CORPUS

Corpus

20 | 2020
Corpus complexes

Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe

Exploring corpus of tweets: from computer processing to complex discursive analysis

Julien Longhi

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/corpus/4567
DOI: 10.4000/corpus.4567
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Julien Longhi, « Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe », *Corpus* [Online], 20 | 2020, Online since 31 January 2020, connection on 08 September 2020. URL : http://journals.openedition.org/corpus/4567 ; DOI : https://doi.org/10.4000/corpus.4567

Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe

Exploring corpus of tweets: from computer processing to complex discursive analysis

Julien Longhi

1 Dans cet article, nous présentons les acquis et développements de plusieurs projets de recherche successifs et complémentaires, menés depuis 2013, ayant pour objectif l'analyse d'un type particulier de données CMC (Computer-mediated communication) : les tweets politiques. Après avoir rapidement présenté ce genre de discours, ses caractéristiques, et les problématiques de recherche entrevues, nous détaillerons la progression construite dans la caractérisation et l'exploitation de ce genre ; l'appréhension et la constitution de ces données sociales en corpus ; la production de résultats scientifiques, et la mise en place de différents types d'exploration de corpus. Nous reviendrons ainsi sur les enjeux de l'exploration de corpus numériques, en prenant en compte à la fois les questions d'acquisition et de pérennisation des corpus, mais aussi des méthodes, outils, et approches pour aborder de tels corpus.

1. De l'analyse des tweets politiques à la constitution d'un corpus de tweets

2 Notre travail sur les tweets politiques est né de la coïncidence de plusieurs facteurs, dont la présentation est importante pour contextualiser à la fois les recherches menées, mais aussi le positionnement théorique, méthodologique, et pratique :

- Dans le cadre d'activités pédagogiques au département Métiers du multimédia et de l'internet de l'IUT de Cergy-Pontoise, la prise en charge d'un module « Ecriture pour les médias numériques » nous invitait à travailler avec les étudiants sur la production de contenus sur différents médias, du blog aux réseaux sociaux. La pratique de ces réseaux était

alors nécessaire, et leur traitement d'un point de vue pédagogique incitait à une caractérisation factuelle des différents supports ou réseaux, puisqu'il n'existait pas de supports pédagogiques sur lesquels s'appuyer ;

- Un intérêt d'un point de vue textuel pour des genres hybrides[1] nous avait conduit à présenter au CMLF 2012 un article sur les « tweets de Mouloud » (Longhi 2012), qui était une rubrique d'un magazine féminin (*Be*) dans laquelle Mouloud Achour proposait des descriptions de l'actualité sous des formes brèves, proches des tweets (mais dans un espace papier) ;
- La rédaction de billets de blog pour une série d'été du *Huffington Post*, « Dis-moi ce que tu tweetes », grâce à l'entremise d'un collègue auprès du responsable des blogs suite au CMLF 2012. Ce travail rédactionnel a à la fois permis de se familiariser avec les pratiques d'écriture sur Twitter, mais a aussi permis de devenir acteur dans le champ de l'analyse politique, dans la mesure où les analyses proposées étaient en lien avec des événements discursifs liés à l'actualité.

3 Ceci nous a amené dans un premier temps à travailler sur le tweet politique en tant que voie d'accès au discours politique. Ceci constituait alors le socle de plusieurs recherches, dont la cohérence réside, après une analyse réflexive a posteriori, dans l'appréhension linguistique de données sociales, constituées en corpus, dont l'analyse, par des méthodes mixtes (quali-quantitatives) permet d'appréhender ce « terrain ».

1.1 Le tweet politique comme genre de discours

4 En 2013, dans un article paru dans l'*Information grammaticale* et intitulé « Essai de caractérisation du tweet politique », nous avons proposé une synthèse des observations formulées dans les articles de blog pour le *Huffington Post*, car la récurrence de certains phénomènes nous avait invité à adopter un regard scientifique sur cet objet. Il s'agissait en outre d'en spécifier la caractérisation dans le cadre du champ politique, puisque si certains travaux existaient à l'époque sur le sujet, ils étaient davantage centrés sur les aspects formels (brièveté), ou sur l'importance de l'environnement technologique dans la production de discours numériques (Paveau 2013a).

5 Un point saillant développé dans cet article était la « condensation sémantique par décontextualisation partielle ou techno-contextualisation » : en effet, si les tweets politiques peuvent fonctionner comme des « petites phrases » (ce qui explique le recours des journalistes à ces messages, notamment en les intégrant dans des articles en ligne, comme « citations »), c'est parce qu'ils ont un caractère polémique sur le plan du discours, et parfois « excessif » sur le plan sémantique, avec le choix de mots qui doivent condenser l'information le plus synthétiquement possible (puisqu'à l'époque un tweet ne pouvait faire plus de 140 caractères ; cette limite est maintenant de 280 caractères). Ce processus de condensation sémantique permet de créer des discours, ou simplement d'en reprendre, et d'en figer les caractéristiques. Une spécificité, importante pour la constitution et l'exploitation de corpus, réside dans le fait que les tweets politiques, qui sont très souvent des citations de prises de parole lors d'émissions radio, TV, ou des discours, ne véhiculent pas avec eux le contexte de production de l'énoncé original (s'ils sont « extraits » de leur interface, on perd les éléments temporels notamment, sauf en utilisant les métadonnées fournies par Twitter) ; mais ils le font d'une autre manière, par le biais des hashtags, qui véhiculent une forme de contexte, de manière stratégique. La décontextualisation s'accompagne

d'une techno-contextualisation, grâce à l'usage des hashtags, qui importent avec eux des éléments contextuels, et peuvent créer des interactions avec les tweets liés aux mêmes contextes. En outre, une des caractéristiques habituellement relevée dans le discours politique est la constitution de l'ethos de la personnalité politique en situation. Chauvin-Vileno (2002 : en ligne) indique (notamment en référence aux travaux de Maingueneau) que « ce qui est en jeu avec l'ethos, c'est l'importance de la représentation de soi et de l'autre dans l'interlocution » Il est utile de distinguer un ethos prédiscursif d'un ethos discursif : l'ethos « repose pour une part sur un savoir préalable des interlocuteurs sur la vie, le caractère, les actions du locuteur », et « ce savoir préalable confère ou non du poids au discours, conditionne la réception ». Ce premier type d'ethos est dit prédiscursif, alors que l'ethos discursif se fonde sur la confiance inspirée par l'orateur par l'effet du discours. L'influence des deux ethos peut se combiner dans les situations concrètes et avoir un degré de pertinence variable selon les types de locuteurs et les genres discursifs considérés. Chauvin-Vileno indique aussi que les images des hommes politiques, créées par la presse, la radio, etc. sont produites et prises dans un circuit discursif (au sens large et non strictement verbal). Les tweets sont le lieu de la construction d'un ethos discursif, ils contribuent à constituer pour les candidats une identité numérique, et visent à la faire percevoir dans les meilleures dispositions. Les discours produits s'intègrent dans des stratégies plus larges de communication politique, de réputation en ligne (e-réputation), voire de marketing viral : ainsi, la constitution de l'ethos peut aller jusqu'à la construction de toutes pièces d'un ethos technodiscursif (par le recours aux moyens de communication numérique), avec une nouvelle grammaire du tweet (par l'utilisation de signes spécifiques au réseau, adaptation à des pratiques ritualisées, comportements en lien avec l'efficacité supposée de certains procédés). C'est ce que nous avions détaillé à propos de certaines personnalités politiques, qui ont bien un compte Twitter, mais ne l'utilisent pas, ou très peu. C'était le cas, lors de nos analyses de 2013, pour certains cadres du PS, récents ministres en particulier. Leur identité numérique était alors prise en charge par le compte Twitter du parti, ce qui crée une forme d'hétérogénéité du dire. En linguistique, la théorie néo-dialogique de l'hétérogénéité discursive a notamment renouvelé la conception classique des différentes formes du discours (Authier-Revuz 1984). Du point de vue de l'écriture du tweet, ces cas témoignent de l'appropriation du @ et son utilisation spécifique à des fins communicationnelles précises. En effet, ce signe permet normalement d'établir une mention, en créant un lien hypertextuel vers le compte d'un autre utilisateur. Par usage, c'est aussi devenu le moyen de citer un utilisateur, de lui répondre : le réseau s'est approprié le @ pour désigner bien souvent une identité virtuelle (le profil de quelqu'un, mais pas ce quelqu'un directement). Ce signe a une autre fonction en début de message : il permet de limiter aux *followers* (suiveurs) d'un compte l'envoi d'un tweet (qui reste public, mais apparaît dans un onglet « tweets et réponses », puisqu'il représente une forme d'adresse).

6 Concernant la constitution de corpus, il est donc important de noter que le tweet politique, loin d'être un simple relais supplémentaire de transmission de l'information, s'avère être une forme originale du discours politique, qui pourrait s'apparenter, s'il se stabilise davantage, à un genre de discours spécifique. En effet, de par ses contraintes matérielles et technologiques, il véhicule des formes qui peuvent devenir des « petites phrases »[2] sur le plan discursif, et s'accompagner d'une intensité sémantique originale, par condensation et décontextualisation partielle, voire une recontextualisation par les moyens technologiques (que nous appelons techno-contextualisation). Ces

caractéristiques s'accompagnent, pour l'utilisateur, de la constitution d'un ethos discursif, qui peut se doubler d'un ethos technodiscursif, grâce au maniement de certains codes propres au réseau. Plus encore, par l'insertion d'arrière-plans aux discours, un ethos prétechnodiscursif peut être véhiculé, dans la mesure où les tweets peuvent mobiliser des prédiscours qui développent, pour l'identité numérique mise en scène, un certain nombre d'éléments présentés comme étant préalables à l'énonciation. Le tweet politique est donc un lieu de renouvellement du discours politique, et un cadre intéressant pour observer certaines mutations des formes textuelles, sémantiques et discursives qui sont produites. Pour le faire de manière plus complète, et complexe, nous avons cherché à analyser ces données à travers la constitution d'un corpus représentatif d'un événement politique. Les élections municipales 2014 ont permis d'avoir un contexte propice pour cela (création du corpus *Polititweets*, décrit dans Longhi 2017).

1.2 La constitution d'un corpus de tweets pour l'analyse du discours : contexte et enjeux

7 La constitution d'un corpus de tweets correspondait à deux objectifs, pensés dès 2014 comme des éléments solidaires d'une même recherche : d'une part se doter d'un corpus pour réaliser une recherche centrée sur le lexique politique, à partir d'analyses d'observables issus des nouveaux moyens de communication ; et d'autre part doter le corpus développé dans le cadre du projet CoMeRe[3] de données textuelles qui n'y étaient pas représentées.

8 Le cadre institutionnel était donc à la fois :

9 1. Le projet « Humanité numériques et data journalisme : le cas du lexique politique » :
Ce projet, soutenu par la Fondation UCP, visait à impulser une recherche sur le thème des humanités numériques, et créer des synergies entre plusieurs acteurs de l'université de Cergy-Pontoise. Ce projet transdisciplinaire a nécessité une phase de développement pour permettre d'amorcer des interactions entre des chercheurs en linguistique (linguistique de corpus, analyse de discours, lexique) et des chercheurs en informatique. Comme nous le verrons par la suite, ce cadre pluridisciplinaire a permis d'ouvrir des voies d'exploration de corpus variées, et de pouvoir mesurer les avantages et inconvénients de telle ou telle approche.

10 2. Le projet CoMeRe décrit sur le site du projet comme suit :
CoMeRe avait pour objectif, à l'horizon 2014-2015, de créer un noyau de corpus de communication médiée par les réseaux (*Computer Mediated Communication* – CMC) en français (voir aussi Chanier et al. 2014). Chaque corpus rassemble maintenant un ensemble de conversations intervenant sur la Toile et les réseaux. Ce projet s'intéressait à une variété de systèmes de communication synchrone ou asynchrone, mono ou multimodaux (éventuellement) : blogues, tweets, SMS / textos, courriels, clavardage, forums, etc. Les corpus et leurs métadonnées ont été structurés suivant des formats standards : TEI (*Text Encoding Initiative*), CLARIN, OLAC.

11 Le corpus *Polititweets*, constitué de 34721 tweets provenant de 205 comptes politiques influents, a ainsi intégré cet ensemble de corpus. La banque de corpus est diffusée en accès libre depuis 2014 sur le site Ortolang. Voici le descriptif donné sur la page d'accueil :

> ORTOLANG est un équipement d'excellence validé dans le cadre des investissements d'avenir. Son but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui :
> - permette, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international ;
> - facilite l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels [...] ;
> - valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics.

12 Pour permettre un usage simplifié et communautaire des données, une licence CC 4.0 est attribuée au corpus, et plus particulièrement une attribution CC BY. Cette licence est explicitée ainsi :

> Vous êtes autorisé à :
> Partager – copier, distribuer et communiquer le matériel par tous moyens et sous tous formats
> Adapter – remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.
> L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.

13 Le recours à l'encodage TEI n'est pas anodin, et les caractéristiques de ce standard méritent d'être précisées. Dans leur ouvrage électronique *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*, Marin Dacos et Pierre Mounier indiquent (2014 : 12) que « la TEI, mais aussi d'autres initiatives de même nature créent progressivement des outils, des méthodes et des espaces partagés entre plusieurs disciplines ». L'enjeu pour nous, et plus largement dans le projet CoMeRe, était donc de pouvoir inscrire la recherche dans la perspective adoptée par la communauté, et notamment dans le cadre plus spécifique de nos projets, en lien avec les standards liés aux humanités numériques. Il est aussi question, avec cette « harmonisation » dans la constitution des corpus, de l'interopérabilité, et de la possibilité de comparaisons des méthodes, des outils, des travaux, etc. D'ailleurs, les auteurs indiquent que ces méthodes « ouvrent la possibilité que se développent des recherches concrètes sur, non pas les outils informatiques dans telle ou telle discipline, mais sur les usages des technologies numériques dans la recherche en sciences humaines, dans sa diversité même. Des problématiques partagées émergent alors, sur les pratiques d'encodage de l'information, sur la structuration, la diffusion et l'archivage des corpus ».

14 Plus concrètement, la TEI peut être définie comme suit[4] :

> La TEI, ou Text Encoding Initiative (initiative pour l'encodage du texte) 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. Le schéma est un centre autour duquel gravitent beaucoup d'activités coordonnées sous forme de comités démocratiques et internationaux pour notamment : conduire la maintenance et la croissance du schéma, rédiger la documentation, développer des outils génériques, assurer le support sur des listes de diffusions, et faire connaître le format.

15 Cette communauté est très active et structurée, et possède un site internet et un manuel d'utilisation très complets et détaillés. Comme le note Meunier (2018, en ligne), ce format permet « une annotation textuelle sophistiquée, il est adaptable et n'empêche aucunement son insertion sur le Web », et en outre les types d'annotations « bénéficient également d'un certain statut consensuel, facilitant la collaboration et les échanges au sein de la communauté académique ». S'il note que « ces formats tendent trop souvent à être lourds et de réalisation coûteuse », nous pouvons ici saluer le rôle structurant des consortiums Corpus Ecrits, puis Corli, qui, en lien avec la TGIR Huma-num, offrent la possibilité de soutiens financiers pour la « finalisation de corpus » (dont nous avons pu bénéficier).

16 Dans le cadre de CoMeRE, le travail d'encodage TEI s'est effectué avec un partenariat européen sur la TEI (groupe d'annotation TEI-CMC) et en relation avec l'infrastructure DARIAH. Ce noyau de corpus devrait être intégré au futur « Corpus de référence du français ». La constitution du corpus pose aussi la question de la méthode et des principes d'analyse pour l'analyse du discours.

1.3 Induction et déduction, contextualisation et extraction : positionnements

17 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 2013b). 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 2013b 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 ». À l'inverse, le type d'approche prônée par Paveau valorise « la capture d'écran », et renvoie à une certaine « conception du discours ». Ces travaux, basés sur des captures d'écran, choisies par les chercheurs comme des exemples « exemplaires », destinés à illustrer un propos, peuvent être critiquées 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.

18 Dans notre recherche, nous adoptons donc une méthode mixte pour appréhender des corpus, 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 (2011, p. 33-34) :

> 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.

19 Dans la littérature, maintenant abondante sur l'analyse des corpus numériques, les approches dépendent en général 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).

20 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 co-textualisée 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.

21 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'aspiration 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) ». La réflexivité de l'approche présentée en 1.2, sur la base des considérations abordées en 1.1, nous permettent de garantir cette prise en compte raisonnée de corpus. Reste alors à aborder la question des outils et formats d'analyse de ces corpus.

2. Comment explorer les corpus : des corpus aux interfaces

22 Les indications proposées par Mayaffre avaient déjà été mises en œuvre dans l'élaboration du corpus *Polititweets*, puisque constitué au format XML-TEI, et déposé sur la plateforme Ortolang[5]. 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.

2.1 Produire des interfaces d'extraction des corpus

23 Ce format XML (« langage de balisage extensible » si nous traduisons) est particulièrement bien adapté à la constitution de documents numériques sur lesquels des analyses pourront être menées, du fait du balisage qui permet l'introduction de métadonnées.

24 Par exemple, dans le tweet suivant :

25 les éléments spécifiques à Twitter sont codés de la manière suivante :

1. Le hashtag :
<distinct type="twitter-hashtag"><ident>#</ident><rs ref=
"https://twitter.com/search?q=%23Bruxelles&src=hash">Bruxelles</rs></
distinct>
2. La @ (mention) :
<addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-
account" ref="#cmr-politweets-p80820758">JLMelenchon</addressee></
addressingTerm>
3. Ou encore le type d'appareil sur lequel a été écrit le tweet :
<f name="medium"><string>Twitter for iPhone</string> </f>

26 Aussi, la nature des différentes données est prise en compte dans le codage TEI : le corpus *Polititweets* permet donc de tenir compte, par le biais d'un codage mobilisant des

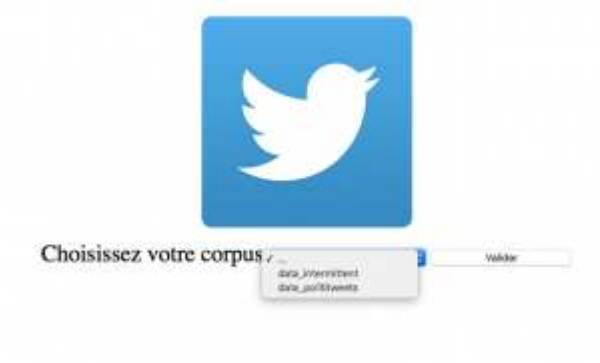
balises TEI spécifiques, des différentes ressources sémiotiques des tweets politiques. Concernant la mise à la disposition de la communauté, si la question de la licence était réglée par le choix opéré, le format XML n'était pas sans poser de problème, puisque ce format ne semblait en effet pas partagé par tous les chercheurs de la communauté (notamment en analyse du discours), et la conversion de ces documents en formats importables dans des outils de textométrie n'était pas considérée comme évidente. Ainsi, une interface[6], 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.

27 Cet outil se présente comme un moteur de recherche :

Image 1. Moteur d'interrogation des corpus

28 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



29 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.

30 L'utilisateur peut ensuite effectuer sa requête, par exemple « démocratie ». En cliquant sur « Valider », les résultats apparaissent : contenu des tweets, auteur du tweet, support de production, et nombre de retweets. Ceci est visible sur la capture d'écran suivante :

Image 3. Résultats obtenus et possibilités d'exports

31 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*.

32 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 » :

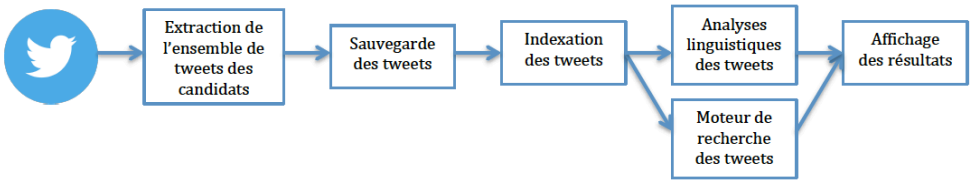
Image 4. Résultats obtenus dans l'interface (corpus) ou en utilisant un logiciel (analyse de similitudes avec Iramuteq)

33 Cet outil constitue donc un premier pas vers l'application #Ideo2017[7] que nous allons présenter : mise à disposition à la communauté des corpus et de l'interface, outil intuitif, aide à la constitution de corpus balisés grâce à la médiation de l'outil. L'objection d'« extraction » (présentée en 1.3 à propos des analyses informatisées des corpus issus des réseaux sociaux) qui provoque une décontextualisation des données textuelle, est néanmoins applicable en partie à cette démarche ; néanmoins, cette critique trouve déjà un certain nombre de réponses concrètes, à la fois par la constitution des champs des balises xml, ou par les exports de l'interface, qui donnent accès à des métadonnées contextuelles, 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.

2.2 Produire des interfaces d'analyse

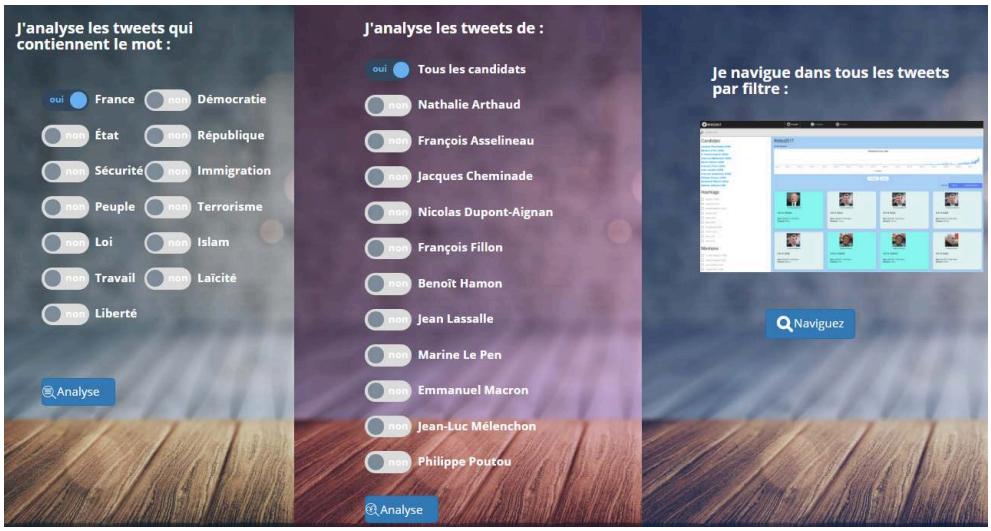
34 L'outil #Idéo2017, élaboré dans le contexte de l'élection présidentielle 2017, est réalisé via plusieurs étapes, comme indiqué dans la Figure 1 : (1) l'extraction de l'ensemble de tweets des candidats, (2) la mise en place d'une sauvegarde des tweets, (3) l'indexation des tweets pour faciliter la 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.

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



35 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 :

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



36 Cette plateforme, conçue comme un prototype opérationnel, a utilisé les fonctionnalités du logiciel Iramuteq, dont le code source est mis à disposition par son concepteur (Pierre Ratinaud). Notamment, l'analyse de similitudes, et la *méthode Reinert*, ont été implémentées, pour représenter les associations de mots (sur la base de l'indice de cooccurrence) et les thématiques (sur la base de la classification hiérarchique descendante).

37 L'analyse *« J'analyse les tweets qui contiennent le mot… »* permet à l'utilisateur de choisir un mot parmi les 13 mots 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*).

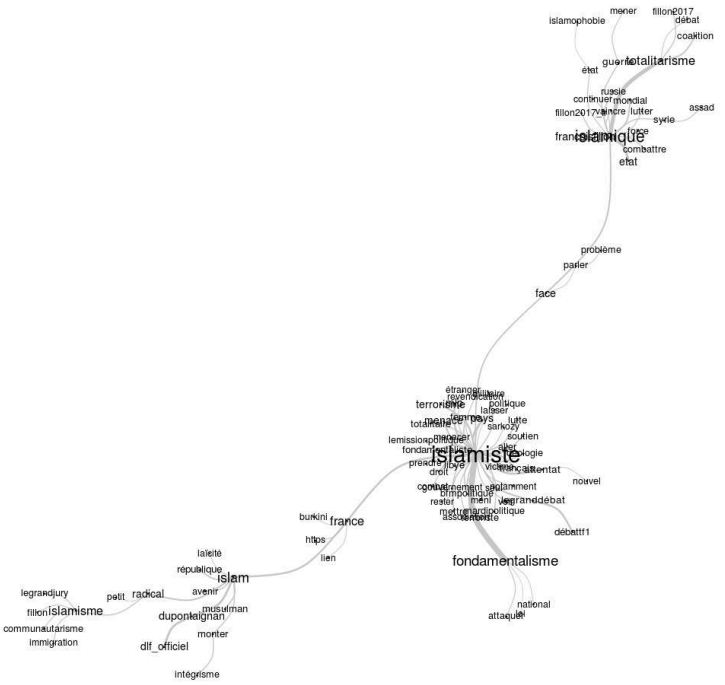
38 L'analyse *« J'analyse les tweets de [candidat] »* permet à l'utilisateur de choisir un candidat parmi les 11 candidats (ou le corpus global des 11 candidats) afin d'analyser ses tweets via les techniques suivantes : les mots les plus utilisés, les thématiques (issues de *la méthode Reinert*), les 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).

39 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 ». Ce retour au texte est d'une grande importance, car il garantit l'équilibre possible entre les analyses quantifiées et les analyses contextualisées ; il permet en retour, par l'exploration du corpus par différentes fonctions, d'expérimenter certains résultats, et d'interagir avec les résultats statistiques, et les données qui composent le corpus.

3. Exemple d'une exploration de corpus : l'ancrage linguistique

40 Avec le contexte politique tendu en France lors de la campagne présidentielle 2017, et les amalgames autour de islam/terrorisme, islam/immigration, etc., l'analyse de la manière dont les candidats à l'élection investissent ce thème était nécessaire. À partir du corpus lié à la recherche de la forme « Islam » (incluant les dérivés[8]), nous avons analysé avec la mise en discours de ce sujet avec les différentes fonctionnalités. Par exemple, pour les mots associés à ce mot (analyse de similitudes), nous obtenions le résultat suivant :

Image 6. Analyse des similitudes du sous-corpus constitué des tweets contenant « islam » et ses dérivés



41 L'Analyse de similitudes (ADS), établit un calcul sur la base d'« un indice de co-occurrence (combien de fois les éléments vont apparaitre en même temps) » pour donner un résultat visuel « où la taille des mots est proportionnelle à la fréquence et où la taille des arêtes et proportionnelle à la force »[9]. Ceci permet de rendre compte visuellement de la fréquence des mots en lien avec les associations spécifiques.

42 On voit que dans ce corpus, « islam » n'est pas le terme le plus fréquent (il n'est pas le plus grand en taille), mais que c'est plutôt « islamiste » (qui concentre ensuite le plus d'associations, avec notamment « fondamentalisme »), et « islamique » (en lien avec « totalitarisme ») ; à gauche de l'image, nous voyons aussi « islamisme », qui apparaît avec « communautarisme » et « immigration ». On se trouve donc face à des processus de construction morphologique(s) par suffixation, qui ne sont pas neutres du point de vue du sens, et de la perception des référents qui sont construits par ces discours :

3.1 « Islamisme »

43 Ce terme est construit sur islam + *isme* : il serait question d'une idéologie en lien avec l'islam, cette idéologie étant thématisée en lien avec l'immigration et le communautarisme. Par exemple (message retweeté par Marine Le Pen) :

Image 7. Tweet de David Rachline

3.2 « Islamiste »

44 On retrouve l'association « fondamentalisme islamiste » qui est l'élément de langage principal de Marine Le Pen à ce sujet. Le *-iste* conduit à la création d'un mot qui « désigne celui qui adhère à une doctrine, une croyance, un système, un mode de vie, de pensée ou d'action, ou exprime l'appartenance à ceux-ci » (TLFI), et le statut d'adjectif crée un lien consubstantiel avec le nom sur lequel porte cet adjectif. L'islam est donc ici véhiculé comme idéologique, et replié sur lui-même, avec « fondamentalisme », qui serait l'idéologie du fondement. L'expression « fondamentalisme islamiste » serait donc une manière de désigner des idéologues de l'islam ancré dans ses fondements. Les liens avec d'autres problématiques sont alors rendus possibles par cette construction, par exemple :

Image 8. Tweet de Marine Le Pen

45 On voit que « communautarisme » intervient ici aussi, avec un rapport de causalité entre les deux notions. Cette construction d'une certaine représentation de l'islam est devenue dans le discours frontiste la grille de lecture principale du phénomène

46 La combinaison morphologique d'un *–isme* et d'un *–iste*, le rôle grammatical de l'adjectif vis-à-vis du nom, et la dépendance ainsi créée, donnent à voir une certaine doxa sur l'islam, présentée comme une évidence, grâce au maniement de la matérialité linguistique. S'il ne s'agit pas ici d'une fausse information, il s'agit d'une manière de présenter le réel qui oriente immédiatement l'interprétation, et contraint l'auditeur/lecteur à une certaine vision du monde.

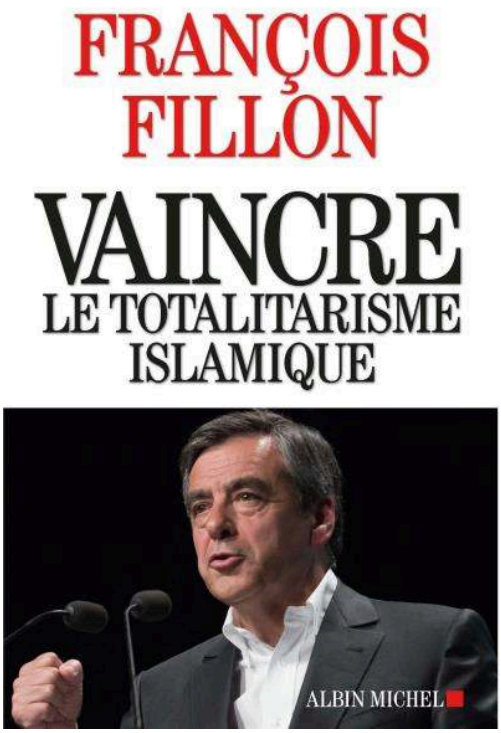
3.3 « Islamique »

47 C'est essentiellement François Fillon qui utilise cette forme, en lien avec « totalitarisme », comme dans cet exemple :

Image 9. Tweet de François Fillon

48 Pour ce dernier emploi, on trouve à proximité dans l'image de l'analyse de similitude, « fillon2017 », ce qui renvoie au hashtag #fillon2017, et montre un élément important de la campagne de François Fillon. Ceci n'est pas sans poser de problème d'information, comme le révèle par exemple L'Obs à la suite de la publication de l'ouvrage du candidat :

Image 10. Couverture de l'ouvrage de François Fillon, et titre d'un article de L'Obs




L'OBS

POLITIQUE MONDE ÉCONOMIE CULTURE OPINIONS DÉBATS TENDANCES VIDÉOS PHOTOS

L'Obs > Politique > Présidentielle 2017

"Islamique", "islamiste" : la fâcheuse confusion de François Fillon

EN BREF

ALERTE INFO Bruno Le Maire promet un déficit public sous la barre des 3% en 2018 et 2019

15:04 Syrie : frappes russes contre Idleb, les plus "intenses" depuis un mois

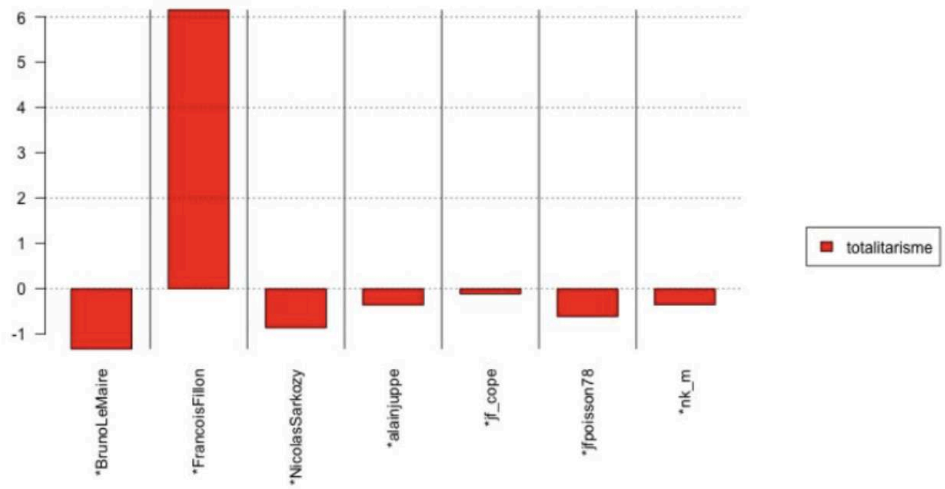
11:48 Egypte : 5 ans de prison pour le journaliste Shawkan, qui devrait être libéré

10:55 "Nous n'avons pas réussi" : échec des pourparlers sur la guerre au Yémen

Surpris par le titre du dernier ouvrage du candidat de la droite à

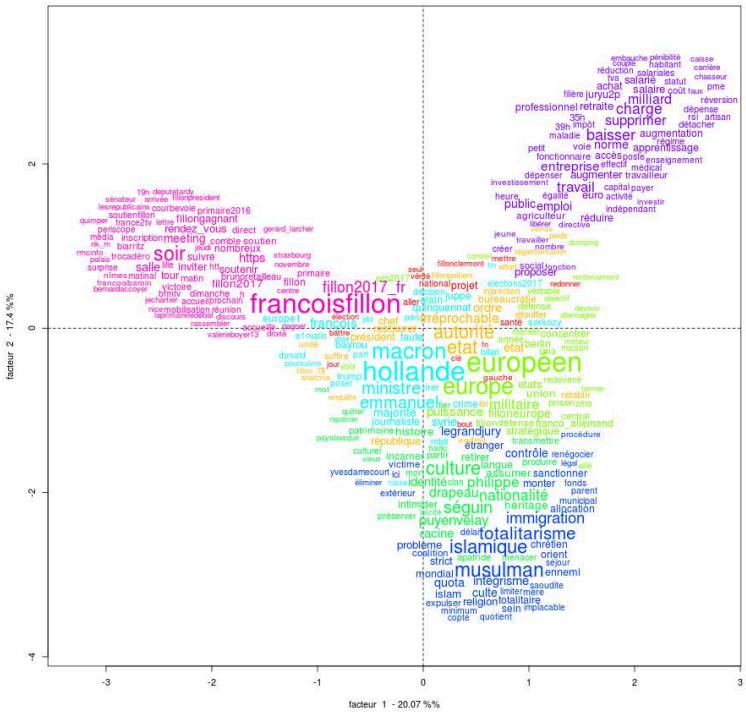
49 Si on retrouve ici un terme en *–isme*, « totalitarisme », islam a ici le suffixe *–ique*, issu du latin *-icus* signifiant « relatif à, qui est propre à » (TLFI). Cet adjectif correspond à un substantif suffixé en *-isme* sans en dériver (et selon le TLFI l'adjectif est postérieur au substantif dans le cas de « islamique »). Le candidat F. Fillon présuppose donc une dimension totalitaire à l'islam, puisque l'adjectif « islamique » (relatif à l'islam) porte sur le nom « totalitarisme ». Il y a en effet une grande différence entre les adjectifs « islamique » (relatif à l'islam) et « islamiste » (lien avec l'appartenance à une doctrine, système de pensée), car c'est ainsi tout l'islam qui est critiqué, et lié au totalitarisme, et non pas seulement une certaine appréhension de l'islam, considéré comme idéologique ou extrême. Notons que cette inflexion sur le caractère totalitaire de l'islam est propre au candidat, et non à son camp politique au moment de l'élection, puisque ce terme est sur-utilisé par le candidat Fillon à la primaire, et sous-utilisé par ses concurrents :

Image 11. Calcul de spécificité du terme « totalitarisme » lors du débat à 7 de la primaire de droite



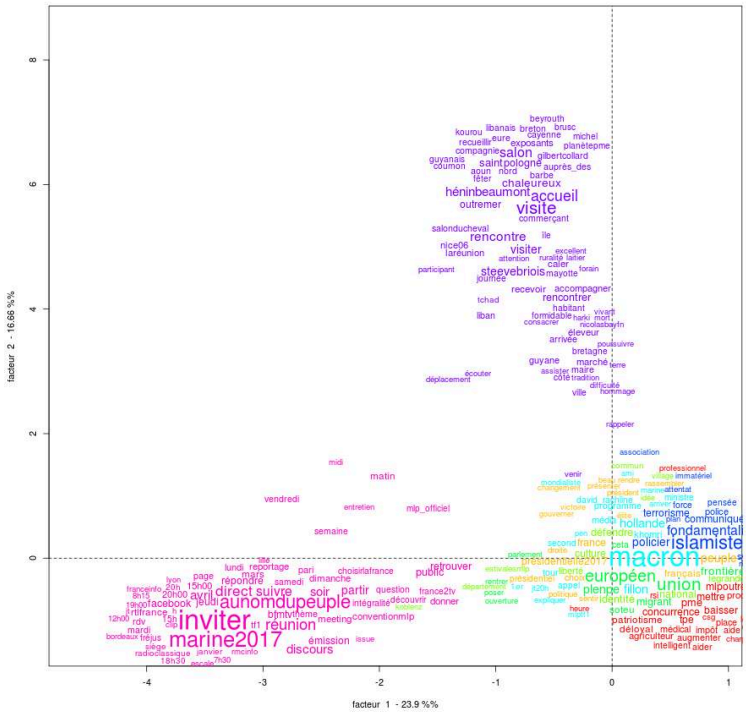
50 À titre d'exemple, pour les deux candidats que nous venons de décrire, Marine Le Pen et François Fillon, l'analyse thématique (méthode Alceste) appliquée au corpus, montre que l'islam s'intègre à une thématique caractérisée par les mots « immigration », « intégrisme », « allocation », « étranger », etc., chez François Fillon (en bleu) :

Image 12. Thèmes dans le corpus de François Fillon



51 ou « policier », « terrorisme », « attentat », chez Marine Le Pen (en bleu également) :

Image 13. Thèmes dans le corpus de Marine Le Pen



52 On se trouve donc face à des mises en discours extrêmement orientées, et connotées, du thème de l'islam, qui se lexicalise de manière spécifique dans des configurations précises. Sur le plan méthodologique, on constate la complémentarité des fonctionnalités de la plateforme #Idéo2017, et la dimension nécessaire processuelle de la représentation des discours : c'est en naviguant par l'analyse des mots, puis des candidats, puis en retournant aux données, que l'on peut mener une étude sémantico-discursive basée sur corpus.

Conclusion

53 La dimension discursive de l'analyse, et l'extension de la problématique aux aspects linguistiques, met en avant la nécessaire prise en compte de la matérialité des discours, et la portée des formes, dans l'exploration de ces corpus de tweets. La « cartographie » proposée par exemple avec l'analyse thématique, permet de voir dans quelles configurations les mots sont utilisés, et ainsi percevoir leur orientation discursive, en les rattachant aux discours dans lesquels ils s'intègrent. Mais ces analyses ne rendent pas pour autant inaccessibles les aspects contextuels des discours, puisque le moteur de recherche de #Idéo2017 permet de naviguer dans la base de tweets, et de retourner dans l'environnement natif des messages. On peut alors retrouver les caractéristiques des approches « contextualisantes » décrites par Paveau 2013b (« *contextualisation technorelationnelle* », « *investigabilité du discours* », « *technoconversationnalité* »). Mais plus encore que les termes conceptualisant l'approche qui justifient la démarche scientifique, ce qui nous intéresse ici, c'est le va-et-vient rendu possible entre les deux modes d'appréhension. Si l'accès au corpus se faisait uniquement par l'interface de Twitter, nous n'aurions pas d'informations sur la spécificité des termes pour les

candidats en étudiés, la manière dont ils en font usage, la manière dont ils se distinguent de leurs concurrents par le sens qu'ils accordent aux termes et à leurs dérivés. Plus généralement également sur les explorations de corpus, nous considérons qu'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), mais qu'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. L'exploration de corpus peut ainsi passer par des nouvelles productions de ressources, outils, interfaces, dont l'interactivité, l'hypertextualité, et l'ouverture, garantissent la validité des corpus, et la possibilité d'y accéder de manière contextuelle, et outillée.

54 Certes, à ce stade, les méthodes adoptées restent classiques, et n'ont pas nécessité d'adaptation. Des comparaisons entre méthodes d'analyses seront nécessaires (la méthode Alceste se fonde sur les unités de contexte, et puisque les tweets sont des messages très brefs, elle fournit probablement ici des résultats proches d'analyses de cooccurrences). Le développement de métriques et l'adaptation des outils seront donc des pistes de travail, afin de prendre également en considération le caractère inévitablement partiel de l'analyse (à la fois du point de vue de la prise de compte des corpus que du point de vue de la diversité des méthodes d'analyse possibles). Mais un tel projet constitue une étape importante dans le développement d'une démarche d'exploration possible, tournée vers la sphère publique et la mise à disposition des outils et méthodes.

BIBLIOGRAPHY

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

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham Ciara R., Hriba L., Longhi J., Seddah D. (2014). « The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres », *JLCL - Journal for Language Technology and Computational Linguistics* 29 (2), 1-30.

Dacos M., Mounier P. (2014). *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*, accessible en ligne sur http://www.institutfrancais.com/sites/default/files/if_humanites-numeriques.pdf.

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

Longhi J. (2012). « Discours, style, format : niveaux de structuration de la textualité des Tweets de Mouloud », *Actes du 3e Congrès Mondial de Linguistique Française*.

Longhi J. (2013). « Essai de caractérisation du tweet politique », *L'Information Grammaticale* 136, 25-32.

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. (2017). « Le corpus Polititweets : enjeux institutionnels, juridiques, techniques et philologiques », *in* Wigwam C.R. & Ledegen G. (dir.) *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Paris, L'Harmattan, coll. « Humanités Numériques », 37-50.

Mayaffre D. (2002). « Les corpus *réflexifs* : entre architextualité et hypertextualité », *Corpus* 1 [En ligne], mis en ligne le 15 décembre 2003, consulté le 8 janvier 2019. URL : http://journals.openedition.org/corpus/11.

Meunier J.-G. (2018). « Le texte numérique : enjeux herméneutiques », *Digital humanities quarterly*, http://www.digitalhumanities.org/dhq/vol/12/1/000362/000362.html.

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

Paveau M.-A. (2013b). « Genre de discours et technologie discursive. Tweet, twittécriture et twittérature », *Pratiques* 157-158, 7-30.

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris : Editions Honoré Champion.

NOTES

1. Des genres qui transcendent les clivages génériques traditionnels, et font interagir les pratiques issues de différents supports : dans le cas des « tweets de Mouloud », le médiatique intégrait la mise en scène de tweets humoristiques, bousculant en retour la distinction entre écrit numérique et papier, en assouplissant en même temps la contrainte imposée par la plateforme Twitter.

2. Ces « petites phrases », appelées ainsi dans le champ politique, peuvent correspondre, selon les cas, aux « formules » telles qu'elles sont thématisées en analyse du discours.

3. https://corpuscomere.wordpress.com. Notons que les membres du projet CoMeRe appartiennent au groupe de travail « Nouvelles formes de communication » du consortium Corpus-écrits. Le projet a reçu l'appui de Corpus-écrits et de Ortolang.

4. https://fr.wikipedia.org/wiki/Text_Encoding_Initiative

5. https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html

6. http://ideo2017.ensea.fr/outil-twitter/index.php

7. http://ideo2017.ensea.fr/plateforme/

8. Cette recherche établie sur la base des radicaux à l'avantage de permettre une étude de la productivité morphologique des termes candidats. Une analyse sémantique pourrait venir compléter les fonctionnalités offertes, mais cela obligerait par exemple à constituer des catégories sémantiques, qui prendraient par exemple en compte « musulman » avec « islam ». Or avec la diversité du discours politique, nous observons une grande variation dans la construction sémantique des discours. Ainsi, une approche statistique centrée sur les formes est privilégiée : les réseaux sémantiques sont néanmoins perceptibles avec d'autres fonctions, comme les thématiques présentées plus loin dans les images 12 et 13 (et qui permettent de prendre en considération les réseaux lexicaux liés à un même thème).

9. https://datahist.hypotheses.org/category/tutoriels

ABSTRACTS

This paper summarizes the conclusions and developments from research projects developed since 2013, about the analysis of a specific type of CMC (Computer-mediated communication) data: political tweets. After characterizing this genre of discourse, and the issues raised, the paper develops the challenges of exploring this kind of corpus; the apprehension and constitution of these social data in corpus; the production of scientific results, and the implementation of different types of corpus exploration. The methods of constitution of corpus, the standardization and the setting in TEI format, the use of tools for analysis of textual data, and the development of platforms, are thus presented, as different points of the same research which aims to characterize and understand a social practice with a scientific method and a civic goal. The exploration of corpus can thus pass through new productions of resources, tools, interfaces, whose interactivity, hypertextuality, and openness, guarantee the validity of corpora, and the possibility of accessing it in a contextual manner, and tooled.

Cet article synthétise les acquis et développements issus de projets de recherche menés depuis 2013 à propos de l'analyse d'un type particulier de données CMC (Computer-mediated communication) : les tweets politiques. Après une caractérisation de ce genre de discours, et des problématiques soulevées, l'article développe les enjeux de l'exploration des corpus de ce genre ; l'appréhension et la constitution de ces données sociales en corpus ; la production de résultats scientifiques, et la mise en place de différents types d'exploration de corpus. Les méthodes de constitution de corpus, la standardisation et la mise au format TEI, l'utilisation d'outils d'analyse des données textuelles, et le développement de plateformes, sont ainsi présentés, comme différents points d'une même recherche qui vise à caractériser et comprendre une pratique sociale avec une méthode scientifique et une portée citoyenne. L'exploration de corpus peut ainsi passer par des nouvelles productions de ressources, outils, interfaces, dont l'interactivité, l'hypertextualité, et l'ouverture, garantissent la validité des corpus, et la possibilité d'y accéder de manière contextuelle, et outillée.

INDEX

Mots-clés: corpus, tweets, textométrie, TEI, plateforme citoyenne
Keywords: corpus, tweets, textometry, TEI, citizen platform

AUTHOR

JULIEN LONGHI

Université de Cergy-Pontoise
AGORA EA7392
Institut universitaire de France