et l'« école ».

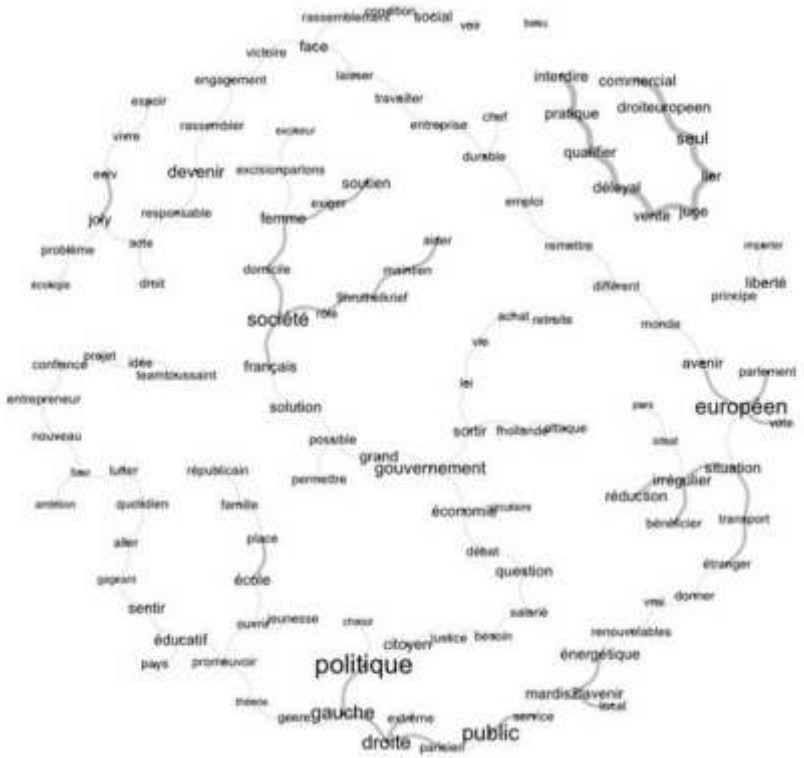
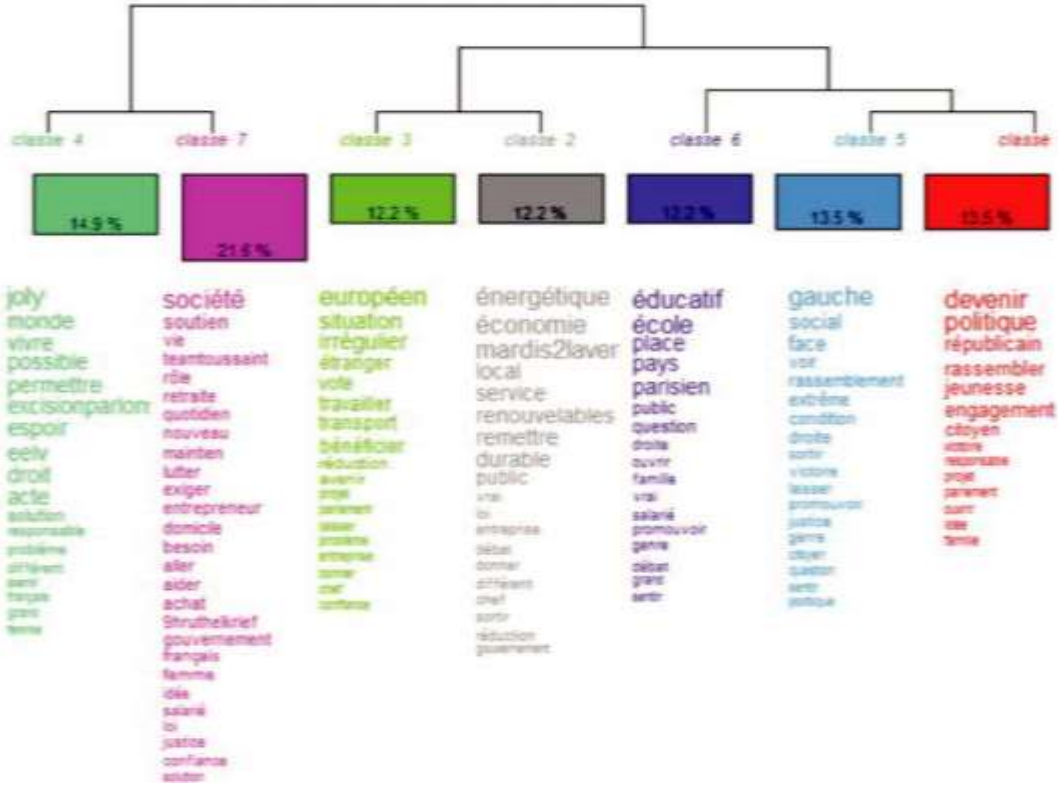


Figure 3. Analyse des similitudes du sous-corpus idéologique


classe 4
classe 7
classe 3
classe 2
classe 6
classe 5
classe 1
14.9 %
21.6 %
12.2 %
12.2 %
13.2 %
13.5 %
13.5 %
joly
monde
vivre
possible
permettre
excisionparlons
espoir
eelv
droit
acte
société
soutien
vie
teamtoussaint
rôle
retraite
quotidien
nouveau
maintien
lutter
exiger
entrepreneur
domicile
besoin
aller
aider
achat
5thruthelikrief
gouvernement
français
femme
idée
salarié
loi
justice
confiance
soutien
européen
situation
irrégulier
étranger
voté
travailler
transport
bénéficier
réduction
avenir
impôt
parlement
laisser
problème
entreprise
donner
cher
confiance
énergétique
économie
mardis2laver
local
service
renouvelables
remettre
durable
public
vrai
loi
entreprise
débat
donner
affirmer
chef
sortir
réduction
gouvernement
éducatif
école
place
pays
parisien
public
question
droite
ouvrir
famille
vrai
salarié
promouvoir
parole
débat
grand
sentir
gauche
social
face
voir
rassemblement
extrême
condition
droite
sortir
victoire
laisser
promouvoir
justice
parole
citoyen
question
sentir
politique
devenir
politique
républicain
rassembler
jeunesse
engagement
citoyen
victoire
résistance
projet
parlement
ouvrir
idée
famille

Enfin, une troisième analyse est possible : issue des travaux de Max Reinert, elle propose une classification hiérarchique descendante. La méthode Alceste (Analyse des lexèmes cooccurrents dans les énoncés simples d'un texte), implémentée dans le logiciel Iramuteq, effectue à partir d'un corpus une première analyse détaillée de son vocabulaire, et constitue le dictionnaire des mots ainsi que de leur racine, avec leur fréquence. Ensuite, par fractionnements successifs, elle découpe le texte en segments homogènes contenant un nombre suffisant de mots, et procède alors à une classification de ces segments en repérant les oppositions les plus fortes. Cette méthode permet d'extraire des classes de sens, constituées par les mots et les phrases les plus significatifs ; les classes obtenues représentent les idées et les thèmes dominants du corpus.

Le résultat sur ce sous-corpus est représenté ci-contre (figure 4).

Les thématiques idéologiques identifiables seraient donc l'éducation, l'énergie, l'Europe, certaines questions de société (retraites, emplois à domicile), et divers points relatifs à la politique « politicienne » (Eva Joly, la gauche, les républicains).

Observons à présent le sous-corpus des tweets efficaces.

Le sous-corpus de tweets efficaces

Ce sous -corpus est constitué de 600 tweets caractérisés ainsi car ayant été les plus retweetés et favorisés. De la même manière, procédons à son traitement dans Iramuteq (tableau 2) :

Tableau 2. Fréquence des mots les plus employés dans le sous-corpus efficace

Mot	Fréquence	Catégorie	Mot	Fréquence	Catégorie
France	67	nom propre	ns	21	nom propre
français	42	adjectif	Pen	19	nom propre
Sarkozy	36	nom propre	grand	19	adjectif
UMP	31	nom propre	beau	19	adjectif
aller	27	verbe	Hollande	18	nom propre
soutien	24	nom	François	17	nom propre
république	23	nom	FN	17	nom propre
Nicolas	23	nom propre			
MLP	22	nom propre			
politique	21	adjectif			

Nous constatons que « France » est le mot le plus employé, ainsi que l'adjectif « français », et les noms propres « Sarkozy » et « UMP ». Le nuage de mots permet de préciser que ce sont « Nicolas Sarkozy » et « France »/« français » qui dominent (figure 5).

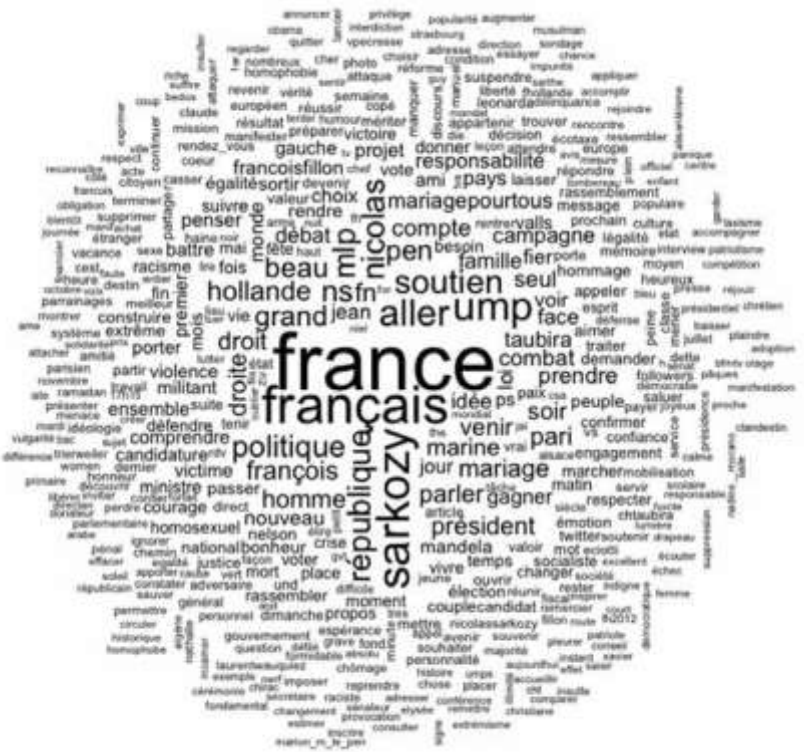


Figure 5. Nuage de mots du sous-corpus efficace

Ce schéma ne permet pas de voir ce qui est vraiment significatif pour ce corpus, aussi l'analyse des similitudes permet là encore de dégager des schèmes discursifs caractéristiques (figure 6).

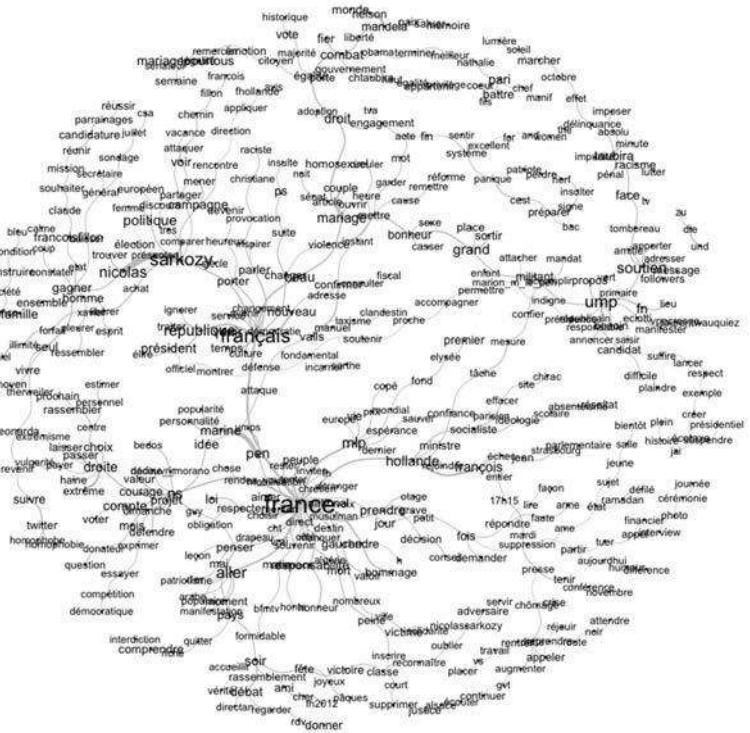


Figure 6. Analyse des similitudes du sous-corpus idéologique

Il semble donc être question de la campagne politique de Nicolas Sarkozy, des Français en lien avec la démocratie et la république, de la France avec les items « Marine Le Pen », « national », « menace » et « voix », ainsi que « mariage », « homosexuel », « loi » et « égalité ». Cela amène à réfléchir sur les thématiques des tweets efficaces, qui semblent concerner Nicolas Sarkozy en contexte électoral, la question de la France en lien avec les probléma-tiques de la nation et du Front national, et dans une moindre mesure le Front national.

Ce résultat est confirmé par l'analyse Alceste (figure 7).

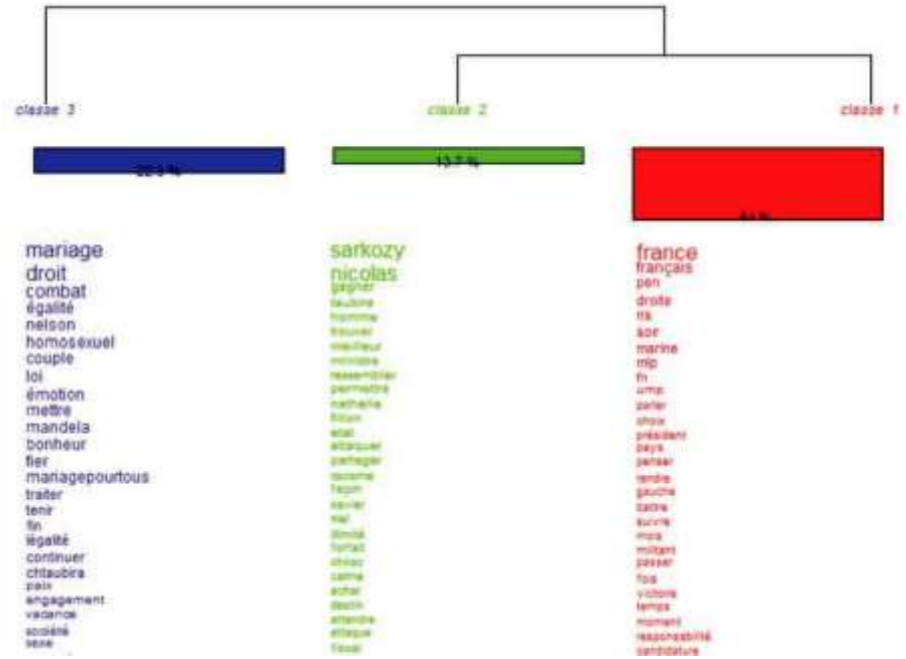


Figure 7. Classification descendante du sous-corpus efficace

Les résultats de ce sous-corpus sont très différents des précédents : alors que le sous- corpus des tweets idéologiques était plus réduit, les résultats étaient bien plus diversifiés. Dans le cas des tweets efficaces, seulement trois thématiques sont identifiées : elles semblent liées à une certaine polémicité, et relativement distinctes l'une de l'autre. L'analyse factorielle des correspondances (AFC) confirme d'ailleurs cela. Cette analyse est très utilisée dans les traitements lexicométriques ou textométriques des corpus, car elle donne une représentation visuelle permettant des interprétations qui mettent en correspondance les parties de corpus ou les sous-ensembles avec les éléments qui composent ces parties :

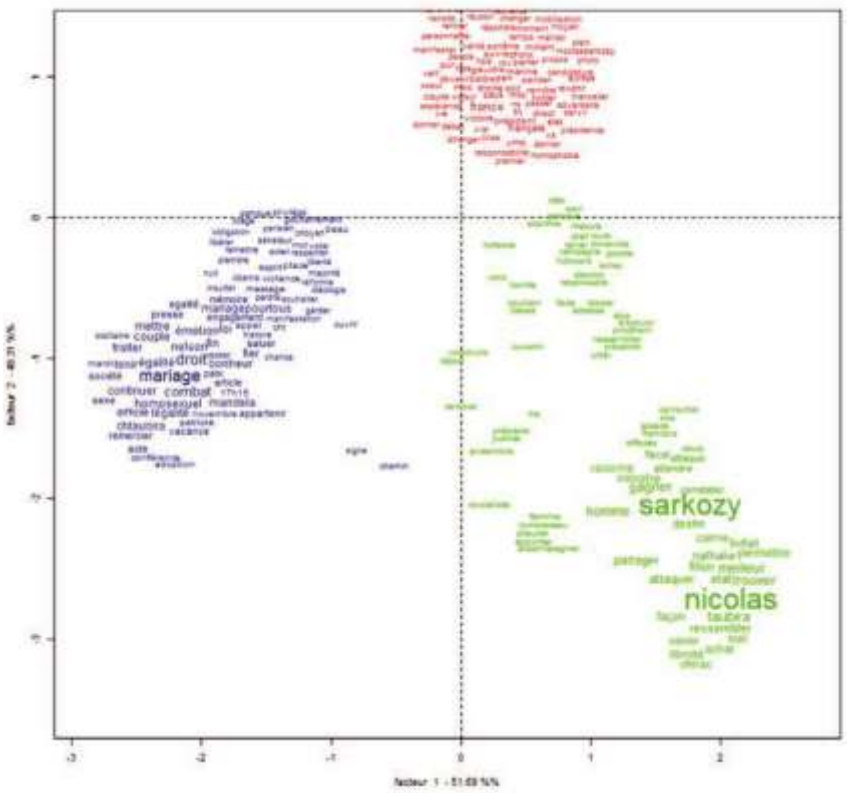


Figure 8. AFC relative à la classification descendante (sous-corpus efficace)

L'AFC peut être définie ainsi :

L'analyse factorielle des correspondances est une méthode statistique qui s'applique aux tableaux de contingence, tels par exemple les tableaux résultant du décompte de différents types de vocabulaire (lignes du tableau) dans les différentes parties (colonnes du tableau) d'un corpus de textes. On commence par calculer une distance (dite distance du χ^2) entre chacune des paires de textes qui constituent le corpus. On décompose ensuite ces distances sur une succession hiérarchisée d'axes factoriels. […] Cette méthode permet d'obtenir des représentations synthétiques portant à la fois sur les distances calculées entre les textes et celles que l'on peut calculer entre les unités textuelles qui les composent. […] L'intérêt principal de l'AFC réside dans sa capacité à extraire à partir de vastes tableaux de données difficilement appréhendables des structures simples qui rendent compte approximativement des grandes oppositions sous-jacentes dans un corpus de textes. (Tutoriel d'André Salem pour Lexico3).

Cette représentation visuelle indique qu'il n'y a pas de recouvrements entre les classes, et que les tweets sont donc rattachés à l'une ou l'autre des trois classes. Cela n'est pas toujours le cas, comme en témoigne la représentation obtenue pour le sous-corpus idéologique (figure 9).

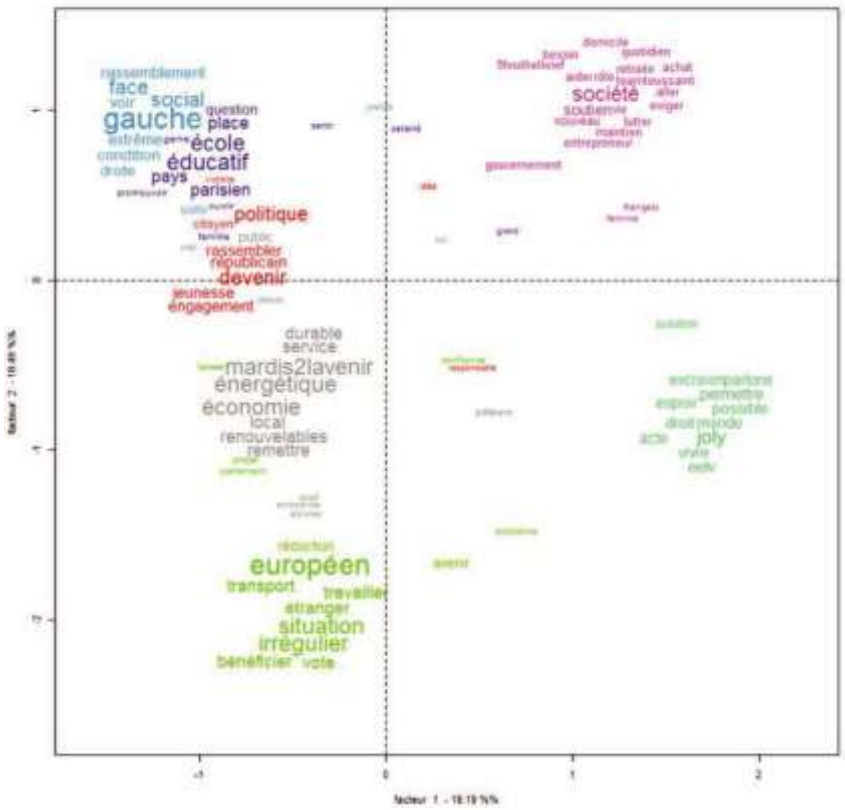


Figure 9. AFC relative à la classification descendante (sous-corpus idéologique)

On distingue en effet des thématiques plutôt indépendantes et d'autres plutôt enchevêtrées, ce qui indique que l'idéologie se distingue non seulement par les thèmes mis en discours, mais aussi par leurs liens possibles, nous y reviendrons.

Le sous-corpus de tweets négatifs

Pour ne pas alourdir le propos, nous ne décrirons pas ici le processus de sélection des tweets négatifs, sujet sur lequel nous renvoyons à Longhi *et al.* (2016). Sur l'ensemble du corpus, 3 190 tweets ont été détectés comme négatifs, grâce à une extraction par Unitex.

À partir de ce sous-corpus, nous avons également procédé au traitement quantitatif des termes présents dans ce corpus (tableau 3).

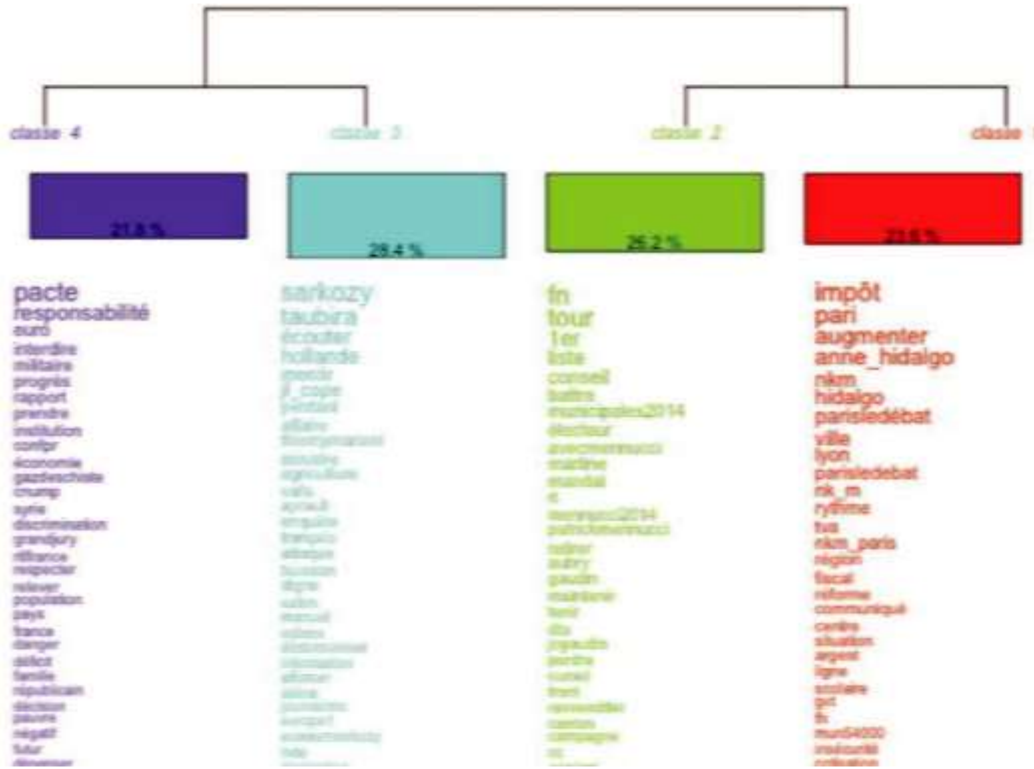
Tableau 3. Fréquence des mots les plus employés dans le sous-corpus négatif

Mot	Fréquence	Catégorie	Mot	Fréquence	Catégorie
politique	63	adjectif	seul	24	adjectif
France	47	nom propre	changer	23	verbe
gauche	40	nom	prendre	22	verbe
aller	39	verbe	Hollande	22	nom propre
français	38	adjectif	voter	21	verbe
parler	29	verbe	vie	21	nom
débat	29	nom	FN	21	nom propre
président	25	nom	élection	20	nom
pays	25	nom	voir	20	verbe
droite	25	nom	UMP	20	nom propre

L'analyse des similitudes permet notamment de voir les environnements lexicaux de « politique », « France », « gauche » ou « français » (figure 10).

Figure 10. Analyse des similitudes du sous-corpus négatif

« Politique » interagit avec « débat », « changer », ou encore « vie », « France » avec « Hollande », « Allemagne » ou « aller », et « français » avec « Sarkozy ». Pour dégager les thématiques de ces tweets, penchons-nous sur l'analyse Alceste (figure 11).


classe 4
classe 3
classe 2
classe 1
pacte
responsabilité
euro
sarkozy
taubira
écouter
hollande
fn
tour
1er
liste
conseil
impôt
pari
augmenter
anne_hidalgo
nkm
hidalgo
parisledébat
ville
lyon
parisledebat

Les thématiques sont elles aussi assez différentes : elles concernent le pacte de responsabilité et l'euro, le débat politicien avec Sarkozy/Taubira/ Hollande, l'élection municipale avec les partis ou candidats impliqués (FN/ tour/liste/conseil), et la question des impôts. Il est en outre assez significatif de constater que ces quatre thématiques sont indépendantes l'une de l'autre (figure 12).

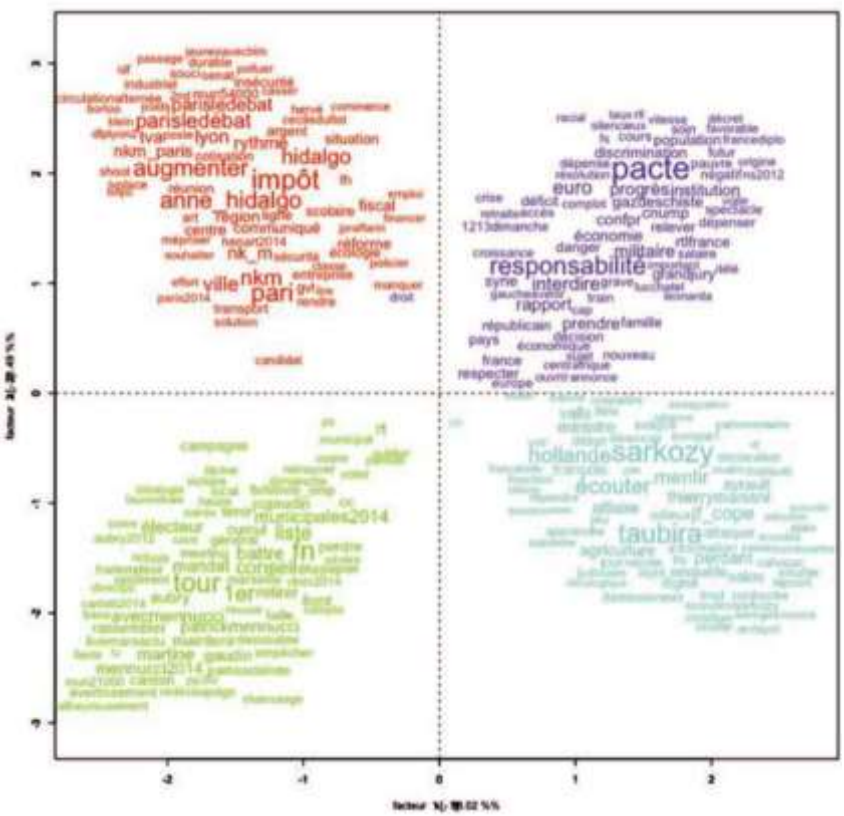


Figure 12. AFC relative à la classification descendante (sous-corpus négatif)

Il est donc intéressant de noter que lorsqu'un tweet fait intervenir une forme de négation (syntaxique, sémantique, énonciative), ce sont des thématiques bien spécifiques, notamment contextuelles et liées aux événements, qui sont mises en discours.

Bilan des études successives

Ces trois analyses successives ont montré plusieurs points intéressants pour répondre à l'hypothèse d'une corrélation entre la forme linguistique des tweets et l'information qu'ils véhiculent, même si elles sont fondées uniquement sur des méthodes fréquentielles et statistiques, et mériteraient d'être complétées par d'autres études. Nous pouvons néanmoins dégager :

— l'importance du terme « politique » dans les sous-corpus idéologique et négatif ;
— l'importance du terme « France » dans les sous-corpus efficace et négatif ;
— la disparité des thématiques identifiées dans ces trois sous-corpus : celles-ci sont plus polémiques dans les tweets efficaces, et contextuelles dans les tweets négatifs.

Cela peut expliquer la manière dont les thématiques se manifestent dans les différents sous -corpus : elles sont relativement indépendantes dans les tweets efficaces et négatifs, mais plus mêlées dans les tweets idéologiques.

Cela invite à penser que les différentes thématiques mises en discours ne le sont pas de la même manière. Sont ainsi confirmés un acquis de la sémantique textuelle, à propos de la corrélation entre le fond et la forme, et la nécessité de porter attention aux formes de sémiotisation opérées grâce aux marqueurs linguistiques.

*
* *

Pour contribuer à la question du partage d'informations dans les réseaux sociaux, nous avons proposé une analyse statistique de sous-corpus dis-tincts, identifiés par des critères linguistiques notamment. Ces analyses ont permis de montrer que malgré quelques similitudes lexicales entre ces sous--corpus, les thématiques qui sont développées sont relativement diffé-rentes : thèmes plutôt conceptuels de la politique, thèmes polémiques, ou thème contextuels. Cela nous conforte dans l'idée de créer des interactions entre les sciences du langage et les sciences de l'information, puisque la forme des tweets entretient des rapports avec l'informativité de ces tweets. Le partage de l'information se retranscrit donc linguistiquement, et la mise en mots contribue à véhiculer certaines thématiques. Cette hypothèse qui semble être confirmée devra être approfondie dans des travaux ultérieurs, notamment en spécifiant les analyses avec d'autres types de sous-corpus.

Références bibliographiques

BEAUD Stéphane et WEBER Florence, 2012. Guide de l'enquête de terrain, 4e édition augmentée, Paris, La Découverte.

DJEMILI Sarah et al., 2014. « What does Twitter have to say about ideology? », in G. Faaß et J. Ruppenhofer (dir.), NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication/Social Media– Pre-conference workshop at Konvens 2014, Hildesheim, Universitätsverlag Hildesheim, vol. 1 : 16-25.

LONGHI Julien, 2013. « Essai de caractérisation du tweet politique », L'information grammaticale, no 136 : 25-32.

LONGHI Julien, 2015. « La théorie des objets discursifs : concepts, méthodes, contributions », mémoire d'habilitation à diriger les recherches, université de Cergy-Pontoise.

LONGHI Julien, 2017. « Le corpus Polititweets : enjeux institutionnels, juridiques, techniques et philologiques », in C. Wigham et G. Ledegen (dir.), Corpus de communication médiée par les réseaux. Construction, structuration, analyse, Paris, L'Harmattan : 37-50.

LONGHI Julien et al., 2014. « Polititweets, corpus de tweets provenant de comptes politiques influents » [en ligne], in T. Chanier (dir.), Banque de corpus CoMeRe, Nancy, Ortolang.fr.

LONGHI Julien et al., 2016. « Extraction automatique de phénomènes linguistiques dans un corpus de tweets politiques : quelques éléments méthodologiques et applicatifs à propos de la négation », in E. Hilgert et al. (dir.), Res per Nomen V. Négation et référence, Reims, Éditons et Presses universitaires de Reims : 517- 535.

MARCHAND Pascal et RATINAUD Pierre, 2012. « L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011) » [en ligne], in A. Dister et al. (dir.), Actes des 11es Journées internationales d'analyse statistique des données textuelles – JADT 2012, Liège, Paris, ILPGA (Lexometrica) : 687-699, disponible sur <http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm> [dernière consultation juin 2017].

SARFATI Georges-Elia, 2014. « L'emprise du sens : note sur les conditions théoriques et les enjeux de l'analyse du discours institutionnel », in J. Longhi et G. -E. Sarfati (dir.), Les discours institutionnels en confrontation. Contributions à l'analyse des discours institutionnels et politiques, Paris, L'Harmattan : 13-46.

RASTIER François, 2004. « Enjeux épistémologiques de la linguistique de corpus » [en ligne], Texto !, juin, rubrique « Dits et inédits », disponible sur <http://www.revue-texto.net/ Inedits/Rastier/Rastier_Enjeux.html> [dernière consultation juin 2017].

RASTIER François, 2008. « Sémantique du web vs semantic web ? Le problème de la pertinence » [en ligne], Texto !, vol. XVIII, no 3, disponible sur <http:// www.revue-texto.net/docannexe/file/1729/ rastier_web_semantique.pdf> [dernière consultation juin 2017].

Ressources

Iramuteq : <http://www.iramuteq.org>.

Lexico3 : <http://www.tal.univ-paris3.fr/ lexico/>.

Tropes : <http://www.tropes.fr>.

Tutoriels pour l'analyse textométrique [Tutoriels] par André Salem : <http://lexicometrica.univ-paris3.fr/numspeciaux/special8/tutoriel1.pdf>.

Wiki TEI : <http://wiki.tei-c.org/index.php/ Main_Page>.