Vers une automatisation de l'analyse textuelle

Nathalie Garric, Hélène Maurel-Indart

Vers une automatisation de l'analyse textuelle

D'après les Journées d'étude « Le style et sa modélisation »,
10 & 11 décembre 2009,
Université François-Rabelais – Tours

Nathalie Garric,
Université François Rabelais – Tours (EA 3850 LLL)

Hélène Maurel-Indart,
Université François-Rabelais – Tours (EA 2115 Histoire des représentations)

Résumé court pour indexation :

À partir de l'ébauche d'un référentiel stylistique, des chercheurs en littérature, en linguistique et en informatique interrogent les faisabilités textométriques de la formalisation et de la reconnaissance du style en corpus.

Résumé :

Le style est-il modélisable en vue de sa reconnaissance automatisée ? À partir d'une définition textuelle du style et d'une ébauche d'un référentiel stylistique, élaborée par l'étude d'un extrait de *La Princesse de Clèves,* littéraires, linguistes et informaticiens tentent de fournir des éléments de réponse à ce questionnement d'actualité. Les solutions d'automatisation proposées interrogent les ressources textométriques, éventuellement associées à d'autres ressources de traitement des données, tout en questionnant la pertinence des niveaux et des unités d'analyse de la textualité. La réflexion se situe essentiellement dans le champ de la linguistique de corpus et adopte une méthodologie contrastive qui vise à évaluer la distance stylistique dans l'intertextualité afin de formuler des jugements d'identité ou d'altérité stylistique.

Mots clé : analyse stylistique, identité textuelle, textométrie, linguistique textuelle, automatisation.

Abstract :

Is it possible to develop a model for automatic recognition of styles? This question has been the focus of literary scholars, linguists, data processing specialists as well as computer scientists who have been tyring to offer solutions to it while referring to the definition of style in text analysis; and to sketch out the stylistics framework based on the study of an extract of La Princesse de Clèves. Automatization or automatic recognition of styles however, raises the question of textometric resources and other related data processing resources. It also raises the question of the relevant units and levels of analysis in relation to text analysis. This volume presents works that have been carried out within the corpus linguistics framework. They utilize a contrastive methodology in order to assess both the similarity and the difference between texts that supposedly use the same style.

Key words : stylistic analysis, textual identity, textometry, textual linguistic, automatisation

Sommaire

Introduction

Nathalie Garric,

Université François Rabelais – Tours (EA 3850 LLL)

Hélène Maurel-Indart,

Université François-Rabelais – Tours (EA 2115 Histoire des représentations)

Les contributions réunies dans ce recueil de textes s'inscrivent dans un projet de recherche intitulé « Analyse textuelle informatisée pour l'identification du plagiat : similitudes et différences, écart et distance ». Cette recherche a été soutenue par l'Institut des Sciences Humaines et Sociales (CNRS) dans le cadre de l'appel à Projet Exploratoire / Premier Soutien de la campagne 2010 (PEPS). Elle a connu différentes manifestations dont l'organisation de deux journées d'étude, « Le style et sa modélisation », tenues à l'Université François-Rabelais de Tours les 10 et 11 décembre 2009[1]. Cette première étape visait le dialogue interdisciplinaire entre spécialistes aux compétences diverses mais tous, qu'ils soient informaticiens, littéraires, linguistes ou juristes, intéressés par le texte.

Les contributions rassemblées ici résultent de ces journées mais nous n'avons pas souhaité les concevoir comme des actes restituant les communications qui avaient été présentées. Lors de ces deux jours, marqués par de nombreux échanges, tous les participants ont appelé, notamment lors de la table ronde qui a clos la manifestation, à la poursuite de la réflexion sous une forme originale que tente de restituer ce recueil de textes.

La démarche inaugurée au cours des journées d'étude et maintenue dans le cadre de cette publication consiste à développer la problématique annoncée sous la forme d'une exploration plurielle à partir d'un objet d'étude commun et partagé. Une première étape de cette publication consistera à définir la problématique de la recherche en s'attachant à montrer sa progressive maturation. Nous présenterons ensuite l'objet d'étude spécifique que Frédéric Calas et Nathalie Garric ont proposé aux différents intervenants pour mener collectivement la réflexion. Il s'agit d'une ébauche d'un « référentiel texte[2] » construit sur la base d'un ensemble de marqueurs textuels hiérarchisés. Il s'appuie sur l'analyse d'un extrait de la *Princesse de Clèves* de Madame de Lafayette. La visée était de faire émerger à partir de cet extrait les besoins linguistiques de l'identification textuelle et de les soumettre à des spécialistes de textométrie utilisant différents logiciels mais également à des spécialistes de l'analyse textuelle et/ou de l'analyse de discours. Nous verrons quelle méthodologie a été adoptée pour sélectionner ces marqueurs textuels et les articuler entre eux. Une fois présenté l'objet d'étude, s'ouvrira dans un dernier temps un ensemble de réactions et de réflexions, suscitées par une série de questions que les chercheurs se

[1] Nous profitons de cette présentation pour renouveler nos remerciements à l'ensemble des participants à ce projet de recherche et pour leur témoigner, au-delà de notre intérêt, de notre plaisir à avoir partager cette collaboration scientifique.

[2] Il s'agit là d'une dénomination provisoire, dans la mesure où ce premier travail ne recense pas l'ensemble des phénomènes stylistiques potentiels offerts par ce texte. L'objet de cette recherche étant notamment de faire évoluer cette matrice.

sont posées concernant à la fois la pertinence de ces marqueurs textuels et les possibilités de les automatiser.

1. Maturation d'un projet de recherche

1. 1. Origine du projet

Le projet présenté est né de recherches d'abord centrées sur la notion de *plagiat* dans un cadre académique strictement littéraire. Ces recherches étaient initialement menées par Hélène Maurel-Indart.

La notion de *plagiat* permet d'interroger le processus de la création littéraire, en prenant à rebours la question de l'originalité en littérature. Le plagiat renvoie en effet à une zone grise qui va de l'emprunt créatif à l'emprunt servile et qui couvre les différents phénomènes d'intertextualité. Dans *Du plagiat,* H. Maurel-Indart élabore une typologie des différentes formes d'emprunt, depuis l'allusion jusqu'à la réécriture littérale, telle que les Oulipiens ont pu la pratiquer, selon une conception récréative ou recréative de la littérature, ce qui a constitué un premier objectif pour la compréhension des phénomènes observés.

Puis, lors d'analyses comparatives des textes, la nécessité de définir un certain nombre de spécificités stylistiques propres à caractériser chacun d'entre eux s'est imposée. L'idée qu'il serait souhaitable de dresser une sorte de carte d'identité textuelle, susceptible de déterminer en quoi consiste l'« essence » d'un texte est apparue. Cette idée trouvait des échos dans le champ de la recherche : des analyses textuelles informatisées, qui nous ont révélé à la fois de nouvelles pistes d'investigation, mais aussi les risques de dérives de ce type d'expérimentation, ont été appliquées à des corpus littéraires. Certes, les travaux d'E. Brunet de 1983 sur « Proust et Giraudoux », ou de 1985 sur « La phrase de Zola » montraient déjà les ressources précieuses de la lexicométrie. Mais l'étude du même auteur (2003) sur « Flaubert traité par Hyperbase », qui définissait une différence de style entre les œuvres de jeunesse et celles de la maturité d'un même auteur par une approche stylistique de nature plus qualitative, a singulièrement retenu notre attention. Ces travaux d'analyse textuelle informatisée marquaient un nouveau pas par rapport aux analyses précédentes, davantage centrées sur le lexique. Or, comme le remarque M. Kastberg (2003), l'analyse morphosyntaxique est peut-être plus significative des choix stylistiques d'un auteur que l'analyse lexicale qui varie surtout en fonction du thème traité, plus qu'en fonction de la manière d'écrire.

Tout récemment, le premier tome de *Comptes d'auteurs, Etudes statistiques de Rabelais à Gracq* d'E. Brunet (2009) a confirmé à quel point l'outil informatique peut apporter sa contribution à l'analyse littéraire : « Enfin la science de la littérature à laquelle nous aspirons tous (à côté de la critique, quelle que soit son obédience), peut s'appuyer sur des faits tangibles, incontestables et vérifiables », déclare H. Béhar dans la préface de l'ouvrage de Brunet.

D'autres expérimentations – dans un contexte plus polémique – ont été effectuées sur les corpus Gary /Ajar et Molière / Corneille. L'ouvrage d'A. Pawlowski (1998) sur *Les Séries temporelles en linguistique, avec application à l'attribution de textes : Romain Gary et Émile Ajar* prend en compte la linéarité du texte qu'il envisage « comme *série d'éléments* ». Il s'appuie ainsi sur la méthode ARIMA fondée par ses prédécesseurs, George Box et Gwilym Jenkins (1970) qui se démarquent de la méthode dite « assembliste ». La démarche consiste à montrer que « le rythme de la série textuelle est produit avant tout par l'alternance des mots très fréquents (ceux que nous

appelons *grammaticaux*) et des mots moins fréquents (ceux que nous appelons *lexicaux*) » (Pawlowski 1998 : 89). En appliquant successivement le filtre *quantité d'informations* et le filtre *distance entre les mots paramètres,* Pawlowski a tenté d'identifier la spécificité textuelle des œuvres respectives de Gary et d'Ajar, mais aussi de Queneau, de Tournier et d'Aragon qui ont été suspectés, lors de la publication des romans signés Ajar, d'en être les véritables auteurs. Cependant, le résultat de l'expérimentation est peu probant, puisque, sur la représentation factorielle, le corpus Queneau superpose le corpus Gary, tandis que celui d'Ajar ne fait que le jouxter. On peine finalement à reconnaître une même matrice à l'œuvre dans les deux corpus romanesques signés respectivement par Gary et par Ajar.

Paradoxalement, ce qui pourrait apparaître comme une impasse laisse penser que différentes approches automatisées sont envisageables et qu'il serait indispensable de les combiner pour obtenir un outil le plus complet possible, compte tenu de la multiplicité des critères textuels à prendre en compte. L'analyse des œuvres romanesques de Gary et Ajar s'imposera effectivement comme incontournable, car elle présente de nombreux atouts. En effet, il s'agit incontestablement d'œuvres du même auteur, appartenant au même genre romanesque et écrites sur une courte période de cinq ans. L'obstacle de « l'aimantation des genres », maintes fois souligné par E. Brunet, n'est donc pas susceptible de fausser l'analyse, ni le risque d'une évolution de l'écriture de l'auteur, de la jeunesse à la maturité. Les éléments à comparer présentent donc une certaine homogénéité et, même si à l'époque les contemporains de Romain Gary n'ont pas reconnu son style dans ses romans signés Ajar, la critique littéraire a, depuis, répertorié des ressemblances fortes aussi bien du point de vue thématique qu'esthétique. On espèrerait qu'un outil informatique soit capable d'identifier une même identité stylistique par delà les différences, moins substantielles.

L'affaire Molière-Corneille est plus complexe puisqu'elle conduit à comparer des œuvres relevant des genres de la comédie et de la tragédie. Or, J-M. Viprey (2003), rappelle, dans cette affaire tumultueuse, l'influence déterminante du genre sur le lexique, puis celle de la prosodie sur le vocabulaire. Une preuve nouvelle est ainsi apportée de la nécessité d'ouvrir la statistique textuelle à l'analyse de la « phraséologie », et non plus seulement au lexique, ce qui implique la prise en compte, comme nous l'avons déjà souligné, d'une multitude de marqueurs textuels.

1. 2. Nouveau contexte de recherche

Si la notion de *plagiat* a pu constituer un objectif premier dans le cadre de ces travaux initiaux, elle a été progressivement intégrée à la notion plus large d'*identité textuelle*, laquelle actualise simultanément celle d'*individuation textuelle*. Ainsi, la recherche qui était initialement développée dans un cadre strictement littéraire s'est vue, au cours du projet, déplacée vers la linguistique et tout particulièrement vers la linguistique de corpus et vers l'analyse linguistique et/ou textuelle du discours. Ce déplacement a bénéficié de certains changements observables dans le champ des Sciences Humaines et Sociales qui ont pu constituer un contexte favorable à son émergence.

Deux ouvrages témoignent de ce contexte : l'un a été publié en 2003, *L'Analyse du discours dans les études littéraires*, sous la direction de D. Maingueneau et R. Amossy ; l'autre en 2005, *Sciences du texte et analyse de discours*, sous la direction de J-M. Adam et U. Heidmann. Le contexte décrit est caractérisé, non seulement par la multiplication des collaborations interdisciplinaires dans le

champ de l'analyse de discours, mais également par une désacralisation de l'objet des études littéraires. Cet état de la recherche, associé aux innovations technologiques numériques qui ouvrent la disponibilité et la circulation des données, a favorisé et renouvelé la réflexion pour une « herméneutique matérielle » (F. Rastier 2001 & 2003) ou une « nouvelle philologie numérique » (J-M. Viprey 2005). La récente publication issue du colloque international *Linguistique et littérature : Cluny, 40 ans après* (2010), sous la direction de D. Ablali et M. Katsberg-Sjöblom, illustre cette volonté d'une convergence entre littéraires et linguistes pour aborder la textualité. Son inscription historique et l'actualité des questionnements 40 ans plus tard marquent que l'interdisciplinarité attendue des uns n'est pas sans résistance des autres.

Dans le contexte décrit, un autre acteur a pris place : l'informatique et tout particulièrement la lexicométrie. Le trio sous-jacent à ce projet de recherche est ainsi identifié et la description qu'en propose E. Brunet (2010 : 127) permet d'en dresser le portrait :

Beaucoup de gens n'aiment pas s'engager dans un tunnel. À plus forte raison si le tunnel est un tuyau où ne passent que des chiffres. Ceux qui abhorrent la statistique peuvent quitter la salle. Ils ont mon temps de parole pour prendre un café. La linguistique et la littérature sont eux aux deux bouts du tunnel. Aux linguistes le premier tiers de mon exposé : occasion pour les littéraires de rejoindre le café. Le second tiers est voué à la littérature : c'est le moment pour les linguistes de prendre l'air. Le dernier tiers est consacré, s'il reste du monde dans la salle, aux deux disciplines ensemble et aux rapports souterrains ou sous-marins qu'elles entretiennent et que les chiffres aident à découvrir.

Le projet exposé se situe dans ce dernier tiers, peu hospitalier, que nous tentons d'établir en espace fédérateur[3]. Ce déplacement a des conséquences immédiates et fortes sur l'objectif et les différents objets de l'analyse et sur le schéma adopté dans cette publication et pour les deux journées d'étude qui l'ont précédée.

[3] Afin de prendre conscience des difficultés inhérentes à ce type de rencontre interdisciplinaire, nous rapportons les réflexions de Pincemin *et al.* (2008). Les auteurs confrontent les fonctionnalités effectivement offertes par le logiciel *Weblex* avec les usages linguistiques qu'en font les linguistes. A partir de plusieurs études de cas, ils parviennent à la conclusion qu'il existe des écarts notables entre les pratiques des linguistes et les disponibilités de l'analyse textométrique : « pour un connaisseur des outils textométriques […] les possibilités de l'outil sont loin d'être toutes exploitées. Certaines sont méconnues, d'autres sont exclues par l'utilisateur en raison de la crainte de maîtriser imparfaitement les implications des calculs statistiques, d'autres sont utilisées partiellement ou rendues inopérantes en raison de propriétés particulières du corpus ». Enfin, les auteurs soulignent que certains besoins détectés par les linguistes pourraient trouver des solutions textométriques mais restent dans l'ombre parce qu'ils ne sont pas exprimés par l'utilisateur. Ces constats mettent en évidence un état de la recherche souvent caractéristique des projets nécessitant des compétences disciplinaires plurielles qui, comme le soulignent les auteurs, se rejoignent sans se confondre.

2. Un nouvel objet : le style et sa modélisation

Quel est donc le nouvel objectif construit ? Il est d'identifier les conditions de la comparaison textuelle dans le but de formuler un jugement d'identité/non identité textuelle. Nous pouvons le schématiser de la manière suivante :

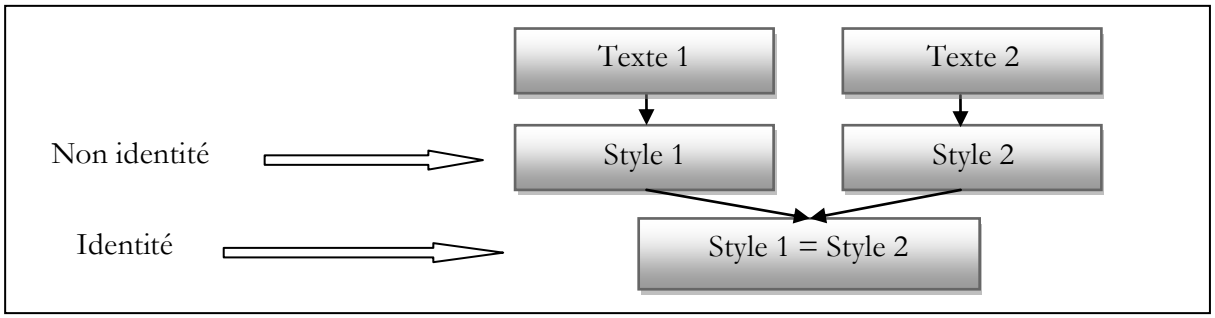

Figure 1 : Jugements de la comparaison textuelle

Étant donné cet objectif, les notions de *texte* et de *style* notamment, mais également les méthodes et les outils utilisés pour mener l'analyse ont dû être réinterrogés. L'œuvre d'un écrivain devient alors un texte et un intertexte, objet empirique méthodologiquement comparable en corpus pour être déconstruit, soumis à des comptages et faire l'objet d'une reconstruction.

2. 1. Redéfinition de l'œuvre littéraire

La première évolution a consisté à accorder, à l'aide de la notion de *texte*, un statut épistémologique à l'œuvre littéraire et simultanément à préciser le cadre théorique de la réflexion. Le texte, dans cette démarche, ne pouvait plus être abordé indépendamment d'une autre notion, celle de *discours*. En effet, différentes dénominations convoquant les deux termes occupent de façon symptomatique le champ de la recherche :

- Analyse de/du/des discours
- Analyse textuelle/des textes
- Analyse textuelle des discours

Les deux premières de ces dénominations scellent la distinction. La dernière, historiquement la plus récente, réunit les deux termes dans une unité syntagmatique. Ainsi, tout aussi nécessaire semble être la distinction introduite ; tout aussi indissociables semblent être les deux réalités désignées. Ce contexte paradoxal justifie probablement l'encre que fait couler la distinction, tout en marquant l'amalgame impossible des deux termes, lequel a pourtant fait partie de leur histoire. J-M. Adam rectifiant, dans sa dernière édition de *Linguistique textuelle* (2004), les formules qu'il avait introduites en 1990 (23) montre la complexité des relations établies entre les deux termes. Les deux formules :

- Discours = Texte + Conditions de Production
- Texte = Discours – Conditions de Production

sont remplacées par le schéma suivant :

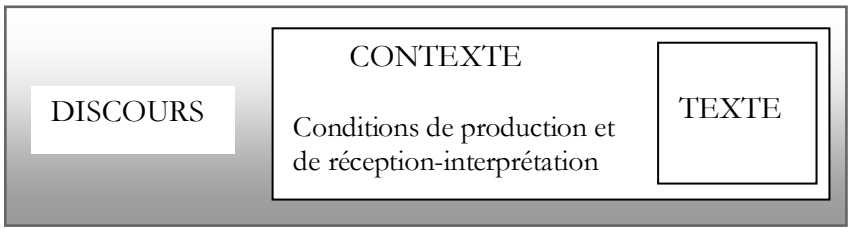

Figure 2 : Le texte et ses extériorités

« Il s'agit d'une formule d'inclusion du texte dans le champ plus vaste de pratiques discursives qui doivent elles-mêmes être pensées dans la diversité des genres qu'elle autorisent et dans leur historicité » (Adam 2004 : 39).

La distinction des deux notions est donc nécessaire notamment pour éviter deux écueils fondamentaux énoncés par J.-M. Viprey (2005) : (1) la définition « d'une science des textes sourde au discours » ; (2) la définition « d'une analyse de discours qui "manque le texte en tant que tel" », l'auteur citant ici G.-E. Sarfati (2003 : 432).

Dès lors, si la notion de *texte* est nécessaire, elle ne doit pas être construite en dissociation d'avec celle de *discours* : le texte, constituant et unité constitutive du discours, existe par lui et celui-ci existe par le texte : aucun n'existe sans l'autre. Se pose alors la question de ce que désigne la notion de *texte*. Nous rapporterons afin d'y répondre deux définitions, celle de J-M. Adam (2004 : 40) et celle de F. Rastier (2001 : 302) :

- « Chaque texte se présente comme un énoncé complet, le résultat toujours singulier d'un acte d'énonciation. C'est, par excellence, l'unité de l'interaction humaine ».
- « suite linguistique autonome (orale ou écrite) constituant une unité empirique, et produite par un ou plusieurs énonciateurs dans une pratique sociale attestée ».

Défini comme objet empirique, le texte n'en reste pas moins « un phénomène construit par une catégorisation métadiscursive » (Cossutta 2004 : 195), d'une part ; et un phénomène variable selon la visée de l'analyse qui détermine différentes formes de corpus intervenant elles-mêmes sur la délimitation textuelle, d'autre part. Le texte est une unité autonome relevant d'une activité interprétative fondée dans la totalité qu'il est, en tant qu'il élabore des cohérences internes construites dans le rapport non seulement local/global mais aussi intra/extra et intertextuel. Un texte se caractérise par des propriétés linguistiques construites à partir de marqueurs de différentes natures et de différents niveaux d'analyse qui trouvent corps ailleurs et autrement sous forme relationnelle, cohésive et contrastive. Il est donc par définition ensemble de régularités et de particularités potentielles diverses dont l'actualisation est contingente au corpus. Le style pourra, selon le corpus effectivement constitué, relever de la similitude ou de la différence.

2. 2. La notion de *style*

La notion de *style* est centrale dans ce projet. Nous abordons cependant avec elle un champ de recherche loin d'être caractérisé par l'unité, comme le montre A. Petitjean (2010 : 245) en concluant le panorama qu'il dresse sur les études du style par ces mots : « le style peut prendre différentes formes de conceptualisation, selon le statut, collectif ou individuel, qu'on lui attribue et duquel dépend la nature des observables ». La tension entre une « stylistique du singulier », définissant un style individuel d'auteur, et une « stylistique du général » (Combe 2002), prônant un style collectif du genre, engendre des positionnements pluriels et divergents.

Étroitement déterminée par la place accordée au texte littéraire et à la stylistique littéraire, la notion de *style* a fait l'objet dans ce projet d'une construction progressive, en relation avec le déplacement rappelé de la littérarité vers la textualité, et la réflexion consacrée à l'élaboration du *Référentiel stylistique*. En parlant de *style*, nous ne faisons pas référence à l'empreinte singulière d'un auteur – encore que la notion d'auctorialité soit liée à celle de style – ou, en d'autres termes, à une appropriation exclusivement individuelle et esthétique de la langue, mais à des spécificités textuelles plus ou moins stables, observables en corpus et construites par lui. Cette conception du style ne se confond pas avec un ornement caractéristique de la littérarité : elle est afférente à tout texte dans lequel le style se réalise en vue d'une finalité et en fonction d'un certain nombre de déterminations portées par la notion de *discours*.

Le travail d'élaboration d'un *Référentiel texte* montre que la notion de *style* est dépendante de celle de *genre*. M. Bakhtine, définissant le genre comme une stabilité relative de l'énoncé « du point de vue thématique, compositionnel et stylistique » (1984 : 269), établit une relation explicite, un « lien indissoluble, organique », entre les deux notions. « Le style entre au titre d'élément dans l'unité de genre d'un énoncé » : il ne peut donc pas être conçu comme manifestation singulière de l'homme puisque sa parole est prise dans les liens dialogiques de l'espace inter-subjectif. Il apparaît ainsi que tout projet de caractérisation textuelle ne peut faire l'économie d'une réflexion sur le genre qui fonctionne à la fois comme composant et interprétant de la textualité. Une autre référence à M. Bakhtine confirme notre propos, tout en explicitant la notion de *textualité* : « les harmoniques dialogiques remplissent un énoncé et il faut en tenir compte si l'on veut comprendre jusqu'au bout le style de l'énoncé. Car notre pensée elle-même – […] – naît et se forme en interaction et en lutte avec la pensée d'autrui, ce qui ne peut pas ne pas trouver son reflet dans les formes d'expression verbale de notre pensée » (1984 : 300).

Le style d'un texte est donc nécessairement tributaire de son genre : il est influencé par ses occurrences génériques, comme l'ont confirmé les travaux d'E. Brunet et de F. Rastier, sans que l'on puisse pour autant, nous semble-t-il, conclure que le style devient tout autre dans la variation générique. Cette relation genre/style permet d'émettre plusieurs hypothèses, parmi lesquelles nous privilégierons la troisième :

- la relation genre/style fonctionne selon une combinaison variable des contraintes du genre
- ou bien on considère que dans cette relation genre/style, il existe un espace de liberté par rapport aux contraintes du genre
- ou enfin, et c'est l'hypothèse médiane que nous privilégions : selon des termes empruntés à P. Charaudeau (1983), la relation genre/style fonctionne dans un espace de liberté surveillée à l'intérieur de l'espace de contraintes génériques.

Cette troisième position rejoint la définition dialectique du style défendue notamment par A. Jaubert (2010) et A. Rabatel (2010) qui permet de penser à la fois la variation et la régularité, la singularité et le social, le particulier et l'universel. La « démarche, qui vise une singularité, ne peut la saisir que si le détail linguistique est rapporté à un ordre régulateur englobant, celui du genre, évidemment historicisé et d'une architecture textuelle qui lui confèrent sa consistance » (Jaubert 2010 : 205). « Le style est donc un des lieux privilégiés d'affleurement de la dynamique de construction/spécification de soi à travers le retravail des formes sociales et culturelles par lesquelles les individus expriment leurs rapports entre eux, leur rapport au monde et leur rapport

au langage, en jouant constamment sur des tensions entre reproduction et innovation » (Rabatel 2010 : 334).

2. 3. Adoption de principes fondateurs

En raison de l'objectif décrit, la perspective lexicométrique et/ou textométrique a été sélectionnée pour mener ce projet de recherche. Lebart et Salem (1994 : 314) la définissent comme « ensemble de méthodes qui permettent d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire à partir d'une segmentation ». À l'aide de comptages d'occurrences d'unités graphiques, elle restitue des décomptes de ces occurrences en référence aux différentes partitions constitutives du corpus. Fondée sur la comparaison statistique et la variation fréquentielle, la procédure lexicométrique répond à certains principes épistémologiques et méthodologiques, décrits notamment par Adam et Heidmann (2003), dans le cadre de l'analyse textuelle des discours, que notre objectif de reconnaissance d'individuations textuelles a identifiés comme fondamentaux. Comme le soulignent les auteurs, adopter la comparaison dite différentielle comme principe heuristique, c'est mettre l'accent sur l'opération de différenciation que constitue tout acte discursif à l'encontre d'un objectif structuraliste d'universalisation de la langue et des textes ; c'est à l'encontre de la tradition littéraire comparatiste, utiliser d'autres textes comme interprétants ou clefs de construction de la textualité sans présumer d'un quelconque rapport hiérarchique entre eux lié à une dévaluation esthétique ou créative de l'un.

On le comprend, la méthode comparative trouve en outre sa justification dans l'assise interdisciplinaire sur laquelle nous fondons notre projet. Et dans cette perspective, en adoptant l'option de la différenciation décrite par U. Heidmann (2005 : 103), « nous nous engageons à construire un axe de comparaison suffisamment pertinent et complexe pour prendre en compte à la fois le trait commun perçu *et* les différences fondamentales des phénomènes à comparer ». Cette volonté est présente dans les éléments d'analyse proposés plus loin qui montrent notamment par l'étude consacrée au discours rapporté, que l'individuation d'un texte ne saurait être identifiée à partir de catégories statiques ou de marqueurs autonomes, puisqu'elle implique une dimension transtextuelle.

La textométrie peut constituer un espace fédérateur entre chercheurs préoccupés par l'objet textuel littéraire mais à condition que les écarts de compétences liés aux spécialités des uns et des autres soient simultanément construits comme objet de réflexion. L'objectif ainsi défini vise à instaurer, pour reprendre un terme emprunté à S. Heiden (2006), une véritable « interopérabilité », non seulement des outils et des données, mais également des compétences.

3. Elaboration de l'outil de travail

L'élaboration du document de travail, adopté comme support des journées d'étude et des contributions ici réunies, ne s'est pas faite sans difficulté. Elle a nécessité, dans le déplacement auquel nous avons fait référence –de la poétique littéraire vers une poétique généralisée pour reprendre une distinction introduite par F. Rastier- une réflexion forte sur les marqueurs dont l'identification ne pouvait se faire qu'en lien avec la notion de *textualité*. C'est pourquoi plutôt que de partir d'un inventaire des différentes unités de l'analyse linguistique, nous avons élaboré le *Référentiel texte* à partir de l'analyse d'un texte, *La Princesse de Clèves* de Madame de Lafayette.

Plusieurs paramètres ont présidé au choix de ce texte. Afin de limiter autant que faire se peut la spirale toujours complexe des arguments de sélection, on s'est arrêté à un consensus autour de trois paramètres :

- le consensus de la critique littéraire sur l'appartenance générique de ce texte à la fiction narrative, c'est-à-dire au roman ;
- le consensus des historiens de la langue sur l'état de langue du texte comme relevant du français moderne ;
- le consensus, à la suite des travaux des sémanticiens, qui recoupent ici aisément celui des littéraires, sur la délimitation relativement aisée des passages dans cette œuvre, en raison de la teneur unitaire de l'isotopie y figurant. C'est ainsi que l'extrait retenu présente une isotopie de la passion saillante, qui a rendu plus repérables les stylèmes et les marqueurs dans ce que nous ne considérons que comme un travail d'approche.

L'objectif, une fois l'analyse stylistique menée, a été de procéder à un travail de déconstruction de cette analyse pour parvenir à l'identification de marqueurs stylistiques que nous avons limités à quelques exemples. Cette démarche pose qu'aucun marqueur n'existe de manière absolue indépendamment de la réception textuelle et surtout de la construction textuelle, autrement dit indépendamment de la mise en série du texte avec d'autres textes. Cette mise en série peut être explicite lors de la construction de corpus, ou implicite lors notamment d'analyses stylistiques qualitatives portant sur un passage comme celle proposée. Elle permet de définir des catégories sémiotiques que nous avons désignées en termes de *stylèmes*. Ces catégories envisagées comme des constructions interprétatives naissent d'une part de corrélations établies dans l'intertextualité à partir des différentes matérialités internes au texte et d'autre part de l'interaction de ces matérialités de nature et de taille variables au cours du processus d'interprétation. La structure du *Référentiel texte* tente de rendre compte de cette complexité de la textualité par la mise en place d'une structure hiérarchique dans laquelle les marqueurs, formes instructionnelles de langue, sont susceptibles de s'associer à des valeurs diverses dans le processus de contextualisation herméneutique.

À partir du document élaboré, des questions ont progressivement émergé lors des discussions et lors de la table ronde, qui ont enrichi, outre les communications et les exposés, les deux journées d'étude sur le style et sa modélisation :

- Pour répondre aux besoins d'une analyse textuelle stylistique, qu'offrent les logiciels existants de lexicométrie ou textométrie ?
- Est-il possible d'automatiser le repérage des indices ? des marqueurs ? des stylèmes ? Quels seraient, le cas échéant, les indicateurs de définition de seuils, si cette notion est nécessaire ?
- Qu'est-ce qui serait automatisable, qu'est-ce qu'il serait possible d'annoter pour guider l'analyse d'un corpus ?
- Quelle serait la pertinence de disposer d'un référentiel (avec ambition ou non d'exhaustivité) d'analyse textuelle en vue de son traitement par les logiciels de textométrie ? Sa structuration hiérarchique est-elle facilitante pour la modélisation ? Jusqu'à quel niveau de hiérarchie peut-on espérer aboutir ?

- En quoi la constitution des corpus est-elle fondamentale dans la démarche ? Est-il possible pour les besoins de l'analyse des marqueurs stylistiques de contraster au sein même d'un texte ? ou le contraste n'est-il pensable et "rentable" que sur des textes externes (intragénériques ou intergénériques) au texte support à l'intérieur d'un grand corpus ? Comment faudrait-il situer l'étape "générique" par rapport à l'étape "stylistique" dans la construction et l'individuation de la textualité ?
- Quelle serait la place d'une analyse stylistique automatisée au sein de l'herméneutique linguistique ?

Les intervenants ont proposé des réponses tantôt consensuelles, tantôt partagées, qui montrent que la réflexion est loin d'être achevée et qu'elle méritera d'être prolongée au fur et à mesure des tests et des études à venir. Michel Bernard envisage une « analyse stylistique outillée » tout en insistant sur la nécessité d'une démarche interdisciplinaire du texte qui associe les fonctionnalités logicielles et les pratiques stylistiques. Etienne Brunet traite les questions soulevées par l'analyse de deux exemples littéraires qu'il soumet aux ressources informatiques, notamment au logiciel *Hyperbase*, en empruntant deux méthodes, l'une documentaire, l'autre statistique. Pascal Marchand interroge les fonctionnalités des outils d'Analyse du discours assistée par ordinateur à partir d'un corpus de politique générale. La notion de *style*, reliée à une évolution chronologique, est traitée conjointement à celle de *genre*, reliée à des déterminations contextuelles, et à celle d'*attitude langagière*, conçue en termes d'adaptation contextuelle. Bénédicte Pincemin aborde les questions soulevées d'un point de vue d'abord pratique. Elle rappelle les fonctionnalités des outils disponibles en insistant sur le gain qu'apporteraient dans le traitement l'introduction de seuils et l'utilisation d'une procédure d'étiquetage des corpus. L'auteur envisage ensuite les enjeux afférents à la pratique herméneutique en lien avec une formalisation du style. François Rastier aborde les différentes questions posées à partir du champ de la linguistique de corpus. L'auteur fonde son argumentation sur une méthodologie contrastive issue de la sémantique différentielle pour construire une coalition de variables destinée à l'interprétation. Max Reinert traite les questions proposées à la réflexion du point de vue de la méthode *Alceste* comme « outil d'aide à la lecture » en proposant auparavant une définition de la notion de *style* et en insistant sur le problème de la place du chercheur dans l'activité d'analyse textuelle.

Avant de présenter l'ensemble de ces réponses, voici le *Référentiel texte* qui a permis de nourrir la réflexion collective.

Bibliographie

Ablali D. et Kastberg-Sjöblom M. (dir.) 2010, *Linguistique et littérature : Cluny, 40 ans après*, Besançon : Presses Universitaires de Franche-Comté.
Adam J-M. 2004, *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.
Adam J-M. et Heidmann U. 2005, *Sciences du texte et analyse de discours,* Genève : Editions Slatkine.
Bakhtine M. 1984, *Esthétique de la création verbale*, Paris : Gallimard.
Box G. et Jenkins G. 1970, *Time series analysis: Fore casting and control*, San Francisco: Holden-Day.
Brunet E. 2010, « Entre linguistique et littérature : un tunnel sous les mots », Ablali D. et Kastberg-Sjöblom M. (dir.), *Linguistique et littérature : Cluny, 40 ans après*, Besançon : Presses Universitaires de Franche-Comté, pp. 127-152.
Brunet E. 2009, *Comptes d'auteurs. Vol. 1 : Etudes statistiques de Rabelais à Gracq*, Paris, H. Champion.
Brunet E. 2003, *Revue Flaubert*, n° 3, http.univ-rouen.fr/flaubert.

Brunet E. 1985, *La Critique littéraire et l'ordinateur / Literary Criticism and the Computer,* Montréal, [2], p. 11-157.
Brunet E. 1983, « Proust et Giraudoux », *Revue d'Histoire Littéraire de la France*, vol. 83, n°5/6, pp. 823-841.
Charaudeau P. 1983, *Langage et discours. Eléments de sémioliguistique*, Paris : Hachette.
Combe D. 2002, « Stylistique des genres », *Linguistique française*, vol. 135, n°1, 33-49.
Cossutta F. 2004, « Catégories descriptives et catégories interprétatives en analyse du discours », Adam J-M., Grize J-B. & Bouacha M-A. (Dir.) *Texte et discours : catégories pour l'analyse,* Dijon : Ed. Universitaires de Dijon, 189-213.
Heiden S. 2006, « Un modèle de données pour la textométrie : contribution à une interopérabilité entre outils », *Proc. of JADT'06 (8èmes Journées internationales d'Analyse Statistique des données Textuelles)*, pp. 487-498.
Heidmann U. 2005, « Comparatisme et analyse de discours. La comparaison différentielle comme méthode », *Sciences du texte et analyse de discours,* Genève : Editions Slatkine, pp. 99-118.
Jaubert A. 2010, « Linguistique et littérature dans le champ des sciences du discours. Vers un nouveau contrat, Ablali D. et Kastberg-Sjöblom M. (dir.), *Linguistique et littérature : Cluny, 40 ans après*, Besançon : Presses Universitaires de Franche-Comté, pp. 197-206.
Kastberg M. 2003, *Encyclopédie de la recherche littéraire assistée par ordinateur* [en ligne], URL : http://www.uottawa.ca/academic/arts/astrolabe/articles/art0032.htm/Comparaison1.htm
Lebart L. et Salem A. 1994, *Statistique textuelle*, Paris : Dunod.
Maingueneau D. et Amossy R. (dir.) 2003, *L'Analyse du discours dans les études littéraires*, Toulouse : Presses Universitaires du Mirail.
Petitjean A. 2010, « Linguistique et littérature : le style en questions, Ablali D. et Kastberg-Sjöblom M. (dir.), *Linguistique et littérature : Cluny, 40 ans après*, Besançon : Presses Universitaires de Franche-Comté, pp. 239-248.
Pincemin B., Guillot C., Heiden S., Lavrentiev A. et Marchellon-Nizia C. 2008, *Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la base de Français Médiéval via le logiciel Weblex*.
Pawlowski A. 1998, *Séries temporelles en linguistique. Avec application à l'attribution de textes : Romain Gary et Emile Ajar*, Paris : Honoré Champion
Rabatel A. 2010, La dialectique du singulier et du social dans l'approche énonciative du style à travers l'articulation des primats et des primautés, des facteurs et des acteurs, Ablali D. et Kastberg-Sjöblom M. (dir.), *Linguistique et littérature : Cluny, 40 ans après*, Besançon : Presses Universitaires de Franche-Comté, pp. 329-340.
Rastier F. 2001, *Arts et Sciences du texte*, Paris : Presses Universitaires de France.
Viprey J-M. 2005, « Philologie numérique et herméneutique intégrative », Adam J-M. et Heidmann U. (dir.), *Sciences du texte et analyse de discours,* Genève : Editions Slatkine, pp. 51-68.
Viprey J-M. 2003, *Morneille, Colière et messieurs Labbé*, http://laseldi.univ-fcomte.fr/morneille.htm, consulté le 11 novembre 2009.

Le style et sa modélisation : éléments d'élaboration d'un référentiel « texte »

Frédéric Calas,

Université Blaise Pascal – Clermont-Ferrand (EA 1002 CELIS)

Nathalie Garric,

Université François Rabelais – Tours (EA 3850 LLL)

Les éléments d'analyse stylistique que nous proposons dans le cadre de cet article sont un support de travail préliminaire en vue de l'élaboration d'un référentiel stylistique. Ils visent à terme une lecture globale de l'objet texte à partir de catégories linguistiques automatisées.

Afin d'élaborer ce document de travail, il nous a paru déterminant d'identifier les catégories que manipule le stylisticien pour construire ses interprétations textuelles en nous appuyant sur une analyse concrète. On a donc procédé à partir d'un passage – qui est un extrait textuel[4] – à la déconstruction systématisée de son analyse, en isolant les procédés qui nous paraissaient constitutifs du passage retenu et donc stylistiquement saillants ou orientés. Ce travail, mené à partir d'un extrait et sans corpus constitué, ne vise aucune exhaustivité des marqueurs envisageables. Il s'agit de montrer, à partir de parcours de lecture spécifiques construits par une procédure en corpus implicite - qui repose notamment sur la connaissance de l'œuvre étudiée et sur celle de textes de déclaration relevant de genres différents, comme la déclaration dans la tragédie classique, le roman par lettres, le sonnet, la comédie – quels sont les différents types de marqueurs de la textualité. La procédure, sous-tendue par une attitude réflexive, a donc privilégié les besoins de l'analyse textuelle sans se préoccuper des faisabilités informatiques[5]. Elle semblait nécessaire afin de parvenir à dresser les impératifs de la reconnaissance de l'identité textuelle, lesquels déterminent les conditions d'un outil adapté qui naîtra d'un travail interdisciplinaire.

Cet article, par un retour réflexif sur l'activité du stylisticien, rappelle d'abord les conditions nécessaires à toute approche stylistique. Il tente ensuite d'identifier dans la matérialité textuelle, contextuelle et intertextuelle, les marqueurs et les constituants supports de l'interprétation. Enfin, il propose une ébauche de référentiel stylistique, limitée à quelques catégories dont l'objectif est essentiellement exploratoire : vérifier l'opérationnalité sur le plan de

[4] Nous reviendrons *infra* sur le choix de ces termes et sur leur justification. Ils sont également liés à la question de la délimitation et des bornes d'un passage. Rappelons ici, sans pouvoir l'utiliser pour l'instant, que certains îlots textuels contiennent des paramètres de frontière forts. Il existe, en effet, en littérature, des unités intermédiaires prédéfinies, comme le paragraphe, la strophe, le sommaire, la scène dialoguée, etc., qui sont délimitées par des indices pertinents, comme la ponctuation, les connecteurs, les anaphoriques, mais aussi les isotopies. D'autres unités sont fixées par la rhétorique comme l'*ekphrasis*, par exemple, et sont également observables sur la base d'indices saillants et récurrents. Ces unités intermédiaires peuvent présenter des caractères textuels forts (séquence descriptive, narrative, dialogale, etc.). Pour obtenir ces unités, on procède par association de marqueurs formant des réseaux de cohérence.

[5] Celles-ci ont fait l'objet des débats qui se sont tenus pendant les deux journées d'étude.

l'automatisation des marqueurs mais également sur celui de l'analyse textuelle de la méthodologie mise en place.

Au-delà de l'objectif restreint à un passage d'une œuvre et non appliqué à une œuvre intégrale, *a fortiori* à une comparaison en corpus, l'ambition de ce travail est d'une part de savoir s'il existe des logiciels qui offrent des outils d'automatisation des marqueurs identifiés, d'autre part, de déterminer s'il est envisageable de travailler sur un panel aussi large de marqueurs que celui convoqué par l'interprétation textuelle en vue de les orienter vers une analyse stylistique.

1. La conduite de l'analyse stylistique

L'analyse stylistique est l'examen des procédés linguistiques mis en œuvre par un écrivain[6], non seulement à des fins communicatives, mais encore en vue de produire un effet esthétique. Elle est sans cesse au service de l'interprétation littéraire du texte, en s'attachant de prime abord aux modalités de l'écriture de l'œuvre, c'est-à-dire à la sélection des mots, des phrases et des procédés rhétoriques au sens large, qui permettent aux auteurs de livrer leur vision du monde, de construire leurs univers et de les faire partager au lecteur.

L'analyse stylistique emprunte à la grammaire, à la linguistique[7] (énonciation, pragmatique, linguistique textuelle, analyse du discours), à la rhétorique, à la poétique et à la sémiotique leurs outils et leurs approches pour décrire l'utilisation qu'un auteur fait de tel ou tel élément langagier. Se pose alors la délicate question de la représentativité d'un procédé et de la valeur qui lui est attachée. Le mot "procédé" est à prendre dans son sens large de "fait observable" à quelque niveau du texte qu'il se présente : procédés de progression textuelle, procédés énonciatifs, lexicaux, grammaticaux et rhétoriques. L'une des difficultés est de définir et de délimiter ces unités, notamment en proposant des catégorisations pertinentes qui non seulement reposent sur les différents niveaux de structuration linguistiques mais également sur les deux faces de ces unités, le plan de l'expression (ou signifiant qui actualise des formes graphiques et/ou phoniques) et le plan du contenu (signifié qui actualise des formes sémantiques). L'analyse repose sur l'articulation des différents niveaux de la matérialité langagière suivants :

[phonique] [graphique] [morphologique] [sémique] [syntaxique] [énonciatif] [pragmatique]

À chacun correspond un trait spécifique générique :

/phonème[8]/ /graphème/ /morphème/ /sème/ /syntagme/ /phrase/ /énoncé/

On part de l'idée que les procédés langagiers ont une valeur[9] ou renferment un potentiel de signifiance identifiable en langue sous la forme d'instructions. L'analyse stylistique s'intéresse aux actualisations de ces instructions dans la totalité textuelle[10] et intertextuelle, prenant ainsi en compte les interactions entre les composants de la textualité.

---

[6] Le travail ne s'applique, à ce stade de la recherche, qu'aux textes littéraires.

[7] On partage le point de vue de Rastier (2001, 3), définissant la stylistique comme le « lieu de rencontre entre la critique académique et les sciences du langage, [elle] est l'endroit privilégié où l'histoire littéraire peut devenir une histoire des formes, des genres et des problèmes esthétiques, en s'appuyant sur l'analyse linguistique des textes. »

[8] Rastier (2007, 33) propose « phème », défini comme « un élément de l'expression, qu'il soit ponctuationnel, phonologique, prosodique, typographie. »

[9] Au sens saussurien du terme.

[10] À la suite de Rastier (2001, 21), on définira le texte comme « une suite linguistique empirique attestée, produite dans une pratique sociale déterminée et fixée sur un support quelconque ». On restreint cependant notre approche, pour l'instant, au support écrit et aux textes littéraires. Le choix du texte littéraire retenu pour l'analyse n'est pas

Toute œuvre appartient à un genre, ce qui revient à dire que toute œuvre suppose un horizon d'attente, c'est-à-dire un ensemble de « règles » servant à orienter la compréhension et l'interprétation du lecteur. Identifier le genre dans lequel s'inscrit l'œuvre est un préalable important à l'étude. En effet, comme le schématise la figure 1, le choix d'un genre littéraire, ou d'un genre discursif de façon plus générale, sélectionne, à un autre niveau que celui du texte, une série de procédés langagiers qui connaissent une réalisation spécifique au sein de chaque genre, engendrant donc des textes différents associables à des styles différents[11].

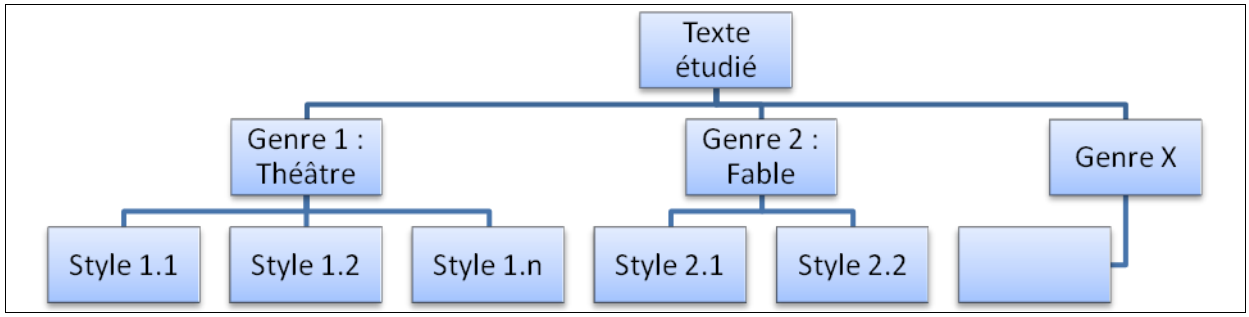

Figure 1 : Déterminations du genre

Ainsi, le thème moral et social du « courtisan » n'est pas traité stylistiquement de la même manière par la comédie de Molière, *Le Misanthrope*, par la fable de La Fontaine, *Les Obsèques de la Lionne*, ni par l'un des *Caractères* de La Bruyère, *De la Cour*. Les codes génériques du genre théâtral, de la fable versifiée ou du genre bref influent, entre autres, sur le traitement de la parole, les positions du narrateur, le traitement du thème, l'ironie ou le comique suivant les cas.

Un autre paramètre important dans le traitement des données est celui du « contexte ». La présence d'un marqueur est – peut-être – en soi significative, mais c'est rarement le cas, ou en tous cas ce n'est pas suffisant. La présence d'un marqueur doit être rapportée au contexte d'insertion de ce marqueur et aux relations qu'il entretient avec d'autres marqueurs, dans le même contexte ou dans des contextes voisins. On entre ici dans une approche « différentielle » du contexte, pour se référer aux travaux de Rastier (notamment Rastier 1987). Reste à définir la taille des contextes et les méthodes de comparaison des contextes et des insertions contextuelles desdits marqueurs. Qui plus est, l'une des nécessités probables sera de faire varier ces contextes en fonction des besoins analytiques. C'est là aussi que se pose la question de la constitution des corpus, et notamment des corpus de référence.

2. Construire l'analyse stylistique

2.1 Les marqueurs linguistiques de l'analyse stylistique

Conduire une analyse stylistique, c'est mettre en relation les procédés relevés les uns avec les autres, pour faire apparaître les enchaînements que le texte unit en profondeur. Cette mise en relation est un programme ou un parcours interprétatif, qui pourrait conduire à l'élaboration d'un

neutre : tout le monde s'accorde pour reconnaître *La Princesse de Clèves* de Madame de Lafayette (1678) comme un texte littéraire, relevant du genre romanesque.

[11] On rejoint ici la perspective ouverte par Pierre Larthomas (1980), sur un exemple développé ressortissant du genre théâtral. Pour suivre les réflexions liées à la constitution d'une stylistique des genres, on lira Seguin (1978) et Guyot (2006).

« cahier des charges[12] » définissant les étapes de ce parcours et les éléments qu'il convient de mettre en relation, ainsi que les niveaux de leur mise en relation[13]. Cette visée requiert de faire émerger de la matérialité textuelle des marqueurs, que l'on peut tenter d'identifier en interrogeant la démarche analytique du stylisticien.

Cette démarche consiste à relever, à identifier et à analyser, le plus techniquement possible sur la base de relevés statistiques (présence *vs* absence, fréquence d'un item), les procédés précis, qui constituent les paramètres fondamentaux des unités « genre », « texte », « discours » (en l'occurrence littéraire). Les procédés à identifier sont nécessairement dépendants de ce que l'on cherche et peuvent être variables selon la conception que l'on se donne du corpus : ce corpus représente-t-il un genre ? Représente-t-il un texte ou un type de textes ? La question au demeurant cruciale de la constitution du corpus ne sera pas abordée de front ici. En revanche, nous ne pouvons négliger de situer le texte dans l'ensemble des faits langagiers. Nous rappelons à cet effet la classification de Malrieu & Rastier (2001, 548) dans laquelle nous introduisons conformément à la description qu'en proposent les auteurs deux niveaux inférieurs :

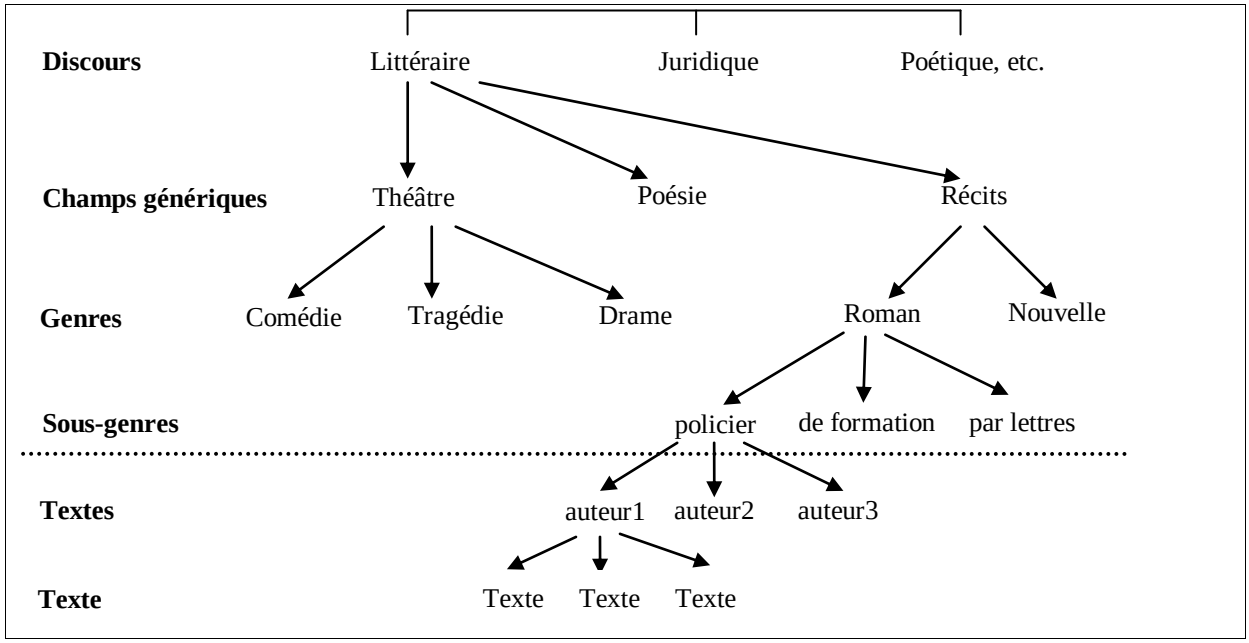

Figure 2 : Niveaux de classification

A partir des marqueurs relevés, il s'agit ensuite de montrer les effets de sens engendrés par ces procédés, d'identifier la valeur du procédé pour l'évaluer quantitativement et qualitativement, ce qui également ne peut être fait (voir *infra*) indépendamment d'une cohérence conférée au texte, celle d'un genre, celle d'un locuteur, celle d'un enjeu discursif, etc.

Quantitativement, c'est-à-dire de manière statistique, il convient de s'interroger sur le degré de présence ou d'absence de telle entrée dans le texte considéré, sur la reprise à haute fréquence, par exemple, de telle figure de style et sur sa répartition dans le texte. L'outil de travail primordial

[12] Dans le cadre restreint du PEPS et des journées d'étude auxquelles il a donné lieu, nous ne sommes pas allés jusqu'à l'élaboration d'un cahier des charges, lequel ne peut s'élaborer qu'en étroite concertation avec les textomètres.
[13] Nous retrouvons ici le questionnement de Pincemin (2007) et de l'ensemble des articles du numéro 6 de la revue en ligne *Corpus*. Pincemin (2007, 5) rappelle que « la question de l'interprétation appelle au plan pratique celle du codage (comment enregistrer, traduire, transcrire) et au plan théorique celle des contextes (quelles sont les unités textuelles et comment entrent-elles en relation). Ces préoccupations sont partagées par l'analyse stylistique systématisée.

repose sur une opération binaire d'opposition et de différenciation des éléments. La répétition (réitération, redondance) d'un élément est un phénomène à observer de très près, car elle se double souvent d'effets d'accumulation dont il conviendra d'observer les divers avatars. La fréquence d'une unité permet de faire apparaître des convergences expressives. La coprésence, tout comme des phénomènes de coréférence avec d'autres marqueurs dans un passage, confèrent à un procédé la signification qu'il déploie précisément dans un texte et constituent simultanément une source d'objectivation de l'analyse.

Qualitativement, il convient d'évaluer la signification d'un procédé remarqué. En effet, toutes les entrées et tous les items du référentiel ne sont pas forcément exploités de la même manière dans le texte. Il faudra choisir à bon escient les rubriques pertinentes pour l'analyse que l'on souhaite faire d'un passage, de l'œuvre entière, de l'ensemble des œuvres d'un corpus. Quel effet dominant est ainsi créé ? Il est tout à fait possible qu'un même procédé serve différents buts. La coprésence, ainsi que les phénomènes de coréférence avec les autres marqueurs du passage, confèrent à un procédé la signification qu'il déploie précisément dans un texte.

2.2 Les marqueurs textuels/génériques de l'analyse stylistique

Comme il a été souligné, un référentiel conçu sur la base des seuls marqueurs linguistiques serait insuffisant pour l'analyse stylistique notamment déterminée par des interprétants génériques. Le travail du stylisticien emprunte à ce niveau deux démarches complémentaires.

L'une **typologique** – qui vise l'identification, au répertoire des genres – littéraires s'entend, si l'on se restreint pour l'instant à ces types de discours et si l'on écarte les genres journalistiques, politiques, juridiques, publicitaires, etc.

L'autre **intra- et intertypologique**, qui étudie la parenté des textes par un travail comparatif (reste à définir s'il est implicite ou explicite), c'est-à-dire si on utilise des caches présentant des textes aux propriétés génériques prototypiques ou non. Dans cette morphologie générale se retrouvent les unités (phonème, morphème, sème, syntagme) ou des unités séquentielles du type de celles que développe Adam (1999).

Rappelons, à la suite de Jakobson (1963) qui proposait de conduire le travail d'analyse en termes de « dominantes » ou à la suite de Rastier (2001), qui suggère de travailler par « faisceaux », que l'identification d'un genre, puis celle d'un texte, suppose que les différents marqueurs soient associés, regroupés, qu'ils forment ce que Garric & Léglise (2005) appellent un « réseau de cohérence », qui alors permettra l'identification ou la refusera. Des indices quantitativement faibles peuvent devenir qualitativement remarquables. Parce qu'ils entrent dans un réseau de cohérence (Garric & Léglise 2005), ils peuvent relever de différents niveaux : un même effet discursif peut trouver ses marqueurs à partir de différentes unités, qui, prises individuellement, peuvent être peu nombreuses, mais saisies dans leur ensemble, engendrent un marquage significatif.

Ce référentiel générique n'est pas explicitement l'objet du travail exploratoire proposé mais le référentiel stylistique ne saurait être utilisé seul, risquant de fausser toute analyse. Le référentiel portant sur le genre est complémentaire et c'est la combinaison des deux référentiels qui confère à l'analyse stylistique son assise. Il pourrait s'apparenter à une sorte de modèle archétypal du genre à étudier qui en dégage les caractéristiques fondamentales et la configuration de ses propriétés.

3. Vers l'élaboration d'un référentiel « texte »

3.1 Choix du texte support[14]

Les jours suivants, le roi et les reines allèrent voir Mme de Clèves. M. de Nemours, qui avait attendu son retour avec une extrême impatience et qui souhaitait ardemment de lui pouvoir parler sans témoins, attendit pour aller chez elle l'heure que tout le monde en sortirait et qu'apparemment il ne reviendrait plus personne. Il réussit dans son dessein, et il arriva comme les dernières visites en sortaient.

Cette princesse était sur son lit, il faisait chaud, et la vue de M. de Nemours acheva de lui donner une rougeur, qui ne diminuait pas sa beauté. Il s'assit vis-à-vis d'elle, avec cette crainte et cette timidité que donnent les véritables passions. Il demeura quelque temps sans pouvoir parler. Mme de Clèves n'était pas moins interdite, de sorte qu'ils gardèrent assez longtemps le silence. Enfin, M. de Nemours prit la parole et lui fit des compliments sur son affliction ; Mme de Clèves, étant bien aise de continuer la conversation sur ce sujet, parla assez longtemps de la perte qu'elle avait faite, et enfin, elle dit que, quand le temps aurait diminué la violence de sa douleur, il lui en demeurerait toujours une si forte impression que son humeur en serait changée.

– Les grandes afflictions et les passions violentes, repartit M. de Nemours, font de grands changements dans l'esprit ; et, pour moi, je ne me reconnais pas depuis que je suis revenu de Flandres. Beaucoup de gens ont remarqué ce changement, et même Mme la Dauphine m'en parlait encore hier.

– Il est vrai, repartit Mme de Clèves, qu'elle l'a remarqué, et je crois lui en avoir ouï dire quelque chose.

– Je ne suis pas fâché, Madame, répliqua M. de Nemours, qu'elle s'en soit aperçue; mais je voudrais qu'elle ne fut pas seule à s'en apercevoir. Il y a des personnes à qui on n'ose donner d'autres marques de la passion qu'on a pour elles que par les choses qui ne les regardent point ; et, n'osant leur faire paraître qu'on les aime, on voudrait du moins qu'elles vissent que l'on ne veut être aimé de personne. L'on voudrait qu'elles sussent qu'il n'y a point de beauté, dans quelque rang qu'elle pût être, que l'on ne regardât avec indifférence, et qu'il n'y a point de couronne que l'on voulût acheter au prix de ne les avoir jamais. Les femmes jugent d'ordinaire de la passion qu'on a pour elles, continua-t-il, par le soin qu'on prend de leur plaire et de les chercher ; mais ce n'est pas une chose difficile pour peu qu'elles soient aimables ; ce qui est difficile, c'est de ne s'abandonner pas au plaisir de les suivre ; c'est de les éviter, par la peur de laisser paraître au public, et quasi à elles-mêmes, les sentiments que l'on a pour elles. Et ce qui marque encore mieux un véritable attachement, c'est de devenir entièrement opposé à ce que l'on était, et de n'avoir plus d'ambition, ni de plaisir, après avoir été toute sa vie occupé de l'un et de l'autre.

Mme de Clèves entendait aisément la part qu'elle avait à ces paroles. Il lui semblait qu'elle devait y répondre et ne les pas souffrir. Il lui semblait aussi qu'elle ne devait pas les entendre, ni témoigner qu'elle les prît pour elle. Elle croyait devoir parler et croyait ne devoir rien dire. Le discours de M. de Nemours lui plaisait et l'offensait quasi également ; elle y voyait la confirmation de tout ce que lui avait fait penser Mme la Dauphine ; elle y trouvait quelque chose de galant et de respectueux, mais aussi quelque chose de hardi et de trop intelligible. L'inclination qu'elle avait pour ce prince lui donnait un trouble dont elle n'était pas maîtresse. Les paroles les plus obscures d'un homme qui plaît donnent plus d'agitation que des déclarations ouvertes d'un homme qui ne plaît pas. Elle demeurait donc sans répondre, et M. de Nemours se fût aperçu de son silence, dont il n'aurait peut-être pas tiré de mauvais présages, si l'arrivée de M. de Clèves n'eût fini la conversation et sa visite.

[14] Madame de Lafayette, *La Princesse de Clèves*, [1678], édition l'Ecole des Lettres, Paris, 1992.

[Justification de la sélection du passage] :

Le passage[15] retenu présente à nos yeux une cohérence forte et une connexité[16] interne et externe également forte ce qui nous a conduit à le sélectionner.

Bornes internes : le passage offre une unité thématique significative, puisqu'on pourrait lui donner un titre (dans un roman balzacien ou hugolien, il pourrait constituer un chapitre, par exemple) « visite de condoléances ». Cette unité se décline à deux niveaux :

celui de la forme ; c'est une séquence « visite de condoléances », située dans la sphère publique (surface) ;

celui du contenu ; c'est une déclaration d'amour indirecte et voilée, située dans la sphère privée (plan visé).

Le texte est encadré par un mouvement de symétrie inversée du passage du « silence » à la parole du duc de Nemours et inversement du passage de la parole au silence chez l'héroïne. Le terme « silence » est ainsi explicitement répété deux fois à 30 lignes d'intervalle. Comme pour une scène de théâtre représentée, le texte signale fortement les entrées et les sorties des personnages-acteurs. Un mouvement vers l'extérieur fait sortir les visiteurs en amont (« il arriva comme les dernières visites en sortaient »), symétriquement un mouvement centripète ramène M. de Clèves, et met un terme à l'entrevue en tête-à-tête (« si l'arrivée de M. de Clèves n'eût fini la conversation et sa visite »). Cette fois, c'est le jeu sur le polyptote « visite(s) » qui assure le bouclage du passage, appuyé par la reprise à l'identique du terme (« conversation »).

Echos et connexité : ce passage si aisément isolable fait écho à d'autres moments significatifs de l'œuvre où se retrouvent des constantes, soit de forme de l'expression, soit de contenu : scène d'entrée dans le monde de la princesse ; scène du bal au Louvre ; scène de l'aveu à son mari ; scène de la visite de condoléances ; écriture de la lettre, ultime conversation chez le Vidame.

[Paramètres génériques] :

Il convient de prendre en considération le genre du texte : le roman. En raison de son aspect multiforme, il est nécessaire de le sous-catégoriser et de parler de « roman d'analyse ». Ainsi, les indices génériques travaillent le texte et partant, l'extrait. Une partie de l'analyse stylistique consistera à faire le va-et-vient entre les contraintes génériques et les « libertés » prises par l'auteur ou entre les attentes génériques et les nouveautés dans le traitement de certaines données. C'est ainsi que toute l'analyse consacrée aux paroles rapportées doit se faire en liaison avec les contraintes génériques propres au roman[17] afin de mesurer quelle est la part d'innovation de *Madame de Lafayette* et quelles sont les spécificités inédites du « psycho-récit ».

[Caractérisation générale] :

Ce passage se caractérise par sa forte cohérence, puisqu'il se lit à deux niveaux à partir du prétexte du « changement » donnant la légende du texte. Ce « changement » d'humeur, d'état et de disposition d'esprit relie explicitement le duc de Nemours (qui connaît une véritable métamorphose depuis sa rencontre avec Mme de Clèves) et celui de la princesse (laquelle est bouleversée par la mort de sa mère à qui elle était très liée). Le thème abordé s'inscrit dans un *topos* littéraire prégnant : la déclaration d'amour. Il joue là encore sur une union d'une forme

[15] Pour faire écho à l'article de François Rastier (2007) intitulé « Passages », nous retenons ce terme et non « extrait » utilisé par la tradition scolaire ou universitaire.
[16] Rastier (2007, 30).
[17] L'analyse est en partie conduite à partir des travaux bien connus de Gérard Genette (1972, 1983), mais aussi à partir des analyses plus récentes de Rosier (1999) et de Rabatel (2003, 2004, 2005) sur la gestion des paroles rapportées.

d'expression (la déclaration – dont les modalités spécifiques renouvèlent le *topos*) et de contenu (la passion – forme exacerbée mais transcendante d'un amour exceptionnel et inédit, dont Mme de Lafayette renouvelle encore une fois le *topos* par le choix d'une situation contraignante : le statut de *Madame de Clèves* et sa haute conception de la parole donnée).

Les sèmes dominants du passage sont déjà présents dans le préambule. On peut isoler :

/amour/ /changement/ /parole/ /peine/ /hyperbolique/ /incapacité/

Il s'assit vis-à-vis d'elle, avec cette crainte[incapacité][hyperbolique] et cette timidité[incapacité] que donnent les véritables passions[amour][hyperbolique]. Il demeura quelque temps sans[incapacité] pouvoir parler [parole]. Mme de Clèves n'était pas moins interdite[incapacité], de sorte qu'ils gardèrent assez longtemps le silence[incapacité][hyperbolique][parole]. Enfin, M. de Nemours prit la parole[parole] et lui fit des compliments[parole] sur son affliction[peine]; Mme de Clèves, étant bien aise de continuer la conversation[parole] sur ce sujet[peine], parla[parole] assez longtemps de la perte[peine][changement] qu'elle avait faite, et enfin, elle dit que, quand le temps aurait diminué la violence de sa douleur[peine][hyperbole], il lui en demeurerait toujours une si forte impression[hyperbolique] que son humeur en serait changée[changement].

3.2 Eléments d'élaboration du référentiel « texte »

Ces considérations introductives, nécessaires à la lecture du passage étudié, étant succinctement rappelées, on présentera les choix méthodologiques retenus pour l'élaboration du référentiel exploratoire.

Le référentiel exploratoire comporte des rubriques qui présentent les procédés constitutifs et spécifiques du « texte ». Il conviendra de les examiner de manière détaillée, car ils forment le matériau brut de l'analyse stylistique. Ces procédés sont constitués de marqueurs que l'on détaille autant que faire se peut et qu'il conviendrait dans un second temps – à l'occasion d'un travail plus fin et plus complexe – de présenter dans un ordre hiérarchisé[18], sur la base d'associations préalablement identifiées.

Dans notre proposition actuelle, les procédés présentés sont identifiés par une catégorie qui reçoit différents marqueurs et que l'on détaille sommairement dans le cadre de ce travail préliminaire. Nous adoptons une présentation hiérarchisée qui progresse d'une catégorie stylistique générique, nommée « stylème », associée à des marqueurs stylistiques, vers des catégories micro-stylistiques (ou catégories linguistiques), associées à des marqueurs micro-stylistiques (ou marqueurs linguistiques, qui trouvent chacun un ensemble d'indices). Selon le degré de complexité du stylème analysé, une ou des catégories de hiérarchie intermédiaire peuvent s'insérer entre les deux niveaux précédents. Nous obtenons une structure à emboîtements, schématisable, et dans laquelle les marqueurs peuvent être plus ou moins nombreux et restent tous susceptibles, à l'exception des marqueurs micro-stylistiques, de donner lieu à une nouvelle catégorie :

[18] Dans les éléments qui suivent un ordre a été retenu. Il mérite cependant discussion et il serait possible de concevoir autrement la hiérarchisation.

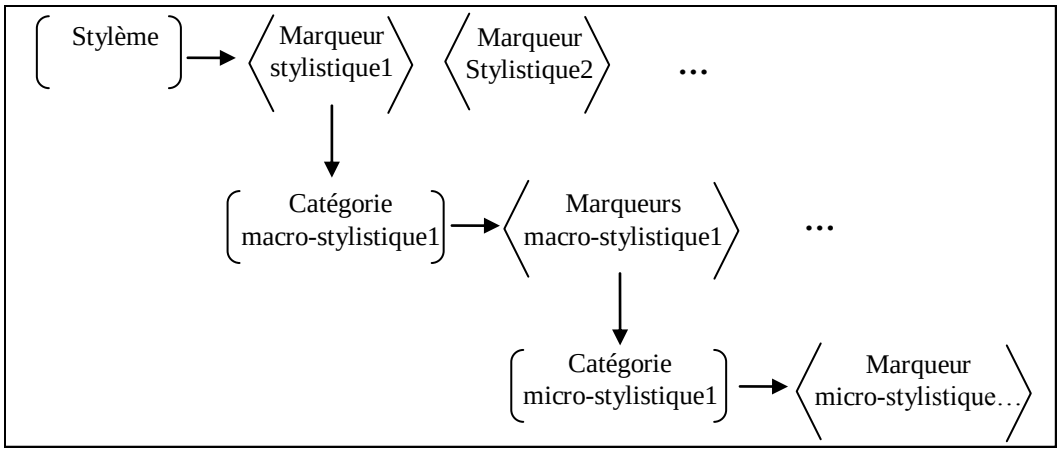

Figure 3 : Déroulé hiérarchique d'un stylème

De par la conception relationnelle et combinatoire du style défendue (voir introduction de ce recueil), aucune unité linguistique ne peut fonctionner comme marqueur dans l'absolu. Un stylème ne peut être constitué que d'unités plurielles – de même niveau ou de niveaux différents – alors identifiées comme marqueurs[19]. Par ailleurs, chacune des unités d'analyse peut être dotée d'une pertinence dans l'unité globale du corpus mais également dans des unités textuelles variables locales, délimitées par et dans le texte lui-même. Par exemple, le *tempo* d'un dialogue sera construit dans le rapport des tours de parole à la séquence dialogique qu'ils constituent. Enfin, le corpus conserve la propriété, par de nouvelles structurations, de devenir unité locale d'un autre corpus.

[19] Voir Molinié (1998, 116-130) sur la caractérisation et le rôle des stylèmes dans l'identification de la textualité d'une part et du style de l'autre. Il rappelle ainsi que « l'identification des stylèmes préalable à l'identification des diverses combinaisons de diverses hiérarchisations stylistiques, ne saurait être menée à bien qu'au prix d'enquêtes sérielles, d'études lancées hypothétiquement sur des modèles stylématiques apparents et provisoires et dont seule la sérialité d'examens appuyés sur toutes les ressources de la répétition permet de tester expérimentalement la plausibilité, la pertinence et la significativité ».

3.3 Analyse sur quelques marqueurs à partir du texte de *La Princesse de Clèves*

Le stylème a été précédemment envisagé en termes de *catégorie stylistique générique*, il ne constitue donc pas une interprétation stylistique finalisée mais une clé possible conduisant à la construction interprétative textuelle. Un même stylème peut donc participer à la définition de plusieurs parcours de lecture en fonction de nombreuses variables, notamment les autres stylèmes actualisés, le genre de l'œuvre, son époque, son auteur, chacune établissant des réseaux intertextuels nécessaires à la textualité.

STYLEME 1

[Stylème] :
[Généralisation]
Ce stylème est construit par des opérations d'identification spécifiques dont la propriété est d'établir une saisie référentielle qui ne permet pas de discriminer le référent d'autres entités. Il peut être construit de manière absolue ou relative, résultant alors d'un processus.
<Marqueurs> :
<Actualisation> <Quantification> <Effacement énonciatif> <Universalisation>

[Macro-catégorie] :
[Actualisation]
<Macro-Marqueurs> :
<Déterminant défini> <Déterminant indéfini>

[Micro-catégorie] :
[Déterminant défini]
<**Micro-marqueurs**> : ..... **Indices**
<Article défini pluriel> ..... *Les*

[Micro-catégorie] :
[Déterminant indéfini]
<Micro-marqueurs> : ..... **Indices**
<Article indéfini> ..... *un, des*
<Déterminant indéfini> ..... *quelque*
<Groupe déterminant> ..... *d'autres*

[Macro-catégorie] :
[Quantification]
<Macro-marqueurs> :
<Déterminant indéfini> <Adverbe> <Pronom indéfini>

[Micro-catégorie] :
[Déterminant indéfini]
<Micro-marqueurs> ..... **Indices**
<Article indéfini> ..... *un, des*
<Déterminant indéfini> ..... *quelque*

[Micro-catégorie] :
[Adverbe]
<Micro-marqueurs> ..... **Indices**
<Adverbe de quantité> ..... *beaucoup de*
<Adverbe de manière> ..... *entièrement*

[**Micro-catégorie] :**
[Pronom indéfini]
<Micro-marqueurs> ..... **Indices**
<Pronom de nullité> ..... *personne*

[Macro-catégorie] :
[Effacement énonciatif]
<Macro-marqueurs> :
<Elocutivité> <Délocutivité>

[Micro-catégorie] :
[Délocutivité]
<Micro-marqueurs> : ..... **Indices**
<Personne 3> ..... *il, elle, lui, elles*
<Présentatif> ..... *Il y a*

[Micro-catégorie] :
[Elocutivité]
<Micro-marqueurs> : ..... **Indices**
<Personne 1> ..... *je, moi, on*
<Verbe de modalité> ..... *vouloir*
<Lexique[affectf]> ..... *aimer, afflictions, passions, violentes*

[Macro-catégorie] :
[Universalisation]
<Macro-marqueurs> :
<Quantification totalisante> <Quantification imprécise> <Sentence>
<Impersonnel>

[Micro-catégorie] :
[Quantification totalisante]
<Micro-marqueurs> : ..... **Indices**

<Déterminant défini pluriel> ..... *les*
<Déterminant[totalité]> ..... *toute*
<Nom[totalité] de> ..... *l'ensemble de*

[Micro-catégorie] :
[Quantification imprécise]
<Micro-marqueurs> : ..... **Indices**
<Article indéfini pluriel> ..... *des*
<Déterminant indéfini[imprécis]> ..... *quelque*

[Micro-catégorie] :
[Sentence]
<Micro-marqueurs> : ..... **Indices**
<Présent générique> ..... *jugent, donnent*
<Infinitif> ..... *éviter, devenir*
<Actualisation générique> ..... *les, la, le*
<Pantonyme> ..... *on, personne, choses, gens*
<Hyperonyme> ..... *les femmes*

[Micro-catégorie] :
[Impersonnel]
<Micro-marqueurs> : ..... **Indices**
<Impersonnel lexical> ..... *il y a, il n'y a*

STYLEME 2

[Stylème] :
[Déséquilibre des voix] :
Ce stylème résulte de phénomènes de décrochages énonciatifs qui associent l'effacement du narrateur et la prise de parole des personnages. Ces phénomènes manifestent la posture des énonciateurs pluriels engagés dans des rapports de domination et/ou d'esquive.
<Marqueurs> :
<Dialogue> <Discours rapporté> <Psycho-récit> <Tempo contrasté> <Dialogisme>

[Macro-catégorie] :
[Dialogue] :
<Macro-Marqueur> :
<Discours direct> <Tours de parole>

[Micro-catégorie] :
[Discours direct]
<Micro-marqueurs> : ..... **Indices**

<Typographie> ..... *Guillemets, tirets cadratins*
<Personnes 1 et 2> ..... *je, tu, nous, vous, on*
<Incise> ..... *verbes*[parole] *+ SN/pro sujet*
<Apostrophe> ..... *Madame*

[Micro-catégorie] :
[Tours de parole]
<Micro-marqueurs> : ..... **Indices**
<Typographie> ..... *retours à la ligne*
<Alternance énonciative> ..... *je/nous/on → tu/vous*
→je/nous/on, adresse directe
<Verbe[parole]> ..... *repartir, continuer, répliquer*

[Macro-catégorie] :
[Discours rapporté]
<Macro-marqueurs> :
<Discours narrativisé> <Discours indirect>

[Micro-catégorie] :
[Discours narrativisé]
<Micro-marqueurs> : ..... **Indices**
<Personne 3> ..... *lui, son, elle*
<Patronyme> ..... *Mme de Clèves, M. de Nemours*
<Nom[parole]> ..... *paroles, compliments, conversation*
<Verbe[parole]> ..... *parler*

[Micro-catégorie] :
[Discours indirect]
<Micro-marqueurs> : ..... **Indices**
<Personne3> ..... *elle, il*
<Verbe[parole]> ..... *dire*
<Complétive> ..... *dire QUE Ph.*
<Concordance temporelle> ..... *aurait diminué, demeurerait*

[Macro-catégorie] :
[Psycho-récit]
<Macro-marqueurs> :
<Narration> <Méta-énonciation> <Analyse>

[Micro-catégorie] :
[Narration]
<Micro-marqueurs> : ..... **Indices**
<Personne 3> ..... *elle, il, lui, ils*
<Personne 1 et 2> ..... *Ø*

<Temps récit> ..... *était, faisait, acheva, demeura*
<déictique> ..... *Ø*
<Référence anaphorique> ..... *les jours suivants*
<Référence absolue> ..... *M. de Nemours, Mme de Clèves*

[Micro-catégorie] :
[Méta-énonciation]
[Micro-marqueurs] : ..... **Indices**
<Nom[parole]> ..... *paroles, conversation, silence*
<Verbe[parole]> ..... *parler, dire*
<Métadiscursivité> ..... *discours, confirmation, répondre, entendre*

[Micro-catégorie] :
[Analyse]
<Micro-marqueurs> : ..... **Indices**
<Modalité épistémique> ..... *croire, sembler, penser*
<Verbe[axiologiques]> ..... *plaire, offenser, souffrir*
<Connecteur> ..... *donc*

[Macro-catégorie] :
[Tempo contrasté]
<Macro-marqueurs> :
<Tour de parole / séquence locale> <Volume>

[Micro-catégorie] :
[Tours de parole]
<Micro-marqueurs> : ..... **Indices**
<Typographie> ..... *retours à la ligne*
<Alternance énonciative> ..... *je/nous/on → tu/vous →je/nous/on, adresse directe*
<Verbe[parole]> ..... *repartir, continuer, répliquer*

[Micro-catégorie] :
[Volume]
<Micro-marqueurs> : ..... **Indices**
<Unité textuelle contenant> ..... *dialogue*
<Tour de parole constitutif> ..... *nombre de phrases/lignes*
<Aspect itératif> ..... *préfixe –RE*
<Aspect duratif> ..... *continuer, longuement*
<Négation verbe[parole]> ..... *sans parler*
<Antonyme nom[parole]> ..... *silence*

[Macro-catégorie] :
[Dialogisme]
<Macro-marqueurs> :
<Renvoi structural> <Renvoi lexical>

[Micro-catégorie] :
[Renvoi structural] :

<Micro-marqueurs> :	**Indices**
<Type textuel>	*maxime La Rochefoucauld, ex : positif/ négatif : « Les femmes croient souvent aimer celui qu'elles n'aiment pas »*

[Micro-catégorie] :
[Renvoi lexical]

<Micro-marqueurs> :	**Indices**
<Unité textuelle contenant>	« *Les passions les plus violentes […] » (Maxime 443)*

[Stylème] :
[Contrariété]
Ce stylème naît de la confrontation inter ou intra-énonciative, il est construit par des manifestations affectives et axiologiques contradictoires, inégalement investies et portées sur l'objet thématique de la prise de parole.
<Marqueurs> :
<Antithèse> <Modalité> <Connecteur> <Rythme>

[Macro-catégorie] :
[Antithèse]
<Macro-Marqueur> :
<Antithèse notionnelle> <Antithèse syntaxique>

[Micro-catégorie] :
[Antithèse notionnelle]

<Micro-marqueurs> :	**Indices**
<Antonyme>	*plaire/ offenser galant, respectueux hardi, trop intelligible*
<Connecteur>	*et*
<Adverbe de comparaison>	*quasi également*

[Micro-catégorie] :
[Antithèse syntaxique]

<Micro-marqueurs> :	**Indices**
<Patron syntaxique>	*reprises syntaxiques et lexicales*
<Connecteur>	*et, mais, ni*
<Adverbe de comparaison>	*aussi*
<Positif/Négatif>	*elle devait… elle ne devait pas*

[Macro-catégorie] :
[Modalités] :
<Macro-Marqueur> :
<Modalité d'énoncé> <Négatif>

[Micro-catégorie] :
[Modalité d'énoncé]

<Micro-marqueurs> :	**Indices**
<Déontique>	*devoir*
<Epistémique>	*croyait*
<Sentence aléthique>	voir ex. en 2.2.2

[Micro-catégorie] :
[Négation]

<Micro-marqueurs> :	**Indices**
<Négation>	*ne, ne pas, ne point, ne rien*
<Négation morpho-lexicale>	*sans répondre, pas maîtresse*

[Macro-catégorie] :
[Connecteur] :
<Macro-Marqueur> :
<Addition> <Opposition>

[Micro-catégorie] :
[Addition]

<Micro-marqueurs> :	**Indices**
<Connecteur d'addition>	*et*

[Micro-catégorie] :
[Opposition]

<Micro-marqueurs> :	**Indices**
<Connecteur d'opposition>	*mais, et + négation*

[Macro-catégorie] :
[Rythme] :
<Macro-Marqueur> :
<Prosodie> <Volumétrie propositionnelle>

[Micro-catégorie] :
[Prosodie]

<Micro-marqueurs> :	**Indices**
<Couplage binaire>	*quelque chose de galant et respectueux/ quelque chose de hardi et trop intelligible*
<Cadence mineure>	*n'eût achevé la conversation et la visite (mineure 9 + 4)*
<Parataxe>	*ponctuation, absence de connecteur interpropositionnel*
<Hypotaxe simple, 1 seul niveau>	*Verbe QUE-P*

[Micro-catégorie] :
[Volumétrie propositionnelle]

<Micro-marqueurs> :	**Indices**
<Décompte syllabique>	
<Isocholie>	*Elle croyait devoir parler et croyait devoir ne rien dire*

6. Conclusion

L'objectif de notre entreprise est de fournir, dans un premier temps, ce référentiel aux textomètres et aux informaticiens, pour étudier comment un certain nombre des éléments présentés dans la hiérarchie descendante retenue peuvent être identifiés par les outils et les procédures logicielles. Cette reconnaissance de marqueurs peut relever de techniques diverses, manuelles ou automatiques, mais elle peut, voire elle doit, s'appuyer sur des opérations statistiques complexes fondées sur des comparaisons statistiques et éventuellement sur un enrichissement progressif des données textuelles. Dans un deuxième temps, sur la base des relevés et sur la base des stylèmes ainsi décomposés, le stylisticien doit pouvoir proposer une interprétation de ces faits et faisceaux de langue et rendre compte des aspects saillants du style d'une œuvre.

7. Bibliographie

Adam, J.-M., (1999) *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

Bernard, M. (2006), « Transcription phonétique des grands corpus littéraires. Les règles du jeu », *Corpus*, n°5, décembre 2006, pp. 143-158, en ligne http://corpus.revues.org/index474.html

Calas, F., (1998) « Les impostures des genres : le cas du roman par lettres », dans *Analyse des discours. Types et genres : communication et interprétation*, actes du colloque international de Toulouse, décembre, *Genres, types, textes*, collection « Champs du signe », sous la direction de Michel Ballabriga, P.U.M., 2001, pp. 359-372.

Calas, F., (1999) « "Petit modèle épistolaire", de la poétique à la stylistique des genres », *Le Français moderne*, Paris, juin, n°1, Tome LXVII, pp. 61 à 80.

Charaudeau, P. & Maingueneau, D., éd. (2002), *Dictionnaire d'analyse du discours*, Paris, Seuil.

De Boissieu, J.-L. et Garagnon, A.-M. (1987), *Commentaires stylistiques*, Paris, Sedes.

Garric, N. & Léglise, I. (2005) « La place du corpus, de l'analyste, du logiciel : exemple d'une analyse de discours patronal à deux voix », *Linguistique de corpus*, G. Williams (eds), Presses universitaires de Rennes, pp. 101-113.

Genette, G. (1972), *Figures III,* Paris, Seuil, coll. « Poétique ».

Genette, G. (1983), *Nouveau discours du récit,* Paris, Seuil, coll. « Poétique ».

Genette, G. (1987), *Seuils*, Paris, Seuil, coll. « Poétique ».

Genette, G. (1991), *Fiction et diction*, Paris, Seuil, coll. « Poétique ».

Guyot, A. (2006), « Stylèmes et corpus génériques : un essai de confrontation au service de la stylistique des genres », *Corpus*, n°5, décembre 2006, en ligne, http://corpus.revues.org/index472.html

Habert, B. (2005) « Portrait de linguiste(s) à l'instrument », *Texto !*, décembre 2005, vol X, n°4, en ligne http://www.revue-texto.net/corpus/publications/habert

Jakobson, R. (1963), *Essais de linguistique générale*, Paris, Minuit.

Larthomas, P. (1980), *Le Langage dramatique*, Paris, P.U.F.

Malrieu, D. & Rastier, F. (2001), « Genres et variations morphosyntaxiques », *Traitement automatique des langues*, vol. 42, n°2, pp. 548-577.

Molinié, G. (1998), *Sémiostylistique,* Paris, PUF.

Pincemin, B., (2007) « Introduction », *Corpus*, «Interprétation, contexte, codage», n°6, décembre 2007, pp. 5-15, en ligne http://corpus.revues.org

Pincemin, B., (2008) « Modélisation textométrique des textes », JADT 2008, 9èmes Journées internationales d'Analyse statistique des Données Textuelles. En ligne.
Rabatel A. (2003a), « Un paradoxe énonciatif : la connotation autonymique représentée dans les "phrases sans parole" stéréotypées du récit », dans « Le Fait autonymique : langage, langue, discours », in *Parler des mots : le fait autonymique en discours*, J. Authier-Revuz, M. Doury S. Reboul-Touré (éd.), Paris, Presses de la Sorbonne Nouvelle, pp. 271-280.
Rabatel A. (2003b), « L'effacement énonciatif et ses effets pragmatiques de sous- et de sur-énonciation », *Estudios de lengua y literatura francesas 14*, Université de Cadix, pp. 33-61.
Rabatel A. (2004), « L'effacement énonciatif dans les discours rapportés et ses effets pragmatiques », dans *Effacement énonciatif et discours rapportés, Langages*, n°156, pp. 3-17.
Rabatel A. (2005), « La part de l'énonciateur dans la construction interactionnelle des points de vue », *Marges linguistiques*, en ligne www.marges-linguistiques.com, pp. 115-136.
Rastier, F. (1987), *Sémantique interprétative*, Paris, PUF.
Rastier, F. (1989), *Sens et Textualité*. Paris, Hachette.
Rastier, F. (1994), « Le problème du style pour la sémantique du texte », Molinié G. et Cahné P. (éds.), *Qu'est-ce que le style ?*, Paris, P.U.F.
Rastier, F. (2001) « Eléments de théorie des genres », *Texto !*, juin 2001, en ligne http://www.revue-texto.net/inedits/rastier
Rastier, F. (2007) « Passages », *Corpus*, n°6, « Interprétation, contextes codage », pp. 25-54.
Rastier, F. (2009) « Hérodias – Intertexte et genèse de formes sémantiques », *Texto !*, vol. XIV, n°3, en ligne http://texto-revue.net
Rastier, F. & Pincemin, B. (1999) Des genres à l'intertexte, *Cahiers de praxématique*, 23, pp. 90-111.
Rosier, L. (1993), « L'incise *dit-elle*, ou l'attribution du dire en discours apporté (le paradigme *dit-il*) », *Actes du XXe congrès international de linguistique et philologie romanes*, Tübingen, Francke Verlag, pp. 656-667.
Rosier, L. (1999), *Le Discours rapporté. Histoire, théories, pratiques,* Paris-Bruxelles, Duculot, coll. « Champs linguistiques ».
Seguin, J.-P. (1978), *Diderot, le discours et les choses*. *Essai de description du style d'un philosophe en 1750*, Paris, Klincksieck.

Spécificités lexicales d'un sous-corpus : quel(s) corpus de référence ?

Michel Bernard

Professeur à l'Université Sorbonne Nouvelle – Paris 3

Résumé :

Les questions soumises à l'examen sont envisagées à partir des disponibilités logicielles textométriques et des pratiques informatiques sur le texte littéraire. Des ressources existent en partie mais elles doivent être associées à une perspective méthodologique et heuristique reposant sur une procédure comparative historicisée et sur une conception hiérarchique des indices textuels.

Mots clés : Analyse stylistique, textométrie, variation, hiérarchie, comparaison textuelle.

Abstract :

The issues addressed here are considered from the point of view of the availability of textometric software and of computer processing of literary texts. Resources do exist but they must be associated with a methodological and heuristic perspective, itself based on a historicized comparative procedure and on a hierarchical conception of textual indications.

Key words: stylistic analysis, textometry, variation, hierarchy, textual comparison

1) Pour répondre aux besoins d'une analyse textuelle stylistique, qu'offrent les logiciels existants de lexico ou textométrie ?

Les logiciels de textométrie permettent de travailler en stylistique dans deux directions :

- Analyse interne des textes : concordances, index, segments répétés, repérage des coocurrences, etc. Il s'agit dans ce cas de retrouver un certain nombre de figures à l'aide de marqueurs lexicaux. C'est directement envisageable pour certaines (la comparaison, par exemple), éventuellement sur la base d'un corpus lemmatisé et catégorisé, mais inenvisageable pour d'autres (la métaphore, par exemple).
- Analyse comparative des textes : la comparaison avec le lexique d'un corpus de référence permet de juger des options d'une écriture en termes d'écarts. La difficulté méthodologique de ce procédé tient au choix du corpus de référence, dont les particularités génériques, stylistiques ou thématiques pèsent statistiquement très lourd sur les résultats.

2) Est-il possible d'automatiser le repérage des indices ? des marqueurs ? des stylèmes ? Quels seraient, le cas échéant, les indicateurs de définition de seuils, si cette notion est nécessaire ?

Un gros travail de recensement des indices, marqueurs et stylèmes reste à accomplir. Il existe certes des dictionnaires et des manuels qui proposent des listes mais aucun de ces catalogues ne se réfère à l'éventualité d'un traitement automatique. A l'opposé, on trouve un grand nombre d'études qui se proposent l'étude d'un phénomème stylistique par des moyens informatiques et statistiques mais ces études restent ponctuelles et ne visent qu'à implémenter le repérage d'un seul trait stylistique (même si, en particulier dans le cadre des études de paternité, plusieurs marqueurs sont analysés).

Il conviendrait donc d'aligner ces deux types de sources, quitte à constater que certains stylèmes n'ont donné lieu à aucune tentative d'automatisation.

L'objectif serait d'établir les définitions des stylèmes directement exploitables pour l'analyse informatique, et donc traduisibles en algorithmes. Il ne faut pas se cacher la difficulté de cette entreprise, qui nécessitera des réévaluations critiques de la terminologie et de la typologie en usage.

La notion de seuil intervient dans le cadre de ces algorithmes. Si l'on prend par exemple le cas de l'allitération, il faudra fixer un degré de concentration des phonèmes consonantiques à partir duquel on estime que le phénomène devient perceptible.

3) Qu'est-ce qui serait automatisable, qu'est-ce qu'il serait possible d'annoter pour guider l'analyse d'un corpus ?

Les logiciels ou fonctionnalités de logiciels à mettre en œuvre devront permettre une utilisation ciblée (c'est-à-dire centrée autour d'un type de phénomène déjà identifié) et une approche heuristique guidée, par laquelle l'automate suggèrera des pistes de recherche, par une analyse du corpus et une détection systématique de tous les phénomènes remarquables (par rapport à des indicateurs préalablement étalonnés). Il faudrait dans cette perspective établir une liste de tests permettant de caractériser un corpus du point de vue stylistique.

4) Quelle serait la pertinence de disposer d'un référentiel (avec ambition ou non d'exhaustivité) d'analyse textuelle en vue de son traitement par les logiciels de textométrie ? Sa structuration hiérarchique est-elle facilitante pour la modélisation ? Jusqu'à quel niveau de hiérarchie peut-on espérer aboutir ?

Un référentiel de ce type faciliterait le travail des développeurs, à condition de pouvoir le formuler en terme opératoires (voir plus haut). Un dialogue entre stylisticiens et informaticiens devra aboutir à une formalisation satisfaisante pour les deux approches. La structuration hiérarchique d'un tel référentiel permettrait, outre une présentation plus claire des résultats, de construire l'assistant décrit ci-dessus, en accumulant des indices par classe. Par exemple, une synthèse automatisée sur le système des temps verbaux dans le corpus devrait s'appuyer sur des

tests de niveau inférieur, sur divers marqueurs temporels, et accumuler à chaque niveau un certain nombre d'indices dont la prise en compte globale permettra de construire une appréciation globale et contrastive.

5) En quoi la constitution des corpus est-elle fondamentale dans la démarche ? Est-il possible pour les besoins de l'analyse des marqueurs stylistiques de contraster au sein même d'un texte ? ou le contraste n'est-il pensable et "rentable" que sur des textes externes (intragénériques ou intergénériques) au texte support à l'intérieur d'un grand corpus ?

Les deux démarches ont leur intérêt, dans la mesure où il est en réalité difficile de les distinguer. Quand on examine un corpus par comparaison avec un « corpus de référence », c'est en réalité ce nouveau corpus, plus volumineux, que l'on étudie. Il faut surtout se défaire de l'idée qu'une collection de textes, quelle que soit sa taille, pourrait constituer une référence indiscutable en matière de vocabulaire ou de style. Chaque corpus a ses spécificités, ses particularités, et toute comparaison sera tributaire de ces tropismes. Comparer une œuvre de Zola avec une collection de romans contemporains, c'est relever tout autant les traits originaux de Zola que les caractéristiques d'une collection dont il faudra interroger les conditions de numérisation et de sélection. Il est important, en particulier, de ne juger de l'effet des marqueurs stylistiques que par rapport à un horizon d'attente que seule l'histoire littéraire peut préciser.

6) Comment faudrait-il situer l'étape "générique" par rapport à l'étape "stylistique" dans la construction et l'individuation de la textualité ?

J'ai proposé de hiérarchiser de la manière suivante les facteurs de variation entre les textes d'un corpus littéraire :

- La langue (y compris les variations diachroniques à l'intérieur d'une même langue)
- Le genre (à hiérarchiser à son tout entre grandes catégories énonciatives : dialogiques, narratifs, etc. et genres de niveaux divers : théâtre, comédie, etc.)
- Le mouvement (ou partis pris esthétiques, ce qui distinguera, par exemple, romans naturalistes ou romantiques)
- La thématique (ce facteur joue plutôt sur le vocabulaire à sémantisme plein que sur les marqueurs stylistiques)
- L'auteur

On pourra bien entendu discuter de l'ordre proposé mais je veux surtout insister sur le fait que l'on ne peut détecter de marque stylistique propre à un auteur (ce que l'on appelle, depuis Buffon, le « style ») qu'après avoir pris en compte tous les autres facteurs. C'est là une difficulté récurrente dans les études d'attribution à un auteur, qui pèchent parfois en attribuant à des options d'écriture personnelles ce qui en réalité relève de contraintes génériques, thématiques, voire linguistiques.

7) Quelle serait la place d'une analyse stylistique automatisée au sein de l'herméneutique linguistique ?

Il me semble important d'insister sur l'impossibilité d'une automatisation complète d'une analyse stylistique, dans la mesure où celle-ci implique une subjectivité du chercheur que la machine ne peut remplacer. Il vaut mieux parler d'analyse stylistique outillée, ou assistée par ordinateur. Un programme peut certes attirer l'attention du chercheur sur certaines caractéristiques du corpus étudié, ou permettre une approche statistique d'un phénomène particulier, mais ces indicateurs ne seront à une véritable analyse interprétative que ce que sont les résultats d'une analyse biologique au diagnostic d'un médecin.

En revanche, la notion d'« outillage », devenue courante aujourd'hui dans de nombreux domaines de la recherche, permet de penser correctement le recours à l'informatique et à la statistique comme une assistance qui, en retour, ne peut laisser intacte la discipline et ses objets conceptuels, en l'occurrence ceux de la stylistique.

Tous des copiateurs

Etienne Brunet

Professeur honoraire de l'Université de Nice

Résumé :

Les littéraires ne font guère confiance à la technologie pour les guider dans le choix ou l'appréciation des œuvres mais ils consentiraient à lui confier un rôle d'expertise analogue au rôle que jouent les empreintes digitales, l'ADN et le carbone 14, pour la signalétique des individus ou l'âge des matériaux. Détecter les mensonges, les supercheries et les plagiats, reconnaître les signatures, fixer les dates, l'ordinateur le peut-il ? La démarche la plus simple est de demander à la machine de comparer l'original et la copie supposée et de relever les passages communs. Mais cette chasse est souvent déroutée par les modifications de détail : les changement de nombre, de genre, de temps, de place, les additions, les suppressions et le recours aux synonymes brouillent la piste. La technique des segments répétés est alors plus efficace si elle porte sur les lemmes et non seulement les graphies, et si les cooccurrences sont prises en compte et non seulement les séquences. Mais même en l'absence de preuves tangibles - sur lesquelles s'appuie la jurisprudence - il y a moyen de suspecter et de dénoncer le plagiat, quand l'analyse globale des éléments linguistiques s'appuie sur une statistique fine, qu'il s'agisse de vocabulaire, de syntaxe ou de rythme du discours.

Résumé court pour indexation :

A partir de deux exemples, on examine les moyens documentaires qui permettent de prouver le plagiat (affaire Vautrin-Griolet) et les analyses statistiques qui étoffent les soupçons (affaire Béyala-Buten).

Mots-clés :

Plagiat littéraire, détection documentaire, présomption statistique, logiciel HYPERBASE

Abstract :

Specialists in literature hardly ever call upon technology to guide them in their choices and appreciations of literary works, but they would [no doubt?] agree to allow it to perform forensic analyses comparable to finger print or DNA testing, and carbon-14 dating, when it comes to identifying individuals or dating materials. Uncovering lies, hoaxes and plagiarisms, recognizing signatures, dating particular elements – is a computer able to perform all of these tasks? The easiest way is to ask the machine to compare the original and the alleged copy and to identify passages common to both. But surveys of this type are often hindered by minor modifications such as changes in number, gender, time, and place; additions, deletions and synonyms throw us off the scent. The search-for-repeated-segments technique becomes more efficient if it tackles lemmas, not simply written forms, and if co-occurrences are taken into account, not simply sequences of words. But even in the absence of any tangible evidence – on which French

jurisprudence relies – there exist ways of detecting and denouncing plagiarism, when the overall analysis of linguistic elements is based on a fine statistical approach, whether it concerns vocabulary, syntax or the rhythm of sentences

Key words: literary plagiarism, documentary detection, presumption based on statistical data, « HYPERBASE » software program

La nature copie sans vergogne. Les races, les espèces, les saisons, les jours ne se maintiennent que par l'héritage, par le retour cyclique, par la transmission du même au même. La culture aussi se reproduit par duplication. Un savoir qui n'est pas copié est perdu. Les grandes avancées ne sont pas tant dans l'invention de savoirs nouveaux que dans la puissance de reproduction du savoir acquis, dans l'extension et la rapidité de la communication, que ce soit la presse de Gutenberg, la photocopieuse, l'ordinateur ou Internet.

Or jamais jusqu'ici la copie n'a été aussi facile ni aussi répandue. Le réseau mondial distribue chaque seconde un torrent de copies : des textes, des images, des paroles, des modes, des rumeurs. Et l'individu armé du couper-coller fait son choix devant l'étal. Il lui arrive même de solliciter son disque dur et de se copier lui-même.

- I –

Mais si l'ordinateur est le fournisseur de propos ou d'idées à répéter, ne peut-il pas aussi fournir l'antidote contre les répétitions, soit pour mieux les camoufler, soit pour mieux les détecter ?

Les littéraires ne font guère confiance à la technologie pour les guider dans le choix ou l'appréciation des œuvres mais ils consentiraient à lui confier un rôle d'expertise analogue au rôle que jouent les empreintes digitales, l'ADN et le carbone 14, pour la signalétique des individus ou l'âge des matériaux. Détecter les mensonges, les supercheries et les plagiats, reconnaître les signatures, fixer les dates, l'ordinateur le peut-il ?

Malheureusement bien des facteurs entrent en jeu qui brouillent la signature : le genre, le sujet, la date, l'environnement et tout cela produit des influences entremêlées qu'il est difficile d'isoler.

D'autre part, les indices qu'on peut tirer d'un texte sont multiples, soit qu'ils appartiennent au vocabulaire, à la syntaxe, à la sémantique, à la métrique, et la convergence des approches n'est pas garantie.

De plus, les ressemblances entre deux textes ne suffisent pas à établir le plagiat : certaines sont involontaires, d'autres sont avouées (la citation, la traduction, le pastiche), certaines portent sur le fond, d'autres sur la mise en forme.

Enfin il s'agit d'un combat inégal, comme celui de l'épée et du bouclier. Le plagiaire peut s'attaquer à n'importe lequel des ouvrages antérieurs (ceux qui ont été publiés mais aussi les manuscrits restés chez les éditeurs). Il a l'initiative, l'imprévisibilité, et toujours un coup d'avance. Le plagié doit soupçonner tous les livres postérieurs (et parfois même les publications

concomitantes, en cas de fuite ou de brevet éventé). Cette dissymétrie est semblable à l'affrontement terrorisme/antiterrorisme ou fraude/amende.

Si la contrefaçon peut être prouvée dans le domaine industriel où toute une législation internationale protège la propriété et les brevets, dans le domaine de la création artistique ou de la culture la difficulté est grande pour authentifier une signature, une voix, un tableau, une sculpture, une ruine, un morceau de musique, un texte, une édition, une école, un genre, une époque.

On illustrera ces difficultés en s'appuyant sur deux affaires récentes : celle du Goncourt obtenu en 1989 par Vautrin pour *Un grand pas vers le Bon Dieu* et inspiré d'une étude universitaire, et celle qui a opposé Calixthe Beyala à Buten pour le roman *Le petit prince de Belleville* publié en 1992 et condamné en 1996 par le TGI de Paris. Pour ces deux cas exemplaires, on aura recours aux ressources de l'informatique en empruntant deux méthodes d'investigation, l'une purement documentaire, et l'autre statistique.

- II –

« Un Vautrin peut en cacher un autre »

Pour le Goncourt *Un grand pas vers le Bon Dieu,* inspiré de deux ouvrages de Patrick Griolet *Cadjins et créoles en Louisiane* et *Mots de Louisiane*, on trouvera dans *le Monde* du 2 décembre 1989 un article exposant le point litigieux, et une présentation plus détaillée dans le journal *Nice-Matin* du 1er décembre. La question devait être soulevée à nouveau dans le *Canard enchaîné* du 20 décembre sous le titre *Un faux pas vers le Bon Dieu* et le lendemain dans le *Quotidien de Paris*. L'affaire est tranchée par les tribunaux, Kiejman, ministre de la justice, étant l'avocat de Vautrin. Vautrin qui a reconnu sa dette n'est pas condamné[20].

L'ouvrage de l'universitaire est en effet assimilé à un dictionnaire où l'information délivrée est utilisable librement. La majorité des termes auxquels Griolet consacre un commentaire d'ordre lexical, orthographique, syntaxique, sémantique ou sociologique se retrouvent dans Vautrin avec les particularités signalées par le chercheur. Mais habituellement un auteur garde son style propre, même lorsqu'il évoque des réalités et des mots étrangers ou techniques.

Vautrin, lui, n'emprunte pas seulement les objets d'une contrée et d'un milieu social et *l'étiquette lexicale* qui est collée dessus, c'est la **langue** même de cette contrée qu'il veut saisir et restituer, et, au lieu de porter sur de simples curiosités accessoires, l'emprunt prend un caractère massif qui englobe tout à la fois : le lexique, la morphologie, la syntaxe et le contenu sémantique. Des pages entières des *Mots de Louisiane* sont l'objet d'une exploitation systématique où non seulement les mots mais aussi les expressions du modèle se retrouvent éparpillés çà et là dans le roman. Ainsi en est-il des expressions nominales qui décrivent la violence:

donner une bûchée, une claquée, un coup de tape, une dandine, un patcharac ici un patcharac là, une poque, une ramasse, une rincée, une rinçure, une roulée, une rousselée, une tatouille, une torgnole, une tripotée, un veux-tu-couri, une vire-tape d'aller et de retour.

Les expressions verbales sont tout aussi pittoresques :

[20] Voir E. Brunet, « Que l'emprunt vaut rin », *French Review,* déc 1990, p. 273-288.

assauter, bûchailler, carder la peau, chapigner, écorcher, esquinter, fendre le biscuit, graisser la coloquinte ou la cagouette, se hartchiner, prendre à la palette, riocher, sacrer une volée, etc..

Or tout cela figure p. 157 de *Mots de Louisiane*.

Plus gravement, des extraits plus larges sont empruntés à la source qui vont de la phrase au paragraphe entier.

Mots de Louisiane		***Un grand pas vers le Bon Dieu***	
Plus la peau était blanche et tendre, Plus elle était aimable	p.22	Plus la peau était blanche et tendre Plus elle était aimable	p.82
Tu peux m'arrêter de t'tiendre les mains	p.24	Elle pouvait pas s'arrêter de tiendre ses mains	p.40
Faut pas bêtiser avec les Indiens	p.31	Faut pas bêtiser avec les Attakapas	p.237
J'sus qu'un bêtiseur	p.31	J'sus qu'un bêtiseur	p.50
J'tiens mon boute	p.36	J'tiens mon boute	p.274
Elle s'met une jolie camisole frayante	p.41	Je t'offrirai une jolie camisole frayante	p.14
Un carencro noir va pas marier une oie farouche	p.43	J'marierai jamais un carencro noir avec une oie farouche	p.144
D'autres sont tombés en chamaille	p.45	d'autres seraient tombés en chamaille	p.139
J'ai hâlé mes chines loin de ces deux fous-fous-là	p.49	il avait hâlé ses chines loin de ces batailleurs	p.61

Figure 1 : Phrases empruntées.

Cadjins et créoles

-Ti connais combien d'Cadjins ça prend pour battre un Cou Rouge?
...Le gros Texien a dit i(l) connaissait pas.
-ça prend treize Cadjins pour batt(re) un Cou Rouge!
Et ça a beaucoup flattéŽl'Texien. Il était fier et satisfait. Et là, là (pe)tit Cadjin a dit:
-ça prend douze Cadjins pour l'escouer en bas d'l'arbre et un pour y casser l'fond ˆ coups de pied après qu'i(l) tombe.

Un grand pas vers le Bon Dieu

-Combien de Cadjins comme moi crois-tu ça prendrait pour battre avec un grand "Américain" comme toi ?
Voicy Smith savait bien sûr pas qui répondre. Le vieux bougu' s'a redressé sur ses ergots. Il sourissait tout drôle. Il a dit:
-ça prendrait treize Cajuns pour battre un Cou-Rouge comme toi! Douze pour le secouer de l'arbre et un pour lui casser le fond du cul à coups de pied après qu'il tombe par terre !

Figure 2 : Longs passages recopiés presque textuellement

-III-

Deuxième cas: l'affaire Beyala-Buten[21]

Dans l'expertise qui précède, l'ordinateur n'a été utilisé que pour ses capacités documentaires. Ayant absorbé le roman de Vautrin, il servait à repérer et à restituer dans le roman tout mot, expression ou passage suspect, formellement désigné et proposé à partir du texte plagié. Alimenté par des soupçons humains et subjectifs (la plupart émanant de l'auteur plagié), il délivrait des preuves matérielles et objectives.

Mais peut-on lui demander plus ? Au lieu de surprendre le voleur sur le fait, la main dans le sac, on aimerait que par des indices indirects et des présomptions statistiques il puisse désigner comme contrefaçon la ressemblance de deux textes. Certes une preuve vaut mieux que deux présomptions mais les faussaires avertis cherchent à éviter les flagrants délits et à multiplier les camouflages pour rendre les preuves directes plus difficiles à administrer. En revanche, comme ils ignorent généralement les techniques lexicométriques, il leur est difficile de se prémunir de ce côté-là.

La méthode la plus courante pour apprécier les faits de similitude est le calcul de la distance intertextuelle. Les deux textes à comparer subissent un examen général de leur vocabulaire et de leur style, et le résultat peut se lire dans une sorte de grille où la distance croît de 0 à 1. Ce test peut porter sur les graphies, sur les lemmes, sur les codes grammaticaux, sur les structures syntaxiques et quand le texte s'y prête sur la prosodie. On reste néanmoins dubitatif quand on n'a que deux textes en présence : celui du plagié et celui du plagiaire. Comme il n'y a pas d'échelle absolue, l'indice laisse le jugement en suspens. Il n'en va pas de même lorsqu'on réunit dans un même corpus d'autres textes suffisamment proches, comme on fait pour les échantillons témoins en pharmacologie. La neutralité de l'environnement fait alors ressortir la proximité suspecte des textes où le plagiat est mis en cause.

Ainsi avons-nous fait dans l'affaire Beyala. Il se trouve que nous disposions d'un vaste corpus constitué autour des écrivains africains d'expression française. Or Calixthe Beyala était représentée dans cette base de 26 romans, et trois de ses titres avaient été retenus : *Le soleil m'a brûlée*, *Maman a un amant* et l'objet du procès : *Le petit prince de Belleville.* Il a suffi d'ajouter à ce corpus la version française du roman de Buten : *Quand j'avais cinq ans, je m'ai tué*, (Seuil, 1981). Or dans tous les tests qui vont suivre on constate la fraternité étroite des deux romans que Beyala a écrits à un an de distance et à peu près sur le même thème*, Maman* et *Belleville,* mais aussi la consanguinité surprenante qui lie Buten à Beyala. Le soupçon de plagiat naît de la ressemblance thématique qui s'exerce sur les graphies et sur les lemmes (graphiques 1 et 2) mais aussi de la similitude dans l'usage de la grammaire et de la syntaxe (graphiques 3 et 4).

[21] Cette affaire est évoquée, comme la précédente et beaucoup d'autres, sur le site de Hélène Maurel-Indart : www.leplagiat.net

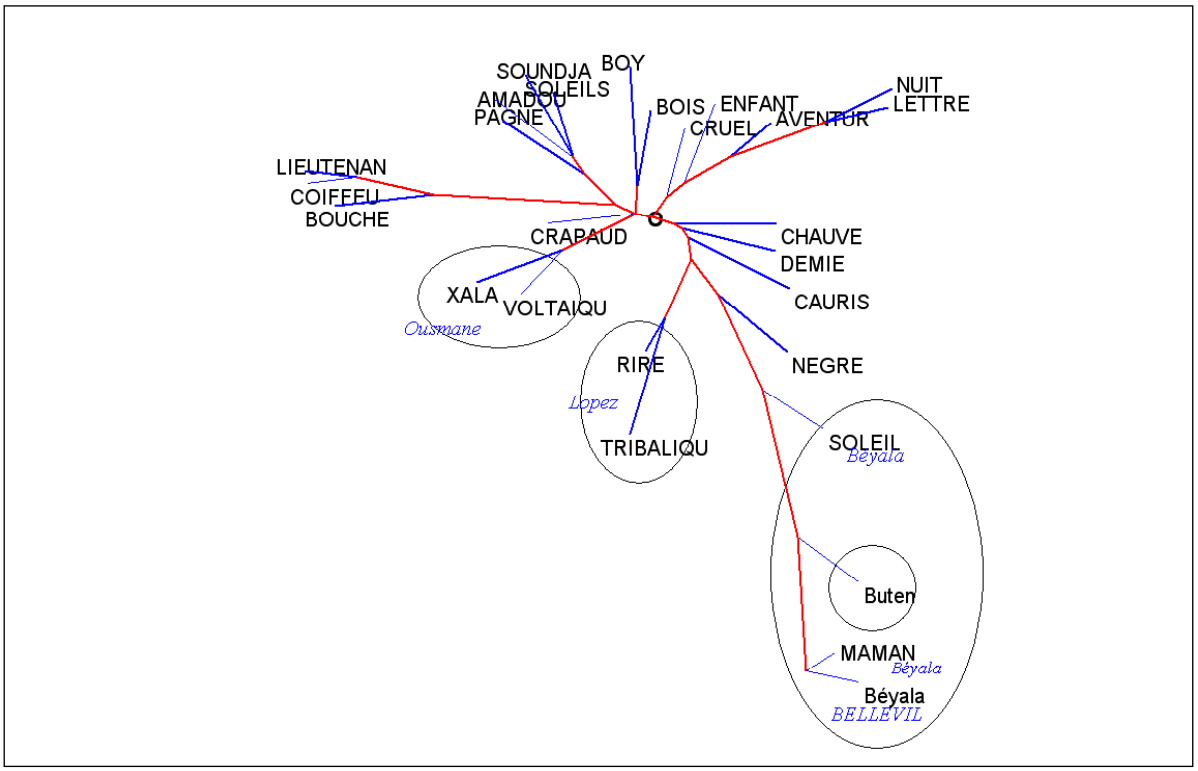

Figure 3 : Distance de Jaccard, sur les graphies : Analyse arborée de 26 romans africains dont 3 de Béyala, le roman de Buten a été rajouté

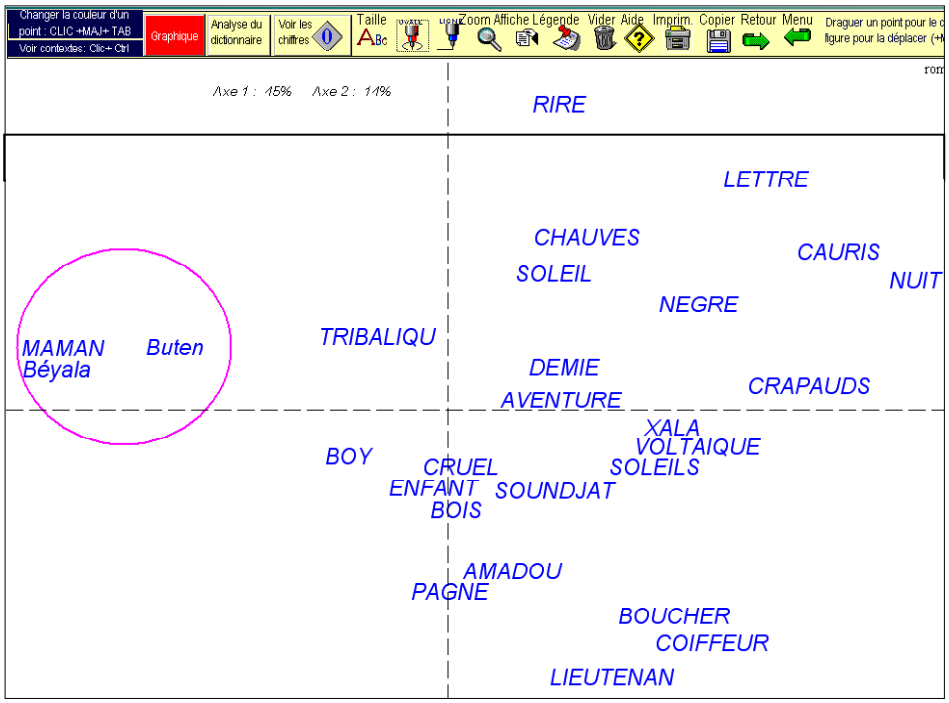

Figure 4 : Analyse factorielle sur lemmes. Distance de Muller

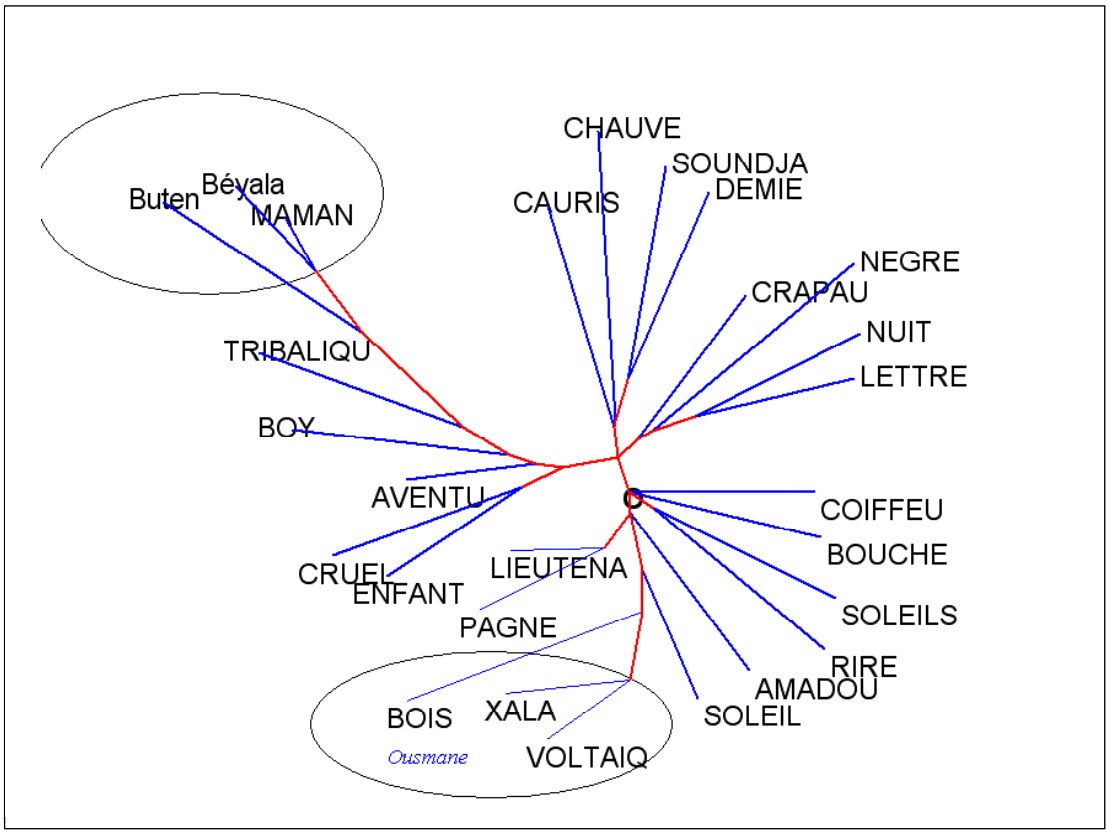

Figure 5 : Analyse de la structure syntaxique

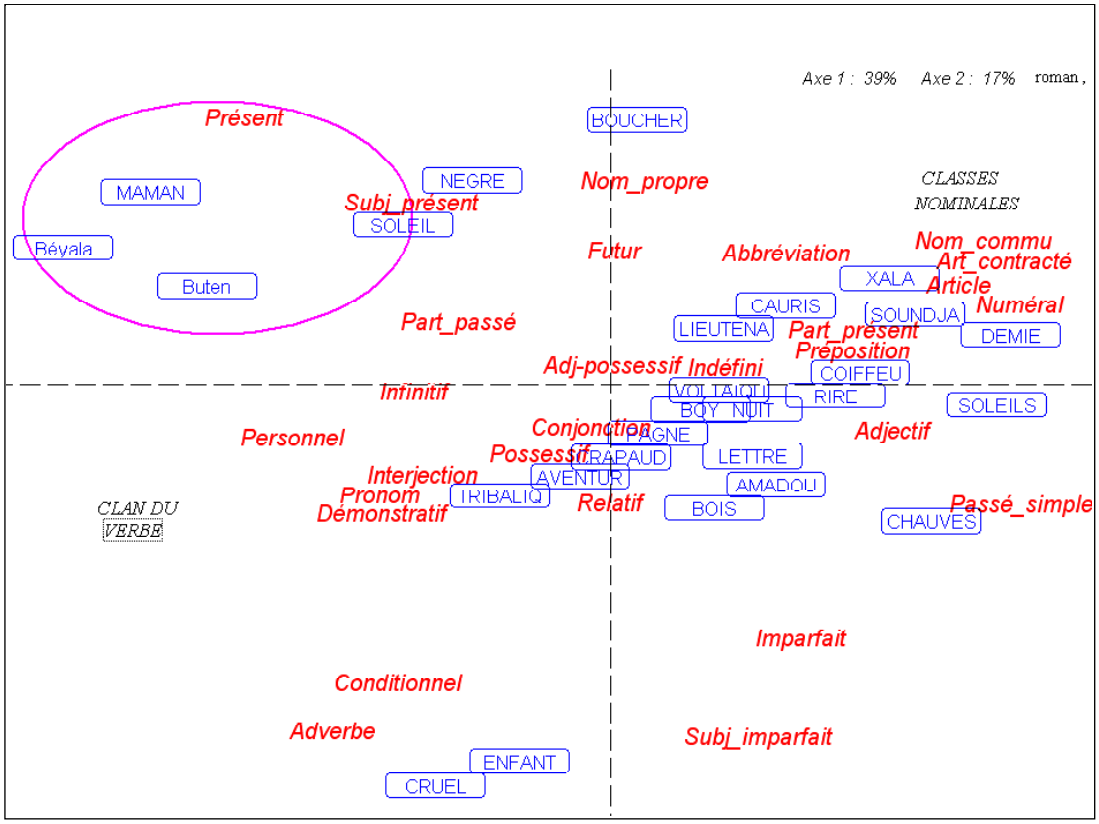

Figure 6 : Analyse factorielle des catégories grammaticales

Les présomptions sont d'autant plus fortes que des outils variés (analyse arborée et analyse factorielle, distance de Jaccard, de Labbé et de Miller) conduisent à des résultats convergents. Partout le nom de Buten voisine avec celui de Beyala.

Mais le logiciel *Hyperbase* (comme aussi *Lexico*) dispose d'une fonction plus précisément adaptée à la recherche des plagiats : le traitement des segments répétés. Si des segments suffisamment longs et nombreux apparaissent identiques dans deux textes indépendants, cette indépendance peut être mise en doute. Pour ce faire, une fois que la liste des segments répétés est établie dans le corpus, une fonction supplémentaire permet de mettre en relief les segments communs à deux textes. Or on constate qu'ils sont plus nombreux à l'intérieur de la triade *Maman*, *Belleville* et Buten.

Cependant des différences de détail diminuent l'efficacité du test. Considérons en effet un extrait de Beyala confronté à un passage copié chez Romain Gary[22].

BEYALA, Calixthe, le Petit Prince de Belleville	GARY, Romain, la Vie devant soi, Mercure de France
J'accompagne M'am faire des courses dans les grands magasins à l'Opéra. Il y a un cirque en vitrine. Les parents viennent avec leurs mômes gratuitement. La vitrine est tout entourée d'étoiles plus grosses que nature. Elles s'allument, elles s'éteignent en un clin d'oeil. Au milieu du cirque, il y a des cosmonautes. Ils vont jusqu'au ciel, ils reviennent sur terre en faisant des saluts aux passants.	J'avais une course à faire dans un grand magasin à l'Opéra où il y avait un cirque en vitrine pour que les parents viennent avec leurs mômes sans aucune obligation de leur part. (…) La vitrine était entourée d'étoiles plus grandes que nature qui s'allumaient et s'éteignaient comme on cligne de l'oeil. Au milieu, il y avait le cirque avec les clowns et les cosmonautes qui allaient à la lune et revenaient en faisant des signes aux passants.

Figure 7 : Exemple de plagiat (sur le site de H. Maurel-Indart)

Les camouflages, transparents à l'œil humain, restent opaques à la machine, qu'il s'agisse de :

- substituer le présent au passé
- substituer le pluriel au singulier (*course(s), magasin(s)*)
- changer la construction (*comme on cligne de l'œil -> en un clin d'œil*)
- changer l'emplacement

[22] Cet exemple est emprunté à Hélène Maurel-Indart. Voir le site cité plus haut.

- utiliser les synonymes :

sans obligation > *gratuitement,*

grandes > *grosses, lune* > *ciel, signes* > *saluts*

En conséquence l'ordinateur, dérouté, ne trouve que 4 segments répétés :

- *les parents viennent avec leurs mômes* (longueur 6)

- *un cirque en vitrine* (longueur 4)

- *entourée d'étoiles plus* (longueur 4)

- *en faisant des* (longueur 3)

Il convient donc d'appliquer le calcul des segments répétés aux lemmes plutôt qu'aux graphies. Au moins les changements de temps, de nombre ou de genre seraient neutralisés. Si l'on procède ainsi, la pêche des segments communs est plus copieuse et plus probante. Étant cliquables, ils servent alors d'hameçon pour appâter d'autres pièces plus intéressantes comme dans l'exemple ci-dessous.

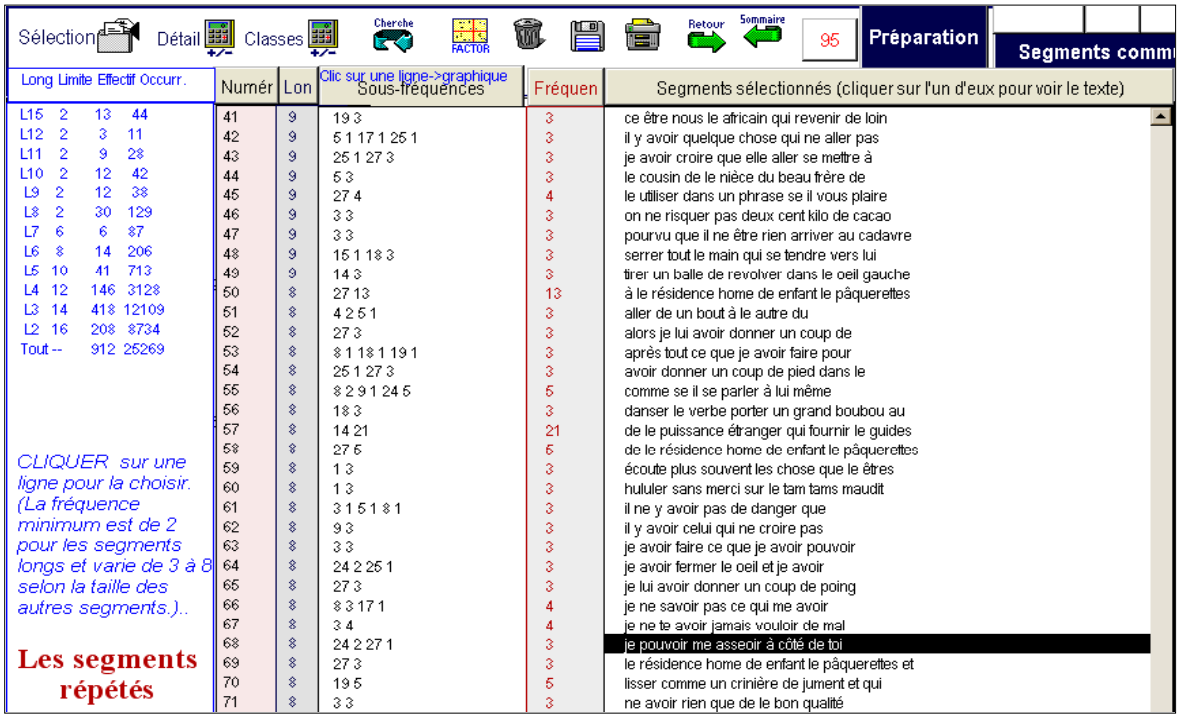

Figure 8 : Les segments-lemmes répétés.

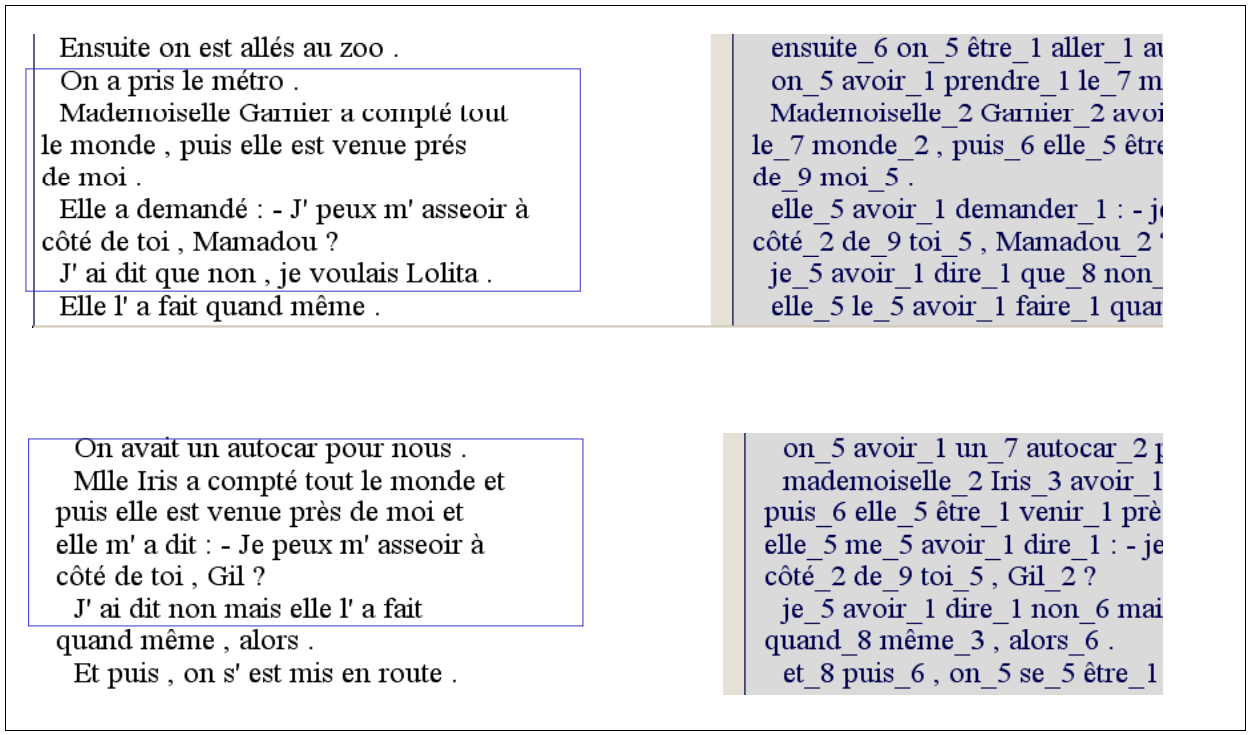

Figure 9 : Quand l'ordinateur renifle le plagiat

- IV –

Un recours paradoxal: Internet

Internet est né de la liberté, mais il pourrait devenir l'instrument du contrôle, de la détection, de la délation et de la répression. Déjà *Google* permet de vérifier ou d'infirmer partiellement les soupçons de plagiat, du moins si la question est bien posée (à partir d'une expression rare et improbable). Ce contrôle est appelé à devenir plus fiable quand les sources écrites auront été copiées à grande échelle et que les moteurs de recherche auront accès à toutes les bases de données et à tous les formats de fichiers. Déjà des logiciels spécialisés sont proposés sur le marché pour la détection et la veille sécuritaire : Plagium, Compilatio.net, Noplagiat.com.

Internet peut aussi jouer le rôle d'un garde-fou préventif. Quand des milliards de textes sont en attente sur le réseau et que des radars électroniques réagissent à la moindre violation, il devient périlleux d'avancer la moindre phrase. La combinatoire des mots n'est pas infinie et l'on risque fort de rencontrer un précédent « quand on naît trop tard dans un monde trop vieux ». Si « le premier qui a comparé une femme à une rose est un génie, et le second un imbécile », il est prudent de recourir à Internet pour éviter d'être le second et prévenir un plagiat éventuel, même involontaire, notamment lorsqu'on choisit un titre. N'essayez pas « Rime et raison » : ce titre a servi des milliers de fois. Même « qui trop embrase mal éteint » a été plusieurs fois défloré. La virginité est devenue très rare et il faut chercher longtemps pour trouver des calembours et des titres originaux[23].

***

[23] Nous sommes médiocrement fier d'en avoir trouvé quelques-uns que le réseau ignorait encore : « *Qui lemmatise dilemme attise* », « *Muller, le lexicomaître* » ou « *Que l'emprunt vaut rin* ».

Entre frilosité et laxisme, toute réflexion sur le plagiat équivaut à penser la nourriture, la culture et la littérature, comme l'admet avec humour Jean Giraudoux : « Le plagiat est la base de toutes les littératures, excepté de la première, qui d'ailleurs est inconnue » (Siegfried). Même s'il n'hésite pas à s'approprier les mythes grecs ou bibliques, quitte à proposer une trente-huitième version d'*Amphitryon* et une énième figure de Judith ou d'Electre, Giraudoux a l'âme trop haute pour s'abaisser à copier. Ses manuscrits montrent qu'il répugne même à recopier ce qu'il a écrit lui-même. Ayant horreur des ratures, il reprend une feuille blanche et improvise une version radicalement différente. On en a compté jusqu'à six pour certaines scènes de sa pièce *Sodome et Gomorrhe*. La *Sodome* de Proust au contraire n'est pas ennemie des ratures, comme l'ensemble de la *Recherche*. Et Proust est plus sensible à la démangeaison du plagiat, même s'il prête ce défaut à Françoise plutôt qu'au narrateur : « Françoise devinait mon bonheur et respectait mon travail. Elle se fâchait seulement que je contasse d'avance mes articles à Bloch, craignant qu'il me devançât et disant : "Tous ces gens-là, vous n'avez pas assez de méfiance, c'est des copiateurs". Et Bloch se donnait en effet un alibi rétrospectif en me disant chaque fois que je lui avais esquissé quelque chose qu'il trouvait bien : "Tiens, c'est curieux, j'ai fait quelque chose de presque pareil, il faudra que je te lise cela." (Il n'aurait pas pu me le lire encore, mais allait l'écrire le soir même.) » (*Le Temps retrouvé, 3e partie).* Chez Proust le plagiat prend la forme honorable du pastiche[24], avoué dans *L'affaire Lemoine*, caché dans la *Recherche* même (notamment dans la caricature des Goncourt aux pages 709-717 du *Temps Retrouvê*). Et il s'en explique dans une lettre à R. Fernandez (1919) : « Le tout était surtout pour moi affaire d'hygiène ; il faut se purger du vice naturel d'idolâtrie et d'imitation. Et au lieu de faire sournoisement du Michelet ou du Goncourt en signant (ici les noms de tels ou tels de nos contemporains les plus aimables), d'en faire ouvertement sous forme de pastiches, pour redescendre à ne plus être que Marcel Proust quand j'écris mes romans. »

[24] Giraudoux ne manquait pas de dons pour le pastiche : on en a un témoignage avec le prix de version grecque, qu'il obtint au Concours général et qui est une brillante imitation de J.Amyot traduisant Plutarque.

Genre, style et attitude à l'égard du langage : tentative de diagnostic automatique sur un corpus politique

Pascal Marchand

Professeur à l'IUT "information & communication" (Toulouse 3) et membre du Laboratoire d'Études et Recherches Appliquées en Sciences Sociales (LERASS)

http://pascal-marchand.fr

Résumé :

Il s'agit de montrer que les outils d'*analyse du discours assistée par ordinateur* (statistique textuelle, analyses morphosyntaxiques) peuvent aider au diagnostic de genre, de style et d'attitude langagière. Le corpus des *déclarations de politique générale* (DPG) des Premiers ministres français de la V^e^ République est analysé dans cet objectif.

Mots-clés : Genre, Style, Attitudes à l'égard du langage, Déclaration de politique générale, Analyse du discours assistée par ordinateur.

Summary:

We want to show that computer-assisted discourse analysis (statistical text analysis, morphosyntaxic analysis) can contribute to the diagnosis of genre, style and attitude towards language. To this end, we analyze the policy statements of French prime ministers throughout the Fifth Republic.

Keywords: Genre, Style, Attitudes towards language, Policy Statement, Computer-assisted discourse analysis**.**

D. Malrieu & F. Rastier (2001, repris et complété par F. Calas, 2009), définissent à propos des textes littéraires, des niveaux de classification en : discours, champs génériques, genres, sous-genres, auteurs et textes. Chaque DPG peut également être définie comme un texte prononcé par un auteur (de 1959 à 2007), dans le sous-genre défini par l'article 49 de la Constitution de la V^e^ République (modifiée en juillet 2008), inclus dans un genre discursif que l'on nommerait « institutionnel », distinct d'autres champs génériques tels que l'interview ou le discours de meeting, mais relevant du discours politique.

Le corpus ainsi circonscrit comporte 231 232 occurrences, représentant 12 742 formes lexicales différentes, dont 5 314 hapax.

Genre et lexicométrie

Le genre est généralement défini comme la convergence entre un critère textuel (lexique, syntaxe…) et un critère contextuel (prescriptions, attentes…). Ainsi, pour D. Malrieu & F. Rastier

(o.c.), « *Les variations morpho-syntaxiques selon les genres sont notables et l'étude des caractéristiques de genre revient à une étude des normes linguistiques* » (2001: 548-577). De la même façon, pour P. Charaudeau, il s'agit d'une « *articulation entre les contraintes situationnelles, les contraintes de l'organisation discursive et les caractéristiques des formes textuelles* » (2002: 280). Ce que confirme F. Calas, lorsqu'il évoque « *un horizon d'attente, c'est-à-dire un ensemble de « règles » servant à orienter la compréhension et l'interprétation du lecteur* ».

En lexicométrie, ces articulations entre le texte et le contexte prennent la forme d'un tableau lexical, dont les lignes sont les unités lexicales issues du texte (segmentation) et les colonnes sont les marques du contexte, codées par l'analyste (partition). « Travailler la parenté des textes » ou effectuer « un travail comparatif », pour reprendre les termes de Calas, revient à établir des liaisons entre les lignes et colonnes du tableau lexical.

Dans le cas présent, on peut construire un tableau lexical croisant le lexique brut (n_i=2097 pour un seuil de fréquence de 11) et les DPG considérées comme unités contextuelles (n_j=36), pour effectuer des analyses statistiques qui établissent de telles liaisons (spécificités, analyse des correspondances lexicales, distances intertextuelles, classifications automatiques…).

La figure 1 rend compte des distances entre les colonnes de ce tableau lexical. L'interprétation des deux axes composant ce plan factoriel peut être guidée par les notions de style et d'attitude langagière, dans les rapports qu'elles entretiennent avec le genre.

Genre et style (premier axe)

Le premier axe (horizontal) peut être considéré comme révélant la chronologie des DPG sur près d'un demi-siècle, puisqu'il ordonne les discours depuis les années 1960 (à droite du graphique) jusqu'aux années 2000 (à gauche du graphique). Nous posons l'hypothèse que cette évolution chronologique traduit un changement dans le style des DPG.

La question est posée de savoir « … de quels marqueurs le stylisticien a besoin pour conduire son analyse. La question est donc d'une part de savoir s'il existe des logiciels qui offrent une telle perspective, d'autre part, s'il est envisageable de travailler sur un panel aussi large de marqueurs en vue de les orienter vers une analyse stylistique ».

Une piste de réponse est donnée par E. Brunet, indiquant que « c'est bien (…) dans la distribution des catégories grammaticales et principalement des classes nominale et verbale, que se manifeste la distinction des styles, des genres et des écrivains » (1983 : 836).

L'analyse morphosyntaxique, telle que la permet le logiciel *Tropes*, permet d'identifier des catégories logico-syntaxiques (verbes, pronoms, adjectifs, modalisateurs, connecteurs, etc.) pour caractériser le style général du texte, ainsi que le proposait P. Charaudeau (1980) : argumentatif, narratif, descriptif, énonciatif.

Appliquée aux DPG, cette indexation automatique permet d'observer que les discours sont passés d'un style argumentatif, dans les années 1960-1970, à un style narratif, dans les années 1990-2000.

En supposant, provisoirement, l'existence d'un genre homogène de la « *Déclaration de politique générale* », on admettrait alors, non seulement que ce genre supporte parfaitement des évolutions liées aux époques et aux thématiques qu'elles impliquent, mais qu'il supporte également des styles logico-syntaxiques différents.

Genre et Attitude langagière (deuxième axe)

Le deuxième axe (vertical) oppose, à quelques rares exceptions, les discours inauguraux (prononcés après la nomination du Premier ministre et dans un contexte de relative popularité ; en haut du graphique) aux discours suivants (souvent prononcés pour faire face à une motion de censure et dans un contexte de relative impopularité ; en bas du graphique). Nous posons l'hypothèse que cette adaptation contextuelle traduit un changement dans l'attitude langagière des locuteurs.

La notion d'*attitude à l'égard du langage* a été définie par J.-L. Beauvois et R. Ghiglione (1981). Reposant sur la distinction **paradigme** / **syntagme** (Saussure, 1916 ; Jakobson et Halle, 1956), elle suppose que les opérations de **choix** (rapport **paradigmatique**) et de **combinaison** (rapport **syntagmatique**) fonctionnent de façon autonome et peuvent faire l'objet d'une centration spécifique sur l'une ou sur l'autre. On définit alors deux registres d'attitude à l'égard du langage : **paradigmatique** et **syntagmatique**, qui opposent **l'univers du lexique** (définition, désignation, objet, référence, norme lexicale) à **l'univers de la pratique du langage** (marques énonciatives, fonctions centrées sur l'interlocution). La manifestation de ces attitudes s'observe au travers d'indices langagiers, et plus spécifiquement d'une **directionalité textuelle *vs* interlocutoire**.

Dans le cas présent, on observe les contributions au deuxième facteur suivantes :

- pour la zone « popularité » (haut du graphique) : *doit, la, et de, de la, républicain, défense, assurer, de l etat, ministre, une, publique, l etat, organisations, société, la qualité, dialogue, de la décentralisation, missions, essentielles, il s agisse, mais aussi, il s agisse de, etat et, sécurité, activités, l indépendance, culture, coopération, qu il s agisse, développement économique*…
  La forte proportion d'articles et de substantifs permet d'inférer une attitude paradigmatique à l'égard du langage ;

- pour la zone « impopularité » (bas du graphique) : *vous, accord, 1982, je, %, 1983, avons, était, 1981, la négociation, gauche, nous avons, en 1983, reprise, avez, vrai, 1987, temps partiel, vous le, huit, monsieur Mitterrand, négociation, nous, industriel, je vous, partiel*…
  La forte proportion de pronoms personnels et de verbes permet d'inférer une attitude syntagmatique à l'égard du langage.

Ici encore, en supposant l'existence d'un genre homogène de la « *Déclaration de politique générale* », on admettrait, non seulement que ce genre supporte parfaitement des inflexions liées aux contextes sociaux, mais qu'il supporte également des attitudes langagières différentes.

Diagnostics de convergence/divergence générique interne/externe

A chaque étape de l'analyse, les opérations en lexicométrie ou en analyse morphosyntaxique, ont permis de mettre en évidence des liaisons entre des facteurs contextuels (chronologie, popularité politique) et des variables textuelles (lexique, syntaxe, directionalité textuelle). Il apparaît que le genre, hypothétique, de la DPG est suffisamment stable pour supporter des contextes différents, qui imprègnent néanmoins le discours dans son style et sa directionalité.

La question se pose alors du diagnostic générique : comment décider qu'un texte particulier est conforme, ou non, au genre ?

L'histoire récente nous permet d'identifier quatre cas problématiques :

1. Un diagnostic de divergence générique interne : La figure 1 permet d'observer que le texte d'A. Juppé (1995b) doit être placé en élément illustratif pour ne pas saturer les deux premiers facteurs. Ce texte, contextuellement défini comme une DPG (49-2), apparaît donc statistiquement comme extérieur au genre défini par l'ensemble des DPG[25].

2. Un diagnostic de divergence générique externe : lorsque N. Sarkozy prononce un discours de « Feuille de route » devant les parlementaires de la majorité le 20 juin 2007, le milieu politico-médiatique diffuse largement l'idée que le véritable discours de politique générale est celui du président et non celui, à venir, du Premier ministre. L'intégration de ce discours à la base de données textuelles dément cette opinion : il ne s'agit pas d'une DPG et la lexicométrie isole ce texte par rapport aux autres[26].

3. Un diagnostic de convergence générique interne : en revanche, la *déclaration de politique générale* que F. Fillon a prononcée, le mardi 3 juillet 2007, s'intègre parfaitement au genre des DPG (figure 1)[27].

4. Un diagnostic de (quasi) convergence générique externe : contrairement à la « feuille de route », le discours de N. Sarkozy devant le parlement réuni en congrès à Versailles, le 22 juin 2009, n'a pas été vraiment présenté comme une DPG. Son intégration à la base de données textuelles permet néanmoins de penser qu'il s'en rapproche : s'il change la structure du deuxième facteur (contextuel), il ne modifie pas le premier (chronologique).

Un certain nombre d'indices peuvent enfin être utilisés pour tenter d'établir un diagnostic de genre : l'emploi de pronoms personnels, de marqueurs discursifs (la négation, par exemple), de marquages sémantiques[28]… Ces indices ne suffisent généralement pas à diagnostiquer une divergence générique. Un indice retiendra néanmoins notre attention : l'utilisation de termes de plus fortes fréquences (banalité lexicale). La figure 2 permet effectivement d'observer que les divergences génériques s'expriment par un déficit (Juppé 1995b) ou au contraire un suremploi

---

[25] A. Juppé, le 15 novembre 1995 :
- Spécificités positives : milliards de francs, sécurité sociale, dette (remboursement de la dette sociale), branches, caisse(s), 1996, 1997, maladie, CSG, réforme, qualité des soins, déficit prévisionnel…
- Spécificités négatives : la France (1 fois), notre pays (0 fois), l'Europe (0 fois), « nous avons » (1 fois).
- Phrase spécifique : « A travers la maîtrise médicalisée des dépenses et la tenue d'objectifs d'évolution rigoureuse tant en médecine de ville qu'à l'hôpital, les professions de santé participeront l'an prochain à hauteur de cinq milliards de francs à l'amélioration du solde de la branche maladie ».

[26] N. Sarkozy, le 20 juin 2007 :
- Spécificités positives : « on » (91 fois ; « on ne peut pas » : 13 fois ; « on va » : 7 fois), souvent associé à la négation (28 fois dont 13 fois : « on ne peut pas ») ; « je » (150 fois et « j' » : 27 fois ; « je veux prendre mes responsabilités » : 14 fois ; « je vais » : 6 fois); négation (« ce n'est pas » : 23 fois) ; « Premier ministre »; « croissance »; « travail » (50 fois : tag?) en cooccurrence spécifique avec « politique ».
- Spécificités négatives : « nous » (« notre pays »; « notre » : 20 fois ; « nos » : 10 fois), « gouvernement », « loi », « projet », « solidarité », « république ».
- Phrase spécifique : « Politique monétaire, politique budgétaire, je ne les jugerai que par rapport à un seul critère : cela récompense le travail ou cela dévalorise le travail. Tout ce qui récompense le travail sera choisi, tout ce qui dévalorise le travail sera écarté ».

[27] Alors même que cette *déclaration* est remarquable par son emploi d'hapax dans une proportion jamais égalée dans la V[e] République (y compris chez R. Barre et M. Rocard). Cet indice, pour être quantitativement faible, est remarquable : contrairement à ce que l'on aurait pu penser, le fait d'employer beaucoup de termes nouveaux, uniques dans le corpus, n'amène pas à sortir du genre.

[28] Les Premiers ministres aiment parfois marquer leur discours par un « tag » : « nouvelle société » (J. Chaban-Delmas=3 fois), « nouvelle citoyenneté » (P. Mauroy=7 fois), « nouvel espoir » (M. Rocard=10 fois), « nouvel exemple français » (E. Balladur=6 fois), « nouvelle démocratie » (A. Juppé=7 fois), « nouvel humanisme » (J.-P. Raffarin=5 fois), … « travail » (N. Sarkozy=50 fois), « crise » (N. Sarkozy=28 fois).

(Sarkozy, 2007) de ces formes de fortes fréquences. D'autres travaux devront approfondir ce résultat.

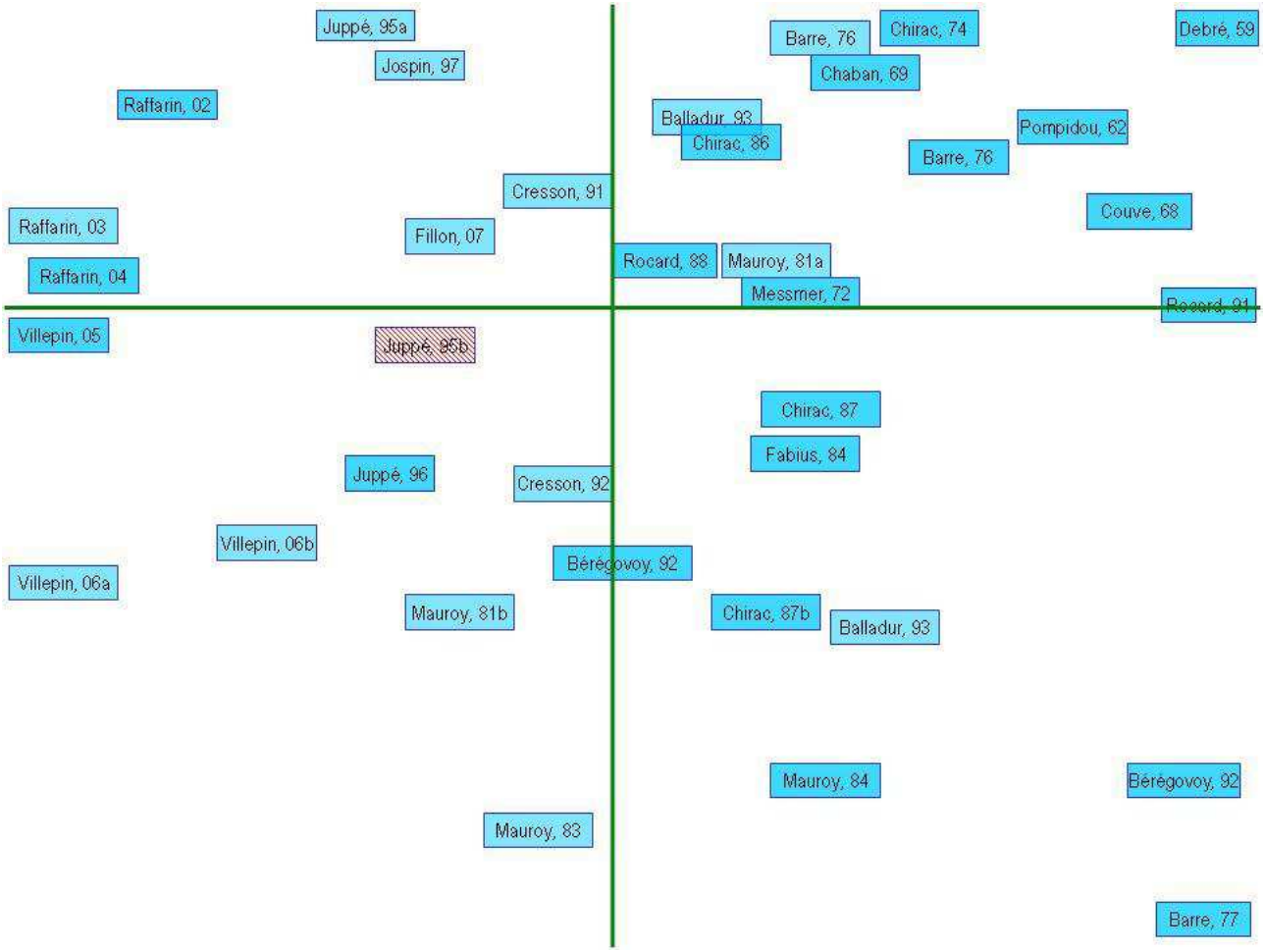

Figure 1 : AFC des Déclarations de politique générale (1959-2007)

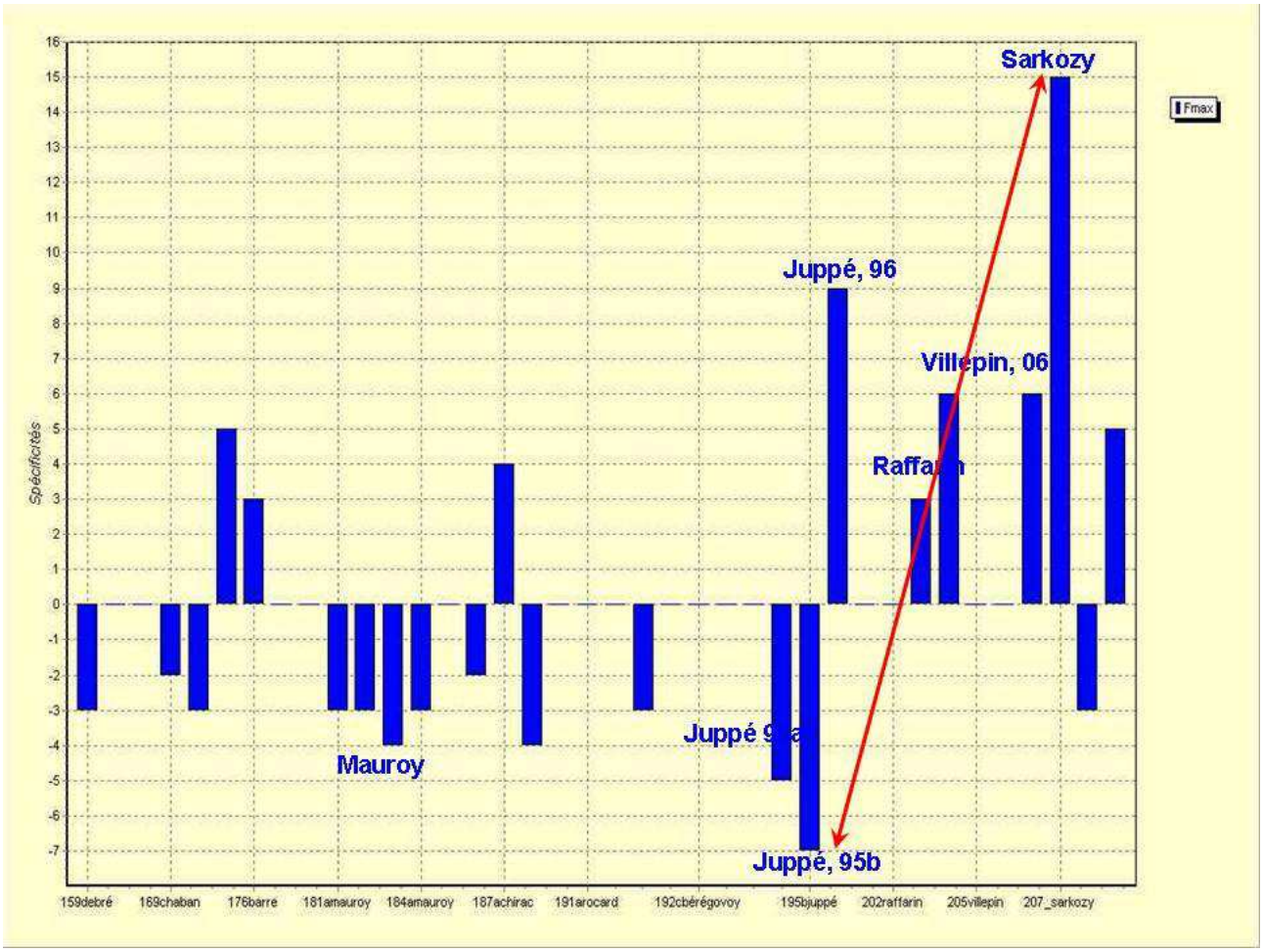

Figure 2 : « Banalité » (Fmax) des Déclarations de politique générale (1959-2007)

Analyse stylistique différentielle à base de marqueurs et textométrie

Bénédicte Pincemin

Chercheur au CNRS & Université de Lyon

Résumé :

Sous la forme de réponse à un questionnaire, ce texte vise à apporter des repères pour une mise en oeuvre de la textométrie en vue d'analyses stylistiques, selon une approche stylistique précisément décrite et illustrée dans un document de travail : "Le style et sa modélisation : document de travail" (Calas & Garric 2009).

Nous considérons d'abord les convergences de fond, les principes en commun, tels que l'observation des contrastes et l'importance de l'interprétation. Puis, concrètement, nous examinons comment une instrumentation actuelle (analyseur TAL dont étiqueteur morpho-syntaxique, moteur de recherche évolué genre CQP, calculs textométriques) pourrait être mobilisée pour alimenter la grille définie dans (Calas & Garric 2009). Complémentairement, nous pointons les parties de la grille dont la formalisation et l'automatisation sont plus difficiles. Certains principes de modélisation peuvent également mériter discussion, si l'on envisage leur déploiement concret : comment définir et opérationnaliser la convergence d'indices permettant d'actualiser un stylème ? quel rôle et quelle structure donner à un référentiel, et comment s'articulerait-il à des analyses textométriques ?

Résumé court pour indexation :

Etant donnée une approche stylistique précisée dans "Le style et sa modélisation : document de travail" (Calas & Garric 2009), nous examinons les apports possibles de la textométrie à l'analyse stylistique de corpus et nous pointons un certain nombre de questions soulevées par la formalisation en référentiel hiérarchique de marqueurs.

Mots-clés :

textométrie, lexicométrie, statistique textuelle, stylistique, linguistique de corpus, interprétation.

Abstract :

As a both quantitative and a qualitative textual approach, textometry could help some stylistic corpus analysis. The stylistic analysis we consider here is the one defined in the working paper "Le style et sa modelisation : document de travail" (Calas & Garric 2009). Concretely, the proposal of this paper is to formalize stylistic elements into a hierarchical structure, from linguistic clues (leaves) to style components (root). We examine how the state of art of corpus linguistics and textometry can provide means to fill in such a referential, for the example given in (Calas & Garric 2009). We also discuss the stylistic modelization, should it be implemented on a large scale : the passage from linguistic clues to stylistic makers ; the role and the composition of the stylistic referential, and in particular how this referential can be linked to textometric analysis.

Keywords :

textometry, statistical text analysis, stylistics, corpus linguistics, interpretation.

1) Pour répondre aux besoins d'une analyse textuelle stylistique, qu'offrent les logiciels existants de lexico ou textométrie ?

Ici comme par la suite nous assimilerons lexicométrie et textométrie : la textométrie est une nouvelle manière de nommer la lexicométrie, pour mieux rendre compte des évolutions récentes de cette discipline (traitement de corpus étiquetés notamment) et éviter de laisser entendre que cette approche se cantonne à une étude lexicale.

Parmi les approches outillées d'analyse textuelle, la textométrie se caractérise par une approche à la fois

1) quantitative, foncièrement basée sur des calculs contrastifs (donnant un rôle déterminant au corpus contextualisant les observations), et
2) qualitative, donnant une place centrale au retour au texte, c'est-à-dire à la possibilité d'observer en contexte les occurrences de tout phénomène repéré par les calculs.

La conception de l'analyse stylistique présentée par les organisateurs de ces journées d'étude est également contrastive, le style étant défini comme « écart manifesté par différentes unités langagières dont la matérialité n'est identifiable qu'en corpus par une méthode comparative ». Et la stylistique est aussi traditionnellement qualitative et interprétative, dans la mesure où l'on reconnaît que la lecture, la réception d'un texte, est une activité non-déterministe : deux lecteurs n'ont pas la même lecture, ni même un même lecteur à deux moments différents ; il n'y a pas de lecture définitive, « épuisant » le texte -surtout s'il mérite une analyse stylistique.

Les logiciels de textométrie réalisent une certaine forme d'instrumentalisation, qui ne se confond pas avec une automatisation, Les données de travail et leur traitement sont construits et explicités (unités d'analyse, textes, corpus), à l'inverse d'un fonctionnement en « boîte noire » coupant court à toute interprétation scientifique. Les observations et les saillances détectées par les calculs statistiques sont relatives au corpus et évoluent dynamiquement avec sa composition, ce qui permet d'échapper à une description plus statique opérée par exemple par un repérage de phénomènes prédéfinis et décrits par des patrons.

Par delà cette affinité de fond, et plus concrètement, les logiciels de textométrie ont plusieurs types de fonctionnalités susceptibles d'intéresser la stylistique : notre description se base ici sur les catégories fonctionnelles présentées dans (Pincemin *et al*, 2010).

Pour la recherche qualitative d'attestations, une fonctionnalité de type *Vocabulaire* (selon les logiciels appelée *dictionnaire*, *index*, *liste*, *t-gen*...) permet d'examiner la diversité et la représentativité des différentes formes de réalisation d'une unité donnée : l'expression privilégie-t-elle ou fuit-elle des formes canoniques, procède-t-elle de façon normative ou diffuse, etc. (Rastier, 2001 : 200) ?

Les logiciels permettent également de s'intéresser à la répartition et à la disposition d'unités au fil d'un texte ou au sein d'un corpus. Au fil d'un texte, on peut considérer le positionnement ou la densité des réalisations de l'unité, avec des fonctions de type *Texte*, *Déroulement* (pour un point de

vue plus global), ou *Concordance* (pour un point de vue plus local). Ce pourrait par exemple être une voie d'approche pour la perception de rythmes, de constructions en miroir. Les fonctions de type *Distribution* traitent de la répartition d'unités entre des parties. Celles-ci peuvent être d'ampleurs diverses, et elles ne sont pas nécessairement continues :épisode, locuteur ou foyer énonciatif, position dans la structure (notamment métrique : pied, rime), type de phrase, etc.

La textométrie sert également des considérations quantitatives. A un premier niveau, on peut déjà opposer la présence et l'absence (correspondant à une fréquence nulle). L'exemple de référentiel nous montre en effet que certains indices pourraient être définis par l'absence de certains éléments : la narration se caractériserait par l'absence de déictiques, l'absence de pronoms personnels de première et deuxième personne. Dans certains cas il semble utile d'aller plus loin que la simple opposition présence *vs* absence, et d'évaluer un degré de présence (par exemple la fréquence 1 reçoit un nom -*hapax*- et s'interprète souvent de façon particulière), voire d'absence (cf. le concept de *nullax*, pour repérer des absences statistiquement significatives). Ce degré de présence ou d'absence peut se mesurer de façon absolue (fréquences) ou relative, avec des modèles simples (fréquences relatives, plus intuitives) ou élaborés (spécificités, plus fiables). Mentionnons deux perspectives pertinentes pour la stylistique : le choix de la base des spécificités (pour contraster finement : par exemple, veut-on mesurer la sur-représentation d'une forme verbale par rapport : à tous les mots ? aux seuls verbes ? Aux autres flexions du même verbe ? etc. cf. (Mayaffre, 2006)) ; et la mise au point de jeux de mesures cohérents pour la caractérisation textuelle (par exemple, des rapports entre catégories de mots -lesquelles-, etc.).

La textométrie outille la cooccurrence : la coprésence marquée de deux unités dans le voisinage l'une de l'autre révèle souvent un lien sémantique et aide au repérage de thèmes, elle rejoint la notion de « réseau de cohérence » (Garric & Léglise 2005). Il reste des questions pour la recherche en textométrie, pour mieux visualiser la cooccurrence (non sans lien avec l'isotopie, voir alors le logiciel ThemeEditor (Beust 2002), qui oblige cependant à ne considérer qu'un sème par mot) et la mesurer (pour deux unités ou plus), mais des fonctionnalités de type *Extrait*, *Cooccurrences* et *Associations* donnent déjà de nombreuses possibilités d'investigation.

2) Est-il possible d'automatiser le repérage des indices ? des marqueurs ? des stylèmes ? Quels seraient, le cas échéant, les indicateurs de définition de seuils, si cette notion est nécessaire ?

et

3) Qu'est-ce qui serait automatisable, qu'est-ce qu'il serait possible d'annoter pour guider l'analyse d'un corpus ?

Selon leur nature, le repérage des indices (tels qu'illustrés dans l'exemple de référentiel de Calas & Garric 2009) est plus ou moins automatisable.

Certains indices semblent généralement repérables sur de simples critères typographiques : une graphie comme *donc*, avec ou sans majuscule ; des ponctuations comme les guillemets, les tirets cadratins ; les retours à la ligne.

Les langages d'interrogation permettent aussi d'accéder à des morphèmes représentables comme des sous-chaînes de caractères (ex. le préfixe *re-*), à des locutions comme *il y a* (avec

éventuellement ses variations sur le temps et le mode), certains langages permettent aussi de repérer des séquences discontinues (typiquement une négation verbale en *ne … pas*, *ne … rien*, etc.).

L'étiquetage du corpus, associé à un moteur de recherche évolué, permet de rechercher des catégories ou des traits. L'étiquetage peut être automatisé en recourant à des outils de Traitement Automatique des Langues : la nature et la qualité de l'étiquetage dépendent de l'état de l'art des techniques. Actuellement, un étiquetage morphosyntaxique est généralement possible, même imparfait il peut enrichir les observations : c'est ainsi par exemple qu'on pourrait repérer les *infinitifs*. Moins évident, l'étiquetage sémantique pourrait être un moyen de repérer un indice comme *lexique/affectif*.

Pour des étiquetages élaborés, une piste intéressante, et qui peut être couplée au TAL, est celle des corpus mutables : il s'agit d'instrumenter efficacement une annotation manuelle du corpus, ou une correction manuelle d'annotation automatique. Des vues en concordance, une mémoire d'analyse, aident à propager un choix d'annotation contextualisé tout en contrôlant sa cohérence. Ce genre d'approche est actuellement à l'étude en textométrie.

En combinant les informations apportées par l'étiquetage du corpus et les possibilités apportées par les langages d'interrogation élaborés, on peut rendre compte de constructions, comme *dire que*, ou *verbe/parole + SN/Pro sujet*.

Mais certains exemples d'indices sont d'un autre degré de complexité :

- *antonyme*, qui n'est pas en soi une catégorie codable dans une étiquette, mais plutôt une relation, qui plus est sémantique et dont la reconnaissance est contextuelle,
- les *reprises syntaxiques et lexicales*, dans la mesure où l'on ne préjuge pas de ce qui est repris,
- le *dialogisme* avec l'identification de *renvois structuraux* ou *lexicaux*, où là aussi toute la difficulté serait d'induire ces renvois (rechercher un renvoi particulier donné serait plus abordable)
- et sans doute tout ce qui tourne autour du *rythme*, bien que des recherches aient déjà été amorcées dans ce domaine.

En ce qui concerne la recherche des marqueurs ou des stylèmes, il est difficile de se prononcer sur la faisabilité de leur repérage automatique, dans l'état actuel de la formalisation, qui ne précise pas complètement comment on passe des attestations des indices à la reconnaissance d'un marqueur, et de la reconnaissance de marqueurs à l'actualisation de stylèmes. Les syntaxes d'interrogation courantes nous proposent deux solutions simples : la conjonction (ET) et la disjonction non exclusive (OU). Ainsi, on pourrait très simplement repérer une catégorie si l'on définit sa réalisation comme l'attestation de tous les marqueurs, ou d'au moins l'un d'entre eux. Mais on perçoit les limites d'une telle solution, trop rigide -soit trop contraignante, soit trop accueillante-. Une voie serait à chercher du côté d'opérateurs originaux comme le CUMUL[29] du moteur de recherche TOPIC de la société Vérity, permettant de parcourir les combinaisons partielles intermédiaires entre le ET et le OU tout en les hiérarchisant (ici, cela mettrait en valeur

29 Le nom de l'opérateur est CUMUL dans les versions francophones du moteur de recherche, et ACCRUE dans les versions anglophones.

les attestations de catégories correspondant à la réalisation d'un plus grand nombre de marqueurs différents).

Une avalanche de questions se présentent encore, à définir : comment rendre compte de l'intensité, la saillance de la réalisation d'un marqueur ? Cela peut passer par de multiples facteurs -la répétition, la diversité, l'originalité...-. Techniquement des seuils pourraient être un élément pour prendre en compte cette dimension d'intensité, mais l'association entre une valeur numérique (fixe ?) et une interprétation (contextuelle ?) reste acrobatique. Et, non moins difficile, la question des contextes, zones sur lesquelles s'élaboreraient et porteraient les catégories : une multiplicité de découpages du texte sont possibles, et les passages, potentiellement chevauchants, ne sont pas donnés, ils sont plutôt au terme de l'interprétation...

Reste aussi à rendre compte de la détermination du local par le global affirmée par la sémantique interprétative :

> [Une conception textuelle de l'isotopie] conduit à un déplacement de problématique. En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire les sèmes. (Rastier, 1987, pp. 11-12)

Le repérage des diverses catégories sur la base de marqueurs repose sur des indices locaux : si l'analyse se base sur un parcours du référentiel arborescent depuis les indices jusqu'aux stylèmes, alors le local occulte le global. Une proposition importante du document de travail (Calas & Garric 2009) permet de réintroduire le global : celle de l'étape de choix du référentiel à appliquer, compte tenu des textes à analyser. Ici, la question aussi peut se poser : faudrait-il instrumenter logiciellement la sélection voire la construction d'un référentiel pertinent ? Une étude de corpus servant par exemple à détecter des indices particulièrement actifs, ou à grouper des indices se comportant en faisceaux convergents, peut contribuer à redonner au global un rôle déterminant dans l'analyse.

4) Quelle serait la pertinence de disposer d'un référentiel (avec ambition ou non d'exhaustivité) d'analyse textuelle en vue de son traitement par les logiciels de textométrie ? Sa structuration hiérarchique est-elle facilitante pour la modélisation ? Jusqu'à quel niveau de hiérarchie peut-on espérer aboutir ?

L'ambition d'exhaustivité, comme celle d'universalité d'un unique référentiel, sembleraient déplacées. Un référentiel pourrait avoir plus modestement une pertinence pragmatique, en proposant des éléments réfléchis et déjà pré-construits, à reprendre, adapter, compléter pour mener une analyse textuelle.

La structuration hiérarchique confère au départ une organisation claire et une bonne lisibilité à la modélisation. Mais la cohérence d'ensemble de l'arborescence peut être difficile à maintenir lorsque la modélisation croît. D'ailleurs, l'arborescence donnée en exemple n'est peut-être déjà pas si limpide qu'il n'y paraît : selon le « niveau » auquel on se trouve, les différentes sous-branches n'ont sans doute pas ici le même type de relation (alors que structurellement c'est

identique) : par exemple le *stylème* [Généralisation] peut être compris comme une composition de plusieurs *marqueurs* en relation de complémentarité (<actualisation>, <quantification>, <Effacement énonciatif>, <Universalisation>), alors qu'à un niveau inférieur, la *macro-catégorie* [Actualisation] est définie par des *macro-marqueurs* plutôt en relation paradigmatique, d'alternative non exclusive (<déterminant défini>, <déterminant indéfini>). Autrement dit, la réalisation du stylème semble davantage reposer sur la réalisation de plusieurs classes de marqueurs différentes, alors que la réalisation de la macro-catégorie pourrait se satisfaire d'attestations relevant d'un seul macro-marqueur. Dans le même ordre d'idées, il n'est pas sûr qu'une augmentation du nombre de niveaux soit un progrès : une structure peu imbriquée et plus souple, en réseau, pourrait être plus appropriée à la description linguistique et stylistique.

Reste aussi, entière, la question de l'articulation entre un tel référentiel et une analyse textométrique : les logiciels « traiteraient » ce référentiel, de quelle manière au juste ? De multiples réponses sont possibles, aucune n'épuisant les attentes d'une analyse stylistique :

1) le logiciel de textométrie peut être utilisé comme un moteur de recherche, pour observer les attestations de tel marqueur, telle catégorie, tel stylème en corpus ;
2) les stylèmes, une fois reconnus, pourraient devenir un niveau de description traitable par un calcul textométrique (observer et caractériser leur récurrence, leur cooccurrence, etc.) : cela supposerait de préciser la construction d'une telle représentation, en particulier comment localiser les stylèmes, voire de réviser le modèle textométrique (prise en compte de types ensemblistes) ;
3) les stylèmes pourraient aussi tout simplement guider l'analyse basée sur un autre niveau de représentation plus classique : suggestion d'unités à observer, éléments de modélisation pour structurer et interpréter des résultats de calculs ;
4) un point de vue inversé pourrait être tout aussi intéressant : à savoir, comment la textométrie peut contribuer à définir des marqueurs ; dans cet esprit, l'enjeu est moins de se doter d'un référentiel, que d 'élaborer des repères méthodologiques pour forger, à partir de chaque corpus, une description adaptée.

5) En quoi la constitution des corpus est-elle fondamentale dans la démarche ? Est-il possible pour les besoins de l'analyse des marqueurs stylistiques de contraster au sein même d'un texte ? ou le contraste n'est-il pensable et "rentable" que sur des textes externes (intragénériques ou intergénériques) au texte support à l'intérieur d'un grand corpus ?

La part statistique des calculs textométrique suppose un corpus, définissant un référentiel de fréquences. Ceci étant, rien n'empêche de calculer une concordance ou des segments répétés sur un seul texte ; même des calculs statistiques peuvent être lancés -sans grande significativité ni fiabilité bien sûr.

Bref, la question de la taille du corpus peut être vue comme une question de « droit » (est-il mathématiquement, statistiquement, légitime de faire tel ou tel calcul, quelle interprétation peut-on donner aux résultats produits) que tout simplement comme une question de (bon) sens : si le

corpus est très petit, la lecture textométrique peut-elle vraiment apporter quelque chose qu'une lecture humaine attentive et experte n'aurait pas décelé ?

Si le texte et corpus est un sonnet, la textométrie aura un intérêt limité. Si le texte et corpus est un roman de Balzac, on peut étudier la composition du roman, sa structuration en chapitres, l'évolution d'unités au fil du texte.

Mais le corpus joue ce rôle décisif de « représenter le monde » au sein duquel se déploie l'analyse ; et il est clair que la pratique herméneutique est fortement intertextuelle : un texte n'est pas reçu isolément, il s'inscrit et prend sens dans une lignée générique, dans une pratique, dans une histoire. Etudier un texte en corpus, c'est se donner un moyen de l'observer avec une profondeur intertextuelle certes artificielle et limitée, mais susceptible de rendre compte d'une part de son fonctionnement intertextuel. Cette conception du corpus est exigeante : concrètement, elle se heurte à la faisabilité de disposer de versions numérisées des textes de l'intertexte idéal visé (pb de droits, coût de numérisation, etc.) ; mais plus encore, la difficulté est théorique : celle de la composition de ce corpus. Tout corpus est un choix, qui peut être objectivé (par son explicitation), mais n'est jamais neutre. Il n'y a pas une solution unique ou définitive, qui permettrait de résoudre la question du corpus, qui peut toujours être réouverte.

Enfin, méfions-nous du « grand » corpus : ce n'est pas la taille qui fait la valeur, c'est l'interprétabilité. Si on me donne un corpus dont je ne connais pas la composition, je ne saurai rien tirer des contrastes observables. (Pincemin 1999).

6) Comment faudrait-il situer l'étape "générique" par rapport à l'étape "stylistique" dans la construction et l'individuation de la textualité ?

Il n'est a priori pas évident de démêler ce qui relèverait du genre et ce qui relèverait d'un style (voire même de définir clairement genre et style) : un style pourrait nourrir un genre, un genre pourrait orienter un style.

Des expériences textométriques ont montré, sur un corpus de littérature française, que le genre était une caractérisation du texte plus forte que l'auteur (cf. travaux de Brunet, par exemple (Muller & Brunet 1988)).

7) Quelle serait la place d'une analyse stylistique automatisée au sein de l'herméneutique linguistique ?

L'automatisation est ici à concevoir comme une instrumentation, mobilisée intelligemment par un stylisticien, et soumettant à son interprétation de nouvelles observations (inattendues, systématiques, à grande échelle).

La rapidité et l'efficacité des investigations en corpus est à la fois un atout et un piège. L'atout, c'est la possibilité démultipliée de confronter à un corpus des hypothèses de modélisation : la théorie peut ainsi s'élaborer dynamiquement, sans attachement excessif à un état de réflexion et d'observation donné. Le piège, ce serait l'illusion que le résultat d'un calcul est en soi une réponse, et l'oubli de la nécessité d'une interprétation englobante, tant dans la conception du traitement

(choix du calcul, des données), que dans une lecture méthodique et synthétique des produits du calcul.

Références

Beust Pierre (2002) - « Un outil de coloriage de corpus pour la représentation de thèmes », *Actes des 6es Journées internationales d'analyse statistique des données textuelles (JADT 2002)*, Annie Morin & Pascale Sébillot (éds), Editions de l'INRIA, pp. 161-172.

Calas Frédéric, Garric Nathalie (2009) - *Le style et sa modélisation : document de travail*, document préparatoire aux journées d'études « Le style et sa modélisation », Université de Tours, 10 et 11 décembre 2009, 30 pages.

Garric Nathalie, Léglise Isabelle (2005) - « La place du corpus, de l'analyste, du logiciel : exemple d'une analyse de discours patronal à deux voix », *in* G. Williams (ed.), *Linguistique de corpus,* Presses universitaires de Rennes, pp. 101-113.

Lebart Ludovic, Salem André (1994) – *Statistique textuelle*, Dunod.

Mayaffre Damon (2006) - « Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République », *Actes des 8es Journées internationales d'analyse statistique des données textuelles (JADT 2004)*, Jean-Marie Viprey (éd.), Presses universitaires de Franche-Comté, Besançon, 19-21 avril 2006.

Muller Charles, Brunet Etienne (1988) - « La statistique résout-elle les problèmes d'attribution ? », *Strumenti critici* vol. III, n°3, Florence, 1988, pp. 367-387.

Pincemin Bénédicte (1999) - « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », *Atelier Corpus et TAL : pour une réflexion méthodologique, 6e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 99)*, Cargèse (Corse, France), 12-17 juillet 1999, Anne Condamines, Marie-Paule Péry-Woodley et Cécile Fabre (éds), pp. 26-36.

Pincemin Bénédicte (2004) - « Lexicométrie sur corpus étiquetés », *Actes des 7es Journées internationales d'analyse statistique des données textuelles (JADT 2004)*, Gérald Purnelle & al. (éds), Presse universitaires de Louvain, Louvain-la-Neuve (Belgium), 10-12 mars 2004, vol. II, pp. 865-873.

Rastier François (1987) – *Sémantique interprétative*, Presses Universitaires de France.

Rastier François (2001) - *Arts et sciences du texte*, Presses Universitaires de France.

Pincemin Bénédicte, Heiden Serge, Lay Marie-Hélène, Leblanc Jean-Marc, Viprey Jean-Marie (2010) - « Fonctionnalités textométriques : Proposition de typologie selon un point de vue utilisateur », *in* Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds), *Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010*, Edizioni Universitarie di Lettere Economia Diritto, Rome, 9-11 juin 2010.

Stylistique et textométrie

Sept questions de principe et d'opportunité

François Rastier

Directeur de recherche, INaLCO,

Équipe de recherche Texte, Informatique, Multilinguisme.

Résumé : En s'appuyant sur les méthodes textométriques et les acquis de la linguistique de corpus, la caractérisation stylistique des œuvres littéraires peut opportunément se renouveler voire se refonder.

Mots clés : Style, linguistique des œuvres, Spitzer, Brunet, interprétation, variables, sémiosis textuelle.

Abstract: By drawing on the textometric methods and the achievements of corpus linguistics, the stylistic characterization of literary works can seize the opportunity to renew itself or even re-establish itself.

Keywords : style, linguistics of literary works, interpretation, textual semiosis.

*

Questions posées par Hélène Maurel-Indart, Nathalie Garric et Frédéric Calas à l'occasion des journées d'études *Le style et sa modélisation*, Université François Rabelais, Tours, 11-12 décembre 2009.

1) Pour répondre aux besoins d'une analyse textuelle stylistique, qu'offrent les logiciels existants de lexico ou textométrie ?

Ces logiciels traitent des mots graphiques (ex. Lexico 3), ou bien les unités étiquetées par un analyseur morphosyntaxique (ainsi Hyperbase utilise Cordial ou Treetagger) ; certains sont aussi capables de traiter une segmentation en unités préalablement encodée dans le corpus, ils peuvent alors opérer sur des unités élaborées (ex. Weblex, Le Trameur).

Les fonctionnalités statistiques de base ont été mises au point dès les années soixante (application de l'écart réduit aux données lexicales, analyses factorielles), et l'essentiel était acquis dans le courant des années quatre-vingts (comme les calculs de spécificités, de cooccurrences, ou de segments répétés).

Depuis une quinzaine d'années, on assiste à une évolution vers la textométrie, qui tient compte des unités de segmentation textuelle (comme le paragraphe) ou des unités sémantiques diffuses comme les thèmes (je pense à la fonction *thème* dans Hyperbase, implantée pour les besoins d'une analyse de corpus romanesques, cf. l'auteur, éd. 1996).

Les besoins sont plutôt méthodologiques et épistémologiques : pour travailler sur les styles, et tout particulièrement les styles littéraires, la stylistique des concours, qui juxtapose des critères grammaticaux et une poétique des procédés, reste d'un faible secours. La notion philologique de corpus lui reste pour l'essentiel étrangère, car les monographies priment, héritage sans doute de la monumentalisation romantique des auteurs.

À l'échelon international, les conceptions dominantes de la littérature, notamment dans les *cultural studies*, qui s'embarrassent peu d'objectivation, n'ont pas permis de faire le lien entre les méthodes de la linguistique de corpus et les études littéraires. Il faudrait pour cela renouer avec une conception philologique des œuvres ; et là encore, ce sont les antiquisants et les médiévistes qui montrent la voie d'une modernisation.

La lexicométrie française s'est en particulier attachée à l'analyse du discours politique (illustrée par le laboratoire de lexicologie politique de l'ENS Saint-Cloud et la revue *Mots*). Certains auteurs, comme Brunet, ont plutôt privilégié les textes littéraires. Mais s'il est déjà difficile de faire admettre une linguistique des textes, qu'en serait-il pour une linguistique des œuvres ? (notons que la séquence « la linguistique des œuvres » ne trouve aujourd'hui aucune occurrence sur Google, contre 49. 500 pour « la linguistique des textes » ; et si « une linguistique des œuvres » trouve une occurrence, c'est je l'avoue sous la plume d'un de mes pseudonymes).

2) Est-il possible d'automatiser le repérage des indices ? des marqueurs ? des stylèmes ? Quels seraient, le cas échéant, les indicateurs de définition de seuils, si cette notion est nécessaire ?

La notion herméneutique d'indice, la notion grammaticale de marque et la notion sémiotique de stylème, que l'on doit à Hjelmslev, ne sauraient donner lieu à des inventaires fermés. Ce sont des réifications *a posteriori* de parcours interprétatifs permettant de localiser tel ou tel signe ou groupe de signes. Ainsi admet-on en grammaire le marqueur zéro, sorte de joker qui supplée l'absence d'un signifiant attendu.

L'éventail des variables d'interrogation dépend de l'état de l'art en textométrie. On peut suggérer d'accroître leur nombre, mais beaucoup de possibilités peu exploitées sont déjà ouvertes. D'autant plus que des variables sans fondement linguistique, qui, plus exactement, échappent à l'imaginaire logico-grammatical, comme les groupements de lettres, les trigrammes par exemple, peuvent se révéler discriminantes.

Si certains types de variables morphosyntaxiques sont beaucoup plus sensibles que d'autres, comme par exemple les temps et les pronoms personnels, les coalitions de variables sont particulièrement révélatrices, pour autant bien entendu que l'on sache les interpréter. Elles pourraient être nommées des *stylèmes*, mais l'inventaire *a priori* de ces coalitions, éminemment variables, me paraît difficile et sans doute superflu.

Telle ou telle spécificité statistique peut être considérée comme caractéristique d'une œuvre ou d'un auteur. Si le relevé de spécificités ne dispense évidemment pas d'une interprétation (qui par malheur, ou par bonheur, n'est pas automatisable), il a tout d'abord une valeur heuristique : comme le Sphinx, le logiciel répond par une énigme.

Gardons-nous de formuler des consignes sous la forme de listes d'instructions. Nous pouvons sans plus proposer des recommandations à caractère déontologique. Par exemple, les seuils à retenir dépendent de la constitution même du corpus : ainsi, dans un corpus homogène monogenre, le seuil d'écart réduit peut descendre à 2 ; dans un corpus de champ générique (comprenant par exemple des romans psychologiques et des romans policiers), il vaut mieux le remonter à 3.

La caractérisation ne peut dépendre d'une seule variable, qu'elle soit locale ou globale. Pour des raisons sémiotiques fondamentales, il faut *a minima* coaliser des variables d'expression et de contenu : tout texte, tout genre se définit par un mode de sémiosis qui détermine une corrélation spécifique entre contenu et expression. Plus les coalitions de variables sont denses et stables, plus la conjecture descriptive produit de nouveaux observables.

Pour aller plus loin, il nous faudrait un programme de recherche systématique sur les coalitions de variables : on en trouve déjà des éléments dans les travaux de chercheurs comme Céline Poudat et Sylvain Loiseau.

Comme le principe même d'un approfondissement théorique se trouve contesté par des auteurs influents (cf. Thomas Pavel, *Le mirage linguistique* ; Antoine Compagnon, *Le démon de la théorie*), la théorie littéraire a de fait rompu les relations épisodiques qu'elle entretenait avec la linguistique et se trouve fort dépourvue aujourd'hui devant les analyses textométriques.

Pour en saisir la portée et en tirer profit, il lui faudrait admettre les points suivants (je résume à grands traits) : (i) La littérature est un art du langage et non l'expression plus ou moins transparente d'une subjectivité de l'auteur ou du lecteur – j'ai pourtant entendu un stylisticien illustre conclure un colloque par ce regret : « C'est dommage, on en est restés au niveau du verbal ». (ii) Les textes littéraires, même contemporains, appellent une étude philologique préalable, et en particulier la constitution des corpus suppose une philologie numérique (cf. l'auteur, 2001a). (iii) Pour ce qui concerne l'herméneutique, laissons ici de côté les développements obscurantistes des herméneutiques d'inspiration heideggérienne. En revanche, l'herméneutique matérielle, appuyée sur la philologie, réfléchit la pratique interprétative, tout à la fois nécessaire et incoercible, pour lui conférer une rigueur critique. Au moment de sa constitution en corps doctrinal, dans l'Alexandrie hellénistique, la grammaire était une discipline auxiliaire pour l'interprétation des textes, littéraires notamment. La linguistique contemporaine a notoirement occulté cette dimension, au profit d'une conception logique de l'interprétation comme transcodage.

Or, voici qu'outre les textes proprement dits, les sorties logicielles demandent à être interprétées : elles soulèvent des difficultés imprévues, parfois insurmontées, qui semblent relativiser les clarifications que l'on peut en attendre. La théorie littéraire ne peut déléguer à la textométrie cette tâche d'interprétation qui exige au demeurant un délicat travail d'explication : comment faire admettre que l'herméneutique des sorties logicielles peut pourvoir la stylistique d'une base empirique ?

3) Qu'est-ce qui serait automatisable, qu'est-ce qu'il serait possible d'annoter pour guider l'analyse d'un corpus ?

Les tâches fastidieuses (relevés, etc.) peuvent être heureusement déléguées, mais rien d'important n'est automatisable. Si le travail reste long – souvent, les thèses durent un an de plus — il fait apparaître de nouveaux observables qui peuvent se révéler précieux.

Certains étiqueteurs morphosyntaxiques peuvent être très utiles. On peut imaginer aussi des étiqueteurs sémantiques basés sur des dictionnaires (cf. le projet Dixem initié par Mathieu Valette et la thèse en cours de Coralie Reutenauer).

La parcimonie reste de mise, car chaque type d'annotation anticipe un type de description, et il serait coûteux de constituer des « ressources » oiseuses, comme on le voit par exemple dans le domaine des ontologies. Aucune annotation n'est utile en soi, à l'exception bien entendu des mentions philologiques.

Gardons-nous de réifier *le* corpus. On ne part pas d'un corpus pour en produire une analyse - du moins cette généralité de méthode ne permet-elle pas de cerner la dimension critique de la description scientifique propre aux sciences de la culture. On part plutôt d'un texte ou d'un groupe de textes, pour trouver le corpus dans lequel il doit être plongé – en d'autres termes, son intertexte : la description « transforme » alors le texte ou groupe de textes en corpus de travail et le corpus de référence en intertexte. En d'autres termes, elle constitue une première globalité (le corpus de travail), puis elle la caractérise pour l'éclairer par une globalité supérieure, celle d'un corpus de référence.

Un corpus n'est pas une macro-unité, mais une collection qui dépend d'un point de vue, d'un faisceau d'hypothèses qui peut voire doit varier au cours de la recherche : le texte devient alors le point de rencontre entre plusieurs corpus génétiques et herméneutiques. Par exemple, tel texte de Rimbaud (je pense à *Marine*) réécrit Banville et Baudelaire, mais aussi Horace et Virgile ; je m'en suis avisé en pratiquant une série d'extensions concentriques du corpus d'étude initial. On ne peut donc tenir pour acquis qu'un texte n'ait qu'un corpus, qui plus est unilingue.

Bref, on renouvelle la lecture d'un texte en le plongeant dans un corpus dont il procède, mais qu'il peut dissimuler. Cela permet en premier lieu de déceler les acceptions lexicales ; par exemple, dans les textes les plus ouvertement nazis de Heidegger, *barbarisch* revêt une acception positive (cf. l'auteur, 2009). La méthode interprétative conduit ainsi à approfondir l'analyse initiale du corpus de travail, pour déterminer un corpus de référence, puis le faire varier de manière critique.

L'*intentio auctoris* reste accessible notamment par le corpus initial de composition, l'*intentio lectoris* par le corpus de lecture, l'*intentio operis* par le corpus dans lequel l'œuvre prend place et dont elle se revendique, fût-ce de façon implicite ou trompeuse. Si l'on convenait de ces conjectures, l'on pourrait en somme redéfinir l'ensemble du processus herméneutique en termes de corpus.

4) Quelle serait la pertinence de disposer d'un référentiel (avec ambition ou non d'exhaustivité) d'analyse textuelle en vue de son traitement par les logiciels de textométrie ? Sa structuration hiérarchique est-elle facilitante pour la modélisation ? Jusqu'à quel niveau de hiérarchie peut-on espérer aboutir ?

Un référentiel peut avoir une valeur didactique, par exemple pour mettre en œuvre une stylistique grammaticale conforme aux programmes de concours. Il risquerait toutefois de rester normatif, car il reste difficile sinon impossible d'anticiper, même en se limitant à des catégories grammaticales : par exemple, la nette corrélation entre l'imparfait et le point-virgule, dans le corpus roman 1830-1970 de Frantext, peut être rapportée à la dominance quantitative des romans psychologiques ; un corpus de polars privilégie en revanche les temps perfectifs et se prive des points-virgules, qui marquent un suspens par trop propice à la réflexion.

Les corrélations entre niveaux et paliers d'analyse mettent les théories linguistiques au défi et revêtent donc une valeur heuristique. Chaque œuvre appellerait son propre « référentiel », dans la mesure où ces corrélations témoignent de la sémiosis textuelle particulière que l'on pourrait nommer le « style » de l'œuvre. Cette singularité ne peut être caractérisée que par contraste. Un corpus de référence est donc nécessaire pour faire apparaître les spécificités statistiques des œuvres. (Un corpus national dans lequel les chercheurs pourraient librement puiser pour constituer des corpus de référence serait bien utile. Rappelons que Frantext n'est pas un corpus disponible et qu'il n'a d'ailleurs pas été constitué pour l'étude de la littérature).

Je plaiderais volontiers pour la méthodologie contrastive, issue de la sémantique différentielle, qui privilégie les comparaisons par auteur, genres, discours. Elle recherche des coalitions de variables linguistiques caractérisantes, en utilisant les méthodes quantitatives, aussi bien sur textes nus que sur textes étiquetés par des analyseurs morphosyntaxiques. Les caractérisations sont globales (classification automatique de textes et de sous-corpus) ou locales (cooccurrents contextuels). Elles tiennent compte de tous les critères disponibles dans l'état de l'art (lexique, mais aussi ponctuation, longueur moyenne des mots, etc.). Hors de la littérature, cette méthodologie a été mise à profit dans le projet européen Princip.Net (Inalco-Magdebourg-Dublin, 2002-2004) et le projet ANR C-Mantic (Ertim-Inalco, Lina, Limsi, en cours).

5) En quoi la constitution des corpus est-elle fondamentale dans la démarche ? Est-il possible pour les besoins de l'analyse des marqueurs stylistiques de contraster au sein même d'un texte ? ou le contraste n'est-il pensable et "rentable" que sur des textes externes (intragénériques ou intergénériques) au texte support à l'intérieur d'un grand corpus ?

Le contraste interne entre parties du texte et le contraste entre textes ne se contredisent en rien : tout dépend du projet descriptif. De fait, il n'y a pas de clôture du texte telle que l'on puisse distinguer avec certitude l'interne de l'externe. Non seulement tout texte s'inscrit dans un corpus, mais il s'écrit à partir d'autres textes qu'il conteste et/ou avec lesquels il rivalise. En outre, la littérature est aussi faite de ce qu'elle ne dit pas, et, si l'on s'en tient aux mots, seule une analyse en corpus peut déceler les mots absents et accuser les singularités des mots présents. Par exemple, l'adverbe *circulairement* dans un poème de Rimbaud, *Marine*, est le seul adverbe du texte, mais c'est aussi un hapax dans la poésie de son temps. Ces deux propriétés se renforcent : comme ce texte est le premier poème français en vers libres (avec *Mouvement*), on peut suggérer que la « libération » du vers s'accompagne de la « libération » du vocabulaire poétique.

6) Comment faudrait-il situer l'étape "générique" par rapport à l'étape "stylistique" dans la construction et l'individuation de la textualité ?

Les genres ont mauvaise presse en théorie littéraire, notamment chez les modernistes ; on concède sans plus qu'ils sont indéfinissables (Todorov, Schaeffer) et l'on rappelle la consigne de Barthes : « se tirer des sociolectes », pour conclure que l'œuvre est à elle même son genre, ce qui perpétue une conception monographique et monumentale de la littérature. Notablement, les théoriciens de la littérature ne se réfèrent jamais aux études empiriques en lexicométrie, textométrie et linguistique de corpus qui depuis plusieurs décennies ont confirmé la prégnance des genres, en littérature comme dans les autres discours (pour une introduction, cf. Malrieu et Rastier, 2001).

C'est là sans doute un effet du romantisme, à présent fort tardif, où nous sommes encore. Le style individuel ne devient d'ailleurs un objet de pensée qu'avec le premier romantisme, et le mot *Stylistik* fut créé par Novalis en 1800. Issue de l'idéalisme romantique et notamment de la lecture de Buffon par les Romantiques d'Iéna, la notion moderne de style reste individualisante, et s'est prêtée à unir l'homme et l'œuvre (comme en témoigne la déclaration aussi apocryphe qu'illustre : « Madame Bovary, c'est moi »).

Les corpus littéraires permettent cependant un point de vue critique sur la monumentalisation des œuvres et des auteurs. Les œuvres ne sont pas pour autant noyées dans un mélange indistinct : comme le sens est fait de différences, différencier systématiquement les œuvres permet en quelque sorte de leur conférer un surcroît de sens.

7) Quelle serait la place d'une analyse stylistique automatisée au sein de l'herméneutique linguistique ?

Nous sommes ici au croisement de la linguistique et de la théorie littéraire : à quelle condition un texte devient-il une œuvre ? Cela dépend de son caractère, qui le rend singulier et irremplaçable, et lui permet ainsi d'ouvrir la tradition interprétative qui peut l'ériger en classique. Si l'on identifie ce caractère au style, une perspective subjectivante peut le rapporter à l'auteur et l'expliquer par sa biographie psychologique, alors qu'une perspective objectivante le rapporte à des formes textuelles particulières. Nous choisissons la seconde, car nous avons à expliquer les œuvres en termes d'œuvres : un auteur, docile reconstruction des biographes, peut sembler compréhensible, mais cette compréhension empathique n'explique rien de son œuvre, où il s'efface non moins qu'il ne s'exprime. Il reste d'ailleurs toujours plus facile de croire comprendre les auteurs que de connaître les œuvres.

Si l'on convient que le style est dans les œuvres et non dans les auteurs, un style n'est peut-être que l'abstraction d'une œuvre. On appellerait *style* ses régularités propres : elle répond de son style, et non l'inverse.

Quand au « style d'auteur », il pose le problème des régularités au sein même des ouvrages d'un même auteur ; on y voit s'établir des lignées stylistiques, les caractères des premières œuvres se développant dans celles qui suivent.

Pour éviter l'involution psychologiste et poser correctement le problème esthétique, distinguons cependant l'identification et la caractérisation, ou si l'on préfère les traits

« morelliens » et les traits « spitzériens ». Morelli, médecin italien, révolutionna à la fin du XIX[e] siècle les attributions de tableaux en décelant des traits, notamment anatomiques, comme les lobes d'oreille, dont la facture caractéristique échappait jusque-là aux faussaires comme aux experts. Quant à Spitzer, on lui a maintes fois reproché de caractériser les œuvres par des traits formels qui paraissaient choisis arbitrairement, mais lui permettaient pourtant d'entrer dans le cercle vertueux d'une interprétation révélatrice.

Ainsi, à l'identification par les caractères morelliens, on peut opposer la caractérisation par les caractères spitzériens : les premiers, répartis régulièrement, se répètent d'œuvre en œuvre et ne se signalent pas par une connectivité sémantique particulière ; les seconds en revanche sont singuliers, d'un haut degré de connectivité, et font l'objet de transpositions à tous les paliers de complexité de l'œuvre : par exemple, l'usage singulier de l'hypallage renvoie chez Borges à une ontologie négative qui commande la structure métaphysique de son œuvre entière (cf. l'auteur, 2001b).

Cette distinction conduit à séparer les traits de facture et les phénomènes de style proprement dits. Certes, un auteur pourrait styliser ses propres traits morelliens, quand par exemple il se parodie lui-même, ou plus profondément quand il élabore son style pour ne plus rien laisser au hasard de l'habitude ; c'est sans doute une des raisons de l'étrangeté de Flaubert.

Une contradiction semble cependant ruiner la reconduction du style à l'auteur : ses œuvres peuvent n'être immédiatement fédérées que par des traits morelliens, et leur dénominateur commun se réduit alors à des traits de facture : le « style d'auteur » serait ainsi *ce qu'il y a de plus superficiel* dans l'œuvre. Les styles des œuvres d'un même auteur peuvent partager des caractéristiques communes, mais pour l'essentiel ils varient avec les genres dont use l'auteur et les divers projets esthétiques dans lesquels il s'engage.

En revanche, le « style d'une œuvre » se définit par les traits générateurs de la structure artistique : ces formes particulières se transposent, tant au plan de l'expression qu'à celui du contenu, tant au palier de la phrase qu'à celui du texte global.

En somme, l'identification morellienne permet d'attribuer une œuvre à son auteur, de l'isoler dans son corpus de référence, mais non de décrire le fonctionnement propre de ses parcours génétiques, mimétiques et herméneutiques. En revanche, la caractérisation « spitzérienne » permet sa singularisation interne et conduit à identifier les contraintes que sa forme artistique exerce sur ces parcours. Elle suppose enfin et permet tout à la fois une interprétation qui parcourt le corpus de référence où l'œuvre se singularise, car les parcours interprétatifs requièrent souvent des interprétants qui sont situés dans d'autres textes.

D'où proviennent les traits morelliens ? Les habitudes pratiques de facture restent ordinairement compatibles avec les normes de langue, de discours et de genre. Elles en exploitent les possibilités, par des choix récurrents au sein d'une norme permissive. Cette sélection du matériau linguistique, discursif et générique constitue une première phase, élémentaire, de la stylisation qui radicalise déjà certaines propriétés systématiques.

Les traits identificatoires les plus efficaces sont des traits morelliens de « bas niveau », comme la fréquence des lettres : comme maints travaux de linguistique quantitative l'ont confirmé, elle permet tant l'identification des auteurs que des œuvres. Même ces traits de « bas niveau » peuvent faire l'objet d'une élaboration stylistique et changer de statut : par exemple, le jeu pathétique de

Perec avec la fréquence des lettres, dans *La disparition*. La fréquence anormale de la lettre *e*, en l'occurrence zéro, est alors promue au rang de trait spitzérien, entendu comme principe organisateur : cette absence évoque alors le deuil d'un orphelin de l'extermination.

Les « traits » spitzériens ne sont pas à proprement parler des traits, du moins au sens atomiste du terme, mais des formes d'organisation, transposables à différents niveaux de complexité, entre lesquels ils établissent des solidarités d'échelle. Ils deviennent ainsi des principes organisateurs de la textualité. Par exemple, l'hypallage chez Borges fait partie des traits spitzériens : le troc indécidable d'attributs qui caractérise les hypallages se transpose au niveau séquentiel (tactique) par des formes en chiasme, au niveau narratif (dialectique) par des récits où les acteurs échangent leurs propriétés, au niveau énonciatif (dialogique) par l'indistinction du lecteur et du narrateur, etc.

Les traits spitzériens restent que je sache propres aux textes littéraires, alors que les traits morelliens se décèlent dans d'autres discours. Ces deux modes de description stylistique s'accordent avec les objectifs d'une linguistique non restreinte. Si, en revanche, outrepassant l'objectif déjà ambitieux d'une individuation, on fixe à la caractérisation le but ultime de conduire à une individualisation, si l'on n'a de cesse de reconduire l'œuvre à l'auteur tel qu'on l'imagine, on la livre aux fades délices de l'empathie universitaire théorisée par une certaine esthétique de la réception.

En somme, les traits morelliens sont des traits d'individualisation de l'auteur, car ils ne dépendent pas du projet esthétique d'une œuvre déterminée : le style d'auteur est ainsi descriptible par des traits morelliens. En revanche, les traits spitzériens individualisent l'œuvre et non l'auteur ; dans la relation entre local et global, ils reflètent l'architectonique de l'œuvre.

Mais que seraient les *traits de contraste*, qui apparaissent par comparaison de l'œuvre avec un corpus de référence, et que je brûle d'appeler les traits *brunettiens*, en hommage à Étienne Brunet. À la différence des traits spitzériens (généralement dégagés par une analyse interne à l'œuvre singulière) et des traits morelliens (propres au corpus d'un auteur), on ne peut objectiver les traits brunettiens que par les méthodes contrastives au sein d'un corpus de référence multiauteurs. Ils sont donc hautement variables, dans la mesure où ils dépendent des critères de constitution de ce corpus de référence. Mais ils sont stables cependant : par exemple, les corpus techniques se signalent par la fréquence insolite de la lettre *d*, sans doute favorisée par la récurrence de syntagmes de forme *N de N*.

Outre qu'ils sont dépendants de l'état de l'art, les trois types de traits dépendent, pour leur mise en relief, du corpus de travail : l'œuvre singulière pour les traits spitzériens ; l'œuvre complète d'un auteur pour les traits morelliens ; un corpus d'œuvres de même genre, de même champ générique ou de même discours pour les traits brunettiens.

Il faudrait un programme de recherche coordonné pour contraster ces trois sortes de coalitions de traits, et je ne formulerai que quelques hypothèses. (i) Les traits morelliens sont uniformément répartis. Ils ne font pas une coalition. En d'autres termes le spectre qu'ils constituent est diffus. (ii) Les traits spitzériens forment une coalition et dessinent donc un spectre formé. (iii) Les traits brunettiens caractérisent le corpus de travail par rapport au corpus de référence : ils participent à la mesure de distances, en général au sein du même genre et du même discours.

Comme les spécificités du texte donnent accès aux trois types de traits, il convient de différencier les méthodologies pour ne pas les confondre. Par exemple, la détection de pastiche fait appel aux traits morelliens, alors que la détection de plagiat pourrait soustraire les traits brunettiens pour accuser les similarités entre textes. Si le pastiche affiche son intertexte, le plagiat le cache. Il pose des problèmes différents de ceux que soulève l'analyse stylistique caractérisante : ainsi, quand Céline, avant même la guerre, plagie les brochures antisémites de la *Propagandastaffel*, il les enrubanne des ficelles de son propre style (au sens morellien), points de suspension et d'exclamation compris.

Pour sa vigueur problématique sinon heuristique, on peut bien entendu conserver la notion de stylistique, mais elle ne sortira pas indemne de sa refondation en linguistique de corpus : d'une discipline esthétique (de tradition philosophique), il lui faudra se transformer, par essais et erreurs, en champ scientifique, pour désigner une typologie des coalitions de traits spécifiants. Elle ne délaissera pas pour autant les problèmes esthétiques – voire éthiques – mais elle les abordera par d'autres voies : par exemple, en contrastant un corpus de témoignages littéraires de l'extermination et un corpus de faux témoignages des plus littéraires, Charlotte Lacoste a pu mettre en évidence que les narrateurs authentiques disent *nous*, car ils expriment la dette des survivants à l'égard des engloutis, et les inauthentiques disent *je*, car ils usent de toutes les grosses ficelles de l'écriture du Moi. Mais nous manquons encore d'une *linguistique des œuvres*.

J'ignore cependant dans quelle mesure la stylistique universitaire actuelle, discipline de compromis qui mêle des critères issus de la grammaire et de la rhétorique, pourra accueillir des traits descriptifs inusités et qui seraient restés inobservables sans les logiciels textométriques, comme par exemple la longueur moyenne des mots ou la position topographique fine.

Le renouvellement des méthodes pourrait menacer des traditions académiques bien établies mais l'on peut souhaiter que les études littéraires, pourtant fort dépendantes du système des concours, se prêtent à une évolution méthodologique d'autant plus nécessaire qu'elles sont menacées de toutes parts.

N.B. — J'ai plaisir à remercier de leurs avis Bénédicte Pincemin et Carine Duteil.

Références

Compagnon, Antoine (2001) *Le démon de la théorie*, Paris, Seuil.
Loiseau, Sylvain (à paraître) Investigating the interactions between different axes of variation in text typology, in Grzybek P. & Kelih E. (éd.), *Text and Language : Structures, Functions, Interrelations.*
Malrieu, Denise, et Rastier, François (2001) Genres et variations morphosyntaxiques, *Traitements automatiques du langage*, 42, 2, pp. 547-577.
Pavel, Thomas (1989) *Le mirage linguistique*, Paris, Minuit.
Rastier, François, éd. (1996) *L'analyse thématique des données textuelles — L'exemple des sentiments*, Paris, Didier.
Rastier, François (2001a) *Arts et sciences du texte*, Paris, PUF, 303 p.
Rastier, François (2001b) Borges et l'hypallage, *Variaciones Borges*, 11, pp. 3-33.
Rastier, François (2009) *Labyrinthe*, n° 33, 2009 (2), pp. 71-108.
Schaeffer, Jean-Marie (1989) *Qu'est-ce qu'un genre littéraire ?,* Paris, Seuil.

Le style et sa modélisation, perspective ALCESTE

Max Reinert

(max.reinert@uvsq.fr)

UVSQ-CNRS UMR8085 (membre associé)

Résumé :

Le point de départ de cette contribution est le « cahier des charges » proposé par les organisateurs des deux journées sur « le style et sa modélisation » à l'Université François-Rabelais de Tours, les 10 et 11 décembre 2009, cahier composé de différentes questions concernant une opérationnalisation informatique pour le repérage des marques de style. Pour des raisons qui touchent une conception du style, et qui apparaîtront au fur et à mesure de ma réponse, je n'ai pu répondre qu'aux deux premières questions. Aussi j'ai ajouté, en troisième partie, un texte sur la méthode d'analyse statistique de discours informatisée dans le logiciel « ALCESTE » pour éclairer la relation entre la conception du style que je soutiens et l'approche exploratoire des discours associée à ce logiciel. Cela étant dit, mon objectif est de montrer que le style d'un auteur ne peut se modéliser sans un engagement de l'analyste dans un sens. Pierre Achard l'a affirmé pour l'analyse de discours, et si je l'affirme ici c'est justement parce que le style, tout comme la logique, est une manière de s'engager dans la recherche d'une sorte de vérité soutenant nos croyances, vérité perçue par l'analyste comme donnant un sens à la succession de signes étudiés… Mais la communiquer nécessite à chaque fois de la réinventer par un nouvel acte discursif dans le respect du texte : c'est ce qu'exprime « s'engager dans un sens » ; elle implique une approche exploratoire serrée du texte en tant que trace signifiante porteuse de tous les engagements passés, ce qui en conditionne le style.

Mots-clés : analyse de discours, statistique textuelle, méthode « ALCESTE », approche exploratoire, style et logique, Lacan, répétition.

Abstract :

My starting point here will be the terms used by the organizers of the two-day conference on « style and how to model it » held at the University of Tours on December 10 and 11, 2009. Among the stipulations as to which topics the conference should cover was a series of questions concerning a possible computer-assisted identification of stylistic markers. For reasons of my own that have to do with a certain conception of style – reasons that will appear more clearly as I proceed with my answer – I was able to answer only the first two questions. I have, therefore, added a third part dealing with the computerized method of statistical discourse analysis known as « ALCESTE », in order to shed some light on the relationship between my conception of style and the exploratory approach to discourse that characterizes this software program. Having said this, my aim is to show that it is impossible to construct a model of an author's style without the analyst committing him- or herself to a particular interpretation of meaning. Pierre Achard claimed this was true of discourse analysis, and if I also hold it to be true of style analysis, it is

precisely because style, like logic, is a way of committing oneself to the search for a truth that sustains our beliefs. This truth is perceived by the analyst as giving meaning to the succession of signs under study, but in order for it to be communicated, it has to be reinvented each time by a new discourse act in harmony with the text: this is the idea expressed by the phrase « committing oneself to a particular interpretation of meaning ». Looking for this truth implies a rigorous exploratory approach to the text as a meaningful trace showing signs of all past commitments, upon which its style depends.

Key words: discourse analysis, textual statistics, « ALCESTE » software program, exploratory approach, style and logic, Lacan, repetition

Les deux premières parties sont introduites par les questions du cahier des charges, la troisième consiste en une brève introduction à la méthode Alceste en tant qu'outil d'exploration des textes (dont la pertinence avec l'analyse du style apparaîtra progressivement)

***A. Question 1** : Pour répondre aux besoins d'une analyse textuelle stylistique, qu'offrent les logiciels existants de lexico ou textométrie ?*

Je répondrai du point de vue de la méthode Alceste. Cette méthode d'analyse statistique de discours apporte une aide à la lecture d'un corpus, et me semble, à ce titre, utile pour l'approche exploratoire du style d'un auteur. Encore doit-on s'entendre sur cette notion. Je la préciserai d'abord à partir de ma lecture de l'*Ouverture des Ecrits* de Jacques Lacan (1966, p 9-10), avant de l'aborder d'un point de vue plus opérationnel.

a. Lacan, dès l'ouverture de ses *Ecrits*, déclare avec Buffon : « *le style est l'homme même...* » (p 9), l'homme *à qui l'on s'adresse...* c'est-à-dire, en tant qu'il est *sujet* (d'une demande inconsciente, d'un désir non reconnu, d'une histoire). Lacan termine cette ouverture (p. 10) par un second aphorisme : *« C'est l'objet qui répond à la question sur le style »*… dont la chute l'isolerait dans un signe à chaque fois renouvelé, signe qui le désigne cependant, mais simultanément, qui le rate dans sa vérité.

Aussi le *style*, si l'on suit Lacan, est ce qui accompagne le mouvement d'un *sujet,* sous l'empire d'une demande inconsciente, d'un désir qui ne lui appartient pas, venant d'un Autre (histoire, traditions, culture, religion, communauté, langue…), pour échoir en un objet qui lui échappe en définitive. Aussi le style semble caractériser pour Lacan à la fois un *sujet* qui se cherche et un *objet* qui se dérobe. Il caractérise surtout ce parcours de leur rencontre ratée. Mais si l'objet s'esquive sans cesse, il choit sous forme de signes renouvelés, car le signe n'est pas l'objet, il n'est que la lueur évanescente d'une vérité voilée le concernant :

> *« C'est l'objet qui répond à la question sur le style, que nous posons d'entrée de jeu.*
>
> *A cette place que marquait l'homme pour Buffon, nous appelons la chute de cet*
>
> *objet, révélante de ce qu'elle l'isole, à la fois comme la cause du désir où le sujet*

s'éclipse, et comme soutenant le sujet entre vérité et savoir. Nous voulons du parcours dont ces écrits sont les jalons et du style que leur adresse commande, amener le lecteur à une conséquence où il lui faille mettre du sien. » (Ecrits, p10)

Donc, si nous entendons Lacan, le style est l'homme même, en tant qu'il est acteur, en tant qu'il est *sujet,* c'est-à-dire, en tant qu'il est parlé par un désir qu'il méconnaît, et qui l'aliène et le soutient dans son engagement ; Il le soutient comme *savoir*, agissant, se répétant, mais inconscient. Pourtant le point de fuite à l'horizon de son parcours dans les signes successifs (évoquant les sémioses de Peirce) implique, en fin de course, la perte de l'objet convoité, ou plus exactement sa métamorphose en *vérité d'un parcours*. Le style est au *sujet,* ce que la logique est au théorème. Après coup, par son expression même dans les signes, si le style du *sujet* rate *l'objet* (absent) de sa quête, cet *objet* sera cependant perpétué comme horizon de *vérité* pour un *sujet* ex-sistant. Se l'appliquant à lui-même, en tant que *sujet* de ses Ecrits, Lacan insiste sur le fait que son propre style lui est imposé par la sorte de logique qui le porte et ce qu'il vise comme vérité… Ce dont il ne peut rien dire d'autre que par ses Ecrits mêmes.

Par cette brève introduction, je désire suggérer seulement qu'il y a *style en acte* par tout ce que l'on ne peut expliciter de son objet de recherche. On est parlé par son style. Mais « on » n'est pas « je » et la trace d'un parcours singulier se confond également avec l'anonymat d'autres traces (Foucault, 1971), donnant leur style à une époque tout autant qu'à un auteur ou qu'à une œuvre particulière. Tout ce que l'on peut en dire reste en deçà de ce qui se joue : le style est l'expression même de ce qui ne peut se dire autrement, et qui vise une vérité du *sujet*. Elle se montre par le style, mais on ne peut la saisir. Selon la formule de Lacan, on ne peut que la « midire » (Séminaire XXI).

Aussi reconnaître un style nécessite d'abord de reconnaître dans le parcours des signes, ce qu'il actualise comme vérité… mais celle-ci ne peut être que sienne… *Aussi les marques d'un style ne deviennent sensibles qu'après une période d'appropriation de cet horizon de vérité voilée (l'obscur objet du désir) qui soutient le sujet.* Et cela dépend de deux activités différentes simultanées... Par exemple, pour ce texte étudié, de ce que Lacan a écrit, et aussi, de ce qu'un lecteur a dû parcourir pour que le style de Lacan n'apparaisse pas comme un simple artifice, mais acquiert la transparence d'une logique.

Une « marque de style » ne peut avoir de sens en dehors de ce parcours. Pour prendre l'exemple de Lacan, il est assez facile de repérer des « tournures de style » que certains de ses élèves ont reprises, et qui ne constituent aucunement, à eux seuls, le style « Lacan ». Prenons l'exemple de la dernière phrase de la citation précédente : « *Nous voulons du parcours dont ces écrits sont les jalons et du style que leur adresse commande, amener le lecteur à une conséquence où il lui faille mettre du sien. ».* Pour supprimer le trait de style à épingler, et qui semble consister en une simple transposition du complément, on pourrait la transformer ainsi : « Nous voulons amener le lecteur à une conséquence où il lui faille mettre du sien (à propos) *du parcours dont ces écrits sont les jalons et du style que leur adresse commande* ». Mais l'expression devient fade, car elle n'est plus compatible avec ce que l'auteur vise comme vérité, le parcours étant antérieur à la prise de conscience de ce qui l'a animé. Ce trait de style de Lacan permet d'appréhender cette antériorité, car justement il s'impose comme trace d'un acte s'originant dans un désir de vérité.

b. Plutôt que par ses figures, qui supposent déjà l'appropriation de formes a priori, et impliquent une modélisation de l'activité, mon hypothèse serait que le style est d'abord sensible à travers des indices concernant *la vérité d'un sujet* (auteur, analyste, lecteur). *C'est tout simplement dire que le style d'un texte ne peut s'appréhender avant de l'avoir lu et interprété.*

L'objet d'une analyse exploratoire type « Alceste » est justement de suggérer des contours à cette activité de lecture, qui permettent de rendre compte d'un mode de présence de l'analyste dans ce qu'il cherche ; par exemple, lorsqu'il interprète les mondes lexicaux comme des registres d'expression visant ce que l'on sent être la vérité d'un auteur… mais celle-ci ne peut que s'interpréter… et elle nécessite une participation de l'analyste, *un engagement* au sens de Achard (1997).

C'est à ce titre que je perçois l'apport de l'analyse « Alceste », comme analyse exploratoire, car l'analyste est le chercheur, et il explore le texte en prenant conscience de ce que peuvent être les registres d'une interprétation possible, à partir des marques explicites pour rendre compte de la sorte de vérité qu'il poursuit... L'apport de la méthode est donc simplement dans cette aide à la lecture, qui permet à l'analyste de repérer à travers les classes statistiques des traces dont la lisibilité dépend en définitive de l'engagement de l'analyste dans une interprétation.

Pratiquement la méthode « Alceste », en tant qu'approche exploratoire, ne permet qu'une aide à la lecture par la présentation statistique (et synthétique) des environnements textuels dans un corpus. Si l'analyste peut leur donner sens, il s'engage vis-à-vis de cette sorte de vérité recherchée par l'auteur. Cette aide à la lecture pourra éventuellement servir à l'orienter vers des figures de style en rapport avec cette vérité pressentie (même si en définitive, celle-ci lui échappe, et demandera de la part d'un lecteur une réinterprétation ou plutôt une réappropriation).

B. Question 2 : *Est-il possible d'automatiser le repérage des indices ? des marqueurs ? des stylèmes ? Quels seraient, le cas échéant, les indicateurs de définition de seuils, si cette notion est nécessaire ?*

Cette possibilité dépend pour le moins d'une réponse positive en (1)… Cela reste très problématique. Elle nécessiterait de plus une modélisation… de l'activité du chercheur, et plus précisément une modélisation de l'objet qu'il vise dans sa lecture, qui n'est pas à confondre avec l'œuvre qu'il étudie. *Il nous semble impossible de modéliser le style d'une œuvre, sans que le chercheur s'engage lui-même dans un sens,* car c'est en tant que passeur d'une vérité entre-aperçue par le lecteur, et prise en charge par lui, que le style devient lumineux, s'exhibe comme logique. D'où l'importance des analyses exploratoires, car en modélisant trop vite, on risque d'enfermer la complexité d'un style dans quelques stéréotypes. Selon mon point de vue, l'automatisation des indices, marqueurs et autres, est illusoire pour approcher ce qui dans le style est vérité vivante… Cela dit, il n'est sans doute pas inutile de s'intéresser à cerner ce qu'un style a parfois de stéréotypé … Mais peut-on alors l'appeler « style »?

C. La méthode « Alceste » comme outil d'exploration des textes

ALCESTE est une méthode informatisée d'analyse statistique de discours. Son principe est à la fois d'une grande simplicité pratique et complexe à justifier conceptuellement. C'est de cette expérience que j'essaie de rendre compte depuis bien longtemps. Il est exclu que je l'approfondisse ici, mais je peux au moins essayer d'en dessiner quelques contours. L'algorithme

consiste à simuler artificiellement la scansion d'une lecture ou d'une énonciation par un simple découpage du texte en unités de texte de « longueur » comparable, d'où l'intitulé de son sigle : *« Analyse des Lexèmes Cooccurrents dans un Ensemble de Segmentations du Texte Etudié »,* étant entendu que la pertinence d'un résultat obtenu à partir d'un découpage arbitraire pose un problème complexe. Ma démarche a été de m'assurer d'abord d'une stabilisation locale des résultats en proposant en standard une double analyse croisée de manière à ne retenir que des classes localement stabilisées. Il s'est avéré que les résultats obtenus par cette double analyse étaient généralement interprétables par les utilisateurs de la méthode. Cela pose le problème du statut de ces résultats dans l'élaboration d'une recherche. Enfin dans un second temps, j'ai pu montrer que la stabilisation des résultats était même plus générale, puisqu'on pouvait l'obtenir entre classes de corpus différents même segmentés selon des longueurs différentes d'unité textuelle. Selon mon hypothèse actuelle, cette stabilisation dépend de lois de productions fractales, assez souvent interprétables en terme de « postures », et saisissables sous forme de « mondes lexicaux stabilisés », qui peuvent être interprétés comme de véritables registres de discours (Reinert, 2003, 2007, 2008). La possibilité de les saisir statistiquement à travers des séquences textuelles d'empan très différent (de quelques lignes, à plusieurs pages, voire dizaine de pages), soulignant la présence de lois fractales dans la distribution du vocabulaire. Ces lois montrent également l'importance des rythmes dans la composition des énoncés.

Un autre fait d'expérience que l'on ne peut négliger est que cette méthode *Alceste* est assez souvent utilisée pour préparer une analyse de contenu, alors qu'elle recouvre une analyse automatique purement formelle, à partir d'un découpage séquencé du texte. Cela peut justifier que les résultats ne soient pas sans rapport avec ce que l'on retient d'une lecture, elle-même rythmée, dynamique. Par là cette méthode peut apporter une aide à l'interprète y reconnaissant les mots de son expérience de lecteur dans les résultats. Mais cela suppose a minima que le corpus textuel soumis à une analyse *« Alceste » constitue, pour cet analyste, un discours sur son objet d'étude*.

L'automatisme d'*Alceste* (paramétrable) est suggéré par son sigle. Quelle que soit la nature du texte étudié, il peut être découpé en segments (de quelques lignes à quelques pages selon la grandeur du corpus), et présenté sous la forme d'un tableau à double entrée entre mots pleins et segments de texte, tableau auquel on applique une analyse statistique particulière (Reinert, 1983). Les résultats se présentent sous forme de classes différentielles de séquences textuelles, représentables en termes de mondes lexicaux.

Par son automatisme même, l'approche *Alceste* ne peut être qu'exploratoire. Sans préjuger des objets intéressant le chercheur, elle offre un cadre indépendant des a priori conceptuels pour élaborer individuellement ou collectivement des interprétations sur les *lois de production du corpus étudié.*

En résumé, reprenons les trois opérations fondant cette méthode d'analyse :

a. *Le découpage du texte en petites unités de contexte.* On sait que *ce découpage relativement arbitraire du texte en unités* de " longueur " comparable (scansion) est un des points les plus controversés de la méthode (J. Jenny, BMS, n°57). Je le situe conceptuellement comme essentiel en ce sens que, sans aucun a priori sur le sens du texte étudié, ce découpage permet d'évaluer les effets d'une scansion quelconque sur l'obtention des mondes lexicaux… avec l'hypothèse forte qu'une stabilisation des

résultats est généralement soutenable.

Cette hypothèse rythmique de la distribution du vocabulaire (voire fractale) est confortée depuis 1990 par la conception d'une double analyse obtenue par variation systématique de la longueur des unités de contexte (*BMS* 1990) et, aujourd'hui, par la mise en évidence de « mondes lexicaux stabilisés » (JADT 2008).

b. *Le choix des mots pleins pour le calcul statistique des classes.* Ce choix, également controversé, n'a pu être remis en question sérieusement alors même que la frontière entre mots pleins et mots outils est très fluctuante (cas de certains adverbes, des locutions prépositionnelles, de certains figements, …). Cela dit ce critère imposerait logiquement une lemmatisation du texte ce qui a été réalisé dans les premières versions du logiciel. Mais l'expérience (et des débats amicaux notamment avec A. Salem) m'ont conduit à expérimenter sur ce sujet délicat, et il semble bien que la lemmatisation peut être omise en pratique… Cela complexifie la notion de « plénitude » qui reste largement problématique, et il n'est pas question de l'approfondir ici. Par contre, le retrait des mots outils du tableau soumis à la classification s'avère être une nécessité. Dans une analyse *Alceste*, c'est la scansion qui les remplace, et un bon indice de cela a été perçu par Pierre Achard, lorsqu'il a retrouvé dans les classes d'une analyse sur les récits de cauchemars, les postures qu'il avait élaborées à partir de son *analyse de discours* des récits de guerre d'anciens appelés en Algérie (Achard, 1991 ; Reinert, 1993 & 2007).

Après ces deux premières opérations, le corpus, en tant qu'il est « lu » comme discours, est représenté par *un tableau de données binaire croisant les « unités de contexte » construites par découpage du corpus (scansion) avec le vocabulaire (l'ensemble des mots pleins de fréquence suffisante).* Du fait de ce découpage, les mots apparaissant dans une même unité de contexte sont susceptibles d'avoir été associés aux aléas statistiques près au même *acte de production* d'un *contenu-énoncé-rythmiquement.*

c. *L'utilisation d'une classification statistique multidimensionnelle de type descendant, pour analyser le tableau de données.* Rappelons que la *classification descendante hiérarchique* mise en œuvre dans l'algorithme d'*ALCESTE* est une méthode d'analyse statistique apparentée à *l'analyse factorielle des correspondances* de J.P. Benzécri dont elle utilise l'algorithme à chaque pas (Reinert, 1983). *Dans les deux cas, l'aspect descendant de la méthode implique que l'analyse porte d'abord sur la forme globale de l'ensemble des distributions plutôt que sur une analyse précise des liens deux à deux (entre deux unités de contexte, ou entre deux mots).* Cette forme globale des répétitions est perceptible statistiquement à travers les notions de *valeurs propres* du tableau de covariance et *de facteurs propres* (ici pour la métrique du chi2, Benzécri, 1973).

Dès les premières *analyses factorielles des correspondances* que Benzécri a introduites au cours des années 1960-1970, les propriétés synthétisantes des premiers facteurs ont été mises en évidence (voir par exemple, l'analyse du graphe de la carte des départements français proposée par L. Lebart in Benzécri & col, « l'analyse des données », tome 2, 1973, p244- 252).

Je considère, pour ma part, la *classification descendante hiérarchique* comme une simple adaptation de *l'analyse factorielle des correspondances* pour le traitement des grands tableaux binaires clairsemés (jusqu'à 40 000 « unités de contexte » par 3000 « mots pleins »)… Elle détermine cependant un changement de point de vue sur la signification d'un facteur. Notamment l'ordre des facteurs est supposé plus stable que le *pourcentage d'inertie extraite*, et ce pourcentage en lui-même n'a d'ailleurs plus de sens du fait de l'absence d'un espace stable de référence.

Du point de vue méthodologique, cette stabilisation des facteurs (à la base également des classifications), *par leur aspect différentiel, est à mettre en rapport avec ce qui se répète rythmiquement dans une énonciation, dont on a dit qu'elle était assez souvent exprimable en termes de changements rythmiques de postures dans la production d'un discours.* Ainsi l'analyse statistique peut les approcher du fait du découpage rythmique du texte. Les *mondes lexicaux* (Reinert, 1993) en sont alors la trace. Cela permet de comprendre qu'une analyse des mots pleins (orientés vers l'objet convoité d'une énonciation) puisse conduire à la différenciation de postures en rapport avec des sujets de l'énonciation en recherche d'issue pour cette sorte de vérité entrevue comme fil conducteur au cours un parcours énonciatif au lieu d'une saisie de l'objet.

En conclusion, si le style épouse ce parcours d'un sujet ratant son objet (« objet petit a » pour Lacan), on peut comprendre en quoi l'approche des mondes lexicaux d'une œuvre peut permettre de travailler sur les traces lexicales de ce désir d'objet (perçu en termes de contenus évanescents à travers la cooccurrence des mots pleins) mêlant l'unité de la quête d'un *sujet* à la multiplicité des postures en métamorphosant sans cesse les apparences — telle Daphné dans les métamorphoses d'Ovide —, ce qui implique des cassures distributionnelles rythmiques dans la composition d'un texte.

Quelques références bibliographiques

Achard Pierre (1997) « L'engagement de l'analyste à l'épreuve d'un événement », *Langage & Société,* n°79, Maison des Sciences de l'Homme, Paris, 5-38.
Achard Pierre (1991) « Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie *», Langage & Société,* 1991, n°55 :5-40.
Benzécri Jean-Paul & Coll. (1973) *L'analyse des Données*, Tomes 1&2, Dunod.
Freud Sigmund (trad. par Jankélévitch), Au-delà du principe du plaisir, *in Essais de Psychanalyse,* petite bibliothèque Payot, 1977
Foucault Michel (1971) L'ordre du discours, Paris : Gallimard.
Lacan Jacques (1966) *Ecrits*, Seuil
Peirce Charles Sanders (trad. G. Deledalle) *Écrits sur le signe*, Seuil, Paris, 1978
Reinert Max (1983) Une méthode de classification descendante hiérarchique. *Cahiers de l'Analyse des Données* 3 : 187-198.
Reinert Max (1990) *ALCESTE*, une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval, *Bulletin de Méthodologie Sociologique* 26, 24-54.
Reinert Max (1993) Les 'mondes lexicaux' et leur 'logique' à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société* 66, 5-39.
Reinert Max (1997) Les 'Mondes lexicaux' des six numéros de la revue 'Le Surréalisme au Service de la Révolution', in *Mélusine, XVI* :270-302, Editions L'Age d'Homme, Lausanne.
Reinert Max (2003) " Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode *Alceste ", Semiotica,* 147, 389-420 (accessible WEB)
Reinert Max (2007) «Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours», *Langage & Société,* n° 121-122 : 189-202. (accessible WEB)
Reinert Max (2008) «Mondes lexicaux stabilisés en analyse statistique de discours», *JADT 2008,* Lyon (accessible WEB)

Conclusion

Les journées d'étude organisées à l'Université François-Rabelais de Tours et les contributions réunies dans ce recueil de textes témoignent de l'intérêt de chacun des participants et des contributeurs pour la création et l'animation de lieux de réflexion et de collaboration scientifiques. Plus encore, certains objets et certaines problématiques de recherche ne semblent pouvoir être construits que dans l'interdisciplinarité et leur étude requiert un espace fédérateur au sein duquel des spécialités différentes mais complémentaires sont susceptibles tout à la fois de constituer richesse et obstacle de la recherche.

L'investissement de chacun des participants justifie également de l'actualité de l'une des visées de ce projet dont l'une des dimensions applicatives concerne, au-delà de la valorisation de notre patrimoine littéraire, le discours scientifique, voire le discours universitaire, qui fait l'objet d'appropriations nouvelles parfois problématiques avec la circulation accrue des écritures de recherche.

Les réponses apportées par les contributeurs de cette publication insistent sur certains points forts :

La nécessité de penser l'informatisation et le style dans un cadre théorique et une méthodologie spécifiques est apparue à l'occasion de ces rencontres : la linguistique de corpus et un traitement contrastif des différents paliers de la matérialité linguistique permettent une nouvelle lecture des textes, appréhendés au plus près de leurs configurations langagières. Ce préalable est une condition fondamentale d'accès à la textualité, une condition de construction des marqueurs, une condition de saisie relationnelle de variables éparses actualisées par le corpus et renouvelées au cours du traitement par ses restructurations plurielles. Texte et textualité ne disposent pas d'autonomie *a priori* mais d'indices de lecture, virtuels et identifiés dans la pertinence des réseaux philologiques naissant du corpus.

Les chercheurs réunis ici s'accordent à penser que la tentation de « tout » attendre de la machine est illusoire : l'automatisation ne peut concerner que certains marqueurs textuels, et qu'une série d'étapes de l'analyse textuelle. Il s'agira donc de travailler au perfectionnement, non pas d'une analyse textuelle informatisée, mais d'une analyse textuelle assistée par l'ordinateur.

Tout d'abord, un travail de préparation des corpus en amont reste indispensable, malgré le coût élevé de sa réalisation. Chaque préparation doit correspondre à une attente particulière de l'analyste. Par exemple, le choix de la lemmatisation ou non du texte dépend étroitement des résultats visés et du type d'analyse attendue. Le choix même du corpus peut aussi déterminer l'issue de l'analyse. Il est donc lié à des interrogations préliminaires. La sélection d'un corpus témoin peut aussi nettement influencer les résultats et remettre en question la légitimité ou *a contrario* renforcer l'analyse.

Si l'intervention manuelle de l'analyste est indispensable en amont, elle l'est aussi en aval. Une fois obtenues les données chiffrées, livrées par l'ordinateur, le travail d'interprétation échoie, non plus à la machine, mais à l'homme. C'est alors au chercheur, et non pas à l'outil utilisé, qu'incombe la responsabilité de l'analyse finale. En conséquence, entre la phase de préparation

évoquée plus haut et celle de l'interprétation, l'outil informatique livre des données dont la nature et le sens sont doublement déterminés par l'intervention du chercheur.

Une autre réserve s'exprime dans les réponses apportées par les contributeurs. Elle consiste à considérer que certains marqueurs textuels ne sont pas totalement automatisables et qu'ils nécessitent non seulement le travail de préparation et d'interprétation, mais qu'ils exigent, au cours même du traitement informatique, des interventions successives sous forme de corrections et d'ajustements progressifs. L'automatisation n'est pas une disponibilité logicielle offerte à l'analyse textuelle, elle est construite par le traitement informatique dont les résultats sont susceptibles à leur tour d'être redéfinis en marqueurs en fonction des propriétés du corpus. La notion de *seuil de pertinence* est apparue nécessaire à la sélection de ces indices virtuels : à partir de quelle fréquence ou de quelle spécificité, certains marqueurs linguistiques deviennent-ils indices ? Cette notion s'est en outre montrée indissociable d'une autre, celle de *seuil de cohérence* : à partir de combien de marqueurs corrélés pour la création d'un procédé textuel ou stylème, une forme linguistique devient-elle indice ?

Autre point fort en jeu dans cette réflexion, celui du référentiel textuel. Conçu comme outil de travail pour établir et construire une véritable réflexion interdisciplinaire, il a atteint son objectif : ouvrir un espace de négociation entre les besoins de l'analyse stylistique et les fonctionnalités logicielles. Son utilité est unanimement reconnue, à condition tout de même qu'il se présente sous une forme suffisamment souple pour être modulé en fonction, d'une part, des types de corpus et, d'autre part, des types d'analyses attendus. Sa structuration hiérarchique – qui rompt avec toute conception du style comme nomenclature de formes et/ou de procédés – lui confère une valeur heuristique compatible avec la notion de *seuil* et susceptible de palier certaines limites informatiques actuelles. Le traitement informatique en fonction de l'état des fonctionnalités disponibles peut déterminer à quel niveau de la hiérarchie textuelle se situer. Plus la tentative d'automatisation porte sur des marqueurs (macro-marqueurs) élevés dans la hiérarchie, plus elle parviendra à corréler des indices de fréquence plus ou moins élevée dont la valeur est construite dans la cooccurrence ou la cohérence textuelle.

La voie reste donc ouverte pour tester la pertinence des marqueurs proposés et la possibilité de leur automatisation, partielle ou totale. L'intérêt des littéraires, des linguistes et des informaticiens se conjugue dans une même volonté d'ajouter de nouvelles fonctionnalités aux outils actuellement existants.