

C o m m u n i c a t i o n e n +



INTRODUCTION À LA RECHERCHE EN SIC

Stéphane Olivesi (dir.)



ANALYSES LOGOMÉTRIQUES ET RHÉTORIQUE DU DISCOURS

Damon Mayaffre

CNRS – UMR 6039 « Bases, corpus et Langage »

L'ensemble des sciences humaines et sociales a affaire aux mots, aux textes, aux discours. Que le texte soit considéré comme archive par l'historien ou source pour le sociologue, que le discours soit perçu comme témoin ou comme acteur pour le politologue, que le langage soit posé comme commencement pour la psychanalyse ou comme fin pour les sciences du langage... : nous travaillons tous *avec* ou *sur* du matériau linguistique. Cette vérité devient truisme pour les sciences de l'information et de la communication pour qui la parole, l'argumentation, la rhétorique sont des objets d'étude.

Partant, la réflexion sur les textes ou les discours – leur composition, les méthodes pour les traiter, leur compréhension ou leur interprétation – n'est pas un luxe mais une nécessité, sauf à s'abandonner à la lecture instinctive, à l'interprétation libre et à une démarche intuitive infrascientifique. Précisément, il ne s'agira dans cette contribution certes pas d'écrire un nouveau précis d'*Analyse du discours* mais d'exposer une méthode d'approche globale des textes, susceptible de décrire le lexique et la grammaire, les réseaux thématiques et les structures rhétoriques; susceptible, surtout, de baliser les parcours de lecture pour objectiver, autant que faire se peut, l'interprétation.

La lexicométrie de la seconde génération – que nous appellerons « logométrie » car elle ne se contente pas de traiter du lexique (*lexi**) mais étend ses procédures à toutes les unités linguistiques jugées pertinentes du discours (*logo**) : mots graphiques, lemmes, co-occurents, codes

grammaticaux, enchaînements syntaxiques, etc. – connaît aujourd’hui un essor certain. Sur les bases posées dans les années 1960-1980, elle prend souffle depuis peu, sous le double mouvement du développement technologique (puissance et généralisation des machines domestiques, performance et ergonomie simplifiée des logiciels, disponibilité des textes numérisés chaque jour plus nombreux) et sur des réflexions épistémologiques abouties, émanant de plusieurs branches montantes de la linguistique : la linguistique de corpus, l’analyse du discours et la linguistique textuelle, l’analyse de données textuelles assistée par ordinateur, la philologie et l’herméneutique numériques.

Pratiquer la logométrie : mode d’emploi

La logométrie est une méthode d’analyse des textes, assistée par ordinateur, qui permet de décrire qualitativement et quantitativement le contenu linguistique d’un corpus. Elle allie en effet des outils de recherche ou de compilation documentaire (concordanciers, recherche de contextes, navigation hypertextuelle) et des outils statistiques et mathématiques susceptibles de caractériser un texte (dictionnaire de fréquences, calcul du vocabulaire caractéristique, distances intertextuelles, accroissement lexical chronologique, etc.). Une des difficultés pour le néophyte, une des forces certainement pour le praticien tient dans le fait que les logiciels proposés sur le marché scientifique croisent ces deux types d’outils. La logométrie met en exergue les traits saillants des discours grâce à une approche *macro* ou *globale* et aux outils quantitatifs, puis permet la lecture et un retour systématique au texte grâce à une vision *micro* ou *locale* et aux outils qualitatifs. Dans le cheminement de l’analyste, trois stations apparaissent incontournables : la constitution du corpus, la définition des unités linguistiques du texte jugées pertinentes, enfin le traitement textuel proprement dit, qualitatif ou statistique.

Les corpus textuels

Depuis quelque temps, particulièrement au tournant des années 2000, la réflexion sur l'objet « corpus » s'impose transversalement dans les SHS comme un des préalables importants aux pratiques scientifiques. On ne citera à titre d'exemple que la création d'une revue entièrement dédiée à cette réflexion en 2001, au titre suggestif, *Corpus*. Dans le domaine logométrique, la réflexion sur les corpus n'est pas seulement importante mais indispensable. En effet, si les contours d'un corpus d'étude en Histoire par exemple ou en Littérature peuvent apparaître théoriques et ajustables au fil de l'analyse, la nécessité d'une saisie physique d'abord et d'un traitement informatique ensuite rend ce contour concret et impérieux en logométrie. Souvent simple potentialité (« tous » les textes susceptibles d'intéresser le chercheur que l'on convoquera ou pas, à discrétion, au fil de l'analyse), le corpus est pour nous un objet physique stable, entré manuellement ou de manière semi-automatique en machine. Souvent mou, le corpus est donc ici un objet dur qui répond à plusieurs caractéristiques.

D'abord, les corpus logométriques sont clos

Les traitements informatiques tournent nécessairement sur des ensembles arrêtés. Une recherche de concordances par exemple convoquera les occurrences d'un mot dans un ensemble de textes donnés (ou plutôt *saisis*) et ignorera évidemment les textes hors de cet ensemble. La limite entre l'intérieur et l'extérieur du corpus est donc tranchée. L'intérieur fera l'objet d'un traitement systématique et exhaustif – systématisme et exhaustivité du traitement constituent une des forces de l'approche assistée par ordinateur – ; et l'extérieur du corpus sera dans un premier temps ignoré.

Surtout, les traitements statistiques proposés exigent une stabilisation de la population traitée. Le décompte d'unité n'a de sens que dans un ensemble précisé. On ne calcule un pourcentage ou un écart réduit que par rapport à un tout intangible. L'ajout d'un texte voire d'un seul mot modifierait mathématiquement l'ensemble de la donnée statistique. Sur ce point, l'exigence statistique rejoint en réalité un

postulat linguistique fondamental : le corpus, en sa clôture, constitue une norme linguistique par rapport à laquelle chaque texte du corpus va se démarquer. Il n'existe en effet aucune norme en Langue, aucune norme linguistique absolue. Le profil d'utilisation d'un mot est non pas dicté par le système mais par la performance qu'un locuteur donné produit dans une situation particulière. La fréquence d'utilisation du mot « France » par exemple sera *quasi* nulle dans un corpus de textes scientifiques mais importante dans un corpus de discours politiques français. La norme est donc toujours *endogène* au corpus ; la norme *est* le corpus. Les régularités ou irrégularités linguistiques d'un texte ne prendront de sens qu'au sein du corpus que l'on aura défini, clôturé et institué en référence.

De ces réflexions, nous retiendrons donc une première idée : la pratique logométrique passe par une délimitation impérieuse du corpus de travail. D'apparence banale, l'idée est contraignante tant techniquement par le travail concret de saisie qu'elle réclame que conceptuellement par l'effort demandé pour circonscrire strictement le périmètre de la recherche.

Ensuite, les corpus logométriques sont contrastifs

Clos, le corpus constitue donc la référence. Partant, le jeu consiste à faire contraster des parties du corpus les unes par rapport aux autres et toutes par rapport à l'ensemble. Pour cette raison, un corpus logométrique est nécessairement contrastif et partitionné.

Sans prétendre être exhaustif, on distingue classiquement deux types de contrastes (qu'il est avantageux de croiser) : les contrastes d'auteurs ou d'œuvres et les contrastes chronologiques. La plupart des corpus rassemble en effet plusieurs auteurs que l'on cherche à caractériser. Ainsi rassemblera-t-on les tragédiens français pour déceler, dans cet ensemble, les caractéristiques de Corneille ou de Racine. Ainsi rassemblera-t-on les discours des présidents français de la V^e République (de Gaulle, Pompidou, Giscard, Mitterrand, Chirac) pour distinguer, sur ce fond, les particularités du parler gaullien ou mitterrandien. De la même manière, lorsqu'un auteur est important,

on recueillera l'ensemble de son œuvre pour caractériser, dans ce tout, tel roman, telle nouvelle, etc.

Par un cheminement tout aussi classique, certaines études rassemblent les productions textuelles d'un pan chronologique donné, pour faire contraster différentes périodes (les années, les semestres, etc.) et déceler les évolutions chronologiques. Ces études diachroniques produisent en général de forts résultats tant les pratiques discursives, particulièrement lexicales, sont soumises à la temporalité, du fait de l'évolution naturelle des locuteurs ou de l'actualité politique changeante qui entoure et détermine la production des discours.

Au final, l'idée de contrastivité n'est pas nouvelle. On ne définit l'Un que par l'Autre, et ne détermine la partie que dans le tout. Ici c'est le corpus qui forme le tout et les textes constituent les parties. Celles-ci, répétons-le, se caractérisent les unes par rapport aux autres, et toutes par rapport à la norme que constitue l'ensemble.

Enfin, le corpus est homogène

Directement contradictoire avec l'exigence précédente, un corpus doit rassembler des textes semblables. La contrastivité doit être tempérée par de l'homogénéité; les textes doivent être différents, certes, mais comparables.

Ainsi tout en étant diachronique, un corpus ne saurait rassembler des textes d'époques trop éloignées sous peine de trouver des résultats inexploitable. De la même manière, un corpus regroupant des locuteurs sans point commun (un auteur politique contemporain et un poète du XVI^e siècle par exemple) apparaîtrait sans intérêt.

Dans ce cadre, les travaux les plus actuels insistent sur la nécessité d'une homogénéité générique. Depuis Bakhtine, certains auteurs (Adam 1999; Rastier 2001) ont en effet montré la prégnance des genres sur nos productions langagières. Bien que négligés, ceux-ci contraignent les discours de manière importante et d'autant plus subtile qu'elle n'est pas toujours consciente. Qu'on le veuille ou non, on ne s'exprime pas de la même façon dans le cadre générique d'un cours universitaire

et dans celui d'une discussion amicale au café du commerce, dans le cadre d'un éditorial et dans celui d'une interview, dans un roman et dans une nouvelle. Aussi dans un corpus mal constitué, le danger existe d'attribuer à une différence d'auteurs ou d'époques ce qui ne relève que d'une différence de genres.

Reste que la tension entre la contrastivité (c'est-à-dire une forme d'hétérogénéité) et l'homogénéité du corpus est le moteur problématique des études logométriques. Trop de variables de contraste empêche de conclure, car l'on ne sait attribuer à l'une ou l'autre variable les résultats des sorties machines ; pas assez de variables ou de contrastes, au contraire, amoindrit l'intérêt de la recherche.

Les unités du texte

Pas plus que les autres logiciels, ceux que l'on utilise en logométrie ne sont intelligents. Il convient de leur indiquer précisément les unités linguistiques que l'on veut chercher et/ou compter dans un texte. Il s'agit tout simplement de définir les *entrées* souhaitées dans le corpus.

L'unité la plus formalisable d'un texte a été pendant longtemps, et reste encore aujourd'hui, le mot graphique. Les mots graphiques apparaissent en effet comme des unités physiques, invariables dès lors que l'orthographe d'une langue est normalisée, et *quasi* universelles, dans leurs principes, au-delà de la différence de langues et d'alphabets. Une segmentation automatique du texte les mettra donc en évidence sans difficulté.

Ainsi, le traitement des textes bruts, desquels on extrait *les concatenations de caractères prises entre deux blancs* (selon la définition des mots graphiques), a longtemps été un horizon indépassable de la lexicométrie. Cet horizon, qui est donc justifié par des raisons pratiques, se justifie aussi théoriquement d'un point de vue linguistique par l'approche matérialiste du sens et du langage qui a présidé à la discipline à ses origines. La lexicométrie originelle s'est en effet méfiée des traitements linguistiques préalables qui manipulaient – trahissaient ? – le texte. Elle a toujours préféré s'appuyer sur des unités matérielles

certes sans grandes pertinences linguistiques mais avérées. L'interprétation sémantique était ainsi clairement renvoyée, en bonne logique, en aval du traitement lexicométrique et non en amont. On prenait la surface du texte dans sa matérialité la moins discutable pour en faire un traitement objectif ; l'exégèse linguistique comme l'interprétation sociolinguistique ne venaient qu'ensuite.

Aujourd'hui néanmoins, les pratiques s'étendent à d'autres unités que la forme graphique. C'est le passage de la *lexicométrie* à la *logométrie* : les lemmes, les codes grammaticaux, les séquences syntaxiques, et finalement toutes les unités linguistiques jugées pertinentes du discours (réseaux thématiques, isotopies, etc.) peuvent espérer être traitées. Cela signifie qu'en amont du traitement statistique, les textes doivent être lemmatisés et étiquetés.

La lemmatisation est une opération linguistique qui consiste à ramener les unités graphiques (notamment toutes les flexions) à leur unité de sens c'est-à-dire aux *lemmes* (les formes canoniques qui servent d'entrées dans les dictionnaires). Ainsi, dans la phrase « je suis parti », le mot graphique « parti » sera ramené à son lemme *partir* (verbe) lorsque le même mot graphique dans la phrase « vive le parti ! » sera identifié au lemme *parti* (nom). La lemmatisation est une opération complexe bien moins évidente qu'il n'y paraît (Labbé, 1999) et peut-être pouvons-nous convenir qu'elle n'est jamais parfaite. Mais les logiciels aujourd'hui la réalisent en quelques secondes avec des pourcentages d'erreurs marginaux (1 % à 2 %), et, dès lors, l'analyste gagnera *a minima*, selon l'exemple donné, à désambiguïser sémantiquement les homographes pour un traitement linguistique plus recevable.

La lemmatisation va souvent de pair avec l'étiquetage morphosyntaxique et le codage grammatical. On aura en effet remarqué que ramener « parti » à *partir* signifie que l'identité verbale (*versus* nominale) de la forme a été reconnue et que, sans doute, le participe passé a été repéré. Ainsi peut-on étiqueter un texte et associer à chaque forme graphique son lemme donc, mais aussi son code grammatical élémentaire (nom, verbe, pronom, etc.), mais encore son genre pour les noms, son nombre, son temps et sa personne pour les verbes, etc. ; certains logiciels vont jusqu'à identifier la fonction grammaticale

des mots dans la phrase (sujet, complément du sujet, complément circonstanciel, etc.).

À partir d'informations grammaticales minimales, il devient encore possible de passer aux structures syntaxiques. Les enchaînements de codes grammaticaux (« pronom + verbe + adverbe », « déterminant + nom + ... », etc.) pourront ainsi être cherchés et décomptés dans le corpus pour en mesurer la régularité d'utilisation. L'idée fondamentale de la logométrie (*versus* lexicométrie) est assez simple : il s'agit de multiplier les niveaux de description linguistique des textes (graphies, lemmes, codes grammaticaux, syntaxe) pour en donner une vision globale. Évidemment, en pareil cas, il est souvent intéressant de croiser ces différents traitements pour chercher/compter par exemple une structure syntaxique dont un des éléments est contraint lexicalement (pronom + le verbe *partir* + adverbe).

Au final, le gain de la lemmatisation et de l'étiquetage est évident. Non seulement la forme graphique très souvent ambiguë, donc dépourvue de sens, se trouve renseignée sémantiquement, mais le traitement du texte est étendu au-delà du simple lexique vers la tonalité rhétorique des discours faite notamment de notes grammaticales et de tournures syntaxiques particulières.

Les outils : articuler quantitatif et qualitatif

Après avoir constitué un corpus adéquat et défini les unités linguistiques pertinentes pour l'analyse, la démarche logométrique mobilise une gamme d'outils de traitement. Il s'agit là du cœur de la démarche. Se distinguent grossièrement deux types d'outils : les outils documentaires d'essence qualitative et les outils statistiques ou quantitatifs.

Les outils documentaires

Les logiciels donnent d'abord libre accès aux textes naturels qui peuvent se lire à l'écran comme ils pouvaient se lire sur papier. Il convient de le préciser tant la discipline a injustement été soupçonnée de déraciner le chercheur du texte d'origine, là où, au contraire, elle s'applique à constamment l'y renvoyer.

La navigation hypertextuelle

Seulement, le texte lu à l'écran est techniquement un *hypertexte* généralisé. Dès lors, les vertus de l'hypertextualité permettent une navigation sans limite dans le corpus pour faciliter l'exploration documentaire. Chaque mot du corpus est en effet relié à un index et à un dictionnaire exhaustifs – comme les seuls noms propres figurent dans les index des ouvrages papiers avec le renvoi aux pages concernées –. Ce dictionnaire exhaustif, que l'on présentera, au choix, selon l'ordre alphabétique des mots ou selon l'ordre hiérarchique de leur fréquence, contient l'adresse électronique de toutes les unités recensées dans le corpus, et, par simple clic, le lecteur sera directement renvoyé aux phrases, aux passages, aux discours contenant l'unité désirée. D'un point de vue épistémologique, on remarquera l'importance de cette démarche élémentaire : aucune sélection de mots n'est *a priori* effectuée, puisque tous les mots sont indexés. La recherche n'est donc pas bornée au départ et s'ouvre, dès lors, sur tous les possibles.

Les concordances

Au-delà de la navigation hypertextuelle, plusieurs outils documentaires permettent une contextualisation linguistique rapide et organisée des unités. Le plus connu est le concordancier (Pincemin, 2006). Il permet d'extraire et de convoquer toutes les phrases du corpus contenant les occurrences d'un mot pour en vérifier l'emploi. La présentation est ergonomisée sous forme de liste exhaustive, centrée sur le mot pour favoriser une lecture systématique. Un système de tri permet d'agencer les phrases selon l'environnement lexical à gauche ou à droite du mot pôle (illustration 1).

Illustration 1. Extrait de la concordance du mot « démocratie » dans le discours de Jacques Chirac (1995-2003), triée à gauche selon l'ordre alphabétique du mot précédant le mot-pôle. Un clic sur une des lignes renvoie l'analyste au plein texte pour une contextualisation élargie (concordance tirée de LEXICO 3).

... pour avoir une plus grande **démocratie** sociale. Là encore, on la vie politique d'une grande **démocratie** ne repose pas seulement ... que je crois qu'une grande **démocratie** doit utiliser le référend à une démocratie, à une jeune **démocratie** en tous les cas de se for sons davantage confiance à la **démocratie** locale, à sa puissance d' consiste à faire confiance à la **démocratie** locale pour saisir les ch ...ons les plus conformes à la **démocratie** et à la dignité de l'hom avec l'objectif de donner à la **démocratie** locale plus de vitalité et

Le retour au texte

Enfin, plus généralement, la logométrie met en scène un *retour au texte* systématique. Chaque outil (les concordances comme nous venons de l'indiquer, mais aussi les outils les plus techniques : graphiques, spécificités, analyse factorielle des correspondances, etc.), nous renvoie au cœur du corpus par simple clic, pour renouer *in fine* avec la lecture naturelle globale du texte, à l'égal de celle effectuée sur papier.

Cette *exigence philologique* apparaît comme une des forces et une des originalités de la logométrie par rapport au traitement automatique des langues (TAL) dont les traitements sont plus désincarnés. L'informatique apparaît ici clairement comme un outil et non comme une finalité. Elle permet d'organiser et de contrôler les parcours de lecture. Elle permet de rechercher et d'extraire de l'information du corpus. Mais l'acte final d'interprétation du chercheur ne peut se faire sans embrasser le texte naturel dans lequel nous sommes toujours replongés.

Les outils statistiques

Le sel des logiciels de logométrie est évidemment le traitement quantitatif dont on ne prétendra pas faire ici le tour. Posons simplement que depuis les années 1960, puis dans les décennies suivantes (Muller, 1977 ; Lebart & Salem, 1994), ce traitement n'a jamais cessé de s'améliorer pour atteindre aujourd'hui des performances remarquables aux vertus herméneutiques avérées. La mise en ligne des manuels d'utilisation des logiciels les plus performants permettra au lecteur de s'en faire une idée plus précise (voir par exemple Brunet 2001).

Les spécificités

Outre le dictionnaire ou index de fréquences dans lesquels sont concentrées toutes les données quantitatives élémentaires du corpus (nombre de mots, fréquences des mots, distributions de ces fréquences dans les différentes parties du corpus...), l'outil fondateur de la discipline est sans doute le calcul des spécificités (Lebart & Salem, 1994). Il permet de repérer, après traitement systématique, le vocabulaire objectivement *spécifique* (caractéristique) d'une partie du corpus par rapport à l'ensemble.

Ainsi, dans le corpus des discours présidentiels de la Ve République, prononcés à la télévision depuis 50 ans par de Gaulle, Pompidou, Giscard, Mitterrand et Chirac, les principaux mots caractéristiques de chaque président (par rapport à l'ensemble) ont pu être repérés. On notera dans le détail, qu'une spécificité peut être négative (il s'agit alors d'un mot *sous-utilisé* par rapport à la norme) ou positive (il s'agit alors d'un mot *surutilisé* par rapport à la norme) (illustration 2).

Illustration 2. Spécificités positives et négatives des présidents de la République (1958-2003). L'indice (ici un écart réduit) positif ou négatif qui suit les mots mesure le degré de surutilisation ou de sous-utilisation... [page suivante]

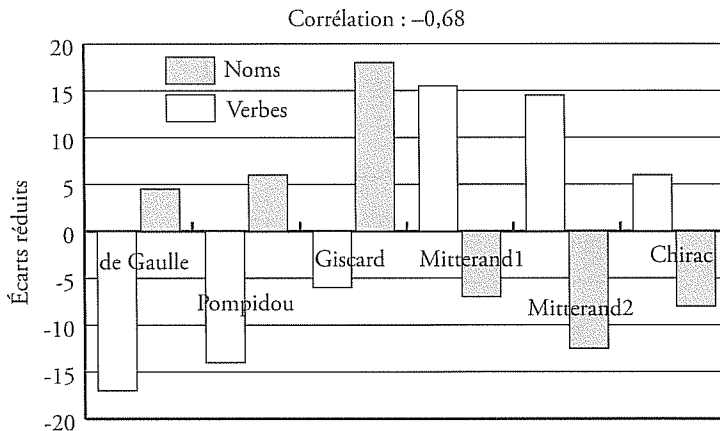
de Gaulle		Pompidou		Giscard		Mitterrand1		Mitterrand2		Chirac	
+	-	+	-	+	-	+	-	+	-	+	-
Algérie (+31)	Je (-37)	de Gaulle (+13)	Je (-18)	Actuel (+30)	Moi (-12)	Je (+33)	Économique (-12)	Je (+31)	Nous (-12)	Naturellement (+23)	Actuel (-11)
Peuple (+21)	Pas (adv.) (-24)	Parisien (+12)	Avoir (v.) (-15)	Situation (+24)	Vouloir (-12)	Monsieur (+19)	Algérie (-9)	Pas (adv.) (+24)	Action (-10)	Aujourd'hui (+22)	Soviétique (-11)
Algérien (adj) (+18)	Avoir (v.) (-22)	Civilisation (+12)	Être (v.) (-13)	Heure (+21)	Nation (-11)	Tchad (+16)	Activité (-9)	Moi (+22)	Concerner (-10)	Jeune (+20)	Français (adj) (-10)
État (+17)	Vous (-18)	Monétaire (+11)	Falloir (-12)	Problème (+20)	Je (-11)	Moi (+15)	État (-8)	Maastricht (+21)	Activité (-9)	Noramment (+20)	Moi (-9)
Régime (+16)	Dire (-16)	Jeunesse (+11)	Faire (-9)	Énergie (+20)	Peuple (-10)	Penser (+14)	Européen (adj) (-8)	Ne (+20)	Problème (-9)	Euro (+20)	Nucléaire (-9)
Univers (+15)	Être (v.) (-16)	Coopération (+10)	Moi (-8)	Événement (+18)	Gens (-9)	Pas (adv.) (+14)	Développement (-8)	Avoir (v.) (+16)	Niveau (-9)	Évoquer (+19)	Question (-9)
Nation (+15)	Ne (-14)	Napoléon (+10)	Pas (adv.) (-8)	Indiquer (+17)	Europe (-9)	Dire (+13)	Problème (-8)	Gorbatchev (+15)	Emploi (-9)	Démocratie (+18)	Communauté (-8)
Atomique (+14)	Falloir (-12)	Autoroute (+10)	Débat (-7)	Question (+17)	Dire (-8)	Chaîne (+11)	Action (-8)	Europe (+14)	Effort (-8)	Mondialisation (+18)	Guerre (-8)
Condition (+14)	Problème (-11)	Conception (+9)	Nucléaire (-7)	Fonction (+15)	Arriver (-8)	Tiers-monde (+11)	Coopération (-8)	Traité (+14)	Français (adj.) (-8)	Réforme (+14)	Peuple (-8)

Totalitaire (+14)	Penser (-10)	Poète (+9)	Emploi (-7)	Emploi (+15)	Référendum (-7)	Vous (+11)	Solution (-7)	Socialiste (adj.) (+13)	Énergie (-8)	Devoir (v.) (+14)	Politique (n.) (-8)
Présent (+12)	Moi (-9)	Bonheur (+9)	Majorité (-7)	Français (adj.) (+14)	Devoir (n.) (-7)	Pershing (+11)	Progrès (-7)	Frontière (+12)	Actuel (-8)	Petit (+13)	Économique (-7)
Coopération (+12)	Important (-9)	Individu (+9)	Démocratie (-6)	Programme (+14)	Intérêt (-7)	Nelle Caléd. (+10)	Monde (-7)	Droit (n.) (+10)	Entreprise (-8)	Essai (+13)	République (-7)
Accomplir (+12)	Emploi (-8)	Nixon (+8)	Chômage (-6)	Pétrole (+14)	Argent (-7)	Nationalisation (+10)	Réforme (-7)	Aimer (+10)	Développement (-8)	Faire (+13)	Attitude (-7)
Soviet (+11)	Société (-8)	Poésie (+8)	Parti (-6)	Niveau (+14)	Guerre (-7)	Nationaliser (+10)	Régime (-7)	Penser (+10)	Région (-8)	Permettre (+13)	France (-7)

Les graphiques de distribution

La logométrie met aussi à disposition nombre de représentations graphiques qui permettent de visualiser le profil d'utilisation d'un mot dans un corpus partitionné. Sur le même corpus présidentiel que précédemment, nous pouvons, par exemple, visualiser la distribution des noms et des verbes pour constater que le phrasé des présidents, longtemps nominal, devient verbal au détour des années 1980 : c'est non seulement le message mais la rhétorique fondamentale des discours qui s'en trouvent modifiés (illustration 3).

Illustration 3. Distributions des noms et des verbes dans le discours présidentiel (1958-2003).



Calcul des co-occurrences et réseaux thématiques

Très précieux pour les recherches thématiques, les algorithmes de logométrie permettent encore de repérer les attirances et répulsions lexicales. Après traitement, tous les mots co-occurents – c'est-à-dire qui apparaissent ensemble au sein des phrases ou des paragraphes – sont repérés systématiquement et présentés par ordre d'affinité. Dans une logique similaire, l'univers linguistique d'un mot-pôle (ici

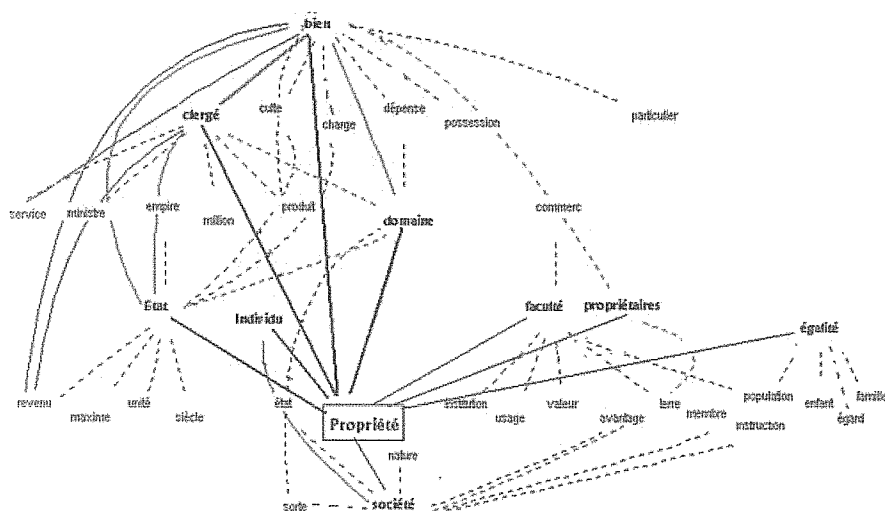
« mondialisation » dans le discours de Chirac [illustration 4]), peut être reconstitué pour appréhender son environnement lexical immédiat : ce sont les traits isotopiques et les thématiques des discours que l'on peut désormais espérer formaliser.

Illustration 4. Environnement lexical du mot « mondialisation » dans le discours de Chirac (1995-2002). Le tableau fait apparaître par ordre hiérarchique les mots qui sont les plus attirés par « mondialisation ». Trois traits isotopiques du discours peuvent ainsi être distingués. Dans un propos assez proche de l'altermondialisme, Chirac (1) dénonce les « dangers » de la mondialisation. Seulement, (2) il juge la mondialisation « inéluctable » et, pourquoi pas, porteur de certains « avantages ». Aussi (3) milite-t-il pour une mondialisation « maîtrisée » (voir Mayaffre 2004 : 133-140). [Illustration tirée d'Hyperbase].

C:\Hyperbas\Elysee.exe			
Environnement lexical de « Mondialisation »			
Cliquer sur un mot pour voir les contextes			
Ecarts réduits	Fréq. dans le corpus	Fréq. dans le texte	Mots (ordre hiérarchique)
8,72	58	9	dangers
7,99	47920	115	la
7,73	18	6	inéluctable
7,47	9	5	maîtrisée
7,24	114	8	effets
6,64	108	7	modèle
5,23	109	5	avantages
5,18	545	8	social
5,07	13	3	porteuse
4,78	75	4	maîtriser
4,61	92	4	exclusion
4,59	371	6	solidarité
4,50	104	4	considérable
4,13	55	3	pauvreté
4,00	66	3	maîtrise

De manière plus visuelle, les logiciels organisent et donnent à voir ces réseaux lexicaux co-occurentiels sous forme de graphes. L'illustration 5, tirée de la dernière version d'Hyperbase, permet de visualiser ainsi les relations privilégiées que « propriété » entretient avec d'autres mots dans un corpus de discours datant du début de la Révolution (1789-1791).

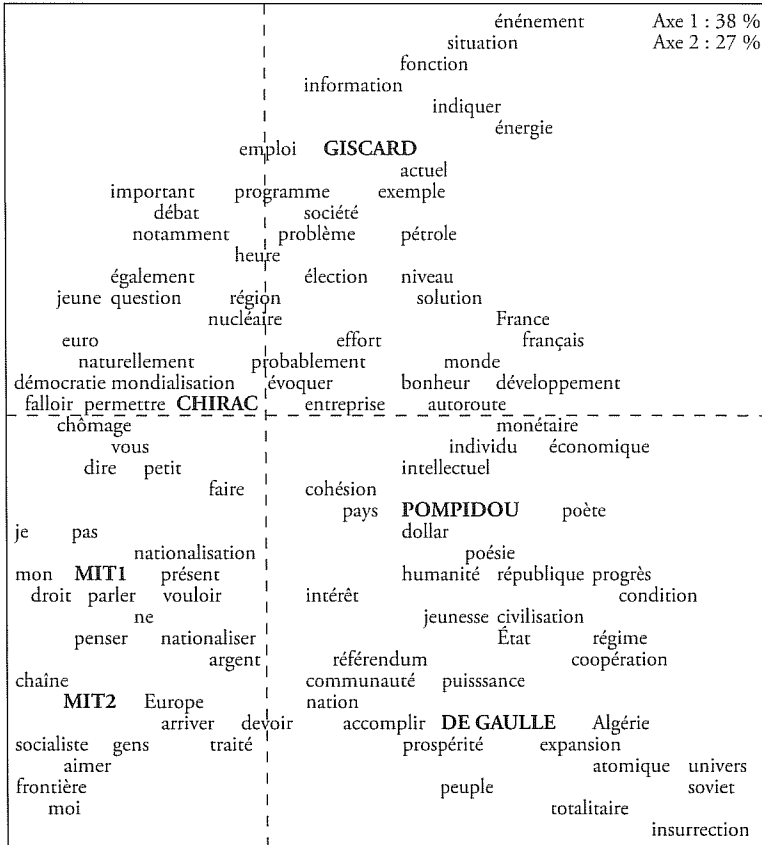
Illustration 5. Graphe de co-occurents: « propriété » dans le discours révolutionnaire (1789-1791). « Propriété » apparaît directement associée à « État », « société », « individu », « clergé », « bien », « domaine », « faculté », « égalité ». À leur tour ces mots sont associés à certains co-occurents très parlants.



Outils macro de classification

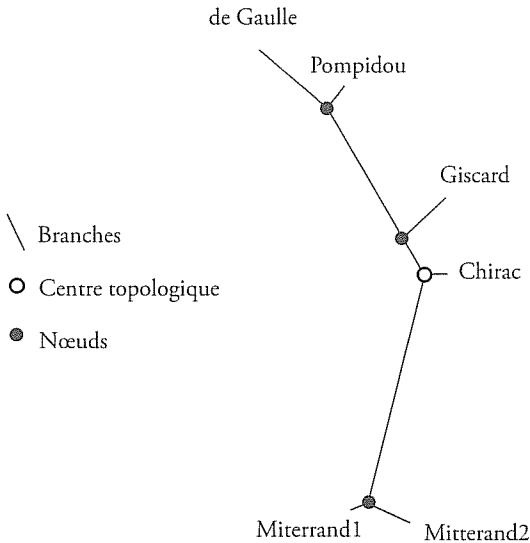
Enfin, le traitement quantitatif prend toute sa puissance lorsqu'il met en œuvre des outils de synthèse susceptibles de traiter une information à la fois massive et plurielle. Le plus connu est l'analyse factorielle des correspondances (Cibois 1994), [illustration 6].

Illustration 6. Analyse factorielle des correspondances du vocabulaire présidentiel. Dans chaque cadran, les présidents dont le vocabulaire est proche sont regroupés. Les mots, responsables de ces regroupements, sont indiqués à proximité des présidents qui les emploient le plus.



Plus synthétique encore, le calcul de la distance intertextuelle et sa représentation arborée permettent de calculer et de visualiser la proximité ou l'éloignement des textes du corpus à partir de la mesure globale de leur vocabulaire commun ou exhaustif (illustration 7).

Illustration 7. Distance intertextuelle et analyse arborée. Aux deux extrémités de l'arbre s'opposent deux paires de textes-feuilles apparentés. Le texte-feuille de Giscard, lui, apparaît sur la branche gaullienne. Chirac, quant à lui, au centre, tient un discours sans parenté directe à égale distance du discours gaullien et miterrandien.



Autres...

Les quelques outils présentés ne prétendent pas recouvrir toutes les fonctionnalités du traitement logométrique ; ils n'en représentent que le socle. En réalité, chaque discipline a spécialisé des fonctions et développé des usages. Les littéraires seront sensibles par exemple au calcul de la richesse du vocabulaire dans la perspective d'études stylistiques ; les historiens seront sensibles au calcul des corrélations chronologiques susceptible de rendre compte des corpus diachroniques ; les linguistes privilégieront des outils à même de traiter l'en deçà des mots (chaîne de caractères, suffixes, préfixes...) ou l'au-delà des mots comme les enchaînements syntaxiques contraints. Tous trouveront dans la logométrie un moyen de mettre en équation le texte, pour repérer, calculer puis représenter des traits objectivement remarquables du corpus.

Comprendre la logométrie : enjeux et attentes

La technicité des outils qui viennent d'être présentés ne doit pas masquer l'effort de réflexion épistémologique, entrepris depuis plusieurs années, notamment dans le cadre de la revue *Mots. Les langages du politique*, sur les enjeux et les limites de la démarche. Après plusieurs décennies de pratique sans doute peut-on théoriser les apports et établir un bilan circonstancié.

Dépasser la lecture empathique des textes

La principale difficulté lorsqu'on se fixe comme objectif l'étude des textes et des discours est l'inévitable confusion entre langage et métalangage, entre l'objet de la recherche (disons pour faire simple : les mots) et l'outillage intellectuel du chercheur (encore des mots !) ou, encore, entre les compétences naturelles du locuteur natif et celles construites de l'analyste. En d'autres termes, le langage nous est si évident que son traitement ne semble pas devoir faire le détour par aucun protocole méthodologique d'analyse. Pourtant, il n'est guère sérieux, dans un cadre scientifique, de s'abandonner à une lecture empathique des textes, à une lecture intuitive, naturelle ou sauvage des corpus. Cette réflexion paraît d'autant plus vraie que les corpus étudiés sont importants (plusieurs milliers de textes, plusieurs millions de mots) dépassant les capacités de synthèse de la mémoire humaine. Dans sa plus simple expression, la science doit expliciter sa démarche pour se mettre sous le coup d'une saine critique ; de plus, elle doit imposer une médiation entre l'objet analysé et le sujet analysant.

L'outil informatique et les algorithmes statistiques de décryptage servent précisément de médiation afin d'objectiver autant que faire se peut l'interprétation des textes. La logométrie, dans sa plus élémentaire ambition, se présente donc comme une méthode qui permet de décentrer l'analyste de ses habitudes de lecture, et de repousser l'entrée dans la subjectivité que toute interprétation requerra. Là où la démarche était 1. Lecture => 2. Interprétation sans autre médiation que la compréhension naturelle de la langue, la démarche devient 1.

Lecture => 2. Décryptage logométrique (index fréquentiel, vocabulaire spécifique, accroissement lexical, calcul de co-occurrences, graphes de distribution, compulsions de concordances, exploration hypertextuelle, etc.) => 3. Interprétation. Ni la lecture naturelle, ni l'interprétation, ni la subjectivité du chercheur ne sont abolies mais elles se trouvent encadrées par une phase de traitements systématiques et exhaustifs de la matière textuelle tels que les logiciels peuvent aujourd'hui les produire.

Valeur descriptive et valeur heuristique du traitement logométrique

Plusieurs décennies après (Robin, 1973) ou (Prost, 1984), on peut reconnaître à la logométrie deux valeurs : une valeur descriptive et une valeur heuristique.

La logométrie sert d'abord à faire une description objective des (macro)-corpus textuels. Les logiciels prennent en compte tous les mots ou toutes les unités linguistiques sans *a priori* sémantique, rhétorique, politique, historique. Ils les classent, les comptent, repèrent ceux objectivement saillants. Ils les donnent à voir sous forme synthétique dans des listes que permet un dépouillement exhaustif. Au fond, la logométrie recompose le texte sous différentes formes (index, tableaux, concordances, graphes) tout en permettant un retour systématique à la forme textuelle d'origine.

La description – ou les moyens de la description – est si puissante qu'elle confine parfois au probatoire. Jacques Chirac aurait utilisé « abracadabrantesque » : un simple clic convoquera la phrase et le discours dans l'océan du corpus chiraquien entre 1995 et 2007. Chirac utiliserait plus souvent « naturellement » que les autres présidents : une simple courbe, après un simple calcul, l'atteste de manière infalsifiable. Favorise-t-il objectivement le présent de l'indicatif dans ses discours ? Les logiciels donneront non pas une réponse intuitive ou impressionniste mais chiffrée difficilement contestable. La description est donc sûre, documentée, chiffrée, comparée au sein d'un corpus ; elle apparaît dans une certaine mesure pouvoir administrer la preuve à des

débats; non pas que les interprétations sociolinguistiques puissent être définitives, mais parce que la base descriptive de cette interprétation cesse d'être sujet à discussion.

Pourtant, la véritable valeur de la logométrie est ailleurs: c'est d'être heuristique. En présentant et lisant les textes différemment, les logiciels nous interrogent différemment, loin des (hypo)thèses convenues. Si la lecture humaine est avant tout syntagmatique (sensible au déroulement), la lecture informatique est paradigmatique (sensible aux parentés). Si la lecture humaine est qualitative, la lecture logométrique est aussi quantitative. Si la lecture humaine est textuelle – au fil du texte –, la lecture logicielle est hypertextuelle – de saut en saut dans le corpus *via* l'entrée désirée. La différence est résumée par (Viprey, 2005) en d'autres termes: la lecture humaine est *linéaire*, la lecture informatique est *tabulaire* et *réticulaire*. La lecture informatique ne vaut pas plus que la lecture humaine – elle vaut même plutôt moins –: c'est le renforcement d'une lecture par l'autre qui est productif.

Les sorties machines, ordonnées et chiffrées du traitement logométrique prennent la forme *d'interrogations*: pourquoi Jacques Chirac utilise-t-il plus souvent « démocratie » – la machine est catégorique et l'œil humain ne pouvait s'en apercevoir – que ces prédécesseurs? Et pourquoi utilise-t-il moins souvent « peuple »? Pourquoi constate-t-on, chez lui, une surutilisation massive des adverbes? Pourquoi son discours est jugé sans parenté – l'indice de distance intertextuelle est sans appel – avec le discours gaullien? Pourquoi?

Dépasser la méthode hypothético-déductive

De la valeur heuristique de la logométrie, de sa puissance interrogative, s'en suit une des ambitions les plus importantes de la démarche: le retournement de la méthode hypothético-déductive habituelle en SHS. Quelle est l'attitude de l'analyste classique face aux textes? Il les lit pour en avoir une *impression*, puis les relit dans l'intention de mettre à l'épreuve ses hypothèses de travail sur le lexique, la grammaire, la rhétorique, le sens. Quoique très répandue, cette attitude est doublement dangereuse.

Le risque, bien connu, est d'abord de toujours trouver ce que l'on cherche. Généralement, la *projection* des hypothèses de travail dans le corpus amène à conclure positivement aux interrogations, comme si nos intuitions étaient toujours les bonnes. Il s'agit en réalité de conclusions ou de constats artefactuels que l'interrogation aura elle-même suscités.

Un autre risque est d'ignorer des faits essentiels du corpus, qui passeront inaperçus faute d'avoir été pressentis comme digne d'intérêt. Obnubilé par le thème de l'insécurité dans le discours de Chirac, le chercheur ignorera le thème – pourtant caractéristique – des jeunes ou de la mondialisation. Et quand bien même ces thèmes émergeraient en cours de lecture, il sera souvent trop tard pour reprendre la lecture du corpus au commencement et mettre à l'épreuve la nouvelle focale.

Quelle est désormais l'attitude d'un analyste outillé par la logométrie face à un texte ? Les informations formelles du corpus remontent en bon ordre (le plus souvent par ordre hiérarchique) pour interroger le chercheur sans *a priori*, bornage ou tabou. Les mots les plus utilisés ou au contraire les moins, ceux objectivement caractéristiques de l'auteur, ceux dont les proximités lexicales sont remarquables, etc. viennent interpeller l'analyste. C'est seulement sur la foi de cette information textuelle dégrossie et triée par ordre de pertinence que l'analyste formulera – et devra formuler ! – des hypothèses qui nécessiteront vérification et retour au texte.

Fondamentalement, nous passons d'une démarche déductive ou *top-down* à une démarche inductive, *bottom-up* ou *corpus-driven*, dans laquelle c'est le texte – dans toutes ses unités et sans sélection ou censure – qui interroge le chercheur et non le chercheur – avec sa part d'aveuglement et de parti pris – qui interroge partiellement et partialement le texte.

Ainsi, non seulement la logométrie permet de contrôler la lecture (qui cesse d'être aléatoire) et les parcours interprétatifs (qui cessent d'être orientés), non seulement elle permet d'interroger le texte avec la rigueur de l'exhaustivité/systématicité, mais elle permet, en préalable, d'encadrer la formulation d'hypothèses de travail, pour en

formuler à la fois de plus sûres et de plus originales. Loin de borner l'analyse, cela signifie que l'idéologie dominante ou les conjonctures scientifiques particulières, qui servent habituellement de matrice inconsciente mais contraignantes à nos interrogations, se trouvent dépassées par une approche à la fois plus objective et plus ouverte des grands corpus textuels.

Connaître les offres logicielles

La principale difficulté peut-être pour les chercheurs est aujourd'hui la diversité des offres logicielles : trop nombreuses, elles finissent par décourager l'utilisateur qui ne sait quels outils choisir. Sans prétendre faire un panorama général des possibilités, quelques indications peuvent être données ici. Sauf exception, les logiciels universitaires, sinon en open source en tout cas en accès quasi gratuit, seront privilégiés, et ceux apportant des gages de qualité et de pérennité pour les années à venir. Par ailleurs, ne perdant pas de vue les exigences d'un public pas toujours averti à la discipline, seront mentionnés les logiciels à l'ergonomie performante qui ne nécessitent pas de compétence informatique particulière.

Lemmatiseurs et analyseurs syntaxiques

Comme on l'a indiqué, le traitement logométrique implique qu'en amont, les textes puissent être lemmatisés, étiquetés, enrichis. Plusieurs offres logicielles existent sur un marché dynamique du fait des capacités toujours plus importantes des machines. Deux logiciels sont non seulement performants, mais ont l'avantage d'offrir des sorties directement articulables sur les logiciels de logométrie. L'utilisateur n'aura donc pas de mal à constituer une chaîne de traitement en trois temps : le texte d'origine (souvent sous format .txt ou .doc) => la lemmatisation et l'étiquetage qui transformeront le texte brut en texte lemmatisé => le traitement logométrique proprement dit.

Le logiciel le plus universel est Tree Tagger. Produit par Helmut Schmid de l'université de Stuttgart, ce logiciel est en *open source*, et se

trouve téléchargeable sur internet (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/). La lemmatisation et le traitement grammatical ne sont pas optimums avec des marges d'erreurs réelles mais le traitement est robuste et a l'avantage de ne jamais bloquer. La force de Tree Tagger est sa gratuité et sa vitesse (quelques secondes pour lemmatiser/étiqueter un corpus de plusieurs millions de mots). Surtout, il fonctionne sur la plupart des langues européennes. Très performant pour l'allemand, il l'est un peu moins pour l'anglais, le français et l'espagnol, et reste à améliorer pour l'italien et le portugais.

Le plus performant des lemmatiseurs pour le français est sans doute le logiciel Cordial Analyseur (commercialisé par la société Synapse Développement) dont le principal inconvénient est le prix (199 euros pour Cordial 2007). S'il ne fonctionne que pour le français, il offre, pour cette langue, une finesse d'analyse remarquable qui explique son adoption par beaucoup de linguistes français. Sa richesse se transforme parfois en défaut puisque non content de lemmatiser et de reconnaître les catégories morphosyntaxiques, instantanément et sans faillir, il met à disposition des utilisateurs des étiquettes sémantiques douteuses ou prétend coder par exemple les registres de langue. La logométrie se méfierait de tels jeux d'étiquettes pour se concentrer sur l'essentiel (graphie, lemme, code grammatical et enchaînement syntaxique).

À côté de ces deux logiciels directement articulables sur les logiciels de logométrie, d'autres logiciels mériteraient d'être mentionnés tels Brill ou Syntex, développé par Didier Bourigault (CNRS – Université de Toulouse le Mirail) dont le traitement des dépendances syntaxiques est d'avenir.

Les logiciels de logométrie

Les logiciels de logométrie sont eux-mêmes très nombreux, fruits d'un développement maintenant pluridécennal. Nous n'en mentionnerons que trois.

HYPERBASE. Conçu par Étienne Brunet à l'université de Nice Sophia Antipolis, Hyperbase apparaît comme un logiciel complet. D'une conception ergonomique classique, il est, sur le marché, le logiciel

qui s'articule le mieux aux deux lemmatiseurs sus-présentés : du texte brut aux sorties statistiques logométriques, l'utilisateur d'Hyperbase se trouve pris par la main, pour un fonctionnement autonome. De plus, le logiciel offre une impressionnante panoplie d'outils statistiques, expérimentée sur des corpus littéraires comme sociopolitiques. Particulièrement adaptée au public non-informaticien, l'utilisation d'Hyperbase essaye d'être intuitive, tendue vers la description-interprétation des corpus textuels SHS.

LEXICO 3. Conçu dès le début des années 1980, Lexico représente peut-être le premier logiciel abouti de lexicométrie. Il est aujourd'hui développé dans sa version 3 par André Salem et son équipe parisienne Syled-Cla2t. Lexico dispose de toutes les fonctions essentielles. Sa première qualité est sa rapidité d'exécution puisqu'il ne faut pas plus de quelques secondes pour traiter de très gros corpus. Son deuxième avantage est son ergonomie moderne. Par un système de glisser-déposer (*drag-and-drop*), les unités (un mot du texte par exemple) sont tirées vers les fonctions pour faire apparaître un graphique, une concordance, une AFC. De plus, un système multifenêtrage rend la session de travail dynamique. La seule faiblesse de Lexico 3 est son incapacité actuelle à traiter les sorties lemmatiseurs. On le favorisera donc seulement pour un traitement des formes graphiques et des segments répétés.

WEBLEX. Conçu par Serge Heiden de l'ENS Lettres de Lyon, Weblex dispose sans doute du moteur de recherche le plus performant, permettant de faire des requêtes complexes (recherche d'expressions régulières, croisement des niveaux linguistiques, choix de l'empan ou de la fenêtre de recherche). Le traitement statistique est poussé, notamment dans le calcul et la représentation des co-occurrences (lexicogramme), mais Weblex n'offre que peu de possibilités de traitements synthétiques (AFC, ACP, Analyse arborée). Le grand intérêt du logiciel est de fonctionner gratuitement en ligne (<http://weblex.ens-lsh.fr/wlx/>). Son principal inconvénient est de ne pas permettre à l'utilisateur de créer ses propres bases de textes sans passer par le concepteur même du logiciel.

Les développements à venir

Conscients de l'éparpillement des forces, particulièrement cruel au moment des mises à jour de chaque produit maison, d'une part, et de la multiplication contre-productive des outils qui égare l'utilisateur d'autre part, les concepteurs des principaux logiciels nourrissent aujourd'hui le projet de fondre dans une « plateforme textométrique » unique et directement accessible sur Internet l'essentiel des fonctionnalités. La conception de cette plateforme fait aujourd'hui l'objet d'un large financement dans le cadre d'un projet ANR « Corpus et outils de la recherche » (2006-2009). Sont impliqués dans le projet, à côté d'Hyperbase, de Lexico 3 et de Weblex, d'autres logiciels à l'identité bien formée : Astartext, DTM, Sato, Xaira.

Outre un développement logiciel important, le projet, piloté par Serge Heiden, espère proposer à terme une synthèse terminologique, un guide d'utilisation pratique et un bilan théorique de la discipline, dont cette contribution peut apparaître comme une version liminaire vulgarisée.

Conclusion : pour une lecture alphanumérique des textes

Deux principes simples président à la démarche logométrique.

Le nombre fait sens. Les régularités et les irrégularités linguistiques, les répétitions, les absences ou les hapax, la richesse lexicale, la distribution étale ou en rafale d'un terme ou d'une catégorie grammaticale, les proximités lexicales ou co-occurrentielles – autant d'éléments non pas intuitifs mais *mesurables* si l'on s'en donne les moyens – font sens. C'est pour cette raison que la logométrie, depuis les premiers travaux dans les années 1960, a établi des algorithmes sophistiqués pour mettre en équation le texte et repérer objectivement grâce à une statistique textuelle de plus en plus précise ce qui relève du simple hasard et ce qui relève d'un choix linguistique significatif. La progression thématique par exemple passe par des *répétitions* rythmées d'indices linguistiques dans le texte ; elle passe par des *corrélations* lexicales que le locuteur

choisit de construire, par des *réseaux* de co-occurrences : la logométrie se propose d'objectiver ces éléments pour contrôler l'interprétation.

Le sens naît en/du contexte. Tout mot doit être contextualisé dans sa phrase ; toute phrase doit être replacée dans son paragraphe ; tout paragraphe situé dans son texte. Plus loin encore : le contexte linguistique est pour nous, non seulement tout le texte, mais le corpus dans lequel chaque texte prend sens (Rastier, 2001). Cette contextualisation, fine ou élargie, ce retour systématique à la phrase, au paragraphe, au texte et au corpus sont favorisés par l'outil, la navigation hypertextuelle, l'extraction de concordances, la lecture de co-occurrences. La logométrie a une finalité herméneutique, et « l'activité interprétative procède principalement par contextualisation » (Rastier, 2001, p. 92) : quel que soit le raffinement des outils statistiques d'analyse qui permettent d'encadrer la démarche, on ne peut faire l'économie d'un retour systématique au texte pour une compréhension-interprétation aboutie.

Répétons pour finir l'essentiel : c'est dans l'intelligence entre ces deux principes (quantification des faits linguistiques/contextualisation linguistique fine ; traitement global et synthétique que permet la statistique/traitement local et particulier que permettent la lecture et le retour au texte) que réside le point fort de la méthode. La logométrie cherche à allier la rigueur mathématique à la posture philologique de l'analyse de texte. Elle cherche finalement à concilier la métrique et le scriptural, les chiffres et les mots. Elle propose, au sens strict, une lecture *alphanumérique* des textes et des corpus.

BIBLIOGRAPHIE

- Adam J.-M. (1999), *Linguistique textuelle. Des genres de discours aux textes*, Paris : Nathan.
- Brunet E. (2001), « Le Logiciel Hyperbase », *Astrolabe*, (revue électronique : <http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm>).
- Charaudeau P., Maingueneau D. (2002), *Dictionnaire d'analyse du discours*, Paris : Seuil.

- Cibois P. (1994), *L'Analyse factorielle*, Paris: PUF.
- Labbé D. (1999), *Normes de saisie et de dépouillement des textes politiques*, Grenoble: Cerat.
- Lebart L., Salem A. (1994), *Statistique textuelle*, Paris: Dunod.
- Mayaffre D. (2004), *Paroles de président. Jacques Chirac et le discours présidentiel sous la Ve République*, Paris: Champion.
- Muller C. (1977), *Principes et méthodes de statistique lexicale*, Paris: Hachette.
- Pincemin B. (2006), « Concordanciers: thème et variations », dans J.-M. Viprey (éd.), *8^e JADT 06* (pp. 773-784), Besançon: Presses universitaires de Franche-Comté, vol. 2.
- Prost A. (1988), « Les mots », dans R. Rémond (éd.), *Pour une histoire politique* (pp. 255-286), Paris: Seuil.
- Rastier F. (2001), *Arts et sciences du texte*, Paris: PUF.
- Robin R. (1973), *Histoire et Linguistique*, Paris: Armand Colin.
- Viprey J.-M. (2005), « Philologie numérique et herméneutique intégrative », dans J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours* (pp. 51-68), Genève: Slatkine.

Fac-similé