Le corpus Polititweets : enjeux institutionnels, juridiques, techniques et philologiques

Julien Longhi

LE CORPUS *POLITITWEETS* : ENJEUX INSTITUTIONNELS, JURIDIQUES, TECHNIQUES ET PHILOLOGIQUES

Julien Longhi Université de Cergy-Pontoise, AGORA

L'analyse du discours politique connaît un renouvellement important, dû notamment aux nouveaux supports et formats d'expression, comme les réseaux sociaux numériques (RSN). Or, ces lieux de production d'écrits sont le plus souvent saisis par des disciplines qui les traitent comme des données sociales, plutôt que comme des discours. C'est diversement le cas pour l'opinion-mining, le data-mining, la fouille de données, etc. Les données textuelles y sont considérées comme des « sacs de mots » (Rastier 2011) et les critères sémiotiques et discursifs sont mis au second plan au détriment de caractérisations qui s'appuient sur des ontologies, des thésaurus, ou des associations sémantiques.

Cet article vise à décrire les enjeux philologiques, herméneutiques, et également institutionnels et interdisciplinaires, de la constitution d'un corpus de tweets politiques. Le corpus *Polititweets* (Longhi et al. 2014 : 34273 messages, 205 utilisateurs) a été élaboré selon le format TEI (avec des pistes d'extension aux formats CMC proposées par un groupe européen qui s'est constitué autour de cette question), afin de tenir compte des éléments spatio-temporels, contextuels, technologiques, interactionnels, thématiques, dialogiques, etc. des messages produits. Il s'agira donc dans un premier temps de décrire le contexte d'élaboration du corpus, la méthodologie et des considérations juridiques. Dans un second temps, nous détaillerons les enjeux philologiques de la constitution du corpus, en explicitant les critères qui ont présidé à sa structuration, pour passer d'une base de données à un corpus au format TEI. Dans un dernier temps, nous décrirons la démarche de mise à disposition du corpus et les questions d'« open access ».

1 CONTEXTE DU PROJET ET METHODOLOGIE

Nous considérons que le tweet politique est un genre spécifique du discours politique (Longhi 2013) et avons donc cherché, par la

constitution de *Polititweets*, à obtenir un corpus qui permette de travailler sur ce genre de discours, de le caractériser, et de l'appréhender sous différentes formes. Au plan institutionnel, ce projet s'est inscrit à l'interface du projet « Humanités numériques et datajournalisme : le cas du lexique », que nous avons porté (soutenu par la Fondation de l'université de Cergy-Pontoise), et d'un projet collectif, intitulé CoMeRe (Communication Médiée par les Réseaux) piloté par Thierry Chanier. Au plan méthodologique, plusieurs points ont dû être traités, notamment après une réflexion sur les aspects juridiques liés au corpus.

1.1 CONTEXTE INSTITUTIONNEL

La constitution d'un corpus de tweets correspond à un double objectif, pensé dès le départ 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 participer dans le cadre du projet CoMeRe à la constitution d'un ensemble de corpus de communications médiatisées par les réseaux.

Le cadre institutionnel est donc à la fois :

1. le projet « Humanité numériques et data journalisme : le cas du lexique politique » : ce projet 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 permis 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 dont le travail sur le traitement des données est reconnu et efficace, mais dont la dimension symbolique et sémiotique des données textuelles n'est pas le centre d'attention. Au regard des recherches des acteurs impliqués (porteur comme partenaires), l'étude est menée à propos du lexique politique.

2. le projet CoMeRe[1] décrit sur le site du projet comme suit :

> CoMeRe a pour objectif, à l'horizon 2014, de créer un noyau de corpus de communication médiée par les réseaux (*Computer Mediated Communication* – CMC) en français. Chaque corpus rassemblera un ensemble de conversations intervenant sur la Toile et les réseaux. Nous nous intéressons 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 seront structurés suivant des formats standards : TEI (*Text Encoding Initiative*), CLARIN, OLAC. La banque de corpus sera diffusée en accès libre en 2014 sur le site Ortolang. L'assemblage des corpus se fera sur les serveurs de la MSH (Maison des Sciences de l'Homme) de Clermont-Ferrand et du Laboratoire de Recherche sur le Langage (LRL). Le travail s'effectue avec partenariat européen sur la TEI (groupe d'annotation TEI-CMC) avec relation avec l'infrastructure DARIAH. Ce noyau de corpus sera intégré au futur « Corpus de référence du français ». 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.

Il s'agit donc d'un travail collaboratif, à plusieurs niveaux (au sein de l'université de Cergy-Pontoise, et dans le cadre plus large de CoMeRe).

1.2 ASPECTS JURIDIQUES

La première interrogation à laquelle nous avons dû répondre concernait la possibilité juridique de constituer un corpus de tweets. Les juristes extérieurs aux problématique des corpus sont en effet souvent réticents *a priori* vis-à-vis de ces projets, et des doutes à propos des droits d'auteur par exemple peuvent être invoqués,

[1] http://corpuscomere.wordpress.com/apropos/

notamment sur des sites d'informations « grand public » tels que ceux de *Wikipedia*[2] :

> Le problème des droits d'auteur s'appliquant à un message sur Twitter est loin d'être évident. Par exemple, si on recopie un tweet d'autrui, on ne peut invoquer le droit de courte citation, car le caractère « court » de la citation se rapporte à la longueur de l'œuvre dont elle est extraite. Les retweets, pour leur part, peuvent même être accusés de violer les droits moraux de l'auteur quand le message est modifié. Mais un tweet n'est pas forcément protégé par le droit d'auteur, car celui-ci ne s'applique qu'aux créations originales. Il est rare qu'un message aussi court puisse être considéré comme une telle création, mais pas impossible (c'est le cas des slogans publicitaires). Twitter lui-même encourage les utilisateurs à placer leurs messages dans le domaine public, ne revendiquant lui-même aucun droit dessus – ce qui lui vaut les félicitations des défenseurs des contenus libres en comparaison de Facebook.

Twitter peut ainsi faire figure d'exception, comme le relève le juriste et bibliothécaire Lionel Maurel sur son blog SI Lex[3] :

> Excellente nouvelle : Twitter ne revendique aucun droit de propriété intellectuelle sur les contenus produits par les utilisateurs du service. On sait que ce n'est pas forcément le cas de tous les services 2.0, qui peuvent se comporter comme de véritables prédateurs de ce point de vue.

Ceci est confirmé par Twitter, que nous avons contacté et qui nous a renvoyé vers plusieurs pages informatives ou de « foire aux questions »[4]. Dans ces pages[5], Twitter explique notamment que « en utilisant nos Services, vous consentez à la collecte et l'utilisation

[2] https://fr.wikipedia.org/wiki/Twitter
[3] http://scinfolex.com/2009/06/14/twitter-et-le-droit-dauteur-vers-un-copyright-2-0/
[4] L'échange avec Twitter a débuté le 17 janvier 2014, et les consultations de ces pages entre janvier et mars 2014. Ces éléments sont consignés dans « Proposition pour l'acquisition d'un corpus de Tweets V5 », document disponible dans le dossier du corpus *Polititweets* lors de son téléchargement.
[5] Par exemple https://twitter.com/tos

(ainsi qu'il est énoncé dans la Politique de Vie Privée) de cette information, y compris le transfert de cette information aux États-Unis et / ou dans d'autres pays à des fins de stockage, de traitement et d'utilisation par Twitter », ce qui se comprend comme l'« astuce » que « cette licence signifie que vous nous autorisez à mettre vos Tweets à la disposition du reste du monde et que vous permettez aux autres d'en faire de même » ; en outre, « Twitter applique un ensemble évolutif de règles sur la manière dont les partenaires de l'écosystème peuvent interagir avec vos Contenus », ou encore « Si vous suivez ces directives, vous n'avez pas à contacter Twitter pour d'autres autorisations d'affichage ou en relation avec les marques déposées. Dans certains cas, l'autorisation du créateur du contenu peut toutefois être nécessaire, les utilisateurs de Twitter conservant les droits sur le contenu qu'ils postent ».

1.3 LE CHOIX DES DONNEES

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 ils étaient cités : cela a conduit à 7087 listes ;
3. Sélectionner parmi ces listes celles qui avaient au moins 6 twittos (utilisateurs de Twitter) et qui contenaient la chaîne de caractère *politic* dans le nom ou descriptif de la liste : 120 listes ;
4. Sur ces 120 listes récupérer les 2934 twittos ;
5. Travail 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 34273 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*. Ces comptes ne sont pas

forcément des individus, mais peuvent être des partis, des influenceurs, ou des comptes satiriques (comme @humourdedroite). Notre critère est en effet celui de l'influence des comptes dans le champ politique, et non la considération institutionnelle ou fonctionnelle de ces comptes.

Cet ensemble de tweets a ensuite dû être mis en forme pour devenir un corpus, afin de répondre aux enjeux institutionnels du projet CoMeRe, mais aussi afin de permettre par la suite des développements de recherches en analyse du discours outillée sur corpus.

2 LA CONSTITUTION DU CORPUS : DE L'EXTRACTION A LA STRUCTURATION DES DONNEES

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

1) qui fait appel à l'API de Twitter : appel d'une dizaine de fonctions de l'API selon nos besoins, et récupération de toutes les informations sous format JSON que nous convertissons[6] ;

2) ceci permet d'enrichir une base de données avec un design de base propre (dizaine de tables, cinquantaine de champs). Ensuite nous avons des programmes qui calculent des indices[7] pour enrichir des champs supplémentaires ;

[6] Nous récupérons pour chaque compte : user_id, screen_name, name, location, url, description, created_at, followers_count, friends_count, time_zone, listed_count, isProtected, favouritescount, lang, isGeoEnabled, isVerified, isTranslator, statusescount, listed_count_community. Pour tous ces tweets : tweet_id, tweet_text, created_at, geo_lat, geo_long, user_id, screen_name, name, source, isTruncated, isFavorited, InReplyToStatusId, InReplyToUserId, InReplyToScreenName, Place, favoritecount, isRetweet, retweetedstatus_id, contributors, retweetcount, isRetweetedByMe, CurrentUserRetweetId, isPossiblySensitive, getIsoLanguageCode, isRetweeted, entities_user_mentions, entities_hashtags, entities_urls. Pour chaque liste twitter ou le compte est présent : idListIMP, name, fullname, slug, description, memberCount, subscriberCount, uri, user_id, count_community, ainsi que tous les identifiants des followers et des followings.

[7] Nous calculons un score local et global d'influence pour chaque utilisateur à partir de la densité et la centralité de son noeud à partir du graphe social étudié. Nous calculons pour chaque hashtag et pour les noms propres et les noms communs (après étiquetage

3) puis export sur mesure, avec les informations stockées, dans le format de données souhaité, comme le format xml dans le cadre de la constitution du corpus.

L'enjeu pour une approche linguistique est donc d'utiliser ce matériau pour élaborer le corpus *Polititweets*.

2.1 CONSTITUTION DU CORPUS : FORMAT TEI

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 ». Toutes ces questions sont centrales dans la constitution de *Polititweets*, et sont l'objet du présent article. Plus concrètement, la TEI peut être définie comme suit :

> 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.

morphosyntaxique dans le texte intégral) un équivalent du TF/IDF. Nous considérons les relations sociales entre chaque utilisateur ciblé ; les Actions sociales globales de chaque tweet récupéré (retweet, mention, reply) ; les Actions sociales locales relative à un utilisateur (voir Alkhouli et *al*. 2015 pour plus de détails sur ces calculs).

> 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[8].

Cette communauté est très active et structurée, et possède un site internet et un manuel d'utilisation très complets et détaillés[9].

2.2 EXEMPLE DE MISE EN FORME D'UN TWEET

Le lecteur pourra abondamment se documenter sur le format TEI en consultant les liens proposés précédemment, aussi nous proposons une illustration. Voici un tweet tel qu'il se présente dans l'interface native de Twitter :

[8] https://fr.wikipedia.org/wiki/Text_Encoding_Initiative
[9] http://www.tei-c.org/index.xml

1 Exemple d'un tweet

Tous les éléments qui ressortent en gras et qui sont précédés de # ou @ sont des technomots (Paveau 2013), c'est-à-dire des mots qui sont enrichis d'une fonctionnalité technologique, et sont cliquables. On y retrouve le hastag (mots, sigle, initiales, précédés du #), permettant de « tagger » le tweet (ici #Bruxelles, #UE, etc. qui sont plutôt thématiques : les hastag peuvent aussi être plus créatifs ou spécifiques au réseau Twitter) ; l'arobase (@) qui peut servir à mentionner un utilisateur (comme @JLMelenchon), où à lui adresser le tweet s'il est placé en début de message. La mention peut avoir plusieurs fonctions, puisqu'il est certes possible de mentionner un compte, mais aussi de s'en servir pour faire porter un propos au compte inséré, avec une forme de discours rapporté. Un tweet peut également comporter une URL (automatiquement réduite).

Tous ces éléments sont donc « codés » grâce au format TEI, afin de baliser les différents aspects sémiotiques et technologiques spécifiques au tweet. Ainsi, notre tweet apparaît comme suit dans le corpus :

```
<post xml:id="cmr-politweets-a448491380948885504" who="#cmr-politweets-p80820758"
  when="2014-03-25T17:07:32" xml:lang="fra">
  <p><distinct type="twitter-retweet"><ident>RT</ident>
      <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
         ref="https://twitter.com/LePG 35022597"
      >LePG</addressee></addressingTerm>:</distinct> A <distinct type="twitter-hashtag"
        ><ident>#</ident><rs ref="https://twitter.com/search?q=%23Bruxelles&amp;src=hash"
        >Bruxelles</rs></distinct>,
        <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
        ref="#cmr-politweets-p80820758">JLMelenchon</addressee></addressingTerm> conclue
    la rencontre-débat sur le <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23GMT&amp;src=hash">GMT</rs></distinct>. Nous
    live-tweetons - <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23UE&amp;src=hash">UE</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23USA&amp;src=hash">USA</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23Europe&amp;src=hash">Europe</rs></distinct> -
      <ref target="http://t.co/hNGmnAagIM">http://t.co/hNGmnAagIM</ref></p>
  <trailer>
    <fs>
      <f name="medium">
        <string>Twitter for iPhone</string>
      </f>
      <f name="retweetcount">
        <numeric value="22"/>
      </f>
      <f name="isRetweet">
        <binary value="true"/>
      </f>
      <f name="retweetedstatus_id">
        <numeric value="448490939338588160"/>
      </f>
    </fs>
  </trailer>
</post>
```

2 Mise en forme du tweet dans le corpus

Ce format XML (« langage de balisage extensible ») 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.

Par exemple, dans le tweet précédemment présenté, nous pouvons repérer un certain nombre de variables codées, telles que :

1. Le hastag : <distinct type="twitter-hashtag"><ident>#</ident><rs ref="https://twitter.com/search?q=%23Bruxelles&src=hash" >Bruxelles</rs></distinct>
2. Le @ (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>

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. Un enjeu de corpus est ensuite la mise à disposition à la communauté, par le choix de licences et d'insertion à des platerformes favorisant cela.

3 MISE A DISPOSITION A LA COMMUNAUTE, OPEN ACCES ET LICENCE APPOSEE

Nous l'avons indiqué, l'objectif du projet CoMeRe est de mettre à la disposition de la communauté un ensemble de corpus issus de nouveaux moyens de communication. Aussi, le projet est doté d'un site internet sur lequel les phases du projet, sa structuration, et sa mise à disposition, sont communiqués. On trouve en outre le corpus sur le répertoire d'Ortolang[10] avec la description en ligne. Des articles ont d'ailleurs été rédigés sur le site pour faciliter l'utilisation des corpus, comme un billet « Comment télécharger un corpus CoMeRe ? »[11] qui donne l'exemple de *Polititweets*. Cette page permet au lecteur d'être orienté sur le site d'Ortolang pour télécharger le corpus. Pour préciser davantage à ce stade le lien entre CoMeRe et Ortolang, voici le descriptif donné sur la page d'accueil[12] :

> 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

[10] https://repository.ortolang.fr/api/content/comere/v1/cmr-polititweets.html
[11] https://corpuscomere.wordpress.com/2015/06/24/comment-telecharger-un-corpus-comere/
[12] https://www.ortolang.fr

> 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.

Enfin, 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[13] :

> 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.

Ainsi, les utilisateurs du corpus peuvent l'utiliser, travailler dessus, proposer des annotations, ajouter des éléments, et republier le corpus, dès lors que les éléments citationnels traditionnels sont garantis. Ceci s'intègre donc dans la démarche adoptée dans le projet CoMeRe à propos de l'open access : accessibilité des données, redistribution et réutilisations possibles, participation ouverte sans restrictions.

CONCLUSION

Nous venons de présenter les lignes principales qui ont présidé à la constitution du corpus *Polititweets* : à la croisée de deux projets de recherche, ce corpus a permis de constituer un terrain d'analyse du discours politique sur Twitter, et de contribuer à l'élaboration d'une banque de corpus de nouveaux moyens de communication. Pour tenir

[13] http://creativecommons.org/licenses/by/4.0/deed.fr

compte de la spécificité des productions (tweets politiques), une réflexion méthodologique et philologique a été menée, notamment pour faire en sorte que le balisage TEI et la structuration répondent aux enjeux technologiques et discursifs des tweets.

Une fois constitué, ce corpus s'intègre à un mouvement de mise à disposition des corpus à la communauté, par l'intégration à la plateforme Ortolang, et l'apposition d'une licence CC 4.0. Les suites du projet, qui seront détaillées dans d'autres articles et travaux en cours de préparation ou finalisation, concerneront l'analyse effective de ce corpus : ces analyses seront rendues possibles par le soin apporté à la constitution du corpus.

REFERENCES BIBLIOGRAPHIQUES

Alkhouli, Abdulhafiz, Vodislav, Dan, Borzic, Boris (2015) : « Algorithms for continuous top-k processing in social networks », *Proceedings of the first International Symposium on Web AlGorithms*, Deauville, France, disponible sur : <hal-01171346>.

Chanier, Thierry, Poudat, Céline, Sagot, Benoit, Antoniadis, Georges, Wigham, Ciara R., Hriba, Linda, Longhi, Julien, Seddah, Djamée (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, Marin, Mounier, Pierre (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.

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

Longhi, Julien (2013) : « Essai de caractérisation du tweet politique », *L'Information grammaticale*, 136, 25-32

Longhi, Julien, Marinica Borzic, Boris, Alkhouli, Abdulhafiz (2014) : *Polititweets, corpus de tweets provenant de comptes politiques influents*, in T. Chanier (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets- tei-v1].

Longhi, Julien, Wigham, Ciara R. (2015) : « Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow », poster présenté à *Corpus Linguistics 2015*, Jul 2015, Lancaster, United Kingdom, accessible sur https://halshs.archives-ouvertes.fr/halshs-01176061.

Paveau, Marie-Anne (9 mai 2013) : 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

Sites référencés :
Corpus CoMeRE : https://corpuscomere.wordpress.com
Licence Creative Commons : http://creativecommons.org
Ortolang : https://www.ortolang.fr/market/home
TEI : http://www.tei-c.org/index.xml