Rapport de Projet Master 1 PluriTAL

Navigations dans Le Monde

encadré par Serge Fleury, maître de Conférences en Linguistique Informatique,

P3 ILPGA

Marianna Kucharski, Ivan Šmilauer, Marguerite Leenhardt

June 12, 2007

Contents

Remerciements

Nous tenons en premier lieu à remercier monsieur Serge Fleury, tout d'abord pour la mise à disposition des données textuelles et de nombreux outils d'analyse issus du développement de son projet *Navigations dans Le Monde*, ainsi que pour son encadrement efficace.

Nous remercions également monsieur André Salem pour nous avoir guidés dans l'utilisation des fonctionnalités du logiciel Lexico3.

Nous adressons également nos remerciements à Christelle Ayache et Olivier Hamon pour leur relecture attentive et leurs conseils sur la structuration de la présente version de notre Rapport de Projet.

Enfin, nous tenons tout particulièrement à remercier nos collègues Nicolas Aubry et Mani Ezzat pour le vif intérêt qu'ils ont manifesté quant à la conduite de notre projet.

Résumé

Ce rapport expose certaines visées applicatives des divers enseignements suivis dans le cadre du Master 1 PluriTAL. Le présent projet, développé par l'équipe MIM, articule différents axes d'analyse des corpus textuels, automatiques, textométriques et linguistiques. Nous rapportons des résultats qui en sont issus, dans l'idée de proposer tantôt une amorce, tantôt un approffondissement de recherches réalisées sur les données que nous traitons.

Après avoir présenté les traitements appliqués aux corpus, le rapport d'activités comprend une exploration lexicométrique, une étude linguistique du point de vue de l'expression de la négation, ainsi que la mise au jour d'éléments pour une analyse interactionnelle, en fonction des données qui s'y prêtent.

Chapter 1

Contexte du projet

1.1 Présentation du projet et des corpus

Le projet Navigations, proposé et encadré par Serge Fleury, se place dans le contexte d'une recherche fondée sur la capitalisation de données collectées par récupération de fils RSS du journal *Le Monde*. L'objectif global est de réaliser une exploration textométrique de ces données textuelles, conjuguée à une analyse linguistique des résultats.

La réalisation du projet Navigations regroupe 8 étudiants, de Paris 3, Paris X et de l'INALCO :

Bounouar BACHA (P3)

Boualem BENMESSAOUD (P3)

Marianna KUCHARSKI (PX)

Marguerite LEENHARDT (P3)

Luiza MAXIM (P3)

Mandana MOGHIMI-CATHEBRAS (INALCO)

Ivan SMILAUER (INALCO)

Marianne YOUNES-MICHIELS (INALCO)

Les compétences de chacun sont mises à profit dans le choix et le traitement des tâches à réaliser. En effet, tant du point de vue des aspects techniques que du point de vue des analyses mises en oeuvre, les profils de cette équipe de projet sont complémentaires.

Pour mener à bien les tâches à réaliser, l'équipe est scindée en modules de travail.

Le présent rapport rend compte des explorations et analyses de l'équipe MIM.

1.2 L'équipe MIM

Représentative de la symbiose des formations conjuguées dans le cadre du Master PluriTAL, l'équipe MIM s'est constituée dans l'idée de mettre en valeur la variété des profils qui la forment.

Marianna Kucharski [1]

Etudiante française, titulaire d'une Licence Sciences Humaines et Sociales mention Sciences du Langage spécialité Traitement Automatique des Langues, obtenue en 2006 à Paris X Nanterre, je m'intéresse particulièrement au traitement robuste de la langue et à ses formalismes logiques, syntaxiques et sémantiques. Initialement destinée à un cursus de Mathématiques, j'ai par ailleurs choisi de poursuivre des études de Linguistique; compétences plurielles que j'ai eu l'opportunité de mettre à profit en découvrant le domaine du TAL.

J'aspire à intégrer professionnellement le domaine du TAL, par le biais d'un Master 2 Professionnel.

Ivan Smilauer [2]

Etudiant tchèque, titulaire d'un Master de Philologie tchèque et française obtenu en 2004 à l'Université Masaryk, Brno (République tchèque), d'un DEA en Linguistique tchèque obtenu à l'INALCO en 2005 et d'une licence de Traitement Automatique de Langues obtenue à l'INALCO en 2006, je suis, parrallèlement à mon initiation au TAL dans le cadre du PLuriTAL, en doctorat en linguistique tchèque à l'INALCO et à l'Université Charles, Prague.

Le sujet de ma recherche de thèse qui est la modélisation de la morphologie flexionnelle du tchèque et son acquisition par les apprenants francophones, m'a amené à m'intéresser aux techniques et outils issus du TAL.

Marguerite Leenhardt [3]

Etudiante française titulaire d'une Licence de Lettres mention Sciences du Langage spécialité Traitement Automatique des Langues Naturelles, obtenue en 2006 à l'Université Paris 3, j'ai tenu à me spécialiser, parallèlement à mon cursus TAL, en Phonétique Expérimentale et Appliquée ainsi qu'en Théories Sociolinguistiques.

Initialement destinée à un cursus de CPGE Littéraire, j'ai fait le choix d'un profil universitaire construit autour de la pluridisciplinarité inhérente à la formation dispensée en Licence par l'UFR ILPGA. Je souhaite intégrer le domaine entreprenarial du TAL à l'issue d'un Master 2 Professionnel.

[1] étudiante en Master 1 Sciences du Langage Mention TAL à l'Université Paris X Nanterre, http://www.u-paris10.fr UFR LLPhi - contact: marianna_kucharski AT hotmail.fr

[2] étudiant en Master 1 Professionnel à l'INALCO, Institut National des Langues et Civilisations Orientales, http://www.inalco.fr - contact: ivansmil AT centrum.cz

[3] étudiante en Master 1 Recherche PLuriTAL à l'Université Paris 3 Sorbonne Nouvelle, http://www.cavi.univ-paris3.fr UFR ILPGA - contact: marguerite.leenhardt AT laposte.net

1.3 Contexte de la recherche

Le travail proposé par Serge Fleury s'appuie sur deux projets en cours [Fleury 2005][4] , dans lesquels s'inscrivent notre recherche.

1.3.1 Phase 1 : première récolte

Le Monde en Surface

Ce projet, débuté en octobre 2005, comporte deux modules :

1. Le premier module, *Fil(s) de presse*, permet, d'une part, de traiter un fil de presse donné, au format RSS; d'autre part, de construire des traitements sur le contenu de ce fil.

2. Le second module, *Archivage des Fils de Presse*, permet un archivage automatique et continu des fils.

A l'issue de ces modules, un corpus de fils RSS résultant d'un archivage heure par heure est à la disposition des étudiants impliqués dans le projet *Navigations*. Une série d'outils appropriés au traitement de ces fils est actuellement en cours de développement[5] .

Le Monde Profond

L'ensemble des versions quotidiennes du journal *Le Monde* ont été régulièrement récupérées sur le site web du journal, aux formats HTML et PDF[6]. C'est la version HTML qui a été utilisée afin d'en produire différents états:

1. un état quotidien des contenus textuels, normalisée au format XML, ainsi qu'une version compatible avec Lexico3;

2. des états statistiques quotidiens.

Les états quotidiens des contenus textuels ont fait l'objet d'une concaténation, afin de produire des corpus chronologiques, en fonction de l'ensemble des dates de récupération.

Le processus d'archivage couvre la période du 12 avril 2003 au 19 septembre 2006[7] .

Une version complémentaire des données disponibles en ligne sur le site du journal *Le Monde* est également mise à profit dans le cadre de ce projet, toujours à partir des fils RSS récupérables. Il s'agit du corpus *Le Monde semi-Profond*, constitué de l'ensemble des contenus textuels des fils RSS - correspondant donc au *Monde en Surface* -, additionné des versions intégrales de tous les articles qui leur sont associés.

[4] cf. synthèse du projet *Navigations* rédigée par Serge Fleury, pour davantage de précisions - URL : http://sfmac.no-ip.com/habilitation/wip/master-2006-2007/navigation-LeMonde-master-2006-2007.htm

[5] URL du projet : http://tal.univ-paris3.fr/filspresse/

[6] http://lemonde.fr/

[7] URL du projet : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm

1.3.2 Phase 2 : seconde récolte

La seconde phase de ce projet est une extension du précédent, faisant suite à l'interruption du *Monde Profond* le 19 septembre 2006[8] .

Le Monde en Surface

Cet aspect du projet recouvre les mêmes données que *Le Monde Surface* présenté dans la Phase 1.

Le Monde Profond

La seconde phase de ce projet s'articule autour de l'archivage en parallèle des fils RSS et des articles complets associés aux items décrits dans les fils.

Le processus mis en place pour la constitution de la version enrichie du corpus, i.e. le corpus *Le Monde semi-Profond*, a été optimisé afin de permettre l'archivage complet des articles associés aux fils, quelle que soit leur longueur. Cette phase d'optimisation a débuté le 20 novembre 2006.

1.4 Les corpus

1.4.1 Les corpus initiaux

Le corpus Présidentielles

Le corpus *Présidentielles-2007*, constitué par Serge Fleury, ne prend en compte que les pages et fils concernant les élections présidentielles 2007[9] . Il est construit par concaténation des fichiers RSS archivés heure par heure et des articles qui leur sont associés.

Le corpus Discours

Le corpus *Discours*, disponible sur le blog de Jean Véronis[10], est le résultat d'un travail réalisé de concert avec la contribution de Damon Mayaffre (CNRS, Nice), de Carine Duteil (CNRS, ATILF) et de Pascal Marchand (IUT, Toulouse). Il regroupe 230 discours, répertoriés en fonction des candidats officiellement en course à l'élection présidentielle. La période de récolte des données s'étend d'octobre 2006 au lendemain du premier tour.

Il faut préciser que pour les candidats sélectionnés au second tour, nous avons pris l'initiative de récupérer des données supplémentaires sur le blog de Jean Véronis; la collecte finale s'étend donc jusqu'au 6 mai : les derniers discours que nous avons récolté sont ceux prononcés à l'issue des résultats.

Jean Véronis met en exergue l'inégalité de médiatisation de telles données. En effet, si les données relatives aux candidats représentant des familles politiques bénéficiant d'un réseau de communication bien assis n'ont pas posé de problème de récupération, les discours des représentants d'organisations politiques dont l'équipe de communication était plus modeste ont été difficiles à récupérer, leurs équipes de campagne ne cherchant pas forcément toujours à

[8] URL du projet : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm
[9] http://www.lemonde.fr/rss/fil/0,57-0,64-823353,0.xml
[10] http://ww.up.univ-mrs.fr/veronis/Discours2007

les diffuser. C'est pourquoi une inégalité flagrante caractérise la proportion des discours collectés par candidat, ce que chacun peut constater en allant consulter le blog de Jean Véronis.

Le corpus Débat

Le corpus *Débat*[11] , récupéré par Pascal Marchand sur la version en ligne du journal *Libération*[12], est une transcription orthographique du débat opposant Ségolène Royal à Nicolas Sarkozy le 2 mai 2007, animé par Arlette Chabot et Patrick Poivre d'Arvor. La version qui nous en a été transmise par Serge Fleury et Pascal Marchand est formatée pour le traitement avec Lexico3, i.e. segmentée par des balises *auteur* spécifiées en attribut par le nom du locuteur, *chabot, poivre, royal, sarkozy.* Chaque tour de parole y est identifié par le marqueur §.

1.4.2 Traitements préliminaires des corpus

Encodage

Étant donné que nous voulons travailler avec des textes en langue française et également par le souci de compatibilité entre les différentes plate-formes utilisées (PC Windows XP, Cygwin, Mac OS X), nous avons décidé d'encoder nos corpus en ISO-Latin1 (ISO-8859-1). Les fichiers texte dont on disposait et qui étaient à la base constitués par la recopie du contenu provenant de l'internet (des pages générées en php pour les discours, un document pdf pour le débat), nécessitaient une modification pour être conformes avec cette norme.

Nous pouvions constater que nos corpus contenaient des caractères incompatibles avec la norme ISO grâce à la fonctionnalité de l'éditeur TextWrangler sous Mac OS X qui ne permet pas d'enregistrer un fichier texte dans un encodage donné si celui-ci contient des caractères qui n'appartiennent pas aux jeux de caractères de ce codage.

En pratique, il était nécessaire de remplacer les caractères propres à la norme Windows-1252 (Europe occidentale) par leurs équivalents en ISO Latin1. Il s'agissait des caractères suivants:

Caractère Win1252	code Win1252	Caractère(s) ISO Latin1	code ISO Latin1
… (Ellipsis)	x85	... (point 3 fois)	x2E x2E x2E
œ (e dans l'o)	x9C	oe (o + e)	x6F x65
’ (Curved quotes)	x92	' (Apostrophe)	x27
€ (Euro)	x80	Euro (E + u + r + o)	x45 x75 x72 x6F
– (Dash ponctuation)	x96	- (Hyphen)	x2D

Figure 1.1: Fig.1 - Remplacement des caractères

[11] transcription réalisée par Littera Stenotypie - Analyse de P. Marchand
[12] http://www.liberation.fr

Une image illustrant la façon de parcourir le fichier texte avec le programme 0xED afin d'identifier les caractères incompatibles avec la norme ISO Latin1 :

Caractère Win1252	code Win1252	Caractère(s) ISO Latin1	code ISO Latin1
… (Ellipsis)	x85	... (point 3 fois)	x2E x2E x2E
œ (e dans l'o)	x9C	oe (o + e)	x6F x65
´ (Curved quotes)	x92	' (Apostrophe)	x27
€ (Euro)	x80	Euro (E + u + r + o)	x45 x75 x72 x6F
– (Dash ponctuation)	x96	- (Hyphen)	x2D

Fig.2 - Caractères incompatibles avec la norme ISO Latin1

Sans avoir effectué ces modifications, nous risquions avoir des problèmes avec les logiciels utilisés. L'étiqueteur TreeTagger, par exemple, n'interprétait pas le signe ´(Curved quotes) comme un séparateur et il considérait les deux formes concaténées comme en seul token (par exemple *cest, lhomme, na*, etc). Ceci aurait été très gênant pour l'exploitation.

Balisage Lexico3

Le corpus **Débat**, tel que nous l'avons obtenu, a été déjà prétraité pour pouvoir être exploité par Lexico3 : les tours de parole ont été marqués par des balises telles que *auteur="locuteur"*, le début de chaque ligne a été marqué par le signe §. Malheureusement, ce balisage n'était pas adapté pour rendre compte de la vraie structuration du débat.

En effet, le balisage des tours de parole n'était pas complet car le document original utilisait deux façons différentes de marquer les locuteurs et seulement une de ces méthodes a été prise en compte lors du balisage initial du corpus. Il fallait donc remplacer les suites § *locuteur :* par *auteur="locuteur"* afin d'unifier le marquage de cette information dans notre corpus. Ceci peut être fait assez rapidement avec la fonction chercher/remplacer d'un éditeur de texte. Nous avons donc remplacé les éléments suivants :

- § Ségolène Royal : 50 occurrences
- § Nicolas Sarkozy : 46 occurrences
- § Arlette Chabot : 3 occurrences
- § Patrick Poivre d'Arvor : 7 occurrences

Le signe § correspondait à la segmentation du débat en paragraphes. Cette segmentation a été introduite lors de la retranscription du débat en forme écrite. En général, un paragraphe correspondait à un tour de parole, mais dans certains cas, plusieurs paragraphes étaient présents au sein d'un tour de parole d'un locuteur. Garder cette information n'était pas trop pertinent car elle correspondait plutôt à la vision de celui qui a transcrit le débat qu'à la structuration réelle

de la communication. Nous avons donc décidé de garder ce signe pour marquer uniquement les tours de parole pour Lexico3 et de segmenter notre texte en phrases (voir la section suivante) qui seront marquées par le signe $.

Le corpus **Discours** était fourni avec un balisage Lexico3 par auteur et par date (les balises *discours2007="auteur"* et *date="aaaammjj"*) que nous avons gardé lors de l'ajout des discours plus récents (jusqu'au 6 mai). En plus, nous avons ajouté (avec une macro emacs) la balise *candidat-date="xx-aamm"* (contenant les initiales des candidats et l'année suivie du mois) qui était utile pour les exploitations longitudinales dans Lexico3.

Pour assurer la cohérence des données linguistiques, nous avons enlevé les titres de chaque discours (cependant nous les avons gardés dans la version XML, voir plus bas).

Un exemple d'un titre : *François Bayrou, Discours au colloque UDF "Développement durable"* (21/10/06).

Pour enlever ceci du corpus, nous avons utilisé une commande Unix, basée sur une spécificité du format des titres (la date parenthésée) qui ne figurait nulle part ailleurs :

```
$ grep -v '\(../../0.\)$' corpus_titre > corpus_sans_titre
```

Fig.3 - Changement du format des titres (commande Unix)

La présence du signe § avait ici le même rôle que dans le corpus Débat, nous avons alors décidé d'effectuer le même traitement (voir la section suivante).

Segmentation en phrases

Quelques uns des objectifs de notre recherche (calcul sur les longueurs des phrases, extraction des phrases contenant un certain motif, etc.) nécessitaient que l'on puisse travailler avec les phrases en tant qu'unités, extraites de nos corpus. Si nous considérons une phrase comme une suite de caractères terminée par un point, nous pouvons baser la segmentation sur ce critère. Par contre, nous ne devons pas oublier qu'un point dans le texte ne signifie pas forcément la fin de phrase.

Contrairement à ce que l'on pourrait penser, la segmentation en phrases est une tâche assez complexe. Nous avons décidé d'écrire un script perl qui traite les cas d'ambiguïté les plus fréquents : le point après une majuscule comme M. pour Monsieur et les trois points à la fin de la phrase. Ceci n'est pas bien sûr exhaustif (nous omettons par exemple le cas ou une phrase serait terminée par une majuscule), les résultats paraissent cependant assez satisfaisants.

Le script *segmentPhrases.pl*[13] prend en argument un fichier texte, le segmente et crée un fichier avec une phrase par ligne. Le script lit le fichier texte caractère par caractère et à chaque occurrence du caractère *pointf*, il décide en fonction des caractères suivants ou précédents s'il faut imprimer le saut de ligne.

Vous pouvez voir ici la boucle de lecture des caractères avec les critères qui servent à la décision (les expressions logiques dans les structures conditionnelles) :

```
while($car[$i] ne "\0")                #lecture du tableau de caractères du fichier
{
    print Sortie $car[$i];             #imprime le caractère courant

    if( ($car[$i] eq "\.") &&              #le caractère courrant est un point et
        ($car[$i+1] ne "\.") &&        #le caractere suivant n'est pas un point et
        ($car[$i-1]=~ /[^A-Z]/) &&
        ($car[$i-1]=~ /[^0-9]/) &&     #le caractère précedent n'est pas majuscule et nombre
        ($car[$i+2]=~/[A-Z]/))         #le 2eme caractère suivant est une majuscule (debut de phrase)
        {
            print Sortie "\n";         #si tout ça OK, imprime fin de ligne et saute l'espace
            $i++;
        }

    if( ($car[$i] eq "\.") &&          #pour les trois points
        ($car[$i-1] eq "\.") &&
        ($car[$i-2] eq "\.")  &&
        ($car[$i+2]=~/[A-Z]/))
        {
            print Sortie "\n";             #si tout ça OK, imprime fin de ligne et saute l'espace
            $i++;
        }

    if(((($car[$i] eq "!") ||          #pour gérer ! et ? (< > a cause des balises XML)
        ($car[$i] eq "\?")) &&
        ($car[$i+2]=~/[A-Z]/) &&
        ($car[$i-1] ne "<")  &&
        ($car[$i+1] ne ">"))
        {
            print Sortie "\n";             #si tout ça OK, imprime fin de ligne et saute l'espace
            $i++;
        }

    $i++;                              #la lecture du car suivant
}
```

Fig.4 - Boucle de lecture des caractères

Nous obtenons donc à la sortie un fichier avec le texte segmenté en phrases (une phrase par ligne). Après avoir enlevé les lignes vides (par exemple avec ce petit script)[14], le fichier est prêt pour un balisage en XML - les balises Lexico3 originales sont restées intactes après la segmentation et serviront comme base pour ce balisage.

Balisage XML Débat

La balisage XML nous permet d'extraire de l'information de nos documents avec des transformations XSL. Ceci est particulièrement pratique notamment pour l'extraction des suites de mots spécifiques par des requêtes Xpath (extraction

[13] cf.version HTML du Rapport
[14] cf.version HTML du Rapport

des patrons) qui est une alternative à des scripts d'extraction en Perl.

L'attribution d'une structure XML au corpus Debat pouvait être faite presque "manuellement". Il s'agissait de convertir (par la fonction rechercher/remplacer d'un éditeur de texte) les balises Lexico en format bien formé XML (c'est à dire par exemple *auteur="locuteur"* par *aut loc="locuteur"*). Nous avons procédé de la même façon pour la génération des balises fermantes correspondantes (c'est à dire remplacer *aut loc="locuteur"* par *aut aut loc="locuteur"*, etc.)

Le document XML bien formé que nous obtenons avec ces opérations en quelques minutes peut être finalement balisé au niveau des phrases - c'est à dire que chaque phrase du fichier segmenté (une phrase par ligne avec les balises séparées de l'autre texte par saut de ligne) peut être enfermée entre les balises *s /s*. Ceci peut être fait par un script[15] qui prend en aurgument le nom du fichier entrée, qui ajoute ces balises à chaque ligne à condition qu'elle ne contienne pas de chevrons (ce qui voudrait dire qu'il s'agissait d'une balise XML), en sortie nous obtenons le fichier balisé.

La version XML du corpus Debat ainsi obtenu est structurée de la façon suivante :

```
<!ELEMENT aut ( s+ ) >
<!ATTLIST aut loc ( chabot | poivre | royal | sarkozy ) #REQUIRED >
<!ELEMENT debat ( aut+ ) >
<!ELEMENT s ( #PCDATA ) >
```

Fig.1 - Boucle de lecture des caractères[16]

Pour les besoins de l'extraction de l'information basée sur des critères grammaticaux, il était nécessaire d'annoter le fichier par un étiqueteur morphologique. Nous avons utilisé TreeTagger qui est simple à utiliser et les résultats sont satisfaisants. Nous pouvons annoter directement le fichier XML, car les lignes avec les balises ne sont pas étiquetées. Nous pouvons appeller TreeTagger par exemple de cette façon (à condition que le path de la commande soit valide) :

```
$ tree-tagger-french fichier-à-étiqueter fichier-sortie
```

Fig.2 - Etiquetage sous TreeTagger

Le fichier que nous obtenons doit être traité par un script de balisage[17] qui structure l'information grammaticale issue du TreeTagger et l'intégre au document pour qu'on obtienne un document XML bien formé. La DTD de ce document est la suivante :

[15] cf.version HTML du rapport pour le téléchargement du script ajouteSent.pl
[16] cf.version HTML du Rapport pour le corpus intégral et la feuille de style associée
[17] cf. version HTML, script TT2xml.pl

```
<!ELEMENT aut ( s+ ) >
<!ATTLIST aut loc ( chabot | poivre | royal | sarkozy ) #REQUIRED >
<!ELEMENT debat ( aut+ ) >
<!ELEMENT s ( e+ ) >
<!ELEMENT e (st, lm, tp) >
<!ELEMENT st (#PCDATA) >
<!ELEMENT lm (#PCDATA) >
<!ELEMENT tp (#PCDATA) >
```

Fig.3 - DTD appliquée à la version XML du corpus Débat

L'élément e contient les éléments st (string), lm (lemme) et tp (type - l'étiquette). Les noms des éléments ont été choisis courts pour diminuer au maximum le volume du document XML. Cela aide beaucoup (au niveau du temps) pendant le traitement du document.

Balisage XML Discours

Pour attribuer une structure XML au corpus **Discours**, nous avons procédé de la même façon que dans le cas du corpus Débat à deux différences près - le titre de chaque discours a été inclus dans la structure XML par un script[18] qui parcourt le fichier ligne par ligne et ajoute les balises *titre /titre* à chaque ligne contenant l'expression régulière (.*0.) : il s'agit de l'indication de la date du discours - par exemple (21/10/06) - qui est unique dans tous le corpus.

Nous obtenons ainsi un fichier XML contenant tous les discours de nos quatre candidats. La DTD de ce document est la suivante:

```
<!ELEMENT cand ( disc+ ) >
<!ATTLIST cand loc (royal | sarkozy | lepen | bayrou) #REQUIRED >
<!ELEMENT disc ( titre, text ) >
<!ATTLIST disc date NMTOKEN #REQUIRED >
<!ELEMENT discours ( cand+ ) >
<!ELEMENT p ( s+ ) >
<!ELEMENT s ( #PCDATA ) >
<!ELEMENT text ( p+ ) >
<!ELEMENT titre ( #PCDATA ) >
```

Fig.1 - DTD appliquée à la version XML du corpus Discours[19]

L'élément p correspond à la division des discours par paragraphes dans leur version publiée sur le site de J. Veronis. La balise *p /p* a été ajoutée au corpus par un script[20] avant la segmentation du fichier en phrases (donc à l'état où il contenait un paragraphe par ligne. Cette manipulation était optionnelle et elle n'aurait servi qu'au moment où nous aurions voulu utiliser la segmentation en paragraphes, ce qui n'était pas notre cas.

Pour l'annotation avec TreeTagger, nous avons décidé d'annoter les discours de chaque candidat à part, sinon la taille du fichier annoté et structuré en

[18] cf. version HTML du rapport pour télécharger le script ajouteTitre.pl

[19] cf. version HTML du Rapport pour télécharger l'archive du corpus Débat avec les 4 candidats choisis

[20] cf. version HTML du rapport pour télécharger le script ajoutePara.pl

XML aurait été trop importante pour une manipulation efficace. Nous avons extrait les discours de chaque candidat à l'aide de la feuille de style XSL ExtrCand2TXT.xsl[21] qui crée un fichier texte avec les discours (sans titres) du candidat spécifié dans la feuille de style.

Le fichier texte obtenu peut être ensuite annoté par TreeTagger et converti en format XML par un script Perl TT2xml2.pl[22] qui structure la sortie du TreeTagger de la même façon que dans le cas du corpus Débat (c'est à dire qu'il crée l'élément e qui contient les éléments st (string), lm (lemme) et tp (type - l'étiquette).

1.4.3 Caractéristiques des corpus

Calculs quantitatifs

Afin de caractériser nos corpus, nous nous sommes intéressés aux données quantitatives telles que, entre autres, le nombre moyen de mots ou de caractères alphanumériques dans une phrase et la longueur moyenne d'un mot en caractères. Nous supposons que ces informations peuvent donner un aperçu de certaines tendances dans les différentes sous-parties de nos corpus. Cependant, nous n'avons pas voulu essayer d'interpréter ces tendances, notre but étant uniquement de fournir des données qui pourraient aider à une telle analyse.

Les outils que nous avions à notre disposition (Lexico3, commande Unix wc -lwc) ne proposent pas de faire ces calculs nous avons alors décidé d'écrire deux scripts Perl : CalcCars.pl[23] et StatsPhMot.pl[24]. Les scripts sont destinés à lire uniquement du texte brut segmenté en phrases (une phrase par ligne).

Le script CalcCars.pl parcourt le fichier passé en argument et imprime un rapport avec les informations suivantes :

```
Nombre de caractères total
Nombre de caractères alphanumériques
Nombre de lignes
```

Fig.2 - Informations extraites avec le script CalcCars.pl

La boucle de lecture du fichier est la suivante :

[21] cf. version HTML du Rapport pour le téléchargement
[22] cf. version HTML du Rapport pour téléchargement
[23] cf. version HTML du Rapport pour téléchargement
[24] cf. version HTML du Rapport pour téléchargement

```
while(!eof(FIC))
{
    $car[$i] = getc(FIC);       #création du tableau pour les caracteres

    if($car[$i]=~ /[a-zA-Z0-0]/)    #si le caractere lu est alphanumerique
    {
        $j++;
    }

    if($car[$i] eq "\n")        #si le caractere lu est fin de ligne
    {
        $k++;
    }

    $i++;                   #tous les caracteres
}
```

Fig.3 - Boucle de lecture du fichier par le script CalcCars.pl

La commande Unix wc -lc peut aussi bien faire cette tâche mais elle ne distingue pas entre les caractères alphanumériques et les autres.
Le script StatsPhMot.pl[25] parcourt le fichier passé en argument et imprime un rapport avec les informations suivantes :

```
Nombre de phrases
Nombre de lignes vides
Longueur moyenne d'une phrase en caractères
Phrase la plus longue en caractères
Nombre moyen de mots dans une phrase
Phrase la plus longue en mots
Nombre de mots
Longueur moyenne d'un mot
Mot le plus long
```

Fig.3 - Informations extraites avec le script StatsPhMot.pl

La boucle de lecture du fichier est la suivante :

```
foreach $ligne()    #LECTURE DU FICHIER LIGNE PAR LIGNE
{
    @slova = ();                    #vider la liste de mots tampon
    $j = 0;                             #initialiser le compteur de mots dans la phrase

    if($ligne !~ /^$/)                      #si la ligne lue n'est pas vide
    {
        $ph[$i] = $ligne;           #tableau de phrases
        chomp $ligne;                       #enlever le car de la fin de ligne
        $phlong[$i] = length($ligne);       #stocker la longuer de phrases sans LF
        &Separate;                  #appel du subprog Separate qui remplace tous les separateurs

        @slova = split(/\s/,$ligne); #créer un tableau de mots à partir de la ligne

        foreach (@slova)            #inserer les mots de la phrase dans la listemots (linéaire)
        {
            $mots[$k] = $_;
            $motslong[$k] = length($_);
            $k++;                   #nb de mots dans le fichier
            $j++;                   #nb de mots dans la phrase
        }
        $phnbmot[$i] = $j;          #nb de mots de la phrase
        $i++;                               #compteur de phrases
    }
    else
    {
        $iV++;    #compteur des lignes vides
    }
}
```

Fig.4 - Boucle de lecture du fichier par le script StatsPhMot.pl

[25]cf. version HTML du Rapport pour téléchargement

Cette boucle structure les données dans la mémoire afin que l'on puisse extraire (par les calculs dans la seconde partie du script) les informations dont nous avons besoin. Nous considérons ici un mot comme une suite de caractères alphanumériques délimitée par des séparateurs définis dans le script)[26].

Graphiques

Cette partie présente les données obtenues par les calculs en forme de graphiques. Les calculs ont été effectués sur le partitionnement suivant :

- Corpus Présidentielles : pres
- Corpus Débat : sarkozy_db et royal_db
- Corpus Discours : bayrou : 42 discours - lepen : 27 discours - sarkozy : 57 discours - royal : 42 discours

Les calculs ont été effectués sur les versions du corpus contenant uniquement le contenu textuel segmenté par phrases.

La taille des corpus Le volume total de nos données - les trois corpus (Présidentielles, Débat et Discours) concaténés - contiennent au total 7 067 386 caractères alphanumériques et 1 621 718 mots.

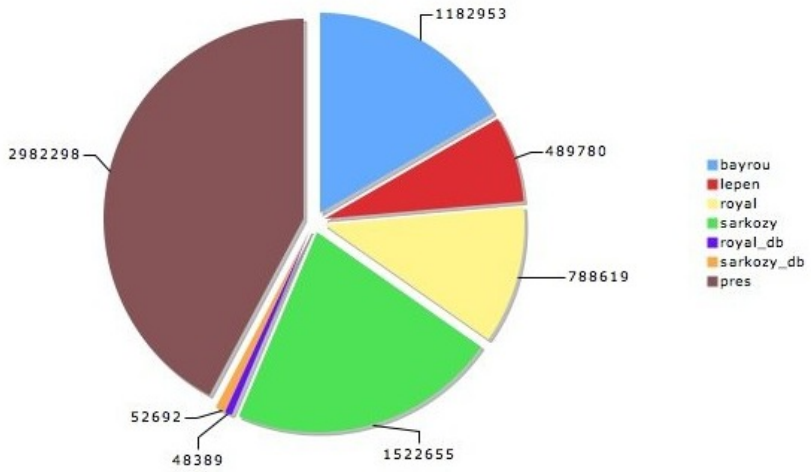

Fig.1 - Le volume des partitions du corpus en caractères alphanumériques

[26]Nous vous invitons à consulter la version HTML du Rapport pour le détail des résultats par corpus.

La longueur d'une phrase En considérant les données obtenues, il ne faut pas oublier que la segmentation des énoncés en phrases (c'est-à-dire, à notre sens, en séquences de caractères terminées par un point sous certaines restrictions) peut être assez arbitraire. Il s'agit notamment de la segmentation des éléments coordonnés qui peuvent être rassemblés dans une seule phrase ou au contraire figurer en tant que phrases autonomes d'après des critères assez vagues.

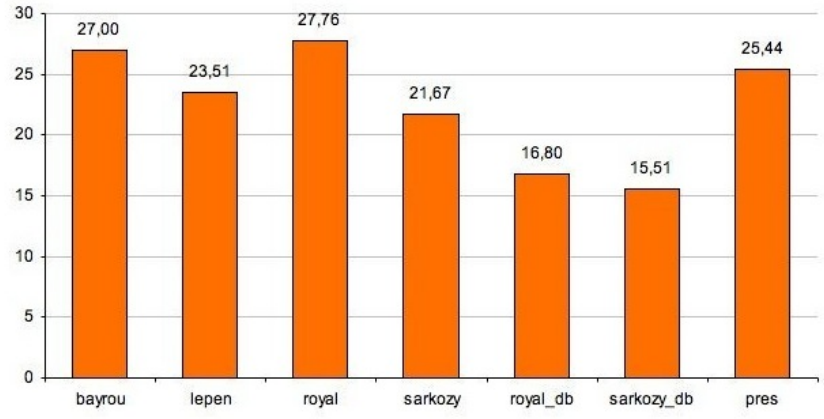

Fig.2 - Le nombre moyen de mots dans une phrase

La longueur d'un mot Les chiffres obtenus (entre 4,3 et 4,7 caractères) reflètent tout d'abord la haute fréquence des mots grammaticaux courts. Les trois mots les plus longs recensés dans nos corpus sont : *intergouvernementale, professionnalisation, professionnalisantes*. Malgré certaines tendances que nous pourrions dégager de nos données, il serait difficile de les mettre en rapport direct avec la complexité des énoncés.

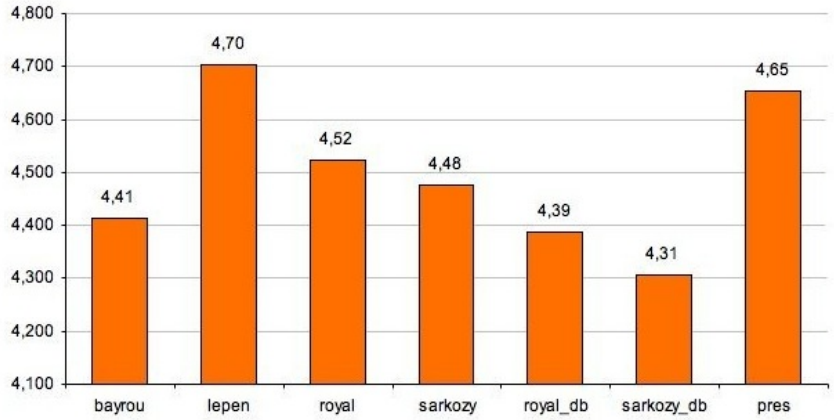

Fig.3 - La longueur moyenne d'un mot en caractères

Chapter 2

Compte-rendu des activités

Trois volets d'étude complémentaires constituent le présent rapport d'activités. Certains phénomènes saillants rendus observables par les divers traitements et analyses appliqués aux corpus de travail sont exposés en détails.
Les références sur lesquelles se fonde notre travail sont plurielles.

Du point de vue des données collectées, nous utilisons essentiellement les corpus issus de la recherche menée par Serge Fleury, présentés plus haut, que nous avons enrichi avec ceux mis à disposition par Pascal Marchand[1] et Jean Véronis[2].

Les outils et traitements mobilisés dans le cadre du projet Navigations sont divers. En effet, nous avons pris le parti d'ainsi mettre en regard les résultats obtenus, outre le fait que certaines opérations de formatage des sous-corpus de travail se sont avérées nécessaires.

Nous utilisons l'outil *Lexico3*[3], développé par le SYLED[4], pour réaliser des statistiques sur les données textuelles.

Nous utilisons également Pajek, logiciel permettant de générer des graphes rendant compte des collocations que peuvent présenter des unités lexicales. De plus, nous faisons appel aux logiciels DICO de Jean Véronis et TagCloudBuilder, permettant de construire des nuages de mots représentant l'importance des fréquences de leurs emplois.

Par ailleurs, nous mettons à profit les acquis des enseignements dispensés en M1 PluriTAL, en particulier les scripts Perl, l'utilisation de documents structurés au format XML ainsi que l'application de requêtes XPath.
Du point de vue des outils théoriques mobilisés, il est fait référence aux auteurs majeurs de la sociolinguistique interactionnelle[5] selon l'axe des enseignements de Luca Greco[6], ainsi qu'aux travaux de Rodolfo Delmonte[7] sur les phénomènes de disfluences[8].

[1] http://pascal-marchand.fr/
[2] http://up.univ-mrs.fr/veronis/
[3] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
[4] http://www.cavi.univ-paris3.fr/ilpga/syled
[5] cf.Annexes - pp.75
[6] Maître de Conférences en Sociolinguistique, P3
[7] http://project.cgm.unive.it/Persone/delmonte.html
[8] cf.Delmonte, Rodolfo (2005): "Modeling conversational styles in Italian by means of over-

2.1 Exploration lexicométrique

Cette partie présente les résultats d'une navigation lexicométrique du corpus Discours, réalisée essentiellement à l'aide de Lexico3. Nous avons également utilisé Pajek pour construire des graphes de mots, ainsi que TagCloudBuilder et Dico pour construire des nuages de mots.

Nous avons choisi d'observer l'emploi des mots "pacte", "programme", "projet" par les quatre principaux candidats à l'élection présidentielle de 2007, François Bayrou, Jean-Marie Le Pen, Ségolène Royal, et Nicolas Sarkozy.

2.1.1 Exploration Lexico3

Introduction

Le principal travail de cette partie consiste en une exploration lexicométrique des unités lexicales **Pacte**, **Programme** et **Projet** dans les différentes parties du corpus Discours, à savoir les parties *discours2007* - regroupant chacune les discours de chaque candidat - et les parties candidat-date - regroupant les discours par mois et par candidat.

Nous nous sommes intéressés à l'emploi de ces trois mots qui jouent, dans le contexte d'une campagne présidentielle, le rôle de synonymes. En effet, l'ensemble des propositions que présente un candidat est habituellement appelé *programme*, et parfois *projet*. Or, nous avons vu apparaître de façon importante au cours de cette campagne 2007, l'appellation *pacte*. Il a semblé intéressant de se pencher sur ces unités - par qui sont-elles employées?, remarque-t-on une évolution positive ou négative au fil du temps de ces emplois? -, ainsi que sur leurs collocations.

Rappelons tout d'abord, par une brève définition, les sens et connotations des termes qui nous intéressent. Définitions du TLFi:

Programme: "1.Ensemble des actions qu'on se propose d'accomplir dans un but déterminé", "POL. Ensemble de projets, d'objectifs politiques (avec l'échéance de leur réalisation)."

Projet: "1. Ce qu'on a l'intention de faire et estimation des moyens nécessaires à la réalisation.", "2. Travail préparatoire, première rédaction. Synon. canevas, esquisse, schéma."

Pacte: "1. DROIT.Convention expresse ou tacite, en principe immuable, entre deux ou plusieurs parties.", "2. Résolution par laquelle quelqu'un décide de rester fidèle à quelque chose."

laps", In DiSS-2005, 65-70.

Le mot *pacte* insiste donc sur l'idée d'accord, de confiance, de fidélité et de valeur "éternelle". Cette nouvelle appellation montre une évidente volonté de changer l'image d'un programme présidentiel, et d'y impliquer l'électeur, afin que celui-ci se sente en confiance avec ses représentants politiques et que son sentiment de distance avec le monde politique soit moindre.

Spécificités

Avant d'étudier le comportement particulier des trois unités lexicales choisies, nous avons fait un petit tour des principales caractéristiques lexicométriques du corpus (PCLC), et notamment des spécificités.

Les PCLC rendent compte, par parties selon la partition choisie, des occurrences, des formes, des hapax et de la fréquence maximale.

Occurrences: nombre d'occurrences des formes répertoriées.

Formes: nombre de formes graphiques différentes.

Hapax: nombre de formes qui n'apparaissent qu'une fois.

Fréquence maximale: nombre d'occurences de la forme la plus fréquente.

Principales caractéristiques de la partition : discours2007

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
"bayrou"	278274	13112	5494	12108	de
"lepen"	109223	12920	6596	5957	de
"royal"	180907	11144	4819	7775	de
"sarkozy"	360372	15082	5761	17082	de

Fig.1 - PCLC par candidat, partition faite avec la balise discours2007[9]

Les valeurs des balises *candidat-date* contiennent les initiales des candidats et l'année suivie du mois. Exemple: "fb-O7O1" correspond à "François Bayrou, janvier 2007". Notons que pour Jean-Marie Le Pen nous avons choisi les initiales "lp".

Une **spécificité** est le rapport entre la fréquence d'une forme dans une partie donnée - i.e. sa sous-fréquence- et sa fréquence dans tout le corpus - i.e. sa fréquence totale. La sous-fréquence de la forme est comparée à la probabilité de voir se réaliser cette forme dans la partie. Cette probabilité est calculée à

[9]cf.version HTML du Rapport pour les PCLC par candidat et par mois, partition faite avec la balise candidat-date

partir d'un modèle hypergéométrique.

Pour résumer, une forme sera dite spécifique positive d'une partie si sa fréquence dans cette partie est "anormalement élevée". De même, elle sera dite spécifique négative d'une partie si sa sous-fréquence est "anormalement faible". Pour plus de précision, consulter le Manuel d'utilisation Lexico3[10].

Par défaut, l'indice de spécificité est calculé pour toutes les formes dont la fréquence est supérieure à 10, avec un seuil de probabilité fixé à 5%. Les tableaux des spécificités générés par Lexico3 présentent, pour chaque forme calculée, la fréquence totale, la sous-fréquence et le coefficient de spécificité (négatif ou positif)[11].

Le tableau des spécificités de Nicolas Sarkozy nous fait remarquer qu'une des formes les plus spécifiques positives de sa partie est le mot "pas"[12]. De plus, on remarque que les mots "pacte", "projet", "programme" ne sont pas des spécificités positives de sa partie, mais des formes spécifiques négatives (coefficients de spécificités respectifs: -2, -3, -5). Il faut tout de même être vigilant sur ce que l'on peut en dire. En effet, les spécificités étant calculées par rapport à un modèle de probabilités hypergéométrique qui tient compte de la position de la partie au sein du corpus, le fait que la partie *sarkozy* soit la dernière peut influencer les données.

Les données concernant la partie de Ségolène Royal mettent en évidence le fait que la candidate emploie de façon très importante le mot "pacte". En effet, il apparaît dans les formes les plus spécifiques positives de sa partie. Le coefficient de spécificité est de +50. De plus, on trouve le mot "programme" en spécificité positive (coefficient +3), ainsi que la forme "projets", au pluriel (coefficient +5).

Quant aux spécificités de François Bayrou, on observe que le mot "projet" est spécifique positif de la partie (coefficient +8). En revanche, les mots "pacte" et "programmes", au pluriel, sont spécifiques négatifs (coefficients respectifs -4 et -2).

Le tableau de la partie de Jean-Marie Le Pen est relativement intéressant puisqu'il nous montre que le candidat emploie de façon anormalement faible les mots "pacte" et "projet", spécificités négatives avec des coefficients respectifs de -11 et -10. Le mot "programme" est en revanche une spécificité positive de sa partie (coefficient +2).

[10] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/manuel3.htm
[11] cf.version HTML du Rapport pour la visualisation des spécificités par candidat
[12] cf.partie Négation de notre projet

Statistiques par parties

Nous nous intéressons ici à la ventilation de nos formes *pacte, programme, projet.*

Lexico3 peut générer des graphes de ventilation, selon la partition choisie, rendant compte des fréquences par partie des formes voulues. Nous avons observé, non pas les formes, mais les groupes de formes. En effet, le groupe *pacte+*, par exemple, contient non seulement la forme pacte, mais aussi les formes *Pacte, pactes, Pactes.*

Partition par candidat

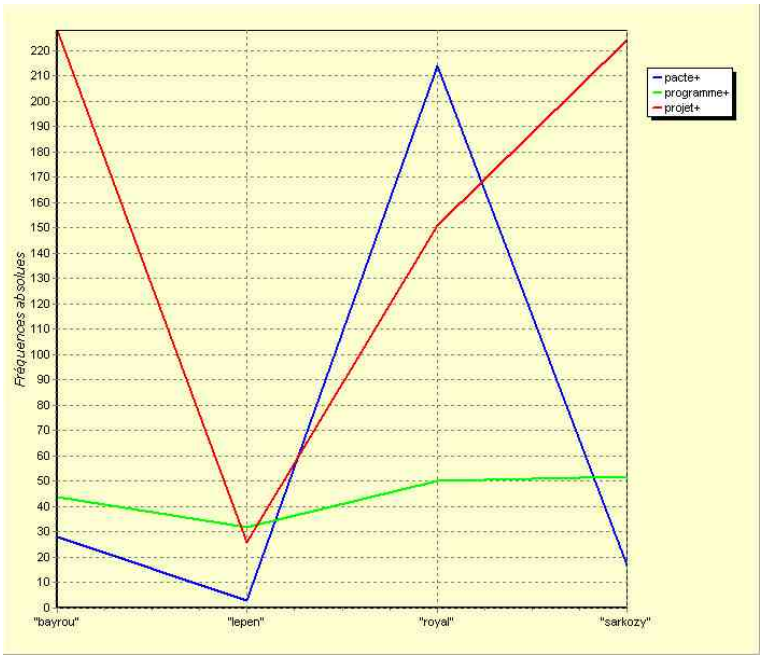

Fig.2 - Fréquences absolues des groupes de formes pacte+, programme+, et projet+, par partie

Ce graphique montre clairement que la candidate Royal est celle qui emploie le plus le mot "pacte". En effet, elle l'utilise abondamment, environ 70 fois plus souvent que les autres candidats.

De plus, on remarque que le mot "projet" est employé très fréquemment surtout par Sarkozy et Bayrou, souvent par Royal et très peu par Jean-Marie Le Pen.

Enfin, le mot "programme" est employé quasiment à la même fréquence par Sarkozy, Bayrou et Royal, un peu moins par Jean-Marie Le Pen. Notons que ce dernier est le candidat qui utilise ces appellations le moins fréquemment.

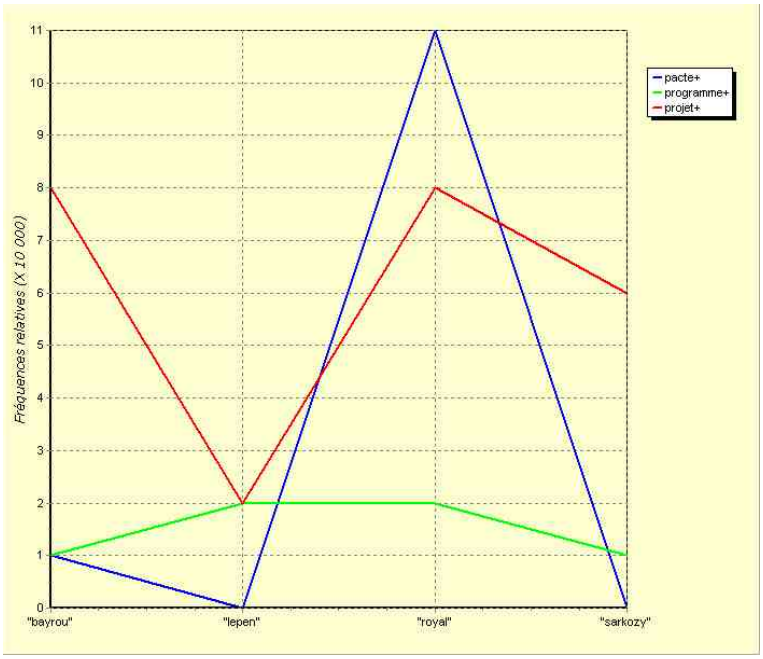

Fig.3 - Fréquences relatives des (groupes de) formes

La fréquence relative d'une forme est sa fréquence dans la partie par rapport à la longueur de cette partie - i.e. le nombre d'occurrences de la forme voulue divisé par le nombre d'occurrences totales de la partie.

On observe que Royal est la seule à utiliser le mot "pacte" extrêmement fréquemment par rapport aux autres formes utilisées. En effet, les occurrences de ce mot sont très rares dans le discours des autres candidats, par rapport aux mots qu'ils emploient.

De plus, relativement aux longueurs des parties, ce sont Bayrou et Royal qui utilisent le plus l'appellation "projet". Encore une fois, on remarque que les trois formes représentent un très faible poids dans le discours de Le Pen.

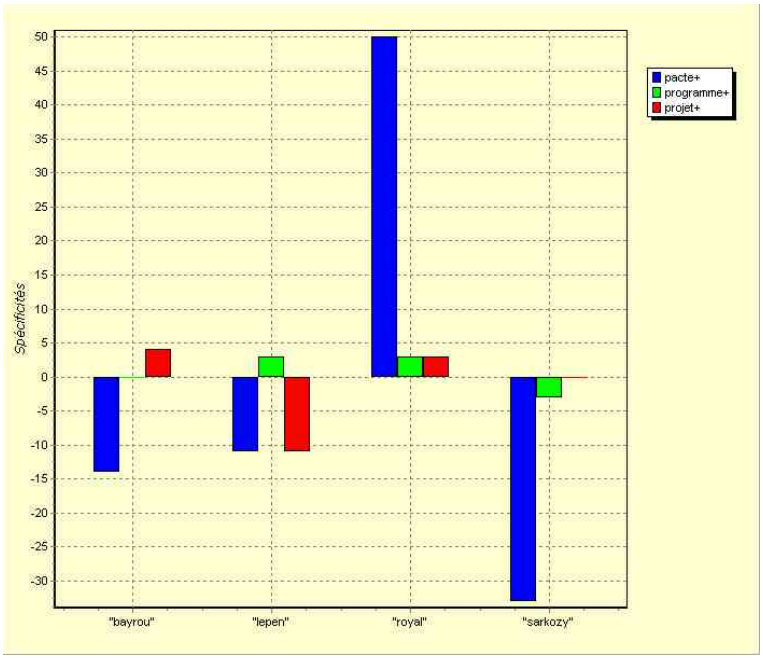

Fig.4 - Spécificités des formes recherchées par partie

Ce graphique nous montre une fois de plus, l'emploi abondant de "pacte" par Ségolène Royal par rapport aux autres candidats.

Partition par candidat et par mois Nous n'avons retenu ici que le graphe des fréquences relatives, plus explicite pour l'observation d'une évolution temporelle. En effet, les parties par mois n'ayant pas nécessairement une taille équivalente, il est plus intéressant d'observer la fréquence d'une forme dans une partie par rapport à la taille de cette partie.

Partitionner le corpus avec la balise *candidat-date* imposant un grand nombre de parties, notre graphe a été agrandi et découpé en quatre parties afin que les données soient lisibles.

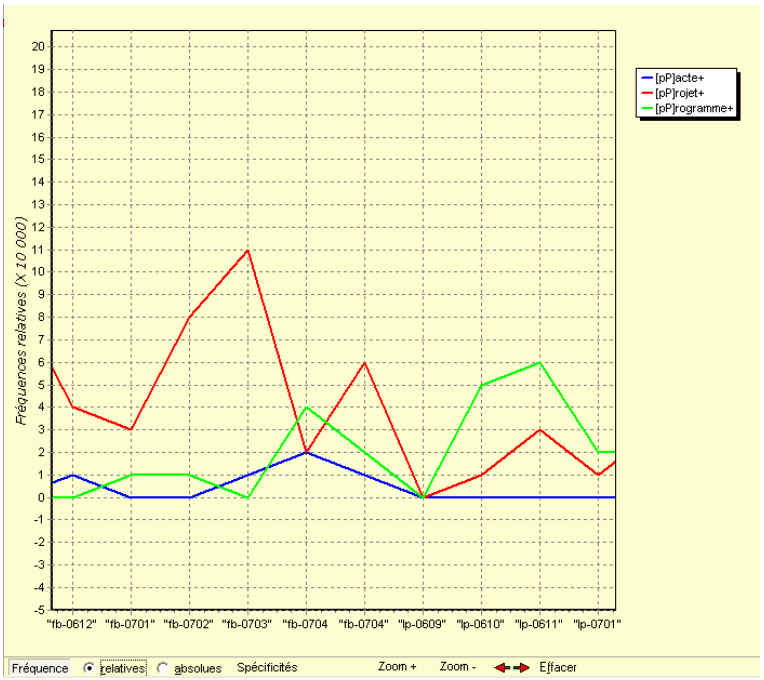

Fig.5 - Parties par mois de François Bayrou (fb)

On observe que l'emploi du mot "projet" chez Bayrou a augmenté de façon importante de janvier à mars. Il a légèrement baissé en avril, mais est resté relativement fréquent. Notons tout de même que le mot "programme" a été davantage utilisé en avril.

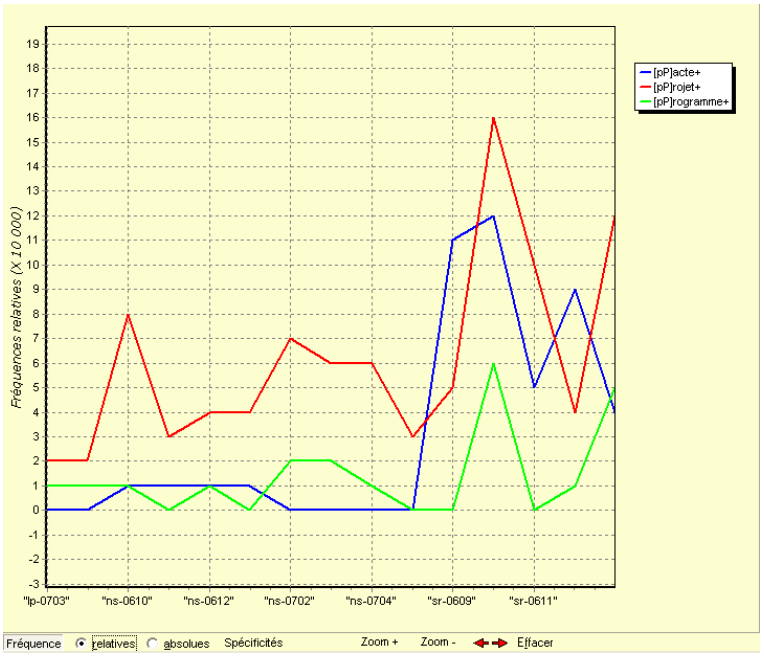

Fig.6 - Parties par mois de Nicolas Sarkozy (ns)

On remarque que le mot "projet" a été utilisé d'avantage à partir d'octobre 2006, et de façon relativement constante ensuite. C'est le mot le plus utilisé des trois par Sarkozy tout au long de la campagne.

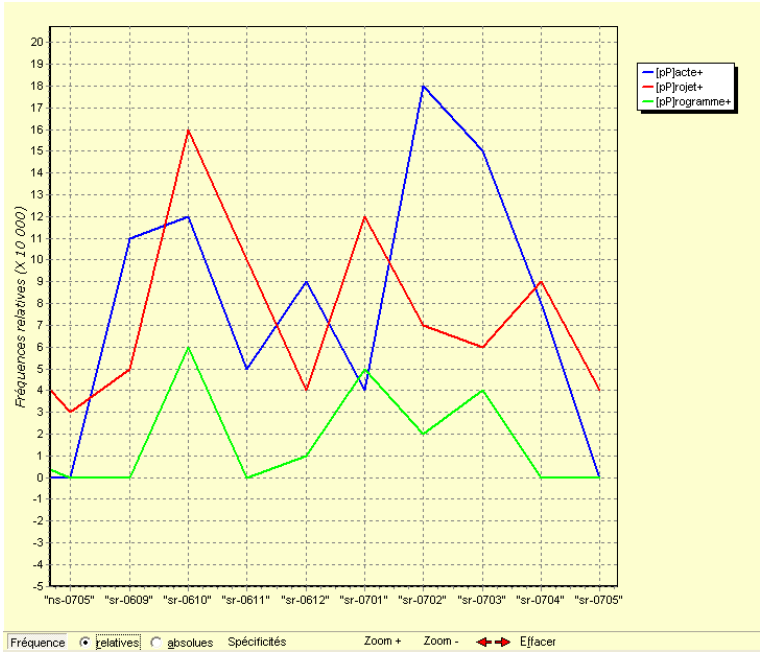

Fig.7 - Parties par mois de Ségolène Royal (sr)

L'observation de ce graphique est intéressante. En effet, on observe que l'emploi de "projet" a été plus fréquent que celui de "pacte" d'octobre 2006 à janvier 2007 (mis à part en décembre, où on peut se demander si l'emploi de "pacte" n'était pas relatif au "pacte écologique" de Nicolas Hulot). C'est à partir de février que la montée de l'utilisation de "pacte" par Ségolène Royal a été fulgurante.

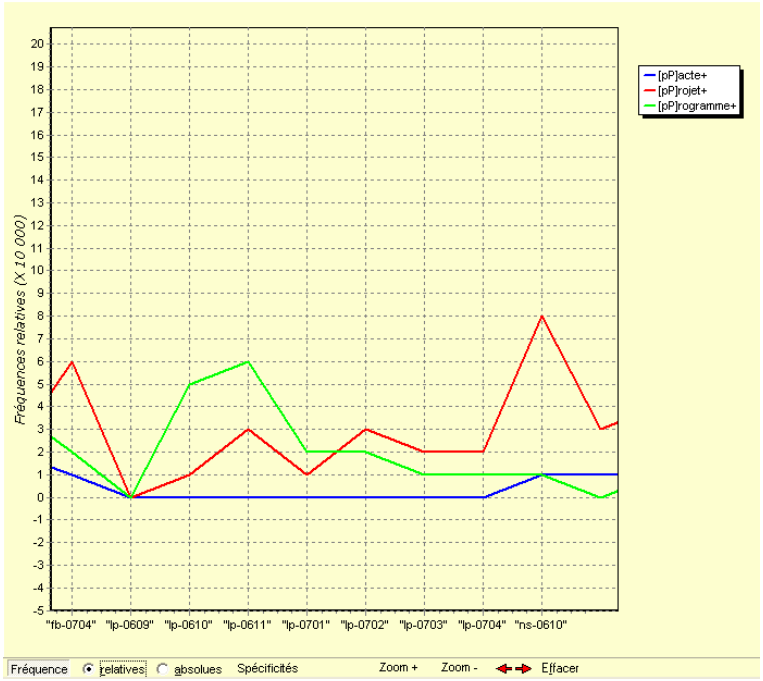

Fig.8 - Parties par mois de Jean-Marie Le Pen (lp)

L'évolution temporelle des emplois pour la partie de Jean-Marie Le Pen montre une fois de plus que ces mots ont été peu utilisés par le candidat. Notons néanmoins une légère augmentation de la fréquence d'utilisation du mot "programme" en octobre et novembre 2006.

Concordances

[13] Une concordance nous permet de rendre compte des différents contextes dans lesquels on peut trouver une forme. Nous avons décidé de trier ces contextes dans l'ordre alphabétique selon le premier mot qui suit la forme observée.

Partition par candidat Concordance de programme+

Pour le candidat Bayrou, "programme" est le plus fréquemment utilisé dans les contextes suivants : "programme commun", "programme politique".

Pour le candidat Le Pen, on observe surtout les collocations: "programme de construction", "programme de redressement", "programme de réformes".

[13]Nous vous invitons à consulter la version HTML du Rapport pour la visualisation des résultats de concordances, trop volumineux pour figurer ici. Nous faisons état des commentaires les concernant.

La candidate Royal emploie le mot le plus souvent dans les contextes suivants: "programme de logement", "programme de la droite", "programme scolaire".

Chez Nicolas Sarkozy, on retrouve les motifs "programme d'armement", "programme de qualité".

On observe que "programme" est souvent suivi d'un syntagme prépositionnel complément du nom, on parle surtout de "programme de...". En revanche, Bayrou produit davantage de constructions de la forme "programme + Adjectif".

Concordance de projet+

Chez François Bayrou, les constructions avec "projet" les plus présentes sont :"projet d'espoir", "projet de résistance", "projet de société", "projet républicain", "projet social". Il utilise également beaucoup "projet pour..." ("...la France", "...le Monde", "...le XXIème siècle").

Dans les discours de Jean-Marie Le Pen, on n'observe pas particulièrement de motifs récurrents contenant "projet".

Chez Ségolène Royal, on trouve essentiellement "projet présidentiel", "projet socialiste", "projet de société", plusieurs occurrences de "projet contre projet(s)", ainsi que "mon projet c'est", et des constructions du type "projet que je..." ("...porterai", "...présenterai").

Nicolas Sarkozy, quant à lui, utilise le plus souvent les constructions "projet collectif", "projet de civilisation", "projet politique", "projet pour la France", "projet présidentiel", "projet républicain".

Concordance de pacte+

Chez le candidat Bayrou, on observe "pacte écologique", "pacte républicain".

Chez le candidat Le Pen, encore une fois, il n'y a aucune construction récurrente de "pacte". Le mot n'apparaît que trois fois dans ses discours.

Ségolène Royal utilise un grand nombre de qualificatifs pour "pacte". On trouve essentiellement "pacte avec les jeunes", "pacte de confiance", "pacte de stabilité", "pacte présidentiel", "pacte républicain", "pacte social", "pacte de la réussite", et aussi "pacte écologique".

Chez Nicolas Sarkozy, on observe "pacte écologique", "pacte européen".

Notons que Royal est la seule à caractériser le mot "pacte" de "présidentiel". De plus, le mot "pacte" est plus souvent utilisé en général dans le contexte "pacte

écologique" de Nicolas Hulot.

Partition par candidat et par mois Concordance de projet+

On observe que, dans le discours de Bayrou, "projet républicain" et "projet de société" apparaissent massivement en mars 2007.

Chez Royal, "projet présidentiel" apparaît essentiellement en janvier.

Sarkozy emploie davantage "projet politique" et "projet présidentiel" en avril.

Concordance de pacte+

L'observation intéressante de cette concordance est le fait que Royal parle de "pacte de stabilité" en octobre, de "pacte social" en novembre et décembre, du "pacte écologique" de Nicolas Hulot en janvier, et introduit de façon massive son "pacte présidentiel" à partir de février. C'est en avril qu'elle commence à utiliser "pacte républicain".

Corpus Présidentielles

Nous avons tenté de mettre en regard les observations obtenues lors de cette exploration lexicométrique du corpus *Discours* avec les données du corpus *Présidentielles*.

Nous avons donc réalisé une brève étude des partitions par mois de ce corpus. En effet, on peut observer certains phénomènes quant à l'évolution temporelle des emplois des mots qui nous intéressent, dans le milieu journalistique, à savoir dans la rubrique "Présidentielles 2007" du journal Le Monde.

Graphiques de ventilation par mois

Fig.1 - Fréquences absolues

"projet" est le mot le plus fréquemment employé. Vient ensuite le mot "programme", puis l'appellation "pacte". De plus, on remarque un accroissement constant de la fréquence de "projet" et "programme", de novembre 2006 à mars 2007. "Pacte", en revanche, voit son emploi augmenter jusqu'en février, et diminuer ensuite.

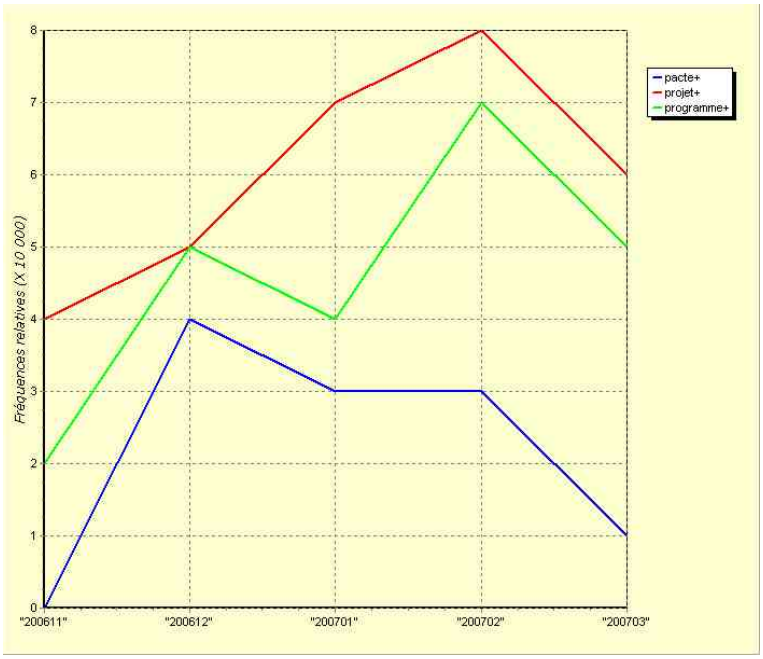

Fig.2 - Fréquences relatives

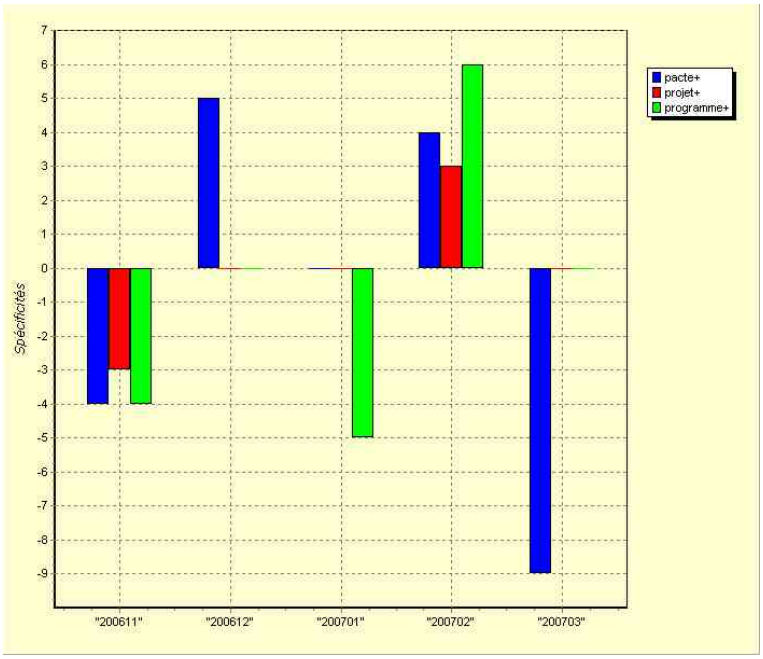

Fig.3 - Spécificités

Ce graphique nous montre que les trois mots connaissent un emploi anormalement élevé en février. De plus, le mot "pacte" est spécifique négatif du mois de mars, et le mot "projet" du mois de janvier.

Concordances par mois [14]

Ces concordances nous montrent tout d'abord que, en décembre 2006 et janvier 2007, on parle essentiellement dans la rubrique Présidentielles du Monde du Pacte Ecologique de Nicolas Hulot. C'est uniquement au mois de février que le "pacte présidentiel de Ségolène Royal" apparaît.

De plus, on trouve à plusieurs reprises, "projet d'espoir" - livre programme de François Bayrou -, et on parle autant de "projet socialiste", que de "projet de l UMP" et de "projet de l'UDF".

Enfin, on observe de façon équivalente "programme socialiste", "programme de l'UMP", "programme de François Bayrou" et "programme du FN" ou "programme de Jean- Marie Le Pen".

Notons que le mot "projet" n'est jamais utilisé dans ce corpus pour caractériser les propositions de Le Pen, et que le mot "pacte" est employé pour définir les propositions de la candidate Royal uniquement.

[14] Nous vous invitons à consulter la versino HTML du Rapport pour la visualisation des concordances, trop volumineuses pour figurer dans cette version. Nous rapportons les commentaires les concernant.

2.1.2 Graphes de mots Pajek

Introduction

Pajek est un logiciel qui permet de générer des graphes rendant compte des attirances cotextuelles d'unités lexicales. Il construit un graphe à partir d'un fichier au format .txt ou .net (Pajek networks) dans lequel les noeuds (mots) et le poids des arcs reliant ces noeuds sont indiqués.

De tels graphes peuvent être très utiles à la visualisation des relations de collocation des unités.

Extraction des patrons

[15] Nous disposons d'un fichier XML par candidat; fichiers que nous avons créés à partir de l'étiquetage du corpus Discours (cf. Partie Corpus du rapport). Ces fichiers sont structurés de la façon suivante:

```
<discours>
<s>
 <e>
  <st>forme</st>
  <lm>lemme</lm>
  <tp>CATEGORIE</tp>
 </e>
 <e>
  ...
 </e>
</s>
<s>
...
</s>
</discours>
```

Fig.1 - Structuration XML du corpus Discours

Nous avons extrait, pour chaque candidat, les patrons NOM ADJ où le lemme de NOM est égal soit à "pacte", soit à "programme", soit à "projet". Voici la feuille de style établie:

[15] cf.version HTML du Rapport pour visualiser les résultats complets des extractions

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
<xsl:output method="text"/>

<xsl:template match="/">
 <xsl:apply-templates/>
</xsl:template>

<xsl:template match="e">
 <xsl:if test="(lm[(text()='pacte')or(text()='programme')or(text()='projet')])
    and(following-sibling::e[1][tp[text()='ADJ']])">
   <xsl:value-of select="lm"/><xsl:text> </xsl:text>
   <xsl:value-of select="following-sibling::e[1]/st"/><xsl:text>
</xsl:text>
 </xsl:if>
<xsl:template>


</xsl:stylesheet>
```

Fig.2 - Feuille de style appliquée à la version XML du corpus Discours

Ces patrons ont été extraits dans le but d'observer les adjectifs attirés par "pacte", "programme" et "projet" pour nos quatre candidats.

Génération des graphes

Nous avons appliqué le programme perl Patron2graphml.pl (cf. Projet BàO, série4)[16] à nos fichiers de patrons afin d'obtenir des fichiers au format GraphMl. Ensuite, nous avons appliqué à ces fichiers résultats la feuille de style GraphML2Pajek.xsl (cf. Projet BàO[17]), pour obtenir des fichiers au format texte, compatibles avec Pajek.

Les graphes pour chaque candidat ont été générés. Nous présentons ci-dessous deux versions pour chaque candidat. La première indique les valeurs des arcs, i.e. la fréquence des patrons. La seconde applique aux arcs une épaisseur différente selon l'importance de la fréquence des patrons. Notons que nous n'avons qu'une version pour Jean-Marie Le Pen, puisque la fréquence des patrons est toujours égale à 1.

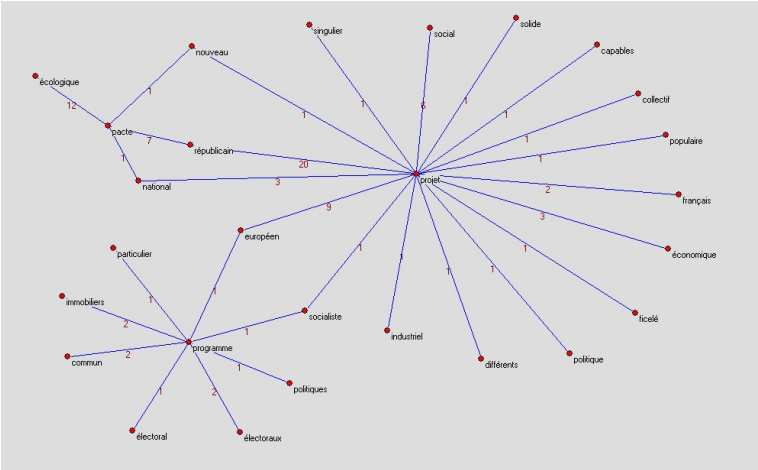

Fig.1 - François Bayrou

[16]http://tal.univ-paris3.fr/plurital/travaux-2006-2007/tr0607-masterproj-sem2.htm
[17]http://tal.univ-paris3.fr/plurital/travaux-2006-2007/tr0607-masterproj-sem2.htm

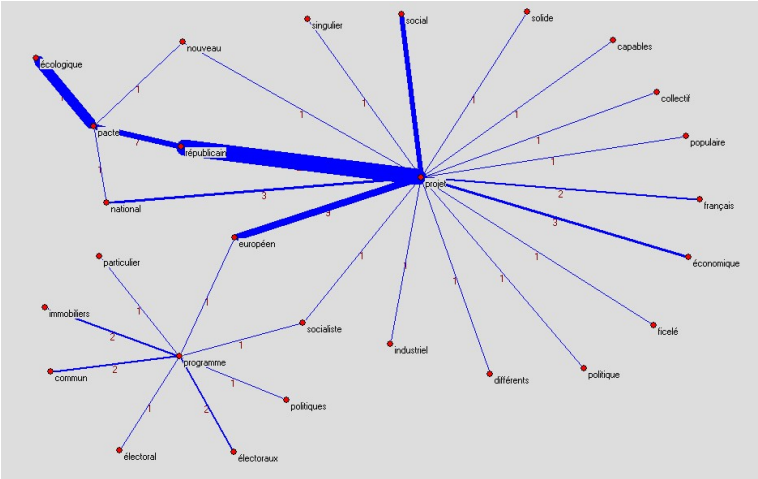

Fig.2 - François Bayrou

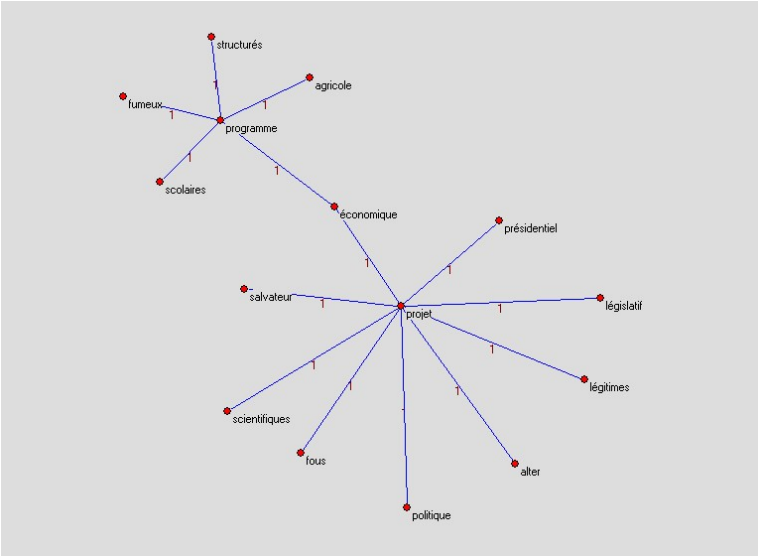

Fig.3 - Jean-Marie Le Pen

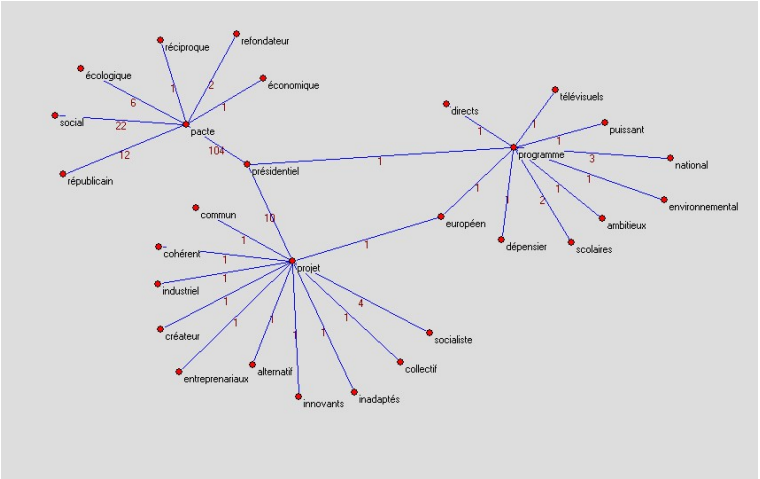

Fig.4 - Ségolène Royal

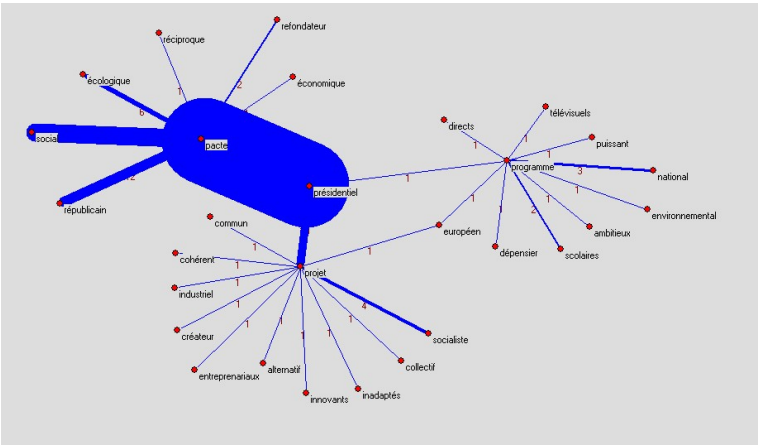

Fig.5 - Ségolène Royal

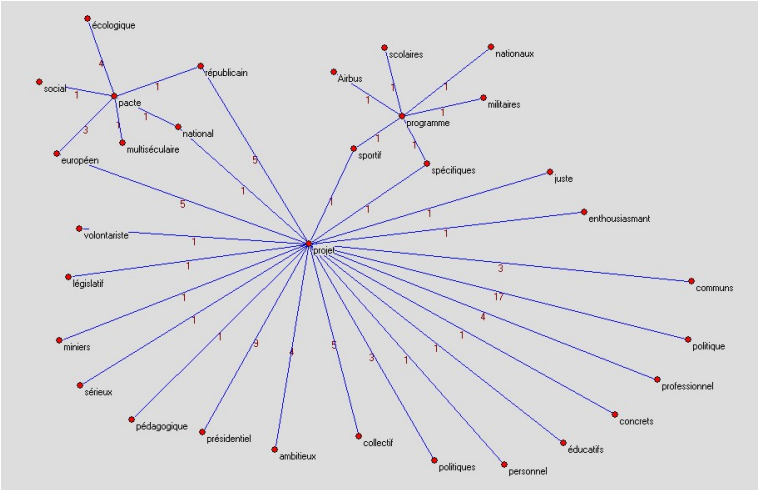

Fig.6 - Nicolas Sarkozy

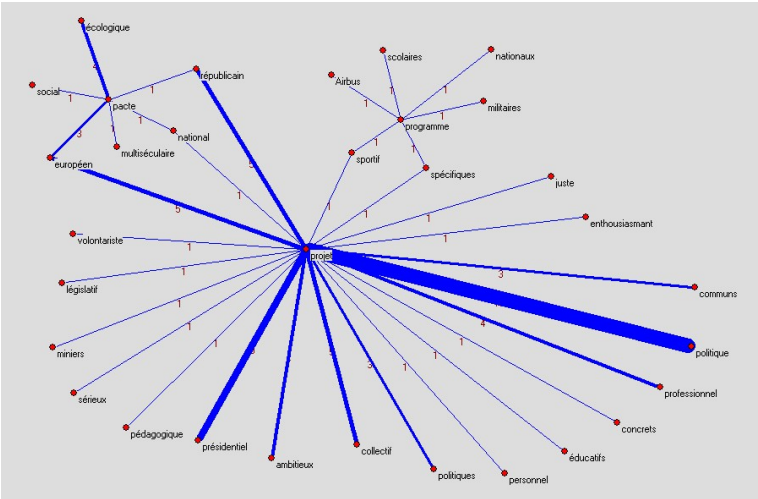

Fig.7 - Nicolas Sarkozy

A première vue, on remarque que les graphes de Nicolas Sarkozy et de François Bayrou se ressemblent quant à leur forme et leur répartition. D'une part, ils ont tous deux un patron dominant, "projet politique" pour Sarkozy et "projet républicain" pour Bayrou, avec une vingtaine d'occurrences. D'autre part, le nom le plus utilisé, à savoir "projet", est celui qui attire le plus d'adjectifs (en moyenne trois fois plus que les noms "pacte" et "programme").

En revanche les graphes de Ségolène Royal et de Jean-Marie Le Pen se distinguent chacun des autres.

En effet, Royal bat de loin le record du patron le plus fréquent, "pacte présidentiel" avec 104 occurrences. De plus, les adjectifs en relation avec le nom qu'elle emploie le plus souvent, à savoir "pacte", sont peu variés mais

très fréquents. Enfin, on observe un équilibre de la fréquence et de la variété des adjectifs attirés par les noms "programme" et "projet". Le graphe de Jean-Marie Le Pen nous montre une fois de plus que les noms "programme", "projet" et (surtout) "pacte" ne font quasiment pas partie de son discours, et qu'aucune construction du type "pacte—projet—programme + adjectif" n'est répétée.

2.1.3 Nuages de mots

Introduction

Les nuages de mots permettent une représentation des mots les plus utilisés.

Nous avons généré des nuages de mots à partir des patrons extraits dans la partie précédente, afin d'obtenir un complément à la représentation des utilisations des adjectifs qualifiant "pacte", "programme" et "projet" par nos candidats.

Nous avons utilisé le logiciel DICO de Jean Veronis qui, à partir d'un fichier texte, génère un fichier dictionnaire répertoriant les différentes formes présentes et leur fréquence correspondante. Le nuage est généré ensuite par le logiciel TagCloudBuilder. Celui-ci prend en entrée le fichier dictionnaire obtenu à partir de DICO et construit un fichier HTML dont le code contient un style - taille et couleur - à appliquer aux mots selon leur fréquence. Ce style est emprunté au Nébuloscope de Jean Véronis. Des liens pointant vers une recherche Google sont automatiquement insérés pour chaque mot.

Il aurait fallu, pour réaliser un "bon nuage", modifier ces liens de telle sorte que, en cliquant sur un mot, on puisse visualiser les différents contextes de ce mot dans le corpus. Or, nous avons choisi de ne pas nous intéresser aux liens puisque, les nuages ayant été établis à partir de nos fichiers de patrons, les concordances des mots n'étaient pas pertinentes ici. Ces nuages nous permettent simplement de visualiser d'une façon différente l'importance des adjectifs liés à "pacte", "programme" et "projet" dans le discours de chacun, bien que la représentation en graphes Pajek constitue évidemment un meilleur outil dans ce cas précis.

Génération des nuages

Fichiers générés par DICO

Forme	Fréquence
PROJET	54
RÉPUBLICAIN	27
PACTE	21
ÉCOLOGIQUE	12
PROGRAMME	12
EUROPÉEN	10
SOCIAL	6
NATIONAL	4
ÉCONOMIQUE	3
COMMUN	2
ÉLECTORAUX	2
FRANÇAIS	2
IMMOBILIERS	2
NOUVEAU	2
POLITIQUE	2
SOCIALISTE	2
CAPABLES	1
COLLECTIF	1
DIFFÉRENTS	1
ÉLECTORAL	1
FICELÉ	1
INDUSTRIEL	1
PARTICULIER	1
POLITIQUES	1
POPULAIRE	1
SINGULIER	1
SOLIDE	1

Fig.1 - DICO pour Bayrou

Forme	Fréquence
PROJET	9
PROGRAMME	5
ÉCONOMIQUE	2
AGRICOLE	1
ALTER	1
EUROPÉEN	1
FOUS	1
FUMEUX	1
LÉGISLATIF	1
LÉGITIMES	1
POLITIQUE	1
PRÉSIDENTIEL	1
SALVATEUR	1
SCIENTIFIQUES	1
SCOLAIRES	1
STRUCTURÉS	1

Fig.2 - DICO pour Le Pen

Forme	Fréquence
PACTE	148
PRÉSIDENTIEL	115
PROJET	25
SOCIAL	23
PROGRAMME	13
RÉPUBLICAIN	12
ÉCOLOGIQUE	6
SOCIALISTE	4
NATIONAL	3
EUROPÉEN	2
REFONDATEUR	2
SCOLAIRES	2
ALTERNATIF	1
AMBITIEUX	1
COHÉRENT	1
COLLECTIF	1
COMMUN	1
CRÉATEUR	1
DÉPENSIER	1
DIRECTS	1
ÉCONOMIQUE	1
ENTREPRENARIAUX	1
ENVIRONNEMENTAL	1
INADAPTÉS	1
INDUSTRIEL	1
INNOVANTS	1
PUISSANT	1
RÉCIPROQUE	1
TÉLÉVISUELS	1

Fig.3 - DICO pour Royal

Forme	Fréquence
PROJET	73
POLITIQUE	17
PACTE	11
PRÉSIDENTIEL	9
EUROPÉEN	8
PROGRAMME	6
RÉPUBLICAIN	6
COLLECTIF	5
ÉDUCATIF	5
AMBITIEUX	4
ÉCOLOGIQUE	4
PROFESSIONNEL	4
COMMUNS	3
POLITIQUES	3
NATIONAL	2
SPÉCIFIQUES	2
SPORTIF	2
AIRBUS	1
CONCRETS	1
ÉDUCATIFS	1
ENTHOUSIASMANT	1
JUSTE	1
LÉGISLATIF	1
MILITAIRES	1
MINIERS	1
MULTISÉCULAIRE	1
NATIONAUX	1
PÉDAGOGIQUE	1
PERSONNEL	1
SCOLAIRES	1
SÉRIEUX	1
SOCIAL	1
VOLONTARISTE	1

Fig.4 - DICO pour Sarkozy

Un exemple de nuage de mots [18]

Fig.5 - Exemple du nuage de mots pour François Bayrou

[18] Nous vous invitons à consulter la version HTML du Rapport pour la visalisation et l'essai des nuages de mots générés ici.

2.2 Phénomènes de la négation

Cette partie est consacrée à l'exploitation de nos corpus du point de vue de la négation. Après avoir présenté brièvement les raisons pour lesquelles nous nous intéressons à ce phénomène, nous montrons quelques données quantitatives obtenues à l'aide d'un script d'extraction et de calcul sur les phrases négatives. La suite de cette section présente les résultats d'une extraction de patrons grammaticaux relevant quelques aspects du phénomène de la négation.

2.2.1 Motivation

Pourquoi pas la négation?

Une observation des spécificités dans le corpus Discours à l'aide de Lexico3 révèle une forte disproportion dans l'emploi de la négation prédicative (identifiée d'après la fréquence des morphèmes *ne* et *n*) entre les différents candidats). Les fréquences relatives des formes *ne, n* et *pas* ainsi que le taux de la spécificité de chacune de ces formes pour chacun des candidats ne peuvent être bien observés que sur les graphiques suivants :

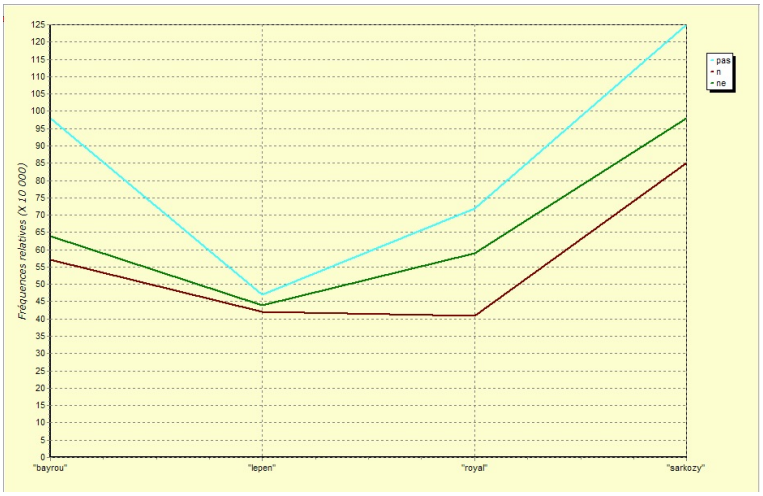

Fig.1 - Formes négatives retenues - fréquences

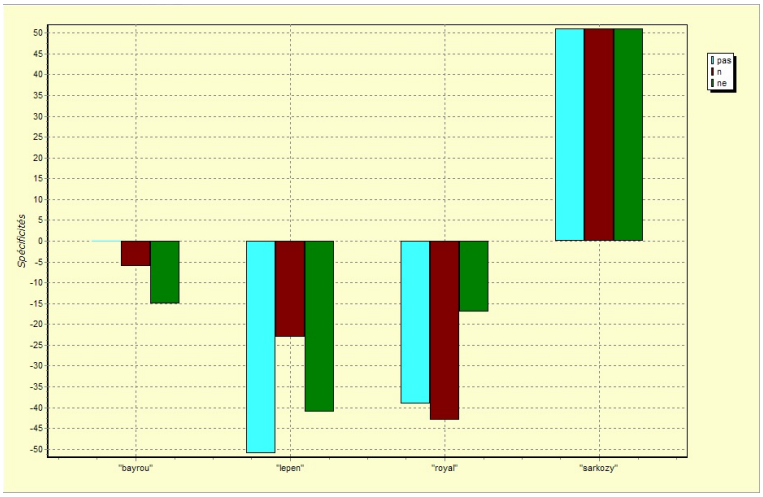

Fig.2 - Formes négatives retenues - spécificités

Nous n'allons pas faire une analyse des différents emplois de la négation qui peuvent être recensés dans notre corpus et nous n'essayerons pas non plus de les interpréter pour arriver à des conclusions qui ne seraient pas banales. Nous n'extraierons et ne présenterons quelques informations que nous ne jugeons pas inintéressantes.

Définitions

Afin d'éviter toute confusion, nous devons définir les termes qui sont utilisés dans cet exposé :

forme négative La suite de caractères *ne* ou *n*, qui sont synonymes. L'avantage de cette forme est qu'elle n'est pas ambiguë, nous pouvons alors baser notre extraction des phrases négatives sur la présence de cette forme dans une phrase donnée.

phrase négative Une suite de caractères terminée par un point (sous certaines restrictions, voir la problème de la segmentation en phrase dans la section Corpus) qui contient au moins une forme négative (morphème négatif) *ne* ou *n*.

Nous n'allons pas distinguer entre les différentes fonctions que peut avoir ce morphème ((1) négative : "Nous ne sommes pas d'accord !"; (2) comparative : "Et en attendant, il faudra que l'Europe se protège et se protège beaucoup plus efficacement qu'elle ne le fait contre toutes ces formes de délocalisations et de destructions d'emplois, j'en fais ici le serment !"; (3) explétive : Pierre a toujours peur que Marie ne s'en aille.

2.2.2 Extraction des phrases négatives

Script d'extraction

Nous étions intéressés par certaines informations liées à l'emploi de la négation (des morphèmes négatifs) dans nos corpus, notamment à la fréquence des phrases négatives et à la fréquence de formes négatives. Nous avons écrit un script Perl ExtraitNeg.pl[19] qui parcourt le fichier (contenant le texte brut segmenté en une phrase par lignes) passé en argument, imprime les phrases négatives extraites dans un fichier et imprime un rapport avec les informations suivantes :

```
Nombre de phrases dans le fichier
Nombre de lignes vides
Nombre de phrases négatives
Pourcentage des phrases négatives
Nombre de mots
Nombre de formes négatives (ne ou n)
Pourcentage des formes négatives (par rapport à la totalité des mots)
Nombre moyen de formes négatives par phrase négative
Phrase avec le plus de formes négatives
```

Fig.1 - Informations sur les phrases négatives extraites

La boucle de lecture du fichier est la suivante :

[19] cf.version HTML du Rapport pour téléchargement

```
foreach $ligne()
{
    @slova = ();
    $j = 0;
    $jNeg = 0;

    if($ligne !~ /^$/)
    {
        $ph[$i] = $ligne;

        # EXTRACTION NEGATION
        if (($ligne =~ /.*\sne\s.*/) || ($ligne =~ /.*\sn'.*/))
        {
            $phNeg[$iNeg] = $ligne;
            $iNeg++;
            print NEG $ligne;
        }

        chomp $ligne;
        $phlong[$i] = length($ligne);
        &Separate;
        @slova = split(/\s/,$ligne);

        foreach (@slova)
        {
            $mots[$k] = $_;
            $motslong[$k] = length($_);
            $k++;
            $j++;

            # EXTRACTION NEGATION
            if (($_ eq "ne") || ($_ eq "n"))
            {
                $kNeg++;
                $jNeg++;
            }
        }

        # EXTRACTION NEGATION
        if (($ligne =~ /.*\sne\s.*/) || ($ligne =~ /.*\sn\s.*/))
        {
            $phnbNeg[$iNeg-1] = $jNeg;
        }

        $phnbmot[$i] = $j;
        $i++;
    }
    else
    {
        $iV++;
    }
}
```

Fig.2 - Boucle d'extraction pour les phrases négatives

Cet algorithme, qui est basé sur la boucle du script StatsPhMot.pl utilisé pour les calculs dans la section Corpus, structure les données dans la mémoire afin que l'on puisse extraire (par les calculs dans la seconde partie du script) les informations qui nous intéressent[20].

Graphiques négation

Les informations extraites à l'aide du script décrit ci-dessus sont présentées à l'aide des graphiques dans la partie qui suit. Comme dans le cas des informations quantitatives concernant la longueur des phrases et des mots, nous n'allons

[20]Nous vous invitons à vous reporter à la version HTML du Rapport pour avoir les rapports d'extraction du script ainsi que les fichiers d'extraction

pas interpréter les résultats.

La présence des formes négatives Le graphique suivant montre le taux de phrases négatives (contenant le morphème *ne* ou *n*) dans les partitions du corpus, c'est-à-dire que, par exemple, 26,36 % des phrases extraites des discours de Bayrou contiennent au moins un de ces morphèmes.

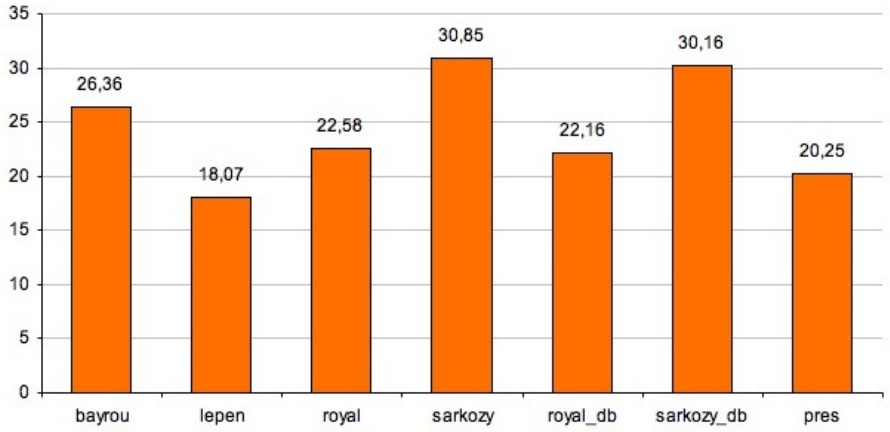

Fig.1 - Le pourcentage des phrases négatives

Le graphique suivant montre le taux de formes négatives (des morphèmes *ne* ou *n*) par rapport aux autres mots, c'est-à-dire que, par exemple, dans les discours de Bayrou, l'ensemble des formes *ne* ou *n* prononcées dans ses discours représentent 1,22 % de l'ensemble de tous les mots employés.

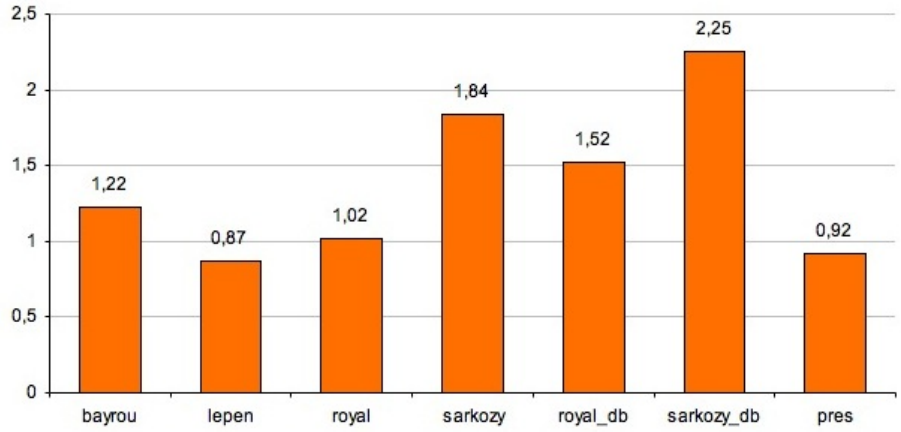

Fig.2 - Le pourcentage des formes négatives

Le graphique suivant montre le nombre moyen de morphèmes *ne* ou *n* pour une phrase négative, c'est-à-dire que, par exemple, une phrase négative dans

les discours de Bayrou contient en moyenne 1,25 formes négatives. Ceci est un indice de la multiple négation utilisée dans une seule phrase (à ne pas confondre avec la double négation).

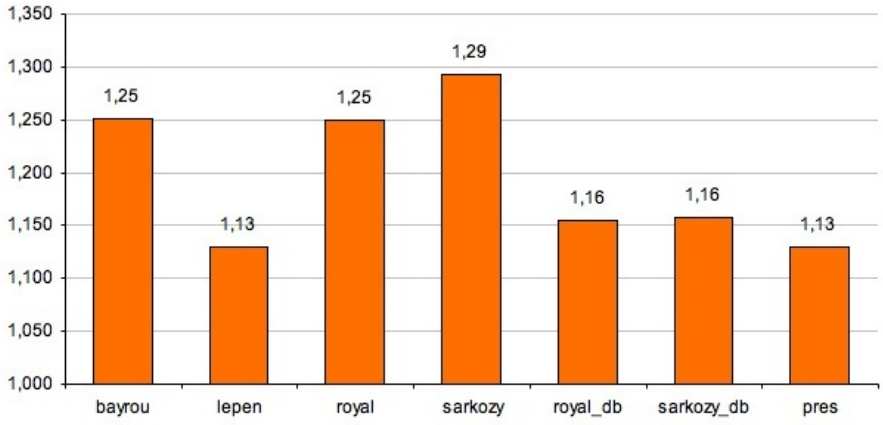

Fig.3 - Le nombre moyen de formes négatives par phrase négative

La longueur des phrases négatives Les deux graphiques suivants montrent la longueur moyenne d'une phrase négative en caractères et en mots en comparaison avec la longueur d'une phrase positive.

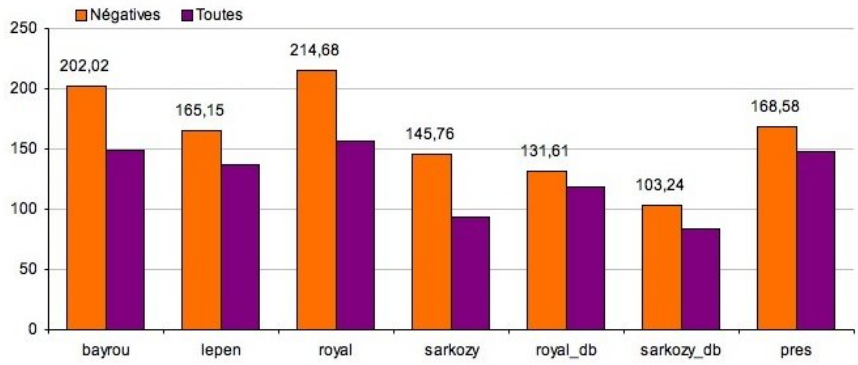

Fig.1 - La longueur moyenne d'une phrase négative en caractères alphanumériques

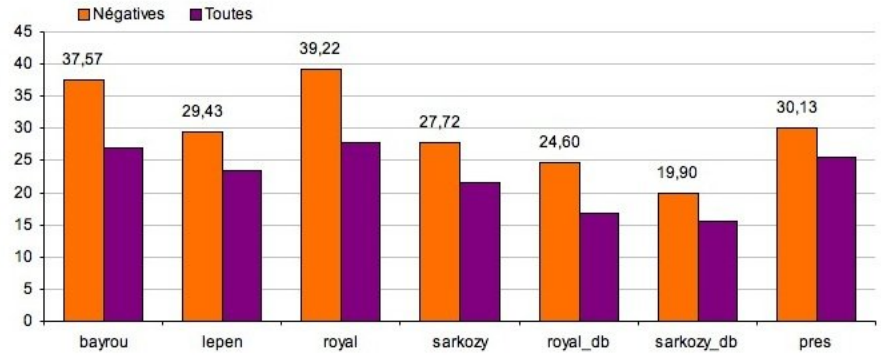

Fig.2 - Le nombre de mots dans une phrase négative

2.2.3 Extraction des patrons

Méthode

L'extraction des patrons et leur visualisation en forme de graphe avec le logiciel Pajek est une façon rapide et efficace d'exploiter un corpus.

En général, un patron peut être défini comme une suite de symboles (ou un motif) qui entre en correspondance avec certains tokens du corpus : par exemple, la suite des symboles *maillot jaune* (qui sont des tokens) est un patron qui est en correspondance avec toutes les suites de tokens *maillot jaune* présentes dans le corpus. L'extraction du patron *maillot jaune* aurait donc comme résultat la liste de tout les occurrences de la suite de tokens *maillot jaune* dans le corpus.

La meilleure façon d'utiliser la méthode d'extraction des patrons est de se servir des suites des étiquettes linguistiques qui sont associées à chaque token (dans un corpus annoté morphologiquement, bien entendu). Ceci nous permet de faire abstraction de la forme des tokens et de travailler uniquement avec les catégories linguistiques. Dans ce cas, une suite de symboles *Nom Adjectif* (qui sont des étiquettes) est un patron qui permet d'extraire du corpus toutes les suites des tokens auxquels sont associées les étiquettes correspondantes, c'est à dire par exemple *maillot jaune, secrétaire nationale, étendards éternels,* etc.

Si nous utilisons des patrons contenant les étiquettes linguistiques, nous sommes bien sûr à la merci des erreurs qui peuvent être faites par un étiqueteur (fausse attribution des étiquettes : du bruit ou du silence), néanmoins ceci est le prix à payer si nous voulons travailler avec des informations métalinguistiques générées automatiquement.

Le programme Pajek peut générer (à partir d'un fichier dans un format spécial) un graphe qui représente un ensemble des noeuds reliés par des arcs. Chaque noeud représente la valeur d'un token (ou de plusieurs tokens si leur valeurs sont égales) qui a été extrait à l'aide d'un patron donné. Les arcs signifient que les deux noeuds qui sont reliés ensemble, sont membres de la suite (du couple) des tokens extraites du corpus. Pour donner un exemple, si avec un patron *Nom Adjectif* nous extrayons du corpus les suites *maillot jaune, maillot blanc et étendards éternels,* nous obtenons un graphe avec 5 noeuds (*maillot, jaune, blanc, étendards, éternels*) avec les arcs entre *maillot* et *blanc*, entre *maillot* et *jaune*, et finalement, entre *étendards* et *éternels*.

Dans cette partie, nous allons utiliser la méthode d'extraction des patrons basée sur des étiquettes linguistiques afin d'exploiter et de présenter certains aspects de l'emploi de la négation dans nos corpus, notamment dans les discours des quatre candidats à la Présidentielle 2007.

Choix de patrons

Afin d'obtenir, un aperçu de l'emploi de la négation dans notre corpus, il fallait réfléchir au choix des patrons (suites des étiquettes linguistiques) qui pourraient donner l'information qui servirait le mieux à ce but. Ce choix doit être motivé, premièrement, par la pertinence des résultats obtenus et leur utilité pour une analyse plus détaillée ; et deuxièmement, par la lisibilité du graphe Pajek. En considérant ces critères, nous avons décidé d'observer deux constructions.

(1) *Pronom personnel + Verbe présent* : La première construction est le prédicat négatif ayant pour sujet un pronom personnel, par exemple : *on n'écoute pas, nous ne sommes pas, il n'est pas, je ne crois pas*, etc. Ceci peut donner une première impression de l'emploi de la négation au présent, bien que nous n'ayons aucune information sur le complément d'objet du prédicat (*on n'écoute pas (QUOI?), je ne crois pas (QUOI?)* et dans le cas des pronoms de la troisième personnes (à part on qui est "omnipersonnels" et *je, tu, nous, vous* qui sont ancrés dans la situation communicationnelle de chaque discours) sur le référent du sujet (*il (QUI?) ne croit pas*). En plus, dans le cas des verbes *être* et *avoir*, nous ne disposons pas de l'information s'il s'agit d'un auxiliaire d'une forme verbale analytique.

(2) *Verbe présent + Verbe infinitif* : La seconde construction, plus spécifique, est le prédicat négatif au présent ayant pour sujet un des trois pronoms personnels *je, nous* ou *on* (pour assurer la connaissance du référent du sujet), suivi par un complément infinitif, par exemple : *je ne peux pas accepter, nous ne pouvons pas ignorer, on ne peut pas travailler*, etc. Même si le résultat est plus spécifique que dans le cas de la première construction, l'information n'est toujours pas complète car nous n'avons aucune information sur les compléments de l'infinitif (*nous ne pouvons pas vouloir (QUOI?), nous ne pouvons pas laisser (QUOI?)*.

Les résultats que nous présentons peuvent servir surtout à montrer les possibilités de cette méthode et à donner une vision globale de l'emploi de la négation (ou plutôt de l'emploi des morphèmes négatifs) dans notre corpus. L'extraction

d'informations plus concrètes (qu'est-ce qui est nié, au fait) devrait être le sujet d'une recherche plus approfondie.

2.2.4 Verbe présent + Verbe infinitif

Outils d'extraction

Pour extraire la première construction, nous avons procédé de la façon suivante. Nous avons utilisé les fichier XML contenant les discours de chaque candidat annotés par TreeTagger et à l'aide de la feuille de style extraitNegPatron.xsl[21] nous obtenons pour chaque candidat une liste de suites de tokens correspondants à la transformation XSL suivante :

```
<xsl:if test="(lm[(text()='ne')])and(following-sibling::e[1][tp[text()='VER:pres']])and
     (preceding-sibling::e[1][tp[text()='PRO:PER']])">
     <xsl:value-of select="preceding-sibling::e[1]/st"/><xsl:text> </xsl:text>
     <xsl:value-of select="following-sibling::e[1]/lm"/>
 </xsl:if>
```

Fig.1 - Feuille de style XSL pour la mise en forme de l'étiquetage TreeTagger

Une partie du fichier de patrons obtenu à partir des discours de F. Bayrou est présenté ici :

```
on écouter
nous pouvoir
nous sommer|être
il être
il être
ils participer
on être
on voter
on être
on être
je avoir
je croire
elle être
je voir
```

Fig.2 - Exemples de patrons (discours de François Bayrou)

Nous allons utiliser les lemmes des verbes (lemmatisés par TreeTagger) pour rendre le graphe final plus clair - si nous nous étions servis des formes fléchies, le nombre de noeuds dans notre graphe pourrait être jusqu'à 6 fois plus importants. En utilisant les lemmes, nous ne perdons pas d'informations car chaque infinitif du verbe sera lié au pronoms personnels qui lui est associé dans le corpus. Cependant nous risquons d'amplifier certains problèmes de l'annotation automatique, car, comme on peut le voir dans l'extrait ci-dessus, certaines lemmatisations ne sont pas correctes ou ne sont pas désambiguïsées comme dans le cas de la forme *sommes* qui est lemmatisée en *sommer — être* (ce qui explique une présence relativement importante du verbe *sommer* dans nos graphes).

[21] cf.version HTML pour téléchargement

Les fichiers obtenus par la transformation XSL doivent être traités successivement par le script Perl patron2graphml.pl[22] et par une transformation XSL graphml2pajek.xsl[23] qui génèrent un fichier en format Pajek. Après quelques manipulations avec le programme Pajek qui servent principalement à l'amélioration de l'apparence visuelle du graphe, nous obtenons les graphes qui sont présenté plus bas.

Les graphes doivent être lus de la façon suivante :
Le graphe du haut contient tous les couples pronom personnel + infinitif du verbe qui figurent dans les constructions *pronoms personnel + "ne" ou "n" + verbe au présent + adverbe négatif ("pas", "plus", "jamais", ...).*

Le graphe du bas lui est identique, le nombre d'occurrences de chaque couple *pronoms personnel + verbe* dans le corpus est exprimé par l'épaisseur de l'arc qui les relie, les arc les plus épais sont aussi les plus foncés.

Exemple : le noeud *je* relié par un arc avec le noeud croire représente la structure *je ne crois (pas, rien, jamais, ...).*

Bayrou

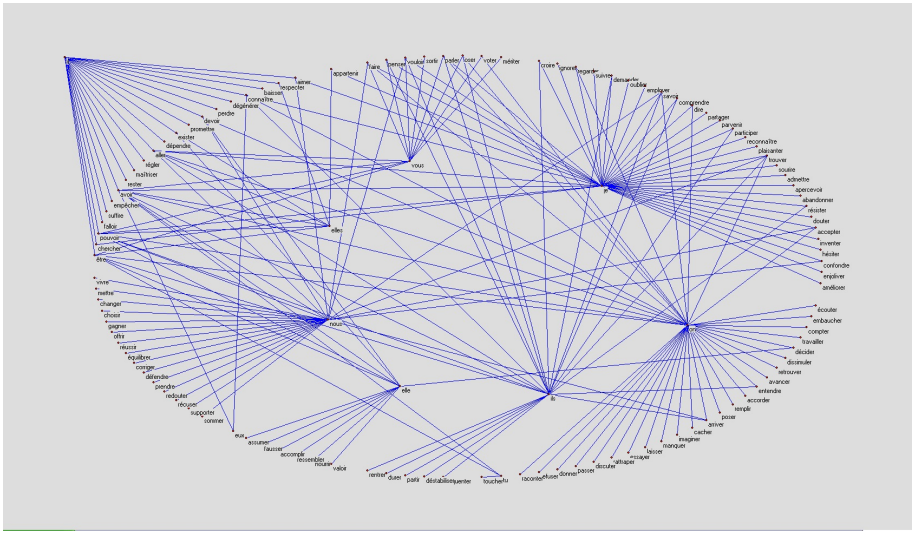

Fig.1 - Graphe 1 Bayrou

[22]cf.version HTML pour téléchargement
[23]cf.version HTML pour téléchargement

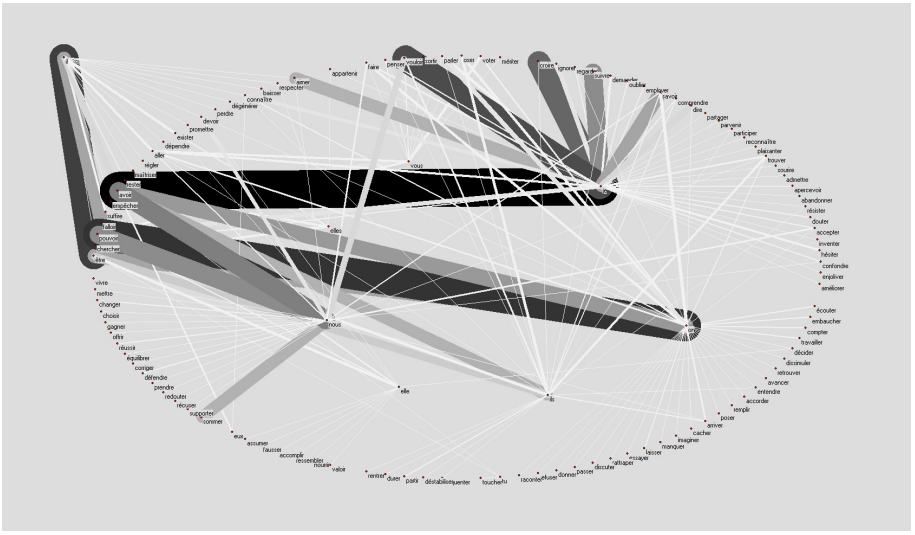

Fig.2 - Graphe 2 Bayrou

Le Pen

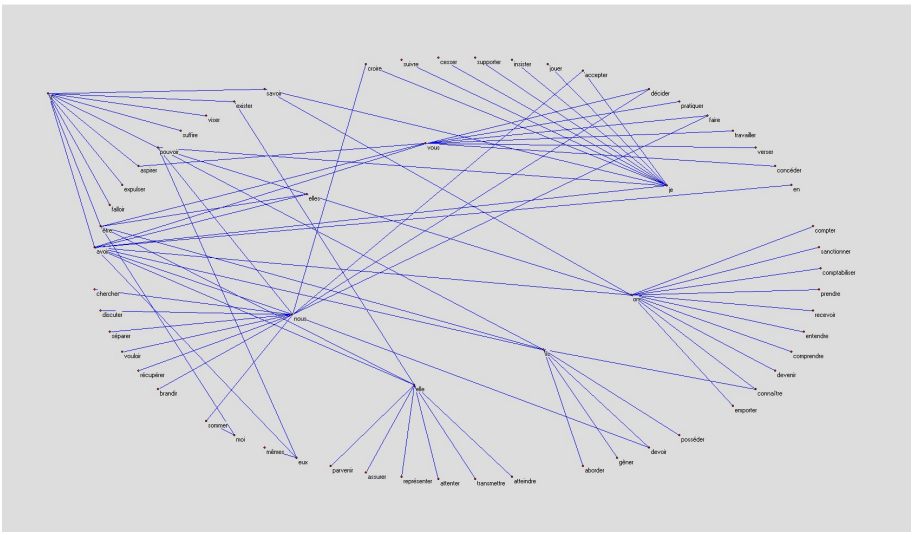

Fig.1 - Graphe 1 Le Pen

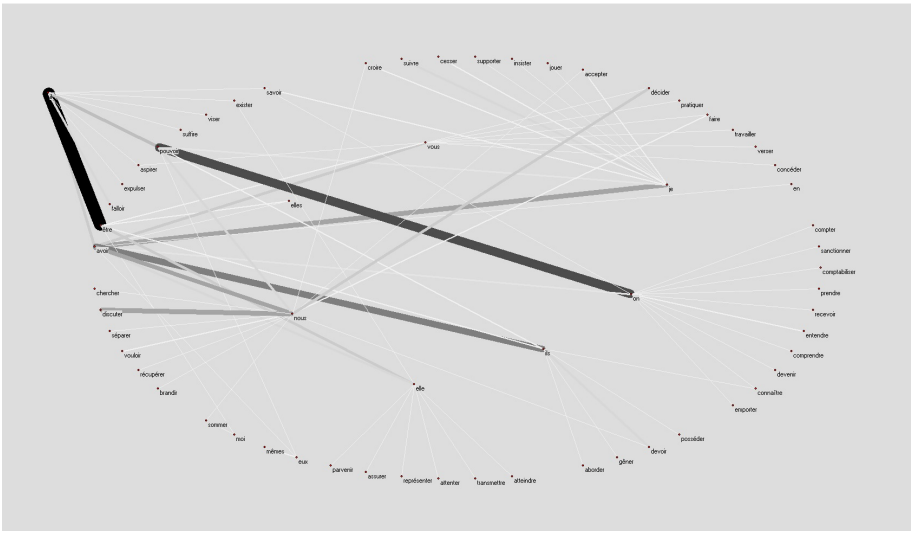

Fig.2 - Graphe 2 Le Pen

Sarkozy

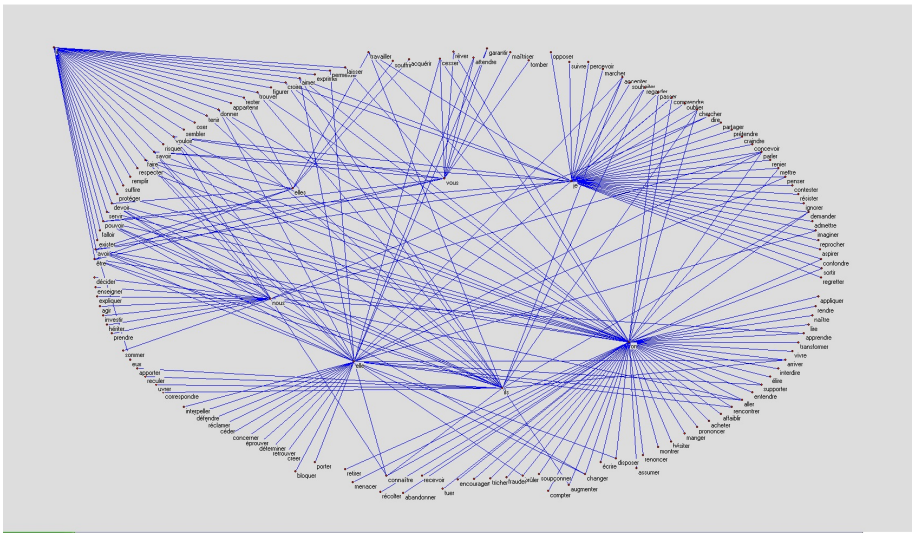

Fig.1 - Graphe 1 Sarkozy

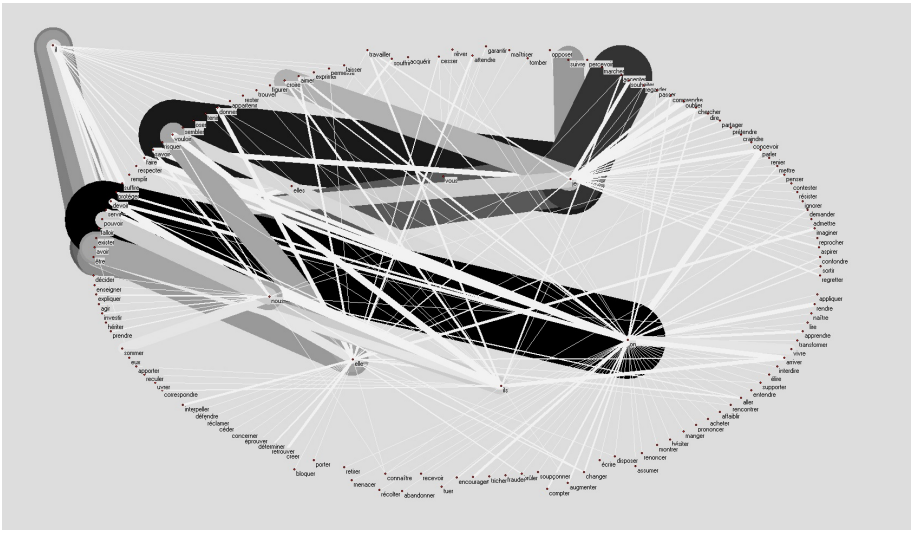

Fig.2 - Graphe 2 Sarkozy

Royal

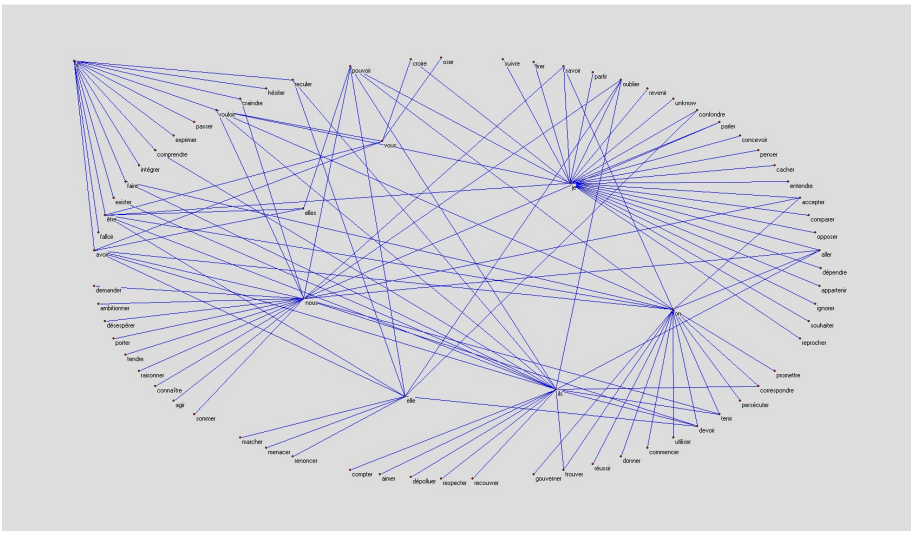

Fig.1 - Graphe 1 Royal

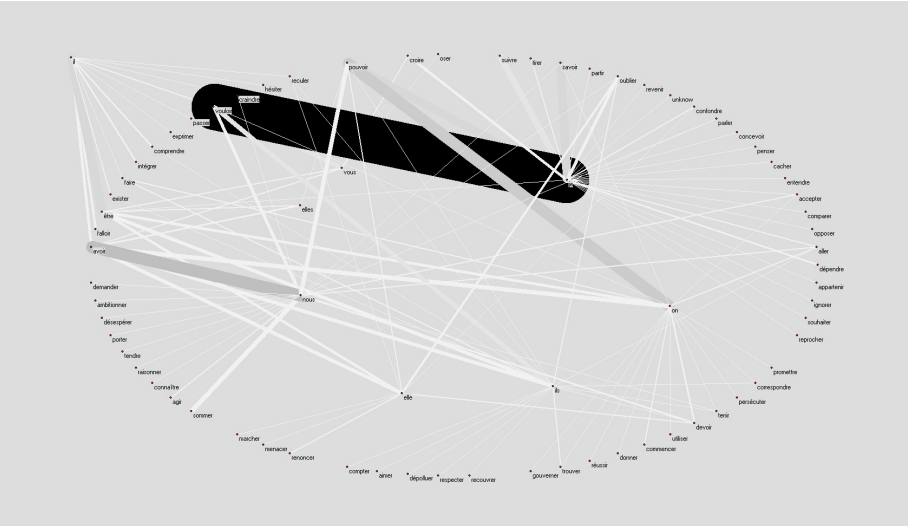

Fig.2 - Graphe 2 Royal

2.2.5 Pronom personnel + Verbe présent

Outils d'extraction

Dans le cas de la seconde structure, nous avons travaillé directement avec les versions texte-brut du notre corpus annoté par TreeTagger, un fichier par candidat. L'objectif était d'extraire le patron **PRO:PER ADV VER:pres ADV VER:infi** et de préparer les fichiers de patrons ainsi obtenus pour un affichage avec Pajek.

L'extraction du patron a été faite par le script ExtraitPatronsLemme.pl[24] qui lit le fichier annoté par TreeTagger passé en premier argument et extrait les patrons inscrits dans le fichier qui est passé en tant que second argument de la commande.

Un extrait du fichier obtenu (fichier qui contient les suites extraites dans les discours d'un candidat) :

[24]cf.version HTML du Rapport pour téléchargement

```
on  ne  pouvoir  pas  gouverner
on  ne  pouvoir  plus  continuer
on  ne  pouvoir  pas  continuer
on  ne  pouvoir  pas  continuer
nous  ne  pouvoir  plus  accepter
on  ne  oser  plus  sortir
nous  ne  pouvoir  pas  continuer
nous  ne  pouvoir  pas  faire
je  ne  vouloir  pas  finir
on  ne  pouvoir  pas  créer
```

Fig.1 - Exemples de suites extraites

Le lemme du verbe prédicatif figure ici pour les mêmes raisons que dans la première structure. Un petit script bash[25] permet à partir de ce fichier d'obtenir trois fichiers ne contenant chacun que les patrons avec un pronom personnel donné (*je, on, nous*), comme par exemple (pour l'extrait du fichier ci-dessus) :

```
on  ne  pouvoir  pas  gouverner
on  ne  pouvoir  plus  continuer
on  ne  pouvoir  pas  continuer
on  ne  pouvoir  pas  continuer
on  ne  oser  plus  sortir
on  ne  pouvoir  pas  créer
```

```
nous  ne  pouvoir  plus  accepter
nous  ne  pouvoir  pas  continuer
nous  ne  pouvoir  pas  faire
```

```
je  ne  vouloir  pas  finir
```

Fig.2 - Exemples d'extraction avec le script bash

Chacun de ces fichiers est ensuite nettoyé par le script EnleveRubish.pl[26]. Ce script, basé sur la segmentation du fichier d'entrée par les espaces, enlève le pronom personnel, le morphème négatif et l'adverbe négatif (pas, plus, ...) pour qu'on obtienne un fichier au format suivant (par exemple pour le pronom personnel, dans l'exemple ci-dessous):

[25]cf.version HTML du Rapport pour téléchargement
[26]cf.version HTML du Rapport pour téléchargement

```
pouvoir gouverner
pouvoir continuer
pouvoir continuer
pouvoir continuer
oser sortir
pouvoir créer
```

Fig.3 - Exemples d'extraction nettoyée

Les fichiers ainsi prétraités peuvent être ensuite convertis en format Pajek, la procédure est identique à la première construction.

Nous obtenons ainsi trois graphes pour chaque candidat. Chaque graphe représente les patrons recensés pour un des trois pronoms personnels.

Les graphes doivent être lus de la façon suivante : un arc entre deux infinitifs représente une structure négative. Par exemple *vouloir* et *faire* reliés par un arc dans le premier graphe de Bayrou (le pronom personnel *je*) doit être lu *je ne veux (pas, plus, jamais, ...) faire* ou *je ne fais (pas, plus, jamais, ...) vouloir*, le choix de la bonne variante est clair dans la plupart des cas.

Bayrou

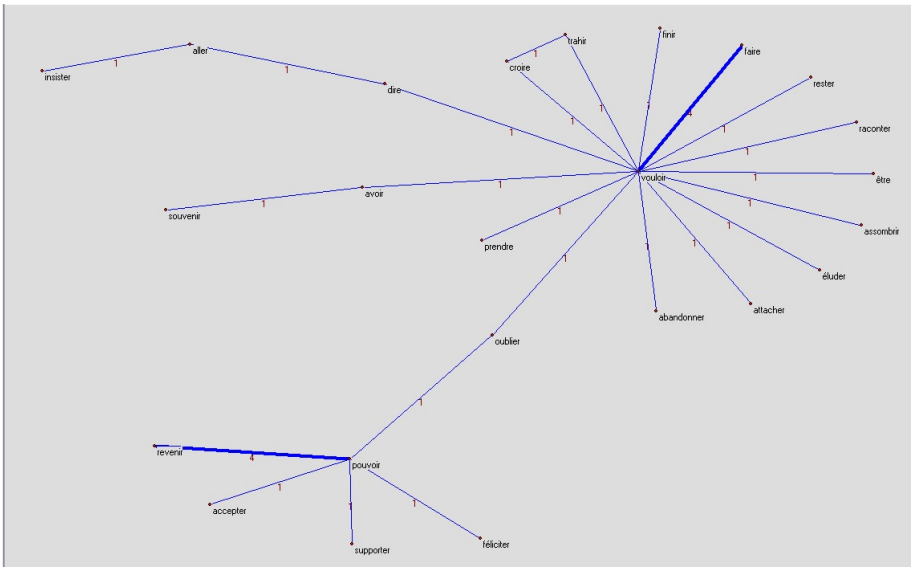

Fig.1 - Graphe 'je' Bayrou

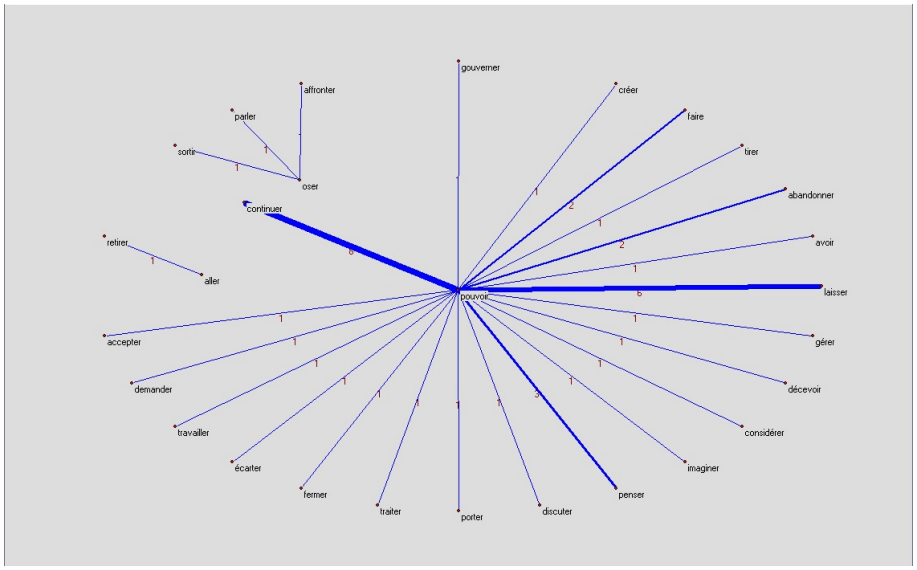

Fig.2 - Graphe 'on' Bayrou

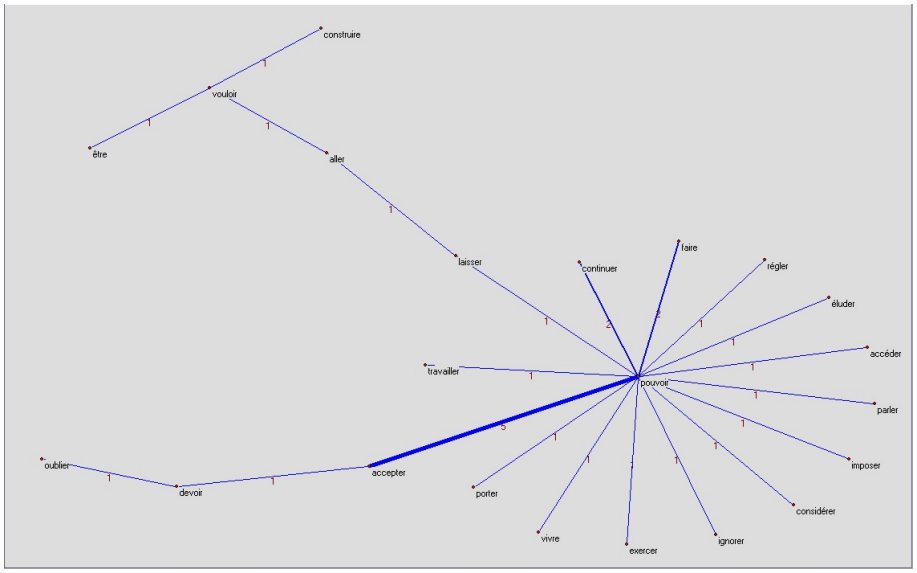

Fig.3 - Graphe 'nous' Bayrou

Le Pen

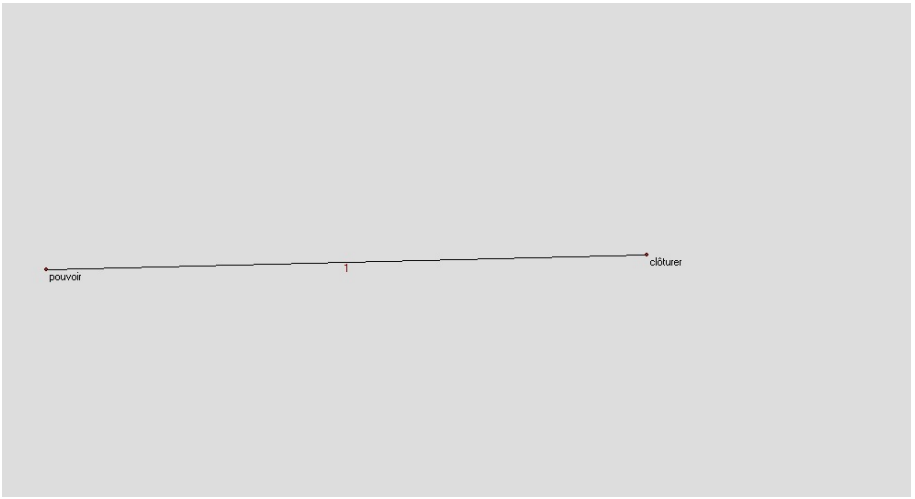

Fig.1 - Graphe 'je' Le Pen

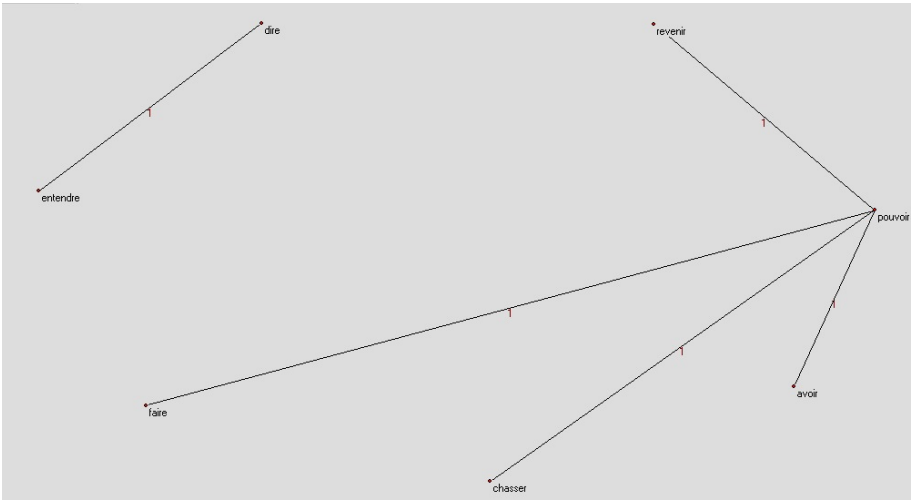

Fig.2 - Graphe 'on' Le Pen

Fig.3 - Graphe 'nous' Le Pen

Sarkozy

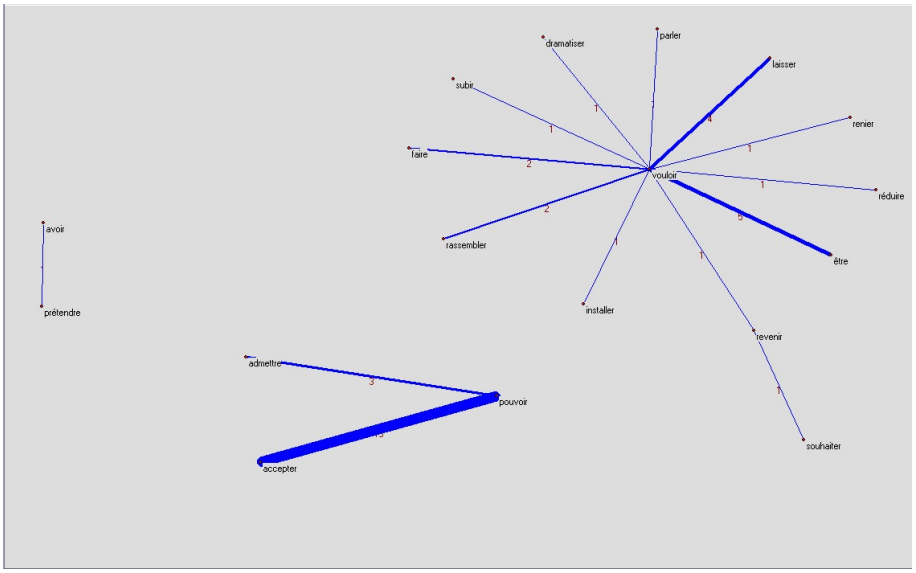

Fig.1 - Graphe 'je' Sarkozy

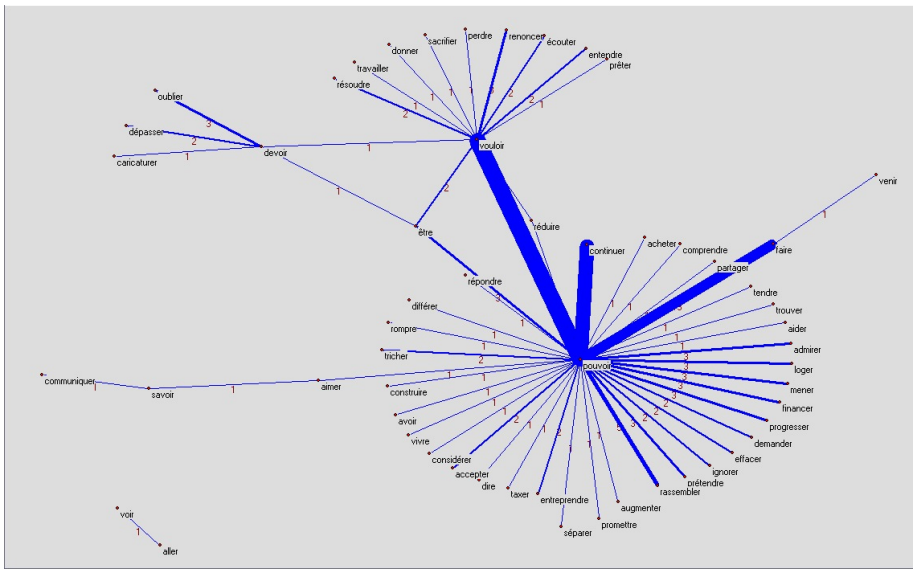

Fig.2 - Graphe 'on' Sarkozy

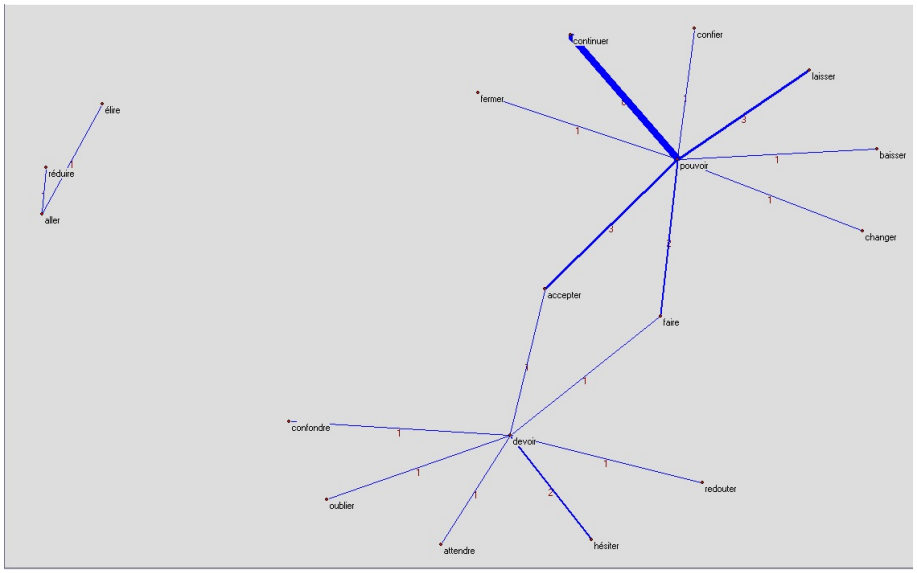

Fig.3 - Graphe 'nous' Sarkozy

Royal

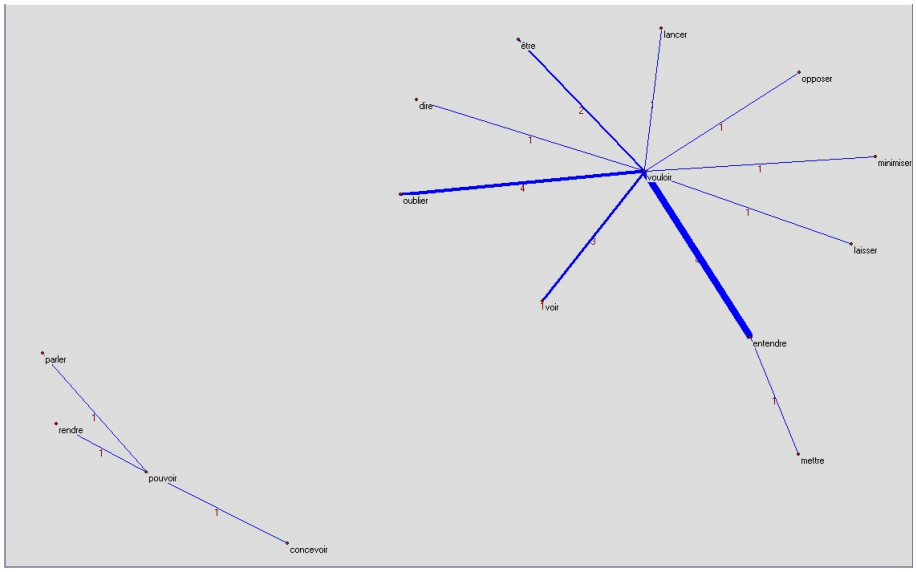

Fig.1 - Graphe 'je' Royal

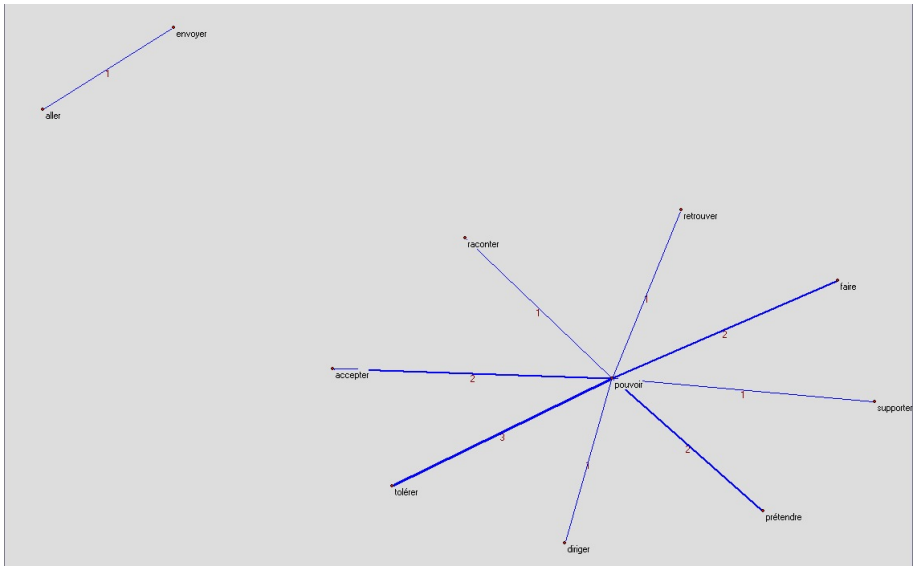

Fig.2 - Graphe 'on' Royal

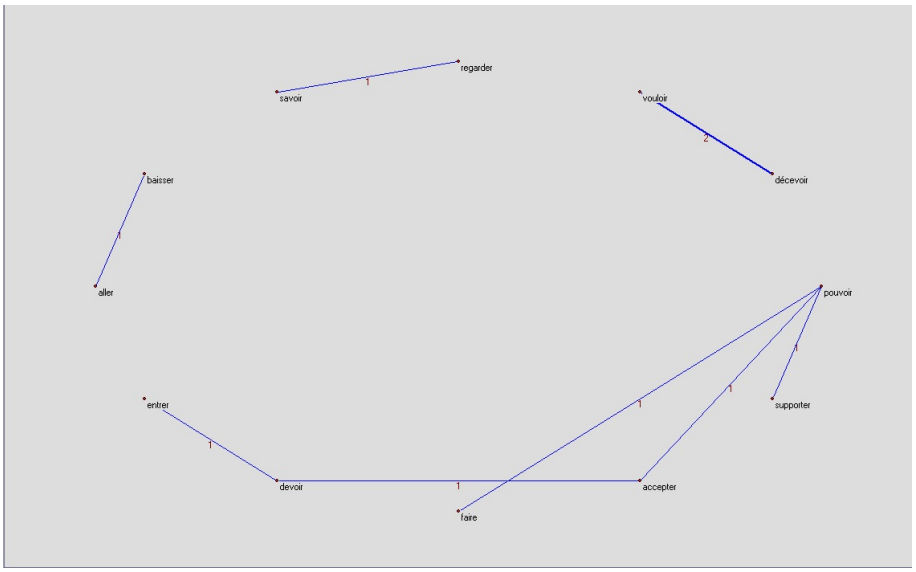

Fig.3 - Graphe 'nous' Royal

2.3 Application d'outils théoriques issus de l'Analyse Conversationnelle

Sur la base de différents axes d'étude du corpus Débat, le cadre théorique mobilisé[27] apparaît comme pertinent afin de fournir une explicitation de certains phénomènes identifiables, en particulier relatifs aux pratiques du *donner-à-voir* mises en oeuvre par les participants.

Le pôle institutionnel de la situation d'interaction induit une asymétrie parmi les membres impliqués. En effet, le statut participatif des interlocuteurs est, en quelque sorte, prédéfini par la forme du débat; on entend par là qu'il y a forcément une asymétrie thématisée dans la répartition des rôles interactifs des intervenants.

Est attendue d'une telle interaction une répartition différente des rôles des participants du fait de leur fonction dans le débat, qui permet de distinguer entre les *régulateurs* –Arlette Chabot et Patrick Poivre d'Arvor– et les *acteurs* –Ségolène Royal et Nicolas Sarkozy– du débat, ce qui est effectivement constatable en observant les données.

2.3.1 Le rôle social de *régulateur* du débat

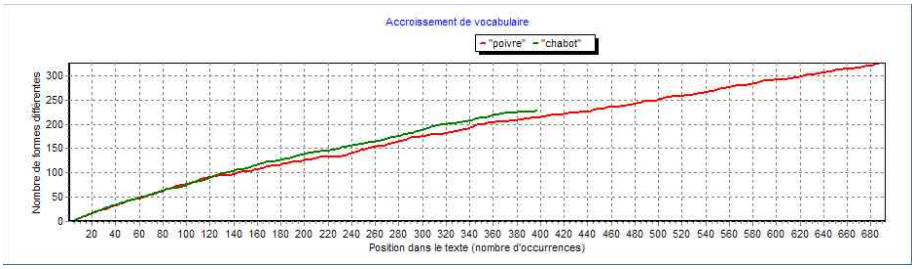

Fig.1 - Accroissement du vocabulaire, locuteurs *poivre* et *chabot*

Principales caractéristiques de la partition : auteur

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
"chabot"	397	228	155	12	vous
"poivre"	682	317	210	36	de
"royal"	11233	2120	1113	438	de
"sarkozy"	12711	2278	1222	539	de

Fig.2 - Principales caractéristiques lexicométriques par locuteur

[27] Nous vous enjoignons à une lecture préalable des Rappels Théoriques (Annexes - pp.75), afin de mieux cerner les notions et les emplois terminologiques convoqués dans cette section.

Accroissement du vocabulaire des régulateurs : indices d'une asymétrie de statut Dans la perspective d'une analyse interactive du débat entre les candidats des formations politiques représentées au second tour des élections présidentielles, les statistiques sur l'accroissement du vocabulaire des régulateurs du débat fournit des résultats on ne peut plus intéressants.

D'après les statistiques sur le vocabulaire des régulateurs illustrées en Figure 1, si Arlette Chabot et Patrick Poivre d'Arvor présentent un accroissement similaire au début du corpus, très rapidement une divergence apparaît.

Le vocabulaire de Chabot s'accroît sur une période d'à peine plus de la moitié du corpus, et s'arrête net, tandis que celui de Poivre ne cesse d'augmenter. Cela peut indiquer plusieurs aspects à prendre en compte du point de vue de leur rôle de régulation et/ou d'implication dans le débat.

En premier lieu, si l'on prend en compte les statistiques sur les formes les plus fréquentes par locuteur pour la tentative d'interprétation de ce graphique d'accroissement du vocabulaire, il faut souligner que Chabot produit avant tout des indexicaux d'interlocution tournés vers les participants au débat, réalisés par le pronom *vous*.

Il est également intéressant de remarquer que, du point de vue de ces statistiques sur les fréquences de formes, Poivre s'ajuste aux fréquences de production de Royal et Sarkozy, en réalisant davantage d'occurrences de la préposition *de*.

Par ailleurs, comme il a été suggéré, de tels résultats semblent devoir être interprétés en termes de différence d'implication des régulateurs dans le débat.

Ce qu'indiquent les courbes de résultats de la Figure 1, c'est d'abord que Chabot fait moins intrusion dans l'interaction qui se déroule entre Royal et Sarkozy, à l'inverse de Poivre, ce qui laisse à penser que le degré d'implication de ce dernier dans le débat est plus important. Cela suggère une asymétrie de statut, que viennent conforter les statistiques sur les nombres d'occurrences et de formes produites par Poivre, globalement deux fois supérieures à celles produites par Chabot.

Du point de vue du lexique observable dans les tours de parole des régulateurs[28], ceux de Chabot présentent des caractéristiques particulières.

En effet, Chabot produit davantage d'énoncés relatifs à la gestion temporelle du débat : le champ lexical exprimant la temporalité, plus précisément lié au minutage, est très présent dans ses productions.

Cet aspect lexical, mis en regard avec une observation partielle des conduites interactives données à voir dans le corpus *Débat*, permet de corroborer l'idée d'une asymétrie de statut, au sein de laquelle Chabot apparaît comme celle qui répartit la parole, en invitant par exemple les uns et les autres à se répondre.

Poivre, quant à lui, apparaît plutôt comme le régulateur thématique, topical, du débat, ce qui tend à donner une amorce d'explication tangible au fait constaté de sa participation plus importante au débat du point de vue bitos de la variété de vocabulaire.

[28] cf.Fichiers XML - Extraction des tours de parole par auteur

En croisant ces trois types de données extraites et calculées sur les corpus, il est possible de mettre en évidence une asymétrie entre Chabot et Poivre quant à ce qu'ils donnent à voir dans leur réalisation du rôle social de régulateur du débat.

Il est entendu qu'une analyse linéaire séquentielle du débat dans son intégralité permettrait une interprétation adéquate – en termes de catégorisation discursive des participants – de l'asymétrie évoquée ici, que nous nous bornons simplement à constater.

2.3.2 Stratégies interactives des acteurs du débat

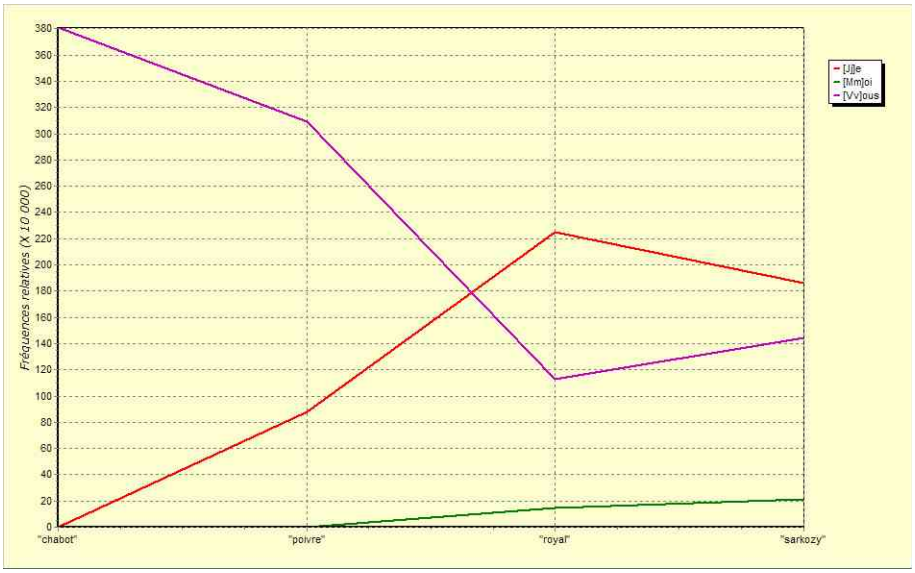

Fig.4 - Graphe de ventilation des formes [Jj]e — [Mm]oi — [Vv]ous, par auteur

Observations sur les indexicaux d'intersubjectivité : indices distinctifs de la gestion de l'interaction Pour le commentaire de ce graphique, on s'intéressera plus particulièrement aux statistiques relatives aux deux principaux acteurs du débat, Ségolène Royal et Nicolas Sarkozy.

Un premier constat concerne la fréquence de production de l'indexical d'interlocution *vous*, davantage employé par Sarkozy. Il faut d'emblée mettre en regard ce résultat avec les observations faites sur le mode d'adressage, partant la gestion de l'interaction, mise en oeuvre par Sarkozy.

En effet, ce dernier, contrairement à Royal, intègre les régulateurs du débat dans le cadre de participation de l'échange qu'il entretient avec son adversaire; il est possible de le constater en regardant de près ses tours de parole : il est le seul des débattants à interpeller les régulateurs du débat.

Cela permet de mettre au jour une différence importante dans les stratégies interactives respectivement adoptées par Royal et Sarkozy.

Tandis que Royal s'adresse préférentiellement – et ce, tout au long du débat – à son adversaire, donc à son interlocuteur direct, sans produire de vocatif

destiné à Chabot ou Poivre, Sarkozy n'hésite pas à les prendre à parti, comme l'illustre l'extrait ci-dessous.

```
§ L'Etat a transféré la compétence de la formation professionnelle aux régions. Entre temps, tenez-vous bien,
Monsieur Poivre d'Arvor, les effectifs de la formation professionnelle dans l'Etat ont augmenté de 60 %. L'Etat a
```

Fig.5 - Exemple d'interpellation de Poivre par Sarkozy

Il est par ailleurs possible de constater que c'est Sarkozy qui produit le plus d'indexicaux personnels, c'est à dire des *je* et *moi*, tout au long du débat.

En croisant ces trois types de données extraites et calculées sur les corpus, il est possible de mettre en évidence une asymétrie entre Royal et Sarkozy quant aux stratégies interactives et discursives qu'ils mettent en oeuvre au cours du débat.

Il est entendu qu'une analyse linéaire séquentielle du corpus dans son intégralité permettrait une interprétation adéquate – en termes de catégorisation discursive des participants – de l'asymétrie ici évoquée, que nous nous bornons simplement à constater.

Alternance de parole : éléments pour une analyse fonctionnelle Il a paru pertinent de s'essayer à une analyse fonctionnelle d'un certain nombre de phénomènes relatifs à l'alternance de parole entre les acteurs du débat, observés au cours de l'étude transversale du corpus *Débat*.

En effet, les phénomènes de chevauchements et d'interruption de parole, observables dans le discours peuvent être conçus comme relevant de l'intention du locuteur, partant, de son parti pris quant à la gestion de l'interaction, constituant à ce titre des éléments privilégiés pour aborder le *donner-à-voir* du statut de *débattant* exhibé par Ségolène Royal et Nicolas Sarkozy.

```
<auteur="sarkozy">
§ Aujourd'hui, c'est le plus faible taux de chômage en France depuis 25 ans. Je ne dis pas que l'on a tout réussi, mais c'est le taux le plus faible. Comment
faire pour aller plus loin? Ce que dit Mme Royal est intéressant. C'est une différence essentielle entre son projet et le mien. Au fond, elle est dans la
stricte logique socialiste du partage du temps de travail. Il y a un temps de travail qui est comme un gâteau, il faut le partager. Elle nous dit que
personne ne travaille plus de 35 heures, ainsi, cela obligera les autres à embaucher. Nulle part ailleurs dans le monde, on ne fait cela. Il n'y a pas un
seul pays, madame, socialiste ou pas, qui a retenu la logique du partage du temps de travail, qui est une erreur monumentale. Les 35 heures n'ont pas créé
d'emplois et ont été responsables de quelque chose de plus grave encore: la rigueur salariale, qui fait que nos salaires sont trop bas. Cela pèse sur le
pouvoir d'achat des Français. Le pouvoir d'achat en berne, c'est moins de croissance. Je propose une autre stratégie, la stratégie qui a marché partout. Vous
citez les démocraties du Nord de l'Europe, c'est ce qu'ils font. Vous avez votre ami Blair au Royaume-Uni, c'est ce qu'il a fait. Zapattero, c'est ce qu'il a
fait. Le travail des uns crée le travail des autres. C'est pourquoi l'Institut Rexecode, organisme indépendant, a noté le projet économique…

<auteur="royal">
§ On connaît la musique! C'est l'organisme du Medef. Vous le savez bien.

<auteur="sarkozy">
§ Savez-vous par qui il est dirigé? Par M. Michel Didier, l'un des économistes que M. Jospin avait nommé en 1998 dans son conseil des experts. L'organisme
que vous venez de contester, son président a été nommé par Lionel Jospin lui-même dans le conseil des experts qui entourait le Conseil des ministres d'alors.
Ce n'est pas gentil pour M. Jospin.

<auteur="royal">
§ Que cela a entraîné un point de croissance en moins. C'est bien l'organisme du Medef. Monsieur Juppé nous a servi cet argument régulièrement. Continuez.

<auteur="sarkozy">
§ Merci de m'y autoriser! Mon projet crée 230 000 emplois de plus.

<auteur="royal">
§ Merci Medef! Non, allez-y, continuez!
```

Fig.6 - Interruption de parole, exemple 1

Dans l'exemple rapporté en Fig. 6, Royal interrompt le tour de parole développé par Sarkozy, en mobilisant une stratégie d'autosélection. Cependant, il s'agit d'un phénomène d'arrogement de parole qualifiable d'abrupt, en ce qu'il ne repose pas sur un point de transition de parole possible. C'est donc un cas de rupture dans la circulation de la parole.

<auteur="sarkozy">
§ Que faites-vous des 35 heures, vous les gardez ?

<auteur="royal">
§ Je l'ai dit. La deuxième loi sur les 35 heures a été une loi trop rigide. Je suis capable de regarder les choses telles qu'elles sont, et la réalité des entreprises telle qu'elle est. J'ai cette responsabilité dans les régions. Nous gérons les aides économiques des entreprises. Tous les jours, je suis auprès des chefs d'entreprise. J'ai vu que la deuxième loi avait été trop rigide. J'ai dit dans mon pacte présidentiel que toute modification du Code de travail se ferait après une négociation entre les partenaires sociaux. Toute modification du Code du travail et pas en assénant un certain nombre de choses. Vous avez vu les réactions.

<auteur="sarkozy">
§ Que veut dire "trop rigide" ? Qu'allez-vous modifier?

<auteur="royal">
§ Les partenaires sociaux se mettront d'accord et discuteront branche par branche. S'il n'y a pas d'accord, il n'y aura pas de nouvelle loi.

<auteur="sarkozy">
§ Que changez-vous? Les 35 heures comme un minimum, je ne les toucherai pas. Je garde les 35 heures comme la durée hebdomadaire.

<auteur="royal">
§ Vous reconnaissez que c'est un progrès social économique important. Je vais vous donner un exemple. J'étais récemment dans une entreprise de haute technologie dans la Creuse qui produit des panneaux solaires. Elle est passée aux 32 heures et a augmenté sa productivité. Contrairement à ce que vous avez dit tout à l'heure, j'en ai parlé à Zapatero et à Prodi. Je suis allée en Suède voir la situation des entreprises. Dans bien des entreprises, quand elles sont performantes sur le plan technologique, elles ont même de durées de travail inférieures aux 35 heures.

<auteur="sarkozy">
§ A 32 heures, ils sont payés combien?

<auteur="royal">
§ Ils sont payés comme aux 35 heures.

<auteur="sarkozy">
§ On n'augmente pas le pouvoir d'achat.

<auteur="royal">
§ Si.

<auteur="sarkozy">
§ Or, il y a un problème considérable de pouvoir d'achat.

<auteur="royal">
§ Laissez les gens libres. Laissez la liberté aux gens. Ne leur imposez pas de travailler plus pour gagner plus. Vous savez ce que c'est que la valorisation du travail ? C'est un travail payé à sa juste valeur. Trouvez-vous qu'il est normal que des salariés commencent leur carrière au Smic à 980 euros nets par mois et terminent… Laissez-moi finir!

<auteur="sarkozy">
§ Qu'est-ce que vous changez dans les 35 heures? On n'y comprend rien.

<auteur="royal">
§ Si, si, vous avez parfaitement compris. Vous faites semblant de ne pas comprendre. J'observe que vous ne reviendrez pas sur les 35 heures. Vous ne les avez pas remises en cause. Elles ne sont pas responsables de tous les maux de la terre, comme le dit également le Medef. J'ai rencontré la présidente du Medef, elle m'a dit : "Revenez sur les 35 heures." Ce n'est pas sérieux! Il y a d'autres sujets sur lesquels discuter. Vous voulez qu'on annule les 35 heures? Elle m'a dit "non". Dont acte. C'est un acquis social important, mais cela a créé des difficultés dans des petites entreprisesparce qu'elles ont été appliquées de façon trop uniforme. Nous rediscuterons des 35 heures pour savoir si, oui ou non, et de quelle façon elles peuvent être généralisées et dans quelles branches. Ce seront les partenaires sociaux qui en discuteront. Je les ai déjà tous reçus les partenaires sociaux. Je n'attends pas d'être élue pour travailler. J'ai rencontré l'ensemble des organisations syndicales et des organisations patronales. Je leur ai dit que la réforme profonde du fonctionnement de la République demain sera une réforme de la démocratie sociale. Je souhaite qu'il y ait davantage de salariés qui adhèrent aux syndicats. Pourquoi ? Dans un pays comme le nôtre où nous n'avons que 8 % de salariés qui adhèrent aux syndicats, alors que dans les pays du Nord de l'Europe 80 % des salariés adhèrent à une organisation syndicale, à ce moment-là il y a un dialogue social constructif qui se crée, des compromis sociaux. Je termine.

<auteur="sarkozy">
§ Dans la fonction publique, autoriserez-vous l'octroi d'heures supplémentaires aujourd'hui interdites, oui ou non ?

Fig.7 - Combat pour la prise de parole, exemple 1

La Fig. 7 donne à voir la façon dont les interactants s'interpellent par des questions, non seulement quant à leurs projets respectifs sur des thèmes électoraux, mais également quant au contenu de leurs énoncés, notamment par des tournures du type : *Que veut dire "trop rigide"?*

Par ailleurs, il faut souligner ici un flagrant indice de combat pour la prise de parole, thématisé par Royal lorsqu'elle dit, par exemple : *Laissez-moi finir!*

De plus, un autre indice pertinent pour la mise en exergue du statut de débattant repérable dans le discours de Sarkozy est la demande d'explicitation à l'interlocuteur, qui contribue à sous-tendre la non intelligibilité de son discours; cela est explicite lorsque ce dernier dit à Royal : *Qu'est-ce que vous voulez changer dans les 35 heures? On n'y comprend rien.*

<auteur="sarkozy">
§ La politique moderne, c'est l'honnêteté !

<auteur="royal">
§ En effet !

<auteur="sarkozy">
§ Les pôles ont été créés en 2004, à la suite du rapport parlementaire remarquable de Christian Blanc. J'étais Ministre des finances, c'est alors que l'Etat les a créés. Cela ne pouvait pas être les régions, car les pôles sont des exonérations fiscales et sociales qui ne peuvent pas être décidées par les régions, mais par l'Etat…

<auteur="royal">
§ Mais qui ne sont pas arrivées…

<auteur="sarkozy">
§ Peu importe.

<auteur="royal">
§ Non, pas "peu importe" !

<auteur="sarkozy">
§ Soyons honnêtes, il y a des choses bien que vous avez faites, des choses bien que nous avons faites…

<auteur="royal">
§ Les exonérations fiscales sont arrivées ?

<auteur="sarkozy">
§ Oui.

<auteur="royal">
§ Ce n'est pas vrai ! Aucune exonération fiscale...

<auteur="sarkozy">
§ Ce n'est pas exact. Cela fonctionne, cela marche et c'est parce que cela marche que vous en parlez. Autrement, vous n'en parleriez pas.

<auteur="poivre">
§ Santé, logement, retraite, ce sont des sujets extrêmement précis.

Fig.8 - Combat pour la prise de parole, exemple 2

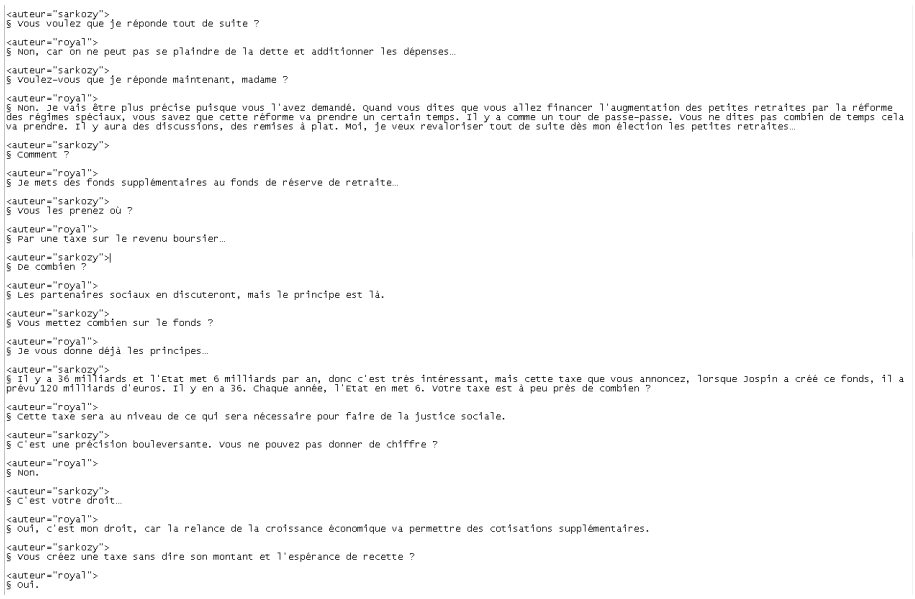

Fig.9 - Combat pour la prise de parole, exemple 3

Les Fig. 8 et 9 sont des exemples évocateurs du combat pour la prise de parole entre Royal et Sarkozy, en ce que les interruptions de parole fondées sur des chevauchements en fin de tour – indiqués dans la transcription par les points de suspension –, sont particulièrement fréquents.
C'est le lieu privilégié du donner-à-voir du rôle de débattant, en ce que les contradictions exprimées entre les interlocuteurs sont thématisées par des tournures exclamatives. Partant, cela est à mettre en corrélation avec le fait que les indices de l'indexicalité marquent une co-construction de l'interaction par les membres en tant que débat : les interactants s'interpellent, se répondent et s'exclament face aux arguments de l'autre, dans le but d'une démonstration des faiblesses des arguments exposés par l'adversaire.

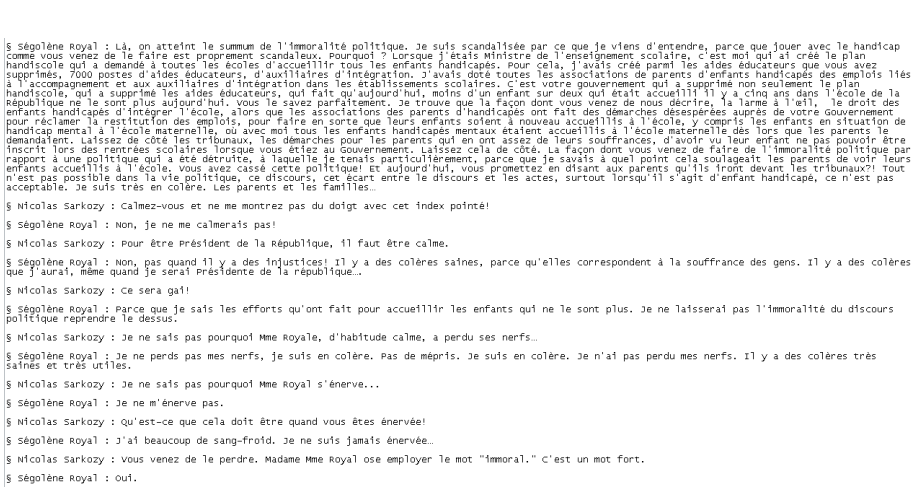

Fig.10 - Interruption de parole, exemple 2

Enfin, la Fig. 10 est un autre exemple d'interruption de parole abrupte. En effet, Sarkozy, alors que Royal développe un tour de parole, s'autosélectionne, ce faisant, s'arroge la parole en commentant la gestuelle de son adversaire : *ne me montrez pas du doigt avec cet index pointé!*, prétexte à une assertion qui tendrait à mettre en exergue la non capacité de Royal à accéder au statut de Président de la République, puisque *Pour être Président de la République, il faut être calme.*

L'intérêt d'une telle séquence, loin de chercher à évaluer le poids des stratégies rhétoriques employées par les interactants, est avant tout la mise en valeur le rôle social de débattant, en ce que par de telles productions, les interactants se catégorisent ainsi.
De tels exemples, fondés sur une étude transversale du corpus, sont manifestes du point de vue la catégorisation en termes de rôle social de *débattant* par les participants Royal et Sarkozy.

Exemple d'un groupe de formes : la fréquence des adverbes comme indice de la complexité du discours Nous nous focalisons ici sur la fréquence d'emploi des adverbes chez les acteurs du débat.

Fig.11 - Interruption de parole, exemple 2

Ce graphe montre une fréquence d'emploi des adverbes nettement plus importante chez Royal que chez Sarkozy. A priori, un tel résultat est difficilement interprétable.

Cependant, sa mise en regard avec les résultats concernant les calculs de fréquence de longueur des phrases lui donne une pertinence quant à l'évaluation de la complexité du discours produit par les interactants.

En effet, ces indicateurs, mis en corrélation, sont l'indice d'un choix discursif, et partant, interactionnel, distinctif du mode d'adressage choisi par Royal et Sarkozy.

De fait, bien que ce dernier formule le plus grand nombre d'occurrences dans le corpus, parallèlement, ses phrases sont plus courtes et il emploie moins

d'adverbes.
Ce qu'il est possible de dire d'après de tels résultats est que les choix discursifs de Nicolas Sarkozy tendent vers moins de complexité que ceux de Ségolène Royal. En somme, les pratiques du *donner-à-voir* mises en oeuvre par les participants

coïcident avec les prérequis d'une telle situation d'interaction : c'est ce qu'il a été tenté de montrer ici.

En effet, tant du point de vue des régulateurs que des acteurs du débat, est donnée à voir une répartition des rôles sociaux dont l'asymétricité peut être conçue comme caractéristique, ce que les traitements statistiques et l'étude transversale du corpus Débat ont contribué à corroborer.

Conclusion

Les trois volets d'étude appliqués aux corpus dans le cadre de ce projet ont permis, outre l'application de divers outils d'analyse, de mettre en valeur les visées applicatives issues de la linguistique quant à l'exploration des données textuelles.

Les résultats obtenus sont légitimés par la mobilisation de traitements robustes des corpus, ainsi que par l'application de fonds théoriques rigoureux.

Nous espérons avoir montré que, loin de se contredire, les traitement robustes et théoriques mis en oeuvre donnent à voir une complémentarité analytique pertinente.

Annexes

[10pt,a4paper]article [latin1]inputenc amsmath amsfonts amssymb Marguerite Leenhardt Rappels théoriques pour l'interprétation des données interactionnelles

2.4 Fonds théorique mobilisé

2.4.1 Apports de la linguistique interactionnelle : le TTS (Turn Taking System, Sacks, Schegloff, Jefferson, 1974)

L'analyse conversationnelle : rappels

L'analyse conversationnelle peut être définie comme l'étude détaillée des méthodes utilisées par les membres pour l'accomplissement interactionnel des pratiques sociales dans plusieurs types de contextes, et entend mettre au point un système d'analyse formelle de la conversation.

Du point de vue de l'analyse conversationnelle, la notion de conversation ne renvoie pas à un genre discursif, mais à la parole en interaction en tant qu'activité sociale ordinaire. Cela permet de considérer comme relevant de la parole en interaction tous les types d'échanges possibles, tels la téléconférence, le discours institutionnel, par exemple. Dans ce cadre théorique, les pratiques sociales ne peuvent exister en dehors de la communauté : elles sont accomplies par les membres, par le biais du langage. Une telle approche doit être qualifiée de praxiologique .

Sacks formule un certain nombre de critères structurels de l'interaction, notamment la structuration des échanges conversationnels en paires adjacentes. Les paires adjacentes sont des objets sociaux , découlant de l'alternance de parole entre les interlocuteurs, rendant possible l'accomplissement d'activités sociales, -telles que salutation/salutation, question/réponse, par exemple-, et qu'il est possible de définir en tant que suite connexe de deux énoncés proférés par deux locuteurs différents, énoncés qui entretiennent une relation de pertinence conditionnelle, i.e. culturellement pertinente, ce qu'explicite Schegloff en écrivant Given the first, the second is expectable.. Ces unités les paires adjacentes apparaissent d'une façon systématique et donnent à la conversation son caractère de phénomène naturel descriptible et observable, la constituant comme une entrée en matière pour étudier l'ordre social. Par ailleurs, les éléments constitutifs de la paire adjacente entretiennent une relation de pertinence conditionnelle, ce qu'explicite Schegloff en écrivant Given the first, the second is expectable., relation culturellement pertinente.

La mise au jour de tels critères constitue une révolution quant aux linguistiques de l'oral, en ce qu'ils sont des indices d'organisation sociale observables en interaction.

Il convient de rappeler que le tour de parole est une unité constamment en

construction : tant qu'autrui ne prend pas la parole, le locuteur en action a la possibilité de développer son tour. Il est par ailleurs possible de cerner les unités de construction du tour de parole en étudiant les moments de transition de parole possibles, qui en permettent une segmentation : ce sont des unités prosodiquement déterminées.

L'analyse conversationnelle envisage le tour de parole comme l'élément constitutif de son objet d'étude. Le tour de parole est une unité non linguistique, qui correspond au temps pendant lequel un locuteur garde la parole ou accomplit une action. Il s'agit donc d'une unité temporelle relevant du verbal et du non verbal. En effet, la longueur d'un tour de parole n'est pas prédictible, puisqu'elle est à chaque fois négociée par les participants : sa durée dépend du temps pendant lequel les participants laissent parler le locuteur en action.

2. Le TTS (Turn Taking System)

Le TTS est le système d'analyse formelle des tours de parole, élaboré par Sacks, Schegloff et Jefferson, qui comprend 13 points :

1. alternance des tours de parole, manifestée par un changement de participants
2. le plus souvent, chacun parle à son tour
3. parfois, même communément, deux locuteurs parlent en même temps, mais pendant des laps de temps réduits de façon générale : on parle de chevauchement de parole; il faut souligner qu'un tel phénomène n'a pas forcément de caractère conflictuel, car un chevauchement peut être une marque d'intersubjectivité, d'accord
4. la circulation de la parole d'un locuteur à l'autre sans silence ou sans chevauchement est fréquente
5. l'ordre des tours de parole, i.e. l'ordre dans lequel les locuteurs contribuent à l'interaction, n'est pas fixé à l'avance et varie constamment
6. la taille des tours de parole n'est pas fixée
7. la longueur de la conversation n'est pas fixée
8. les productions, du point de vue des topics ou thèmes abordés, ne sont pas fixées
9. la distribution des tours de parole n'est pas fixée
10. le nombre de locuteurs peut varier
11. la parole peut être continue ou discontinue
12. des techniques d'allocation des tours de parole sont identifiables, en tant que techniques d'attribution de la parole mises en oeuvre par un participant pour devenir le locuteur en action 13) les unités de construction des tours de parole sont de nature variée, verbales ou non verbales

L'analyse conversationnelle détermine deux pôles dans le continuum desquels il est possible de situer le contexte, émergeant au fil de l'interaction. D'une part un pôle ordinaire, au sein duquel le statut des participants est équivalent, symétrique. D'autre part un pôle institutionnel, caractérisé par l'asymétrie de statut des participants. Ces pôles sont à concevoir moins comme des unités discrètes que comme des unités continues, car peuvent se présenter des difficultés dans la détermination absolue de l'ordinarité ou de l'institutionnalité des contextes.

L'analyse formelle des tours de parole permet une description à l'aide d'unité plus larges, dénommées séquences, qui sont le squelette de l'interaction. Il est ainsi possible de distinguer des séquences d'ouverture et de clôture, ainsi que d'autres types de séquences constitutives du corps de l'interaction.

2.5 Les processus de contextualisation verbaux : une analyse des choix linguistiques (Gumperz, 1999)

La sociolinguistique interactionnelle s'est préoccupée d'intégrer les dimensions pragmatiques et interactionnelles dans l'analyse des faits de variation sociale. Elle s'attache donc à décrire la signification pragmatique des variables en analysant la manière dont elles contribuent à l'interprétation des énoncés dans l'échange conversationnel, se centrant ainsi sur la dynamique des échanges verbaux, en particulier au sens contextuel des expressions verbales.

Gumperz met en avant la notion de choix des participants dans la situation de communication, en ce qu'ils ont le choix entre différentes possibilités de réalisations linguistiques, qui sont fonction du contexte, mais aussi fonction des présupposés culturels et de l'expression personnelle des interlocuteurs. Partant, le choix du locuteur est envisagé en tant que choix social transmis aux auditeurs, qui, à leur tour, choisissent dans leur inventaire linguistique. Il en découle que c'est en fonction de l'objectif communicationnel qu'est effectué un regroupement de variables donné. Gumperz définit la notion de contextualisation en tant que c'est le procédé par lequel nous évaluons les sens du message et les structures séquentielles de la conversation relativement à certains aspects de la structure superficielle du message, appellés 'indices de contextualisation'.

Pour Gumperz, la situation d'interaction renvoie à des codes sociaux qui règlent l'échange et permettent une interprétation des actes de parole. En effet, les codes sociaux mobilisés mettent l'accent sur des traits conceptuels particuliers connaissances partagées, croyances, intentions, présupposés dont le fondement peut être social ou culturel.

Le travail d'analyse, du point de vue de la contextualisation, consiste à retrouver dans le texte produit par les interactants des traces de leur compétence sociale d'interprétation, qui doivent être co-interprétées pour identifier une construction interactive du contexte. Dans le texte résultant de l'activité interactive, l'analyse doit mettre en évidence les traces laissées par cette co-interprétation.

2.6 La catégorisation discursive (Goffman, 1981, Mondada, 1999)

Les catégories sociales données à voir dans l'interaction par les productions verbales des interactants ne lui préexistent pas : elles sont produites en discours. De fait, il est possible de distinguer des collections de catégories sexe, nationalité, ethnie, langue, par exemple -. Un locuteur peut être catégorisé en mobilisant plusieurs collections, sa description est donc potentiellement infinie. Cependant, une seule catégorie suffit généralement à sa définition, en ce qu'elle est la catégorie la plus pertinente quant à la situation de communication : c'est la règle de l'economie, qui met en évidence le fait que la catégorisation ne répond pas à l'exigence de donner une définition référentiellement exacte, mais à celle d'offrir une description pertinente pour l'activité en cours et le contexte dans lequel elle s'inscrit. La règle de la cohérence peut alors être mobilisée, en ce que les autres membres de l'interaction peuvent être catégorisés en recourrant aux catégories de la collection alors introduite.

Un certain nombre de notions issues de ces principes complètent l'analyse catégorielle du discours.

En premier lieu, la notion d'appariemment ou d'inférence. En effet, sur les catégories se fondent de nombreuses inférences, car elles constituent une sorte d'archive des connaissances que les membres ont de la société à laquelle ils appartiennent, s'inscrivant dans un ensemble de représentations partagées. Il convient de préciser que la négation de la notion d'inférence peut être mobilisée en contexte : en effet, un locuteur peut discursivement mettre en valeur une catégorie, un comportement social contradictoire avec l'ensemble d'inférences légitimes quant aux représentations culturellement partagées.

En second lieu, la notion de paires standardisées. De ce point de vue, il est entendu que la catégorisation d'un des membres entraîne la catégorisation de son interlocuteur. Pensons par exemple aux paires parent/enfant, médecin/patient.

Enfin, la notion de boucle métaénonciative. L'on parle de boucles métaénonciatives pour mettre en avant le fait que les catégories sont travaillées, ajustées en contexte. Cela est permis par les tours de parole : la catégorie peut se définir localement.

Il s'agit en somme d'identifier vers quelles catégories s'orientent les interactants afin d'accomplir ensemble l'organisation de la conversation. Il faut préciser qu'un même individu, en fonction de l'interaction, peut donner à voir des rôles sociaux différents. Dans des types d'activités et des contextes d'interaction différents, les interlocuteurs peuvent construire la pertinence des catégories de façon variable, allant de pair avec une gamme de comportements diversifiés.

Le problème de la catégorisation des interactants est avant tout un problème pratique qui s'impose à eux dans le cadre de l'organisation de l'interaction. L'on dira alors que les participants s'orientent vers des catégories pertinentes en con-

texte qui garantissent l'ordre de l'interaction. Du point de vue de l'analyste, il s'agit de savoir comment les interactants sélectionnent les catégories pertinentes dans le cours de leur activité. En effet, les catégories pertinentes ne sont pas toujours explicitement thématisées dans l'interaction. Cependant, la catégorisation des locuteurs peut être liée à l'activité en cours, et ils peuvent passer d'une catégorie à une autre au cours de l'interaction, les pertinences étant localement définies, comme on l'a évoqué plus haut.