Du corpus réflexif au corpus réfléchi : La plateforme #Idéo2017 pour extraire contextuellement les pratiques citationnelles et analyser la circulation des discours politiques sur Twitter

Julien Longhi

DU CORPUS REFLEXIF AU CORPUS REFLECHI : LA PLATEFORME #IDEO2017 POUR EXTRAIRE CONTEXTUELLEMENT LES PRATIQUES CITATIONNELLES ET ANALYSER LA CIRCULATION DES DISCOURS POLITIQUES SUR TWITTER

Julien LONGHI
Université de Cergy-Pontoise, laboratoire AGORA (EA7392),
Institut universitaire de France (IUF)

Introduction

L'analyse des réseaux sociaux numériques mobilise des approches scientifiques, voire des positionnements académico-institutionnels, différents, qui pourraient schématiquement être regroupés en deux catégories : « extraction vs contextualisation » (Paveau 2013). Bien sûr, les tenants de chacune des positions valorisent leurs pratiques, en définissant parfois de manière un peu caricaturale celle des autres. Par exemple, Paveau 2013 écrit que « les études existantes en français traitent souvent les énoncés sur les réseaux avec les concepts et outils de l'analyse du discours hors ligne, et travaillent sur le discours de manière logocentrée », précisant que « les énoncés sont extraits des environnements numériques et présentés traditionnellement sous forme de liste ou d'énoncé individuel, leur matière verbale étant seule prise en considération ». Ce type d'approche valorise « la capture d'écran », et renvoie à une certaine « conception du discours ». D'un autre côté, ces travaux, basés sur des captures d'écran, choisies par les chercheurs comme des exemples « exemplaires », destinés à illustrer un propos, peuvent être critiqués par les tenants de l'« extraction » pour leur manque de représentativité, les difficultés à généraliser les conclusions, ou simplement les difficultés méthodologiques de choix des données et la validation des résultats.

Dans la mesure où les deux approches présentent à la fois des intérêts et des limites, et plutôt que de situer le débat d'un point de vue théorique, dogmatique, ou encore partisan, nous souhaiterions illustrer la possibilité offerte par la plateforme #Idéo2017[1] de combiner extraction et contextualisation. Développée dans le cadre de l'élection présidentielle française en 2017, cette plateforme offrait[2] au grand public un moyen d'accéder aux tweets produits par les comptes officiels des 11 candidats, ainsi que des fonctionnalités d'analyses issues de la textométrie. Dans un

[1] La plateforme est accessible via ce lien : http://ideo2017.ensea.fr/plateforme/

[2] Une autre plateforme a ensuite traité des législatives, à partir des comptes Twitter des principaux partis ; puis la plateforme #quinquennat a prolongé ces possibilités, avec les comptes des principales personnalités du début du quinquennat.

premier temps, nous présenterons quelques enjeux en lien avec la constitution et l'analyse de corpus en contexte numérique, notamment dans la perspective de saisir la circulation des discours et la prise en compte du discours rapporté (DR) ou partagé (DP). Nous présenterons ensuite certaines spécificités de la plateforme #Idéo2017, puis nous présenterons quelques exemples qui illustrent l'intérêt des choix techniques opérés pour l'analyse du discours rapporté et de la circulation de la parole en contexte politique.

1. Extraire ou contextualiser le discours numérique : des nouvelles pratiques d'analyse pour des nouveaux observables

Comme rappelé en introduction, différentes approches existent en matière de recherche sur les discours numériques. Les approches dépendent notamment des ancrages scientifiques qui sont convoqués par les chercheurs : interfaces et dispositifs pour les sciences de la communication, interactions et pratiques langagières pour les analystes de discours ou interactionnistes, thématiques et spécificités statistiques pour l'analyse des données textuelles, etc. Ces approches conditionnent également le rapport aux « données » : à propos de tweets politiques (produits par les comptes officiels des candidats à une élection, par exemple), est-on face à des discours (voire des « technodiscours »), face à des données (éventuellement textuelles) que l'on peut regrouper en corpus ? ou même face à des « data » que des algorithmes peuvent traiter de différentes manières pour dégager des tendances, des comportements, ou des évolutions ? Le travail du chercheur n'est pas le même selon les réponses que l'on apporte à ces questions, et s'il n'existe pas selon nous de bonne ou de mauvaise réponse, il s'agit de trouver une voie qui permette à la fois d'appliquer des traitements statistiques de nature à répondre à la problématique du projet de recherche (fournir aux citoyens des outils objectifs les aidant à appréhender les discours politiques en contexte électoral) et aux exigences imposées par le matériau complexe qu'est le tweet politique (Longhi 2013, 2017).

1.1 Discours rapporté ou « technodiscours » rapporté : quelques clarifications

La première clarification concerne la caractérisation spécifique du discours numérique comme « technodiscours » (Paveau 2014 en ligne et mis à jour par la suite), et l'établissement de concepts nouveaux forgés pour pallier aux insuffisances des analyses « logocentrées ». Le technodiscours rapporté est défini comme suite (Paveau 2014 : en ligne) :

> Le technodiscours rapporté est une forme numériquée (Paveau 2015 [2013]) de discours rapporté, ce dernier étant défini comme « opération métadiscursive de représentation d'un acte d'énonciation par un autre acte d'énonciation » (Authier-Revuz 2001 : 192). Dans l'opération de technodiscours rapporté, le dispositif du discours citant / discours cité, fondateur des descriptions traditionnelles du discours rapporté hors ligne (Authier 1992-1993, 2001, Rosier

2008), et maintenu à propos des corpus en ligne (von Münchow 2004), est en partie ou totalement pris en charge par un outil technologique (Paveau 2015).

Paveau (*ibid*.) explique en outre que « les paroles d'autrui, produites en un temps t et un espace e1 du web 2.0, sont rapportées en un temps t+1 sur un espace e2, via des outils de partage de contenu, activés la plupart du temps par des technosignes », et que la distinction énonciative du discours rapporté hors « est assurée en partie par le dispositif technologique », concluant que « le technodiscours rapporté est une forme de discours rapporté native du web ». Dans la même voie, Bibie-Emerit (2015) résume que la notion de technodiscours rapporté « oblige à resituer le discours numérique natif dans les réseaux sociaux » car « il n'existe pas dans le web 2.0 de lieu étanche au partage de contenu et les internautes, au travers du technodiscours rapporté mais aussi au travers de leurs traces numériques, passent d'un réseau social à l'autre continuellement et entrainent les autres dans ce mouvement » (p.188).

Considérant ces propositions, nous retenons donc qu'il faudra prendre en considération les spécificités technologiques des réseaux, Twitter dans notre cas, pour analyser le matériau langagier, et pour cet article le discours rapporté et la circulation de la parole. Néanmoins, nous pensons qu'il est possible de se doter d'une conception du corpus, et plus généralement de la production scientifique, qui permette de garder la trace de ces spécificités, tout en développant dans le même temps des analyses outillées. Celles-ci fournissent des hypothèses, font ressortir des régularités invisibles « à l'œil nu », et donnent donc au chercheur des informations de nature à élargir son point de vue initial. De manière plus anecdotique (nous ne pourrons pas développer cela ici), nous considérons également que le vocabulaire traditionnel de l'analyse du discours suffit à mener des analyses discursives sur les corpus issus du web (et il n'est pas nécessaire de préfixer tous les termes de *techno*-) puisque les mécanismes, même augmentés d'une dimension technique, conservent un fonctionnement appréhendable du point de vue des catégories de l'énonciation notamment.

1.2 Discours rapporté, discours partagé, hypertexte

Bien que notre propos soit plutôt orienté sur le discours numérique, et qu'il aborde comme cas d'étude la circulation des discours et la reprise de propos en contexte politiques, il est nécessaire d'apporter quelques clarifications sur la thématique du discours rapporté. Nous nous référons notamment aux travaux de Simon (2017), Grossmann et Rosier (2018) et Grossman (2019).

D'un point de vue général, les discours hypertextualisés auxquels nous pouvons être confrontés dans les messages (tweets) que nous analyserons, « soulèvent deux questionnements » dont l'articulation permet de mesure la portée discursives des phénomènes à analyser : « la problématique de renouvellement des pratiques d'écriture et de lecture des discours numériques et la problématique énonciative et argumentative d'influence de ces pratiques, tournée vers la mobilisation stratégique de discours autres » (Simon 2017 en ligne). La lecture de ces discours numériques doit donc être

prise en compte dans la méthode d'analyse que nous proposons, et l'influence possible de ces discours doit pouvoir être mesurée, et analysée objectivement.

Concernant la circulation des discours sur les réseaux, nous suivons Grossman et Rosier (2018) qui considèrent « les rapports entre DR et hypertextualité […] dans une approche continuiste des pratiques numériques qui reconfigurent des pratiques et des genres anciens ». Pour eux, le DR a « toujours été lié à des modifications technologiques ». Ils mobilisent les trois hypothèses terminologiques suivantes, qui distinguent :

1) Rapporter : « l'activité signalante de la distinction d'un discours citant et d'un discours cité », constatant que « la citation se porte bien sur la toile et participe de l'éthos des internautes » ;

2) Partager : « l'intention pragmatique et la visée du discours, à dimension argumentative (témoignage, preuve…). Ce partage indique le statut discursivo-social de l'énonciateur.e rapportant à la façon (c'est une hypothèse) d'un marqueur évidentiel qui atteste à la fois de sa légitimité à faire circuler des discours et de sa légitimité en tant qu'énonciateur.e en marquant la manière variée (perception, inférence, indice) dont il a eu accès à la source de ce qu'il/elle diffuse » ;

3) Augmenter : « configuration numérique qui associe un DR et une pratique de DP. Un DR minimal partagé, sans être commenté, de par sa circulation est déjà en soi augmenté ».

Cette notion de discours rapporté (DR) adoptée est très ouverte, et définie comme « l'ensemble des procédés permettant de signaler, d'introduire un discours, écrit ou oral ou polysémiotique, émis par un énonciateur différent de l'énonciateur principal ». Nous nous inscrirons dans cette voie, en prenant la circulation des discours politiques sur Twitter d'une manière ouverte. Grossmann (2019) précise que le terme polysémiotique « rend compte du fait que dans l'univers numérique d'aujourd'hui, le discours rapporté ne concerne plus seulement le texte écrit, mais très souvent, des images, du son, des vidéos » (ce qui sera le cas dans les exemples que nous présenterons). Un point important à garder à l'esprit dans le développement de la méthode que nous présenterons est que la notion de discours rapporté « ne doit pas être essentialisée, non seulement parce qu'elle recouvre des phénomènes très différents, mais aussi parce que son invention même et son histoire compliquée, telle qu'elle a été retracée par Rosier (1999), l'enracine dans des pratiques discursives spécifiques et répond à des besoins hétérogènes ». Il conclut son étude sur le discours rapporté partagé sur la toile en le définissant « comme un plurisystème complexe, qui fonctionne « en 3 D » [qui] associe un discours primaire, qui mobilise tous les types du DR classique, à des discours secondaires (insérés ou mis en arrière-plan) qui peuvent eux-mêmes comporter des éléments citationnels. Cette combinaison permet également d'articuler étroitement deux fonctions principales : une fonction de représentation du discours d'autrui […] ; une fonction évidentielle, exercée le plus souvent par le texte ou l'extrait inséré, mis en arrière-plan ou indexé ».

La conception du corpus que nous allons adopter devra donc nous permettre de prendre en compte la circulation du discours, les processus énonciatifs à l'œuvre dans leur production comme leur lecture, et la mesure de leur influence.

1.3 Constituer des corpus réflexifs : une première alternative

Pour constituer un corpus qui prenne en compte les spécificités du web 2.0, les propositions de Mayaffre (2002 : en ligne) sur les corpus *réflexifs* (qu'il positionne « entre architextualité et hypertextualité ») sont intéressantes :

> C'est ce co-texte qui doit, autant que faire se peut, désormais se trouver intégré dans le corpus lui-même. Ou, autrement dit, les macro-corpus en embrassant la plupart des discours ou textes d'un sujet donné, d'un locuteur donné, d'une période donnée compteront automatiquement le co-texte des textes qui le composent : le co-texte des textes du corpus sera le corpus. L'avantage est évident. Il ne sera plus nécessaire de sortir du corpus pour comprendre et interpréter ses composants. Et l'analyse contextualisée ou cotextualisée de chacun des textes se fera grâce à une navigation interne au corpus et non sur la base de ressources extérieures arbitrairement et subitement convoquées.

Il explicite qu'ainsi « corpus et archive pourront se confondre en grande partie », et que le « travail même d'archive sera partie intégrante du travail de saisie et de constitution du corpus ». Cette conception du corpus est particulièrement intéressante car, si on ne peut certes pas épuiser l'absorption de pages web, de posts, de tweets, etc., lors de la constitution d'un corpus, on peut entrevoir la possibilité de collecter les liens, les réponses, les références, etc., afin d'enrichir le discours initialement ciblé de tous les discours qui l'environnent, concrètement ou potentiellement. C'est pour cela que Mayaffre explique que « les corpus réflexifs devront être organisés techniquement comme des *hypertextes* : chaque texte constituant devra être relié aux textes considérés comme parents », et propose des pistes d'encodage « sous la norme SGML (Standard Generalized Markup Language) et ses applications HTML (Hyper Text Markup Language) ou XML (Extensible Markup Language) », qui « apparaissent *a priori* comme les plus simples et les plus universels pour créer ces liens hypertextuels sur de grosses bases de données, tout en permettant un traitement lexicométrique traditionnel à un niveau de granularité plus fin (habituellement le mot) ». Cette réflexion à propos des grandes bases de données textuelles peuvent être transposées dans le cadre du web 2.0 et des réseaux sociaux numérique, même si la réflexivité est augmentée dans ces environnements avec l'adjonction potentielle de nouveaux éléments.

Ces indications avaient déjà été mises en œuvre dans l'élaboration du corpus *Polititweets* (Longhi *et al*., 2014). Ce corpus a en effet été constitué

au format XML-TEI, et déposé sur la plateforme *Ortolang*[3]. Néanmoins, nous avons pu constater une difficulté dans la communauté pour l'utilisation de ce type de corpus pour les analystes du discours, même ceux qui utilisent des logiciels de textométrie. Le format XML ne semblait en effet pas partagé par tous les chercheurs, et la conversion de ces documents en formats importables dans des outils n'était pas considérée comme évidente. Ainsi, une interface[4], développée dans le cadre d'un stage de M2 sciences du langage de Abdelouafi El Otmani (El Otmani et Longhi, 2016), propose une médiation avec ces fichiers, et donc une prise en main plus aisée par les usagers des logiciels. Cet outil se présente comme un moteur de recherche :

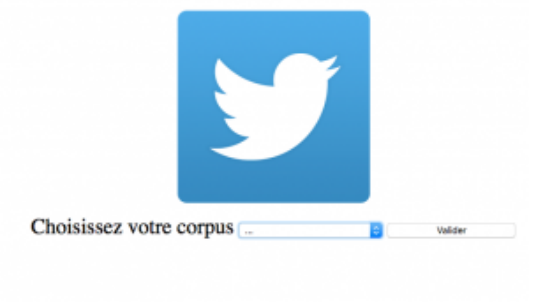

Image 1 : Moteur d'interrogation des corpus

Il convient en premier lieu de choisir le corpus souhaité (d'autres corpus ont été produits par la suite) :

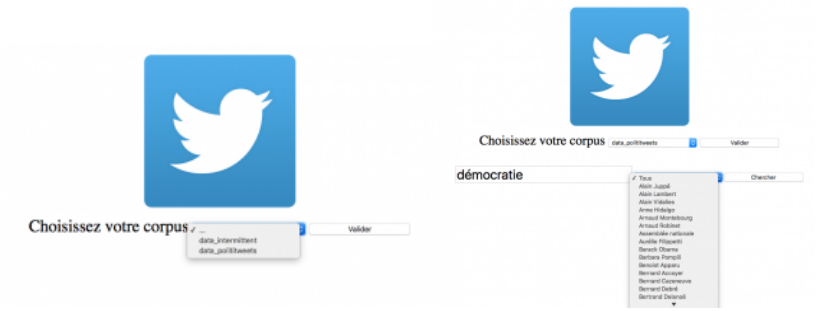

Image 2 : Sélections par métadonnées dans les corpus

Dans notre cas, nous choisissons *Polititweets*. L'utilisateur peut choisir de faire une recherche dans tout le corpus, ou de se focaliser sur un compte twitter spécifique.

L'utilisateur peut ensuite effectuer sa requête, par exemple *démocratie*. En cliquant sur « Valider », les résultats apparaissent : contenu des tweets,

---

[3] Le corpus est accessible via ce permalien : https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html

[4] L'interface permettant de traiter le corpus est accessible sur le site du projet #Ideo2017, via ce lien : http://ideo2017.ensea.fr/outil-twitter/index.php

auteur du tweet, support de production, et nombre de retweets. Ceci est visible sur la capture d'écran suivante :

Rang	Tweet	Auteur	Outil	Retweet
1	Référendum suisse : une bonne nouvelle pour la démocratie ! http://t.co/ZpDXt60w2e #ChVote	Marion Le Pen	HootSuite	56
2	RT @AvecMenucci: .@patricknemucci : "nous avons un problème de démocratie dans cette ville" #Vallier #Menucci2014	M-A Carlotti	Twitter for iPhone	10
3	@LutteFerrari @eturiellel !U E déteste et méprise la démocratie. Vivement le 25 mai!	Dupont-Aignan	Twitter for BlackBerry®	36
4	Les Suisses nous donnent une leçon de démocratie . Maîtriser ses frontières c est maîtriser son destin de peuple libre.	Dupont-Aignan	Twitter for iPad	119
5	La démocratie locale au secours d'une démocratie nationale chancelante? Message centrale des électeurs à un gouvernement sourd et aveugle?	Eric Woerth	Twitter for iPhone	18
6	Voir ou revoir le débat "Trop ou pas assez de démocratie" de la 23e Journée du Livre politique auquel j'ai participé http://t.co/8hJNrHUvFL	Franck Riester	web	4
7	Je participe aujourd'hui à la 23ème Journée du livre politique à l'@AssembleeNat sur le thème "Trop ou pas assez de #démocratie"	Franck Riester	Twitter for iPhone	5
8	"#QAG La théâtralisation est forte. La démocratie,c'est la confrontation mais il doit y avoir des règles. C'est notre responsabilité à tous"	Alain Vidalies	Twitter for iPad	1
9	"Dans une démocratie, on a besoin que le débat ait lieu. Des incidents de ce type nourrissent uniquement l'antiparlementarisme." @itele	Alain Vidalies	Twitter for iPhone	2
10	#DirectAN L'Assemblée nationale adopte la loi sur la formation professionnelle, l'emploi et la démocratie sociale par 52 voix contre 2.	Alain Vidalies	Twitter for Android	7
11	#DirectSénat le #sénat adopte le projet de loi relatif à la #formation professionnelle, à l'#emploi et à la démocratie sociale.	Alain Vidalies	Twitter for iPad	8
12	#urgence #démocratie "@stephanesg: Dimanche un seul vote pour ne pas rendre Grenoble aux acolytes de Carignon : la liste d'@EricPiolle"	François de Rugy	Twitter for iPhone	7
13	#démocratie #qualitédevie #urbanisme #transports #culture #innovation ... Vous avez envie que #Nantes respire ? Dimanche votez écologiste !	François de Rugy	Twitter for iPhone	6
14	RT @TeamGianneslni: Qd le MairePS19e donne des leçons démocratie et ds le même tps n'invite pas les élusUMP aux vœux police #Paris2014 http…	Jean-François Lamour	Twitter for iPhone	8
15	RT @le1SemeavecNKM: .@anne_hidalgo vante les mérites de la démocratie de proximité devant la presse mais ne daigne pas se présenter au cons…	Jean-François Lamour	Twitter for iPhone	7
16	RT @AminneBouabbas: @amarcireau @emmanuelmaurel @fxbourmaud @JeromeGuedj @mulienemann social-libéral vous voulez dire ? La social-démocratie…	MN Liennemann	Twitter for iPad	2
17	Boycott de l Assemblée. Pas d UMP, pas de bruit. La démocratie du vide. L'image du pouvoir. l'insulteur fier comme un torero sans taureau.	Gilbert Collard	Twitter for Android	76
18	Que ceux qui ont bafoué le vote des Français sur la Constitution européenne s'abstiennent de donner des leçons de démocratie: chut !#PS #UMP	Florian Philippot	Twitter for iPhone	222
19	Une pensée pour la démocratie malmenée et pour les agents de la Poste confrontés au quotidien à ces petits sectaires http://t.co/jRqSI1CYN8	Florian Philippot	web	43
20	Montebourg par son mépris pour les Suisses vient de dévoiler tout le mépris de la Caste pour la démocratie et la souveraineté populaire #RTL	Florian Philippot	web	114
21	Bravo la Suisse ! Une vraie démocratie !	Florian Philippot	Twitter for iPhone	167
22	Ce matin, je signe la Charte anticor pour Sevran:-) #transparence #démocratie	Clémentine Autain	Twitter for Android	15

Image 3 : Résultats obtenus et possibilités d'exports

Le menu en haut de la page permet de produire des exports sur mesure pour 2 logiciels d'analyse de données textuelles, *Lexico3* et *Iramuteq*.

En choisissant par exemple *Lexico3*, sans nettoyer les liens, on obtient un corpus qu'il ne reste plus qu'à copier et utiliser pour une analyse dans le logiciel. En faisant de même avec *Iramuteq*, après analyse dans le logiciel, on obtient facilement par exemple l'analyse des similitudes, qui rend compte des cooccurrences de *démocratie* :

Image 4 : résultats obtenus dans l'interface (corpus) ou en utilisant un logiciel (analyse de similitudes)

Cet outil constitue donc un premier pas vers l'application #Idéo2017 que nous allons présenter : il permettait déjà la mise à disposition à la communauté des corpus et de l'interface, consistait en un outil intuitif, et constituait une aide à l'élaboration de corpus balisés grâce à la médiation de l'outil. L'objection d'« extraction » reste néanmoins présente, même si des champs des balises xml, ou dans les exports de l'interface, donnent accès à des métadonnées telles que l'utilisateur, le type de matériel sur lequel le tweet a été produit, ou encore le nombre de retweets.

Pour tenir compte de l'équilibre entre extraction et contextualisation, il faudrait que l'interface développée soit interactive, et permette un retour

aux données natives, et même à l'interface Twitter. C'est ce que propose la plateforme #Idéo2017.

2. La plateforme #Idéo2017 : un corpus réfléchi propice à l'analyse du discours rapporté

L'outil #Idéo2017, élaboré dans le contexte de l'élection présidentielle 2017, est réalisé via plusieurs étapes, comme indiqué dans la Figure 1 : (1) l'extraction de l'ensemble de tweets des candidats, (2) la mise en place d'une sauvegarde des tweets, (3) l'indexation des tweets pour faciliter la recherche dans l'ensemble de tweets, (4) l'application d'un ensemble d'analyses linguistiques sur les tweets, (5) la mise en place d'un moteur de recherche sur l'ensemble de tweets, et (6) l'affichage des résultats sur une page web[5].

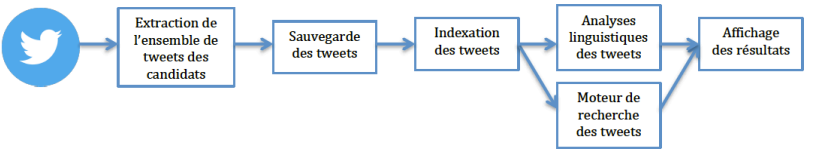

Figure 1 : Chaine de traitement de l'outil #Idéo2017.

Le choix de l'affichage des résultats permet de concurrencer la démarche traditionnelle avec ce type d'outils d'analyse textuelle : extraction de corpus, mise en forme, formatage, balisage, puis usage d'un logiciel. Ici, tout ce travail est réalisé en amont, et l'utilisateur, en cliquant sur les fonctionnalités, a directement accès aux résultats. Comme indiqué sur l'image suivante, qui reproduit la page d'accueil de la plateforme, trois volets sont proposés à l'utilisateur :

[5] L'objectif de rendre les résultats compréhensibles à un public qui n'a pas de connaissances en linguistique n'est pas facile, et nous avons détaillé ailleurs (Longhi 2018) le travail d'accompagnement effectué en parallèle du développement de l'interface : rédaction de billets de blog (sur *The Conversation* et *Huffington Post*) ; réalisation, avec l'aide de projets étudiants, de vidéos décrivant l'utilisation de la plateforme ; documentation du site avec les liens vers les articles en lien avec le projet (déposés dans HAL). Certains de ces contenus détaillaient en outre les fonctionnalités/algorithmes utilisés. En parallèle, un travail de diffusion des savoirs linguistiques et des pratiques d'analyse ont été diffusées dans une séquence du MOOC « Usages du Web », dans le chapitre 8 « Appréhender les enjeux du numérique ». Ce cours a été largement diffusé grâce à sa mise en ligne sur FUN MOOC : https://www.fun-mooc.fr/courses/course-v1:u-cergy+156002+session02/about.

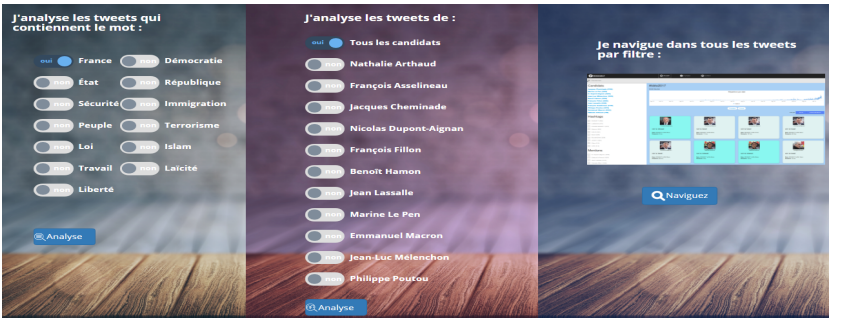

Image 5 : Page d'accueil de la plateforme #Idéo2017

L'analyse *"J'analyse les tweets qui contiennent le mot…"* permet à l'utilisateur de choisir un mot parmi les 13 mots[6] qui sont souvent employés dans les débats politiques (Alduy 2017). Cette entrée donne accès à quatre analyses possibles : l'usage de ce mot par les différents candidats (sur/sous-emploi et fréquences de la forme exacte), les mots associés à ce mot (analyse de similitudes, basée sur les cooccurrences entre les mots), l'emploi de ce mot et ses dérivés par les différents candidats (analyse basée sur les racines des mots : par exemple *islam*, *islamisme*, *islamiste*, seront regroupés sous la forme /islam/), et le nuage de mots. Ces analyses sont en fait des résultats produits grâce au code du logiciel d'analyse textuelle *Iramuteq*, issus de calculs qui portent dans le logiciel des noms plus techniques (l'*analyse de similitude* devient par exemple *les mots associés à ce mot*).

L'analyse *"J'analyse les tweets de [candidat]"* permet à l'utilisateur de choisir un candidat parmi les 11 candidats (ou le corpus global des 11 candidats) afin d'analyser ses tweets via les techniques suivantes : les mots les plus utilisés, les thématiques (issues de *la méthode Reneirt*), les relations entre les mots, le nuage de mots, les spécificités des différents candidats (possible si l'utilisateur a choisi d'analyser tous les candidats en même temps).

Enfin, le moteur de recherche permet à l'utilisateur de faire des recherches sur toute la base des tweets, grâce à un outil appelé *ElastricSearch*. Il permet aussi de récupérer les tweets sous forme de vignettes, cliquables, et dont le lien permet de retourner au message original dans Twitter. C'est donc le moyen de retourner aux données natives, que l'on peut ainsi « contextualiser ».

Si nous recherchons par exemple les tweets de Philippe Poutou, nous pouvons avoir la liste de tous les tweets (742) postés (entre le 1er novembre de 2016 et la date de consultation, l'extraction étant interrompue à l'issue de l'élection), en format texte, ainsi que des analyses statistiques fondées sur ce corpus :

---

[6] Le terme *mot* a été choisi car il semblait plus compréhensible, pour le grand public, que des termes issus des sciences du langage, comme *lexie* par exemple. Aussi, nous continuerons dans cet article à utiliser *mot*.

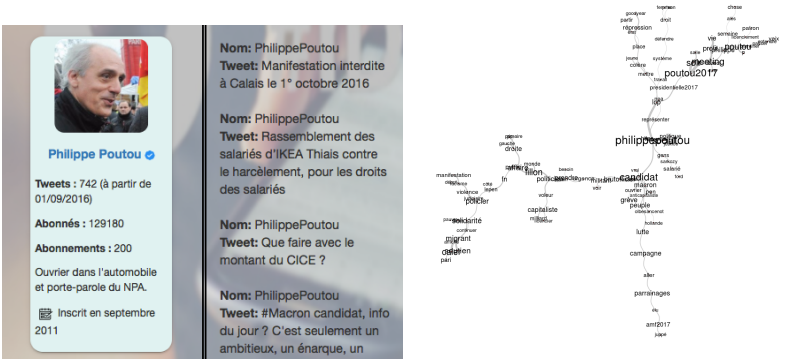

Image 6 : Résultats offerts par #Idéo2017

Certes, on peut considérer cette pratique comme une simple extraction, qui témoignerait d'une approche logocentrée. Néanmoins l'utilisateur/trice a ensuite la possibilité de recontextualiser le discours, et de le situer tel qu'il a été produit. Par exemple, considéons qu'il/elle s'intéresse à ce message :

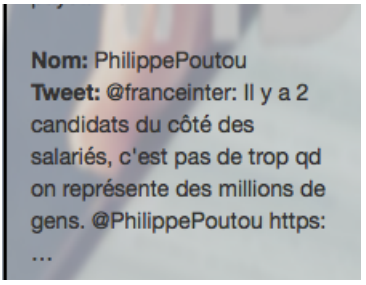

Image 7 : visualisation d'un message potentiellement intéressant

C'est un tweet de @franceinter retweeté par Philippe Poutou (donc un tweet circulant, pour rejoindre la thématique du volume). L'utilisateur/trice peut alors aller dans la partie « moteur de recherche », pour retrouver le tweet en question :

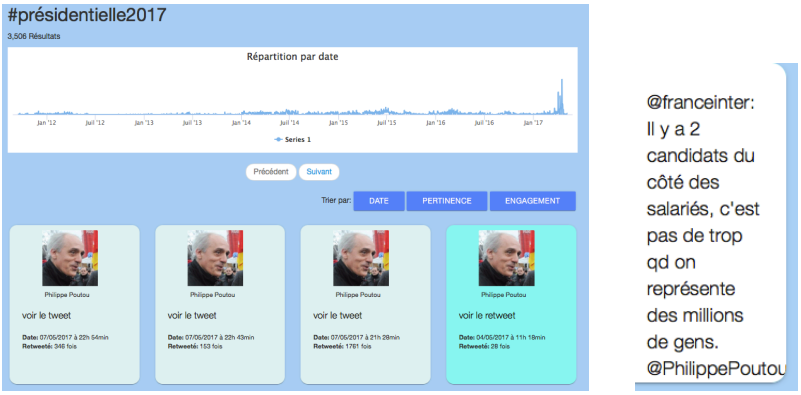

Image 8 : exploration par le moteur de recherche et obtention du résultat

Il/elle peut ensuite l'ouvrir, par un simple clic, dans l'interface de Twitter, soit dans son environnement natif :

Image 9 : prise en compte du message dans l'environnement Twitter

On peut alors retrouver les caractéristiques des approches « contextualisantes » décrites par Paveau 2013 (« *contextualisation technorelationnelle* », « *investigabilité du discours* », « *technoconversationnalité* »). On voit sur cette image le lecteur de vidéo, qui propose une interview (le contenu textuel du tweet, le discours rapporté, est en fait un extrait de cette vidéo, qui est une interview de Philippe Poutou). Mais plus encore que les termes conceptualisant l'approche, ce qui nous intéresse ici, c'est le va-et-vient rendu possible entre les deux modes d'appréhension. En effet, l'analyse contextualisée nous semble ici trouver un intérêt car elle s'est focalisée sur un message qui contenait le mot *salarié*, élément visiblement saillant dans le corpus de Philippe Poutou (repérage statistique et en collocations). Si l'accès au corpus se faisait uniquement par l'interface de Twitter, nous n'aurions pas d'informations sur la spécificité de ce terme pour le candidat en question, la manière dont il en fait usage, comment il se distingue de ses concurrents par le sens qu'il lui accorde. Le discours rapporté à l'ère du numérique ne doit selon nous pas être analysé avec des nouveaux concepts, qui « augmentent » (avec un *techno*-) les concepts traditionnels de l'analyse du discours, mais être analysé avec les concepts traditionnels selon une nouvelle manière de suivre les observables. S'il est impossible d'épuiser l'analyse d'un fait de discours, qui va sans cesse être repris, commenté, cité, modifié (si le compte change d'avatar, si les fonctionnalités changent), il est néanmoins possible, et c'est l'enjeu de notre recherche, de fournir des moyens d'accès à ces observables, et des représentations intelligibles de ces particularités.

Conclusion : Des nouveaux observables aux nouveaux résultats de la recherche, comment repenser la « production » des savoirs ?

Les propositions formulées dans cet article sont de nature à rendre saillants des sujets d'analyse non perceptibles à l'œil nu, et de motiver l'étude de certains phénomènes, donc la complexité ne serait pas forcément apparente à première vue. Un des intérêts de la démarche qui articule des approches dites quantitatives et des approches dites qualitatives est donc l'identification des objets discursifs pertinents, pour pouvoir discrétiser, dans l'ensemble des débats ou problèmes, les objets discursifs dont l'analyse pourra apporter des éclairages inédits et nécessaires. Certes les analyses quantitatives conduisent à l'effacement d'une quantité considérable de ressources sémiotiques (avatars, médias partagés dans les tweets, interactions, chronologies), mais nous avons montré qu'il est ensuite possible de les réinscrire dans l'enrichissement des données, par le biais d'un retour dans les documents originaux, ou encore l'intégration des métadonnées (par l'intermédiaire de l'usage du moteur de recherche : soit directement par les métriques proposées, soit indirectement par le retour à l'interface Twitter). Ces enrichissements permettent en quelque sorte de contextualiser les données : la contextualisation des données est le « principal "garde-fou" contre toutes les généralisations hâtives ou indues » (Beaud et Weber, 2012, p.217). En ce qui concerne la circulation des discours dans les espaces numériques, et la question du discours rapporté, il faut prendre en compte la complémentarité des deux démarches, qui indépendamment ne peuvent pas rendre compte de la réalité discursive, mais qui peuvent se compléter pour permettre d'appréhender leur nature dans ces espaces en perpétuelle modification. C'est donc par les « corpus réfléchis » (un corpus qui peut effectuer un « retour sur lui-même », soit dans sa matérialité formelle, soit dans son environnement natif) que l'on peut concevoir un nouveau moyen d'analyse du discours numérique, et en particulier du discours rapporté à l'ère du numérique. L'intérêt d'une telle plateforme est aussi, pour l'étude du discours rapporté, de tenir compte du constat d' « un principe général : celui de la porosité des énonciations, entre narration et dialogue, entre formes du DR qui alternent dans les mêmes énoncés » (Rosier 2008, p.16).

Bibliographie :

Alduy, C. (2017) : *Ce qu'ils disent vraiment. Décoder le discours des présidentiables*, Paris : Seuil.

Beaud S., Weber F. (2012) : *Guide de l'enquête de terrain*, Paris : Éditions La Découverte.

Bibie-Emerit, L. (2015) : *Description du discours numérique : étude des bouleversements linguistiques du web 2.0 au travers de l'exemple des*

souhaits d'anniversaire sur Facebook, Thèse de doctorat, Université Michel de Montaigne - Bordeaux III, <tel-01442467>.

El Otmani, A., Longhi, J. (2016) : Outil d'analyse de tweets : http://ideo2017.ensea.fr/outil-twitter/index.php

Grossmann, F. (2018) : « Discours rapporté versus discours partagé : convergences, différences, problèmes de frontières », Conférence invitée dans le cadre du colloque Ci-dit, Université libre de Bruxelles. Le discours rapporté à l'ère numérique : du discours cité au discours partagé, Université Libre de Bruxelles, Jun 2018, Bruxelles, Belgium, hal-02004746.

Grossmann, F., Rosier, L. (2018) : « Du discours rapporté au discours partagé. Analyser les usages du discours rapporté hypertextualisé ». In Simon J. (éd), Le discours hypertextualisé. Espaces énonciatifs mosaïques, Presses universitaires de Franche-Comté, collection Annales littéraires (version en ligne consultée sur Academia).

Longhi J. (2018) : « Le projet #Idéo2017 : Quelles implications du/de la chercheur-e en tant qu'acteur-trice potentiel du changement social ? Exemplification à partir du discours politique numérique », *Cahiers de linguistique*, n°44/2, p.103-116.

Longhi J., Marinica C., Borzic B., Alkhouli A. (2014) : *Polititweets,* corpus *de tweets provenant de comptes politiques influents*. In Chanier T. (éd), Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets- tei-v1]

Longhi, J., Marinica, C., Hassine, N., Alkhouli, A., Borzic, B. (2017): "The #Idéo2017 platform", 5th conference CMC and Social Media Corpora for the Humanities, Bolzano, Italy, 3rd and 4th October 2017 – Conference proceedings: https://cmc-corpora2017.eurac.edu/proceedings/cmccorpora17-proceedings.pdf

Mayaffre, D. (2002) : « Les corpus *réflexifs* : entre architextualité et hypertextualité », *Corpus* [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003, consulté le 08 janvier 2019. URL : http://journals.openedition.org/corpus/11

Paveau, M.-A. (2013) : « Analyse discursive des réseaux sociaux numériques », *Dictionnaire d'analyse du discours numérique, Technologies discursives*, [Carnet de recherche], http://technodiscours.hypotheses.org/?p=431, consulté le15/09/2017

Paveau, M.-A. (2014) : « [Dictionnaire] Technodiscours rapporté », *Technologies discursives*, 31/12/2014, https://technodiscours.hypotheses.org/606.

Rosier, L. (2008) : *Le discours rapporté en français*, Paris, Ophrys, « L'essentiel français ».

Simon, J. (2017) : « Présentation », *Semen* [En ligne], 42 | 2017, mis en ligne le 24 août 2017, consulté le 09 août 2019. URL : http://journals.openedition.org/semen/10607.