ÉTUDE QUANTITATIVE DES CHANGEMENTS ESTHÉTIQUES ET DES VARIATIONS GÉNÉRIQUES CHEZ TROIS GRANDS ÉCRIVAINS : ANALYSE LEXICOMÉTRIQUE D'UN CORPUS LITTÉRAIRE

Margareta KASTBERG-SJÖBLOM
ILF-CNRS, BCL (UMR 6039), Université de Nice

SOMMAIRE

La notion de genre reste encore aujourd'hui l'institution première du code littéraire, bien qu'elle ait souvent été discutée et remise en question. Les théoriciens la considèrent avec réserve, affirmant que chaque genre littéraire en englobe plusieurs, et les hésitations terminologiques manifestent ce caractère "d'appartenance multiple et emboîtante" de tout écrit littéraire. En effet, la codification des genres n'est pas chose aisée ni stabilisée. Le système traditionnel nous propose – ou nous impose – selon le code générique institutionnel, certaines classifications reconnues : romans, nouvelles, essais, etc. Mais, cette distinction des genres transmise par la critique littéraire "traditionnelle" est-elle réellement pertinente ?
Pour répondre à cette question essentielle, nous avons choisi de recourir à l'outil informatique et aux méthodes de la linguistique quantitative qui montrent bien que les genres existent, qu'on le veuille ou non, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence de différentes typologies de textes. L'analyse lexicométrique valide cette idée, l'opposition générique est extrêmement claire et permet de définir des caractéristiques génériques en s'appuyant, non sur des valeurs anthropologiques ou sociales, mais sur les propriétés mêmes des textes.
Le présent exposé propose d'étudier les variations et les oppositions génériques chez plusieurs grands écrivains français : Julien Gracq, Gustave Flaubert et J.M.G. Le Clézio, et en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. L'œuvre de ces écrivains présente une riche variété de textes et se décline en différents genres, ayant des caractéristiques typologiques bien distinctes. Bien que les auteurs évoquent souvent une "écriture unique", déclarent n'appartenir à aucun groupe et tentent même de transgresser un système social établi, les différentes typologies de textes existent et ces variations sont à observer à tous les niveaux.
L'analyse du corpus en situation permet en effet d'abord de caractériser la structure du vocabulaire. Le rythme de la narration est ensuite corrélé à l'analyse de la longueur des mots et des phrases. Enfin, l'étude de la distance lexicale met en évidence des aspects sémantiques et thématiques révélateurs et permet aussi de constater d'importantes variations typologiques. Ces différentes analyses mettent donc bien en exergue l'opposition entre les différents genres, toujours présente et souvent même prépondérante dans toutes les différentes analyses statistiques.

1. Le corpus

Plusieurs écrivains français ont mis en question ou refusent même le cloisonnement en genres, parlant d'une seule et unique écriture. Parmi ces auteurs certains ont une large production qui se décline en plusieurs genres littéraires. Nous nous intéresserons ici à trois d'entre eux : Julien Gracq, Jean-Marie Le Clézio et Gustave Flaubert.
Un point commun entre ces trois écrivains très productifs est la difficulté de classer leurs ouvrages dans des genres littéraires traditionnels. Notamment la critique gracquienne évite parfois complètement de prendre position ou d'effectuer une classification générique quelconque en qualifiant tous les textes de palimpsestes.
Notre corpus Gracq englobe pratiquement la totalité de sa production avec 17 ouvrages et rassemble plusieurs genres littéraires, notamment les essais littéraires qui sont richement représentés dans ce corpus :

Romans : *Au château d'Argol*, *Un beau ténébreux*, *Le rivage des Syrtes*, *Un balcon en forêt* et *La presqu'île*.
Critiques, Essais ou mélanges[1] : *André Breton. Quelques aspects de l'écrivain*, *Préférences*, *Lettrines 1*, *Lettrines 2*, *Les eaux étroites, En lisant, en écrivant*, *La forme d'une ville*, *Autour des sept collines* et *Carnets du grand chemin*.
Poèmes en prose : *Liberté grande*.
Théâtre : *Le Roi-pêcheur* et *Penthésilée*.

La production littéraire de Le Clézio est vaste, s'étend sur plus de quarante ans et englobe plusieurs genres littéraires. Le corpus est constitué de 31 livres ; tout d'abord des six premières œuvres, classées, par leur style particulier et innovant, comme appartenant à l'École du "nouveau roman" : *Le procès-verbal*, *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre* et *Voyages de l'autre côté*. Les neuf romans qui suivent cette période, considérés par les critiques comme plus "traditionnels", sont les suivants : *Désert*, *Le chercheur d'or*, *Voyage à Rodrigues* (écrit sous forme de journal personnel), *Angoli Mala*, *Onitsha*, *Etoile errante*, *La quarantaine*, *Poisson d'or* et *Hasard*. *Mydriase* et *Vers les icebergs* sont difficiles à classer dans un genre précis, ce sont plutôt des récits poétiques. Le corpus inclut ensuite les recueils de nouvelles : *Mondo et autres histoires*, *La ronde et autres faits divers* et *Printemps et autres saisons*. Les essais littéraires sont de différentes époques. *L'extase matérielle* et *L'inconnu sur la terre* traitent de thèmes généraux tandis que *Trois villes saintes* et *Le rêve mexicain ou la pensée interrompue* s'intéressent exclusivement à la culture amérindienne. Celle-ci constitue également le principal intérêt des ouvrages à vocation ethnologique, *Les prophéties du Chilam Balam* et *La fête chantée*, tandis que *Sirandanes* s'intéresse à la culture de l'île Maurice. Sont inclus en outre dans le corpus deux livres pour enfants : *Voyage au pays des arbres* et *Pawana* ; la seule biographie *Diego et Frida*, et le récit de voyage *Gens des nuages*.
La production de Gustave Flaubert s'étend sur la moitié du XIXème siècle et ce corpus inclut 15 livres ; des romans comme *Madame Bovary*, *L'éducation sentimentale*, *Salammbô* ou *Bouvard et Pécuchet*. Nous y trouvons aussi *Les trois contes*, le dialogue poétique et philosophique de *La tentation de Saint Antoine* dans ses trois versions, le récit de voyage *Par les champs et par les grèves,* les *Mémoires* et les *Souvenirs* ainsi qu'une partie de la *Correspondance*.
Ce grand corpus a été numérisé et traité par le logiciel *Hyperbase*, version 6.0. et il contient 4.121.141 occurrences et 79.833 formes (dans la version qui s'appuie sur les formes graphiques) réparties sur les soixante-trois œuvres du corpus.
Le traitement lexicostatistique automatisé permet un certain nombre d'analyses qui ouvrent la voie à des interprétations et à des études différentes de ce corpus, basées sur des données impartiales, et non sur des critères subjectifs.
C'est en premier lieu à travers une étude sur la structure lexicale du corpus que nous pouvons observer l'influence de la riche variation typologique des textes.

2. La structure lexicale

Les différentes recherches sur la structure lexicale offrent la possibilité, indépendamment du contenu lexical, de situer, de distinguer et de comprendre la structure formelle des textes afin de pouvoir comparer différents discours, genres, époques ou auteurs, au niveau exogène aussi bien qu'au niveau endogène, ainsi que les parties de l'œuvre d'un écrivain ou de tout autre producteur de texte ou de parole. Ces recherches, qui au fond sont très proches de la lexicométrie traditionnelle, permettent aussi d'étudier l'évolution dans le temps.
Les calculs effectués par le logiciel *Hyperbase*, utilisé dans cette étude, permettent de mesurer l'étendue des textes dans le corpus en prenant en compte des contraintes statistiques. Les calculs du poids relatif, c'est-à-dire l'espérance mathématique de l'événement : occurrence d'un mot dans le texte considéré (P) et non-occurrence de ce mot dans le même texte (Q=1-P), permettent l'emploi des lois classiques de la lexicométrie, principalement la loi normale et la loi binomiale (Muller 1977 : 159-169), et elles servent aux calculs de pondération dans les différents traitements statistiques.
L'étude de l'accroissement lexical détermine l'apport du vocabulaire au fil du temps ; cet accroissement est, pour un segment déterminé du texte, le nombre d'unités nouvelles, c'est-à-dire n'ayant pas été employées antérieurement, qui apparaissent dans ce segment. Pour effectuer

[1] Gracq donne souvent la dénomination de "fragments" à ses ouvrages.

cette mesure, on découpe le corpus en tranches. La représentation graphique ci-dessous rend compte de l'accroissement du vocabulaire dans l'ordre chronologique.

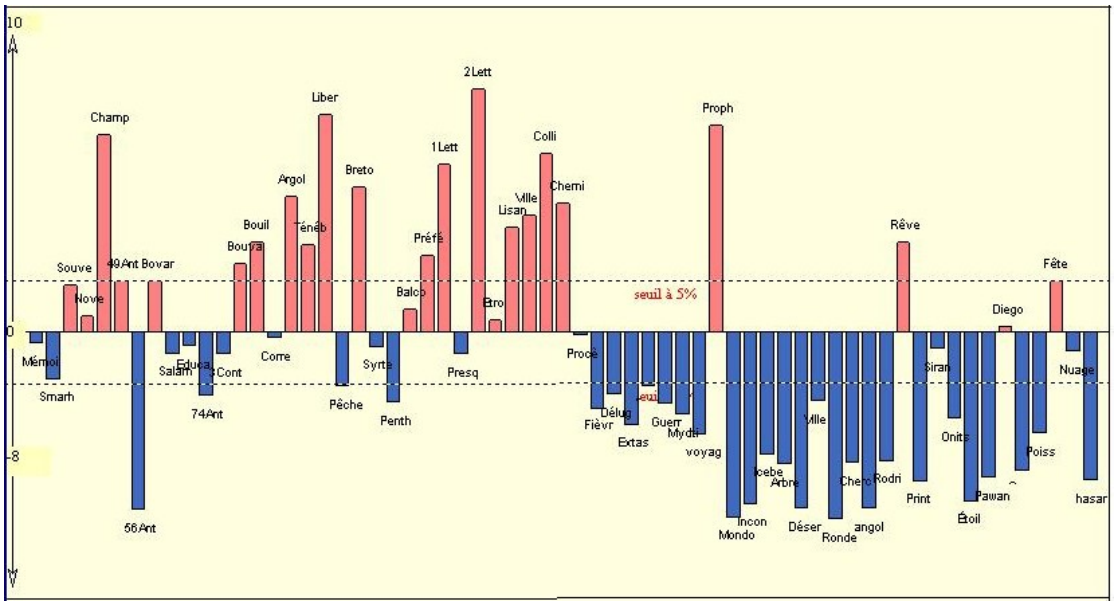

Figure n°1 : Accroissement lexical du corpus

Le graphique qui, de gauche à droite, s'oriente selon la chronologie, avec le corpus Flaubert suivi par celui de Gracq et de Le Clézio[1] nous permet de constater que les écarts autour de la moyenne, l'axe horizontal, sont de très grande ampleur, avec des ruptures et des reprises. Le seuil à 5 % est dépassé de nombreuses fois, avec des "pics" importants, dans le sens positif aussi bien que négatif.

L'étude de l'accroissement fait en effet très clairement apparaître l'opposition générique très importante du corpus. Dans la première partie, chez Flaubert, le vocabulaire ne croît de façon significative qu'une seule fois avec l'unique récit de voyage *Par les champs et par les grèves.*

Parmi les ouvrages de Gracq, notons que le récit *Liberté grande* et les essais comme *Lettrines I et II* ainsi que *Autour des sept collines* et *Carnets du grand chemin*, introduisent régulièrement de nouveaux thèmes dans le corpus. Mais le plus frappant est peut-être l'extraordinaire impact de l'apport lexical qui advient avec l'introduction du monde amérindien dans le corpus, ici avec l'essai littéraire *Le rêve mexicain* de Le Clézio. Notons aussi l'introduction d'un nouveau genre à ce corpus ; celui de l'ouvrage ethnologique qui fait appel à un apport lexical massif dans *Les prophéties de Chilam Balam*.

Nous pouvons en effet constater que dans ce corpus l'opposition des genres est plus importante que celle des trois auteurs. Il n'y a pas de limite nette entre l'œuvre de Flaubert et celle de Gracq. En revanche, la partie qui couvre les ouvrages de Gracq est nettement plus riche en apport de vocabulaire que celle de Le Clézio.

Par ailleurs, l'étude de la richesse lexicale[2] a montré que Gracq semble avoir un usage plus riche du vocabulaire tandis que Le Clézio s'exprime en règle générale avec un vocabulaire plus restreint. C'est dans la partie gracquienne que nous constatons des valeurs excédentaires et plus précisément dans la dernière partie et dans les essais, les pièces de théâtre étant forcément pauvres, qui témoignent d'un vocabulaire dont la richesse augmente vers la fin de l'œuvre.

3. Le rythme du récit

La ponctuation est essentiellement d'ordre syntaxique et doit être comprise comme l'écrit Nina Catach[3] :

[1] Il convient, avant d'interpréter cet histogramme, de souligner le fait que ce corpus n'est pas chronologique, et en postposant Le Clézio à Gracq cette étude désavantage évidemment le dernier auteur.
[2] Cf. M. Kastberg Sjöblom (2006) p. 50-57.
[3] N. Catach (1994) p. 48.

> "… associant à la fois : une suite de mots (aspect constructif), un message (aspect actuel), une substance et une forme intonatives (mélodie expressive et aspect intonatif) et un sens (contenu du message, résultant de l'ensemble des données précédentes). La ponctuation tient dignement son rôle à ces différents niveaux de la syntaxe."

Dans le langage écrit, la fin de la phrase est généralement représentée par le point, les points de suspension, le point d'interrogation, le point d'exclamation.
L'emploi des signes de ponctuation varie selon le genre de discours, l'époque et l'auteur. La ponctuation peut également connaître de grandes variations chez un même écrivain. L'étude de la distribution des signes de ponctuation permet de définir la longueur de la phrase et ainsi le rythme du récit. Le logiciel *Hyperbase* permet de recenser les différents signes de ponctuation de manière aisée et fiable et permet ainsi de connaître la longueur de la phrase. La longueur de la phrase et les caractéristiques de ses segments dépendent étroitement de sa complexité. Le recensement des signes de ponctuation faible permet de connaître la segmentation interne de la phrase et d'étudier sa complexité. Toutefois, dans cette étude nous nous bornons au recensement des signes de ponctuation forte. Cette distribution n'est pas régulière à travers le corpus, les bâtons excédentaires en signes de ponctuation révèlent donc une phrase courte et brève tandis que les bâtons bleus, déficitaires témoignent des phrases longues.

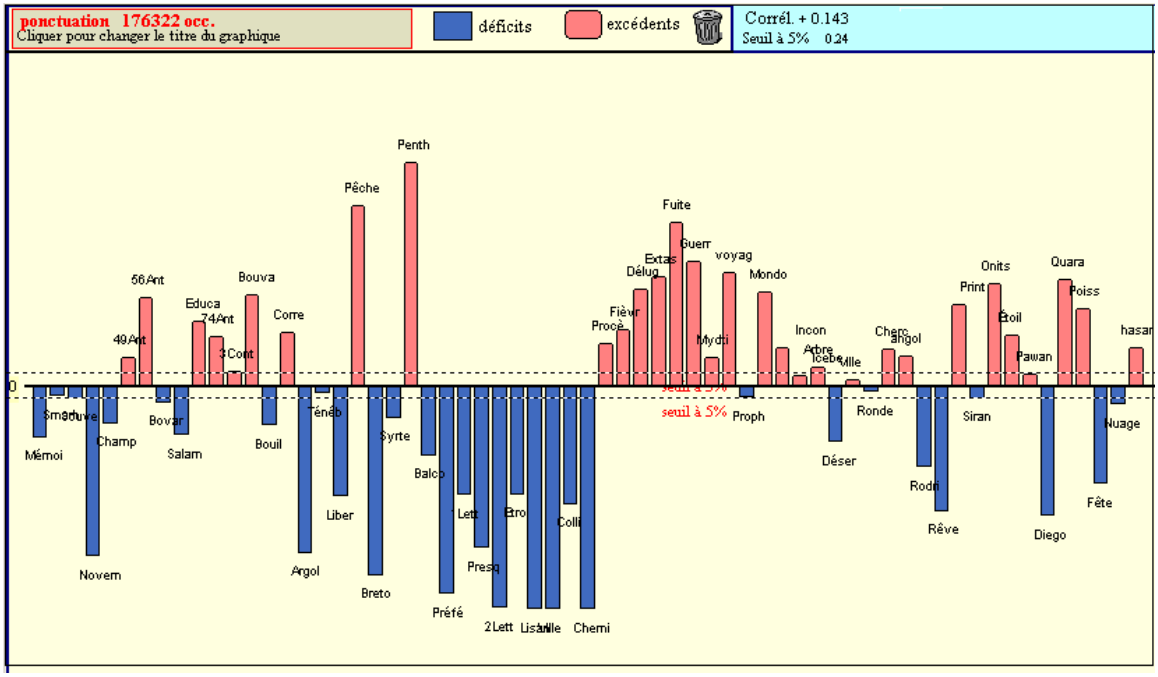

Figure n° 2 : Distribution relative des quatre signes de ponctuation forte dans le corpus

L'histogramme distingue bien les différences entre les auteurs : Flaubert semble abandonner une phrase longue et complexe pour des phrases plus courtes, tandis que le récit gracquien est plus généralement caractérisé par une phrase longue. Quant à l'écriture leclézienne on distingue les trois parties caractéristiques de son œuvre, si souvent commentées par les critiques.
Néanmoins, c'est peut-être le facteur du genre qui est le plus révélateur de cet histogramme. Le début romanesque de Flaubert est caractérisé par des phrases longues, tandis que des ouvrages comme *La tentation de Saint Antoine* font appel à une phrase plus courte, tout comme les romans tardifs et la correspondance. La phrase de Gracq est longue, à l'exception des pièces de théâtre *Le Roi-pêcheur* et *Penthésilée*. Avec Le Clézio on introduit à ce corpus le genre particulier du nouveau roman qui est caractérisé par une phrase courte et brève. Lorsque Le Clézio abandonne ce genre pour un style romanesque plus "traditionnel" la phrase reste généralement courte. C'est d'ailleurs une des caractéristiques du style leclézien, apprécié intuitivement par les lecteurs. C'est dans les ouvrages ethnologiques, les essais littéraires et la biographie *Diego et Frida* que nous trouvons des phrases longues correspondant à un style bien plus intellectuel.
En effet, nous avons pu observer, dans les diverses analyses sur le rythme de la phrase que Le Clézio exploite différentes techniques et différents styles d'expression selon les époques et selon

les différents genres littéraires et que les résultats qui en découlent manifestent des variations non négligeables.
Le facteur prédominant de ces divergences semble en effet être celui du genre. Les mots courts dominent l'œuvre romanesque tandis que les mots longs se trouvent dans les ouvrages ethnologiques et dans les essais. C'est également à l'intérieur de ce genre que nous trouvons les phrases les plus longues qui toutefois ne manifestent pas de complexité particulière, préférant la coordination à la subordination[1].
Les analyses que nous avons effectuées jusqu'à présent ont en commun de ne pas considérer le mot en soi, mais des liens statistiques qui donnent à voir des réseaux signifiants indépendants de l'interprétation du contenu. Dorénavant, nous nous intéresserons au contenu du discours qui implique la signification des mots et les différentes catégories lexicales.

4. La distance lexicale

L'étude de la distance lexicale permet de comparer différentes œuvres par le vocabulaire qu'elles partagent et celui qui les sépare. Il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de repérer ceux qui partagent des thèmes semblables.
L'analyse arborée ci-dessous tient compte de la distance lexicale en s'appuyant sur les occurrences (le calcul N)[2].

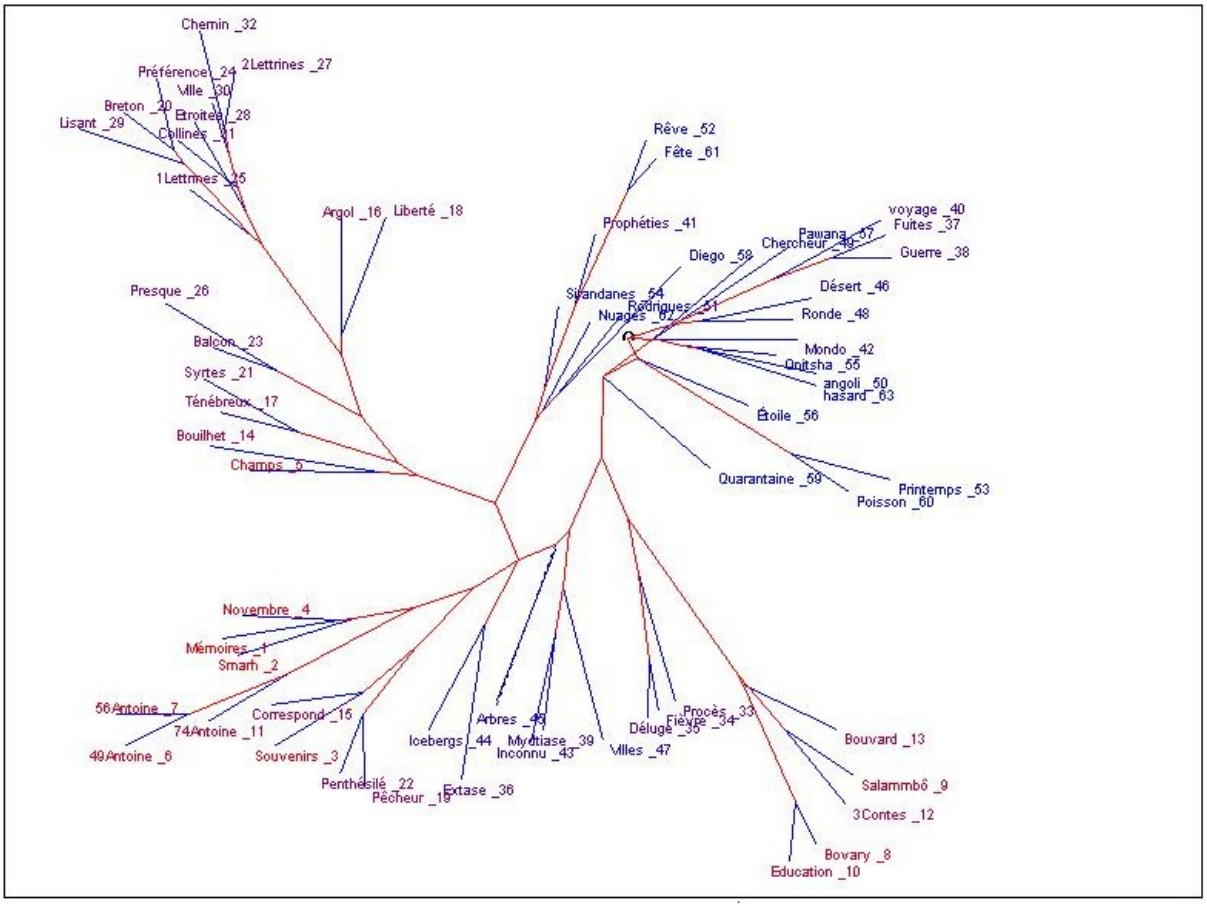

Figure n°3 : Analyse arborée de la distance lexicale s'appuyant sur les formes graphiques prenant en considération la fréquence (N)

L'arbre nous permet ici de constater premièrement la division entre les trois œuvres. L'œuvre gracquienne se trouve dans la partie supérieure gauche du graphique, l'œuvre de Flaubert dans la partie inférieure, tandis que l'œuvre leclézienne est opposée aux deux autres dans la partie supérieure droite du graphique.
En effet, il s'agit de trois écrivains très différents et ce graphique rend bien compte de ce fait. Il s'agit aussi d'époques différentes ce qui influence évidemment cette analyse.

[1] Cf. M. Kastberg Sjöblom (2006) p. 108-127.
[2] E. Brunet (2006) p. 52.

Néanmoins, l'opposition des genres reste un facteur très important. À l'intérieur de chaque œuvre, nous pouvons bien distinguer chez Gracq les essais des romans. Quant aux pièces de théâtre *Le Roi-pêcheur* et *Penthésilée*, la différence de genre les pousse dans le camp Flaubert, tout en bas de l'arbre. Les livres de Flaubert se divisent aussi selon le genre. Le graphique met en relief les ouvrages qui partagent soit le même thème comme les trois versions de la *Tentation de Saint Antoine*, soit le même genre comme les romans *Madame Bovary*, *L'éducation sentimentale* et *Salammbô*. C'est aussi le lien du genre qui rapproche l'écriture personnelle des *Souvenirs* et de la *Correspondance*.
La structure lexicale de l'écriture leclézienne est également déterminée par le genre. Les ouvrages ethnologiques comme *La fête enchantée* et *Le rêve mexicain* ou la biographie ou les essais littéraires tardifs sont bien séparés de l'œuvre romanesque regroupée à droite du graphique. Il est toutefois intéressant de pouvoir constater que certains ouvrages de Le Clézio transgressent aussi la "frontière leclézienne" pour se trouver au milieu des ouvrages de Flaubert. Il s'agit d'un côté des essais poétiques (*L'inconnu sur la terre* et *Mydriase*) et de l'autre côté des essais littéraires écrits avant la rencontre avec le monde amérindien comme *Parmi les icebergs* et *L'extase matérielle.*
Nous trouvons ici également bien regroupés les premiers livres de Le Clézio appartenant à l'école du "nouveau roman" : *Le procès-verbal*, *La fièvre* et *Le déluge*. Ce style particulier d'écriture semble, selon cette analyse, beaucoup plus proche du style de Flaubert que celui qu'emploie Le Clézio dans ses autres ouvrages.

Conclusion

Ainsi, la numérisation et l'analyse lexicométrique de la quasi totalité des textes de trois grands écrivains français nous ont permis de mettre en exergue non seulement les changements esthétiques, mais surtout l'importance de l'opposition générique qui s'observe à tous les niveaux de l'écriture : dans la structure, dans la morphologie, dans la syntaxe aussi bien que dans le vocabulaire.
Bien que ces auteurs, et les époques soient très différents et que l'on pourrait porter à cette étude une critique sur leur comparabilité, la division générique ressort de l'analyse comme le facteur prépondérant.
Cette analyse souligne en effet la fragilité des études qui s'appuient sur la distance intertextuelle pour résoudre des problèmes d'attribution ou de datation de textes. Ces études se révèlent parfois être totalement biaisées justement par la division générique ou typologique.
Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans les différents textes qui forment le corpus relativement hétéroclite de ces trois auteurs.

BIBLIOGRAPHIE

ADAM, J.-M., GRIZE, J.-B., & BOUACHA, M. A. 2004. *Textes et discours : catégories pour l'analyse*, Dijon, PU Dijon, Collection Langages EUD.
ADAM, J.-M. 2005. *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*, Paris, Armand Colin.
BRUNET, E. 1988. *Le vocabulaire de Victor Hugo*, Paris-Genève, Champion-Slatkine.
BRUNET, E. 2003. Flaubert traité par Hyperbase, Rouen, Revue *Flaubert* n° 3.
BRUNET, E. 2006. *Hyperbase, Manuel de référence*, Nice, CNRS.
CATACH, N. 1994. *La ponctuation*, Paris, PUF.
KASTBERG SJÖBLOM, M. 2006. *L'écriture de J.M.G. Le Clézio – Des mots aux thèmes*, Paris, Honoré Champion, Collection "Lettres Numériques".
KASTBERG SJÖBLOM, M. 2004. Comment l'ordinateur peut-il servir dans l'étude stylistique d'un texte littéraire et de quelle façon l'analyse de la distribution des parties du discours peut-elle contribuer à la compréhension des textes ?, in M. Ballabriga & F.-Ch. Gaudard (éds.), *Champs du Signe*, Toulouse, Editions Universitaires du Sud, pp. 119 -152.
MALRIEU, D. & RASTIER, F. 2001. Genres et variations morphosyntaxiques, in A. Martin Municio (éd.), *Actas del segundo seminario de la escuela interlatina de altos estudios en linguística aplicada, Matemáticas y tratamiento de corpus*, San Millán de la Cogolla, 19-23 septiembre de 2000, , Logroño, Fundación San Millán de la Cogolla.
RASTIER, F. 2001. *Arts et Sciences du texte*, Paris, PUF.