L'Herméneutique numérique

Damon Mayaffre

L'Herméneutique numérique

de Damon Mayaffre

Centre national de la recherche scientifique, UMR 6039 (Nice, France)

L’analyse de textes assistée par ordinateur (ADT) n’a plus à présenter les fondements de sa méthode. Depuis plus de trente ans maintenant, théoriciens et praticiens l’ont abondement fait. Ouvrages spécialisés, monographies, actes de colloques ou articles témoignent du souci scientifique pointu et sans cesse renouvelé tant d’un point vue linguistique, statistique ou informatique des fondateurs de l'ADT. Depuis 1990, la tenue de journées internationales à Barcelone, Montpellier, Rome, Nice, Lausanne, Saint-Malo, a permis de faire le point sur les progressions rapides d'une discipline d'avenir (1). Ici même dans *l'Astrolabe*, les longs articles d’Etienne Brunet rappellent les tenants et les aboutissants des traitements automatisés et quantitatifs des textes.

A vrai dire, toutes les objections ont déjà été adressées et toutes ont été dépassées. La plus importante limite est en train d’être levée. Elle concernait la vanité des traitements (statistiques ou non), portant sur des corpus bruts; traitements accusés de désincarner les textes en travaillant sur des unités (le mot graphique) sans réelle pertinence linguistique. Mais aujourd’hui, pour le français, un logiciel de lexicométrie comme Hyperbase associé au lemmatiseur Cordial permet non seulement d’analyser la surface matérielle ou graphique du texte mais encore de traiter du texte lemmatisé, de ses codes grammaticaux, de ses combinaisons et enchaînements syntaxiques voire de ses isotopies sémantiques. Tant et si bien que le monde étroit et contestable de la *lexico-métrie* de la première génération semble dépassé pour s’ouvrir sur l’univers nouveau d’une *logo-matique* pleine et entière qui embrasse le texte, *qualitativement et quantitativement*, dans toutes ses dimensions linguistiques, *lexicale, grammaticale, syntaxique, sémantique, rhétorique*.

L’heure est donc plutôt au bilan critique et pragmatique, c’est-à-dire à une réflexion sur les apports épistémologiques réels, objectivement constatés, de l’ordinateur pour les différentes sciences humaines dans leur rapport fondamental aux textes. Nous le ferons du point de vue de l’historien qui, comme le littéraire, se nourrit de corpus textuels (2), et dont la pratique quotidienne, l’herméneutique, se trouve revisitée par la révolution numérique. Ce bilan épistémologique se fera non pas *in abstracto* mais sur la foi d’une analyse d’un important corpus de discours politiques français de l'entre-deux-guerres (832 discours, 1.600.000 mots, représentant une cinquantaine de livres de poche), entreprise durant cinq années dans le cadre d’une thèse de l’Université française. Il se trouve aujourd'hui confirmé, par l'étude d'un corpus plus vaste encore des discours des présidents de la Ve République française (1958-2002, 3000 discours, 6.000.000 mots) dont les premiers résultats sont à paraître.

L'Ordinateur: un outil probatoire

Face à un corpus textuel, pour peu que celui-ci soit un peu important, l'ordinateur se révèle une aide précieuse là où l'œil et la mémoire de l'homme touchent à leur limite. Dans le cadre des démarches hypothético-déductives qui dominent les sciences humaines, il devient un outil probatoire intéressant. Le chercheur s'en servira en effet avant tout pour mettre à l'épreuve ses hypothèses. La *philologie numérique* permettra d'entrer avec souplesse dans une grande masse de textes, de la lire et la relire, de la décrire de manière contrôlée et systématique, amenant la rigueur qui souvent fait défaut dans les sciences molles, pour convaincre de la pertinence d'une thèse et pour en administrer irréfutablement la preuve.

Lire et naviguer

Les corpus textuels électroniques traités par les logiciels d'ADT proposent d'abord de nouveaux parcours de lecture qui enrichissent la trajectoire linaire habituelle de la lecture naturelle des textes. En effet, le premier mouvement - celui qui autorise ensuite tous les autres - des logiciels français comme Hyperbase, Lexico , Sphinx, consiste à indexer, sans exclusive, la totalité des unités linguistiques jugées pertinentes (une chaîne de lettres, le mot, le lemme, le segment, la catégorie grammaticale, etc.)

Partant, il devient loisible aux chercheurs de repérer ces unités, de les trier, de les compter et, pour ce qui nous concerne ici, simplement, *de les lire*, soit par navigation hypertextuelle en nous projetant par simple clic de souris dans leur texte d'origine, soit par convocation des phrases ou des paragraphes dans lesquels elles figurent, sous forme de concordanciers exhaustifs aisément consultables.

La lecture naturelle, au fil continu du texte, reste évidemment toujours possible, mais les facilités d'une lecture électronique sont le plus souvent appréciées (3). Par sauts de puce donc de texte en texte, on se reportera ainsi à tous les passages contenant «patrie» dans les discours de François Mitterrand, exactement - la rapidité en plus - comme l'on entre parfois dans un ouvrage papier *via* l'index des noms propres. De la même manière, par une compilation exhaustive, on consultera la liste de tous les paragraphes contenant un verbe performatif conjugué au présent dans les discours de De Gaulle, pour une lecture systématique des contextes d'utilisation de ces verbes.

Répétons-le, ce qui est remarquable ici - outre la rapidité - c'est que l'indexation du texte n'est pas réduite, comme dans un livre, aux noms de lieu ou aux noms propres: elle est *intégrale* (c'est-à-dire sans *a priori)* permettant une navigation à l'infini, n'interdisant rien et autorisant tous les parcours de lecture imaginables jusqu'à épuisement du champ des possibles: certains chercheurs pourront aborder le texte par l'étude de tel mot («patrie» ou «nation», «guerre» ou «paix», «libéralisme» ou «communisme»), d'autres par celle des catégories grammaticales, les verbes ou les noms, les adverbes ou les déterminants... Tel chercheur ne pensait devoir étudier que les noms suffixés en -isme et s'aperçoit en cours d'analyse de la pertinence d'étudier les adjectifs suffixés en -iste. Celui-ci encore veut consulter la troisième personne singulier des verbes conjugués au futur; celui-là consultera seulement (mais tous) les subjonctifs, etc.

Comprendre et interpréter

Mais lire c'est aussi comprendre, c'est-à-dire, selon François Rastier, interpréter (4): la lecture hypertextuelle et les concordanciers modifient notre rapport sémantique au texte. Si les sémanticiens reconnaissent volontiers qu'un mot n'a pas à proprement parler de sens mais seulement des emplois, ils n'ont pas toujours eu les moyens matériels de compiler ces emplois. A grande échelle, seul le *Trésor de la langue française* (15 volumes, Editions du CNRS puis Gallimard, 1971-1994) a entrepris un vaste dictionnaire qui n'a plus la prétention de définir le sens littéral des mots mais d'en montrer les usages, grâce aux immenses ressources de la base de données de textes, *Frantext*, composée de la plupart des oeuvres littéraires françaises du XVIe siècle à nos jours (5).

Plus modestement, mais au quotidien, la démarche du *TLF* peut être reproduite au sein de n'importe quel corpus numérique. Ainsi sur notre corpus de textes politiques et dans le cadre de la sémantique historique ou tout simplement d'une compréhension de l'entre-deux-guerres, l'utilisation - c'est-à-

dire le *sens attesté* - de «fasciste» dans le discours politique français a été étudiée (6). Dans une masse de textes représentant cinquante livres de poche, toutes les phrases - sans oubli possible - contenant le mot ont été extraites. Et la consultation systématique des contextes des onze occurrences dans les discours de Léon Blum pour l'année 1934, par exemple, est instructive (illustration 1).

Figure 1. Contexte de «fasciste» dans le discours politique français (L. Blum, 1934)

-Interprètes de la volonté du peuple, nous sommes résolus, sur le terrain parlementaire comme sur tous les autres, à barrer la route à l'offensive outrageante de la réaction FASCISTE (Blum 1934, occ. 1)

-La réaction FASCISTE ne passera pas. (Blum, 1934, occ. 2)

-Les deux commissions d'enquête vont prendre à leur tour des vacances, normales cette fois et légitimes. Le caractère véritable de l'émeute FASCISTE est définitivement démontré et l'ensemble des dépositions recueillies corrobore avec la clarté la plus puissante des conclusions auxquelles nous étions parvenus. (Blum, 1934, occ. 3)

-Installé au pouvoir par l'émeute FASCISTE, le gouvernement suspend ainsi le retour de l'émeute sur ceux qui oseraient entraver son œuvre. (Blum, 1934, occ. 4)

-Le Bloc National ou l'Emeute FASCISTE, voilà le choix dans lequel Doumergue enferme la France (Blum, 1934, occ. 5)

-Le pays veut la République contre l'émeute FASCISTE (Blum, 1934, occ. 6)

-Le désir d'unité, la volonté d'unité, latents sans doute depuis de longues années, sont apparus brusquement à la surface sous le choc de l'agression FASCISTE du 6 février (Blum, 1934, occ. 7)

-Et ces messieurs du Tiers-parti nous expliquent en effet que la principale vertu de l'opération qu'ils préconisent sera de soustraire les bons Français à l'obligation de choisir entre le bloc purement réactionnaire ou FASCISTE –représenté jusqu'à nouvel ordre par le gouvernement et la majorité actuels- et un bloc révolutionnaire (Blum, 1934, occ. 8)

-Mais qu'est donc la journée du 12 février sinon la mobilisation spontanée de la France républicaine et socialiste contre l'attentat FASCISTE du 6 février (Blum, 1934, occ. 9)

-Et, par conséquent, le mouvement républicain du 12 février n'ayant été qu'une riposte et un coup d'arrêt à l'émeute FASCISTE, ce sont les amis et les adversaires de l'émeute fasciste qui… (Blum, 1934, occ 10)

-Ainsi l'émeute FASCISTE est devenue comme le bras séculier auquel recourt le gouvernement légal de la République (Blum 1934, occ. 11)

D'évidence le mot n'est pas utilisé dans son acception actuelle. Et il paraît certain que le leader socialiste ne se fait pas, en 1934, une claire idée de ce que représente le danger fasciste. Le fascisme, c'est dans un cadre franco-français, *grosso modo*, «la réaction» (la «*réaction fasciste*») incarnée par le pâle gouvernement Doumergue après la manifestation (qualifiée «d'émeute») du 6 février 1934: nous sommes loin d'une théorie du totalitarisme.

De là, il est par exemple possible de réévaluer le concept de Front populaire défini comme un rassemblement des forces de gauche contre le fascisme mais en réalité, pour Blum, simple Cartel des gauches contre la Droite conservatrice. De là encore, par un élargissement chronologique de l'étude,

et dans une approche diachronique, il serait intéressant de suivre, pas à pas, l'évolution de l'utilisation de la notion chez Blum entre 1934 et 1940, et de manière plus générale chez les hommes politiques français, des années 20 dans un sens initial et naïf («la réaction»), jusqu'à l'après-guerre dans l'acception actuelle du mot (le totalitarisme, l'impérialisme, l'antisémistisme…), en passant par un infléchissement qu'il devient intéressant de dater (seulement après la guerre? dès la guerre d'Espagne? en 1938?)

Pour finir, précisons encore que les concordanciers rendent facile la nécessaire mise en comparaison systématique de l'utilisation que fait un auteur (ici Blum) avec celle que font d'autres auteurs (Thorez pour le PCF par exemple (Mayaffre, 2000, p. 407-483) à une même époque: s'il n'est pas rare que les hommes politiques utilisent les mêmes mots, la plupart des différends politiques résident dans l'utilisation fine que chacun peut en faire pour en donner un sens différent et justifier des actions opposées.

Au fond, la démarche et le questionnement que nous venons de détailler sont très classiques (lire, contrôler, comparer, comprendre, interpréter, et, en Histoire, éviter les anachronismes sémantiques). Seulement ici, l'ordinateur nous donne, grâce à un relevé exhaustif et une lecture systématique, les moyens certains de nos ambitions scientifiques.

Tester, attester, déduire

L'ordinateur s'avère ainsi un puissant outil documentaire, de dépouillement et d'exploration. C'est pourquoi, il peut faire merveille dans une démarche hypothético-déductive et trouve sa pleine dimension dans une fonction probatoire. Il permet en fait de tester des hypothèses de travail et d'attester de leur validité en contrôlant la présence, l'absence, l'importance ou encore le sens d'un mot, d'un lemme, d'un thème.

Dans la dernière monographie sur Léon Blum, son biographe, voulant absolument dresser le portrait d'un chef charismatique, affirme à propos du dirigeant socialiste: «*Il a le ton de la conviction: il sait, il est persuadé, il a la preuve*» (7). Affirmation (gratuite) que nous nous sommes permis de tester et qui s'est avérée péremptoire. En effet, le relevé exhaustif et impartial que fait l'ordinateur du vocabulaire (particulièrement, le relevé des verbes) prouve absolument l'inverse. La rhétorique de Blum est précisément celle de l'hypothétique, celle du doute dans laquelle les «*je crois*», les «*je pense*», ou les «*j'espère*» l'emportent statistiquement sur toute forme de conviction clairement affirmée. Il s'agit même là de la première caractéristique du discours blumien par rapport à celui d'autres locuteurs: la force de sa rhétorique est d'exposer en pleine lumière les doutes voire la faiblesse du locuteur ajoutant ainsi au pathos d'un discours dominé par l'affectif (Mayaffre 2000, p. 166-219) (8).

Dans un gros corpus, les impressions linguistiques, au fil de la lecture, sont souvent trompeuses; comme les impressions météorologiques le sont au fil des années. Les littéraires (pas plus que les météorologues) ne peuvent éternellement se satisfaire d'à-peu-près; ils doivent se donner les moyens de prouver ce qu'ils avancent. Dans une grosse masse de textes, une lecture superficielle et orientée aura toujours les moyens de trouver un exemple pour illustrer n'importe quelle thèse: il y a bien des «j'affirme» dans le discours de Blum; seulement ils sont rares. Pire: après qu'un auteur a concentré son attention (et attiré celle des lecteurs) sur un phénomène de discours, il y a de fortes chances qu'il remarque ledit phénomène et s'y arrête (et le lecteur à sa suite), fut-il minoritaire. Nous touchons là au cercle vicieux des conclusions artefactuelles qui encombrent, avec la force des fausses évidences, nos études littéraires.

Bref, dans un grand corpus, seul l'ordinateur pourra contrôler (ici, en l'occurrence, mesurer) le degré de significativité d'un phénomène et pourra garantir la représentativité d'un exemple. Au passage, c'est l'usage des citations qui se trouve ainsi corrigé. Dans les études littéraires, la citation prend trop souvent, comme le dénonce le grand historien français Antoine Prost (9), une dimension probatoire, lorsque sa valeur est seulement illustrative. Une citation peut illustrer une thèse non la démontrer. Un extrait, en lui-même, à lui seul, ne prouve rien - si ce n'est la présence ponctuelle d'un phénomène -, seule sa représentativité compte. Et comment, par une lecture naturelle ou intuitive, garantir qu'une phrase est représentative dans un corpus de plusieurs centaines de milliers de phrases?

Les vertus de l'ordinateur jusqu'ici évoquées (lire, tester, attester et par là déjà comprendre et interpréter) paraissent aussi simples qu'indispensables. *A minima*, il s'agit de doubler la lecture flottante habituelle par une lecture assistée. Comme assistant, l'ordinateur fait preuve à la fois de rigueur et de souplesse. Rigueur par l'exhaustivité et la systématicité de l'indexation, donc des explorations, donc des relevés d'information. Souplesse car l'ordinateur peut balayer le texte en quelques secondes, avancer et revenir en arrière sans se lasser, surfer sur la vague d'informations sans se laisser laminer par elle. Et notons que ce gain de souplesse apparaît aussi important que celui que le *codex* a donné, au début de l'ère chrétienne, au *volumen* (10): là où il fallait dérouler l'ensemble du rouleau pour trouver une information, il a été possible de feuilleter le codex pour reprendre tel ou tel passage. Mais là où il est devenu (et reste) possible de feuilleter à l'aveuglette un livre, nous pouvons aujourd'hui repérer n'importe quel passage et nous y reporter directement, par un index exhaustif, tel un marque page généralisé, *via* d'innombrables clefs d'accès (chaîne de caractères, mot, lemme, expression, cooccurrence, catégorie grammaticale, fonction grammaticale, bi-code syntaxique, tri-code…)

Lecture assistée donc, lecture contrôlée aussi pourrait-on dire, tant ce qui contraste entre la lecture numérique et la lecture naturelle, c'est le contrôle de la démarche grâce à la médiation que l'ordinateur apporte ou impose entre soi et le texte c'est-à-dire entre le chercheur et le sens. Pourtant, jusqu'ici, cette nouvelle philologie numérique ne révolutionne rien. Elle systématise des procédures traditionnelles et les rend possibles à grande échelle, mais ne modifie pas au fond l'herméneutique classique. Il est en revanche une deuxième dimension de l'ADT, beaucoup plus ambitieuse, que l'on se propose maintenant d'aborder: sa vocation heuristique.

L'Ordinateur: un outil heuristique

Se donner les moyens d'explorer le texte pour trouver ce que l'on recherche ou ce que l'on pressent est utile. Se laisser interpeller par lui pour découvrir ses éléments saillants que l'on ignore devient passionnant. De l'hypothético-déductif en vigueur nous passons à un positivisme-inductif original. La démarche épistémologique face au texte se trouve donc inversée: là où traditionnellement le chercheur interrogeait le texte sur la base d'hypothèses de travail construites, c'est le texte qui interroge le chercheur sans tabou et sans *a priori*. Par une lecture différente (hypertextuelle plutôt que linéaire, nous l'avons vu, mais aussi paradigmatique plutôt que syntagmatique, quantitative plutôt que qualitative), l'ordinateur voit autre chose pour déranger nos certitudes et élargir l'horizon étroit de nos *modes* (aux deux sens du terme) d'interrogation.

Mesurer

L'indexation du texte en machine décompose sa structure syntagmatique (quitte plus tard à la reconstruire pour étudier par exemple les régularités des enchaînements syntaxiques). Aussi l'ordinateur pourra-t-il lire le texte sur un axe paradigmatique inhabituel par le biais notamment de vastes dictionnaires de fréquences où les mots, les lemmes ou toutes autres unités linguistiques seront triés, par exemple, par ordre alphabétique.

En l'état ces dictionnaires ne présentent qu'un intérêt limité: lourds à manipuler puisqu'ils comptent autant d'entrées que d'unités linguistiques du corpus (très rapidement plusieurs dizaines de milliers), ils ne font état que de la présence d'un terme et donnent sa fréquence d'utilisation en valeur absolue dans le texte. C'est pour aller plus loin que la branche la plus active de l'ADT, longtemps appelée lexicométrie, a mis au point depuis plusieurs décennies des coefficients statistiques susceptibles de déterminer la «valeur» des fréquences enregistrées (11). Il s'agit donc non seulement de *compter* les unités linguistiques mais de les *mesurer* (comparer) pour *donner sens* aux fréquences trouvées.

Est-ce un complexe du littéraire à l'égard du scientifique? Peut-être: sans faire le culte du chiffre, il semble qu'il est garant d'une certaine objectivité, d'une certaine neutralité dans la description du monde, des choses, d'un corpus. Et répétons que cela nous paraît particulièrement incontestable lorsque ce corpus est un peu vaste. A vrai dire, les études qualitatives, elles-mêmes, ne font pas l'économie des ordres de grandeur pour décrire leur objet, mais elles en font un usage anarchique. Et au fil des démonstrations l'on trouvera des jugements, sans garantie mathématique, du type «il y a

beaucoup de noms et de déterminants dans la prose de Flaubert», «Racine, en vieillissant, utilise de moins en moins le lexique de l'amour», «Proust compose de longues phrases»… «Beaucoup», «de moins en moins», «longue»: *quid* ou plutôt *quantum* exactement.

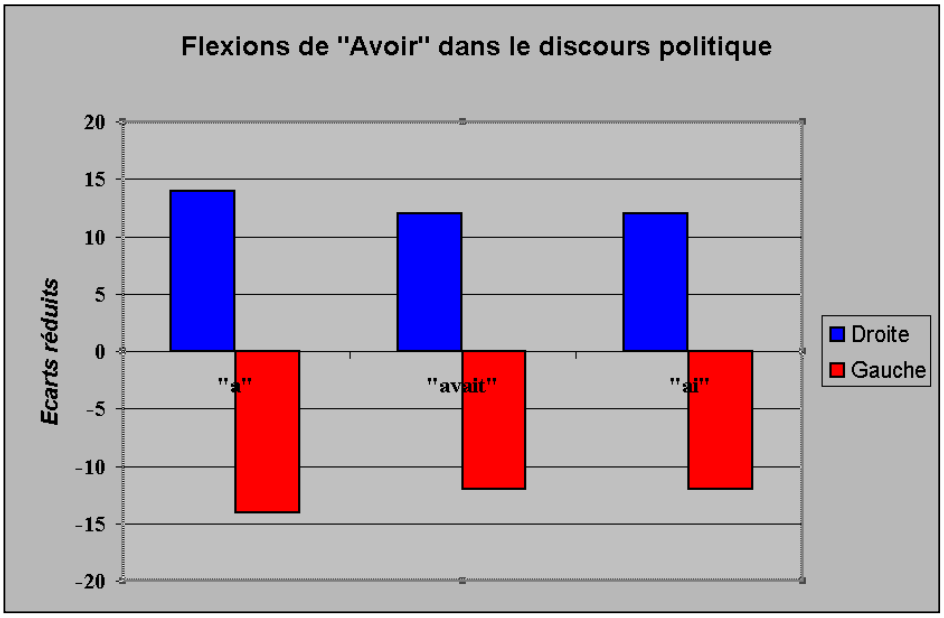
Dans le corpus donc, l'ordinateur indexe puis compte toutes les unités linguistiques pour nous faire une description précise et originale du contenu des textes. Surtout, lorsque le corpus est contrastif et se décompose en plusieurs parties (par exemple s'il rassemble plusieurs auteurs ou plusieurs œuvres du même auteur) l'ordinateur nous indique les unités linguistiques *spécifiques* (sur-utilisées ou sous-utilisées) de chacune des parties par rapport à l'ensemble. Ainsi, entre mille exemples, l'on saura du premier coup d'œil que ce qui caractérise grammaticalement le discours de Chirac par rapport à ses prédécesseurs à la présidence de la République (De Gaulle, Pompidou, Giscard et Mitterrand) est la sur-utilisation des adverbes, ou que Giscard d'Estaing sur-emploie le mot «énergie» dans des proportions statistiques considérables.

Cette description quantitative certaine et irréalisable à l'œil nu, nous amène ainsi à voir différemment le corpus. Surtout, parce que la totalité, sans exclusive, des unités linguistiques est pesée et soupesée, le pointage des caractères discriminants nourrit intelligemment nos investigations en soulevant des questions objectives dont non seulement on ignore la réponse mais dont on ne présentait pas forcément la pertinence heuristique.

Questionner

En effet, lorsque l'ordinateur décrit la partition de la Droite dans le concert politique français de l'entre-deux-guerres son attention (faite donc d'indexation et de mesures) s'arrête sur un phénomène linguistique étonnant. La caractéristique majeure des locuteurs de Droite par rapport à ceux de la Gauche est la sur utilisation des mots «a», «avait», «ont», en d'autres termes des flexions du verbe-auxiliaire «avoir». Ce phénomène est difficilement perceptible par la lecture naturelle tant le verbe-auxiliaire «avoir» se retrouve partout (y compris dans le discours de la Gauche) et se niche dans l'intimité syntaxique des textes. Pourtant l'écart statistique, ici mesuré en écart réduit, est important comme le montre simplement l'illustration 2:

Figure 2. Flexions de «Avoir»: distribution politique Droite/Gauche

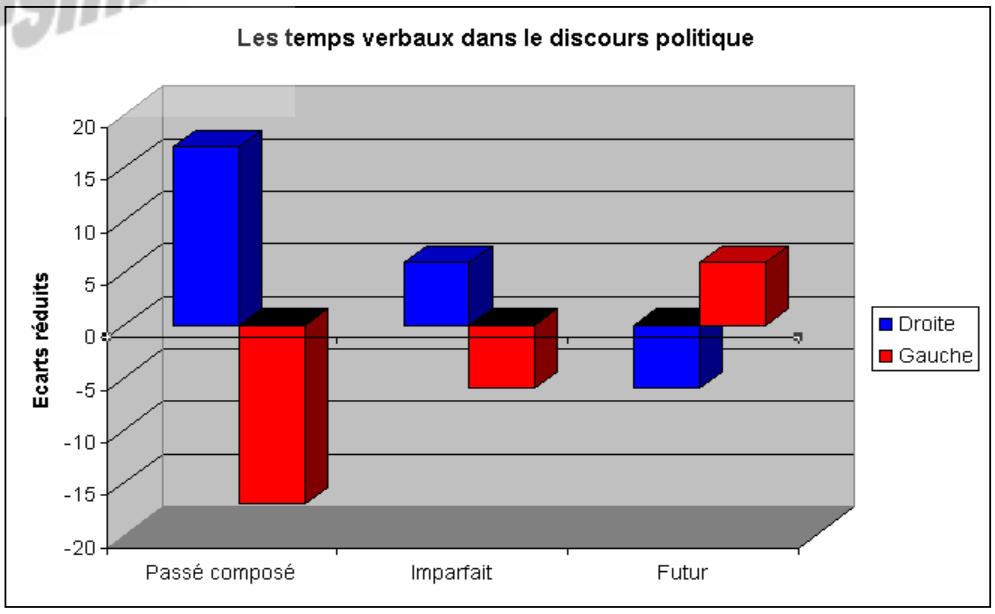


Ainsi, un constat quantitatif sans appel possible a été établi par l'ordinateur et il prend rapidement la forme d'une interrogation. Pourquoi un locuteur de Droite, ministre ou simple journaliste, député ou chef de parti, devant l'Assemblée nationale ou dans la presse, en 1929 ou en 1936, décline-t-il ses discours sur le mode du verbe-auxiliaire «avoir»? Pourquoi? L'analyste ne peut se dérober au questionnement objectif de l'ordinateur et n'aura de cesse d'essayer d'interpréter ce phénomène

dûment attesté.

En l'occurrence, nous nous sommes d'abord assuré (lecture systématique des contextes) qu'il ne s'agissait pas d'une redondance du verbe mais seulement de l'auxiliaire pour conclure que le discours de Droite n'est pas un discours du possédant mais un discours conjugué au passé composé ou au plus-que-parfait. Du reste, l'étude grammaticale effectuée par le couple logiciel Hyperbase-Cordial confirme ce résultat (illustration 3).

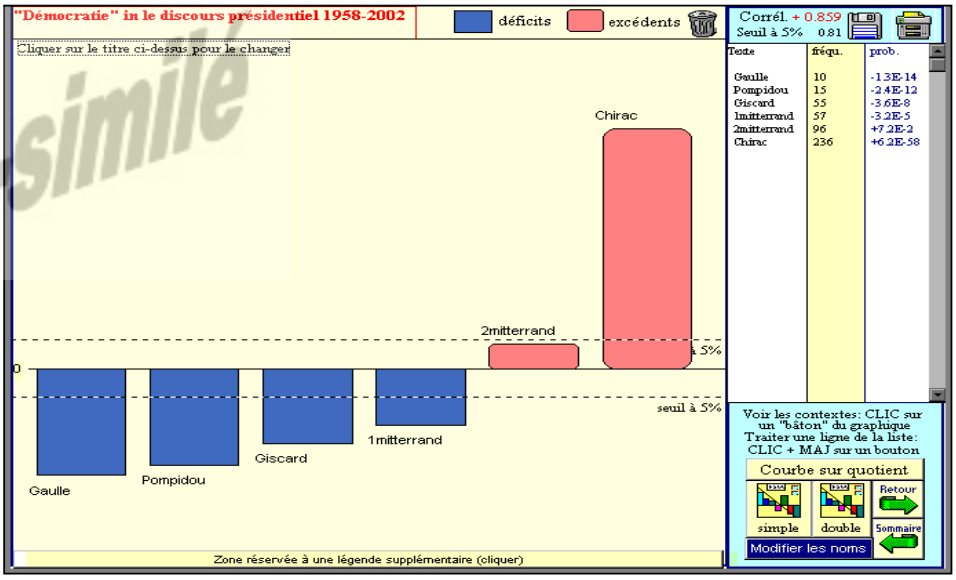
Figure 3. Les temps verbaux : distribution politique Droite/Gauche



Le discours de droite est un discours du constat. Quelles que soient les conditions de production, il est le plus souvent conjugué au passé (lorsque celui de la gauche est plutôt conjugué au futur). Le conservatisme pour ne pas dire le passéisme de la Droite se trouve inscrit dans la structure grammaticale des discours et sans doute dans la structure mentale des locuteurs. Plus précisément, dans l'entre-deux-guerres, la droite modérée, républicaine, orléaniste, dépassée par la crise de 1929 et par la montée du fascisme, semble épuisée et montre son incapacité à se projeter dans le futur. Elle semble condamnée dès le début des années 30 à s'abîmer dans le vichysme pour totalement disparaître de l'échiquier politique français (Mayaffre 2000, p. 225-291). Après Vichy, puis le monopole gaulliste, il faudra attendre Pinay (1952) voire Giscard d'Estaing (1974) pour la voir renaître.

De la même manière lorsque l'on compare les discours des cinq présidents français de la Ve République, l'ordinateur nous amène à réfléchir sur un phénomène de grande ampleur que l'on n'avait pas soupçonné. En effet, le mot qui connaît la plus importante et la plus régulière des progressions sur près de cinquante ans est «démocratie» (illustration 4).

Figure 4. Distribution de «Démocratie» dans le discours présidentiel (1958-2002)



A quoi attribuer ce phénomène qui ne peut être dû, - la statistique est formelle -, au simple hasard linguistique? J. Chirac serait-il plus démocrate que V. Giscard d'Estaing? Le F. Mitterrand du second septennat (1988-1995) plus que celui du premier septennat (1981-1988)? Une question est posée sur un phénomène lexical important de corpus, dont nous ne pouvons plus nous détourner tant elle apparaît statistiquement pertinente?

Induire

Si nous avons développé deux exemples et laissé le second sans réponse interprétative, c'est pour mieux insister sur l'inversion épistémologique de la démarche heuristique. Lorsque que le littéraire avait affaire à un grand corpus, il le lisait avec un questionnement précis, une hypothèse de travail établie, sans quoi il se perdait dans cette lecture. Pourtant, le danger de ces questionnements *a priori* ou exogènes est double et doublement cruel.

Le premier inconvénient a déjà été évoqué: le danger est grand en effet de projeter ses hypothèses de travail sur le texte, c'est-à-dire d'induire automatiquement, par nos questionnements, des réponses que l'on peut qualifier d'artefactuelles. Très concrètement, on peut s'étonner que dans les études traditionnelles, le chercheur confirme le plus souvent ses hypothèses pour rarement les infirmer: sommes-nous donc, tous, toujours, si pertinents face à un texte? Ou arrivons-nous toujours à tordre la réalité des textes pour la faire correspondre à nos (hypo)thèses de travail? En matière de justice, on sait qu'une instruction à charge conclut souvent à la culpabilité du prévenu lorsque qu'une instruction à décharge l'innocente; il convient de pouvoir juger sur des éléments objectifs et non de *pré-juger*. En analyse de texte, le risque est réel de toujours finir par trouver ce que l'on cherche (c'est-à-dire au fond ce que l'on sait déjà ou croit savoir), alors qu'une heuristique bien menée consiste à pouvoir découvrir ce que l'on ne sait pas.

Le deuxième danger est plus évident. Une hypothèse de lecture - quand bien même se trouverait-t-elle être pertinente - en oblitère toujours d'autres, plus pertinentes peut-être. Lorsqu'on lit un discours politique dans l'optique d'y rechercher la pensée économique de l'auteur, il devient difficile, humainement, de se concentrer sur sa pensée politique. Cette dimension du discours risque donc de passer à tout jamais aux oubliettes. Quand bien même l'œil aigu du chercheur la jugerait pertinente à mi-lecture, celui-ci acceptera-t-il de reprendre, de relire, de retraiter toute la première partie de l'œuvre pour approfondir l'analyse? Et si l'on acceptait de le faire, qu'en serait-il d'une troisième grille de lecture possible (la pensée sociale par exemple), puis d'une quatrième (l'écologie), puis…?

En d'autres termes la méthode hypothético-déductive est dangereuse autant par le risque de projection de réalités artefactuelles sur le texte que par l'oblitération de faits réels trop nombreux pour pouvoir être tous embrassés par la mémoire humaine. Sans doute nécessaire, elle ne peut être considérée comme une panacée.

Avec la lexicométrie, ce sont les informations objectivement pertinentes du corpus qui remontent, en bon ordre, jusqu'au chercheur. D'une certaine manière, l'ordinateur dégrossit le texte pour attirer notre attention sur les faits objectivement marquants (la sur-utilisation de l'auxiliaire avoir et du passé composé par la Droite, le sur-emploi des adverbes par Chirac, la progression spectaculaire de «démocratie» dans le discours présidentiel français depuis 1958). Dans une logique quantitative, ces faits sont rangés par ordre décroissant pour souligner d'abord ceux qui sont statistiquement incontournables jusqu'aux derniers, réels, mais plus discrets. Tant et si bien qu'aucun trait linguistique réellement discriminant ne peut échapper à l'analyste.

Le chercheur ne se précipite plus sur le texte sans autre médiation que ses pré-jugés. Ce sont les informations (lexicales, grammaticales, syntaxiques…) du texte indexé, trié, mesuré qui viennent l'interpeller dans ses compétences interprétatives. Si l'interprétation pourra être l'objet de discussions, les bases descriptives et interrogatives seront elles irréfutables: la subjectivité des chercheurs s'en trouve ainsi incontestablement repoussée pour être cantonnée au niveau de l'interprétation, alors qu'elle sévissait déjà au niveau descripto-interrogatif. Positivisme (les éléments discriminants et objectifs sont bien là, dans la lumière crue d'une description quantitative) et induction interprétative deviennent alors les caractéristiques de *l'herméneutique numérique*.

Conclusion

Insensiblement, depuis quelques décennies, l'ordinateur renouvelle notre rapport au(x) texte(s), notre rapport à la lecture, notre rapport au sens. Pour François Rastier, l'enjeu est plus important encore. Par ses possibilités (mémoire infinie, exhaustivité des relevés, souplesse d'utilisation), par sa façon d'appréhender la matière linguistique (approche paradigmatique, approche quantitative, hypertextualité) l'ordinateur redéfinit l'objet de la linguistique. Pour le linguiste français, cet objet n'est pas le mot ou la phrase (pure création des grammairiens) (p. 30) mais le texte voire le corpus textuel dans lesquels les mots et les phrases prennent leur sens; sans lesquels rien ne peut faire sens. Or très vite, dès que le texte est long et que le corpus compte quelques éléments, les régularités linguistiques, les irrégularités, les redondances, les raretés, les hapax, les présences, les absences, etc. ne peuvent être perçus sans la médiation de l'outil informatique.

Surtout, l'ordinateur nous interpelle dans nos pratiques herméneutiques et heuristiques. La lecture, la compréhension, le questionnement, puis l'interprétation des textes gagnent aujourd'hui en rigueur pour sortir du tout-subjectif. Dans un retournement spectaculaire, la démarche inductive complète la démarche déductive, le positivisme se substitue au constructionnisme. Et l'analyse se trouve libérée des grilles de lecture pré-construites (nos hypothèses imposées toujours par l'idéologie scientifique dominante) qui corsetaient notre description du texte dans un savoir déjà-là, réduisaient le champ d'interrogation de notre recherche dans nos *a priori* et limitaient par là-même l'amplitude de nos interprétations.

Sur des critères bien définis et stabilisés, notamment quantitatifs, l'ordinateur décrit le texte et nous apostrophe, sans pré-jugé, sur des phénomènes linguistiques (spécificité, constellation lexicale, rafale…) objectivement discriminants. Dès lors non seulement l'ordinateur se révèle un outil documentaire susceptible de mettre à l'épreuve nos hypothèses et d'aller traquer les preuves de nos interprétations, mais un outil heuristique susceptible de reculer l'horizon de nos investigations.

Notes

1 - Les dernières JADT se sont tenues en mars 2002 à St Malo (*JADT 2002, 13-15 mars 2002*, Saint-Malo, IRISA-INRIA, 2 volumes, 848 p.). Les prochaines doivent avoir lieu en Belgique au printemps 2004. Les actes des JADT 1998 (Nice), 2000 (Lausanne), 2002 (St Malo) sont publiés en version papier mais aussi en version électronique par la revue *Lexicometrica:* http://www.cavi.univ-paris3.fr/lexicometrica/

2 - On sait en effet que l'Histoire naît avec l'archive et que l'historien (contrairement au paléontologue ou au sociologue) travaille (presque) toujours sur des sources écrites.

3 - Pour une réflexion sur la lecture hypertextuelle, on se reportera aux travaux du laboratoire parisien «Communication et Politique» (UPR 36 - CNRS) dirigé par Georges Vignaux; équipe «Hypertextes et textualité électronique»: http://lcp.damesme.cnrs.fr/ On lira dans *l'Astrolabe* les articles de Sophie Marcotte et Christian Vandendorpe.

4 - On lira particulièrement le chapitre III, «Philologie numérique», de F. Rastier, *Arts et sciences du texte* (pp. 73-98).

5 - Pour plus de renseignements sur Frantext se connecter au site: http://www.inalf.fr/frantext si vous êtes abonné, ou sinon, au site: http://www.inalf.fr/atilf

6 - Pour cette étude sémantique de «fascisme», voir D. Mayaffre, «La Construction du sens en politique: le cas de "fascisme" dans le discours politique français des années 30», *Cahiers de la Méditerranée*, no 61, décembre 2000.

7 - I. Greilsammer, *Blum*, Paris, Flammarion, 1996, p. 261.

8 - Pour un autre exemple frappant de l'utilisation de l'ordinateur comme redresseur de contre-vérités historiques, voir D. Mayaffre, «History and Information Technology: The French are way behind», *Lexicometrica*, 2001.

9 - A. Prost, «Les Mots», in R. Rémond, *Pour une histoire politique*, Seuil, 1988, p. 258-259.

10 - On lira l'article électronique de Roger Chartier, «Du Codex à l'Ecran: les trajectoires de l'écrit», *Solaris*, no 1, 1994: http://grimmy.info.unicaen.fr/bnum/jelec/Solaris/d01/1chartier.html

11 - On se reportera évidemment à l'ouvrage de référence de Charles Muller, *Principes et méthodes de statistique lexicale*, ou à celui de L. Lebart et A. Salem, *Statistique textuelle*.

Références bibliographiques

Jadt 1995, III Giornate internazionali di Analisi Statistica dei Dati Testuali, Université degli Studi di Roma, CISU, 1995.

Jadt 1998, 4e Journées internationales d'analyse statistique des données textuelles, Nice, Université de Nice-Sophia-Antipolis-CNRS, 1998 (les articles sont également disponibles sous forme électronique: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/JADT1998.htm).

Jadt 2000, 5e Journées internationales d'analyse statistique des données textuelles, mars 2000, (les articles sont également disponibles sous forme électronique: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/tocJADT2000.htm).

Jadt 2002, 6e Journées internationales d'analyse statistique des données textuelles, Saint-Malo, IRISA-INRIA, 2002 (les articles sont également disponibles sous forme électronique: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/tocJADT2002.htm).

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

D. Mayaffre, *Le Poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres*, Paris, Champion, 2000.

D. Mayaffre, «History and Information Technology: The French are way behind», *Lexicometrica*, revue de lexicométrie sur l'internet, 2001: http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3/dm2001.htm

Ch. Muller, *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977; réédition: Champion, 1992.

F. Rastier, *Arts et sciences du texte*, Paris, PUF, 2001.

2002

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

De la lexicométrie à la logométrie

Formalisation et quantification des textes

Quelques obstacles historiques et épistémologiques dans le développement de l'analyse de texte informatisée