Visualisation chronologique des analyses ALCESTE : application à Twitter avec l'exemple du hashtag #mariagepourtous

Pierre Ratinaud[1]

[1] Université de Toulouse - LERASS – ratinaud@univ-tlse2.fr

Abstract

We propose in this paper a way to visualize temporality of « lexical worlds » revealed by ALCESTE analyses. We use as an example a corpus of tweets that consist of the indexing of 9 months of the hashtag "#mariagepourtous." These visualizations present a chronological reading of the appearance and strength of expression of the different lexical classes. They simplify the temporal reading of these data and highlight elements of interpretation. They are also a tool for the comparison of different corpora.

Résumé

Nous proposons dans cet article des modes de visualisation de la temporalité d'expression des mondes lexicaux mis en évidence par les analyses de type ALCESTE. Nous utiliserons comme exemple un corpus de *tweets* correspondant à l'indexation durant 9 mois du *hashtag* « #mariagepourtous ». Les visualisations proposées permettent d'avoir une lecture chronologique de l'apparition et de la force d'expression des différentes classes lexicales. Ces modes de visualisation simplifient la lecture temporelle de ces données et mettent en évidence des éléments d'interprétation. Elles sont également un outil intéressant dans le cadre de la comparaison de corpus.

Mots-clés : analyse ALCESTE, chronologie, visualisation, *twitter*, IRaMuTeQ

1. Introduction

Les analyses ALCESTE permettent de mettre en évidence les « mondes lexicaux » mobilisés dans les corpus. Quand il s'agit de travailler sur des corpus pour lesquels la composante longitudinale est centrale, les modes de restitutions des résultats traditionnellement utilisés (profils des classes et AFC) ne semblent plus adaptés car ils ne permettent pas une lecture efficiente de la temporalité d'expression des différents champs lexicaux. L'objectif que nous poursuivons ici est de proposer des modes de visualisation de ces résultats qui conduisent à une lecture rapide et précise de la chronologie liée à ces corpus. Un exemple de domaine dans lequel ce problème est particulièrement saillant est celui de l'analyse des conversations sur *twitter*. Nous utiliserons ici un corpus de *tweets* composé des messages contenant le *hashtag* #mariagepourtous et émis sur une période de 9 mois, du 7 novembre 2012 au 31 juillet 2013. Après avoir rappelé le contexte dans lequel se sont déroulés les débats autour du mariage pour tous, nous présenterons le corpus et l'analyse à laquelle il a été soumis. Les résultats apparaissent alors sur des frises temporelles qui permettent de visualiser 3 types d'informations conjointement : l'effectif des classes, l'effectif des *tweets* par jour et la sur-représentation des dates dans les classes. Enfin, nous utiliserons ces représentations pour comparer le corpus #mariagepourtous à des corpus construits sur la même période à partir de l'indexation des *hashtags* #mariagegay et #mariagehomo.

2. Du contexte au corpus

2.1. Le contexte

Le 7 novembre 2012, le conseil des ministres français adopte le projet de loi sur le mariage pour tous dont l'objectif est de donner le droit de mariage et d'adoption aux homosexuels. Ce texte est une promesse de campagne du nouveau président François Hollande. De décembre au printemps, la France connaîtra de nombreuses manifestations de soutien et d'opposition au projet. Le texte sera débattu à l'assemblée à partir du 29 janvier 2013 pour finalement y être voté en seconde lecture le 23 avril 2013. Le premier mariage homosexuel a été célébré fin Mai 2013. Cette période sera particulièrement marquée par l'amplitude du mouvement d'opposition à la loi, réuni sous l'appellation « la manif pour tous » et par la violence, à la fois physique et verbale, associée à cette opposition. L'illustration 1 propose une lecture chronologique des événements de cette phase reportés sur un graphique des fréquences des tweets que nous analyserons.

2.2. Le corpus

De façon à suivre les conversations sur *twitter* liées à cette thématique, nous avons utilisé l'outil *yourTwapperkeeper*[1] pour indexer les *tweets* contenant le *hashtag* #mariagepourtous sur la période allant du 7 novembre 2012 au 31 juillet 2013. Parmi les différentes techniques disponibles pour étudier ce média, l'indexation par *hashtag* est la plus appropriée pour cibler une thématique (Rieder, 2010). Le corpus comprend 1 260 092 *tweets*. Le mode de fonctionnement de *twitter* fait qu'une partie importante de ces *tweets* (61,28 % dans ce cas) relève de ce que l'on appelle des *re-tweets*. Il s'agit d'une pratique qui consiste à répercuter auprès de ses *followers*[2] le message posté par un autre utilisateur. L'information contenue dans un *re-tweet* est donc la réplication exacte de l'information contenue dans un *tweet* précédent. Cette redondance n'a pas trop d'intérêt lorsqu'il s'agit de déterminer les thématiques abordées autour de l'utilisation d'un *hashtag* particulier. Elle serait même contre-productive dans le cadre des analyses auxquelles nous allons procéder. En effet, la classification utilisée repère rapidement les ensembles hyper-homogènes formés par les groupes de *re-tweets*, conduisant à l'apparition d'un très grand nombre de petites classes qui rendent plus difficile la perception des mondes lexicaux mobilisés. Si l'objectif est de déterminer les proportions exactes de chaque type de discours présents dans ce type de corpus, il est préférable de réinjecter les *re-tweets* a posteriori[3]. Nous ne travaillerons donc que sur les *tweets* originaux. Le corpus se compose alors de 453 107 *tweets*. Le taux de *re-tweets* étant relativement constant dans le temps et ne variant pas non plus avec la fréquence des *tweets* (voir illustration 2), ce corpus nous paraît tout à fait représentatif de l'ensemble.

Nous avons procédé à un nettoyage de ces textes, notamment en retirant systématiquement les URLs qui peuvent y apparaître et le *hashtag* qui a servi à l'indexation. Dans ce format, le corpus est composé de 6 474 386 occurrences (fmax : le (276 770)). Ce corpus contient 73 380 hapax (53.36% des formes - 1.13% des occurrences). Nous utiliserons la journée comme unité de temps. Le corpus s'étale alors sur 266 jours.

[1] https://github.com/540co/yourTwapperKeeper

[2] Sur twitter, on appelle *followers* les utilisateurs qui suivent les *tweets* d'un compte particulier.

[3] Cette phase ne sera pas abordée ici.

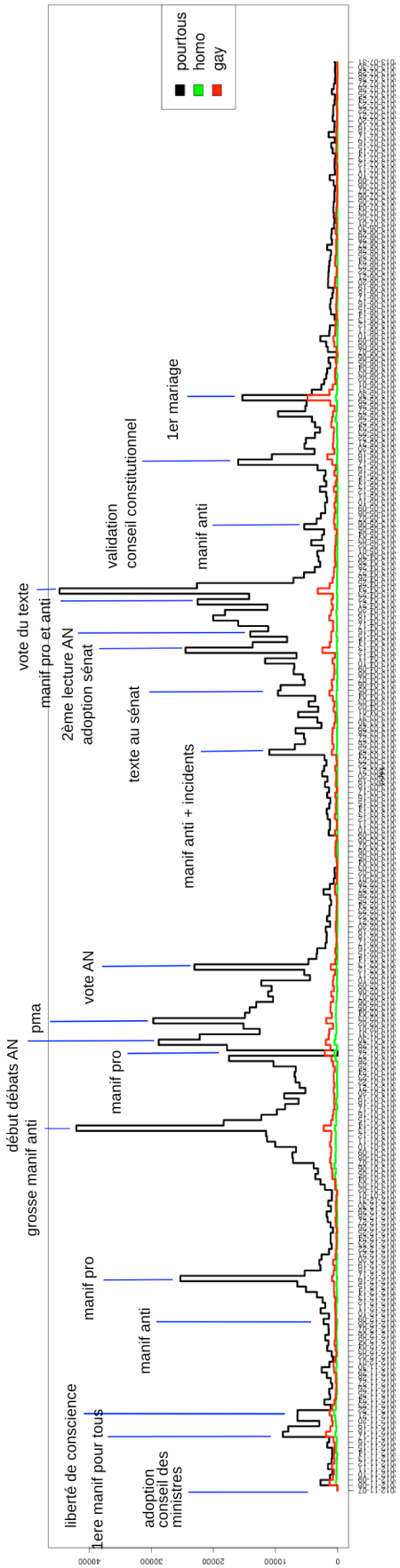


Figure 1. Fréquences des tweets liés aux hashtags #mariagepourtous, #mariagehomo et #mariagegay du 7/11/2012 au 31/07/2013

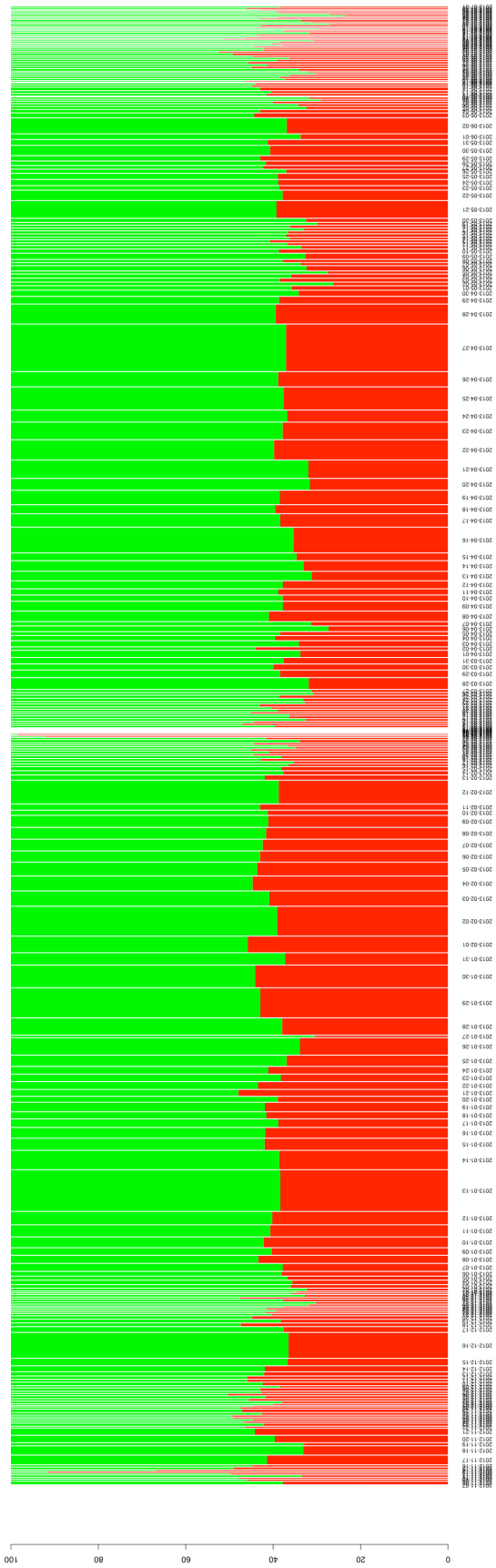


Figure 2. Evolution de la proportion de tweets originaux (en rouge) et de retweets (en vert) du 7/11/2012 au 31/07/2013 pour le hashtag #mariagepourtous. La largeur des colonnes est proportionnelle à la fréquence des tweets

3. Analyse

Nous avons soumis ce corpus à une analyse ALCESTE (Reinert, 1983, 1990) avec le logiciel IRaMuTeQ (Ratinaud et Déjean, 2009 ; Ratinaud et Marchand, 2012). L'analyse consiste en une classification hiérarchique descendante unique sur un tableau de présence/absence croisant sur les 453 107 *tweets* avec les 8 000 formes pleines les plus fréquentes. Après plusieurs tests, nous avons retenu une classification pour laquelle nous demandions 70 classes terminales. Nous avons conservé celles qui regroupent au moins 10 500 *tweets*. La classification retient au finale 88,7 % des tweets du corpus.

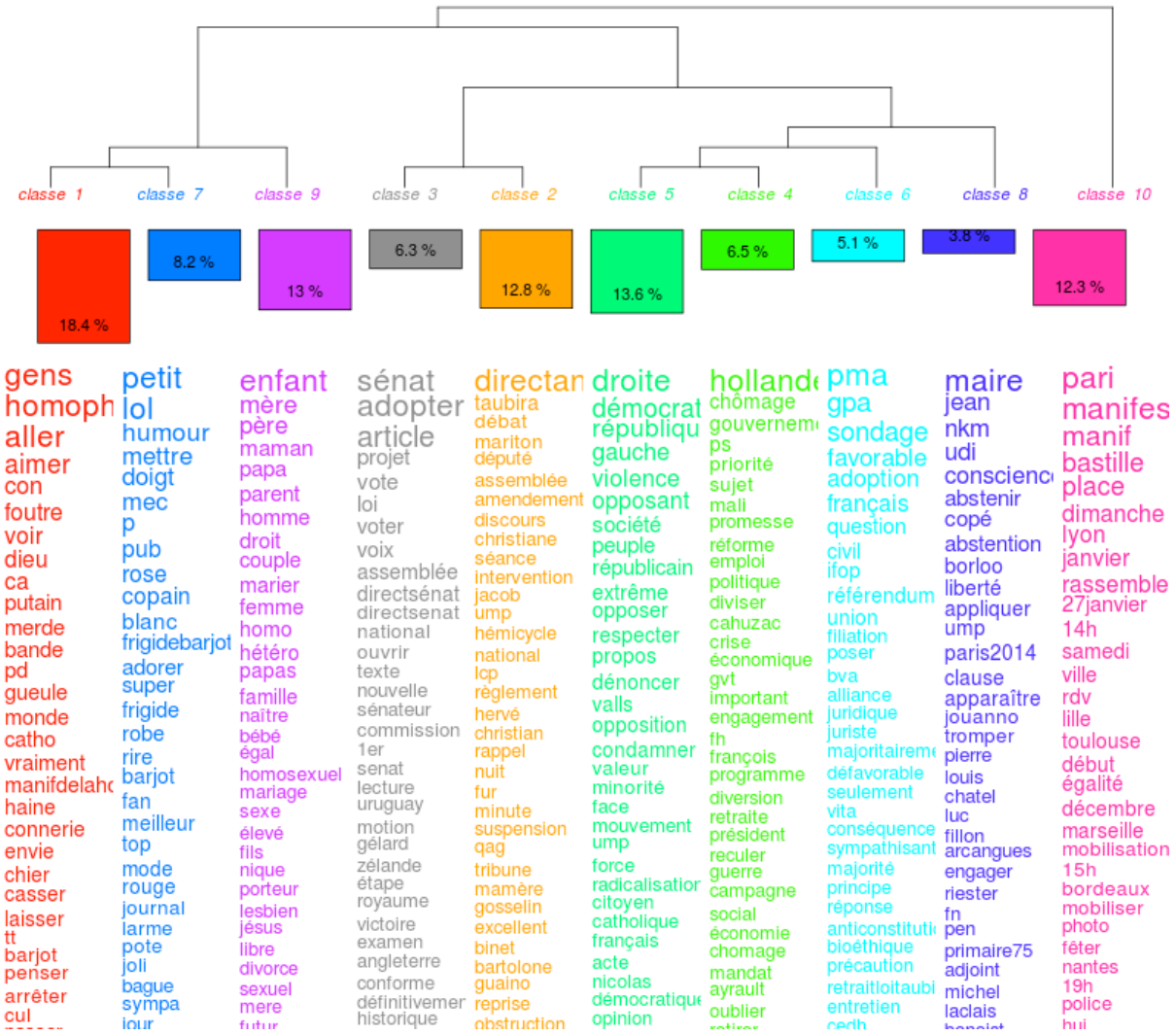


Figure 3. Dendrogramme de la classification, taille des classes et liste des mots les plus caractéristiques de chaque classes (par ordre décroissant des chi2 de liaison aux classes)

L'illustration 3 permet une lecture rapide du profil de ces classes et nous pouvons les résumer ainsi (de gauche à droite sur l'illustration 3) :

La classe 1 regroupe 18,4 % des tweets et correspond à la discussion directe et aux interpellations agressives entre les partisans et les opposants au projet. Le vocabulaire relève souvent du familier ou du vulgaire. Cette classe est fortement marquée par des pronoms

personnels (je, me vous, on, ils, tu...) et semble dominé par un discours d'opposition à la « manifpourtous ».

laissez les gens s aimer bande de pas humanisés ça vous fou quoi qu ils se marient ça va pas tuer le monde comme la bombe

a tous les gens qui se disent homophobe la loi du [mariagepourtous] est passée et allez vous faire foutre

j emmerde les homophobes enfin on avance en france les gens vont peut être enfin évoluer et arrêter leurs conneries[4]

La classe 7 regroupe 8,2 % des tweets. Elle fait référence à la partie « people » du traitement médiatique du mouvement social et de ses acteurs, notamment celui portant sur Frigide Barjot. Il s'agit aussi en partie d'interpellations à propos du mariage pour tous. C'est également une classe du « je ».

je mets les doigts partout pourquoi pas dans une bague les pro inspirés yann barthès petit journal

j y pense avec le mariagegay ça va êtres le bordel pour le jeux des 7 familles p humour blague lol mdr

frigide barjot grillée les doigts ds le nez le petit journal du 16 04 cc

La classe 9 regroupe 13 % des tweets et traite de la place de l'enfant dans les couples.

un enfant a droit à un papa et une maman va dire ça aux orphelins aux mères et pères célibataires tussa tussa tf1

ds ce cas les enfants ont aussi le droit d avoir 2 papas ou 2 mamans ou 1 père blanc et 1 mère noire ou inverse mariage gay

pour le droit de l enfant à être élevé par son père et sa mère ou à défaut par un homme et une femme droitsdesenfants

La classe 3 représente 6,3 % du corpus. Les tweets font références ici aux débats du Sénat.

le sénat adopte le 1er article du projet de loi sans modifier le texte de l assemblée nationale

ça y est a 179 voix ct 157 l article 1er du projet de loi concernant le [mariagepourtous] est adopté par le sénat ce soir directsénat

1er article du projet de loi mariage pour tous adopté par le sénat

La classe 2 regroupe 12,8 % des tweets et porte sur les débats à l'Assemblée Nationale.

brillante christiane taubira introduit le débat sur le [mariagepourtous] à l assemblée nationale directan

pour avoir enduré toutes ces heures de débat et h mariton on peut bien béatifier christiane taubira non directan

a vendre 128 dvd interventions députés ump débat mariagepourtous bonus lefur mariton poisson en boucle encore une fois directan

La classe 5 est composée de 13,6 % des *tweets*. Ces textes dénoncent les violences verbales et physiques liées à ce contexte et commentent différentes dimensions politiques (rôle de la droite dans l'opposition, rapport à la démocratie...).

la démocratie selon le peuple de droite la gauche au pouvoir est illégitime éditodumonde

[4] Nous proposons pour chacune des classes trois messages parmi les plus caractéristiques de la classe. Les mots en rouge sont des formes pleines présentes dans le profil des classes.

après l adoption de la loi par les 2 assemblées de la république la droite continue à éructer contre déni de la démocratie

la violence prend le pas sur le débat républicain selon des méthodes coutumières aux mouvements d extrême droite lyon

La classe 4 regroupe 6,5 % des tweets et interroge l'intérêt de cette loi dans un contexte de crise.

non cumul des mandats nbre des députés priorités de hollande chômage prise d otages guerre au mali on s en tape

au lieu de parler du chômage et de la compétitivité le gouvernement préfère parler du [mariagepourtous] affligeant ps hollande

chômage crise économique la france aux abois mais pendant ce temps le mali et le [mariagepourtous] monopolisent les médias hollande

La classe 6 représente 5,1 % des messages. Elle porte sur les thématiques de la procréation médicalement assistée et de la gestation pour autrui, notamment par le biais des *tweets* commentant les sondages sur ces questions.

sondage ifop 39 des français favorables au mariage adoption pma gpa

sondage les français ne veulent pas de filiation fiction non adoption pma gpa

nouveau sondage bva 43 en faveur du les français découvrent l enfumage adoption pma gpa

La classe 8 regroupe 3,8 % des tweets. La thématique est ici la liberté de conscience (pour les maires d'appliquer la loi et pour les députés UMP et UDI de voter le texte). Il est aussi question du vote pour la direction de l'UMP.

jean bizet françois hollande avait parlé de liberté de conscience pour les maires qui refusent le directsénat

jc fromentin maire de neuilly sur mediaparlive je n engage pas le groupe udi chacun a une liberté de conscience

ump mariagegay abstention luca philippe nkm lellouche le maire pour riester apparu

Enfin, la classe 10 porte sur les manifestations et regroupe 12,3 % des *tweets*.

paris jour j début manifestation 14h place de la bastille mobilisation rt pmlive mon16dec

60 000 pers à la manifestation aujourd hui à paris police c est peu pour une manif nationale rdv le 13 janvier

dimanche à la manifestation pour l égalité des droits 14h place denfert rochereau à paris

Notons pour finir que certaines de ces classes semblent contenir encore énormément de variabilité.

4. Représentation de la chronologie

Le mode classique de représentation graphique de ce type de résultat passe par l'utilisation d'une analyse factorielle des correspondances. L'illustration 4 montre la projection des 266 dates sur le plan factoriel construit à partir de l'analyse menée sur le tableau de contingence croisant les formes pleines et les classes terminales. Bien que ce plan permette une première analyse, l'unicité des points marque ici fortement la sur-représentativité d'une date dans une classe, mais cache l'évolution des classes dans le temps. Nous pouvons par exemple voir sur la gauche du graphique (qui correspond à la position de la classe sur les manifestations) des dates qui correspondent effectivement à des manifestions (19 janvier et 16 décembre par exemple).

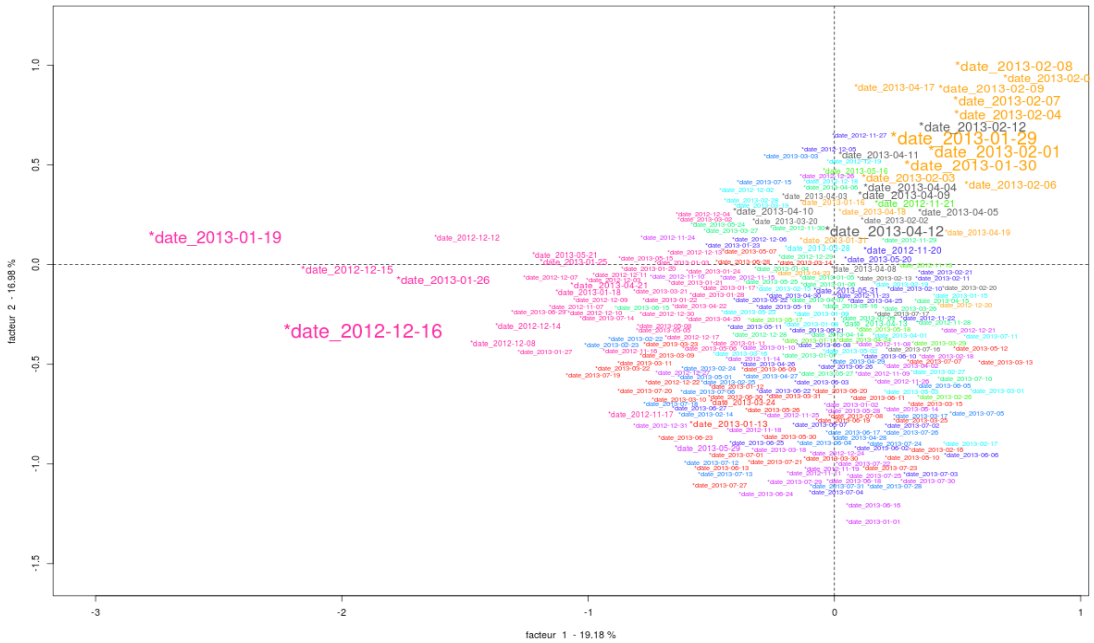
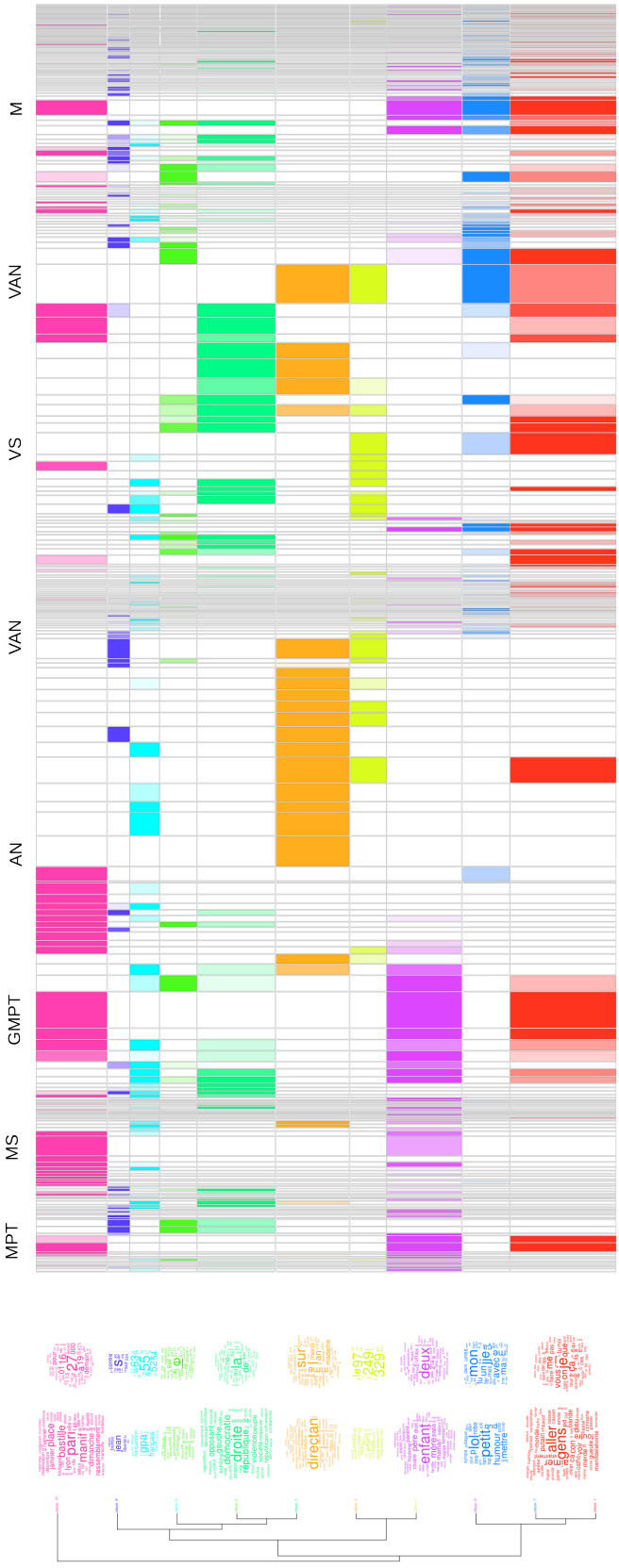


Figure 4. Plan factoriel des variables « dates »

L'illustration 5 projette sur le dendrogramme de la classification la sur-représentativité des dates dans les classes. Pour chacun des jours de l'indexation, nous utilisons le seuil de significativité du chi2 de liaison de la date aux classes pour déterminer l'intensité de la couleur. Les cases blanches correspondent à des seuils supérieurs à 0.05. Les classes les plus foncées correspondent à des seuils inférieurs à 0,00001. Un gradient de couleur est établi entre ces deux valeurs. Une couleur foncée indique donc une forte sur-représentation des *tweets* relevant de la thématique pour le jour considéré. Par ailleurs, cette représentation permet à tout moment de l'analyse de garder en tête la proportion de chaque classe (par la hauteur des lignes) et l'importance des communications de chaque journée (par la largeur des cellules). Ce graphique conduit en première intention à prendre conscience de la nature sociale des interactions sur *twitter* : les ensembles de cellules peu épaisses (notés 1, 2 et 3 sur le graphique) correspondent en effet à des périodes de vacances scolaires (1 : vacances de noël, 2 : vacances d'hiver, 3 : vacances de printemps). Dans ces moments, l'intensité des *tweets* diminue fortement.



MPT (17/11/2012) = 1ère manifestation des opposants (« manifpourtous »)

MS (16/12/2012) = manifestation de soutien au projet

GMPT (13/01/2013) = grande manifestation d'opposants

AN (29/01/2013) = début des débats à l'assemblée nationale

VAN (12/02/2013) = premier vote à l'assemblée nationale

VS (12/04/2013) = Vote au sénat

VAN (23/04/2013) = second vote à l'assemblée nationale

M (29/05/2013) = premier mariage homosexuel à Montpellier

Figure 5. De gauche à droite - Dendrogramme de la classification, nuage des formes pleines, nuage des formes supplémentaires et graphique de la sur-représentation des classes par jour. La hauteur des lignes est proportionnelle à la taille de la classe. La largeur des cellules est proportionnelle à la fréquence des tweets sur la journée. L'intensité des couleurs est proportionnelle à l'intensité de la sur-représentation de la classe.

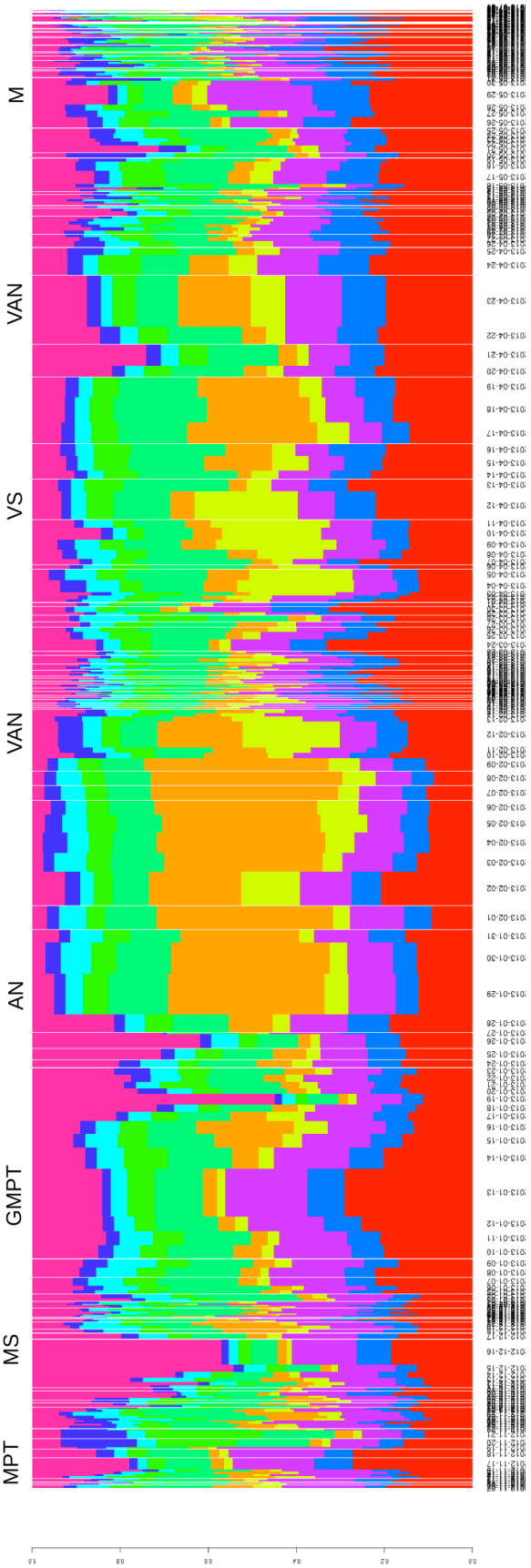


Figure 6. Proportion de chacune des classes de la classification pour chacun des jours du 7/11/2012 au 31/07/2013. La largeur des barres est proportionnelle à la fréquence des tweets. Les couleurs et l'ordre des classes reprennent les couleurs et l'ordre de l'illustration 3.

L'illustration 6 permet de visualiser la présence quotidienne de chacune des classes de discours. Toutes les thématiques sont en effet présentes à chaque moment de l'indexation. Ce qui change, c'est l'intensité de leur expression.

De façon à analyser cette chronologie, nous distinguerons 4 périodes :

La première période commence au début de l'indexation et s'arrête au début des débats à l'assemblée (AN dans le graphique). Dans cette période, les manifestations pro et anti s'enchaînent. Cela est clairement visible sur la distribution de la classe 10 sur l'illustration 5 (classe sur les manifestations en rose ; première ligne). Nous pouvons alors remarquer que les manifestations des opposants conduisent systématiquement à une sur-apparition de la classe 1 (classe de la dénonciation de l'homophobie en rouge ; dernière ligne). Ce n'est pas le cas pour les manifestations de soutien au projet (période notée MS et juste avant AN). L'illustration 6 montre également clairement ce phénomène : les manifestations de soutien ont conduit à une intensité de communication sur ce *hashtag* bien supérieure à celle atteinte lors des manifestions des opposants. C'est également dans cette période que les arguments développés par les opposants sur la place de l'enfant dans le couple sont sur-représentés (classe 9 en violet, 3ème ligne en partant du bas).

La seconde période est celle des débats à l'assemblée (de AN à VAN). Elle est évidemment dominée par la classe sur les débats à l'assemblée nationale (classe 2 en orange). Cela est particulièrement visible sur l'illustration 6. On peut voir sur ce graphique que la classe des débats à l'assemblée, qui est une des plus importantes de l'analyse, se concentre effectivement sur les deux périodes de débat à l'assemblée.

La troisième période couvre les débats au sénat et la seconde lecture à l'assemblée (du premier VAN au second VAN). Dans cette période on retrouve bien sûr les classes des débats à l'assemblée et au sénat (classe 3 en jaune), mais également la classe sur la dénonciation des violences (classe 5 en vert foncée, 5ème ligne). C'est en effet dans cette période qu'une partie du mouvement de la « Manif pour tous » se radicalise et que des agressions homophobes sont perpétrées et médiatisées. La position de la droite dans le mouvement d'opposition, au côté des partis et des groupuscules d'extrême droite est débattue.

Enfin, la quatrième période court du vote à l'assemblée à la fin de l'indexation. Cette période est marquée par la sur-apparition de la classe liée aux traitements satyriques et people de l'opposition au projet (classe 7 en bleu foncée, deuxième ligne en partant du bas). On note également sur la première ligne les dernières manifestations des opposants qui conduisent une nouvelle fois à une sur-représentation de la classe 1, comme lors de la première période. Le premier mariage organisé à Montpellier (noté M) donnera lieu à des manifestations et à la sur-représentation des discours sur la place de l'enfant dans le couple.

5. Intérêt pour la comparaison des corpus

Nous avons soumis deux autres corpus de tweets, prélevés sur la même période, au même traitement. Le tableau 1 présente succinctement ces corpus. Il s'agit des *tweets* contenant les *hashtags* #mariagehomo et #mariagegay

	Nombre de tweets	Taux de retweets	Nombre de tweets dans le corpus texte	Occurrences	Hapax
#mariagegay	137501	66,13 %	42069	632739	18407 - 52.03 % des formes - 2.91 % des occurrences
#mariagehomo	22223	67,36 %	6855	100953	6559 - 55.38 % des formes - 6.50 % des occurrences

Tableau 1 : présentations des corpus #mariagegay et #mariagehomo

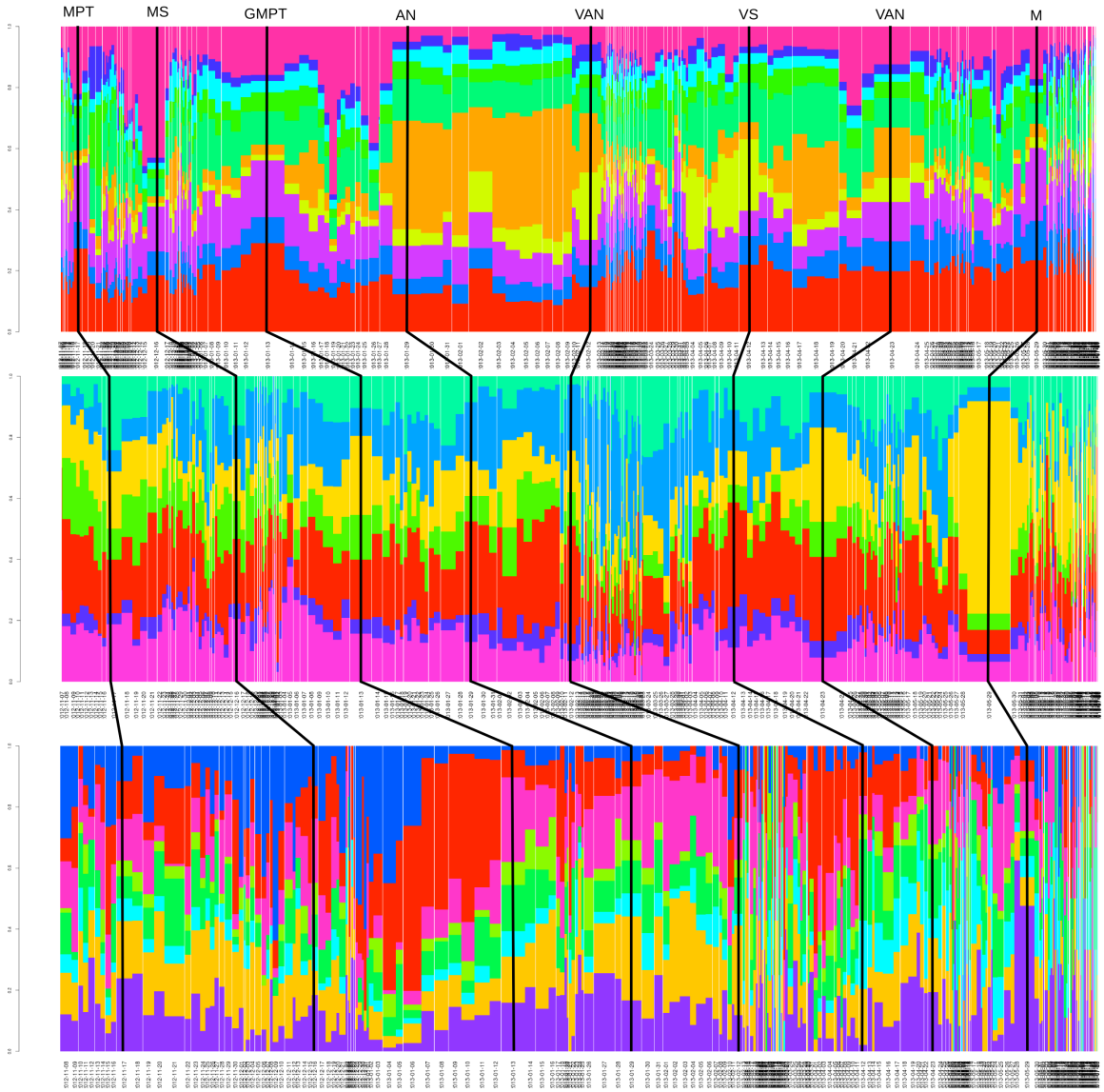


Figure 7. Comparaison des corpus #mariagepourtous, #mariagegay et #mariagehomo

La première information apportée par l'illustration 7 est que ces trois corpus ne se situent pas dans la même « temporalité », si l'on rapporte celle-ci à la fréquence quotidienne des tweets utilisant chacun de ces *hashtags*. Par exemple, le traitement des débats (entre AN et VAN) n'a pas du tout la même importance dans le corpus #mariagepourtous que dans les corpus #mariagehomo et #mariagegay. Proportionnellement, ces deux hashtags ont été plus utilisés dans la première période de l'analyse précédente (période des manifestations) que celui du #mariagepourtous. Le détail des classifications semble attesté d'un investissement différent de ces *hashtags*, ce qui a été montré dans d'autres analyses (Cervulle et Pallier, 2014).

6. Conclusion

Tous les textes ont une chronologie, un début et une fin. De ce fait, tous les corpus peuvent être étudiés dans leurs rapports à une temporalité. Le développement des réseaux socio-numériques et l'accès facilité aux bases de données textuelles, en permettant la constitution de corpus de grande taille, rendent encore plus prégnante la nécessité de la prise en compte de la dimension temporelle de la production textuelle. Les modes de représentations que nous proposons ici, qui reposent sur la projection de classifications ALCESTE sur des échelles chronologiques, mettent effectivement en évidence les variations dans l'expression de classes de discours. Elles permettent de compléter la lecture figée que nous imposent les dendrogrammes et les analyses factorielles. Dans l'exemple que nous venons de développer, la visualisation de cette temporalité rend saillante une corrélation forte entre les variations lexicales et les agendas sociaux, politiques et médiatiques.

Références

Cervulle M. et Pallier F. (2014). #mariagepourtous : Twitter et la politique affective des hastags. *Revue française des sciences de l'information et de la communication*. 4. Retrieved from http://rfsic.revues.org/717.

Ratinaud P. et Déjean S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. *Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009)*. Toulouse - Le Mirail.

Ratinaud P. et Marchand P. (2012). Application de la méthode ALCESTE à de “gros” corpus et stabilité des “mondes lexicaux” : analyse du “CableGate” avec IRaMuTeQ. In *Actes des 11eme Journées internationales d’Analyse statistique des Données Textuelles* (pp. 835–844).

Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données.*VIII, (2), 187-198.

Reinert M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologique.* 26. 24-54.

Rieder B. (2012). The refraction chamber: Twitter as sphere and network. *First Monday.* 17. (10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4199/3359.