De la lexicométrie à la logométrie

Damon Mayaffre

De la lexicométrie à la logométrie

de Damon Mayaffre

CNRS, Université de Nice (France)

- Le Tournant de la lemmatisation
  - Lemmatiser = dégrouper
  - Lemmatiser = regrouper
- La Forme graphique: pis-aller ou choix théorique?
  - Les Limites techniques des années 1980-2000
  - Un choix théorique
- De la logométrie

Introduction

Depuis sa constitution à la fin des années 1960, la lexicométrie politique a connu en France une heure de gloire pour aujourd'hui s'essouffler. Longtemps soutenue institutionnellement par le laboratoire «Lexicométrie et textes politiques» de l'ENS Saint-Cloud ou par la revue *Mots/Ordinateurs/Textes*, elle souffre de la disparition *ab intestat* du premier et de la reconfiguration éditoriale du second (1). Dans le même temps, notons que le développement d'une lexicométrie littéraire, dont on date les premiers balbutiements dès l'après-guerre avec les travaux de Busa (voir l'historique dans Busa, 1998) puis de Guiraud (Guiraud, 1954), a connu le même désenchantement puisque après les analyses fondatrices de Charles Muller sur le théâtre classique (Muller, 1967) ou d'Etienne Brunet sur l'ensemble de la littérature française (Brunet, 1981), peu de littéraires aujourd'hui étayent leur exégèse des textes par un traitement informatisé et quantitatif des corpus (2).

Cependant, cet essoufflement paraît devoir être éphémère tant il est paradoxal dans le panorama scientifique actuel: en effet les objections longtemps dirimantes contre l'approche lexicométrique nous semblent aujourd'hui en grande partie dépassées, et le désamour intervient étrangement au moment où les hypothèques sont pour l'essentiel levées. Matériellement, par exemple, la disponibilité de textes numérisés de plus en plus nombreux et de bonne qualité éditoriale, sous un format universel XML, non seulement favorise mais réclame une approche automatique et quantitative. Là où les chercheurs étaient arrêtés dans leur premier mouvement par la fastidieuse saisie numérique des textes, ils se trouvent aujourd'hui noyés par des données textuelles informatisées de plus en plus vastes et immédiatement disponibles sur le Web ou ailleurs. Concomitamment, le développement d'outils lexicométriques toujours plus puissants rend possible le traitement de ces macro-corpus textuels. Longtemps limitées aux traitements d'ensembles de 250.000 ou 500.000 occurrences, les capacités des logiciels sont sans cesse repoussées, pour donner à l'outil - nécessaire supplétif à l'œil ou à la mémoire humaine à partir d'un certain volume - toute sa raison d'être. Bref, la disponibilité et l'abondance des corpus numérisés d'une part, l'amélioration des capacités des logiciels d'autre part sont la condition *sine qua non*, désormais remplie, du redémarrage de la linguistique quantitative assistée par ordinateur.

Les obstacles techniques ou matériels levés, c'est sur la plus profonde objection contre le traitement lexicométrique que nous voulons ici nous arrêter, celle remettant en cause fondamentalement la validité linguistique de l'approche, c'est-à-dire sa pertinence scientifique. Très vite, en effet, les linguistes ont souligné la vanité du traitement lexicométrique car celui-ci s'arrêtait à la matérialité graphique des textes. De fait, le «mot», pris dans sa définition la plus restrictive, ne recouvre pas une réalité linguistique opérante pour permettre la compréhension des textes. Ainsi la lexicométrie peut être soupçonnée de permettre, au mieux, une description du contenu matériel «de surface» des textes, et aucunement d'en recouvrer le sens. Au fond, elle serait un gadget coûteux en temps, sans grande pertinence scientifique.

Le débat sur la lemmatisation des textes, *c'est-à-dire sur l'unité d'indexation et de décompte en lexicométrie*, n'est pas nouveau puisque, dès l'origine, Charles Muller l'a mis en chantier (Muller, 1963). Mais après d'autres, dans le domaine de la littérature française (Muller, 1963, 1967, 1968, 1977; Brunet, 2000 et 2002), dans celui de la littérature latine (Mellet et Purnelle, 2002; Longrée, 2002) ou dans celui du discours politique (Labbé, 1990 et ici même 2002), nous voulons montrer que les progrès récents des logiciels de lemmatisation, articulés à la nouvelle génération des logiciels de lexicométrie aboutissent à une mutation et à une amélioration de nos

pratiques statistiques sur les textes: c'est ce que nous appelons le glissement de la lexicométrie originelle vers une logométrie pleine et entière, susceptible de renouveler la discipline. Cette amélioration est d'ores et déjà effective pour le français depuis quelques années grâce au développement d'Hyperbase, souvent présenté dans *l'Astrolabe* (Brunet, 2001 et 2003) (3) et qui traite sans difficulté les sorties du lemmatiseur Cordial (4); annonçons simplement ici que ce bond qualitatif pour le français apparaît à ce point porteur (Mayaffre, 2004; Kastberg, 2005) que le logiciel niçois s'articulera au cours de l'année 2005 avec le lemmatiseur polyglotte Tree Tagger (5) pour permettre le traitement logométrique de textes anglo-saxons et romans.

Mais que l'on ne s'y trompe pas cependant: le propos n'est pas de renoncer au traitement lexicométrique sur textes bruts, il est de compléter ce traitement par une analyse complémentaire sur textes lemmatisés. Mieux: cet article voudrait rappeler que le traitement des textes lemmatisés qui ouvre la voie à des analyses grammaticales ou syntaxiques nous semble indispensable, mais qu'il ne peut se faire qu'à condition de garder accès au texte réel, natif, brut, que le locuteur/scripteur a effectivement émis.

Le Tournant de la lemmatisation

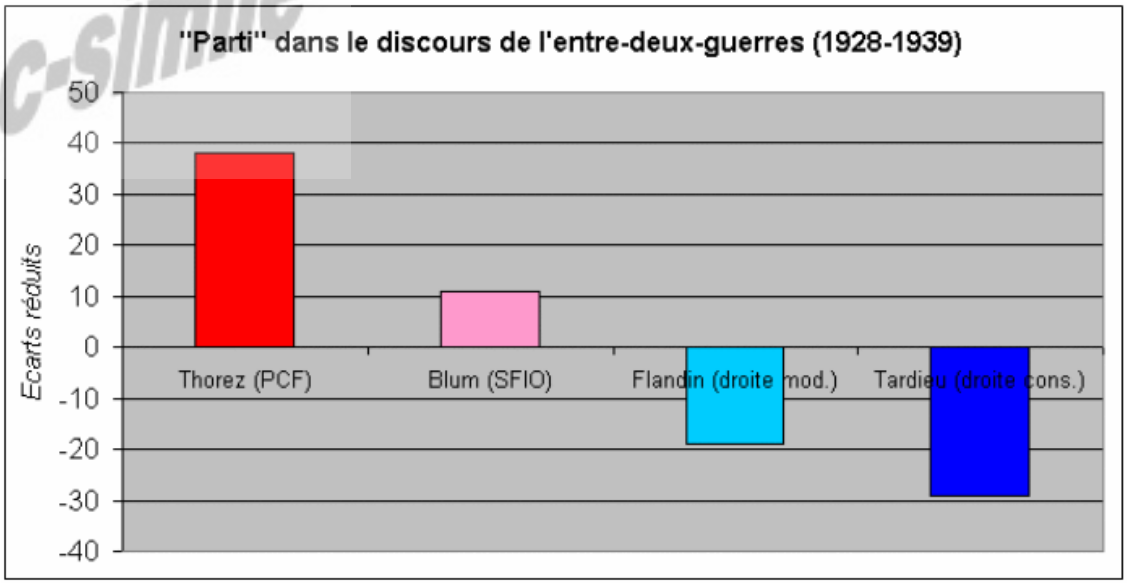
Sans doute est-il superflu de rappeler les limites des traitements lexicométriques sur formes brutes. Depuis la littérature inaugurée par Charles Muller (Muller, 1963), les exemples fourmillent pour illustrer l'aporie linguistique que constitue l'appréhension d'un texte par ses mots pris dans un sens graphique, visuel, matériel, informatique d'*une concaténation de lettres comprise entre deux blancs*. Nous ne ferons ici que survoler un problème rebattu depuis quarante ans pour l'illustrer par quelques exemples massifs pris dans nos propres travaux.

D'évidence, deux opérations, *a minima*, doivent être envisagées pour que l'unité d'indexation et de décompte cesse d'être une unité matérielle, aveugle sémantiquement (le «mot»), pour devenir une unité de sens pertinente linguistiquement (le «lemme»): les dégroupements et les regroupements linguistiques. La première opération - le dégroupement appelé aussi «désambiguïsation» - consiste dans sa plus simple expression, à séparer les homographes pour les rattacher à leur vocable respectif. La seconde opération consiste schématiquement à regrouper sous un lemme unique (en français l'infinitif pour les verbes, le masculin singulier pour les noms, etc.), les formes différentes - classiquement les «flexions» - signifiant la même chose ou se rattachant au même signifiant. *Grosso modo*, la *lemmatisation* d'un texte tient aujourd'hui dans ces deux opérations dont nous ne prétendons pas épuiser le sens (6).

Lemmatiser = dégrouper

Qu'il nous soit permis de donner un exemple de la nécessité de désambiguïser le texte; exemple non pas en marge mais central dans notre thèse (Mayaffre, 2000). Après la saisie laborieuse de plusieurs centaines de discours politiques, après un traitement lexicométrique approfondi, la caractéristique première du discours de gauche *versus* le discours de droite dans l'entre-deux-guerres (1928-1939) nous est enfin apparue: la gauche sur-utilise, comme tout premier mot, le mot «parti», alors que la droite, symétriquement, le sous-utilise (figure 1) (7).

Figure 1: Répartition de la forme «parti» dans le discours politique de l'entre-deux-guerres (comparaison gauche/droite).



Cette conclusion statistique est une conclusion majeure de notre travail (*ibid*.: 65-101): les chiffres ici sont éloquents et il s'agit de la toute première sortie-machine d'un traitement statistique important. Elle permet de proposer des interprétations linguistico-politico-historiques intéressantes. En effet, en France, la gauche, depuis 1789, se construit sur la base d'une identité partisane, là où la droite se construit sur la base d'une identité a-partisane qui prétend se confondre avec la nation, au-delà des classes sociales ou des partis. Au fond, c'est un rapport différent au Politique qui se joue ainsi, la gauche assumant son rôle de parti(e) politique sur l'échiquier, la droite refusant son rôle, déniant par là sa nature. De profondes raisons historiques peuvent être avancées pour expliquer le phénomène. On sait par exemple le complexe politique de la droite sous la République depuis sa prise de position originelle en faveur de la monarchie durant l'été 1789; complexe qui peut se traduire par ce déni linguistique et politique (*ibid*.: 289-292). Nous ne pousserons pas ici plus loin l'analyse, notons simplement qu'aujourd'hui encore, en 2005, il est remarquable de voir à gauche, en France, un *«Parti» communiste* (PCF) et un *«Parti» socialiste* (PSF), alors que la droite ne compte pas de «parti» mais une *«Union» pour un mouvement populaire* (UMP) et une *«Union» pour la démocratie française* (UDF).

Cinq ans après ce travail doctoral, il n'est pas question de remettre en cause l'interprétation donnée, mais l'honnêteté invite à préciser qu'elle repose, tout entière, sur une description linguistique dangereusement bancale du corpus. En effet, dans le traitement lexicométrique effectué, la forme «parti» qui a été indexée et décomptée et sur laquelle repose toute l'analyse, regroupait indifféremment - faute de lemmatisation - le substantif «parti» et le participe passé du verbe partir!

Cet exemple est parlant et il pourrait être généralisé. Selon Dominique Labbé, près d'un tiers de la composition des textes français est homographe et, selon Charles Muller, ce taux varie en fonction des productions mais ne descend jamais au-dessous de 15 %. Refuser la lemmatisation, c'est admettre qu'une bonne partie du traitement quantitatif que l'on voudrait objectif - les chiffres donnent cette impression d'objectivité - compte ensemble torchons et serviettes, au motif qu'ils revêtent la même apparence graphique. Le discrédit est particulièrement important car, là où certaines pratiques se contentent de l'intuition pour analyser les textes, le traitement quantitatif aspire à la froide objectivité: il ne saurait donc se faire sur des unités impertinentes linguistiquement.

Par ailleurs, toujours dans le cadre des *dégroupements*, mais au-delà des homographes, il convient de dire un mot des formes contractées car elles représentent une surface non négligeable des textes. (Effectivement la contraction, qui répond avant tout à une logique économique de la langue, concerne, par définition, des unités très souvent utilisées comme «du» ou «au»; unités fréquentes qui pèsent donc dans les décomptes). Ainsi «du» doit être dégroupé en «de le». Non par luxe bien sûr, mais par souci d'équité car sinon une discrimination linguistique artificielle entre la formule féminine «de la» et la formule masculine «du» serait arbitrairement créée. Dès lors, les conséquences mathématiques seraient automatiques: on trouverait beaucoup plus de «la» dans les textes que de «le» (dont une partie serait fondue dans «du»), nous laissant imaginer à une féminité du discours. De la même manière, il convient de dégrouper «au(x)» en «à le(s)». Et ainsi de suite.

- **Lemmatiser = regrouper**

La pertinence des *regroupements* est moins directement évidente et surtout moins innocente comme nous le

démontrerons plus bas. Mais elle reste difficilement contestable dans certaines de ses tâches élémentaires. «Bel» et «beau» ou «nouvel» et «nouveau» doivent être regroupés dans l'index des formes sous une seule entrée. Ils n'ont pas à être comptés distinctement, sinon leur poids se trouverait divisé par rapport au poids des autres adjectifs tel «laid» ou «ancien» qui ont - du point de vue quantitatif - l'avantage d'avoir une forme unique. «Je» et «j'» ont-ils besoin d'être distingués lorsque les autres pronoms personnels ne souffrent pas de diamorphisme? Et le constat sur «je» et «j'» est valable pour tous les cas d'élision («le» et «l'», «de» et «d'»); cas très fréquents tant ils affectent toujours les mots les plus usités de la langue, ceux dont l'usage répété invite à la distorsion ou la contraction.

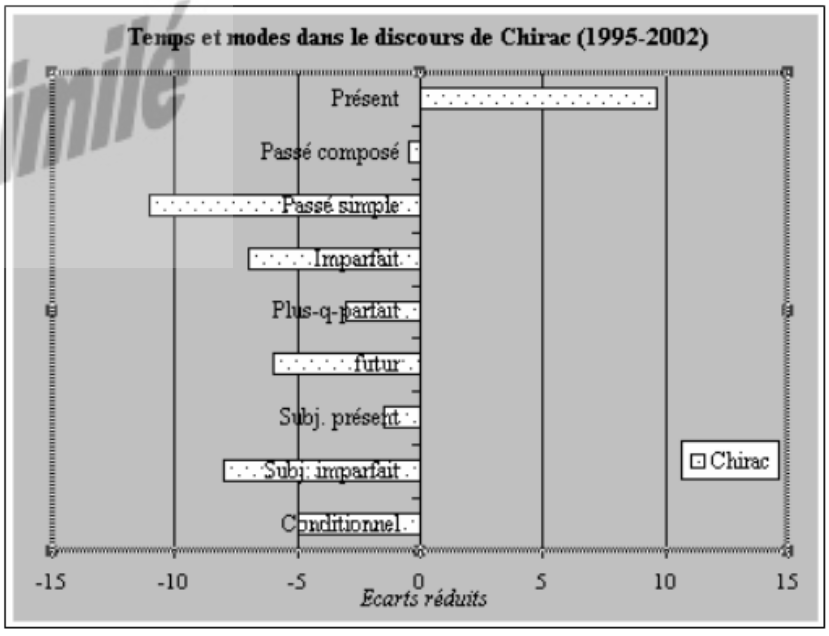
Dit autrement, souvent deux lexies sont strictement synonymes et seules des contraintes morpho-linguistiques expliquent leur diversité graphique: dans ces conditions comment une approche qui prétend traiter objectivement le texte pour en retrouver le sens peut-elle justifier de les considérer séparément, au risque de diluer leur poids ou leur fréquence dans le corpus par rapport à d'autres mots concurrents qui ne connaissent pas de dispersion graphique?

Par ailleurs, le *regroupement* de «clin d'œil» en un seul vocable n'a guère besoin d'être justifié puisque «clin» n'est pas même une entrée des dictionnaires. Compter «clin» indépendamment reviendrait à considérer quelque chose qui n'existe pas sémantiquement. Et ce constat caricatural sur «clin d'œil» doit être généralisé à toutes les lexies composées («chemin de fer», «pomme de terre», «président de la république») qui ne constituent qu'un seul vocable et dont il ne serait pas seulement absurde mais dangereux sémantiquement de compter les différentes composantes. Certes pour les lexicométriciens, le traitement des segments répétés (Salem, 1987) est un moyen de repérer les lexies composées, mais la réponse est imparfaite car, s'il représente un moyen de retrouver les lexies composées, le traitement des segments répétés n'a pas vocation à retrancher ces dernières du dictionnaire pour éviter de créer des entrées non pertinentes dans l'index des formes et le tableau des fréquences.

Plus généralement, enfin et surtout - mais répétons que nous n'entendons pas épuiser le problème ni présenter tous les cas de figure où la lemmatisation en général et les regroupements en particulier paraissent indispensables -, l'ensemble des flexions de la langue - flexions adjectivales, nominales ou verbales avant tout - peuvent être rassemblées. C'est le plus gros travail de la lemmatisation (8). Il apparaît en effet souvent très utile de décompter ensemble, par exemple, toutes les flexions du verbe dire («dis», «dit», «dira», «dirions», etc.), pour mesurer l'emploi de ce verbe chez un locuteur, lorsque le décompte de ses différentes flexions dilue quantitativement son usage dans le texte et que cette dilution, qui plus est, ne se fait pas équitablement entre tous les verbes (je «dis», tu «dis» = une forme / je «suis», tu «es» = deux formes).

Et c'est ici que la lemmatisation des textes est révolutionnaire et que, devenue accessible à tous, elle ouvre la voie à de riches pratiques qui nous paraissent susceptibles de refondre la discipline. Ramener une forme fléchie à son lemme implique une reconnaissance d'informations linguistiques essentielles. Ramener «dira» au verbe dire, c'est savoir (9) que cette graphie dans le texte est un verbe et que celui-ci prend cette forme à *la troisième personne du futur*. Dès lors, les critères quantitatifs - dont nous n'avons plus besoin de souligner le raffinement aujourd'hui - peuvent s'appliquer à ces informations linguistiques essentielles et nous pourrons analyser l'usage statistique des verbes *versus* les noms, mais aussi l'usage des personnes verbales (ici la troisième du singulier), mais encore du temps (ici le futur) dans le corpus. Ainsi, entre autres exemples, l'étude contrastive du discours des différents présidents de la Ve République a permis de montrer que Chirac - par rapport à de Gaulle, Pompidou, Giscard et Mitterrand - conjuguait massivement voire exclusivement son discours au présent de l'indicatif: la communication instantanée, immédiate, dans le *ici et maintenant* du discours est une caractéristique majeure de la parole politique contemporaine qui privilégie le phatique et le performatif sur la narration (figure 2; pour plus de détails: Mayaffre, 2004: 112-120) (10).

Figure 2 : Temps et modes dans le discours de Chirac (1995-2002) par rapport au discours présidentiel (1958-2002).



En bref, la lemmatisation permet d'une part de décompter des unités à la pertinence linguistique plus avérée (les lemmes qui renvoient à des vocables), et d'autre part de compléter le traitement purement lexical traditionnel par un traitement statistique d'autres régularités linguistiques tels les codes grammaticaux, les modes, les temps, le genre, le nombre, etc. Dans ces conditions, il apparaît aujourd'hui difficile de se priver d'un tel enrichissement.

La Forme graphique: pis-aller ou choix théorique?

Pourtant, pendant trente ans, dans le domaine du discours politique, les traitements sont restés, le plus souvent, à strictement parler *lexi*-cométriques. Il s'agissait de traiter le texte brut ou natif, c'est-à-dire des *lexies* au sens de formes graphiques comprises entre deux blancs ou plus précisément entre deux caractères définis par le chercheur, puis par l'ordinateur, comme délimitateurs (les blancs donc mais aussi les ponctuations voire les apostrophes et les tirets). (Pour le détail lire nécessairement: Collectif Saint-Cloud, 1974 et 1985; Lafon, 1984).

Les Limites techniques des années 1980-2000

Deux raisons ont empêché pendant des années d'aller au-delà du texte brut. D'abord, aucun logiciel de lemmatisation ne paraissait suffisamment performant pour ramener un texte fléchi à un standard lemmatisé sans erreurs majeures et une fastidieuse relecture manuelle (11). Toujours lourd dans l'approche lexicométrique, le fardeau de la saisie puis de la mise en forme des textes devenait insupportable pour celui qui entendait traiter des gros corpus et aboutir dans un délai raisonnable à des sorties machine fiables. Aujourd'hui les lemmatiseurs-étiqueteurs affichent, dans leurs fonctions de base (désambiguïsation des homographes, regroupement flexionnel des noms ou des verbes, reconnaissance des catégories grammaticales élémentaires), un taux de réussite proche de 100 %: pour ces fonctions de bases quelques erreurs subsistent mais aucune susceptible de fausser le traitement statistique que nous opérons pas la suite; les vérifications manuelles restent utiles mais non plus obligatoires (12).

Ensuite, aucun logiciel de lexicométrie sur le marché scientifique en France ces dernières années (Lexico, Hyperbase, Weblex, Sphinx, etc.) n'était susceptible de traiter les sorties des lemmatiseurs de manière simple et automatique. Ainsi la chaîne de traitement n'était pas établie. Nous trouvions des lemmatiseurs d'un côté et de l'autre des outils statistiques ou lexicométriques performants, mais l'articulation restait douloureuse pour l'utilisateur non informaticien, et réservée à quelques spécialistes (par exemple Labbé, 1990-b). Dans ce cadre technique limitatif des années 1980-2000, la plupart des chercheurs - à l'exception déjà notée de Charles Muller, Etienne Brunet, Dominique Labbé ou Sylvie Mellet - ont dû se contenter du traitement des textes bruts et, comme souvent en pareilles circonstances, ils eurent tendance à faire d'un pis-aller technique une loi scientifique, d'une limitation pratique une théorie: l'engagement pour une indexation minimale devint en effet la règle absolue, en partie controuvée, mais rarement dépassée de la communauté (Collectif Saint-Cloud, «L'indexation minimale. Plaidoyer pour une non-lemmatisation», ENS de Saint-Cloud; communication au

«Colloque sur l'analyse des corpus linguistiques: Problèmes et méthodes de l'indexation minimale», Strasbourg, 21-23 mai 1973; et Collectif Saint-Cloud, 1985).

Un choix théorique

Pour comprendre cette réalité, il faut ajouter qu'aux obstacles purement matériels guère recevables aujourd'hui, se sont combinées des raisons plus profondes, qu'il convient de ne pas ignorer tant leur acuité, notamment sous la plume de Maurice Tournier, reste d'actualité; et la nouvelle génération de chercheurs amenée sans doute à verser naturellement dans la logométrie sur textes lemmatisés gagnera à reprendre ici quelques contributions fondatrices de la discipline (voir notamment Tournier, 1985 et 1987). Si le traitement lexicométrique fut pendant longtemps un dogme dont le Laboratoire de Saint-Cloud a été le gardien, c'est qu'en effet seule la lexie (ou item formel) objectivement attestée dans le texte est une entrée *fiable* - précisément parce qu'elle est attestée -, *stable* - la convention graphique est bien établie -, et *naturelle* - aucun filtre, traitement ou pré-traitement n'intervient entre le locuteur-scripteur et l'analyseur. Lemmatiser un texte, c'est ramener son vocabulaire à un lexique, au sens où J. Picoche entend ces deux mots, en se proposant «d'appeler *lexique* l'ensemble des mots qu'une langue met à la disposition des locuteurs, et *vocabulaire* l'ensemble des mots utilisés par un locuteur donné dans des circonstances données.» (Picoche, 1977: 45).

Or pour Maurice Tournier et les analystes du discours politique, le lexique n'est qu'une construction souvent politique, arbitraire, historique: «Le mot hors situation n'est qu'une vue de l'esprit» (Tournier, 1987: 5). De la même manière qu'il n'y a pas de Langue en soi, il n'existe pas de lexique en lui-même ou de «*norme de sens*» neutre, universelle, anhistorique (Tournier, 1985: 483) que l'on pourrait plaquer automatiquement aux textes. Le renvoi des unités d'un texte à une entrée canonisée du dictionnaire apparaît même particulièrement dangereux voire définitivement rédhibitoire dans le cadre du discours politique, car celui-ci est présenté, précisément, comme un lieu de tension sémantique, de construction et de reconstruction du sens. Et le travail de l'analyste est *précisément* d'étudier cette construction notamment en articulant le traitement linguistique à un savoir historique et à une approche idéologique. Dès lors, faire précéder l'analyse que l'on se propose de faire, d'une «normalisation» sémantique déjà toute faite revient à donner les conclusions du travail avant même l'analyse! Il y a là une entorse épistémologique dans le procès de la démarche, évidente et insurmontable:

> Il n'est aucun acte de langage [...] qui soit neutre, il n'existe pas de noyau de sens à l'état pur ni de valeurs universelles incarnées dans les mots [...] Voilà pourquoi, il n'y a de «lexique» étudiable pour le lexicométricien, de mot hors-jeu, c'est-à-dire de langue en soi, toute innocente, à laquelle les discours concurrents pourraient faire appel au titre d'arbitre. Tel est le second principe réducteur de la lexicométrie: tout mot en tout texte participe à un acte politique, dès lors qu'il sert à stabiliser ou à déstabiliser ne serait-ce que le langage. L'analyste ne peut renvoyer à aucune norme extérieure au vocabulaire des textes étudiés et doit en conséquence s'en tenir à des corpus clos. (Tournier, 1987: 5).

Puis d'autres arguments viennent s'ajouter. Désambiguïser la forme «parti» en verbe et en substantif apparaît nécessaire et élémentaire, disions-nous. Mais qu'aura-t-on fait au juste? Dans notre quête du sens, rien peut-être d'important en donnant pourtant l'impression dangereuse d'avoir fait l'essentiel. «*Il prit le parti de ne rien dire*» et «*il adhéra au parti*»: dans les deux cas «parti» renvoie au substantif (*versus* le verbe), mais sans signifier pourtant la même chose. Plus subtilement: «*le parti communiste*» et «*le parti des crève-la-faim*» renvoient sans aucun doute au même sème général (regroupement de personnes porteuses d'une revendication) mais dans un premier cas nous avons affaire à une forme politique institutionnelle, dûment établie, clairement organisée, dotée d'une structure, d'une histoire, d'un programme, d'une idéologie; dans l'autre cas, nous avons, au fond, tout l'inverse, un magma politique dont la principale caractéristique est de n'être pas organisé jusqu'à son impuissance politique. Alors faut-il créer (et décompter) deux vocables? Si oui, qui décidera par exemple si le «Parti» radical-socialiste, au contour si flou dans l'entre-deux-guerres, renvoie au magma ou à la forme organisée?

Le danger de la lemmatisation est là: avec elle, nous nous imaginons sauvés de la noyade linguistique, alors même que, se croyant à l'abri, l'on reste dangereusement au milieu du guet. «La lemmatisation ne résout rien et empire tout» (Tournier, 1985: 485) conclut Maurice Tournier non sans pertinence: elle ne résout pas la question du sens - indécidable hors contexte et surtout pas par une machine - et travestit déjà le texte effectivement émis que l'on se proposait d'étudier. Reprenons notre exemple pour forcer le trait: «*il faudra consulter les partis politiques*», et «*le Parti a décidé*» (que nous écrivons avec une majuscule pour mieux nous faire comprendre). Avec la lemmatisation, pluriel et singulier seront ramenés à un lemme unique, qui deviendra la seule entrée dans l'index et la table des fréquences: c'est là l'objectif même de la lemmatisation. Pourtant, ici, la principale information sémantique tient dans l'usage différent du nombre. Dans le premier cas, nous aurons sans doute affaire à la description d'un système politique républicain qui compte nécessairement plusieurs partis parlementaires, dans le second cas, nous aurons peut-être affaire à un système totalitaire qui compte un seul

parti, un parti-famille ou parti-caserne, monolithique, quasiment déifié. Loin d'aider à attraper le sens politique du texte, la lemmatisation, ici, en complique l'accès. Et les tenants des formes graphiques savent convaincre en multipliant les exemples où l'usage du pluriel et du singulier - imprudemment amalgamés dans la lemmatisation - font sens et n'ont pas à être confondus: «libertés» et «Liberté», «peuples» et «Peuple», «démocraties» et «Démocratie», etc.

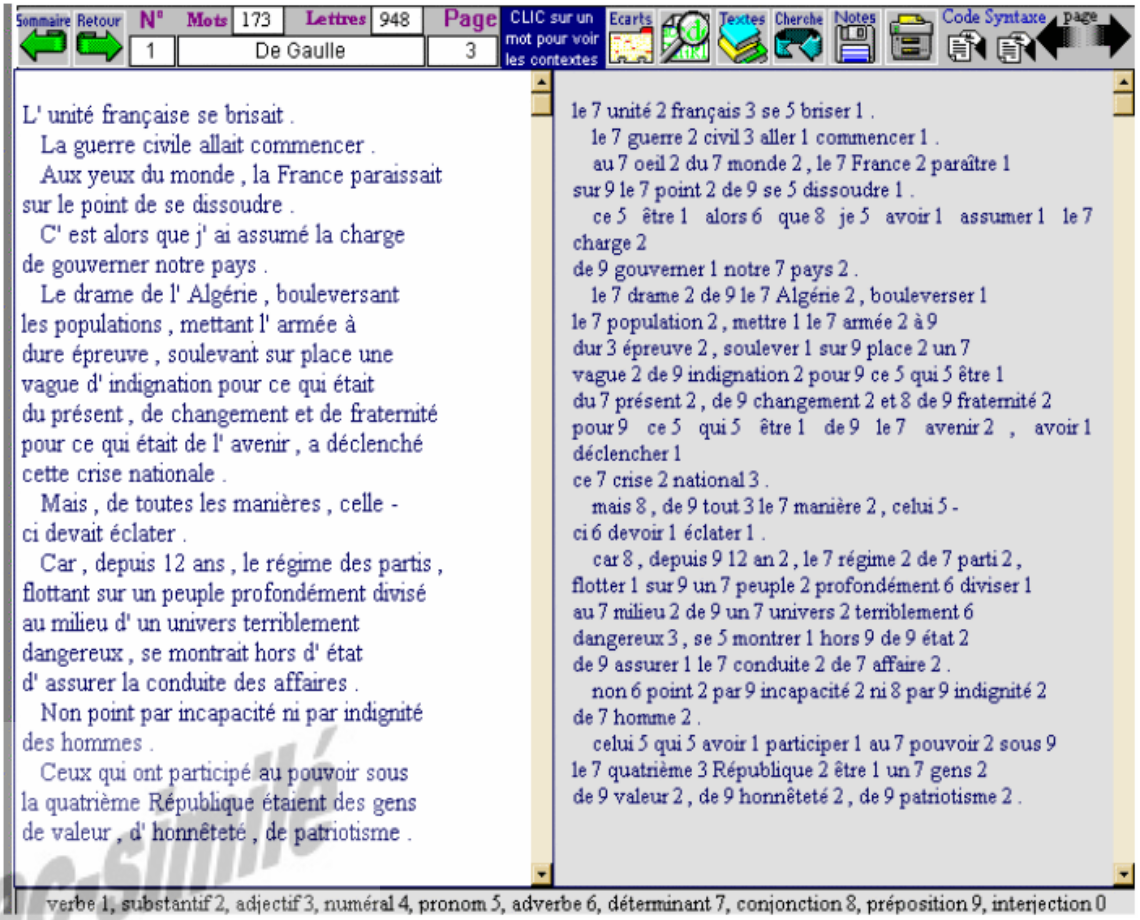
Pour les formalistes donc, lemmatiser un texte est périlleux et parfois contre-productif. Surtout, c'est aller à l'encontre même de l'approche lexicométrique qui entend déconstruire le plus objectivement possible un texte pour accéder à son sens. Pour Tournier, la philosophie de la lexicométrie politique se trouve à l'exact opposé de la logique dictionnairique que reprend la lemmatisation. Le discours est l'endroit où la langue est travaillée par l'idéologie, où les mots sont enjeux ou leur sens est en jeu, là où le dictionnaire, qui canonise le sens, enregistre de manière figée et naïve ce travail. Dans ces conditions, le lexicométricien ne peut faire référence à aucun dictionnaire sans remettre en question le fondement même de son travail.

De la logométrie

A reprendre ainsi l'ensemble du débat entre lemmatiseurs et formalistes, notre conclusion apparaît plus simple qu'un jugement de Salomon. Pour Tournier qui opte pour la forme graphique, l'alternative serait forcément cruelle et castratrice: «Oui je l'avoue, les arguments se contrebalancent. L'un (*le lemmatiseur*) crie à la trahison de la langue, l'autre (*le formaliste*) à la trahison du texte [...] Avouons-le ensemble une bonne fois, il y a effectivement trahison des deux côtés.» (Tournier: 1985, p. 487). Pour Muller, qui opte pour le texte lemmatisé, le choix résolu en faveur de la lemmatisation apparaît comme nécessaire mais pécheur: «J'ai dit les objections purement linguistiques qui me font préférer les péchés de la lemmatisation à la pureté du formalisme.» (Muller: 1984: IX). Dès lors le chemin paraît balisé. Il est tout aussi impossible de renoncer aux formes graphiques, que de s'en tenir à elles; impossible de se priver des richesses de la lemmatisation - notamment en ce qu'elle permet l'étude des caractères grammaticaux des textes - que d'oublier le texte natif. Il ne peut s'agir de «renvoyer les disputeurs dos à dos» (Tournier, 1985: 487), mais de les faire se tenir côte à côte, main dans la main.

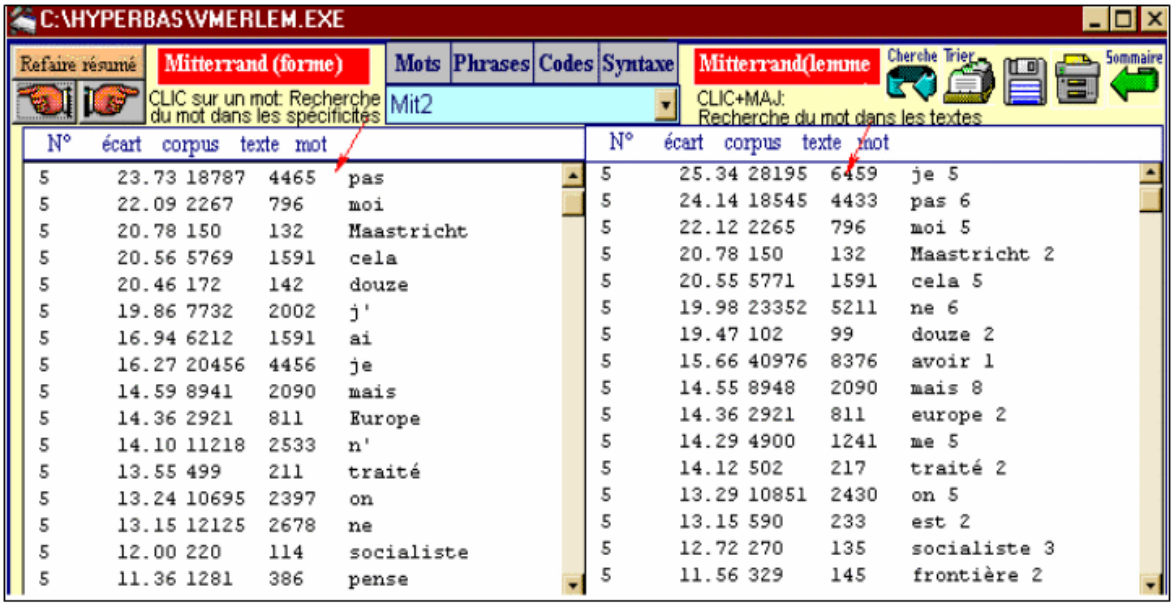
Evidemment cette conclusion apparaîtrait angélique si elle n'était pas techniquement possible et d'ores et déjà effective. Hyperbase s'applique aujourd'hui dans toutes ses fonctionnalités à décliner ses traitements simultanément sur le texte brut et sur le texte lemmatisé. L'ensemble de l'ergonomie du logiciel a été conçu à cet effet (lire nécessairement ici même, Brunet, 2003). Dès l'activation du bouton «Lecture» qui permet la consultation du texte du corpus, la fenêtre de l'écran est divisée et l'on lira en parallèle le texte brut et le texte lemmatisé (figure 3).

Figure 3: Lecture parallèle du texte brut (partie gauche) et du texte lemmatisé (partie droite) dans Hyperbase (ici allocution de de Gaulle, le 13 juin 1958).



Bien sûr le bouton «Spécificités», essentiel dans le traitement statistique des textes, donne à voir dans la colonne de gauche les lexies spécifiques et dans la colonne de droite les lemmes spécifiques - suivis de leur code grammatical (figure 4).

Figure4: Lecture parallèle des formes spécifiques et des lemmes spécifiques dans Hyperbase (ici les spécificités de Mitterrand par rapport au discours présidentiel moyen (1958-2002)).



La distance intertextuelle, les recherches thématiques, les AFC, les analyses arborées seront calculées ou construites parallèlement selon les formes graphiques, les lemmes ou les catégories grammaticales. Les fonctions documentaires traditionnelles en lexicométrie, comme la recherche de concordances, de contextes, de

co-occurrences, etc., fonctionneront aussi, indifféremment, sur le texte natif et sur le texte transformé et étiqueté.

A vrai dire, aujourd'hui, il n'est plus question de faire un choix exclusif entre formes et lemmes mais de savoir comment organiser techniquement les différentes entrées dans le corpus pour combiner les différents points de vue sur le texte. Hyperbase a choisi une étude parallèle en juxtaposant les analyses et en adaptant, à moindre frais, la statistique *lexicale* traditionnelle à une statistique *grammaticale* et *syntaxique*. Mais d'autres voies sont possibles pour croiser ces différents niveaux de traitement, et demandent à être explorées: il s'agit là d'un des principaux programmes de l'analyse des données textuelles contemporaine (Pincemin, 2004).

Conclusion

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.

Un texte est un tout dont le fonctionnement linguistique est complexe; son traitement ne saurait se borner à un seul point de vue. Finalement, le traitement logométrique cherche à reproduire autant que possible l'indescriptible acte de lire, qui conjugue l'appréhension du matériel graphique et sa complexe compréhension linguistique, sans jamais les faire divorcer.

Notes

1 - On notera néanmoins la création par André Salem en 1997 de la revue *Lexicometrica* (http://www.cavi.univ-paris3.fr/lexicometrica/) dédiée entièrement à l'analyse des données textuelles et particulièrement aux traitements lexicométriques de textes politiques.

2 - Ici, on notera la création en 2000 de *l'Astrolabe* qui se fixe comme objectif d'explorer les relations entre littérature et informatique (http://www.uottawa.ca/academic/arts/astrolabe/presentation.htm).

3 - Hyperbase est conçu et développé par Etienne Brunet et produit par l'UMR 6039 Bases, Corpus et Langage (CNRS-Université de Nice).

4 - Cordial est produit et développé par Synapse Développement (http://www.synapse-fr.com/).

5 - Tree Tagger est développé par Helmut Schmid de l'Université de Stuttgart (http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html).

6 - Pour le détail de ces opérations, consulter Muller, 1977: 3-46, et Labbé, 1990.

7 - Le graphique se lit ainsi: par rapport à la norme du corpus (composé de nombreux discours politiques de l'entre-deux-guerres), les deux locuteurs de gauche sur-utilisent «parti» à hauteur d'un écart réduit de +39 pour Thorez (PCF) et de +11 pour Blum (SFIO); et les deux locuteurs de droite sous-utilisent la forme à hauteur de –19 pour Flandin (droite modérée) et de –29 pour Tardieu (droite conservatrice).

8 - Il s'agit même de son sens restreint originel. Au départ, «lemmatiser» un texte consistait avant tout à ramener les formes fléchies à leur entrée canonique du dictionnaire.

9 - Nous n'entrerons pas ici dans la compréhension de ce savoir; la plupart des lemmatiseurs se réfère à un dictionnaire (par exemple Cordial), mais d'autres utilisent une approche distributionnelle ou statistique (par exemple Tree Tagger) pour aborder le texte.

10 - Le graphique se lit ainsi: par rapport à l'usage présidentiel moyen sous la Ve République (de Gaulle, Pompidou, Giscard, Mitterrand, Chirac entre 1958 et 2002), Jacques Chirac sur-utilise le présent de l'indicatif à

hauteur d'un écart réduit de + 9; en revanche il sous-utilise tous les autres temps ou modes (passé composé = -1; passé simple = -12; etc.)

11 - C'est ce que confesse Maurice Tournier en 1987: «[La lexicométrie sur textes lemmatisés] ne permet pas l'étude de gros corpus, car elle exige en préalable à tout traitement et dans l'absence d'analyseurs automatiques ultra-perfectionnés, un énorme investissement en pré-analyses de toutes sortes, décryptant pas à pas les énoncés et faisant appel pour chaque désambiguïsation à une multiplicité de modèles linguistiques [...]» (Tournier, 1987: 2). Précisément, notre propos est de montrer que ces «analyseurs automatiques ultra-perfectionnés» qui faisaient défaut dans les années 1980 sont désormais disponibles.

12 - Des progrès restent évidemment à faire, notamment dans la reconnaissance des *fonctions* grammaticales. Plus loin, Cordial prétend reconnaître les isotopies sémantiques mais sans y arriver.

Références bibliographiques

E. Brunet (1981), *Le Vocabulaire français de 1789 à nos jours d'après les données du «Trésor de la Langue Française»*. Vol. I: 852 p.; vol. II: 518 p.; vol. III: 453 p. Genève-Paris: Slatkine-Champion.
-(2000), «Qui lemmatise dilemme attise», *Lexicometrica*, no 2,
http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.html
-(2001), «Le Logiciel Hyperbase», *L'Astrolabe*,
http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm
-(2002), «Le Lemme comme on l'aime», *in* Morin A. et Sébillot P. (éd.), *JADT 2002, 6e Journées internationales d'analyse des données textuelles*, Rennes, IRISA,, vol. 1, p. 221-232.
-(2003), «Statistique et lemmatisation. L'exemple de Rabelais», *L'Astrolabe*,
http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm

R. Busa (1998), «Dernières réflexions sur la statistique textuelle», *in* S. Mellet (éd.), *JADT 1998, 4e Journées internationales d'analyse des données textuelles*, UNSA-CNRS, Nice, p. 179-183.

Collectif Saint-Cloud (A. Geffroy, P. Lafon, M. Tournier) (1974), «L'Indexation minimale. Plaidoyer pour une non-lemmatisation», ENS de Saint-Cloud. Communication au «Colloque sur l'analyse des corpus linguistiques: Problèmes et méthodes de l'indexation minimale», Strasbourg, 21-23 mai 1973.

Collectif Saint-Cloud (P. Lafon, J. Lefèvre, A. Salem, M. Tournier) (1985), *Le Machinal. Principes d'enregistrement informatique des textes*, Paris, Klincksieck.

P. Guiraud (1954), *Les Caractères statistiques du vocabulaire*, Paris, PUF.

M. Hug (2002), «Désambiguïsation automatique d'homographes verbe/nom», in Morin A. et Sébillot P. (éd.), *JADT 2002, 6e Journées internationales d'analyse des données textuelles*, Rennes, IRISA, vol. 1, p. 371-379.

M. Kastberg Sjöblom (2005), *L'Ecriture de J.M.G. Le Clézio - Des mots aux thèmes*, Paris, Honoré Champion (à paraître).

D. Labbé (1990-a), *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahiers du CERAT.
-(1990-b), *Le Vocabulaire de François Mitterrand*, Paris, Presses de la FNSP.
-(2002), «La Lemmatisation des grandes bases de textes», *L'Astrolabe*,
http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm

P. Lafon (1984), *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion, 1984 (avec préface de Charles Muller).

D. Longrée (2002), «Spécificités stylistiques et distributions temporelles chez les historiens latins: sur les méthodes d'analyse quantitative d'un corpus lemmatisé», communication présentée au colloque «2e Journées de la linguistique de corpus», Lorient, 12 au 14 septembre.

D. Mayaffre (2000), *Le Poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres (1928-1939)*, Paris, Champion.
-(2004), *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République*, Paris, Champion.

S. Mellet (2003), «Lemmatisation et encodage grammatical: un luxe inutile?», *Lexicometrica*, numéro spécial http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1/spec1-texte2.htm

S. Mellet et G. Purnelle (2002), «Les Atouts multiples de la lemmatisation: l'exemple du latin», in Morin A. et Sébillot P. (éd.), *JADT 2002, 6e Journées internationales d'analyse des données textuelles*, Rennes, IRISA, vol 2, p. 529-538.

Ch. Muller (1963), «Le Mot, unité de texte et unité de lexique en statistique lexicologique», *Travaux de linguistique et de littérature*, Université de Strasbourg, II, 1, p. 155-175.
- (1967), *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse (réédition: Genève-Paris, Slatkine-Champion, 1979).
-(1968), *Initiation à la statistique linguistique*, Paris, Larousse (particulièrement les pages 133-155).
-(1977), *Principes et méthodes de statistique lexicale*, Paris, Hachette université (particulièrement les pages 3-46).
- (1984), «De la lemmatisation», préface à Lafon P., *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion.

J. Picoche (1977), *Précis de lexicologie française*, Paris, Nathan.

B. Pincemin (2004), «Lexicométrie sur corpus étiquetés», in *Le Poids des mots, JADT 04*, Louvain, UCL-Presses universitaire de Louvain, vol. 2, p. 865-874.

A. Salem (1987), *Pratique des segments répétés. Essai de statistique textuelle*, Paris, Klincksieck.

M. Tournier (1985), «Sur quoi pouvons-nous compter? Réponse à Charles Muller», *Etudes de philologie et de linguistique offertes à Hélène Nais*, *Verbum* (numéro spécial), Presses universitaires de Nançy.
- (1987), *La Réduction: principe de lexicométrie politique*, brochure de URL «Lexicométrie et textes politiques», 14 pages.

2005

Voir dans l'encyclopédie de *l'Astrolabe*:

L'Analyse par tableaux. II – Applications à la poésie

L'Herméneutique numérique

La Lemmatisation des grandes bases de textes

Le Logiciel Hyperbase

Statistique et lemmatisation