SHS Web of Conferences **20**, 01021 (2015)
DOI: 10.1051/shsconf/20152001021

Dans un corpus hybride : les messages twittés, l'intertextualité et la formule

A hybrid corpus: twittes, intertextuality and the formula

Daniela Virone[1,a] et Mirko Lai[1]

[1]*Università degli studi di Torino, Italie*

Résumé. L'article propose une réflexion pratique et méthodologique sur l'exploitation d'un corpus de twittes, considéré comme un corpus complexe pour ses caractéristiques particulières (dont la présence des métadonnées) et la possibilité de le mettre en relation avec des corpus plus traditionnels. Le modèle d'analyse quantitative et qualitative expérimenté sur le débat autour du mariage homosexuel en France en 2013 et en particulier sur la formule « mariage pour tous », ici mot-dièse et formule, veut poser les bases pour de nouvelles méthodes d'exploitation des données en analyse du discours.

Abstract. Within this article, we propose a practical and methodological reflection on the analysis of a Twitter corpus, that we consider as a complex corpus because of its particular features (the presence of text and datasets). We will link this corpus to two more traditional ones: a parliamentary debate and some newspaper articles about gay marriage in France. A quantitative and qualitative model analysis has been experimented on a corpus related to the French debate centered on the formulation "marriage pour tous" (marriage for all). , We treat the hashtag as a 'formula'. We want to build a model for new methods on discourse analysis.

1 Introduction

Conduire une recherche dans le domaine de l'analyse du discours signifie souvent aujourd'hui avoir affaire à des corpus de plus en plus grands et complexes qu'il faut organiser, rendre intelligibles et enfin exploiter. L'analyse de ces corpus ne peut pas être menée de façon traditionnelle (avec la seule annotation par exemple), il faut trouver une nouvelle méthodologie permettant de les analyser et d'en tirer des données utiles à la recherche.

L'étude sur la formule[b] « mariage pour tous », que nous sommes en train de mener au sein de notre doctorat de recherche en analyse du discours, nous a confronté à cette situation, celle de l'exploitation d'un macro-corpus hétérogène qui se partage en trois sous-corpus différents : les discours institutionnels, les articles de presse et les messages Twitter, trois genres qui font de notre grand corpus de recherche un ensemble plurisémiotique (au niveau des médias utilisés : la voix, l'écrit, l'écran (K. Lund, et al., 2010[6]). Nous considérons aussi notre corpus comme un corpus hybride (D. Mayaffre, 2005 [10] et 2002 [11]), qui pose le problème de la relation entre un corpus 'innovant' et des textes plus traditionnels, tels que des articles de presse ou des interventions parlementaires. Nous jugeons chaque typologie textuelle comme indépendante des autres et soumise aux contraintes spécifiques de son genre (donc analysable en soi-même), mais aussi comme appartenant à un ensemble structuré qui est le macro-corpus : en effet il s'agit, pour nous, de faire dialoguer ces trois typologies, en mettant en place des solutions qui permettent d'analyser les trois

[a] Auteur de correspondance : daniela.virone@unito.it

[b] La formule est « un ensemble de formulations qui, du fait de leurs emplois, à un moment donné et dans un espace public donné, cristallisent des enjeux politiques et sociaux que ces expressions contribuent dans le même temps à construire » (Krieg-Planque, 2009 [5]) (§2)

corpus et de repérer les relations entre eux, tout en respectant leurs particularités et différences, mais aussi la grande quantité de données qu'ils mettent à notre disposition.

Au travers de cette étude, nous essaierons non seulement de construire un cadre méthodologique, mais surtout d'illustrer une méthodologie d'analyse d'un corpus, laquelle permet d'exploiter un corpus tiré du réseau social Twitter et de mettre ce corpus en relation avec des genres textuels traditionnels. Notre démarche développe une méthode de travail pour l'exploitation de grandes bases de données (qui ne comprend pas seulement l'analyse quantitative) afin de saisir comment le praticien de l'analyse du discours du discours peut approcher les nouveaux genres textuels que le Web propose. Croiser les données du corpus Twitter avec les données des corpus traditionnels nous permet de mener une réflexion sur la langue et les structures complexes mises en œuvre dans l'argumentation dans les médias traditionnels et sur la toile.

Après avoir dédié une première section au cadre théorique de référence, nous expliquerons les particularités et les contraintes du corpus Twitter du point de vue de la collecte, du stockage et de l'exploitation des données et, enfin, nous proposerons une analyse diachronique et synchronique de la formule « mariage pour tous » (MPT) dans le réseau Twitter qui le met en relation avec les corpus du débat parlementaire[c] et des articles de journal[d]. L'étude synchronique de la formule permettra de mettre en évidence les structures linguistiques qui l'accueillent ainsi que la façon dont elle partage l'espace publique (virtuel ou pas) ; tandis qu'une étude sur son origine permet d'expliquer comment et pourquoi à un certain instant le hashtag #mariagepourtous (désormais #mpt), qui appartenait à une petite communauté d'usagers, est devenu formule. Les trois corpus, s'entrecroisant continuellement, créent, à notre avis, un espace de circulation de la formule qu'il vaut la peine d'étudier.

2 Cadre théorique de référence et méthodologie suivie

La « notion de formule » d'Alice Krieg-Planque offre une base théorique pour notre réflexion linguistique (A. Krieg-Planque, 2009 [5]). Celle-ci attribue l'étiquette de formule à toute locution présentant le propriétés suivantes: caractère figé (la formule est connue de tous dans sa forme cristallisée), inscription discursive (c'est grâce à ses nombreux emplois dans toutes sortes de discours que cette séquence se fige et devient un enjeu), aspect polémique (la polémique concerne souvent la signification même de l'unité lexicale), référent social et historique (tout le monde emploie la séquence sans qu'il y ait besoin de l'expliquer). Or, le syntagme nominal MPT, apparu en France au cours du débat sur l'extension du droit du mariage aux couples homosexuels réunit précisément toutes ces propriétés (D. Virone, 2015 [13]). Notre but est celui d'étudier cette formule dans l'espace public, considérant que, comme le dit Krieg-Planque, les discours politiques, médiatiques et institutionnels produits autour de la formule « sont à la fois l'instrument et le lieu des divisions et des rassemblements qui fondent l'espace public » (Krieg-Planque, 2009 [5]). Pour savoir comment la formule « mpt » nait et agit dans l'espace public français, nous avons décidé de prendre en considération trois corpus (le débat parlementaire, les articles de presse et les messages échangés surTwitter) afin de suivre, d'une part, la diffusion et l'évolution de la formule en temps réel et, d'autre

[c] Le corpus contient toutes les interventions des députés et sénateurs lors des passages parlementaires de la loi dite du mariage pour tous, en Assemblée Nationale du 27 janvier au 12 février2013, au Sénat du 4 au 12 avril 2013 et en deuxième lecture en Assemblée du 15 au 23 avril 2013. Il a été collecté à partir de la section dédiée dans les sites de l'Assemblé Nationale (http://www.assemblee-nationale.fr/14/debats/index.asp) et du Sénat (http://www.senat.fr/seances/comptes-rendus.html). Puis il a été stocké sur des fichiers au format txt, exploités avec l'aide de l'outil d'analyse linguistique Ant.conc (http://www.laurenceanthony.net/software.html).

[d] Le corpus est composé des articles de journaux qui contiennent l'expression « mariage pour tous » dans le titre ou dans le corps de l'article, à partir de la première attestation en 2010, jusqu'à juin 2013 et téléchargés depuis la base de données *Factiva* (http://new.dowjones.com/products/factiva/) et des moteurs de recherche internes du journal *Le Monde* (http://www.lemonde.fr/recherche/?keywords=&qt=recherche_globale), dont les articles ne sont pas affichés sur le site Factiva. La recherche a été faite par mots clés : « mariage pour tous ». Puis il a été stocké sur des fichiers au format txt, exploités avec l'aide de l'outil d'analyse linguistique Ant.conc (http://www.laurenceanthony.net/software.html).

part, de pouvoir disposer d'un corpus représentatif de la quasi- totalité du débat, des acteurs et des espaces discursifs impliqués dans celui-ci , enfin nous avons collecté les trois corpus pour en faire un seul corpus réflexif[e] (D. Mayaffre, 2009 [10]). L'analyse quantitative d'un macro-corpus réflexif permet d'en tirer des réflexions linguistiques sur le dialogue entre ses composantes, qui se font écho à travers les mots et les lemmes, ainsi qu'à travers les structures argumentatives choisies. Par conséquent, on devrait s'attendre aux mêmes structures argumentatives, voire aux mêmes arguments dans les twittes, que dans les articles ou les interventions partageant une opinion.

L'analyse des grandes bases de données a été réalisée à travers des outils d'exploitation quantitative et qualitative (notamment avec la logométrie, une méthode d'analyse assistée par ordinateur (P. Marchand, P. Ratinaud, 2014 [8] et 2014[9])). Dans ce cas, la démarche utilisée le plus souvent est inductive, c'est-à-dire que les résultats découlent de l'observation des données qui émergent du texte. Nous avons ajouté à cette méthode une démarche déductive (P. Charaudeau, 2009 [2] et D.Maingueneau, 1991 [7]) , c'est-à-dire que dans notre recherche on fait une hypothèse d'usage de la formule « mariage pour tous » que les medias traditionnels et les twittes pourront confirmer ou pas. En particulier nous hypnotisons que la présence de la formule véhicule les mêmes structures linguistiques et les mêmes argumentations dans les trois types de discours concernés.

En ce qui concerne Twitter, nous nous référons aussi aux travaux sur les hashtags et leur diffusion virale, en particulier nous partageons la théorie de Cunha (E.Cunha et al., 2011 [4]), qui étudient les caractéristiques des langues naturelles : ils font l'hypothèse que les « hashtags peuvent effectivement servir de modèles pour caractériser la propagation des formes linguistiques » (« hashtags may effectively serve as models for characterising the propagation of linguistic forms » (ibidem [4])), c'est-à-dire que le hashtag est désormais l'un des modèles d'innovation de la langue, on y trouve des formes nouvelles (ou des néologismes) qui sont susceptibles de passer dans la langue courante en peu de temps, mais qui peuvent également être oubliées aussi vite. Ce qui nous a intéressé, c'est le passage du hashtag/néologisme à la formule.

3 Twitter: un corpus innovant

Twitter est un réseau social très utilisé partout dans le monde et dont la particularité est la brièveté des messages (140 caractères) mais surtout l'utilisation des mots-dièse ou hashtags : on peut définir un hashtag comme un mot ou une locution, où les mots sont tous unis et qui est introduit par un symbole de dièse (#). Du point de vue de l'argumentation, on peut considérer le hashtag comme un signe qui permet de s'insérer dans le même discours et de créer des communautés virtuelles[f].

Nous avons créé notre corpus à travers le filtre du mot-dièse #mpt : de cette manière nous avons mis en place une base de données de 254.366 twittes, avec exclusion des retwittes (qui s'élèvent à 398.803 messages). Les twittes ont été postés par 50.827 usagers dans un laps de temps de 32 mois – de décembre 2010 à juillet 2013.[g] Ils ont été stockés sur une base informatique qui en permet l'exploitation, après avoir été lemmatisés avec le logiciel TreeTagger[h], ce qui nous a permis de

[e] D. Mayaffre parle de corpus réflexif, entendant « par réflexivité du corpus le fait que ses constituants […] renvoient les uns aux autres pour former un réseau sémantique performant dans un tout (le corpus) cohérent et auto-suffisant » (D. Mayaffre 2009[10]).

[f] Les usagers créent continuellement des hashtags, mais une partie d'entre eux seulement est destinée à la notoriété. Comme les néologismes, ils ont besoin d'un temps pour s'affirmer. Ils connaissent une négociation entre les usagers et leurs diffusion dépend de l'usage qu'en font ces derniers et du nombre de personnes touchées par l'innovation. « In the context of the network theory we indicate two moments on a novelty propagation process: the precise time of the innovation and a later point when some individuals have accepted the innovation» (E.Cunha et al., 2011 [4]).

[g] Les twittes ont été téléchargés à partir du moteur de recherche interne au site de micro-blogging Twitter.com et suivant toutes les politiques de l'entreprise (Twitter®) en matière de traitement des données. En ce qui concerne l'échelon temporel choisi : la date de décembre 2010 correspond à la première apparition de l'hashtag, la date finale correspond à la fin du débat politique et médiatique.

[h] Produit par: Helmut Schmid, TC project at the Institute for Computational Linguistics of the University of Stuttgart. Reperable sur le site http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

disposer des informations grammaticales sur le corpus. Un autre logiciel, MongoDB[i], a été utilisé pour effectuer des analyses quantitatives[j] et des recherches particulières sur les données à disposition, comme on le verra dans la section suivante. Pour une meilleure mise à profit, le corpus a été partagé en deux parties : la première, qui va de décembre 2010 à juillet 2012, compte 3528 twittes postés par 1130 usagers, c'est-à-dire 0,2% du total des twittes. Ce premier ensemble nous a permis de travailler sur l'analyse diachronique et donc sur l'émergence et l'affirmation de la formule sur Twitter. La deuxième tranche, qui va d'août 2012 à juin 2013, comprend la partie la plus importante des twittes (250.823 messages postés par 50.513 usagers).

En ce qui concerne les métadonnées, nous avons retenu les dates de parution des messages, les «user-Id » (le nom d'utilisateur avec lequel l'usager est connu sur le réseau[k]), et les hashtags qui accompagnent le mot-dièse #mpt et nous avons utilisé ces données pour des recherches particulières axées sur les métadonnées. Nous disposons aussi du nombre des URL, des vidéos et des images qui accompagnent les twittes. La présence de ces métadonnées fait du corpus Twitter un corpus complexe (K. Lund, K. Becu-Robinault, 2010 [6]) en soi, car toutes ces données dialoguent avec le texte et nous permettent de le mettre en relation avec le débat extérieur au site.

Une première analyse des métadonnées, par exemple, a permis de construire un graphique qui montre la scansion temporelle des messages sur toute la période (cf. **Figure 1**). Nous avons pu mettre en relation ces données avec les dates qui ont caractérisé le débat hors Twitter et en effet nous avons observé que les pics des twittes accompagnent les jours de plus grande exposition médiatique du sujet : les manifestations (notamment la Manif pour tous du 13 janvier 2013), le débat à l'Assemblée Nationale (janvier-février 2013), l'approbation de la loi (23 avril 2013). Cette première analyse nous a permis d'inscrire le débat Twitter dans un contexte plus ample.[l]

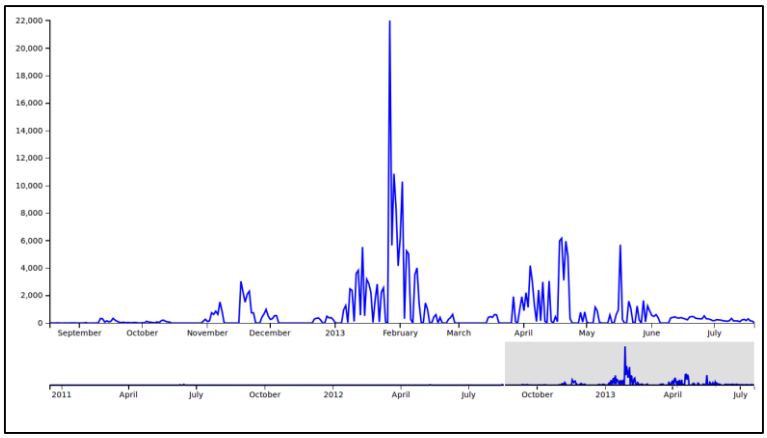

Figure 1. Graphique montrant le nombre des twittes postés de septembre à juillet 2013

Une autre recherche permet de voir quels sont les usagers les plus actifs ou les plus retwités, ce qui est intéressant du point de vue de la circulation de l'information sur le réseau et propose aussi une

[i] Il s'agit d'une base de données qui permet de stocker les données lemmatisées et de les analyser avec des demandes faites à l'aide d'un algorithme. © 2015 MongoDB, Inc. sur www.mongodb.org

[j] La logométrie nous propose une analyse qualitative et quantitative à l'aide de logiciels tels que IRaMuTeq (Ratinaud et Marchand, 2012 [8]). Le logiciel d'analyse des données permet de faire face à la quantité des textes à analyser.

[k] Dans la plupart des cas, l'« user-id » permet de repérer des informations sur le sexe ou le métier de l'usager à travers son profil sur Twitter, mais aussi son appartenance à une association ou sa sympathie pour un parti politique (étant donné que les informations partagées sur le réseau correspondent à des profils réels). Dans le cadre de cet article, toutefois, nous n'avons pas la possibilité d'exploiter ces données.

[l] Elle permet aussi de faire des analyses thématiques des mots sur une base temporelle comme le fait Ratinaud (P. Ratinaud et P. Marchand, 2014 [9]).

lecture de l'espace public français, où plusieurs acteurs (les hommes politiques, des associations, les célébrités) font partie du débat. Comme on peut le voir dans les tableaux ci-dessous (cf. **Figure 2** et **Figure 3**), les usagers les plus actifs sont des associations (fandetv, LeMariagepourtous, Pridemap) ou des personnes (jrossignol, Hirschfeld_J) tandis que l'humoriste Michaël Youn est le personnage le plus retwité, une célébrité donc, dont les deux twittes ont eu un écho énorme.[m] On y trouve aussi le seul twitte de Christiane Taubira et celui du Conseil Constitutionnel qui annonce la constitutionalité de la loi. L'analyse des noms des usagers nous a permis aussi de faire des observations en ce qui concerne l'origine de la formule (§4).

user	tweets
fandetv	4277
cutesmilingcat	2194
jrossignol	1718
Hirschfeld_J	1710
LeMariagePrTous	1129
Yasmilady	1084
JackyMAJDA	953
Pridemap	913
jsherpin	893

Figure 2. Tableau montrant les 9 usagers qui ont posté le plus de twittes.

user	tweets	retweets
MichaelYoun	2	4354
eliodirupo	1	1100
Conseil_constit	1	917
kavanaghanthony	1	612
ChTaubira	1	479
farrugiadom	1	417
AmandineDu38_	1	399
lebonlebon	1	327
youssouphamusik	3	976

Figure 3. Tableau montrant les 9 usagers les plus retwittés

Enfin, une analyse des mots-dièse associés (cf. **Figure 4**) au #mpt nous permet d'en tirer quelques réflexions : les hashtags #DirectAn et #DirectSenat concernent le débat parlementaire, tandis que #manifpourtous se réfère aux manifestations contre la loi. Si on isole les twittes avec le seul hashtag #DirectAn, par exemple, on peut s'attendre à des commentaires sur le débat à l'Assemblée Nationale, ce qui nous permettrait de mettre en relation les thèmes exploités dans ce corpus avec le corpus du débat parlementaire.

mariagepourtous	250,235
DirectAN	25,100
manifpourtous	14,031
DirectSenat	5,055
UMP	4,831
MariageGay	4,115
PMA*	3,361
homophobie*	3,101
LGBT*	2,861
GPA	2,851
Hollande*	2,626
Taubira*	2,391
PS *	2,098

Figure 4. Tableau montrant les hashtags les plus utilisés avec le mot dièse #mpt. Les occurrences marquées avec astérisque (*) sont des cas de variation orthographique qui ont été regroupés.

4 Une analyse synchronique des données

Mais notre but final est d'analyser le texte : un exemple d'exploitation est donné par le graphique ci-dessous (cf. **Figure 5**) qui représente les mots les plus utilisés dans le corpus, partagés par thématique abordée à travers un algorithme (V. Blondel, et al., 2008 [1]). qui compte les cooccurrences en

[m] Le message le plus retwitté appartient à Michael Youn : « En ce jour de manif un seul slogan : Mieux vaut un mariage gay qu'un mariage triste ».

combinaison avec la fréquence des mots.[n] Ce graphique montre quels sont les thèmes approchés par les internautes qui, encore une fois, peuvent être croisés avec les données tirées des autres corpus.

Figure 5. Les quatre nuages des mots les plus fréquemment en cooccurrence. On peut observer que les 'nuages' proposés indiquent, en sens contraire aux aiguilles d'une montre, à partir du premier en haut à gauche : 1. Le thème des droits des enfants et de la famille, 2. Le projet de loi, 3. Les manifestations, 4. L'approbation de la loi et le débat parlementaire.

Toutefois, ce genre d'analyse ne nous dit rien sur les structures linguistiques qui hébergent la formule/hashtag alors que nous sommes en train de rechercher aussi les éléments de la chaîne syntagmatique entourant le plus fréquemment la formule. Il est intéressant de vérifier si ces éléments cooccurrents de la formule, que nous avons identifiés en dehors du corpus Twitter (corpus journalistique/corpus débat parlementaire) se retrouvent également dans l'espace virtuel. Pour ce faire nous nous sommes concentrés en particulier sur la place de l'hashtag dans les twittes et sur celle du syntagme nominal « mariage pour tous »dans les titres des articles des journaux, les deux étant liés par la brièveté de la formulation linguistique[o].

Dans notre article précédent (D. Virone, 2015 [13]) nous avions analysé la locution « mpt » à l'intérieur des titres du *Monde* et du *Figaro* pour savoir quelles étaient les fonctions syntactiques y recouvertes par la formule[p] : dans cette étude nous observions qu'elle se trouvait en particulier dans des syntagmes prépositionnels. Si, du point de vue sémantique, les syntagmes les plus récurrents à la

[n] Une fois éliminés les URL, les hashtags, les mentions, les nombres, les tokens non lemmatisés, les prépositions, les adverbes et les déterminants, nous avons créé un réseau (network) qui affichait 35.146 nœuds et 344.425 liens, nous avons retenu les 4 groupes qui avaient le nombre majeur de nœuds (les autres ne représentaient que 1% du total des mots). Le network ainsi composé affiche 18.581 nœuds et 328.607 liens, c'est à dire 52,87% et 95,41% du total des nœuds et des liens. Pour terminer, nous avons bâti le nuage des mots en utilisant les 20 premiers mots de chaque communauté. La taille est proportionnelle au degré du nœud (V. Blondel et al., 2008 [1]. Le logiciel est repérable sur le site https://sites.google.com/site/findcommunities/

[o] Il est évident que nous ne considérons pas les titres et les twittes comme égaux, étant données les différences de buts, de structure etc... , toutefois notre hypothèse de travail vise à trouver des contiguïtés entre les sous-corpus abordés.

[p] Du point de vue formel dans les titres, nous distinguons 4 typologies de structures : 1. Une structure avec MPT comme occurrence isolée 2. La formule se trouve dans un syntagme prépositionnel : le SP est formé par un SN + P + MPT. a. Du point de vue sémantique, la formule peut désigner le projet de loi contre lequel on manifeste (« *contre le MPT* »; « *opposants/opposition au MPT*»), ou que quelqu'un soutient (*Les partisans du 'MPT' défilent à Paris*). b. Dans d'autres cas, elle est complément du nom, précédée par la préposition '*sur*': Débat + *sur* + MPT ; Loi + *sur* + MPT, Texte +*sur*+ MPT. 3. Elle est utilisée dans les mécanismes de composition des mots : préfixe +MPT . 4. Elle entre dans des séries productives, issues d'une composition/décomposition ou recomposition de la formule…(D. Virone, 2015 [13]).

gauche de MPT font référence aux aspects juridiques (les mots les plus fréquents sont *loi* et *projet*), dont le but est simplement d'indiquer le domaine référentiel auquel renvoie la formule, en revanche les prépositions employées (essentiellement *pour* ou *contre*) renvoient directement à la polarisation de l'opinion publique, et permettent de constituer deux camps opposés. Nous avons donc utilisé cette méthode pour l'analyse de l'hashtag dans le corpus Twitter. Comme nous venons de le dire, une observation des formes lexicales à gauche de l'hashtag #mpt (cf. **Figure 6**) a mis en évidence la fréquence des mots liés aux champs sémantiques de la loi (*projet de loi*) et aux champs de l'opposition/soutien de/à la loi (*opposants, contre, anti, pour*). Une recherche par syntagmes a relevé une importante présence des locutions où l'hashtag est un syntagme prépositionnel.

Nombre des twittes	**Syntagme**
2817	la loi MPT
898	la loi sur le MPT
648	le débat sur le MPT
523	la manif sur le MPT
420	je suis pour le MPT
315	le projet de loi MPT
264	les antis-MPT
221	la manif contre le MPT
209	les opposants au MPT
208	pour ou contre le MPT
127	je suis contre le MPT

Figure 6. Tableau indiquant le nombre des twittes contenant les syntagmes les plus utilisés

Cette liste semble valider l'hypothèse formulée, c'est à dire que les twittes présentent les mêmes structures que dans les titres des articles et permettent à la fois de thématiser le hashtag/ formule (*le projet de loi, la loi*…) ou de proposer une polarisation de l'espace public virtuel et réel d'une communauté et d'un individu (*la manif contre/ je suis contre*). En troisième lieu nous observons que l'hashtag est utilisé aussi pour la composition (*anti-#mariagepourtous*).

On a analysé également les occurrences qui se trouvent à droite de l'hashtag, tout en observant qu'en général celui-ci est placé en fin de phrase[q]. Une première lecture donne comme résultat que le verbe *être* à la troisième personne du singulier est le plus fréquent (9205 occurrences de : *est, c'est* et *n'est* confondus[r]), c'est pour cette raison que nous avons voulu rechercher les éléments que les twittes proposent après celui-ci. En effet, ce verbe peut être utilisé comme auxiliaire mais aussi comme prédicat et, en ce cas, le syntagme qui le suit peut être attributif, c'est-à-dire qu'il peut être utilisé pour attribuer une qualité au sujet par le biais d'un adjectif positif ou négatif, ou d'un nom donnant lui aussi une valeur au sujet : ce procédé est donc une façon d'exprimer une opinion. Par ailleurs, le MPT en tant que formule est considéré comme un conteneur vide de sens (D. Virone, 2015 [13]) , donc cette typologie de structure sert au locuteur pour expliquer la formule à partir de son propre point de vue. Cela explique pourquoi on trouve ce genre de structure attributive dans les trois corpus.

Pour donner un exemple pratique de la mise en relation des trois corpus, nous avons réuni dans le tableau ci-dessous (cf. **Figure 7**) quelques-unes des propositions qui suivent le verbe *être*. En particulier nous avons voulu rechercher la présence des mêmes mots, ce qui signifie que du point de vue argumentatif on retrouve encore les mêmes idées mises en circulation par les médias ou par les hommes politiques et qui sont partagées par les usagers de Twitter.

[q] Les mots les plus utilisés après l'hashtag sont :4716 *est*, 4414 *pour*, 4014 *pas*, 3614 *c'est*, 3415 *danger*, 3280 *je*, 1816 *vous*, 1453 *sont*, 1143 *nous*, 963 *contre*, 875 *n'est*

[r] Dans le corpus Twitter nous avons sélectionné (après avoir éliminé les messages où *être* est suivi par un participe) 1322 messages avec l'hashtag suivi par 'c'est', 1859 avec 'est' et 301 à la forme négative.

Corpus Parlement	Corpus Presse	Corpus Twitter
Le mariage pour tous est un nouveau pas vers la liberté (sénat)		#mpt est aussi une liberté #mpt est la liberté de choisir de se marier ou pas c'est pas du tout une imposition ! #homophobie #mpt est une liberté supérieure à la liberté d'expression
Le mariage pour tous est uniquement un slogan, (1fev,3seance, Ass)	3 janvier 2013, « Le terme de mariage pour tous est un slogan publicitaire »	#mpt est un slogan mensonger #mpt est juste un slogan stupide ?
Le mariage pour tous est un grand projet (30jan2seance,Ass)		#mpt est le grand projet maçonnique casser la famille les valeurs ancestrales de l'humanité*
Le mariage pour tous est un message fort de tolérance envoyé à toute la société (30jan,2seance,Ass)		#mpt est un message fort de tolérance pour lutter contre les discriminations chantal #guittet #directan
Le mariage pour tous est une évidence (30jav,2seance,Ass)	« Le mariage pour tous est une évidence » Le Parisien, 1 avr 2013 (titre)	#mpt est une évidence , les opposant disent 'ils ont pas besoin de ça pour s'aimer' dans ce cas , les hetero non plus #mpt est une évidence (3) #mpt est une évidence contre laquelle on ne peut lutter . même les whatfor le savent depuis tjs #lamournapasdeloi #mpt est une évidence , l'autre option est impossible ! URL #mpt est une telle évidence pour moi que les seules raisons que je puisse comprendre contre . #pensées 5/7 #mpt est un artifice, comme le mariage républicain, en aucun cas une évidence. #mpt est une évidence, les opposant disent 'ils ont pas besoin de ça pour s'aimer' dans ce cas , les hetero non plus nécessité #directan
	« Le mariage pour tous est un faux problème? » La Nouvelle République, 25 nov 2012 (titre)	#mpt est un faux problème . en 2014 , qui pourra encore voter ump ?
	« Le mariage homosexuel sous le paquet cadeau d'un mariage pour tous est une fausse bonne idée. Sous couvert de générosité … » VosgesMatin, 30 nov 2012	#mpt est une bonne idée . #arméededumbledore (4)
	«Le mariage pour tous est un progrès humain» La Dépêche du Midi, 16 Dec 2012 Nîmes Dumas : « Le mariage pour tous est un progrès pour l'enfant » Midi Libre, 28 Jan 2013 'Le mariage pour tous est un grand progrès' L'Indépendant, 20 avr	#mpt est un progrès ? #mariagegay* #mpt est un progrès de la liberté et de la tolérance contre l'homophobie et une vision conservatrice de la société #mpt est contraire au progrès et un recul de la société #mpt est un progrès considérable , dit @fhollande . #confpr #mpt est un progrès sociétal et tout progrès juste et qui renforce l'égalité et la liberté est encouragé par la gauche. #ps #mpt est un progrès pour personne pas même une avancée pour les homos futurs prisonniers d'une normalité impossible à gérer*

	2013	#mpt est un progrès pour toute la société et pas juste les homos #mpt est un réel progrès #jtpm
	« Voeu des élus: Le mariage pour tous est une avancée » Ouest France, 24 Dec 2012 « … Jean-François Macaire, dans un communiqué. Le mariage pour tous est une grande avancée sociétale et je soutiens sans… » La nouvelle république, 11 jan 2013	#mpt est une avancée ; » #mpt est une avancée majeure de notre sociéte civilsée #mpt est une avancée démocratique #fb URL … #mpt est une vraie avancée pour les droits de l'homme et de la femme #mpt est , avant toute consideration , une avancée sociale pic . twitter . com/rdsypn3s #mpt est une grande avancée de société n'en déplaise @frigidebarjot @christineboutin #mpt est une avancée quant à l'idée de famille aujourd'hui encore très ancrée dans un modèle patriarcal rigide et bourgeois #mpt est une avancée en matière d'égalité URL
	« Famille pour tous : "Le mariage pour tous est un prétexte" » LePoint.fr, 22 Avr 2013	#mpt est un prétexte pour flinguer la gauche et cacher la corruption ump URL #copé #sarkozy #mpt est un fallacieux prétexte ! #mpt est un prétexte . #mpt est un prétexte | URL #famille # c'est la compétence de l'adulte qui élève l'enfant qui compte #mpt est plutôt un prétexte pour introduire l'idéologie du gender et détruire la famille , n'en déplaise à #harlemdésir

Figure 7 : Tableau montrant l'utilisation des mêmes jugements de valeurs sur le « mpt » dans les trois corpus. Les cases vides dénoncent l'absence de la formulation recherchée dans l'un des corpus. Les twittes indiqués avec le symbole de l'astérisque (*) ont une valeur argumentative opposée à celle des autres énonciations.

Le tableau montre que la structure linguistique du verbe *être* suivi par un syntagme nominal ou adjectival est très performante sur Twitter, où la brièveté de l'expression linguistique appelle à une surutilisation des structures attributives, mais il montre également qu'elle est utilisée aussi dans la presse comme dans le débat parlementaire, ce qui nous dit quelque chose sur les choix discursifs des journalistes et des hommes politiques, lesquels aiment s'en servir comme d'un procédé argumentatif étroitement lié à la nature polémique de la formule (D. Virone, 2015 [13] et Krieg-Planque, 2009 [5]). De plus, le repérage dans les trois corpus des mêmes mots est aussi très intéressant du point de vue de la circulation des idées. En effet, il est impossible de savoir qui a utilisé une formulation en premier, par contre il est évident que l'expression d'une opinion n'est jamais sans conséquences : une telle argumentation, quand bien même elle serait fautive, une fois mise en circulation, se répand dans l'espace public et intègre et alimente le débat par sa seule présence.

5 Une analyse diachronique

Du point de vue diachronique la mise en relation des données Twitter avec la presse nous a permis de repérer les origines de la formule[s]. Grace à une simple observation des données nous avons tiré des conclusions intéressantes : de l'analyse des métadonnées, par exemple, nous avons pu constater que parmi les usagers les plus actifs (qui ont utilisé l'hashtag dès le début) il y a Gilles Bon Maury et Jean Luc Romero, deux députés du PS, ce qui fait croire que les deux hommes politiques ont 'fait passer' l' hashtag dans le milieu de la gauche, laquelle s'est emparée de la locution lui donnant la légitimation dont elle avait besoin pour devenir formule (D. Virone, 2015 [13]).

[s] Pour faire cette analyse nous n'avons considéré que les twittes de la première période (décembre 2010 à aout 2012).

user	tweets
JeanLucRomero	135
GekkoHopman	112
ProjetEntourage	101
Engagement31	96
JeromePasanau	88
Yagg	79
Funny_Fog	76
Pascal_Lelievre	75
unevisionautre	73
GillesBonMaury	72

Figure 8. Premier twitte contenant le hashtag #mpt et liste des usagers les plus actifs dans la période décembre 2010 - juillet 2012.

En revanche, une analyse qualitative et diachronique de l'hashtag nous a permis de tracer son histoire et le suivre depuis sa naissance (le premier twitte, d'un personnage marginal et méconnu, date de décembre 2010) jusqu'à son affirmation sur le réseau en janvier 2012 (cf. **Figure 8**). En particulier une exploitation du réseau Twitter a fait ressortir comment le hashtag a été accepté dans la petite communauté Twitter, une base de réflexion théorique intéressante sur la diffusion des néologismes dans et hors le réseau. La mise en relation constante du corpus avec les articles de presse a permis de tirer des conclusions sur l'origine de la formule MPT, ce qui permet entre autres choses d'éclaircir la relation que le Web 2.0 entretient aujourd'hui avec l'actualité politique et sociale.

L'hashtag #mpt, en effet, est resté relégué dans la petite communauté des usagers de Twitter qui débattait sur l'opportunité de son utilisation[t] pendant un an et demi. Il n'avait alors aucune diffusion publique (en 2011 on retrace 7 occurrences dans le corpus des articles, toutes à l'intérieur des discours rapportés) ; en janvier 2012, le hashtag est finalement accepté par la communauté et dans les mois suivants il sera de plus en plus utilisé par la presse (19 occurrences entre janvier et aout 2012) mais il ne sortira définitivement de sa crèche qu'en septembre 2012 quand la politique (le PS) légitime la locution, avec l'usage qu'en fait Christiane Taubira dans une interview au quotidien *La Croix*. On a raison de croire que c'est avec le passage dans la presse traditionnelle que l'hashtag #mtp est devenu la formule MPT.

Si l'analyse synchronique, dont on a précédemment parlé, permet de réfléchir sur la base des chiffres qui découlent des recherches effectuées sur une grande quantité de données, l'analyse diachronique, faite sur une petite partie du corpus et menée de façon qualitative, semble être beaucoup plus subjective car elle n'offre à l'analyste d'autre clé de lecture que son observation et sa capacité de mettre en relation des mots avec la réalité qui les a générés. Toutefois cette lecture permet quant à elle d'observer des détails, qui, autrement, peuvent échapper au regard. C'est pour cette raison que la méthodologie de l'analyse quantitative doit toujours être accompagnée d'une analyse qualitative qui se charge de tracer de plus près l'histoire des mots ou des textes qu'on est en train d'étudier.

[t] Gilles Bon Maury écrit : *« Le mariage est une institution et n'a donc aucune sexualité "mariage homosexuel" ne veut rien dire.#mariagepourtous* ». Mais il y a du désaccord dans la petite communauté, manduette 77 répond: *« Oui mais c'est pas le principe d'un #mariagepourtous ce matin, mais un vrai #mariagehomo ».* Une réaction ironique: *« Marions donc les chevaux avec les fantômes #mariagepourtous »* (argumentation qui sera reprise par des journalistes). En général les usagers sont favorables. @Arrandine pose la question des trend twitter: « *#mariagegay ou #mariagepourtous je m'en fous mais c'est pas en ayant deux hashtag qu'on sera dans le trends! »*, la reponse: *« T'as raison il faut donc en retenir un et fidèle à notre projet et à nos valeurs. #mariagepourtous »*. Encore Gilles Bon Maury le 6 janvier 2012: « *Le #mariagepourtous n'est pas un choix culturel, c'est un droit. Par pitié, cessons nous d'utiliser (sic) l'expression "mariage gay". Les institutions n'ont pas de sexualité #mariagepourtous »*. A juillet il semblerait que le hashtag a désormais gagné, encore Jean Luc Romero écrit: « *La ministre #Bertinotti l'assure. Le #mariagepourtous sera voté en 2013. Demain mobilitions-nous (sic) pour la Marche des Fierté LGBT »*.

6 Conclusions

Pour conclure, on a essayé avec cet article de poser des bases pour une pratique d'exploitation des données Twitter, une base de données en évolution et qui permet de se confronter aux nouveaux genres textuels que la toile a générés. On a cherché à mettre en relation ce corpus avec d'autres corpus, montrant comment trois genres textuels différents peuvent dialoguer entre eux par l'utilisation d'une méthode de travail qui ne se base pas seulement sur l'observation directe des données dont on dispose. Traiter des corpus à travers l'outil informatique permet d'avoir une approche globale, mais il faut aussi savoir quel est le but poursuivi. Dans cette perspective méthodologique, il est très important que l'analyse quantitative soit accompagnée, sinon précédée, d'une analyse qualitative des données mais surtout il est nécessaire d'établir un dialogue entre les corpus, aussi bien dans une perspective contrastive, faisant ressortir les différences et les ressemblances, que dans une perspective linguistique, qui ne s'intéresse seulement aux mots comme unités de signification, mais à leur récurrence dans des structures syntactiques et dans des points différents d'une structure textuelle qui a ses propriétés rhétoriques et pragmatiques et avec lesquelles les expressions linguistiques entrent en relation.

7 Bibliographie

1. V. Blondel, J. Guillaume, R. Lambiotte, and E. Mech, "Fast unfolding of communities in large networks," J. Stat. Mech, 2008.
2. P. Charaudeau, *Dis-moi quel est ton corpus, je te dirai quelle est ta problématique*, Corpus En ligne, **8**, 2009
3. F. Chiusaroli, *Scritture brevi oggi. tra convenzione e sistema*, in F. Chiusaroli, F. M. Zanzotto, Scritture brevi di oggi, Eds. Università Orientale di Napoli, pp. 4–44 (2012).
4. E.Cunha, G. Magno, G. Comarela, V. Almeida, M.A Goncalves, F. Benevenuto*, Analyzing the dynamic evolution of hashtags on twitter: a language-based approach*, in LSM, Portland, Oregon: Ass. pour Comput. Ling., pp. 58–65 (2011).
5. A. Krieg-Planque, *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique*, PU de Franche-Comté, coll. « Annales littéraires », (2009)
6. K. Lund, K. Becu-Robinault., *La reformulation multimodale et polysémiotique comme aide à la compréhension de la physique*, in A.Rabatel (ed.), Analyse sémiotique et didactique des reformulations, Besançon, (2010)
7. D.Maingueneau, *L'analyse du discours. Introduction aux lectures de l'archive*, Paris, Hachette Université, (1991)
8. P. Marchand, P. Ratinaud, *Analyse lexicométrique des tsur le #mariagepourtous*, in "Comprendre les mondes sociaux, (2014)
9. Marchand P., Ratinaud P., *Application de la méthode ALCESTE à de « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ*, in : JADT 2012, Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles, ,Liège, p. 835-844, (2012)
10. D. Mayaffre, *Les corpus politiques : objet, méthode et contenu. Introduction*, Corpus En ligne, **4** (2005)
11. D. Mayaffre, *Les corpus réflexifs : entre architextualité et hypertextualité*, Corpus En ligne, **1** (2002)
12. P. Ratinaud, *Visualisation chronologique des analyses ALCESTE: application à Twitter avec l'exemple du hashtag #mariagepourtous* , Université de Toulouse, LERASS, ratinaud@univtlse2.fr (2014)
13. D. Virone, *La formule mariage pour tous dans la presse*, in P. Paissa, F. Rigat, M.B. Vittoz (eds), Hommage à Mariagrazia, Ed Dell'Orso, Alessandria, (2015)