Exploration textométrique de l'œuvre de M. Barrès

Virginie Lethier
Université de Franche-Comté/ ELLIADD

Résumé : Cette contribution se donne pour objectif de mettre en relief les apports d'une analyse assistée par les outils informatiques et statistiques à l'analyse du discours littéraire. Héritière de la statistique lexicale et de la lexicométrie limitées à l'entrée de l'unité graphique, la textométrie permet désormais un accès renouvelé au texte et à la pluralité de ses dimensions, jusqu'alors occultées par les artefacts du régime matériel de l'imprimé. Il s'agira ici d'appliquer des méthodologies d'exploration des contrastes grammaticaux et de la co-occurrence à un corpus numérique des œuvres de M. Barrès.

Mots-clés : Corpus, textométrie, analyse du discours littéraire

Resumo : Esta contribuição tem como objectivo realçar os resultados de uma análise assistida pelas ferramentas informáticas e estatísticas a análise do discurso literário. Descendente da estatística lexical e da lexicometria, limitadas a entrada da unidade gráfica, a textometria permite, desde os anos 2000, alcançar a pluralidade das dimensões do texto desde então ocultadas pelos artefactos do processo de impressão. Pretende-se aqui, aplicar essas metodologias de exploração dos constrangimentos gramaticais e das co-ocorrência num corpus numérico das obras de M. Barrès.

Palavras-chave : Análise do discurso literário, corpus, textometrie

Abstract : The purpose of this paper is to highlight the contributions of the textual statistical methods, which produce the emergence of the linguistic, textual and discursive reliefs of the text itself in order to use them as clues for the continuation of the discourse analysis. Exploring a digitalized corpus of the novels produced by the French writer M. Barrès (1862-1923), we aims to illustrate how textometry overtake the traditional entrance of the vocabulary used in lexicometry and its ways to approach textuality.

Key words : Digitalized corpus, lexical statistics, discourse analysis

Introduction

L'analyse des données textuelles assistée par les outils informatiques, initiée en France dans les années 1950, a fait l'objet d'un intérêt tout particulier de J. Peytard. Familier des travaux pionniers de Pierre Guiraud (1954) et de Charles Muller (1964) qu'il cite d'ailleurs dans « D'une Sémiotique de l'altération » (1993), Jean Peytard soutient, à la fin des années 1970, le développement

des recherches textuelles informatisées au Grelis (Groupe de Recherches en Linguistique, Informatique, Sémiotique). On jugera également de cet intérêt, alimenté par ses échanges avec J.-Ph. Massonie[1], en se référant aux pages clôturant le chapitre intitulé « Exemplifications et prospectives des textes littéraires » de l'ouvrage *Discours et enseignement du français* (1992) qu'il publie avec S. Moirand. Dans ces quelques pages, sont interrogés avec acuité les apports de l'outil informatique à une analyse sémiotique du texte littéraire.

Vingt années se sont écoulées depuis cette publication, durant lesquelles l'analyse de données textuelles a tiré profit du développement de la micro-informatique et de l'augmentation des capacités de traitements des logiciels pour opérer un changement de paradigme porté en germes dans la littérature du champ dès la fin des années 1980.

Ce sont les potentialités de l'analyse de données textuelles telles que permises par le passage d'une approche *lexicométrique* à une approche *textométrique* que nous nous proposons d'illustrer dans cette contribution, en rendant compte d'une recherche en cours sur les productions littéraires de Maurice Barrès. Le corpus sur lequel se fonde cette étude résulte en l'occurrence de la mise en série de neuf romans produits par cet écrivain entre 1888 et 1922 : *Sous l'œil des barbares (1888), Un homme libre (1889), Le Jardin de Bérénice* (1891), fondant la trilogie *Le Culte du Moi ; Les déracinés*, (1897), *L'appel aux soldats* (1900), *Leurs figures* (1902). A ces tomes de la trilogie *Le Roman de l'énergie nationale*, s'ajoutent les romans *Colette Baudoche. Histoire d'une jeune fille de Metz* (1909), *La Colline inspirée* (1913) et enfin *Un jardin sur l'Oronte* (1922).

Après avoir opéré un retour réflexif sur la pratique d'analyse des données textuelles assistée par informatique, nous proposerons une exploration de l'œuvre barrésienne focalisée sur les saillances grammaticales et des phénomènes co-occurrentiels, constituant deux nouvelles voies d'accès à la textualité.

1. Retour sur la pratique d'analyse des données textuelles assistée par informatique

Si cette contribution ne saurait être le lieu d'une exposition des principes statistiques qui fondent l'analyse de données textuelles assistée par informatique[2], nous souhaiterions proposer un certain regard sur le statut et le rôle des outils informatiques convoqués dans cette étude.

Des travaux pionniers de Pierre Guiraud aux développements les plus récents de la textométrie, et ce en dépit de divergences d'objets, d'unité de mesure, de gestion du co(n)texte et d'algorithmes, l'emploi de l'outil informatique correspond tout d'abord à la mise en œuvre d'une approche herméneutique contrôlée et *différentielle* des données textuelles. Segmentant, indexant, et comptabilisant les formes d'un corpus avec une rigueur et une constance auxquelles ne peut prétendre l'être humain, l'outil informatique permet d'obtenir des relevés fiables et exhaustifs, servant de base à une approche des contrastes observables au sein d'un corpus. Ces contrastes sont entendus comme autant d'*entailles* où se joue l'accès à la thématique d'un texte, à

sa structure stylistique, aux stratégies énonciatives et socio-discursives qui le déterminent. Science de l'écart par excellence, la statistique sert ainsi une lecture des lieux de variance et de rupture du texte. L'intérêt des méthodes statistiques, et notamment de l'analyse factorielle des correspondances, pointant le « là où ça varie », n'a pas manqué d'être repéré par Jean Peytard. Ce dernier souligne avec justesse que l'intérêt principal de ces méthodes est de contraindre « le chercheur à revenir au texte » (Peytard, 1992 : 214) et cite par ailleurs un passage de Jean-Pierre Massonie que nous reproduirons ci-dessous :

> […] la statistique devient donc l'art et la manière de poser des questions nouvelles à l'endroit qu'il convient dans le texte…Non de résoudre des problèmes, mais de les faire naître par une lecture nouvelle, en indiquant où on doit relire (Massonie, 1990 : 102).

Si nous nous permettons de reproduire ce long passage, c'est qu'il formule de façon très nette les apports profondément heuristiques de l'outil statistique. Ce dernier n'acquiert sa pleine dimension que lorsqu'il est convoqué pour sa capacité à faire émerger du corpus lui-même des parcours de lecture, et que les observations quantifiées, envisagées comme des pistes et non comme des résultats, alimentent la poursuite de l'analyse.

Simple auxiliaire du geste de lecture, en ceci que cet acte implique un mouvement de compréhension et d'interprétation restant nécessairement à la charge de l'analyse, l'outil informatique autorise désormais d'interroger des dimensions de la textualité dont l'examen était jusqu'ici impraticable. Sous l'effet conjoint des progrès technologiques importants ayant eu cours dans la dernière décennie du XX^{e} siècle et d'un dialogue resserré avec la linguistique textuelle[3], un profond renouvellement de l'analyse des données textuelles s'est en effet opéré, marquant le passage d'une démarche dite « lexicométrique » à une démarche « textométrique », dont il s'agit à présent de présenter les possibilités.

2. Analyse de la distribution des parties du discours

Une première évolution significative du passage de la lexicométrie à la textométrie est à trouver du côté de la diversification des unités désormais prises en charge par l'approche statistique. L'approche textométrique se distingue en premier lieu de l'approche lexicométrique, procédant à des relevés quantitatifs sur la seule surface graphique du texte, par sa prise en charge enrichie des unités linguistiques, grammaticales et morpho-syntaxiques tissant le texte. Le développement de logiciels automatiques d'annotation perfectionnés tels que *Cordial*[4], dont les sorties sont directement exploitables par le logiciel de traitement hypertextuel et statistique *Hyperbase*[5], permet de développer des études systématiques et exhaustives des lemmes, des codes grammaticaux et des enchaînements syntaxiques.

La première classe d'analyse destinée à illustrer cette évolution sera ici relative à la distribution des parties du discours dans le corpus des romans de M. Barrès. Comme l'a montré P. Guiraud (1954), les parties du discours variant en fonction des époques, des auteurs et des genres, et constituent des traits encore plus discriminants que le vocabulaire. Convenons d'emblée que l'identification automatique des parties du discours est nécessairement imparfaite : elle présente néanmoins l'avantage d'être cohérente. De plus, le

taux d'erreur inhérent à toute lemmatisation automatique ne saurait suffire à remettre en cause la validité des tendances lourdes dégagées par une analyse statistique d'un grand ensemble de données (Rastier, Malrieu, 2001).

L'examen du classement hiérarchique des catégories grammaticales attestées dans les productions de M. Barrès permet en premier lieu d'observer la richesse en substantifs (24,80%) de notre corpus, observée traditionnellement chez les grands écrivains du XIX^{e} siècle. En comparaison, les verbes et les déterminants représentent respectivement 16% et 15% du corpus. Les catégories des pronoms et des prépositions comptabilisent chacune 12% des effectifs. On notera enfin la domination de la catégorie des adjectifs (7%), particulièrement usités au XIX^{e} siècle, sur celles des adverbes et des conjonctions (5% chacune) et le nombre dérisoire des interjections.

Pour observer comment se distribuent, dans les romans de M. Barrès, les parties du discours, nous avons soumis à un programme d'analyse factorielle des correspondances une vaste matrice présentant, en lignes, les parties du discours identifiées par *Cordial* (adjectif, adverbe, conjonction, déterminant, interjection, préposition, pronom, substantif, verbe) et en colonnes, les romans composant le corpus. Aux intersections des lignes et colonnes de cette matrice, sont enregistrés les effectifs de chaque partie du discours dans chaque roman.

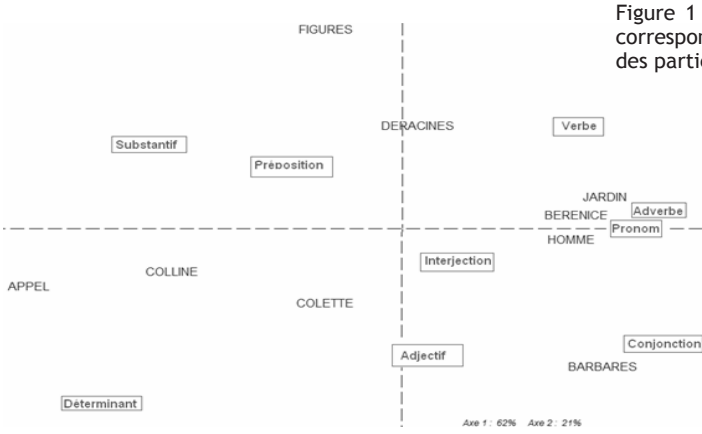

Figure 1 : Analyse factorielle des correspondances de la distribution des parties du discours

Sur la visualisation d'analyse factorielle des correspondances issue de cette procédure (figure 1), le premier facteur (62% de la variance totale) rend compte d'une bipolarisation, traditionnellement observée dans les corpus clos (Brunet, 1985 ; Kastberg, 2002), entre la catégorie nominale et la catégorie verbale. A gauche, le substantif polarise les prépositions et les déterminants, privilégiés dans *La Colline inspirée*, *Colette Baudoche*, *L'appel aux soldats* et *Leurs Figures*. A droite, le verbe attire les adverbes, les pronoms et les conjonctions. C'est dans cette zone que s'ancrent les tomes du *Culte du moi* et *Le Jardin sur l'Oronte*, qui partagent de sous-employer de façon très significative les substantifs (en écarts réduits : -12,3 pour SOB, -13,7 pour HL, -9,1 pour JB ; -8,6 pour JSO), catégorie dont on a pu observer la prégnance dans les récits traditionnels (Brunet, Kastberg 2003).

Ce faisant, l'axe 1 pointe le passage d'un style à tendance verbale, prévalant dans le *Culte du Moi*, à un style excédentaire en substantif, qui est caractéristique des productions ultérieures de M. Barrès. Cette évolution, si elle a été observée chez un certain nombre d'auteurs dont V. Hugo (Brunet, 1988) dont la fréquence d'emploi des substantifs s'intensifie avec le temps, ne peut cependant être rapportée à la seule chronologie. Les valeurs excédentaires de la catégorie verbale dans *Un jardin sur l'Oronte*, dernier roman de M. Barrès, en témoignent et tendent à confirmer que la cause de l'évolution stylistique pointée par l'axe 1 est à chercher du côté des déterminations génériques.

De surcroît, il convient de remarquer la position singulière de l'adjectif sur l'axe 1. Il était en effet attendu que le profil des adjectifs soit proche de celui des substantifs. Or, comme on peut l'observer sur le graphique ci-dessous, les profils distributionnels des deux catégories divergent. Là où M. Barrès active un style à tendance nominale, l'adjectif connaît des sous-emplois significatifs, indices d'un style aride culminant dans *Leurs Figures*. A l'opposé, alors que les substantifs sont déficitaires dans *Le Culte du Moi*, les substantifs y abondent ; le même phénomène est observable pour *Colette Baudoche*.

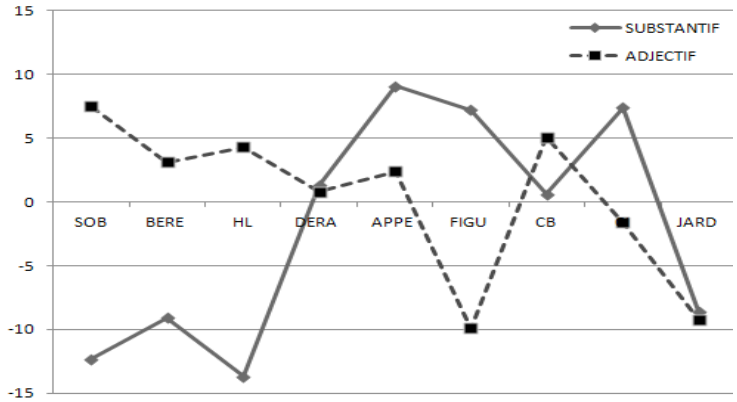

Figure 2 : Distribution des adjectifs et des substantifs dans le corpus

Comme l'indique le retour au texte, les valeurs excédentaires des adjectifs dans la trilogie *Le Culte du Moi* sont à relier à un effet de style à l'œuvre dans cet extrait d'*Un Homme Libre* :

> D'abord un vaste territoire, mon tempérament, produisant avec abondance une belle variété de phénomènes, rebelle à certaines cultures, stérile sur plusieurs points, où des parties sont encore à découvrir, pâles, indécises et flottantes.

Entre autres pistes d'analyse ouvertes par la visualisation AFC considérée (figure 1), on s'attardera sur la ventilation des conjonctions, catégorie que le deuxième facteur participe à situer dans la partie inférieure du graphique, au même titre que les déterminants et les adjectifs. Les conjonctions donnent à voir une intéressante évolution de la syntaxe de M. Barrès et convergent avec le profil des substantifs, précédemment signalé, pour marquer les spécificités de la trilogie *Le Culte du Moi* et d'*Un Jardin sur l'Oronte* dans l'œuvre barrésienne.

Comme on peut l'observer ci-dessous (figure 3), seuls ces romans présente un sur-emploi significatif des conjonctions dans le corpus :

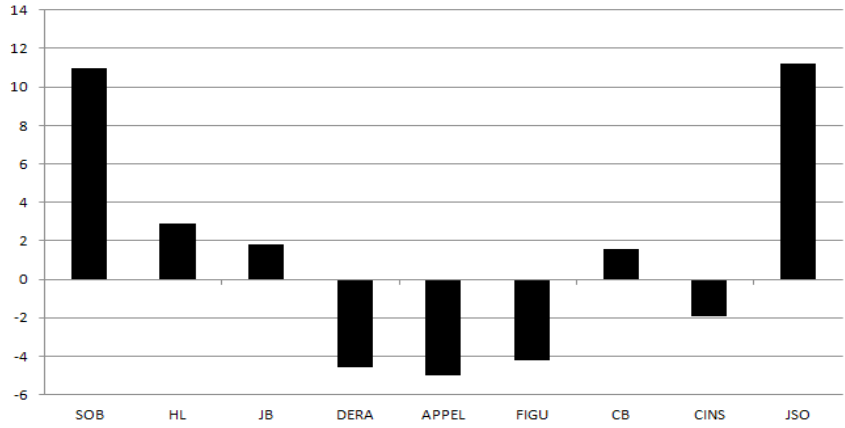

Figure 3 : Distribution des conjonctions (en écarts-réduits)

Ce graphique donne ainsi à lire un abandon progressif des phrases complexes, qui s'opère dès le deuxième tome du *Culte du Moi* : les valeurs excédentaires d'emploi des conjonctions décroissent en effet au fil des tomes de la trilogie (en écarts-réduits : +11,3 pour SOB ; +2,7 pour HL, +1,6 pour JB) jusqu'à devenir non significatives dans les productions ultérieures de M. Barrès. Un retour aux phrases complexes clôture néanmoins l'œuvre barrésienne.

L'examen systématique des types de conjonction employés permet de compléter ces observations : les conjonctions de subordinations[6] (*comme, lorsque, pourvu, puisque, quand, que, quoique, si)* apparaissent systématiquement sur-employées dans les tomes du *Culte du Moi* et dans *Un jardin sur l'Oronte*. Elles sont en revanche sous-employées dans les autres romans. –

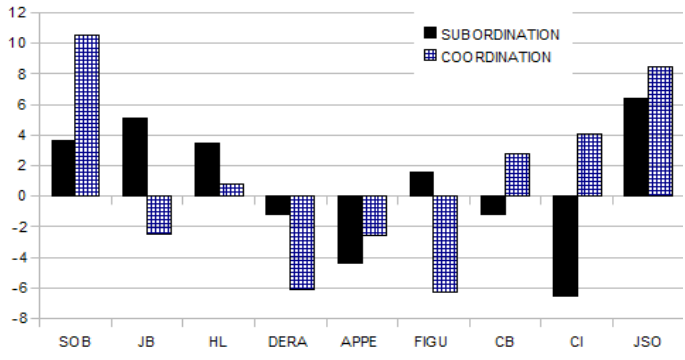

Figure 4 : Distribution des conjonctions de coordination et de subordination (en écarts-réduits)

Concernant les conjonctions de coordination, après une valeur très excédentaire dans *Sous l'œil des barbares*, celles-ci connaissent un profil décroissant, extrêmement marqué dans le *Roman de l'énergie nationale*, pour tendre à s'imposer dans les derniers romans publiés par M. Barrès. Le profil distributionnel de la conjonction copulative *et*, dont on remarquera qu'il représente 76% des effectifs total du groupe des coordonnants, explique cette évolution. Un retour au texte pointe que cette forme se fait le ressort privilégié d'une rhétorique poétique, particulièrement prononcée dans le premier et le dernier roman de M. Barrès. On jugera du fonctionnement de la conjonction *et* dans l'extrait de *Sous l'œil des Barbares* reproduit ci-dessous :

> En sorte que cette constatation même n'est qu'un lieu commun et cet enseignement une vieillerie surannée, et que rien ne vaut que par la forme du dire. Et cette forme, si belle que les plus parfaits des véritables dandies ont frissonné, jusqu'à la névrosthénie, de l'amour des phrases, cette forme qui consolerait de vivre, qui sait des alanguissements comme des caresses pour les douleurs, des chuchotements et des nostalgies pour les tendresses et des sursauts d'hosannah pour nos triomphes rares, cette beauté du verbe, plastique et idéale et dont il est délicieux de se tourmenter, on l'explique, on la démonte...

3. Analyse de la distribution des verbes du corpus

Ne pouvant donner suite à toutes les pistes d'analyse ouverte par l'analyse factorielle de la distribution des parties du discours dans l'espace de cette contribution, nous choisirons de revenir sur la catégorie verbale, sur-employée dans la trilogie du *Culte du Moi* et *Le Jardin sur l'Oronte*. Il peut être jugé surprenant que la catégorie verbale, qui se fait généralement l'indice d'un récit traditionnel dans les corpus littéraires, soit spécifique à cette trilogie par laquelle M. Barrès entend rompre avec les codes du roman et instaurer un nouveau sous-genre romanesque[7], celui du roman *métaphysique,* dans l'espace duquel domineraient les *idées*, et par suite, les substantifs.

Pour en savoir plus sur les verbes utilisés par M. Barrès dans ses productions littéraires, nous avons donc constitué un dictionnaire des verbes du corpus établi à partir du relevé des formes lemmatisées. Le dictionnaire des verbes tient ainsi compte de la dispersion des formes verbales (liée aux temps, aux modes, aux personnes et aux genres) qu'il réunit sous le même lemme.

Nous avons ensuite soumis aux procédures de l'analyse factorielle des correspondances les 400 verbes (lemmes) les plus fréquents du corpus, en vue d'observer les verbes spécifiques à chaque partition du corpus.

Si les verbes sont généralement moins sensibles aux thématiques que la catégorie nominale et subissent essentiellement les contraintes de la structure du récit, on appréciera sur le graphique ci-dessous que la ventilation des verbes fait émerger un regroupement cohérent des trilogies et de la chronologie. Dans la partie supérieure du graphique, dominent nettement les verbes de parole *(annoncer, déclarer, dire, répondre, répéter, raconter, affirmer, crier),* et les verbes se rapportant à une activité physique (*marcher, courir, pousser, jeter*).

L'influence des faits thématiques, dans la partie inférieure du graphique, n'échappera pas aux familiers de *La Colline inspirée* et du *Culte du Moi*. Dans cette zone inférieure, les verbes de perception abondent *(ressentir, entrevoir, sentir)*, de même que ceux renvoyant à une activité psychique *(songer, rêver, imaginer, considérer)*, ainsi que les formes verbales témoignant d'une *sensibilité* physique (*souffrir, pleurer, sourire*). On notera avec attention que si les verbes renvoyant à l'existence (*être, exister, développer*) sont propres à cette zone inférieure, la partie droite n'héberge que les verbes pouvant être rapportés au début du processus (*naître)*, tandis que la fin de celui-ci (*mourir*) se localise à gauche, là-même où s'inscrivent les derniers romans de Barrès, aux côtés des verbes *disparaître, cesser, abandonner*, mais aussi *prier*.

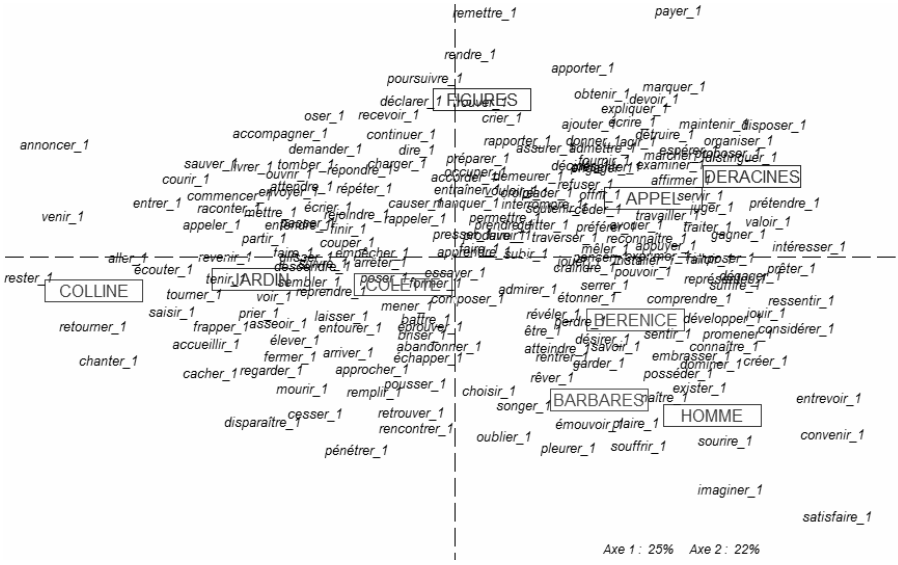

Figure 5: Analyse factorielle des verbes (lemmes) les plus fréquents

Cette analyse permet ainsi de remarquer que si la catégorie verbale est spécifique, dans le corpus considéré, aux tomes du *Culte du Moi* et au *Jardin sur l'Oronte*, les types de verbe utilisés se distinguent nettement de ceux d'un récit traditionnel pour évoquer le profil d'un roman d'analyse, portant les traces d'un retour sur la vie psychique, affective et physiologique au détriment des verbes d'action et de parole.

4. Exploration de l'organisation fine du vocabulaire

En vue d'illustrer une seconde évolution caractéristique du passage de la lexicométrie à la textométrie, nous centrerons le dernier temps de cette contribution sur l'exploration du vocabulaire.

Traditionnellement, la statistique lexicale et l'analyse lexicométrique s'attachent à mesurer le profil distributionnel d'une forme lexicale en procédant à différentes analyses comparatives basées sur les normes endogène ou exogène

du corpus. Les observations ainsi objectivées sur les formes lexicales sont précieuses, mais présentent l'inconvénient majeur d'être déco(n)textualisées, et donc déchargées de leur sens qu'il s'agit de reconstruire par un mouvement essentiel de retour au texte.

La lexicométrie, statistique de l'*occurrence*, fait désormais place à une statistique de la *co-occurrence*, c'est-à-dire de la rencontre de deux unités linguistiques au sein d'un contexte linguistique délimité. L'outil informatique n'est dans ce cadre plus uniquement convoqué pour opérer des relevés de fréquence d'une forme au sein d'un corpus, mais pour mesurer la co-présence des unités. À l'aune du rapport fréquentiel entre deux items co-présents dans le corpus au sein d'une fenêtre co-textuelle délimitée, émergent les phénomènes dynamiques d'associations et de répulsions lexicales tissant le texte. Or, c'est par ses voisinages et ses antagonismes lexicaux, autrement dit par le co(n)texte, que se construit le sens d'un mot, comme l'ont très clairement formulé M. Demonet *et al*. (1975), dont les travaux pionniers portent en germes l'exploitation statistique de la co-occurrence[8]. Effective depuis une dizaine d'années, suite aux contributions majeures de J.-M. Viprey (1997), S. Heiden (2004), et d'E. Brunet (2006), l'exploitation des mesures de co-occurrence marque un profond bouleversement de l'accès à la sémantique pour l'analyse statistique.

Pour témoigner de cette évolution profonde de l'analyse statistique des données textuelles, nous procéderons ici à l'analyse généralisée des cooccurrences attestées dans un corpus résultant de la mise en série des seuls tomes de la trilogie du *Culte du Moi*. Nous utiliserons pour ce faire le logiciel *Astartex*[9]. Notre objectif, ici, n'est pas de rechercher les co-occurrents d'un terme pivot, mais de rendre compte du réseau des co-occurrents de chaque forme lexicale du corpus. L'enjeu de cette classe d'explorations est de mettre au jour, de façon synthétique et globale, l'organisation *non-séquentielle et réticulaire* (Viprey, 2006) du texte. Dans l'esprit de la linguistique distributionnelle inaugurée par Z. Harris (1952), nous visons à mettre en relief des unités de forte équivalence distributionnelle, constituant ce que J.-M. Viprey nomme une *isotropie* et qu'il définit comme un réseau co(n)textuel commun, ou si l'on préfère, un réseau de profils lexicaux collocatifs. Il est important d'insister sur le fort positionnement méthodologique qui préside au concept d'isotropie : *a contrario* d'une projection instable de catégories du *lexique* construites *a priori*, fondées sur le sentiment littéraire et linguistique de l'analyste, le concept d'i*sotropie* relève d'une approche centrée sur le vocabulaire dont l'organisation émerge du texte lui-même. Le concept d'isotropie, au même titre que la démarche d'observation généralisée des cooccurrences, – par opposition à la recherche des co-occurrents d'un terme pivot –, s'ancre donc dans une perspective méthodologique fondamentalement heuristique.

Sur le plan méthodologique, l'examen de la co-occurrence généralisée passe par la construction d'une vaste matrice où figurent en lignes et en colonnes les items les plus fréquents du corpus. En l'occurrence, la visualisation présentée ci-dessous (figure 6) résulte du croisement des 200 substantifs avec, à leur intersection, le nombre de co-occurrences lignes-colonnes dans une unité de contexte déterminée. Le paramétrage de la zone à explorer autour des

formes interrogées susceptibles de fournir une activité pertinente sur l'activité coocurrentielle a été fixé à un empan de 20 mots à gauche et à droite dans les limites de la phrase. Ce paramétrage, comme tout paramétrage, est arbitraire, puisque nul ne voudrait soutenir que l'activité co-occurrentielle s'arrête dans ces bornes artificielles et typographiques. Il nous semble néanmoins justifié par la gestion équilibrée qu'il instaure entre le silence et le bruit dans une perspective d'abord lexicale (et non par exemple micro-syntaxique). La proximité isotropique entre les items est, dans cette perspective, l'indice de profils lexicaux proches. Selon le même principe, le plus ou moins fort éloignement de deux items traduit une dissimilarité des co(n)textes d'occurrences.

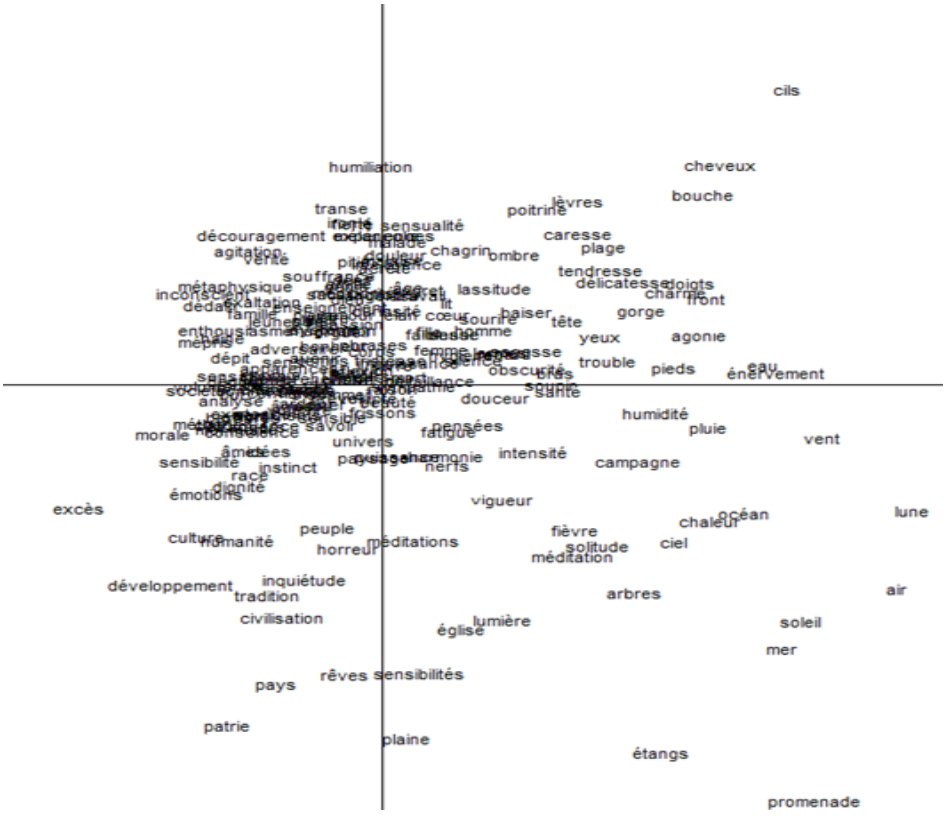

Figure 6 : analyse micro-distributionnelle du vocabulaire du *Culte du Moi*

Sur l'axe 1 du graphique ci-dessus synthétisant l'organisation micro-distributionnelle du vocabulaire de *Le Culte du Moi*, on observera que le vocabulaire abstrait cède progressivement la place au vocabulaire concret. Aux formes marquant des points de fixation de la pensée barrésienne (*métaphysique, exaltation, inconscient, soi-même, sensibilité*) succèdent, à droite, les formes renvoyant au corps humain (*bouche, cheveu, front, figure, poitrine, lèvres, gorge*) et aux éléments naturels *(ciel, océan, soleil*). Sur l'axe 2, structurant verticalement la configuration fine du vocabulaire, l'individu,

à travers le vocabulaire de l'activité psychologique, physiologique et émotive (*souffrance ; exaltation, découragement, humiliation*), s'oppose au collectif à travers les formes *patrie, pays, humanité, civilisation, peuple, race*, elles-mêmes associées à *excès*, à *inquiétude* et *horreur*. Le quadrant inférieur droit du graphique, où s'articulent les éléments lexicaux de la nature à ceux de la *méditation*, suggère que les éléments naturels sont le lieu et le moteur d'un dialogue de l'âme avec l'univers dans le *Culte du Moi*. L'inscription de la forme *promenade* dans cette zone contribue à préciser son sens chez M. Barrès : plus que le temps d'une activité physique, la promenade se fait le temps de l'activité méditative et du retour sur soi[10].

On s'attardera enfin sur les liens de proximité topographique, et par suite sémantique, entre les formes *fièvre*, *méditation*, *vigueur*, situées dans ce même quadrant inférieur droit. On pourrait être en effet étonné de ne pas rencontrer la forme *fièvre* dans le quadrant supérieur droit, là même où s'inscrivent les formes se rapportant au corps. La *fièvre*, moins qu'un signe physiologique, une simple élévation de température, est, dans la trilogie, une *énergie* favorisant la reconstruction du Moi à laquelle participe la nature.

On comparera la structure fine du vocabulaire du *Culte du Moi* (figure 6) à celle du *Roman de l'énergie nationale* (figure 7), obtenue par l'application des mêmes procédures de l'analyse factorielle des correspondances aux 200 substantifs les plus fréquents de ce nouveau corpus. Les pôles isotropiques qui émergent sur ce graphique diffèrent radicalement de ceux de la figure (6). La partie inférieure du graphique est ainsi intégralement structurée autour du vocabulaire politique. Ce dernier est concret (*ministère, république, parlement, gouvernement*) dans la partie gauche du graphique, où l'on devine d'ailleurs certaines lexicalisations par la proximité de leurs constituants : ainsi les items *président*, *conseil*, *commission* semblent pouvoir être interprétés, sans extrapolation, comme la trace des groupes nominaux « le président du conseil », « le président de la commission ». Dans le quadrant inférieur droit, le vocabulaire politique tend à l'abstraction et à la conceptualisation. Ce faisant, le graphique pointe que la *conscience* est désormais liée à la *patrie*, à la *nation*, confirmant la dimension idéologique de la trilogie. De même, l'*énergie* n'est plus personnelle mais résolument nationale et patriotique. L'infiltration dans l'œuvre littéraire de la thèse politique de l'*enracinement* contribue par ailleurs à structurer le quadrant supérieur gauche du graphique : on remarquera ainsi la proximité topographique d'*âme*, de *terre*, et de *sens*, du *passé* et du vocabulaire de la famille (*famille, femme, mère, fils, enfant*).

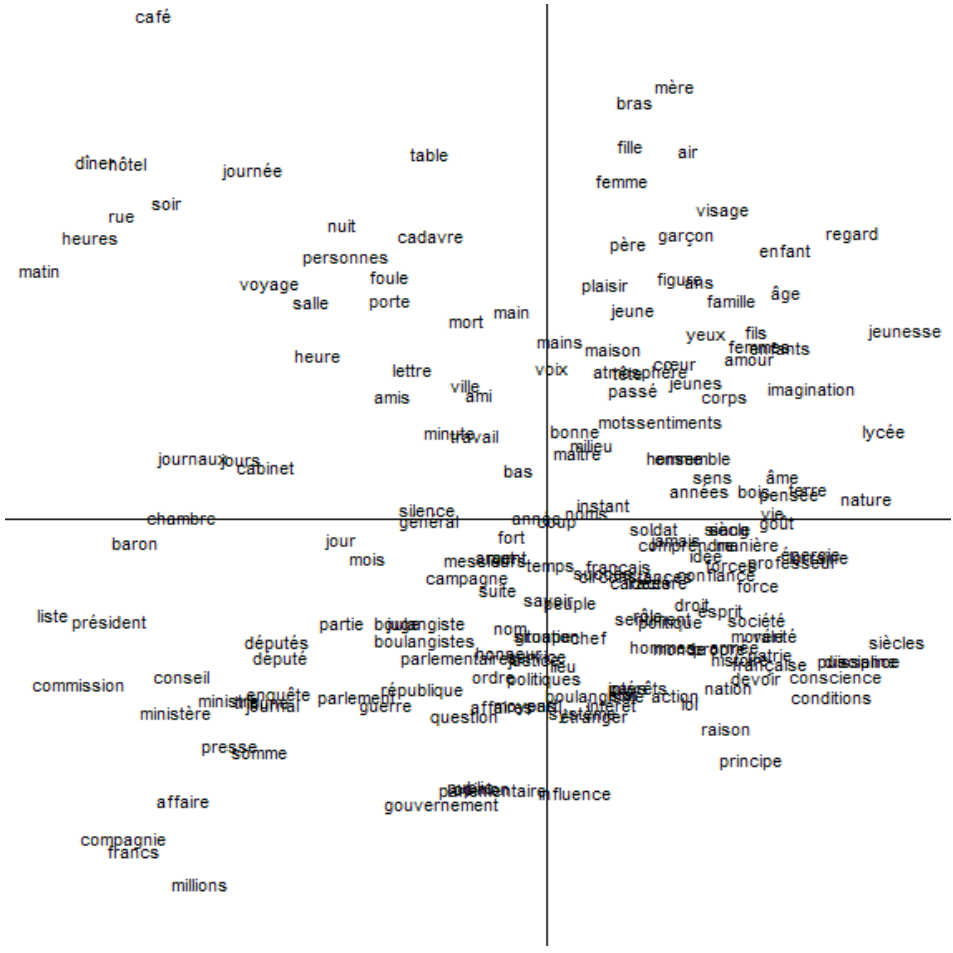

Figure 7 : Analyse microdistributionnelle du vocabulaire du *Roman de l'énergie nationale*

Perspectives

Cette contribution, mettant en relief la puissance heuristique de l'outil statistique à partir d'une étude d'un corpus littéraire, avait pour objectif de montrer les voies ouvertes par la textométrie. Alimentée par les progrès technologiques et portée par une ouverture aux questions herméneutiques et à un effort de définition de l'objet *texte*, la textométrie permet désormais de faire émerger du corpus lui-même un parcours de lecture fondée sur des entailles dont l'examen systématique et cohérent était jusqu'ici hors d'accès. En ouvrant l'accès à l'étude des codes grammaticaux, des structures syntaxiques, des phénomènes de co-occurrence généralisée (mais aussi d'autres indices tels que le rythme du discours, les rimes, etc.), toutes les composantes de l'écriture sont désormais interrogeables. Le champ d'investigation des *entailles* voit ainsi son terrain considérablement élargi.

Visant à illustrer certaines potentialités de la textométrie, le parcours du corpus barrésien proposé ici n'a d'autre prétention que de montrer un certain mouvement d'exploration, par lequel l'outil interpelle l'attention du chercheur sur la matérialité textuelle et lui suggère des pistes d'analyse émergeant des saillances du corpus lui-même. Chacune de ces pistes appellent ensuite de recevoir des analyses fines, où l'acte d'interprétation prend sa pleine dimension.

Notes

[1] Fondateur à Besançon du laboratoire « Mathématique Informatique Statistique ».
[2] Nous renvoyons ici le lecteur à l'abondante littérature disponible sur ce sujet, et notamment à l'ouvrage de référence de Lebart & Salem (1994).
[3] Cf. Adam J.-M. (2006).
[4] *Cordial* est un logiciel développé par la société *Synapse développement*.
[5] *Hyperbase* est un logiciel développé par développé par E. Brunet au sein de l'UMR 6039 « Bases, corpus, langages » (Université de Sophia-Antipolis, Nice).
[6] En raison de l'homographie des éléments de cette liste (*que, comme, quand, si)*, on a veillé à ne convoquer ici que les formes correspondant aux conjonctions par une recherche sur les lemmes.
[7] Cf. M. Barrès, dans l'« Examen des trois romans idéologiques », au sujet de la trilogie du *Culte du Moi* : « J'ai fait de l'idéologie passionnée. On a vu le roman historique, le roman des mœurs parisiennes ; pourquoi une génération dégoûtée de beaucoup de choses, de tout peut-être, hors de jouer avec les idées, n'essayerait-elle pas le roman de la métaphysique ? »
[8] On citera également l'entreprise de « topologie exhaustive des réseaux signifiants » expérimentée par C. Condé et L. Follet (1993) basée sur le recensement automatique des récurrences d'un lien de proximité entre deux lexèmes.
[9] *Astartex* est un logiciel développé au sein du pôle 4 « Archives, Bases, Corpus » de la MSHE par J.-M. Viprey.
[10] Remarquons d'ailleurs que la forme verbale *promener* constituait également une entaille dans la précédente analyse, de par son profil distributionnel singulier synthétisé dans la figure 5. *Promener* était en effet le seul verbe d'activité physique et/ou motrice à s'inscrire dans une zone où dominaient les verbes d'activité psychique et émotive.

Bibliographie

Adam, J.-M. 2006. « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles » in *JADT 2006*. URL: http://lexicometrica.univ-paris3.fr/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf

Brunet, E. 2006. « Navigations dans les rafales » in *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*. Besançon : Presses Universitaires de Franche-Comté, pp.15-29.

Brunet, E. 1985. *Le Vocabulaire de Zola*. Genève-Paris : Slatkine-Champion.

Brunet, E. 1988. *Le Vocabulaire de Victor Hugo*. Genève-Paris : Slatkine-Champion.

Brunet, E. 1999. « Ce que disent les chiffres », in *Nouvelle Histoire de la langue française*. Paris : Le Seuil, pp. 675-727.

Condé , C., Follet, L. 1993. « D'informatique et d'Apollinaire », in *Mélanges offerts à Jean Peytard*, tome 1, pp. 291-314. Besançon : Annales Littéraires de l'Université de Franche-Comté.

Demonet, M., Geffroy, A., Tournier, M. et al. (1975 [1978]) *Des tracts en mai 68. Mesures de vocabulaire et de contenu*. Paris : Presses de la Fondation nationale des sciences politiques.

Guiraud 1954. *Les Caractères statistiques du vocabulaire*. Paris : Presses universitaires de France.

Harris, Z. 1952. « Discourse Analysis » in *Language* 28, 1, pp. 1-30.

Heiden, S. 2004. « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex » in *Actes des 7es Journées internationales d'analyse statistique des données textuelles (JADT 2004)*. Louvain : Presses universitaires de Louvain, pp. 577-588.

Kastberg Sjöblom, M. (2006) *J.M.G. Le Clézio – Des mots aux thèmes*. Paris : Honoré Champion.

Lebart, L., Salem, A. (1994) *Statistique textuelle*. Paris : Dunod.

Madini, M. 2010. « Quelques « lieux de rencontre » de Jean Peytard », in *Semen*, 29, [En ligne]. URL : http://semen.revues.org/8862

Massonie, J.-P. 1990. *Analyse informatisée des textes*. Besançon : Annales littéraires de l'Université de Franche-Comté.

Peytard, J. 1999. « Écriture et pointillés de sens : lecture-analyse de deux pages de Proust (*La Fin de la jalousie*) », in *Semen*, 11, [En ligne]. URL : http://semen.revues.org/2911

Peytard, J. 1993. « D'une sémiotique de l'altération », in *Semen*, 8, [En ligne]. URL : http://semen.revues.org/4182

Peytard, J. & Moirand S. 1992. *Discours et enseignement du français. Les lieux d'une rencontre*. Paris : Nathan.

Rastier, F., Malrieu, D. 2001. « Genres et variations morpho-syntaxiques », in *TAL*, 42, 2, Jussieu : CNRS/ Association pour le Traitement automatique des Langues. pp. 547-577.

Viprey, J.-M. 1997. *Dynamique du vocabulaire des* Fleurs du mal. Paris : Seuil.

Viprey, J.-M. 2006. « Structure non séquentielle du texte » in *Langages*, 163 : «Unités du texte». Paris : Larousse, pp. 71-85.