2. Une méthode : la logométrie

La méthode utilisée dans notre recherche depuis 12 ans souffre depuis ses origines d'un défaut d'état-civil. Encore aujourd'hui, l'incertitude planant sur son nom est le principal handicap à sa diffusion. Or, loin de clarifier la situation, sans doute avons-nous contribué à la confusion terminologique existante en proposant en lieu et place des termes déjà sur le marché scientifique (*linguistique quantitative, lexicologique quantitative, lexicométrie, textométrie, grammaticométrie, statistique lexicale ou statistique textuelle, analyse des données textuelles* ou *analyse statistique des données textuelles*, etc.) le terme « logométrie » {**11** ; **16** ; **18** puis systématiquement ensuite[26]}. Nous reprendrons ici ce vocable, non par souci d'originalité, mais parce qu'il nous semble le plus ouvert par son premier formant (logos) et le plus précis par son deuxième formant (métrie). *Prendre la mesure du discours* (politique) par tous les moyens possibles, quantitatifs et qualitatifs, et dans toutes ses dimensions linguistiques, sans présager des dimensions utiles et celles impertinentes, tel est l'objectif non borné de notre travail[27].

La méthode utilisée avec assiduité depuis 12 ans souffre surtout, dans l'usage vulgaire, d'une réduction voire d'une perversion de ses objectifs. Assignée seulement à administrer la preuve chiffrée grâce au décompte des mots, elle manque l'essentiel, se met sous une saine critique de scientisme et se condamne à n'être qu'un gadget convoqué par les littéraires, les historiens ou les politologues en mal de critères objectifs (*i.e.* quantifiés) pour sceller leur démonstration.

La logométrie que nous avons essayé de promouvoir est autre chose : c'est selon une expression martiale utilisée dans {**32**}, le bras armé de cette herméneutique numérique appelée modestement de nos vœux dès 2002 {**11**}[28], qui voit aujourd'hui le

[26] Dans le détail en 2002 {**11**} a été essayé « logomatique ». En 2004, c'est l'adjectif « logométrique » qui apparaît dans le titre de {**16**} et « logométrie » qui est utilisé au cœur de {**II**}. En 2005 {**18**} nous titrons « De la lexicométrie à la logométrie ».

[27] En anticipant sur le développement à venir, rappelons la définition de la logométrie donnée dans {**18**} : « Ce que nous appelons Logométrie, c'est un ensemble de traitements documentaires et statistiques du texte qui ne s'interdit rien pour tout s'autoriser ; qui dépasse le traitement des formes graphiques sans les exclure ou les oublier ; qui analyse les lemmes ou les structures grammaticales sans délaisser le texte natif auquel nous sommes toujours renvoyés. C'est finalement un traitement automatique global du texte dans toutes ses dimensions : graphiques, lemmatisées, grammaticalisées. L'analyse ainsi portera sur toutes les unités linguistiques de la lettre aux isotopies, en passant par les n-grams, les mots, les lemmes, les codes grammaticaux, les bi-codes ou les enchaînements syntaxiques. ».

[28] L'expression *herméneutique numérique* a été utilisée en 2002 dans le titre de l'article {**11**}. Notons néanmoins qu'avec la naïveté qui convient aux travaux anté-doctoraux, nos deux premiers articles dès 1997 posent dans leur introduction {**1** : 159} et {**2** : 9} la question *herméneutique* ; et nous espérions dans la conclusion de notre thèse en 1998 avoir établi une *herméneutique rigoureuse* {**I** : 750}. Cette posture herméneutique qui se renforcera ensuite, nous l'avons dit, à la lecture de linguistes comme [Rastier 2001], [Viprey 2005] ou [Adam 2008], tient initialement beaucoup aux écrits de Jacques Guilhaumou qui en France a constitué « *l'AD en discipline interprétative à part entière* » [Guilhaumou 2006 : 12 et 25] ; sur le *tournant herméneutique* décisif (quoique remis en cause) de l'Analyse du discours à la française, lire particulièrement [Guilhaumou 1993].

jour sous des formes variées, et apparaît comme l'enjeu des décennies futures pour les sciences du texte. A ce titre, l'objectif premier de la logométrie n'est pas d'administrer la preuve mais de mettre en place un protocole de lecture pour baliser, au sein de corpus tels que définis plus haut, des parcours interprétatifs contrôlables. La logométrie est un mode de compréhension générale du corpus alliant descriptions macro susceptibles par exemple de produire, globalement, des typologies textuelles {**I** : 743 ; **II** : 45, 91, 115, 125, 202, 204, 231 ; **14** ; **16** ; etc.} et descriptions micro susceptibles de traiter, localement, de phénomènes linguistiques pointus du texte (par exemple l'usage des adverbes et de « naturellement » chez Chirac {**II** : 106-109, 160-170}). C'est une approche intégrale du texte alliant, comme indiqué *supra*, posture paradigmatique (création de dictionnaires, d'index, de tableaux de *sélection*) et posture syntagmatique (création de concordanciers, étude de motifs syntaxiques, prise en considération des *combinaisons*). C'est un effort systématique de déconstruction du texte qui permet de traiter statistiquement les régularités et les phénomènes linguistiques saillants, combiné à un effort de (re)contextualisation – jusqu'au retour au texte intégral – pour en construire le sens. Finalement, c'est une lecture révolutionnaire mais non destructrice qui cherche à adjoindre à la lecture *naturelle, linéaire, qualitative, traditionnelle* du texte – si ce n'est que cette lecture se passe sur écran et non sur papyrus, sur parchemin ou sur papier–, une lecture *hypertextuelle, quantitative, tabulaire, réticulaire* que seul autorise le numérique {**24**}. En un jeu de mots douteux proposé dans {**33**}, c'est une lecture totale car *alpha-numérique* du corpus usant dans un même mouvement de l'*alphabet* et du *nombre*, des chiffres et des mots.

2.1. Des lexies (lexi) au discours (logo*) : retour sur les unités textuelles*

« Logométrie » s'est imposée à nous *versus* la lexicométrie traditionnelle en poussant l'étude des formes graphiques (lexies* simples) ou même du lexique (lexico*) vers des réalités linguistiques d'ordre grammatical, syntaxique, sémantique ou encore rhétorique qui composent, ensemble, le discours (ou logos*).

Ce mouvement, en lui-même, n'est guère original[29] et ne mériterait pas d'être mentionné s'il ne constituait, au quotidien, la principale mutation de nos pratiques notamment de notre premier ouvrage issu du doctorat {**I**} à notre second sur le discours présidentiel {**II**}; dans cette mutation l'influence de Sylvie Mellet rompue aux corpus latins lemmatisés / étiquetés du Lasla a été déterminante [voir par exemple Mellet et Purnelle 2002], ainsi que celle d'Etienne Brunet dont le logiciel s'articule depuis 1998 aisément sur plusieurs lemmatiseurs (Winbrill, Tree Tagger, Cordial) ; et cette mutation symbolise grossièrement notre trajectoire interdisciplinaire d'abord centrée sur les sciences historiques et par là peu soucieuse de la qualité du matériau linguistique traité (le mot), puis, à partir de notre recrutement par la section 34 et notre intégration à l'UMR, *Bases, Corpus et Langage,* centrée sur les sciences du langage[30].

[29] puisque [Muller 1963] renonçait déjà au texte brut au début des années 60.

[30] Rappelons en effet à cette occasion qu'après un cursus universitaire complet en Histoire, notre doctorat {**I**} a été encadré par un historien (R. Schor) et soutenu devant un jury exclusivement –à l'exception d'E. Brunet– composé d'historiens (S. Berstein, E. Brunet, J. Guilhaumou, G. Pervillé, D. Peschanski, R. Schor). Si notre thèse a su convaincre, ensuite, le jury de recrutement de la section 34 du CNRS, et la collection *Lettres numériques* de Champion, elle apparaît au détour de certaines phrases bien innocente linguistiquement. On lira par exemple avec le recul nécessaire la section III de l'*Introduction*

(. . .)

L'enjeu des analyses logométriques, simple à exprimer, difficile à mettre en pratique, apparaît dans le quotidien de notre recherche : articuler traitement quantitatif et traitement qualitatif, vision macroscopique du corpus et vision microscopique, dénombrement global – qui signifie décontextualisation et appréhension paradigmatique – et lecture locale – qui signifie recontextualisation et appréhension syntagmatique. C'est dans le va-et-vient critique entre un index de fréquences, une liste de spécificités, un tableau chiffré de distances intertextuelles d'un côté, et le texte de l'autre dans sa chaîne entière (texte intégral), comme dans ses extraits (concordances, fenêtres de contextualisation fixes ou coulissantes, etc.) que réside la force heuristique de la logométrie.

L'importance – pour ne pas dire le primat – du traitement quantitatif n'est pas une obsession scientiste visant à objectiver autant que possible le « signifiant » du corpus ou les « données » textuelles. Elle se trouve justifiée par la théorie rastiérienne de la détermination du local par le global [Rastier 2001]. En tant que globalités qui informent ses parties, les corpus textuels traités, par leur taille même, peuvent difficilement se passer d'un traitement macroscopique d'ordre quantitatif ; sauf à prétendre rendre compte des tendances profondes de masses de plusieurs centaines de textes et plusieurs millions de mots par de simples impressions. Plus simplement, nous l'avons dit, en tant que collections de textes ou que *séries*, les corpus ont tout à gagner d'une approche statistique à même de révéler les régularités / irrégularités linguistiques qui les animent à côté des accidents ou des épiphénomènes.

Pourtant, contrairement aux traitements TAL, les dénombrements proposés et, plus généralement, le traitement informatique mis en œuvre n'entendent pas se substituer à l'acte traditionnel de lecture : si le processus interprétatif est outillé par la statistique, il ne peut se réaliser sans le retour systématique au texte et une intimité avec l'œuvre. (Précisons seulement que ce retour au texte et cette intimité avec l'œuvre sont eux-mêmes favorisés par l'informatique et les outils de navigation, de lecture ou de recherche hypertextuelles).

Dès lors, la logométrie apparaît comme une *discipline* non pas au sens académique de champ de connaissance ou de branche d'activité, mais en tant qu'ensemble de règles à suivre garantissant la rigueur des résultats obtenus. Sa pratique, au-delà du va-et-vient fondamental mentionné entre la lettre et le chiffre ou la présence constatée et la distribution calculée, demande de la part du chercheur une posture adéquate face au corpus et de la part des outils mobilisés des fonctionnalités et une ergonomie *ad hoc*.

La posture, d'abord, allie deux principes complémentaires : la centralité du texte et le décentrement du lecteur ; ou, en termes généraux, la centralité de l'objet et le décentrement du sujet. Si le texte dans l'évidence de sa matérialité est central – au commencement et à la fin du traitement–, sa lecture ne saurait être im-médiate. La lecture naturelle du texte se trouve encadrée ou médiatisée par la statistique et l'informatique afin de retarder l'entrée dans la subjectivité interprétative {idée exprimée dès 2, puis régulièrement ensuite}. L'affirmation est triviale : dans les sciences, l'objet (ici le corpus textuel) ne peut être appréhendé par le sujet (ici le lecteur-expert) sans une mise à distance, c'est-à-dire, concrètement, sans la médiation d'un outil (ici les logiciels de logométrie, ailleurs le microscope ou la lunette). En d'autres termes, la proximité vantée entre le chercheur et l'œuvre ne saurait être confondue avec une promiscuité ; pas plus que l'outillage qui sert à mettre à distance l'objet n'est pour nous une finalité

qui nous déracinerait du corpus. Et précisons avec force que ce décentrement du sujet par rapport à son objet, nécessaire partout, nous semble particulièrement indispensable en matière textuelle, tant le texte donne l'impression trompeuse d'être naturel, et le sens l'illusion d'être immédiat.

L'outil, ensuite, en tant que froid prolongement de la main et du cerveau ou en tant que médiateur, prend, dans ces conditions, un rôle non négligeable. Aussi avons-nous toujours illustré nos démonstrations par les fonctionnalités des logiciels utilisés et plus particulièrement par l'ergonomie d'Hyperbase que nous espérons avoir contribué à améliorer. Toute l'ergonomie d'Hyperbase, de l'historique page d'accueil aux fonctions les plus récentes de l'hiver 2008-2009, est tendue vers l'organisation de la double lecture logométrique : lecture quantitative (« fonctions statistiques ») et lecture qualitative (« fonctions documentaires ») (illustration 4)[46].

Illustration 4 : Page d'accueil d'HYPERBASE (base « Republic »)

Dans le détail du logiciel, chaque outil statistique permet par simple clic de retourner aux textes pour engager une lecture traditionnelle, de même que la lecture traditionnelle du texte plein est outillée par des liens nous renvoyant au dictionnaire de fréquences, à la liste des spécificités, etc. Théoriciens et praticiens du domaine s'accordent pour penser que c'est dans cette mise en scène logicielle, d'apparence technique, que se joue l'essentiel. Des outils logométriques pointus statistiquement mais qui gêneraient le retour au texte amèneraient le parcours interprétatif dans des impasses ; pire encore, favoriserait les sur-interprétations (d'une AFC par exemple)

[46] A côté de ce premier choix ergonomique fondamental (la séparation des fonctions documentaires et des fonctions statistiques), rappelons le second choix ergonomique important : la juxtaposition/comparaison du texte brut et du texte lemmatisé/étiqueté.

voire les mésinterprétations (d'une arborée par exemple). Inversement, ces mêmes logiciels, s'ils n'offrent que des possibilités de recherches documentaires et de lecture, sans instrumentation statistique, seraient aussi inopérants qu'une loupe pour observer l'univers. Sur de gros corpus, les recherches documentaires focalisées en effet sont parfois trompeuses et contreproductives : attestation (locale) n'est pas raison (globale) ; et présence ne veut pas dire significativité. Bref, si l'on veut bien considérer avec [Bachelard 1934 : *Première partie, introduction*] que l'outil est toujours et seulement une *théorie matérialisée*, c'est sans doute par ses outils que le logométrie traduit le mieux sa différence théorique avec la linguistique textuelle traditionnelle d'un côté et le TAL de l'autre. Et pour cette raison, sans avoir de compétences informatiques propres, nous ne sommes pas et ne deviendrons pas insensible aux développements informatiques qui sont menés dans l'UMR 6039 (refonte et réécriture d'Hyperbase) et au sein de projets tel l'ANR Textométrie 2007-2010 (création d'une plateforme textométrique accessible en ligne).

2.3. Administration de la preuve et contrôle de l'interprétation

Lorsqu'elles ne sont pas supercheries, les pratiques statistiques qui prétendent être inférentielles ou probatoires sont des pratiques dangereuses en matière textuelle. Plus généralement, au-delà de notre domaine, de la statistique lexicale et des inférences aveugles et automatiques (attribution d'auteur, arbitrage judiciaire de lettres anonymes, etc.) qu'elle peut parfois entraîner, les méthodes les plus rigoureuses doivent renoncer, dans les Sciences humaines et sociales, à vouloir administrer la preuve là où elles nourrissent seulement la démonstration.

Pour cette raison, nous avons plusieurs fois répété {dès **1** , **2** et **3**, puis **11**, **15**, **18**, **21**, **26**, **33**} que la logométrie a une valeur heuristique plus que probatoire et que son objectif est de contrôler l'interprétation plus que d'objectiver ou d'inférer « le » sens.

Si nous avons parfois insisté, au début de notre recherche, sur la précision descriptive susceptible de trancher des débats historiques sur le discours politique contemporain, la dimension heuristique et la finalité herméneutique se sont imposées chaque année un peu plus au cœur de notre démarche[47].

Renoncer à prouver paraîtra suffisamment modeste ; prétendre mieux interpréter reste un programme ambitieux.

La pratique logométrique amène en effet à réfléchir à l'herméneutique – au sens linguistique réduit d'interprétation des textes et non dans celui philosophique large d'interprétation du monde–. En lisant Schleiermarcher, Szondi ou Bollack, nous nous sommes affronté à son mur c'est-à-dire à son cercle : l'acte interprétatif met le tout et ses parties dans un rapport sans commencement ni fin, le premier expliquant les secondes, la somme des secondes expliquant le premier. De ce cercle aporétique, aucun auteur ne prétend sortir, et notre *position pratique* décrite dans {**24**} a été inspirée par Heidegger :

[47] En conclusion de {**I** : 753-754} nous avons en effet narré l'accident scientifique à l'origine de notre conversion à la logométrie ; et cette narration insiste beaucoup sur la dimension probatoire du traitement. Nous rappèlerons encore *infra* le pouvoir descriptif de la logométrie, mais préférons insister désormais sur sa vertu heuristique.

« l'essentiel […] n'est pas de sortir du cercle, mais d'y entrer de la bonne manière » [Heidegger cité sous des formes différentes par Szondi 1989 : 10 et 105]

Pour notre part, nous entrons dans le cercle herméneutique par le bas c'est-à-dire par le corpus ; par sa matière attestée plus que par son idée supposée, par son expression observable plus que par son contenu (déjà) interprété, par son appareil formel plus que par son dispositif informel : la logométrie est un matérialisme textuel ou une herméneutique matérielle ; c'est une démarche à dominante inductive qui s'appuie sur les éléments positifs du corpus ; et nos pratiques sont essentiellement émergentistes.

La modestie de ces affirmations, moins glorieuses qu'une profession de foi théorique ou que ces démarches hypothético-déductives qui font la part belle à l'intuition du chercheur, nous est imposée de toute part, c'est-à-dire aux trois niveaux de réflexion qui sont synthétisés dans cet essai : réflexion sur l'objet (les corpus textuels), réflexion sur la méthode (la logométrie), réflexion sur la fin (le langage politique).

En amont d'abord, nous l'avons dit, le corpus est décrit comme la matrice active du sens et non comme son réceptacle passif. C'est pour cette raison qu'il convient moins de l'interroger *top down* que de le laisser nous interroger *buttom up*. La dynamique du corpus – pour qu'elle reste une dynamique – ne doit pas être corsetée par des hypothèses de travail trop contraignantes : celles-ci, répétons-le, président à la sélection des textes mais ne doivent pas gouverner l'analyse linguistique. Supposer de manière transcendantale un sens, c'est le plus souvent déjà l'importer là où nous disions qu'il réside de manière immanente et toujours renouvelée dans le corpus, sa composition, son organisation, sa réflexivité. Si le sens naît du bouillonnement contrastif ou réflexif du corpus, il convient de le laisser émerger du tréfonds plutôt que de l'hypothéser de l'extérieur ou du dessus. Notre linguistique de corpus est *corpus-driven* et non *corpus-based* [Tognini-Bonelli 2001]. Affirmer que le sens n'est pas une donnée objective des textes mais le produit de parcours interprétatifs au sein du corpus nous parait suffisamment subversif pour la sémantique traditionnelle pour que ces parcours interprétatifs eux-mêmes ne se chevillent pas sur le matériel textuel et se trouvent investis, informés, surplombés à tout moment par l'hypothèse ou la thèse du chercheur. Les analyses de discours, notamment, qui selon [Sarfati 2003 : 432] risquent de *manquer le texte en tant que tel* pour favoriser d'un côté les conditions de production des textes et de l'autre, pour ce qui nous intéresse ici, « le point de vue du chercheur sur les textes » ne nous apparaissent pas sans danger.

En aval ensuite, nous le verrons, le discours politique (très) contemporain est un matériau brûlant que l'on a tout intérêt à refroidir. Dans l'analyse, le citoyen perce sous le chercheur, et les sacro-saintes hypothèses scientifiques de travail sont le plus souvent pré-jugés partisans, *a priori* politiques et reflet de l'idéologie dominante du moment : à tout prendre, nous les jugeons non pas comme l'expression de la liberté créatrice du chercheur mais comme l'expression de son aliénation. En tout cas, la projection des hypothèses de travail dans l'objet de recherche est un risque classique en SHS que nul ne peut ignorer ; c'est le danger majeur en analyse du discours politique que nous devons sérieusement prendre en compte. Dans le détail, il est possible de montrer {33} que ce danger se décline de deux manières : trouver toujours ce que l'on cherche (établissement de conclusions artefactuelles ; la thèse est induite par l'hypothèse) ; ne

pas trouver ce que l'on ne cherche pas (ignorance d'éléments importants du corpus, non reconnus faute d'avoir été pressentis comme dignes d'intérêt)[48].

Au milieu donc, enfin, la méthode logométrique est heuristique et a comme principal objectif de favoriser, par l'organisation des parcours de lecture, l'émergence du sens. Mieux : par le contrôle des parcours interprétatifs – notamment en initiant le mouvement *par le bas*, et en chevillant l'interprétation *sur le* texte – la logométrie entend favoriser non seulement l'émergence du sens mais l'émergence des hypothèses de travail qui nous mènent au sens ; il s'agit donc moins de contrôler l'interprétation en tant que telle que de maîtriser le *processus* interprétatif, toutefois dans son premier et si fondamental mouvement d'entrée dans le cercle.

C'est en effet le plus souvent du traitement systématique et exhaustif logométrique – là où la lecture humaine est aléatoire, partielle, partiale; déjà orientée – que surgissent les hypothèses de travail. En bon ordre – de manière hiérarchique – le traitement logométrique fait remonter du corpus des caractéristiques majeures qui deviennent autant d'interrogations. Les milliers de sorties machine qui ont été consultées depuis 12 ans se sont transformées en effet en autant d'hypothèses de travail, de problématiques, d'interrogations. Entre mille, donc, illustrons le procès de la démarche par deux exemples, le plus lointain avec des conclusions arrêtées (c'est-à-dire un processus interprétatif poussé à son terme), et le plus récent avec des conclusions encore en devenir.

(i) Après comptage exhaustif et systématique de tous les mots d'un corpus contrastif de l'entre-deux-guerres {**I** : 229 et ss.}, le traitement des spécificités a fait ressortir qu'une des premières caractéristiques lexicales ou graphiques du discours de droite *versus* le discours de gauche dans les années 30 est le mot « a ». Pourquoi ?

C'est de ce constat que l'on peut difficilement imaginer plus épuré, c'est de cette interrogation que l'on n'envisage pas plus élémentaire que vont naître les hypothèses de travail non seulement sur la fonction du verbe-auxiliaire *avoir* dans la langue française mais sur son usage en discours ; usage bizarrement inégal selon l'appartenance gauche / droite des locuteurs dans le discours politique des années 30. Le processus heuristique est ainsi – et pas autrement – engagé, et de questionnement en questionnement, de vérification en vérification, de traitement en retraitement, nous avons pu montrer que la droite des années 30 pensait ou parlait au passé composé (lorsque la gauche parle notamment au futur). Penser et parler au passé composé (et aussi au plus-que-parfait et à l'imparfait) est une constatation non triviale pour la France de l'avant-guerre si l'on veut bien s'essayer à une interprétation socio-linguistique puis historique. La droite de l'avant-guerre tient sans toujours s'en rendre compte un discours passéiste ou un discours de l'accompli, un discours du bilan ou un discours du constat[49] au détriment de toute projection ; et ce discours du constat est d'autant plus douloureux que la droite des années 30 pose ses yeux nostalgiques sur un monde qui s'écroule. De fait, le malaise républicain de la droite, certes originel, prend une acuité nouvelle avec le 6 février

[48] Inversement, [Rastier 2009] constate : « Comme souvent, quand l'instrumentation permet un nouveau rapport à l'empirique, on ne trouve plus ce que l'on cherche, et l'on trouve ce qu'on ne cherchait pas ».

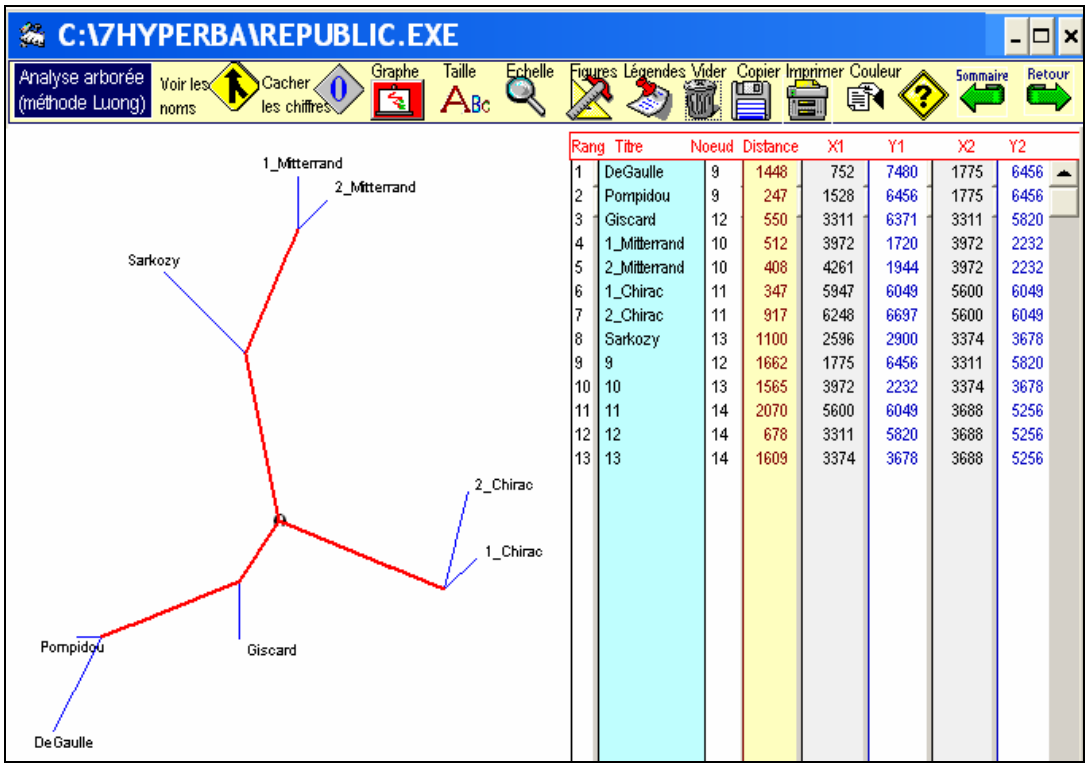
[49] Sur-utilisation massive du passé composé donc, sur-utilisation aussi du plus-que-parfait et de l'imparfait, sur-utilisation encore du gallicisme « il y a », ou « il y avait », sur-utilisation enfin des dates : telles sont les sorties-machine sur lesquelles s'initie l'interprétation {**I** : 229 et ss.}.

1934, le danger communiste qui s'incarne dans le Front populaire ou la confrontation / répulsion / attirance avec les fascismes : sans horizon politique la droite se retourne alors vers de vieux modèles bonapartistes hérités du siècle précédent. Surtout, la crise économique mondiale qui touche la France à partir de 1930, ébranle les certitudes économiques d'une droite française fraîchement convaincue par le capitalisme industriel : sans horizon économique, la droite se retourne ici encore vers un modèle économique agraire et paternaliste hérité du XIX^{ème} siècle. Bref, c'est une droite rétrograde, qui apparaît épuisée idéologiquement, qui voit arriver la guerre, et qui s'abandonnera *quasi* unanimement dans la réaction et le vichysme. Plus loin donc, Vichy, son idéologie, son discours n'apparaissent plus comme des accidents de l'histoire issus de la défaite militaire mais comme l'aboutissement logique d'une classe politique incapable d'imaginer – de dire, de *conjuguer* – l'avenir, et qui se tourne, dès les années 30, vers un *passé* césariste et ruraliste re-*composé*[50].

(ii) Sur le corpus contrastif des présidents de la République de de Gaulle à Sarkozy[51], le global et puissant calcul de la distance intertextuelle [cf. *Corpus* 2, 2003], montrera, dans des publications encore à paraître, que le discours de Nicolas Sarkozy est relativement proche du discours de François Mitterrand (illustration 5). Pourquoi ? (Et on comprend que l'interrogation, surgie du corpus et non imposé à lui, est ici d'autant plus intéressante que le constat est contre-intuitif au regard de nos *a priori* sur la différence de culture des deux hommes, différence d'appartenance politique, différence de génération, etc.).

[50] Toujours plus loin dans le processus interprétatif, résumons encore une autre conclusion forte tirée, de proche en proche, du constat de la sur-utilisation du passé composé dans le discours de la droite. Le pétainisme comme aboutissement logique du discours passéiste d'un Flandin ou d'un Tardieu ne peut plus être situé dans la généalogie tripartite de René Rémond. Pour [Rémond 1954-1982], le vichysme est l'héritier de l'extrême-droite *légitimiste* : il apparaît pourtant ici comme l'héritier d'une droite *orléaniste* qui refuse le présent et l'avenir. Dit autrement, la tripartition de René Rémond, qui s'applique à dresser trois généalogies à droite, en les supposant étanches, ne tient pas pour la période ; et sans doute pas plus pour les autres. Cette conclusion a été discutée, puis validée, par les historiens de notre jury de thèse.

[51] Le corpus Sarkozy est en cours de constitution et de traitement, au moment où ces lignes sont écrites seuls les discours des 18 premiers mois de sa présidence ont pu être saisis et traités. Les résultats présentés ci-dessous demanderont confirmation.



Rang	Titre	Noeud	Distance	X1	Y1	X2	Y2
1	DeGaulle	9	1448	752	7480	1775	6456
2	Pompidou	9	247	1528	6456	1775	6456
3	Giscard	12	550	3311	6371	3311	5820
4	1_Mitterrand	10	512	3972	1720	3972	2232
5	2_Mitterrand	10	408	4261	1944	3972	2232
6	1_Chirac	11	347	5947	6049	5600	6049
7	2_Chirac	11	917	6248	6697	5600	6049
8	Sarkozy	13	1100	2596	2900	3374	3678
9	9	12	1662	1775	6456	3311	5820
10	10	13	1565	3972	2232	3374	3678
11	11	14	2070	5600	6049	3688	5256
12	12	14	678	3311	5820	3688	5256
13	13	14	1609	3374	3678	3688	5256

Illustration 5 : Distance intertextuelle calculée sur les lemmes. Représentation arborée selon la méthode Luong

Faut-il envisager une explication d'ordre générique ? Sarkozy reprendrait à Mitterrand un mode de communication politique spécifique tel que les interviews à bâton rompu et un genre de discours particulier. L'hypothèse de travail ici est forte car elle poserait à nouveau la question de la prégnance du genre sur une production discursive ; prégnance définitivement démontrée par plusieurs auteurs dont Brunet dans le domaine littéraire, mais que nous nous sommes permis de nuancer – voire de contredire – dans le discours politique en montrant en français {**17**} et en portugais {**29**} que le facteur idéologique était plus important que le facteur générique, et que la *formation discursive* façonnait davantage le discours politique que la norme du genre (voir *infra 3.1.1. Thorez* versus *Blum ou la question du genre*).

Faut-il envisager plutôt une explication d'ordre thématique ? Le rapprochement des deux discours viendrait du fait que Sarkozy n'hésite pas à utiliser des thèmes et un vocabulaire de la gauche mitterrandienne. Et si tel était le cas, l'analyse devra être poussée pour mesurer jusqu'à quel point l'emprunt lexico-politique est important. L'idée de rupture discursive sarkozienne que nous avons nous-même reprise {**27**, **34**, ***B***) devra être alors revisitée : las de rupture, il s'agirait alors d'une continuité insoupçonnée avec un président inattendu.

Faut-il avoir recours à des explications rhétoriques et à une catégorie d'analyse récemment (re)théorisée comme le *registre* [Gaudin et Salvan 2008] ? Pour convaincre, Sarkozy et Mitterrand n'hésiteraient pas à utiliser, l'un comme l'autre, le ton polémique – discours énonciativement tendu avec prise en compte et prise à partie de la pensée de l'adversaire – là où les autres présidents favorisent la tonalité didactique – discours universalisant, détendu, avec effacement de l'interlocution.

Plus simplement et avant tout, quels sont les mots qui expliquent ce rapprochement global imprévu remonté des profondeurs du corpus ? Sont-ce quelques grands substantifs (« pays », « politique », « gouvernement » etc.) ? Sont-ce plutôt des verbes, notamment les énonciatifs, les modaux et les (semi-)performatifs tels « penser », « dire », « vouloir », « falloir » ? Sont-ce encore les pronoms et d'abord le premier d'entre eux « je », lorsqu'un de Gaulle ou un Pompidou utilisent plutôt un « nous-la-France » ? Etc.

Nous laissons ici volontairement ces questions ouvertes pour témoigner du processus heuristique engagé et parce qu'à ce stade nous ignorons la réponse, mais les hypothèses de travail imposées par le traitement logométrique du corpus et non projetés par le pré-jugé de l'analyste ne manquent pas. Et si les conclusions que nous serons amené à tirer resteront à coup sûr discutables, les interrogations qui ont émergé là où elles sont habituellement importées, auront quelque légitimité (et quelque obligation) à avoir été posées.

En de simples mots, et à l'heure de décrire sans maquillage notre pensée, nous avons la faiblesse de croire qu'il est fondamentalement différent, d'un côté, d'hypothéser par exemple une proximité entre les discours de Sarkozy et ceux de de Gaulle et de chercher à confirmer ou infirmer – confirmer le plus souvent ! – notre géniale intuition, et de l'autre côté, de constater (ici grâce au calcul de la distance intertextuelle) un certain rapprochement entre deux locuteurs (ici Sarkozy et Mitterrand) et de réfléchir, sur la base de ce constat empirique impérieux, sur les raisons linguistiques et socio-linguistiques de ce rapprochement[52].

Le cercle herméneutique est toujours un cercle – c'est ce qui fait sa fertilité–, mais nous y entrons différemment. Dans le premier cas, il s'agira d'objectiver (de démontrer) la subjectivé (l'hypothèse de travail). Dans le second cas, il s'agira de subjectiver (d'interpréter) l'objectivité (un constat linguistique empirique dûment attesté). Impasse symétrique, pensera-t-on, dont les écueils symétriques seraient soit la conclusion artefactuelle induite par l'hypothèse, soit le bond interprétatif qui nous projette à un moment hors du simple constat[53]. Impasse dissymétrique, affirmons-nous, car seul le second cas aura fait avancer une description reproductible de l'objet.

S'il est probable que les interprétations de tel phénomène linguistique en effet varieront – les nôtres ont toujours été très engagées –, posons *a minima* que la communauté scientifique pourra désormais réfléchir sur des bases descriptives communes : sur-utilisation matérielle du passé composé dans le discours de la droite de l'entre-deux-guerres ; proximité discursive attestée entre Mitterrand et Sarkozy ; sous-utilisation massive du vocabulaire marxiste chez Blum dans l'entre-deux-guerres ; affaiblissement du vocabulaire bolchevik et apparition du vocabulaire jacobin dès 1931

[52] Fondamentalement différent d'un côté d'hypothéser – non sans *a priori* historique – une rupture politico-discursive du PCF en 1935 avec le Front populaire, et de voir de l'autre côté émerger un tournant discursif dès 1931, 1932 ou 1933 et de proposer une interprétation de cette mutation précoce (*cf.* une des principales conclusions de notre thèse {**I** : 407-418 , particulièrement l'AFC p. 477}, voir aussi *infra 3.2.1*). Fondamentalement, différent de pré-supposer – non sans *a priori* citoyens – un pétainisme inavoué chez Sarkozy et de voir émerger, par un traitement systématique des co-occurrences, l'utilisation simultanée de « travail » « famille » et « patrie » dans le discours, et de proposer une interprétation de cette co-occurrence {**30**,**31**, **32**}. Etc.

[53] Bond interprétatif ? Dans ce deuxième cas, [Peschanski 1989 : 8] parle de la nécessité de « franchir le rubicon de l'interprétation »

chez Thorez ; nominalisation systématique du discours chez Giscard et pronominalisation du discours chez Mitterrand ; sur-emploi de « naturellement » et des adverbes chez Chirac ; co-occurrence statistique de « travail », « famille », « patrie » dans le discours de campagne du candidat Sarkozy en 2007 ; spécificité positive spectaculaire de « je » chez Blum puis chez Mitterrand ; inclassabilité lexicale et discursive du discours de Jospin en 2002 (*versus* son discours antérieur entre 1997 et 2001) ; montée en puissance des segments répétés de longueur importante dans le discours de Chirac au fil de sa présidence (1995-2007) et sur-emploi constaté du présent de l'indicatif par rapport aux autres présidents ; longueur objective de la phrase chez Royal (*versus* la phrase chez Sarkozy) , etc[54].

Résumons donc pour conclure : la statistique ne prouve rien : une *probabilité* d'emploi n'est pas, par définition, une *preuve*. Elle a vocation non à inférer mais à interroger de manière inductive. Du reste, l'inférence est toujours chose délicate dans une démarche inductive : d'un constat empirique, même bien établi, il est risqué de tirer une loi ou une relation de cause à conséquence définitives ; en matière textuelle, il est souvent imprudent de conclure du général le systématique[55]. En revanche, l'induction s'arme volontiers de la statistique. La remontée d'hypothèses de travail ou de constats, sur lesquels l'on commencera à réfléchir, ne peut en effet s'envisager que si les moyens de description de l'existant (les régularités et irrégularités linguistiques du corpus) sont sûrs et solides ; et sur nos grands corpus de textes, souvent inaccessibles à l'œil et la mémoire humaine, nous envisageons mal de moyens d'exploration plus réglés que l'informatique (dépouillement systématique et exhaustif du matériau brut ou qualifié du texte) et plus solides que la statistique (dénombrements, indices, étude de la macro et de la micro distributions des termes, repérage des récurrences ou des absences, des co-occurrences, etc.). Reste enfin que si le traitement quantitatif a la lourde tâche, par le compte rendu qu'il donne du corpus, de contrôler la description et d'initier le processus interprétatif, c'est par le retour au texte que ce processus pourra aboutir. *Le nombre fait sens* (déchiffrement quantitatif des saillances linguistiques du corpus) ; *le sens naît en/du (con)texte* (lecture contrôlée de la chaîne du texte) : entre ces deux affirmations, que l'ergonomie d'un logiciel comme Hyperbase met minutieusement en forme, se situe notre pratique quotidienne.

[54] On l'a compris, nous énumérons ici en vrac quelques constats linguistiques qui nous ont servi d'*entrées* dans le processus interprétatif, et ont abouti, au terme de ce processus, à des interprétations tranchées. Pour épuiser le sujet, répétons donc après {**I** : 754} et {**II** : 177}, la règle herméneutique intangible fixée dans notre recherche : *ne jamais rien dire qui ne soit initié par un constat logométrique issu d'un traitement systématique du corpus*. Dans le détail de nos ouvrages, équivalents à 1100 pages tapuscrites, nous ne voyons que 5 pages {**II** : 177-182} qui contreviennent à cette déontologie ; et nous nous excusions auprès du lecteur de cette contravention.

[55] Précisons par là, qu'en matière textuelle, nous nous méfions des corpus dits échantillonnés. (Un échantillon de texte(s) ?). Lorsqu'un corpus – exhaustif si possible – arrive à témoigner de lui-même, ne lui demandons pas de témoigner, en plus, d'autres réalités. Autrement dit, « le texte est unique en son genre » selon l'expression de Riffaterre rapportée par [Adam 2008 : 13] : il ne peut que difficilement *être représenté* par autre chose que lui-même ; ni *prétendre représenter* autre chose que lui-même.

Conclusion

L'engagement méthodologique au cœur ou à la périphérie de chacune de nos productions scientifiques depuis 12 ans et que nous entendons poursuivre dans l'encadrement de la recherche paraîtra naïf.

Les textes sont des objets si complexes qu'aucun formalisme ne pourra jamais prétendre les épuiser. Surtout, les textes ont une dimension esthétique : leur science est pour beaucoup aussi un art qui en appelle autant au sentiment qu'à la raison.

Dans ces conditions, s'enfermer dans une pratique logométrique rigoureuse convoquant la machine et les mathématiques, l'informatique et les statistiques peut apparaître réducteur voire contre-productif. De fait, il faut admettre avec le philosophe ou le poète, avec certains historiens travaillant sur textes rares comme certains littéraires sur pièces précieuses, l'inutilité de passer certains corpus textuels en machine pour leur meilleure compréhension.

Pourtant, on ne saurait ignorer que les sciences humaines souffrent dans leur ensemble d'un déficit et de méthode et de crédibilité dans les résultats obtenus. Cela affecte particulièrement les sciences ayant affaire aux textes et à leur interprétation.

La raison de ce déficit méthodologique, auquel nous sommes prêt à répondre par un excès de méthode, est connue. La croyance dans la transparence du texte et l'immédiateté du sens demeure une croyance bien établie. En Histoire, par exemple, l'apparente évidence du contenu dénoncée déjà par [Robin 1973] est toujours aveuglante pour un locuteur-analyseur natif, et les méthodes de traitement trop contraignantes peuvent apparaître superfétatoires : c'est le constat que nous avions pu tirer en français {**3**} et en anglais {**8**} juste après notre thèse {**I**}. Les échanges que l'on a pu avoir récemment grâce à Jean-Philippe Genet {**33**} autour du projet ANR ATHIS (Atelier international Histoire et Informatique – 2006-2009) ne démentent pas aujourd'hui le *French way behind* que nous regrettions[56]. Plus fondamentalement, la subjectivité peut être érigée, parfois à juste droit, en système dans les humanités ; ici en système indépassable d'interprétation des textes. Tout formalisme peut alors être perçu comme de l'objectivisme, et l'objectivisme caricaturé en scientisme : la logométrie, nous le savons, n'échappe pas à cette critique. Plus simplement, enfin, le texte ou le discours peut être considéré seulement comme un média, comme un support, comme un moyen, et non comme un acteur social en lui-même, un *événement* [Guilhaumou 2006], un objet scientifique, une fin ; et au titre de simple média qui transmettrait le monde là où, pour nous, il le construit, convenons qu'il ne nécessite pas de méthode d'analyse plus approfondie qu'une lecture immédiate et intuitive.

Dans ce cadre, et après avoir plaidé de manière complexe les vertus de la logométrie, revenons pour conclure au plus simple. Deux raisons prosaïques mais fondamentales président à notre engagement méthodologique passé et à venir, pour faire de lui une évidence : la taille des corpus traités ou à traiter (cf. *supra* 1. *Un objet :*

[56] Concrètement, c'est ce retard qui nous a poussé à postuler en section 34 *Langue, Langage, Discours* du CNRS et aujourd'hui définitivement à soutenir notre HDR en Sciences du langage. Notons au passage que la section 34 a fait alors preuve d'ouverture pluri-disciplinaire ; après le recrutement de Jacques Guilhaumou en 1986, c'était la deuxième fois de son histoire – et dernière fois jusqu'à nouvel ordre – qu'elle recrutait en son sein un chercheur dont la thèse avait été soutenue en Histoire.

les corpus textuels numériques) et leur nature politique (cf. *infra* 3. *Une fin : le langage politique*).

Les petits recueils de petites phrases ou les opuscules de *formules* sont l'objet de journalistes ou d'érudits [sur les *formules* voir l'ouvrage remarquable de Krieg 2009, compte rendu dans {**h**}]. Notre objet et nos préoccupations sont autres. Nous avons en effet, du mémoire de Maîtrise en 1995 à la thèse que nous co-dirigeons aujourd'hui, cherché à appréhender de gros corpus, souvent exhaustifs, avec l'espoir d'en rendre compte non seulement ponctuellement et localement mais systématiquement et globalement ; *a minima*, avec l'espoir d'y naviguer sans s'y noyer. (Et devant l'inflation textuelle que produit chaque jour notre société numérique, la taille de nos corpus ne devrait pas diminuer). Dès lors, l'assistance de la statistique et de l'informatique – la logométrie – ne se vit pas comme une option mais comme un impératif.

Surtout, le discours politique n'est pas un discours comme les autres (lire immédiatement 3. Une fin : le langage politique). Sa fonction sociale l'oblige. Le discours politique a moins une dimension esthétique qu'une dimension pragmatique et plus précisément propagandique. La pertinence d'un discours politique se mesure à son efficacité pratique, jusqu'à sa force perlocutoire. *Et la redondance est la condition de cette efficacité*. La construction de l'*ethos* du locuteur politique par exemple passe par une performance linguistique à chaque fois rééditée de tribune en tribune, de discours en discours, de paragraphe en paragraphe. Les thématiques – en période électorale notamment – pour s'imposer en *agenda setting* doivent être développées puis reprises avec des répétitions qui ressemblent parfois à des bégayements. Les mots, tout simplement, qui apprivoisent la réalité et l'enferment, ne s'imposent comme référents sociaux ou comme *praxèmes* qu'à l'occasion d'une récurrence d'emploi (*récurrence combinée* dans le cadre des co-occurrences) proche du martèlement. A côté de la taille des corpus embrassés, c'est donc la redondance intrinsèque du discours politique – redondance supérieure à celle du discours littéraire par exemple ; redondance que l'on préfèrera contrôler plutôt que ressentir par ailleurs – qui fait de la logométrie, pour nous, non pas un luxe mais une nécessité.

(. . .)

(. . .)

Au final, ces rappels sur la banalité du discours chiraquien visent seulement à souligner un élément trop souvent oublié dans le traitement quantitatif – ou dans l'idée que l'on se fait des traitements quantitatifs. La logométrie met en valeur, le plus souvent, les récurrences importantes c'est-à-dire les éléments saillants d'un discours. Parfois, de manière symétrique, ce sont les absences remarquables qui sont exhibées (les spécificités négatives par exemple) afin de proposer une caractérisation en creux du discours. Pourtant parfois, c'est la banalité quantitative qui fait sens et offre à l'analyste des chevilles interprétatives ; comme essayera de le montrer Julien Bonneau dans sa thèse, à la suite de certains travaux saint-clousiens, le vocabulaire dit *banal* est un axe de recherche important qu'il conviendra de retravailler[71]. Pour plaire au plus grand nombre, Jacques Chirac semble s'être appliqué à utiliser dans une juste mesure un vocabulaire commun et des traits rhétoriques partagés. Là où d'autres – et nous pensons particulièrement au discours sarkozien (cf. *infra* immédiatement) – tiennent un discours typé, qui fait clivage ou dissensus, Chirac a choisi le discours moyen du consensus.

3.1.3. Sarkozy *versus* Royal

On pointera une difficulté importante dans l'étude précédente. Si le corpus présidentiel fait contraster des locuteurs, il fait contraster du même chef des périodes historiques différentes. A strictement parler, il est impossible d'imputer à l'identité politique des hommes ce qui relève peut-être de la chronologie d'une époque. Deux variables (le temps et les présidents) se superposent strictement sans que l'on puisse isoler l'une de l'autre.

Notre étude doctorale, elle, présentait l'avantage de comparer des locuteurs dans une même période donnée (1928-1939). Néanmoins, un autre paramètre parasitait l'analyse : la période étudiée est si lourde d'événements (crise de 29, arrivée d'Hitler au pouvoir, 6 février 1934, Front populaire, guerre d'Espagne, Munich, début de la 2ème guerre mondiale) qu'il devenait difficile de la traiter en synchronie. Si l'on veut bien mesurer, par exemple, l'évolution du discours communiste sous le coup des événements, on conclura à l'existence non pas d'un discours mais de deux discours successifs, le premier d'essence bolchevick, le second d'essence jacobine {**I** : *Les mutations du discours de Thorez* : 407-446 ; **10** ; voir aussi *infra* 3.2.1. *Thorez (1930-1939) : continuum lexical et hiatus chronologique*}. De la même manière, il n'existe pas un discours de Tardieu mais trois successivement entre 1928 et 1939, qui épousent la trajectoire politique de l'homme : un discours orléaniste (1928-1932), un discours bonapartiste (1932-1934) et un discours légitimiste (1934-1939){**I** : *Tardieu ou la tentation autoritaire* : 607-740}.

Bref, les études logométriques exigent des corpus contrastifs dont le tout constitue la norme par rapport à laquelle les parties comparées se caractérisent. Seulement, il est difficile d'isoler les variables principales de contraste des variables secondaires qui affectent le traitement. Ainsi veut-on contraster Chirac et Jospin pour comprendre ce qui différencie encore le discours de gauche et le discours de droite à la fin du XXème siècle ? Las, nous contrastons (aussi, du même coup) génériquement le discours d'un président de la République d'un côté et le discours d'un Premier ministre

[71] Après Charles Muller, Pierre Lafon par exemple notait dés le premier numéro de *Mots*, l'importance des *formes de bases* dans le discours politique [P. Lafon, *Mots*, 1, 1980 « Sur la variabilité de la fréquence des formes dans un corpus, pp. 127-165].

de l'autre {**14**}. Ainsi veut-on comparer de Gaulle à Mitterrand pour comprendre deux personnalités historiques majeures ? Las, nous comparons (aussi, du même coup) les années 1960 aux années 1980, et l'ORTF qui diffuse le discours gaullien à la télévision berlusconnienne[72] qui médiatise pour partie le discours mitterrandien.

Quoique restée superficielle car réalisée à chaud, l'étude du discours de la campagne électorale 2007 {**27**, puis **30**, **31**, **32**, ***B***} a le mérite de neutraliser au mieux les différentes variables qui informent un discours, pour espérer isoler, seule, l'identité discursive des locuteurs. La chronologie en effet est stricte et identique pour tous les orateurs étudiés : 5 mois entre janvier et mai 2007. Le genre est stable et bien identifié : il s'agit toujours de discours publics tenus lors de meetings électoraux. Le statut des locuteurs quoique différent dans le détail est identique pour l'essentiel : ils s'agit officiellement de candidats à la magistrature suprême.

Au-delà de l'image générale que l'on peut dresser de l'échiquier politique français de Buffet à Le Pen (cf. ci-dessous, par exemple, l'illustration 13 sur l'identité lexicale de chacun à partir de 30 mots significatifs), cette étude a permis de décrire pour la première fois le discours sarkozien. Les premiers éléments que nous avons pu alors mettre à jour ont été confirmés depuis aussi bien dans l'ouvrage de [Calvet et Véronis 2008] que par les études récentes de [Marchand http://pascal-marchand.fr/spip.php?rubrique4], le numéro spécial de [*Mots* 2009 {**B**}] ou l'ouvrage collectif de [Perrineau (éd.) 2008]. Surtout, comme mentionné plus haut, nos conclusions sur le discours du candidat-Sarkozy semblent pouvoir être reprises pour l'essentiel pour les discours du président-Sarkozy que l'on se propose d'étudier dans les prochaines années.

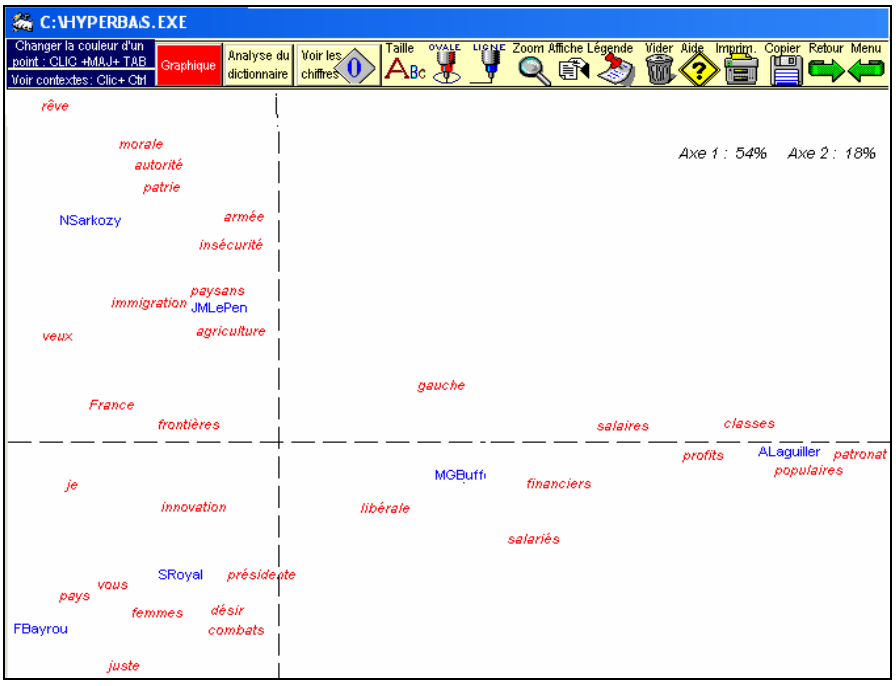


Illustration 13 : AFC de 30 mots durant la campagne 2007 (graphe non commenté ici)

[72] Nous avons noté ailleurs {**II** : 254} que la libéralisation ou privatisation des chaînes de télévision a été marquée en France par le rachat inaugural, en 1986, de la 5ème chaîne par Silvio Berlusconi ; suivra, en 1987, le rachat de TF1 par le milliardaire Bouygues.

A l'inverse de celle de Chirac, l'identité discursive de Sarkozy est forte et volontairement marquée. Sarkozy signe ses discours quantitativement et qualitativement, quand une forme d'anonymat rhétorique avait fini par s'imposer au plus haut sommet de l'Etat. Trois traits linguistiques méritent d'être rappelés, car ils accréditent l'idée de ruptures discursives – sinon de ruptures politiques –, et érigent ainsi le parler sarkozien en objet d'étude privilégié pour l'avenir.

(i) A l'analyse, le discours de Sarkozy est d'abord celui de l'extrême lexical ou de l'hyperbole lexicale {27} là où le discours républicain traditionnel est habituellement tempéré. Explicitement, Sarkozy s'en prend aux *euphémismes* qui seraient le propre de la pensée unique et finalement responsables de la mort de la politique en tant qu'action déterminée sur le monde. Plusieurs fois en meeting, Sarkozy répète ce paragraphe, dans lequel, notons-le au passage, une habile confusion est entretenue entre le dire et le faire :

> Un jour j'ai utilisé le mot « racaille » en réponse à l'interpellation d'une habitante d'Argenteuil [...]. On me l'a reproché. C'est mépriser la jeunesse que de lui parler par euphémismes sous prétexte qu'elle ne serait pas capable de regarder la réalité en face. Quels éducateurs serons-nous si nous nous laissons aller à ces petites lâchetés ? Si les multirécidivistes n'ont rien à craindre ? Si les mineurs peuvent se livrer aux pires excès sans être punis ? Si nous apprenons à nos enfants que l'âge excuse tout ? Si les voyous ne peuvent même pas être appelés des voyous ? (30 mars 2007, Nice ; 19 avril, Marseille, etc.)

Dès lors, la logométrie ne manque pas de repérer dans les discours la récurrence de mots forts que Sarkozy s'applique à répéter et qui discrimine son discours par rapport aux autres candidats (ou aux autres présidents) plus timides lexicalement : « haine », « détestation », « rêve » (verbe et nom), « capitalisme », « polygamie », « voyous », « assistanat », « excision », « morale » (nom et adjectif), « viol », « barbarie »[73]. Lexique de fer donc contre la langue de bois ou de coton supposée de ses adversaires : ce fut la première conclusion de notre étude.

(ii) Le deuxième trait frappant du discours de Sarkozy a déjà été noté : il cherche le *dissensus* lorsque le discours habituel recherche le *consensus* ; il marque les clivages lorsqu'un Bayrou durant la campagne ou un Giscard durant sa présidence par exemple aspirent à l'unanimité ; il incrimine lorsqu'une Royal ou un Pompidou concilient. En ce sens, il s'approche plus du discours de l'extrême droite que celui de la droite traditionnelle. Ce dissensus recherché peut être repéré par le jeu de vocables antinomiques souvent axiologiques (« bon » / « mauvais » par exemple) ou par un vocabulaire ou des thématiques traditionnellement polémiques dans le débat politique français (la « sécurité » ou « l'insécurité », « l'immigration » ou « l'identité », l' « autorité », la « morale », la « religion »). Mais il est surtout important de noter que ces clivages souvent manichéens s'appliquent à traverser – à abolir – les classes sociales pour se situer dans le seul *domaine moral* : il y a là un distinguo important avec le discours d'extrême gauche (notamment de l'entre-deux-guerres), lui aussi dissensuel mais d'une toute autre manière. Chez Sarkozy, en effet, la ligne de fracture ne passe pas entre l'ouvrier et le patron, mais au sein même des classes populaires entre « celui qui se lève tôt » et l' « assisté » ; elle ne passe pas entre le patron et les salariés, mais au

[73] Nous ne reproduisons pas ici les indices statistiques. Le lecteur aura compris que l'analyse repose sur les mots spécifiques de Sarkozy par rapport à notre corpus de références (cf. pour le détail {27}).

sein même du patronat entre le bon patron et le « patron voyou » ; elle ne passe pas entre les Français et les immigrés, mais au sein même des immigrés entre l'immigré intégré et « l'immigré clandestin », « l'immigré sans papier » ou celui qui « égorge son mouton dans son appartement »[74]. Cette série de mises à l'index, transclassiques et bornées au domaine moral, donne au discours de Sarkozy un aspect violent mais reste électoralement efficace : on remarquera en effet que l'incrimination pointe toujours une minorité voire une marginalité, voire une déviance que la quasi totalité du corps électoral réprouve : le voyou, le polygame, l'exciseur, l'assisté sont des boucs émissaires idéaux. Enfin, si le discours de Sarkozy est un discours dissensuel, alors la négation en est sa clef de voûte. La logometrie classe en effet le « ne… pas » (et le « non ») comme le premier facteur quantitatif discriminant du parler sarkozien et en donne ainsi sa grille de lecture rhétorique (« il ne faut pas… », « vous ne pouvez pas… », « je ne veux pas… », « il n'est pas vrai que…. »). Le discours et la pensée de Sarkozy se construisent en s'opposant ; contredisent avant de dire ; nient avant d'affirmer ; interdisent avant d'autoriser ; se nourrissent du conflictuel et des antagonismes dans un premier temps, pour mieux paraître rassembleurs dans un second.

(iii) Faute d'épuiser entièrement un discours complexe, un troisième et dernier trait majeur mérite d'être relevé dans notre étude naissante: Sarkozy s'applique à tenir un discours *oralisé*. C'est au fond sa première signature qui contraste aussi bien avec Royal lors du second tour de l'élection présidentielle qu'avec ses prédécesseurs à l'Elysée lors de leurs allocutions[75]. Même en campagne électorale où l'on sait que le texte est préparé ligne à ligne, et parfois récité à la virgule près d'un meeting à l'autre, Sarkozy feint l'improvisation, la spontanéité, le dialogue direct. C'est peut-être à cette aune simple que l'on peut mesurer le mieux l'importance de la plupart des phénomènes discursifs et décrypter le plus simplement le parler sarkozien. L'hyperbole lexicale mentionnée, par exemple, a pour fonction de donner au discours un aspect entier et spontané, loin d'une langue administrative ou technocratique contrôlée, qui qualifierait peut-être la « racaille » de « contrevenant à l'ordre public ». L'omniprésence du « je », si caractéristique, montre un orateur *in praesentia*, debout face à son auditoire, et qui assume – incarne – directement son propos, sans se réfugier derrière un texte impersonnel ou un collectif politique sans forme. La récurrence des interrogations – le

[74] Lors de sa principale émission électorale à la télévision, Sarkozy a ainsi pu déclarer le 5 février 2007 sur TF1 à 21 heures : « Je suis le premier homme politique de droite à dire qu'il faut une immigration choisie, mais je vous dis aussi une chose avec la plus grande force, personne n'est obligé d'habiter en France et quand on habite en France, on aime la France et on la respecte… [Interruption] Non ce qui rend raciste monsieur, c'est cette espèce… [Interruption] Non, si Le Pen dit le soleil est jaune, je ne vais pas être obligé d'arriver en prétendant qu'il est bleu. Personne n'est obligé, je répète d'habiter en France, mais quand on habite en France, on respecte ses règles, c'est-à-dire qu'on n'est pas polygame… [Interruption]. J'y viens, on n'est pas polygame… [Interruption]. Je vais y venir. On ne pratique pas l'excision sur ses filles, on n'égorge pas le mouton dans son appartement et on respecte les règles républicaines. »

[75] Notons ici au passage que la question de l'oralité des discours est une limite importante de notre travail que l'on aimerait essayer de lever dans les années prochaines. Jusqu'à maintenant nous travaillons uniquement sur des textes issus, pour beaucoup, d'une retranscription officielle, disponible à ce titre sur le site de l'Elysée, de Matignon ou les sites officiels de campagne (ou encore dans le JO pour le discours ministériel ou parlementaire). Nous écoutons en général ces discours à la télévision, à la radio ou en vidéo mais sans affronter la question de la fidélité de la transcription ; sans étudier jamais, par exemple, la prosodie des discours ; ou encore sans analyser la gestuelle de l'orateur en meeting ou à la télévision.

point d'interrogation est hautement spécifique du discours de Sarkozy – a encore pour vocation de simuler l'interaction orale avec l'auditoire et un dialogue supposé : à moindre frais, par l'auto-interrogation, Sarkozy simule le débat participatif. Les redondances et les répétitions, dont la forme la plus aboutie sont les anaphores rhétoriques lourdes à force d'être nombreuses dans le discours[76], font ici aussi penser à un discours qui se construit sur le champ, dans une forme d'improvisation bégayante. Le « on » encore – deuxième caractéristique quantitative après le « ne… pas » – qui méritera un développement particulier dans la lignée des travaux d'Irène Tamba, pourrait être ramené, encore, à cette dimension, si l'on veut bien le considérer comme une forme relâchée et orale du « nous » académique, etc.

Hyperbole lexicale, conflictualité ou dissensus, oralité et forme relâchée : bien sûr, il ne s'agit là que d'un résumé des traits quantitatifs d'un discours qui reste, à ce stade, plus un sujet de recherche pour les années à venir qu'un sujet épuisé. Mais le discours de Sarkozy nous parait d'ores et déjà un objet d'étude privilégié en matière d'identité discursive tant celle-ci parait marquée et en rupture avec l'identité discursive policée des présidents successifs ou des derniers présidentiables. Et ici concluons que cette identité discursive épouse et construit une identité politique tout aussi marquée. En termes rhétoriques, dans la lignée des travaux de [Amossy 1999 et 2006], de [Maingueneau 2004] ou récemment du [n°3, 2009, *Argumentation & Analyse du discours* : « Ethos discursif et image d'auteur »], on notera que l'efficacité du propos sarkozien tient dans l'adéquation remarquable entre l'éthos discursif que construit le discours et l'éthos pré-discursif (disons, l'image politique préalable) que Sarkozy s'est forgé depuis plusieurs années[77] : durant la campagne, les « je veux » du candidat par exemple, qui avec le « ne… pas », et le « on » restent la signature linguistique principale, renvoient et alimentent l'image du ministre de l'Intérieur qui a pris en charge fermement le problème de l'insécurité entre 2002 et 2007. Les « ne…pas » renvoient au père sévère ou au premier flic de France sachant poser des interdits dans une société qualifiée depuis mai 68 de laxiste et déliquescente. Plus généralement, ce phrasé direct, rapide[78] et cette langue relâchée, ce lexique extrême et cette pensée décomplexée, ces thématiques aux confins de l'extrême droite comme l'immigration ou l'identité nationale, ces valeurs assumées comme « l'ordre », « l'autorité », la « morale »[79] renvoient au corpus idéologique de cette nouvelle droite dynamique, populaire ou populiste, bonapartiste ou autoritaire qui en Europe (Berlusconi, Aznar) ou aux Etats-Unis (Bush) a su séduire, au-delà des personnes âgées et des retraités, des classes aisées et de l'élite conservatrice, la plèbe au début du XXI^{ème} siècle[80].

---

[76] Ici aussi l'approche quantitative est parlante : durant la campagne électorale près d'un 1/3 des phrases prononcées sont des reprises anaphoriques.

[77] [Amossy 1999 : 147] parlerait aussi d'adéquation entre *éthos discursif* et *éthos institutionnel*.

[78] Avec le point d'interrogation (?), le point (.) est hautement spécifique de Sarkozy *versus* les autres candidats à la présidentielle de 2007. La conclusion est facile à tirer : Sarkozy fait des phrases plus courtes que les autres, en moyenne de 22 mots *versus* 33 mots pour Royal. Ce même rapport d'un tiers de moins a été aussi constaté entre la phrase sarkozienne et la phrase gaullienne, pompidolienne ou mitterrandienne.

[79] Sur le lexique et la thématique sarkozienne voir {**27**, mais aussi **30**, **31** et **32**, ***B***}

[80] La sociologie électorale a bien identifié le vote sarkozien. Il s'agit d'abord des personnes âgées : 44 % des 65-74 ans ont ainsi voté Sarkozy dès le premier tour (contre 31 % de la population globale = +13). Il s'agit ensuite des classes très aisées : 56 % des personnes aux revenus supérieurs à

comme une « méthode interprétative par excellence » [Guilhaumou 2006 : 40] dont le but est d'encadrer la subjectivité créative du lecteur.

Un objet, une méthode, une finalité : les trois entrées dans notre travail passé constituent un programme de recherche d'avenir.

(i) Le corpus reste en grande partie aujourd'hui à modéliser ; du reste, il s'agit là du projet collectif de notre laboratoire de rattachement, *Bases, Corpus et Langage*. Le passage au numérique en effet permet et exige de reprendre une réflexion aussi ancienne que l'étymologie du mot ; et ce passage – une révolution culturelle au-delà des simples avancées techniques – est aujourd'hui à peine engagé. Dans notre domaine, comme le notent Adam et Viprey, dans leur appel à contribution pour [*CORPUS* 2009. *Corpus de textes, textes en corpus*], il semble qu'il y ait actuellement un écart entre des pratiques surabondantes qui se réclament des corpus (textuels) et une théorisation encore déficitaire. Victime de son succès ne serait-ce que par la vulgarisation du syntagme « linguistique(s) de corpus », la notion de corpus reste à travailler. Particulièrement, nous semble-t-il, l'heure est de savoir si la linguistique accepte le corpus – notamment les corpus numériques – comme un objet ; non comme un moyen, un champ d'expérimentation, un réservoir d'exemples mais un objet linguistique *en lui-même* ou *pour lui-même*, *dans lequel* ou *par lequel* s'épanouit le sens ; bref, le corpus comme condition du déploiement de la sémantique interprétative. Si l'on connaît des linguistes assez hardis pour déclarer le texte comme un des objets de la linguistique, peu, nous semble-t-il, revendiquent le même statut pour le corpus. De fait, objecte-t-on, le corpus n'est pas le produit naturel du locuteur mais celui de l'analyste. Mais le texte n'est-il pas souvent le fruit des éditeurs ? La phrase, le fait du grammairien ? Du reste, si les corpus sont bien artefactuels – des objets construits –, certains le sont sous l'autorité de l'auteur lui-même, comme cette sélection éditoriale de discours de [Chirac 2007], ou le site officiel de l'Elysée qui met plus ou moins en valeur tels ou tels discours du président : ici, c'est toute la réflexion autour de la tension entre *archive*, *corpus* et *discours* qui peut être reprise [Guilhaumou, Maldidier, Robin 1994]. La linguistique de corpus en tout cas, telle que nous l'entendons, serait celle qui prendrait à bras le corps la « corporalité », c'est-à-dire les effets de sens, de cohésion, de cohérence du corpus, ou les conséquences que produit l'immersion du mot, de la phrase, du texte dans un corpus ; les effets réellement produits par *la mise en corpus* [Adam et Viprey 2009 : 18]. Les pistes sont certes aujourd'hui tracées mais l'essentiel reste à réaliser en la matière. La notion de *corpus réflexif* notamment que nous avons avancée pour régler le régime entre textes dans le corpus, c'est-à-dire au niveau de l'intertextualité, se trouve retravaillée actuellement par exemple par [Florea 2009] à un niveau intratextuel : cela réclame confirmation. Plus généralement, les réflexions sur l'architextualité – notion floue depuis Genette mais que des macro-corpus numériques structurés pourraient matérialiser – réclament sans aucun doute un approfondissement, notamment à la lumière des avancées méthodologiques de l'ADT auxquelles nous espérons contribuer, et de celles de la philologie numérique.

(ii) La méthode utilisée n'a pas d'âge. Dans ses fondements, nous ne pouvons pas la dater. A ce titre, il est peut-être ambitieux de vouloir la renouveler. Les concordanciers par exemple sont presque aussi vieux que les livres et sont répandus en Occident dès le XIII^{ème} siècle. Les dictionnaires alphabétiques des formes remontent eux à l'Antiquité. Plus récemment, la création d'index (des noms propres, des lieux, des

notions) est devenue la norme dans l'édition moderne. Quant au décompte des unités textuelles, il a sans doute toujours été un fantasme pour l'interprète, et les ordres de grandeur ont toujours servi à décrire le texte comme ils décrivent le monde[121]. De manière moins générale, on considèrera que la logométrie – anciennement la lexicométrie – est solidement constituée en corps de doctrine dans les années 1980[122]. Il s'agit donc seulement aujourd'hui de faire fructifier l'héritage et de prolonger un appareillage informatique et statistique qui a fait la preuve de son efficacité, mais adapté essentiellement d'une part au corpus brut et d'autre part au schéma d'urne. En une phrase qui résume ces deux limites : le texte était jusqu'ici avant tout considéré comme *un sac de mots graphiques* là où on peut le considérer désormais, aussi, comme *un espace d'unités linguistiques établies*. Nous l'avons dit, en effet, deux pistes de recherche doivent être poursuivies. D'abord, sous le double effet de l'enrichissement des données que permettent les nouveaux formats de saisie des textes, comme XML, et le perfectionnement des étiqueteurs automatiques ou des analyseurs morpho-syntaxiques, c'est l'unité de décompte qui demande à être redéfinie. Ou plutôt (car nous avons une idée des unités quantifiables d'un texte), l'enjeu est de pouvoir mettre en correspondance ou en dialogue les traitements quantitatifs que l'on est capable aujourd'hui de faire sur les différents états de texte (texte brut, texte lemmatisé, texte étiqueté / grammaticalisé) : nos modestes propositions dans {**23**} demandent prolongement. Ensuite sous l'influence de la linguistique textuelle, les traitements quantitatifs doivent s'affronter à la linéarité du texte, c'est-à-dire que des traitements quantitatifs susceptibles de prendre en considération *l'organisation spatiale* du texte doivent prendre le relais des approches probabilistes traditionnelles. Ces deux programmes de recherche (diversification des unités textuelles / approche topologique des textes), quoiqu'indépendants l'un de l'autre, se trouvent exposés conjointement dans la proposition d'avenir de [Mellet et *alii* 2009 : 109] :

> Nous partons du postulat qu'un texte est d'abord un ensemble (E) d'unités linguistiques qui ne sont pas indépendantes les unes des autres, et qui est muni d'une structure ou, plus exactement, de plusieurs structures imbriquées dont l'union constitue cet ensemble.

(iii) Le langage politique enfin est un sujet inépuisable que nous n'avons jusqu'ici qu'effleuré dans notre travail et que les années futures ne suffiront pas à épuiser. Aussi loin que remonte la *polis* : le *logos*. Aussi loin que remonte la démocratie : le discours, la Pnyx, l'agora. Mieux, Athènes la machiste et l'esclavagiste n'a pas inventé la démocratie : elle nous a légué la logocratie c'est-à-dire une organisation sociale et un système de gouvernement des hommes par le discours (même

[121] Concordandiers, dictionnaires des formes, index, traitement quantitatif : inutile d'insister sur l'apport de l'ordinateur pour ces pratiques ancestrales. Du texte à l'hypertexte, de la lecture à l'hyperlecture, du décompte manuel à une statistique textuelle moderne : les avancées sont décisives.

[122] On ne reconstituera pas ici une généalogie précise. On sait seulement l'importance des travaux de [Muller 1973 et 1976], des thèses de [Brunet 1978] et [Lafon 1984], de la création du laboratoire de Saint-Cloud et des 10 premiers numéros de la revue *Mots*, du rôle des logiciels historiques comme Lexico ou Hyperbase. C'est peut-être la thèse d'Etat de [Salem 1993], reproduite pour partie dans [Lebart et Salem 1994] qui scelle la discipline. Rappelons que cette généalogie – à laquelle il faudrait intégrer l'œuvre majeure et déterminante de Benzécri, puis la littérature récurrente autour des JADT – est essentiellement française et que le monde anglo-saxon apparaît bizarrement en retard au début du XXI^{ème} siècle en matière de statistique textuelle, comme le montre l'artisanat des travaux actuels de [Biber 2004], [Partington et Morley 2004] ou [Teubert 2009].

sophistique), la raison (même déraisonnable), la rhétorique (même manipulatrice). Socialement, notre objet d'étude n'est donc pas anecdotique, et ce n'est point s'aveugler sur le *linguistic turn* que d'affirmer qu'il est constituant et constitutif de la vie des sociétés ; particulièrement sans doute de nos sociétés contemporaines où la doxa se trouve à chaque instant reconstruite par les discours que véhiculent sans cesse la radio, la télévision, internet. Même dans le champ étroit du seul XX^{ème} siècle français, même travaillé par de nombreux auteurs avant nous, le sujet reste donc en devenir, ne serait-ce que parce qu'il est un objet en constant renouvellement. Renouvellement en quantité d'abord comme l'atteste par exemple la production quotidienne voire bi-quotidienne d'un discours par l'Elysée (messages, allocutions, interviews, tribunes, etc.). Renouvellement en qualité surtout comme semble l'attester les constats diachroniques que nous avons déjà pu faire sur les dernières décennies, et dont l'intrigue contemporaine – notamment le parler Sarkozy qui s'annonce original – continue de nous tenir. Objet majeur donc qui appelle approfondissement, le langage politique l'est enfin, aussi, pour nous, parce qu'il s'agit d'un objet linguistique privilégié que les linguistes négligent encore et qu'on gagnera dès lors à appréhender. Rares sont en effet les problématiques du texte et du discours, voire celles de la langue qui ne prennent une dimension particulière dans le langage politique. Le dialogisme, la polyphonie, la circulation des discours par exemple, peu évoqués jusqu'ici dans notre travail, ne prennent-ils pas une lumière plus crue dans l'arène politique que partout ailleurs ? L'argumentation dans le discours et dans la langue, toujours en marge de nos réflexions, connaît-elle un champ d'étude plus pertinent que le langage politique ? Et encore, n'est-ce pas de manière sur-qualifiée que le langage politique peut témoigner de la performativité du langage, de la subjectivité dans la langue, de l'intentionnalité des discours ? La question générale, qui traverse toute les linguistiques, de la *référence* ne prend-elle pas une résonance particulière dans des textes politiques parfois instituants, souvent déterminants ? Ou enfin (sans avoir l'idée d'être exhaustif) : les problématiques de l'énonciation (de manière concrète la valeur par exemple d'un *nous* (chez Thorez) ou d'un *on* (chez Sarkozy)) ou plus généralement l'approche pragmatique ne s'exemplifient-elles pas mieux qu'ailleurs dans le *hic et nunc* si contraignant d'un locuteur politique ?

Et gageons que cette acuité linguistique n'est pas étrangère à l'importance sociale mentionnée précédemment : le discours politique est pour nous un programme de recherche fascinant, gros de recherches doctorales à encadrer, parce que le locuteur politique actualise la langue et le monde dans un même mouvement en parlant. Sa performance linguistique et sa performance politique ne font qu'une.