Débats sur la toile

Jessica Mange[1], Pascal Marchand[2], André Salem[3]

[1] IUT GEA - Boulevard Maréchal Juin - 14032 Caen Cedex

[2] LERASS – 115 D route de Narbonne - BP 67701 - F-31077 TOULOUSE

[3] SYLED-CLA2T, Sorbonne nouvelle – Paris 3, 19 rue des Bernardins, 75005 Paris

Abstract

We have analyzed a corpus of postings on a French website forum dealing with the referendum about the European Constitution. It was gathered over a 6-week period before the May 29, 2005 ballot. We first showed what makes this kind of corpus specific and outlined the research openings it raises in the field of intertextuality. Several textometric methods of analysis, applied to various sections of the corpus, shed a quantitative light on the debate and help clarify its dynamic.

Résumé

On analyse un corpus rassemblant les contributions de participants à un forum ouvert sur un site web français, à propos du projet de référendum sur la constitution européenne durant les six semaines qui ont précédé le scrutin (29/05/05). On commence par montrer la spécificité de ce genre de corpus et les directions de recherche qu'il ouvre sur les questions de l'intertextualité. Plusieurs méthodes d'analyse textométriques opérant sur des partitions différentes du corpus permettent ensuite d'éclairer le débat sous un angle quantitatif et d'en mettre à jour la dynamique.

Mots-clés : textometry, intertextuality.

1. Introduction

À la suite d'une prise de position en faveur du «oui » de son rédacteur en chef, au cours de la campagne précédant le référendum sur le *traité établissant une constitution pour l'Europe*[1] et des réactions suscitées chez ses lecteurs par cette prise de position, l'hebdomadaire *Télérama* ouvrait sur le web, entre le 22.04.05 et le 22.06.05, un forum de discussion[2] sur le thème : *l'Europe : oui ou non* ?

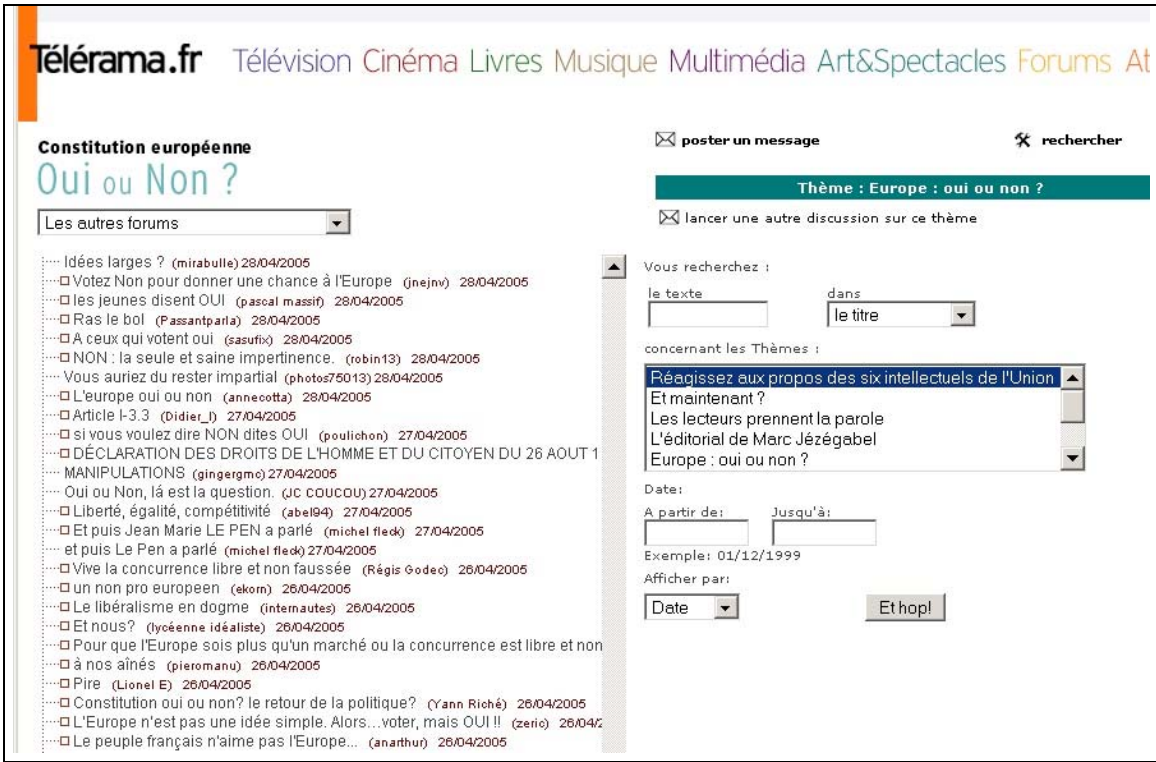
Tout au long de cette période, 2002 messages ont été publiés sur ce forum par des internautes désireux d'exprimer leur propre opinion ou de réagir à des idées exprimées par d'autres. Il s'en est suivi un riche débat que nous nous proposons de soumettre ici à une étude textométrique. Le but serait de mettre en évidence, à partir de ce corpus, la dynamique d'un débat mené par des citoyens, et non par professionnels du débat politique ou journalistique, d'en dégager l'évolution chronologique ainsi que les principaux arguments en faveur du « oui » et du « non ».

[1] Le 29 mai 2005, les citoyens français ont été appelés à se prononcer sur un *Traité établissant une constitution pour l'Europe.* La campagne qui a précédé le vote a donné lieu à une polémique très vive. Contrairement à ce qui s'est passé dans d'autres pays de l'Union européenne, les électeurs français ont finalement rejeté la proposition.

[2] http://forums.telerama.fr/forums.asp?forum=147

2. Le corpus *Forum*

Le corpus compte 336 080 occurrences au total pour 19 370 formes. Les messages sont généralement assez courts, avec une moyenne de 74 occurrences. Rares sont ceux qui dépassent les 500 occurrences, mais les plus longs peuvent exceptionnellement dépasser les 2000 occurrences. En s'appuyant sur les dates de publication, il a été possible d'introduire des balises indiquant le jour et la semaine de la mise en circulation du message.



Extraits de messages balisés

§ <jour=02> <sem=17><opinion=non><loc=non/17> votre édito et bien d'autres ne font que renforcer mon envie de crier :- peuples d'europe, réveillez-vous ! la démocratie se meurt et cette constitution sera son tombeau. - avec ce traité, plus de - dynamique européenne-, plus de - projet européen -. citez-moi les articles où l'on parle des peuples, d'un peuple européen, d'un projet européen, d'un espace européen... allez, je vous laisse avec votre conscience \...\

§ <jour=03> <sem=17> <opinion=oui><loc=oui/17> les réactions sur le referendum, sur les forums, dans les cafés, à la maison le montrent, la politique n'est pas morte. mais quelle politique? et chacun de voir dans la constitution les prémisses d'une politique libérale ou sociale. pourtant c'est une constitution pas un vote pour un choix politique, plutôt une manière de s'unir différente (c'est néanmoins un choix politique sur la construction européenne). \...\

Figure 1 : *États successifs du corpus* **Forum**

Pour comparer les arguments des deux partis en présence, on a procédé à un codage des messages[3]. Après lecture de chacun des messages[4], nous avons introduit une balise <OPINION=oui> devant les messages des internautes qui appelaient à se prononcer positivement lors de la consultation à venir, une balise <OPINION=NON> devant les messages qui prônaient au contraire le vote négatif. Seuls 71 messages (3,5%) n'ont pu être clairement attribués à l'un des deux camps, nous les avons éliminés du corpus pour ces premières expériences. Les messages conservés se répartissent en 1 000 « pro-oui » (48,1% des messages et 146 784 occurrences) et 1 006 « pro-non » (48,4% des messages et 180 906 occurrences).

On trouve sur la figure 1 un aperçu du corpus sous sa forme *native* (i.e. tel qu'il se présente sur le web) et un exemple de messages munis des balises qui permettront l'exploration textométrique du corpus.

3. l'intertextualité au bout du clavier

Le genre *dialogue sur un forum web*, présente une originalité certaine par rapport aux débats de différents types (face à face télévisuels, correspondances écrites, etc.) qui ont été étudiés jusqu'à présent à l'aide des méthodes textométriques[5]. Pour pouvoir s'exprimer sur un forum, il faut être capable de rédiger des interventions à partir d'un clavier d'ordinateur. La pratique du traitement de texte, inséparablement associée à l'accès au forum, permet aux participants de reprendre plus facilement que dans un débat sous forme orale les séquences de textes produites par les locuteurs précédents pour les citer en les approuvant ou en les critiquant, en les reproduisant ou en les modifiant à l'aide de simples *copier/coller*.

Sur un forum, les participants s'expriment ou réagissent à une ou à plusieurs interventions déjà *postées* sur le site, sans que l'on puisse toujours savoir s'ils ont pris le temps de lire, et a fortiori de comprendre, les interventions qui ont précédé leur propre message.

Le forum constitue un outil privilégié pour observer la circulation des mots, des séquences de mots, expressions, concepts créés pour l'occurrence ou mobilisés à partir de la communication socio-politique du moment. Il peut constituer un laboratoire d'expériences extrêmement précieux pour étudier la circulation des unités textuelles entre locuteurs et à travers les périodes temporelles, pour observer de ce qu'on appelle la *dimension intertextuelle* du discours.

Les segments répétés du corpus *Forum*

Cette circonstance se révèle très importante lorsque l'on soumet des textes recueillis sur un forum à des études textométriques. Le repérage des *segments répétés*[6] (séquences composées

[3] Dans cette préparation du corpus, les caractères majuscules ont été ramenés à la minuscule correspondante, les balises introduites prennent les valeurs suivantes :

<jour=jj>	indique le jour de l'intervention [entre 2 et 41]
<sem=ss>	indique numéro de la semaine de l'intervention [entre 17 et 22]
<opinion=xx>	indique l'opinion globale émise par le participant [*oui* ou *non*]
<loc=oo/ss>	combine l'opinion et le numéro de la semaine

[4] Remarquons que l'opinion exprimée par chacun des messages peut-être identifiée, dans un grand nombre de cas, par le repérage automatique de segments comme : *j'appelle à voter /je vote/ je voterai oui (*resp. *non)* ou *des/les partisans/tenants du oui (*resp. *non).*

[5] Un numéro de *Langage et Société* a été consacré aux écrits électroniques (Fraenkel et Marcoccia, 2003). Pour des analyses textométriques sur ce sujet, cf., par exemple, les analyses sur le débat télévisé *Mitterrand-Chirac* pour les présidentielles de 1988 (Salem, 2004).

[6] Pour un exposé sur les segments répétés, cf., par exemple, Salem (1987).

de plusieurs formes attestées à plusieurs endroits dans le corpus) du corpus ***Forum*** fait apparaître un nombre important de longues séquences répétées dont on vérifie facilement qu'elles sont dues à la reprise intégrale par certains des intervenants de longues séquences déjà produites, par eux-mêmes ou par d'autres participants au forum.

La forme la plus intéressante de circulation de telles séquences est sans doute celle qui résulte de la formation de concepts propres à la période dans laquelle le débat prend place comme :

la notion de service public	14	occ.
la libéralisation des services	12	occ.
fonctionnement du marché intérieur	11	occ.
cohésion sociale et territoriale	6	occ.
sous protectorat militaire américain	6	occ.
un marché intérieur où la concurrence est libre et non faussée	4	occ.

ou de séquences, beaucoup plus fréquentes, permettant de désigner les acteurs et les enjeux du débat en cours, comme :

les partisans du oui	78	occ.
les partisans du non	61	occ.
si le oui l'emporte	47	occ.
les tenants du non	15	occ.
si le non l'emporte	14	occ.
les tenants du oui	13	occ.

Mais beaucoup de ces répétitions relèvent plutôt de la reprise d'un texte précédent à des fins polémiques. Le discours ainsi rapporté est parfois encadré par des guillemets mais, le plus souvent, reproduit sans marques de présentation particulières.

> § <jour=17><sem=19><opi=non><loc=non/19> cher vous .... avez-vous lu la constitution j'en doute ! pour l ' irak on était seul , on a fait des adepte , en 1789 sur les droits de l' homme , on était seul aussi et vous connaissez la suite . \
>
> § <jour=17><sem=19><opi=oui><loc=oui/19> bonjour, pour l ' irak on était seul, on a fait des adepte, et ça a servi a quoi ? (a part montrer qu ' on a une grande gueule) . puisque vous parlez de l ' irak, imaginons ce qu ' il se serait passe avec le tec:

Souvent, les fautes d'orthographe conservées intactes dans la réponse témoignent, s'il en était besoin, du rôle du *copier/coller* dans la genèse de la réponse.

> § <jour=03> <sem=17> <opi=oui><loc=oui/17> juste un petit détail qui a son importance... il faut peut-être préciser une chose : les français n'ont jamais été européen... il est plutôt conservateur, xénophobe, pédant, arrogant... les français voudraient que l'europe soit française un point c'est tout... ils n'aiment pas les compromis, la discussion... ils préfèrent les débats caricaturaux, les invectives, les violences et le chantage /…/
>
> § <jour=03> <sem=17> <opi=non><loc=non/17> - : les français n'ont jamais été européen... il est plutôt conservateur, xénophobe, pédant, arrogant.. - vous en avez d'autres en stocks, des principes racistes du genre ? a priori, généralement ce sont les anglais que l'on taxe de conservatisme et les allemands que l'on taxe de xénophobie. - ils n'aiment pas les compromis, la discussion. ils préfèrent les débats caricaturaux, les invectives, les violences et le chantage...et vous en arrivez à cette conclusion parce

> que votre europe moisie, votre europe anti-sociale, ne fait pas l'unanimité ? c'est cela, votre sens de la discussion, du débat ?

Certains messages font intervenir des séquences empruntées au document qui fait l'objet du débat, en citant de larges extraits du projet de constitution :

> § <jour=08> <sem=17> <opi=oui><loc=oui/17> /…/ je le répète, je ne suis pas scandalisée qu'un état intervienne financièrement pour sauver une entreprise et des emplois, même si ça - fausse - la concurrence. article i-44 les états membres qui souhaitent instaurer entre eux une coopération renforcée dans le cadre des compétences non exclusives de l'union /…/

On trouve aussi des interventions reproduites pratiquement à l'identique par un même participant, à quelques jours d'intervalle, sans doute dans l'espoir d'étendre l'impact du premier billet à quelques lecteurs supplémentaires auxquels il aurait échappé lors de sa première publication.

> § <jour=09> <sem=18> <opi=oui><loc=oui/18> fraîchement de retour en france après avoir passé une année à voyager entre le timor oriental et istanbul, je ne comprends pas bien le débat qui anime la france...
>
> § <jour=12> <sem=18> <opi=oui><loc=oui/18> de retour en france après avoir passé une année à voyager entre le timor oriental et istanbul, je ne comprends pas bien le débat qui anime la france...

Dans les deux interventions partiellement reproduites ci-dessus, le premier mot du texte a été modifié mais les 351 formes qui suivent sont absolument identiques, ce qui fournit une preuve que la seconde intervention ne procède pas d'une nouvelle rédaction.

4. Une approche chronologique

Comment évolue le vocabulaire des participants tout au long du débat ? Pour répondre à cette question, nous allons considérer une partition du corpus en 6 semaines[7]. La figure 2 montre le plan correspondant aux deux premiers facteurs issus de l'analyse factorielle des correspondances (AFC) réalisée à partir du décompte des formes dont la fréquence est au moins égale à dix occurrences dans l'ensemble du corpus[8]. Comme on le voit sur cette figure, les semaines consécutives occupent sur les axes des positions globalement proches. Le premier axe (horizontal sur la figure) restitue intégralement la chronologie du corpus. Cela traduit le fait, souvent constaté dans le cas de l'étude des séries textuelles chronologiques ou simplement de corpus longitudinaux, que le vocabulaire employé par les participants évolue progressivement dans le temps[9].

Après identification du schéma général de l'évolution progressive vocabulaire au fil des semaines, il devient possible de caractériser chacune des parties (ou groupe de parties) par le vocabulaire et les séquences qu'elle privilégie et par celui qu'elle évite. Par exemple, les

[7] Cette partition peut être réalisée en sélectionnant la clé <sem=ss> introduite plus haut.

[8] Pour une présentation de l'analyse factorielle des correspondances appliquée aux corpus textuels, cf., par exemple, Lebart & Salem (1994).

[9] On reconnaît ici le schéma général de l'analyse des *séries textuelles chronologiques*. Par suite de l'inversion des axes 2 et 3, dont les valeurs propres correspondantes sont très proches, c'est la projection sur les axes 1-3 qui fournit la représentation classique *en parabole* qui caractérise habituellement l'effet Guttman produit par l'évolution progressive du vocabulaire. Signalons de plus que la même analyse réalisée à partir des interventions regroupées par jour d'émission fournit des résultats semblables sur les journées, bien que l'effet chronologique soit un peu moins net, comme on peut s'y attendre s'agissant d'une typologie réalisée à partir de fragments de taille plus faible. Pour des compléments sur les *séries textuelles chronologiques*, on consultera Salem (1993).

unités textuelles les plus employées dans les deux premières semaines se révèlent être les formes qui réfèrent en principe à l'objet du débat : *constitution, traité, traité constitutionnel, l'union, politique*. Ces formes sont nettement moins utilisées dans les semaines qui correspondent à la fin du débat sur le forum.

Dans ces dernières périodes, on note au contraire une plus grande utilisation des termes : *tu, vous, dimanche, docteurs, patient, remède*. Que l'on peut interpréter comme la marque d'un discours plus polémique, plus centré sur l'imminence de l'échéance électorale et dans lequel les enjeux principaux ont déjà été décrits.

Le recensement des segments répétés les plus longs (par exemple, ceux dont la longueur dépasse dix formes) dans les textes relevant des différentes semaines indique également que le nombre des longues citations, auxquelles les internautes ont souvent eu recours dans les premières semaines du débat, diminue fortement dans les dernières parties.

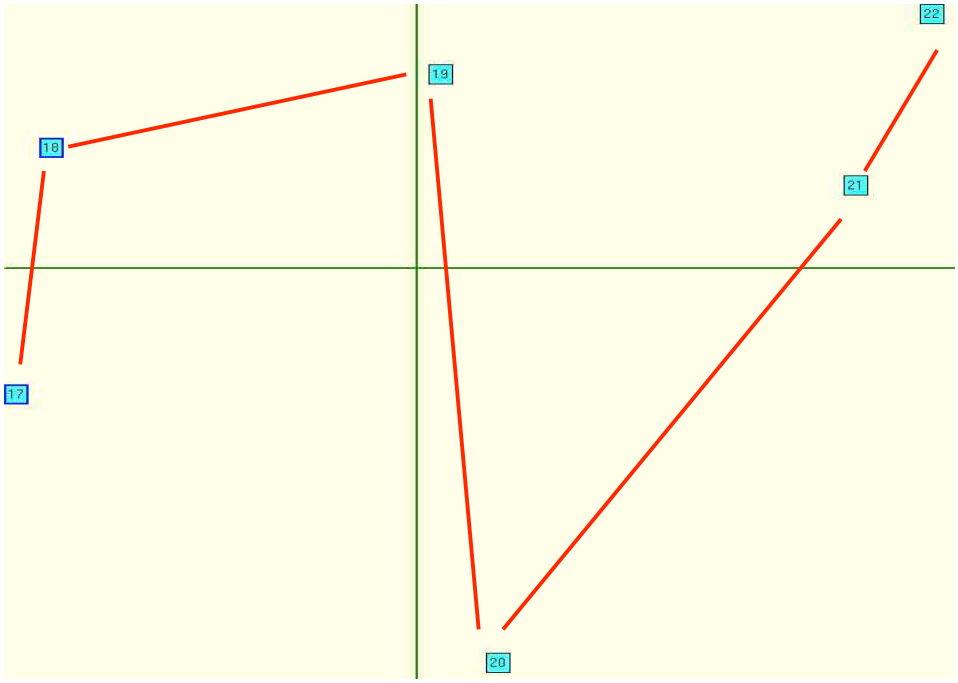


Figure 2 :
Analyse des correspondances réalisée à partir du tableau 6 semaines x formes de fréquence ≥ 10

5. Les *pour* et les *contre*

La deuxième dimension du corpus qu'il nous faut explorer concerne les différences dans l'emploi du vocabulaire que l'on peut constater chez les internautes selon qu'ils ont choisi l'une ou l'autre des réponses à la question posée par le référendum à venir.

Parmi les formes et les segments caractéristiques des partisans du *oui* on relève :

> *vous, gauche, fabius, compromis, de gauche, nons, l ue , que vous , monde, france, tenants du non, votre, villiers, extrême gauche…*

Les partisans du *non* privilégient de leur côté les termes :

> *constitution, partisans du oui, je voterai non, euros, les partisans du oui, appelle à voter non, démocrate, médias*

À partir des vocabulaires dégagés comme caractéristiques pour les deux groupes de participants, nous allons construire deux unités textuelles que l'on appelle *types généralisés* ou *tgen*[10]. Sur la figure 3, on a représenté la ventilation de ces deux *tgens* que l'on appellera *tgens spécifiques* dans les textes correspondant aux 6 semaines qui composent le corpus.

Le tgen *SP-oui*, plus sombre sur la figure 3, rassemble toutes les occurrences du corpus qui correspondent à une forme spécifique pour le groupe de participants qui se prononce pour le *oui*. Le tgen *SP-non*, affiché dans une couleur plus claire, rassemble celles des occurrences qui correspondent au contraire à des formes spécifiques pour le groupe des participants qui se prononce pour le *non*.[11]

Comme on le voit sur cette figure chacun des deux camps connaît un moment d'expression plus marquée des termes qui font sa spécificité (la semaine n°18 pour le camp du *non*, la semaine n°21 pour celui du *oui*). Autre fait remarquable, le camp du *non* délaisse dans cette même semaine 21 le vocabulaire spécifique notamment mis en place lors de la semaine 18. La semaine 21 apparaît comme un moment très important dans l'affrontement entre les deux camps qui fournissent durant la semaine considérée des prestations équivalentes du point de vue du volume de leurs contributions.

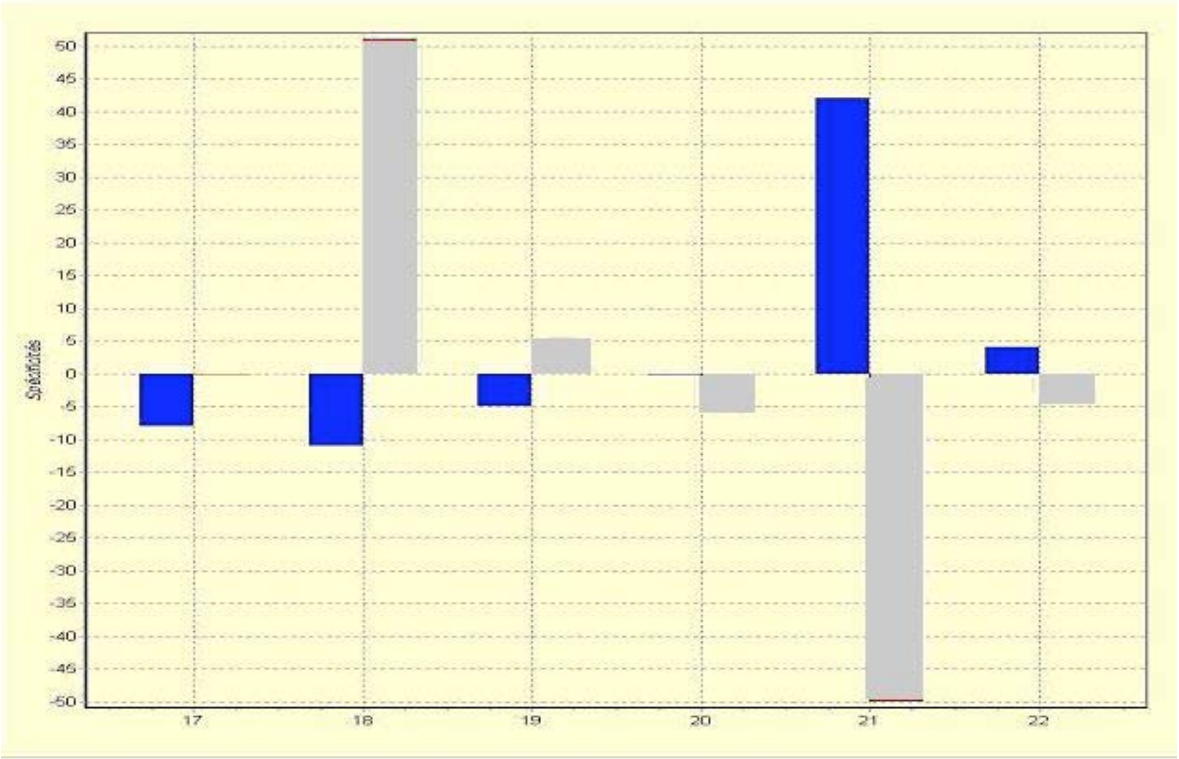


Figure 3 :
*Ventilation des tgens spécifiques **SP-oui** et **SP-non** dans les 6 semaines du corpus*

[10] Le Tgen « ensemble d'occurrences sélectionnées parmi les occurrences du texte » généralise la notion de type unité dont on peut recenser les occurrences dans le texte. Sur les types généralisés, on consultera Lamalle & Salem (2002).

[11] Nous avons sélectionné, pour cette expérience les unités qui ont un indice de spécificité supérieur à 4 (i.e. celles pour lesquelles la méthode des spécificités renvoie une probabilité inférieure à 1/10000^{e}).

6. Analyser le débat

Nous allons tenter d'approfondir ces constats à partir d'expériences portant sur une partition du corpus en 12 parties (6 semaines x 2 opinions), c'est-à-dire que chaque partie rassemble dorénavant des messages produits au cours d'une même semaine par des internautes d'opinions identiques. L'analyse des correspondances présentée sur la figure 4 a été réalisée à partir de cette dernière partition et ne retenant que les formes dont la fréquence est supérieure ou égale à 10. On retrouve sur ce graphique l'évolution chronologique vue au point précédent, déclinée cette fois en deux sous-ensembles dont chacun correspond à l'une des opinions exprimées.

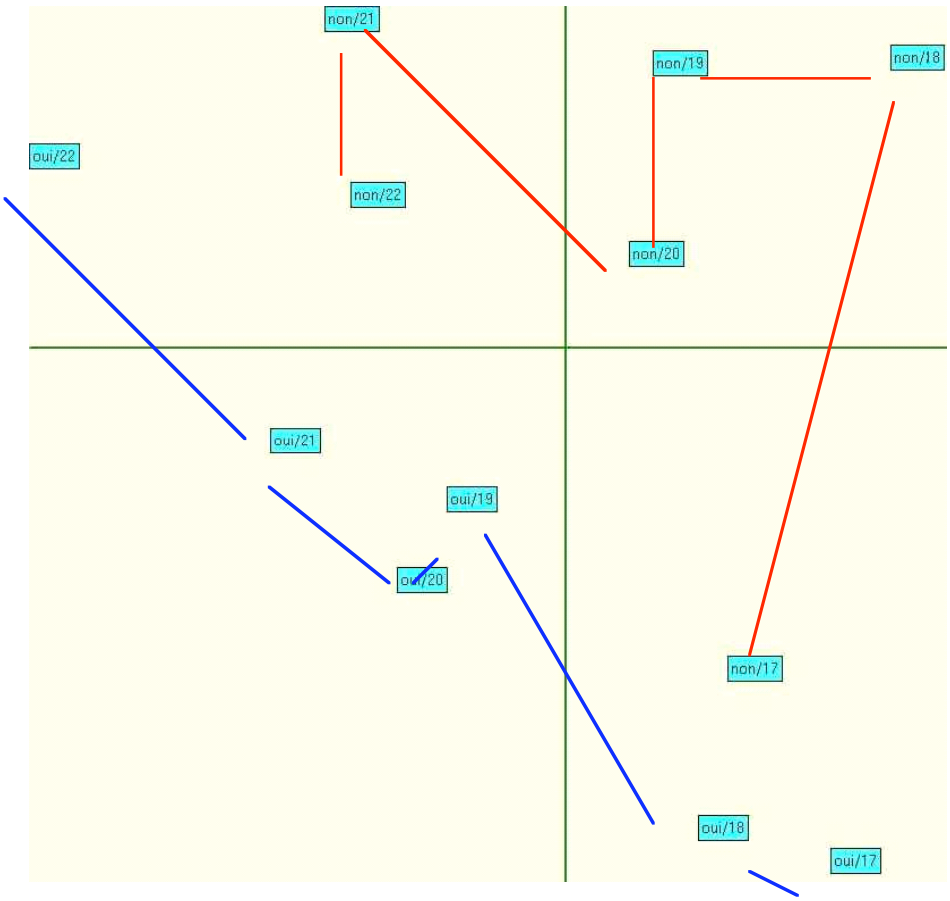


Figure 4 :
Analyse des correspondances réalisée à partir du tableau
12 semaines/opinions x formes de fréquence ≥10

La disposition des parties que l'on observe sur la figure 4 suggère l'hypothèse d'une proximité initiale du discours des deux camps (semaine 17) qui cesse dès la semaine 18 du fait d'un renouvellement important dans le discours du *non*. Les deux séries chronologiques évoluent ensuite vers la gauche du graphique sans jamais se mélanger à nouveau.

Sur la figure 5, on montre la ventilation des deux *types spécifiques* (***Sp-non***, formes spécifiques pour le camp du non et ***SP-oui***, formes spécifiques pour le camp du oui), introduits plus haut, dans la dernière partition que nous venons de constituer. Les parties

correspondant aux discours des participants qui se prononcent pour le *non* sont représentées sur la gauche de la figure, celles qui correspondent aux opinions positives sont regroupées sur la droite.

Le constat fait plus haut à propos de la ventilation de ces *tgens* (paroxysme pour le *non* en semaine 18, forte poussée en revanche pour le vocabulaire du *oui* en semaine n°21) se trouve à la fois confirmé et précisé. Après la semaine 18, le tgen spécifique du *non* connaît comme on le voit un affaiblissement progressif dans les parties qui correspondent au non. Les termes qui participent au tgen ***Sp-non*** sont également moins utilisés au fil du temps dans les parties qui correspondent au discours du *oui*. On tentera donc de vérifier l'hypothèse suggérée par l'analyse de cette figure : les participants au forum qui se prononcent pour le oui évitent de plus en plus le vocabulaire spécifique du camp adverse à mesure que le débat se développe.

Pour ce qui concerne le tgen spécifique des partisans du *oui*, la situation est différente. Le camp du *oui* utilise de plus en plus son vocabulaire spécifique, lequel est plutôt sous-utilisé par les partisans du *non* tout au long de la période considérée sans que l'on puisse déceler une évolution chronologique dans dette évolution. Par ailleurs, on peut remarquer que le camp du *non* utilise de moins en moins le vocabulaire spécifique du camp adverse.

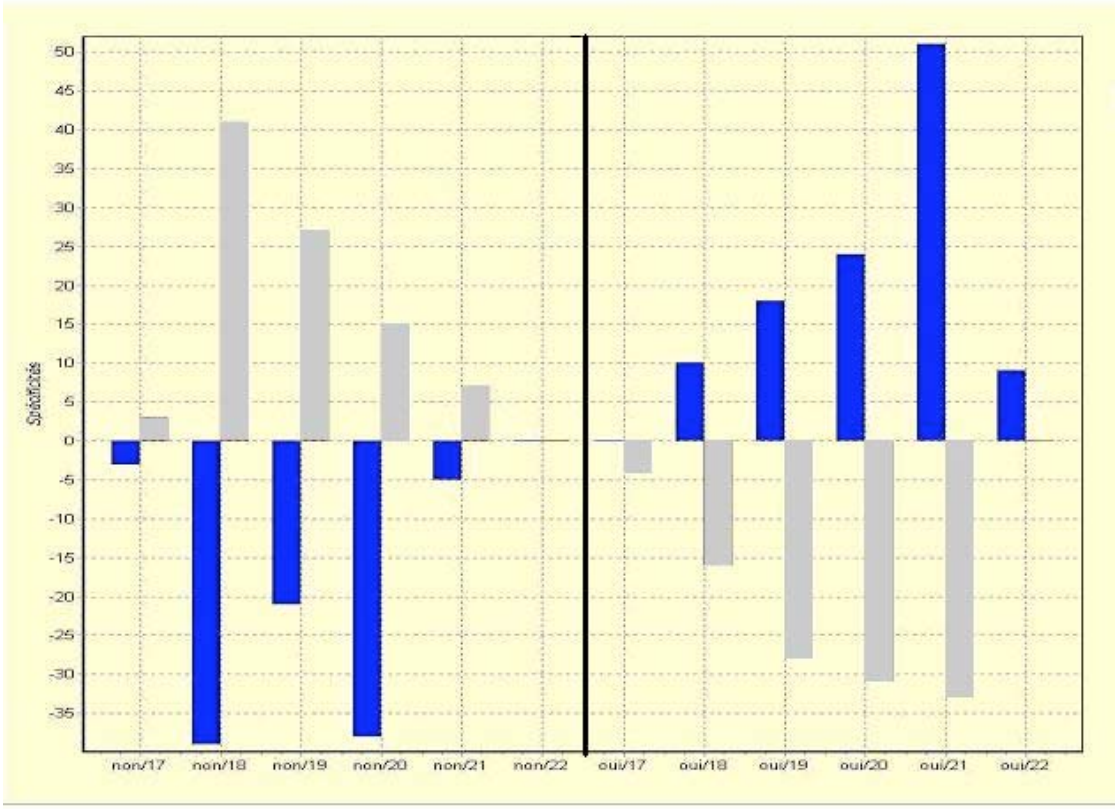


Figure 5 :
Ventilation des tgens spécifiques ***SP-oui*** *et* ***SP-non*** *dans les 6 semaines du corpus*

Perspectives

Comme nous l'avons vu, le dépouillement chronologique au fil des semaines fait apparaître très clairement l'évolution du vocabulaire du débat que nous avons recueilli sur le forum. Il nous permet de distinguer des moments forts de cette évolution.

L'introduction d'un repérage des opinions en *pour* ou *contre*, effectué par un humain, nous permet d'affiner la description de cette évolution pour chacun des groupes de participants. On peut envisager à l'avenir d'assister cette tâche en chargeant des procédures automatisées de trancher dans un nombre de cas important.

Ces premiers résultats nous encouragent à explorer, à d'aide des mêmes méthodes, plusieurs autres directions de recherche que nous avons seulement esquissées ici : l'évolution de la dynamique de l'argumentation à l'intérieur de chacun des groupes d'opinion, celle de l'autodésignation et la désignation de l'autre par chacun des groupes de locuteurs ainsi que celle des classes d'arguments que l'on peut constituer à partir d'outils de classification divers.

Références

Fraenkel B. et Marcoccia M. (éds) (2003). Écrits électroniques : échanges, usages et valeurs. *Langage et Société 104*.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des 6e Journées d'analyse des données textuelles,* IRISA, Rennes.

Marchand, P. (2004). *Psychologie sociale des médias*. Presses Universitaires de Rennes.

Marchand P. (2005). Le grand oral de Dominique de Villepin. *Bulletin de Méthodologie Sociologique,* 87 : 80-85.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris, Dunod.

Salem A. (2004). Introduction à la résonance textuelle.In *Actes des 7e Journées d'analyse des données textuelles,* Presses universitaires de Louvain.