Peut-on refuser les genres littéraires ? Etude quantitative d'un corpus informatisé

Kastberg Sjöblom, Margareta

CNRS-ILF, Bases, Corpus et Langage (UMR 6039), Université de Nice, France

Littérature et linguistique

La notion a souvent été discutée et remise en question. Pourtant les études ont montré 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 l'écrivain contemporain J.M.G. Le Clézio, en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives.

L'analyse du corpus en situation montre en effet que la structure lexicale, la morphosyntaxe et la sémantique varient avec les genres. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques.

Introduction

La notion de genre, et notamment de genre littéraire, reste aujourd'hui encore 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 : la nouvelle peut se présenter en effet sous forme de fable, de lettre, de poème en prose etc. Les hésitations terminologiques (nouvelle, conte, récit…) 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. ; classification souvent mise en question et critiquée par les écrivains eux-mêmes.

Qu'on le veuille ou non, les études ont montré que les genres existent, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence des différentes typologies de textes. Par ailleurs, dans l'étude quantitative – lexicométrique ou logométrique 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.

L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques. Cette opposition générique dans les analyses de statistique lexicale est si forte qu'elle empêcherait même de fonder de grands espoirs sur les méthodes quantitatives pour attribuer un texte à un écrivain plutôt qu'à un autre. Un excellent exemple de la force de ce clivage générique est celui de certaines tragédies de Molière que l'on a attribué à Corneille (Brunet, 2000).

Ces variations, indiscutables et déjà bien documentées par ailleurs, sont-elles observables également à l'intérieur d'un corpus ou dans l'œuvre d'un seul écrivain ? Comment évoluent-elles ? Les oppositions génériques sont tout d'abord à constater dans la structure du vocabulaire et dans son évolution ; c'est l'étude de la richesse lexicale, de la diversité du vocabulaire, de l'accroissement lexical ainsi que des hapax qui permet de tirer des conclusions sur ce phénomène.

L'étude des parties du discours et de la syntaxe à travers une analyse "grammatico-métrique", possible grâce aux versions lemmatisées et étiquetées du corpus, permet de relever aussi certains aspects morphologiques et syntaxiques qui différencient les types de textes.

L'opposition générique opère aussi au niveau du contenu lexical et thématique d'une œuvre. L'étude de la distance lexicale entre les différents livres du corpus ainsi que celle des spécificités lexicales, mettent en exergue les variations thématiques, ou isotopiques, récurrentes dans ce corpus "multigénérique".

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. Dans cet exposé nous nous intéresserons à une seule monographie celle de J.M.G. Le Clézio, un de nos plus grands écrivains contemporains.

Le Clézio s'est lui-même intéressé à tout le procédé de la création littéraire et ses idées se traduisent souvent par le refus de certaines normes littéraires, refus se présentant comme une contestation sociale. Accepter les conventions du roman, ou de tout autre type d'écriture présentait, surtout au début de sa création, pour l'écrivain le risque de s'enfermer dans un système sociopolitique, dans un cloisonnement conventionnel des genres qui le dérangeait au plus profond. Tout au long de sa production littéraire, Le Clézio a en effet tenté des expériences en transgressant les catégories et les genres, notamment celui du roman.

"Tout ce qu'a écrit Le Clézio, écrit Michelle Labbé (Labbé : 1999, p. 17), du moins jusqu'à *Désert*, contient le roman de sa lutte contre le roman, sa quête de l'écriture, la grande histoire d'amour de l'œuvre. Il ne propose pas de théorie structurée sur la création romanesque ni de critique de forme sur le roman dit "traditionnel", comme ont pu le faire ses contemporains N. Sarraute, A. Robbe-Grillet, J. Ricardou ou Ph. Sollers, mais plutôt des réflexions fréquentes, récurrentes et dispersées."

On trouve en effet ces réflexions sur la littérature dans toute l'œuvre leclézienne, aussi bien dans les articles et les essais que dans les préfaces, les nouvelles, les romans et même les épigraphes aux chapitres de romans. Pourtant, même si l'auteur veut transgresser un système social établi, les différentes typologies de textes sont présentes dans l'œuvre leclézienne et des variations génériques sont à observer à tous les niveaux.

La production littéraire de Le Clézio est vaste, s'étend sur plus de quarante ans et englobe plusieurs genres littéraires. Le corpus numérisé et employé dans cette étude contient 2.281.659 occurrences et 51.009 formes (dans la version qui s'appuie sur les formes graphiques) réparties sur les trente-et-une œuvres du corpus.

Celui-ci est constitué 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*.

Ce grand corpus a été numérisé et traité par le logiciel *Hyperbase*, version 5.5., élaboré au sein du laboratoire CNRS "Bases, Corpus et Langage" de Nice. 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.

La structure lexicale

Les différentes recherches sur la structure lexicale offrent la possibilité, indépendamment du contenu lexical, de situer, de distinguer et comprendre la structure formelle des textes afin de pouvoir comparer différents discours, genres, époques ou auteurs différents au niveau exogène aussi bien qu'au niveau endogène, 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.

Les graphiques suivants permettent de constater une des caractéristiques de notre corpus ; le premier histogramme regroupe les 100 plus hautes fréquences, c'est-à-dire les mots les plus fréquents : mots-outil etc. et rend compte de leur distribution ; le deuxième illustre la distribution des hapax dans les différentes œuvres du corpus:

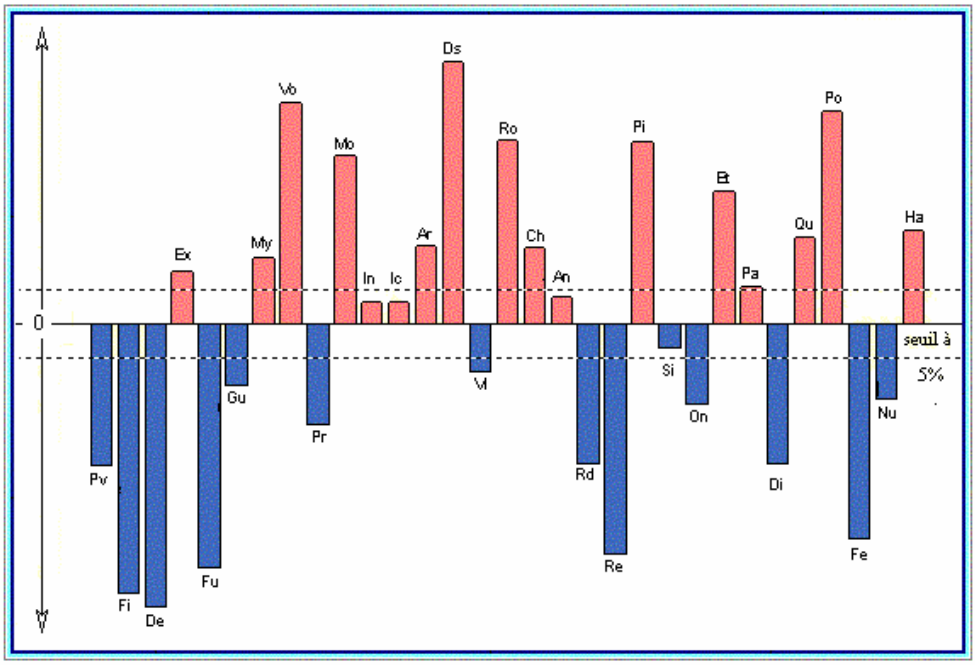


Figure n°1 : La distribution des plus hautes fréquences à travers le corpus

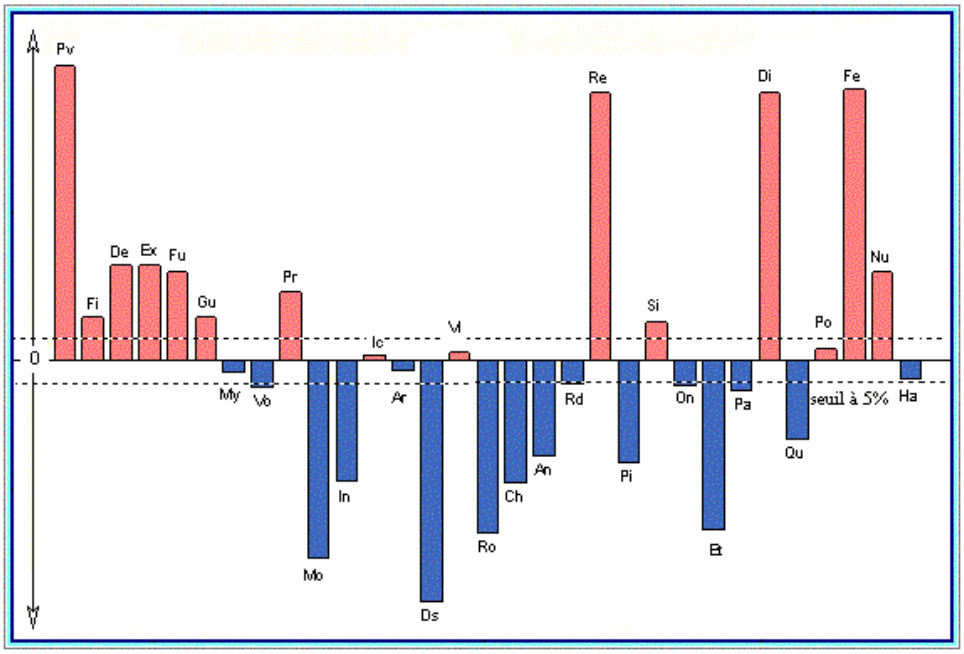


Figure n°2 : La distribution des hapax à travers le corpus.

L'accueil fait aux hapax est en effet déterminé par le genre littéraire. Les taux négatifs, à quelques exceptions près, sont à trouver dans les romans et dans les nouvelles. Dans les autres ouvrages ainsi que dans les œuvres inspirées par l'école "nouveau roman", les hautes fréquences sont importantes. Les histogrammes illustrent parfaitement l'inversion du mouvement dans la distribution de fréquences. Il est aisé de constater que les livres qui contiennent le plus d'hapax sont les plus "pauvres" en hautes fréquences.

L'étude la plus traditionnelle en lexicométrie est peut-être celle du rapport entre le nombre d'occurrences (N) et le nombre de vocables (V). Ce rapport donne une idée du nombre de mots différents comparé à l'étendue des textes et il permet, les valeurs correctement pondérées, de mesurer la richesse lexicale. L'analyse de la richesse lexicale des différents ouvrages reflète aussi l'influence du genre dans lequel il s'inscrit. Notre corpus ne fait pas exception à cette règle, déjà bien documentée par ailleurs. En effet, les caractéristiques des différents genres se retrouvent dans notre corpus. Les romans et les nouvelles présentent le vocabulaire le plus "pauvre" tandis que les essais, les ouvrages ethnologiques et les récits de voyage offrent le vocabulaire le plus "riche". Dans ces derniers ouvrages, nous pouvons également noter la même tendance à la hausse de la richesse lexicale vers la fin de l'œuvre.

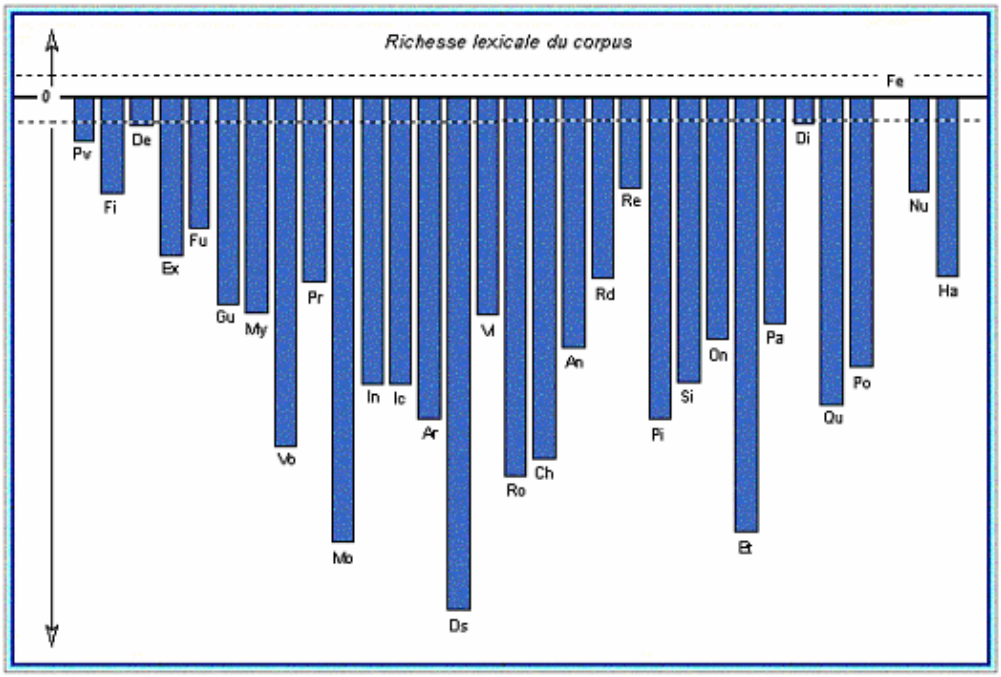


Figure n°3 : La richesse lexicale calculée sur l'étendue relative des textes, suivant la loi binomiale

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

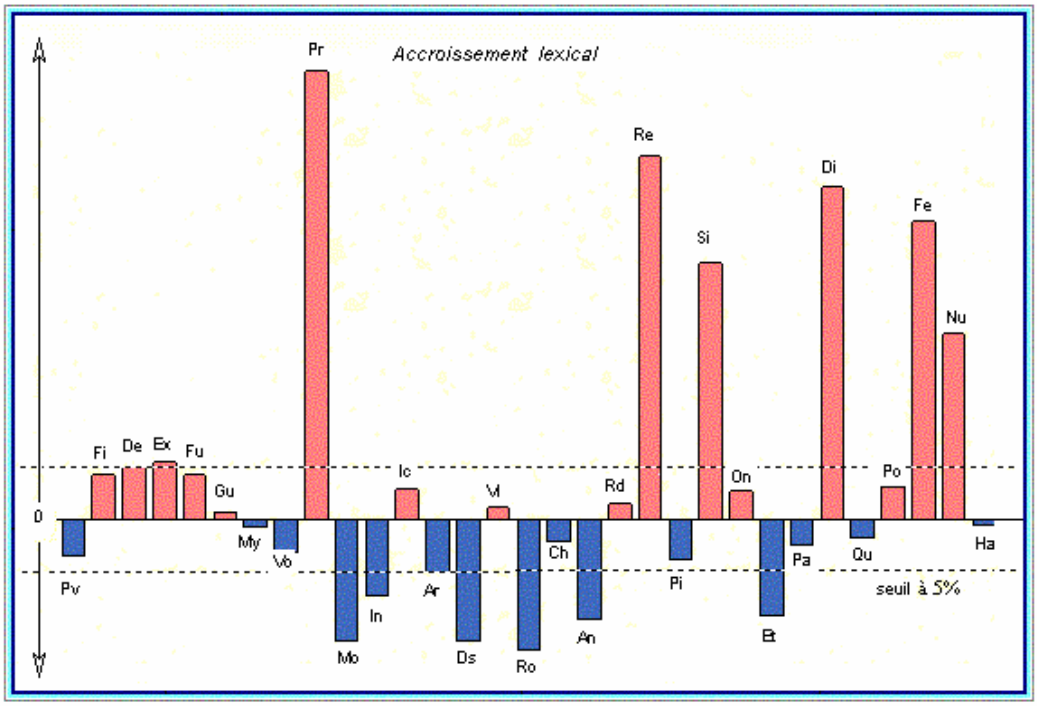


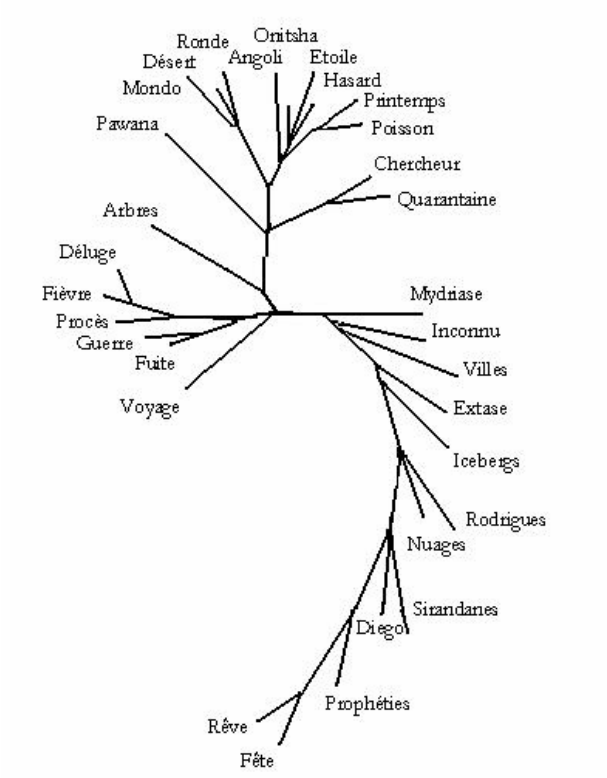
Figure n°4 : Accroissement lexical du corpus

Le graphique qui, de gauche à droite, s'oriente selon la chronologie, 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, comme dans l'étude de la richesse lexicale et des hapax, l'opposition générique très importante du corpus.

En outre un deuxième facteur se superpose au facteur générique : le facteur chronologique, qui divise l'œuvre de Le Clézio en trois périodes principales. Nous avons pu constater la courbe récurrente d'un vocabulaire qui croît de manière significative au début de l'œuvre et qui décline brusquement à partir de la fin des années 1970, pour s'accroître de nouveau vers la fin de l'œuvre sans que ces dernières valeurs atteignent les apports de la période initiale. La chute que nous avons observée dans nos différents histogrammes correspond bien à la rupture dans l'écriture de notre auteur, si souvent évoquée par les critiques littéraires. Enfin, la troisième période apporte des thèmes nouveaux à partir de 1987, sans pour autant présenter des apports lexicaux très importants, sauf quand le genre l'impose, comme dans les ouvrages ethnologiques, dans les essais et dans la biographie.

Dans l'étude de la distance lexicale il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de repérer ceux qui partagent des thèmes semblables. Nous avons trouvé, comme auparavant – dans les analyses structurelles et stylistiques du corpus –, des oppositions fortes entre les différents genres littéraires et un regroupement des livres appartenant à une même variété générique. La division tripartite à l'intérieur du genre romanesque écarte les ouvrages initiaux inspirés de l'école "nouveau roman" des autres – en indiquant que le changement thématique intervient déjà avec *Voyages à l'autre côté* – pour ensuite distinguer les autres ouvrages romanesques en un deuxième et un troisième regroupements du corpus qui tiennent bien compte de la chronologie de l'œuvre et de son évolution.



Globalement, l'analyse de la structure lexicale du corpus permet de constater en premier lieu le rôle très important du genre littéraire. Les essais, les ouvrages ethnologiques et la biographie présentent une richesse lexicale avec une grande spécialisation du vocabulaire ainsi que des apports lexicaux importants dans notre corpus. En deuxième lieu, les différentes analyses mettent en évidence le facteur chronologique et l'évolution de l'œuvre. Les résultats confirment les intuitions contradictoires que peut avoir le lecteur de Le Clézio : d'un côté celle d'un vocabulaire riche, de l'autre celle d'un style pauvre, d'une écriture quelque peu répétitive. La bipolarité de la structure lexicale confirmée par l'analyse statistique, avec un

vocabulaire qui tend soit vers l'abondance soit vers le dépouillement, est le fidèle témoin du paradoxe de l'écriture leclézienne et oppose ainsi le genre "nouveau roman" au genre "roman traditionnel".

Ces oppositions observées à l'étude de la structure lexicale d'un corpus ne sont pas indépendantes de la syntaxe. Par exemple le déficit dans les basses et moyennes fréquences n'est pas un choix délibéré mais la conséquence d'un emploi intensif des mots grammaticaux – qui se concentrent dans les fréquences très élevées. C'est donc un choix syntaxique – dont nous relevons les effets dans le lexique. Il semble que Le Clézio fasse moins appel à un style recherché au point de vue de la syntaxe dans les ouvrages où il emploie beaucoup d'hapax, comme dans les ouvrages ethnologiques où la richesse d'hapax correspond souvent à la découverte d'une nouvelle culture. Inversement, dans les livres qui sont pauvres en hapax, comme dans les romans de la fin des années 1970, la richesse en hautes fréquences pourrait être un indice d'une plus grande complexité de la syntaxe.

L'opposition générique est en effet également à constater au niveau syntaxique et l'analyse quantitative de la distribution des différentes parties du discours constitue une base impartiale et concrète pour permettre à l'étude syntaxique d'un corpus de taille.

Les parties du discours

La distribution des parties du discours dans les ouvrages littéraires n'est pas constante, elle est fortement influencée par l'époque, l'auteur et le genre de discours, et l'emploi des catégories grammaticales dans un texte donné peut constituer un indice très révélateur.

En effet cette distribution, qui manifeste peut-être des choix plus subtils que celui du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, peut apporter à l'analyse des éléments nouveaux. Il s'agit en réalité de choix inconscients faits par l'auteur lors de la création et de l'élaboration d'un texte qui permettent au chercheur de distinguer des divisions grammaticales caractéristiques et personnelles.

Désormais la quantification et la lemmatisation des corpus ouvrent la voie à cette analyse. Elles demandent l'accès à la forme canonique du mot, au lemme, et ne peuvent guère se fonder sur la distribution des effectifs d'un corpus s'appuyant sur la forme graphique. C'est la lemmatisation qui permet d'étiqueter le corpus selon les catégories grammaticales et de classer les éléments du vocabulaire selon leur appartenance à une catégorie spécifique.

Les codes grammaticaux fournis par l'étiqueteur morphosyntaxique au cours de l'opération de lemmatisation "automatique" constituent ici un outil indispensable (Kastberg Sjöblom, 2002, pp. 80 - 88).

Le corpus *Le Clézio* a été traité avec la version d'Hyperbase lemmatisée selon le programme Cordial 7 qui aboutit au bout du traitement à quelque 200 codes grammaticaux différents, en utilisant toutes les combinaisons possibles. Nous en avons extrait les 11 catégories fondamentales parmi celles que propose le programme Cordial : verbes, substantifs, adjectifs, déterminants, pronoms, numéraux, interjections, prépositions, adverbes, conjonctions et délimiteurs (signes de ponctuations). Pour une vision synthétique des accords qui lient les codes grammaticaux et les différents sous-corpus, nous avons recours à l'analyse factorielle de la liste de fréquences de ces différentes classes du corpus :

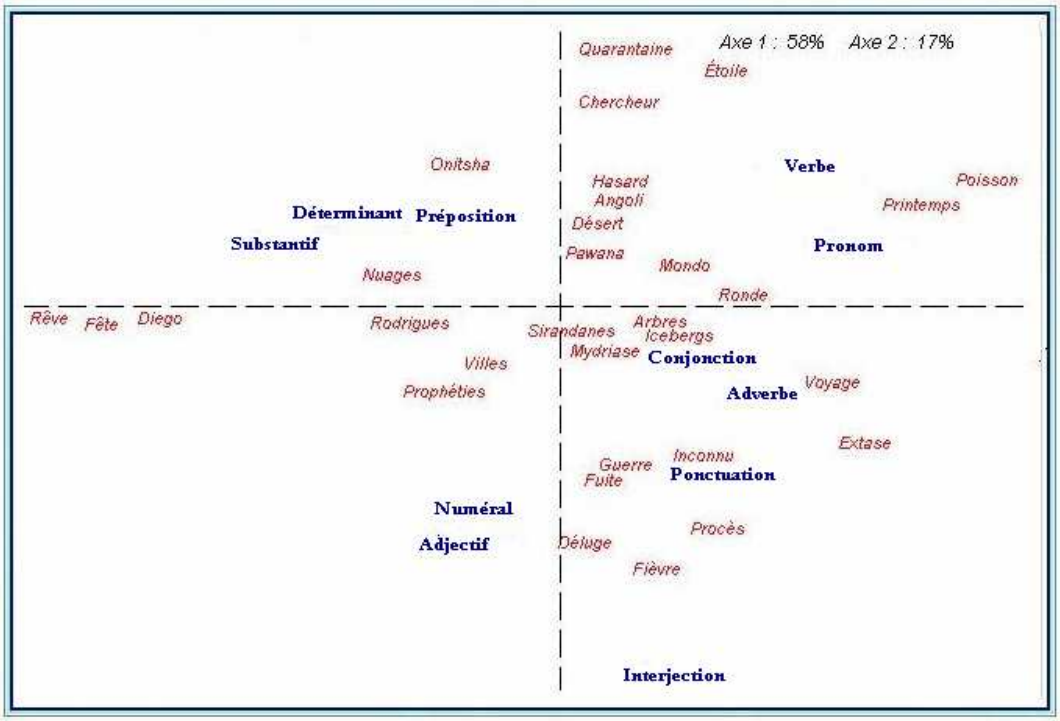


Figure n°5 : Analyse factorielle de la distribution grammaticale selon la lemmatisation par Cordial

Nous voyons que le premier facteur oppose la catégorie verbale à la catégorie nominale. Le substantif à gauche attire les prépositions, les déterminants et les adjectifs, tandis que le verbe en haut à droite attire les pronoms et les adverbes. Cette bipolarité que nous pouvons observer des catégories des substantifs et des verbes chez Le Clézio n'a pourtant rien d'original : elle a été observée dans bien d'autres corpus : Étienne Brunet l'a bien remarquée dans ses diverses études et il souligne également le rôle important de l'opposition des genres littéraires (Brunet 1985 : 155). De ce point de vue, l'œuvre de Le Clézio s'inscrit tout à fait dans la dynamique générale de la littérature française.

Le second facteur parcourt la chronologie de l'écrivain du bas vers le haut du graphique. Les premiers ouvrages, *Le procès-verbal*, *La fièvre*, *Le déluge, La guerre* et *Le livre des fuites* se trouvent en bas du graphique autour des catégories secondaires qui témoignent d'une écriture foisonnante (adjectifs, adverbes et interjections). Les derniers romans, *La quarantaine*, *Poisson d'or*, *Etoile errante*, *Le chercheur d'or* et *Hasard*, se situent en haut du tableau autour des catégories fondamentales, témoignant peut-être d'un assagissement de l'écriture, d'un travail de simplification de style.

L'analyse factorielle rend ici également aussi compte de l'opposition générique. Les ouvrages ethnologiques se regroupent à l'extrême gauche du graphique, les premiers romans appartenant à l'école du "nouveau roman" en bas à droite, tandis que les œuvres fictionnelles se trouvent au centre supérieur du tableau. Les ouvrages qui se trouvent au milieu sont les plus courts, tous genres confondus.

La tendance générale de la distribution des autres catégories grammaticales, les adjectifs, les verbes, et les adverbes etc., met presque toujours en évidence les mêmes phénomènes et les mêmes oppositions. Nous observons systématiquement en premier lieu l'opposition des genres littéraires ; les romans sont riches en verbes mais pauvres en substantifs tandis que les biographies et les ouvrages d'ethnologies, par exemple, sont tous très riches en adjectifs et en substantifs et pauvres en formes verbales. Ces typologies de texte, avec leurs nombreuses descriptions, sont nettement favorables à l'adjectif. Les essais, souvent de caractère poétique, le sont aussi, mais il s'agit là plutôt d'un effet de style. En outre apparaît généralement l'évolution chronologique de l'œuvre. L'adjectif notamment, après avoir été excédentaire au début de l'œuvre, régresse au fur et à mesure que l'œuvre progresse dans le temps et reflète ainsi un changement de style.

La distribution des verbes ne fait pas exception à cette tendance et même à l'intérieur de cette catégorie nous observons des phénomènes identiques. Le logiciel Hyperbase permet désormais de distinguer et de regrouper les sous-catégories de verbes de façon automatique (Kastberg Sjöblom, 2002, pp. 96–103). L'analyse regroupe les verbes selon leur statut de principal ou d'auxiliaire, selon le mode, selon le temps exprimé ou bien selon la personne.

C'est en effet le verbe qui gagne le plus à l'étiquetage: si l'on s'intéresse à son contenu sémantique, le regroupement des formes d'un même verbe offre un avantage indéniable, et si l'on étudie un temps verbal et qu'on regroupe les verbes qui partagent le même codage et le même temps, l'avantage est encore plus net. Lorsqu'il s'agit de l'emploi des verbes, on a tout lieu de penser qu'un écrivain y porte attention. Le choix qu'il fait du passé ou du présent, de

la première ou de la troisième personne, a des conséquences importantes pour la conduite du discours et une telle décision ne saurait être inconsciente. Le système verbal s'étageant sur plusieurs plans : le mode, le temps et la personne (sans compter le nombre, l'aspect et d'autres paramètres), on pourrait isoler successivement les trois plans principaux (les seuls que relève *Cordial*), ou bien les croiser et, par exemple, consacrer une ligne du tableau à la troisième personne du pluriel du présent de l'indicatif des verbes auxiliaires (croisement de 5 variables). En étudiant ensemble, en tant que variables indépendantes, les modes, les temps et les personnes, on se donne le moyen de repérer lequel de ces trois paramètres est le plus discriminant, mais aussi quelle interaction s'exerce entre les uns et les autres.

Les modes du verbe français se distinguent selon la tradition et les textes officiels en cinq ou en six classes : infinitif, participe, subjonctif, impératif, indicatif et conditionnel.

La distribution des différents modes dans notre corpus est celle que l'on trouve dans pratiquement tous les corpus littéraires, c'est-à-dire avec un indicatif qui domine largement (63,3%), et les participes et les infinitifs qui occupent à peu près un quart du groupe (respectivement 19,2% et 14,4%). Quant aux autres modes, ils sont minoritaires (conditionnel 1,6%, impératif, 0,8% et subjonctif 0,7%). L'analyse factorielle nous permet de situer les différents ouvrages de notre corpus par rapport à la distribution des modes :

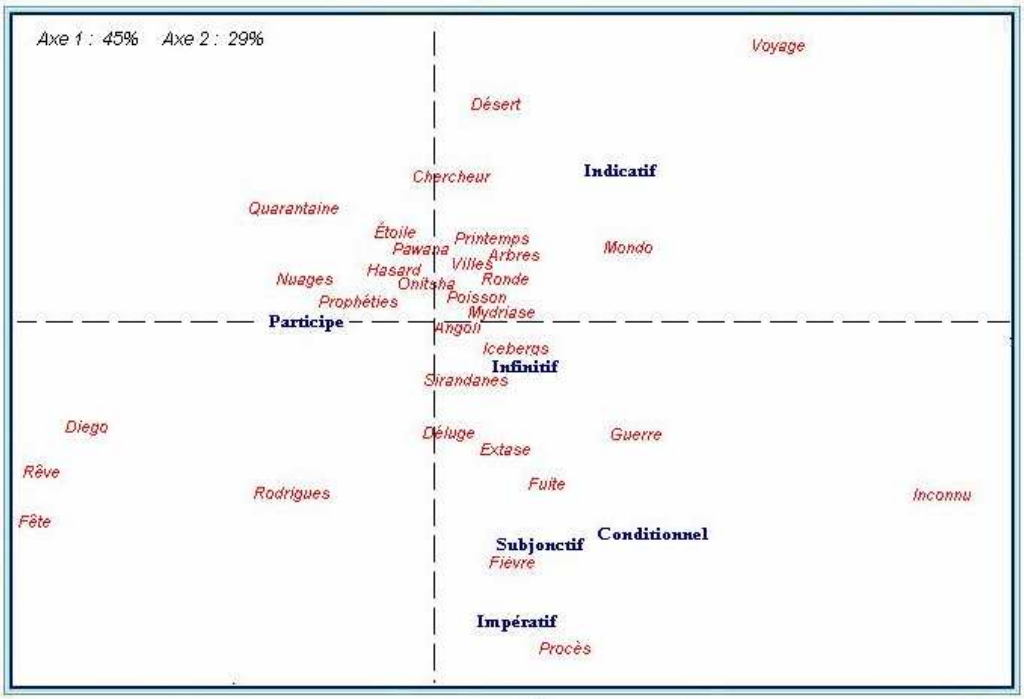


Figure n°6 : Analyse factorielle des modes verbaux

Le premier axe du graphique rend compte de l'opposition des genres littéraires, qui semble avoir une influence importante aussi sur l'usage des différents modes verbaux. Les

romans “traditionnels” se trouvent relativement au milieu du champ, dans la partie supérieure entre l’indicatif, qui est le mode traditionnel du récit, et le participe, qui est ici la trace des temps composés, en particulier celle du passé composé. Le subjonctif, le conditionnel et l’impératif appartiennent aux livres inspirés de l’école “nouveau roman”, regroupés en bas et à droite du graphique. Les essais tardifs, *Le rêve mexicain* et *La fête chantée*, ainsi que *Diego et Frida*, se trouvent ensemble éloignés du reste, en bas et à gauche. L’axe vertical reflète la chronologie de l’œuvre et rend bien compte de l’évolution dans l’écriture leclézienne. Les effets du style et la langue souvent recherchée du début de l’œuvre – se manifestant par le recours à des modes comme le subjonctif, le conditionnel ou l’impératif – seront abandonnés en faveur d’un style moins recherché, d’un récit plus traditionnel et une simplicité voulue de l’écrivain, privilégiant l’indicatif.

L’indicatif, mode du récit par excellence, domine effectivement le récit leclézien, mais la distribution interne montre toutefois des variations relativement importantes :

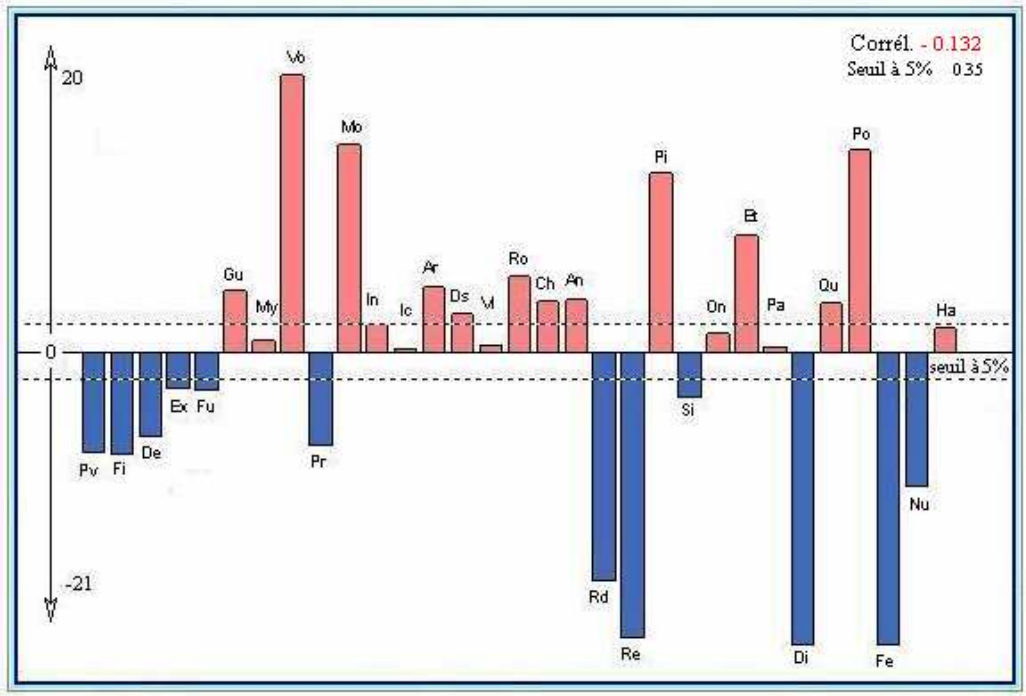


Figure n°7 : La distribution relative de l’indicatif dans le corpus (écarts réduits).

L’histogramme met en relief l’opposition des genres littéraires que nous avons constatée dans l’analyse factorielle. Les romans et les recueils de nouvelles privilégient l’indicatif – à l’exception des premiers romans et de *Voyage à Rodrigues* – tandis que ce mode est déficitaire dans les essais, les ouvrages d’ethnologie, la biographie et le récit de voyage. Nous pouvons aussi observer la tendance chronologique de cette distribution, avec des valeurs négatives de plus en plus importantes au fur et à mesure que l’œuvre progresse. C’est dans ces

livres que nous avons pu observer auparavant les déficits les plus grands par rapport à la catégorie des verbes dans son ensemble.

L'étude des temps verbaux reflète également l'opposition générique dans notre corpus. L'analyse factorielle des temps de l'indicatif (présent, imparfait, passé simple et futur) rend compte des rapports qui lient les temps et les différents ouvrages du corpus :

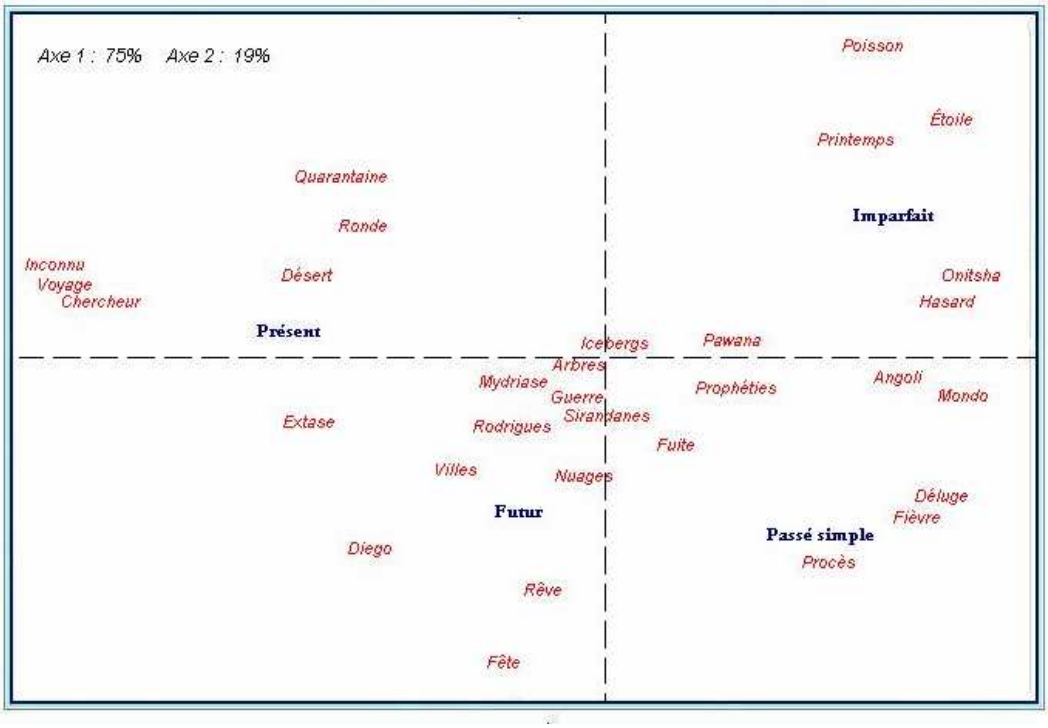


Figure n°7 : Analyse factorielle des temps verbaux

Le premier axe de l'analyse oppose à nouveau les genres littéraires (bien que la division soit moins nette que dans les analyses précédentes) ; l'imparfait – du côté droit du graphique – est attiré par les romans de la deuxième période de Le Clézio. Le passé simple est employé dans la période "nouveau roman", les ouvrages ethnologiques semblent favoriser le futur.

Quant au présent, sa position est plus difficile à expliquer, il semble que ce temps soit beaucoup employé par Le Clézio dans les romans de la deuxième période de l'œuvre. Le deuxième facteur de l'analyse factorielle rend compte de la temporalité et de l'évolution chronologique de l'emploi des différents temps verbaux chez Le Clézio. Nous trouvons les premiers livres en bas du graphique, une grande partie des ouvrages au milieu et tout en haut du tableau sont rassemblés les derniers romans : *Hasard*, *La quarantaine*, *Poisson d'or*, *Onitsha* et *Etoile errante*.

L'usage des différents temps verbaux dans un corpus est en effet un facteur qui – à part sa fonction première : nous situer dans le temps – est souvent déterminant pour le style d'un

écrivain, et change avec l'évolution d'une œuvre littéraire. Dans l'œuvre leclézienne, nous avons pu constater que l'emploi du verbe change au fur et à mesure que l'œuvre progresse et que la fréquence des verbes est plus ou moins dominante selon l'époque ou les genres littéraires. Nous avons également vu qu'à l'intérieur de la catégorie verbale il y des variations importantes quant au mode aussi bien qu'à l'emploi des différents temps verbaux qui reflètent bien le changement perpétuel et la recherche de renouvellement de notre écrivain, tout en gardant certaines constantes qui contribuent à donner au récit leclézien son caractère particulier, redondant, incantatoire et mystérieux.

Dans notre corpus, ce deuxième critère, morphologique, montre que la première période "nouveau roman" se démarque grammaticalement toujours du reste par son usage important du substantif et de l'adjectif, mais aussi par l'emploi de l'impératif et, paradoxalement pour une écriture expérimentale, par l'usage de formes temporelles très traditionnelles comme le passé simple. La rupture bien connue de l'œuvre leclézienne entraîne un changement vers une écriture qui privilégie l'action et par conséquent les catégories verbales, notamment les formes conjuguées à l'imparfait et les temps composés. L'étude de temps verbaux et de l'usage très personnel qu'en fait Le Clézio permet de mieux cerner une technique qui consiste à donner au récit cette valeur universelle tant appréciée par ses lecteurs.

Une écriture qui change est une des caractéristiques fondamentales de notre corpus. En effet, il n'y a pas de "stabilisation" du style mais, au contraire, des écarts grandissants chez le Clézio. Toutefois, bien que les procédés morphosyntaxiques ne soient pas statiques, que les techniques d'expression changent, qu'elles évoluent et qu'elles soient constamment mises en question, c'est l'opposition générique qui reste prépondérante.

Enfin, troisième critère, l'étude du contenu du discours qui implique la signification des mots, les différentes catégories lexicales ainsi que l'étude des spécificités – positives et négatives – permettent de dégager les caractéristiques d'une œuvre et de son évolution.

Sémantique

Aucun lecteur de Le Clézio n'est surpris par les résultats de l'étude statistique des thèmes de l'œuvre : certaines thématiques sont très importantes, la nature – terrestre et marine –, les couleurs ; d'autres comme le milieu urbain, les parties du corps, les insectes et le

minuscule sont très présentes au début de l'œuvre mais perdent de l'importance au fur et à mesure que l'œuvre progresse ; de plus leur importance varie selon le genre de texte. Dans une perspective endogène, les variations à l'intérieur du corpus sont toutefois importantes et l'étude statistique des spécificités permet de les cerner.

Le milieu maritime est un décor de prédilection des récits lecléziens. Déjà *Le procès-verbal* se déroule dans une ville au bord de mer que l'on s'accorde à identifier comme étant Nice et le port d'attache du bateau *Azzar* sur lequel se déroule la fiction du dernier livre du corpus, *Hasard*, est bien Villefranche-sur-mer, à quelques kilomètres seulement de Nice.

La répartition chronologique des emplois du mot *mer*, qui est le premier de la liste de spécificités, résume à elle seule l'évolution de la structure thématique de la nature :

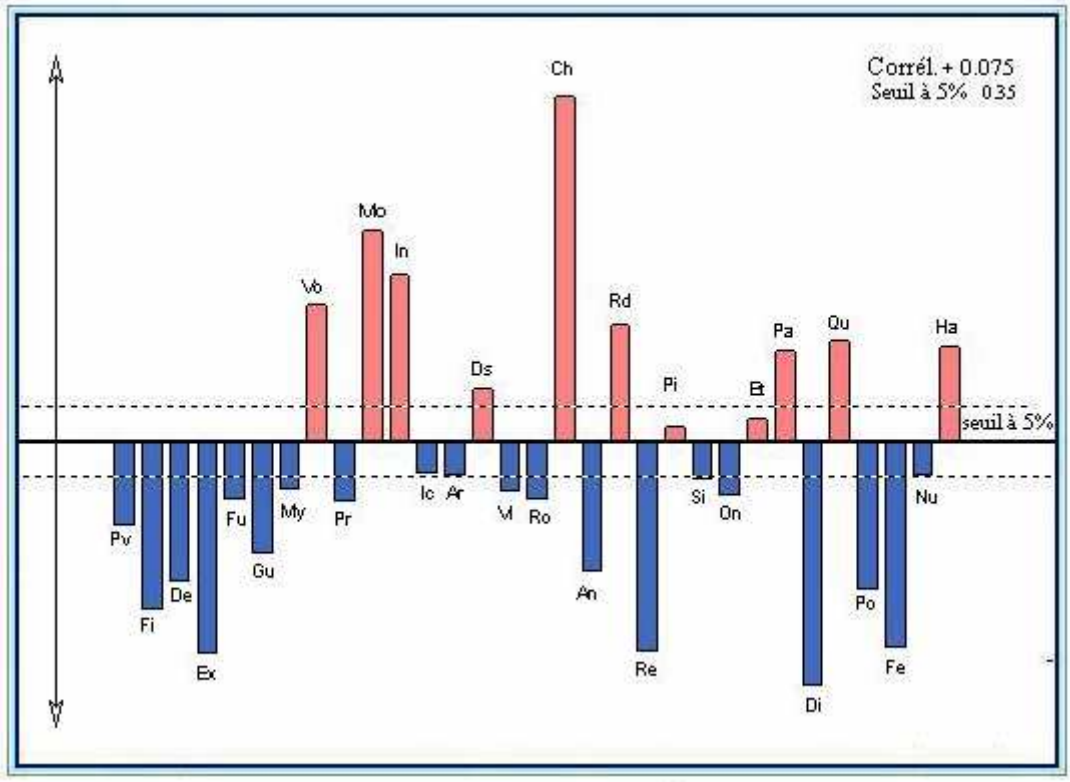


Figure n°8 : Le mot *mer* dans le corpus

Considérons *la mer* comme un thème lecléziens ; nous ne le ferons pas à partir d'une liste projective de lexèmes (*océan*, *vague*, *bateau*, etc.…) mais en prenant pour base une hiérarchie endogène de vocables, avec l'aide du logiciel Hyperbase. Ce logiciel permet en effet d'effectuer une recherche de contextes automatique où chaque occurrence d'une forme est montrée dans son contexte (ici le paragraphe). Par exemple voici ci-dessous quelques-uns des 3367 contextes obtenus pour *la mer*.

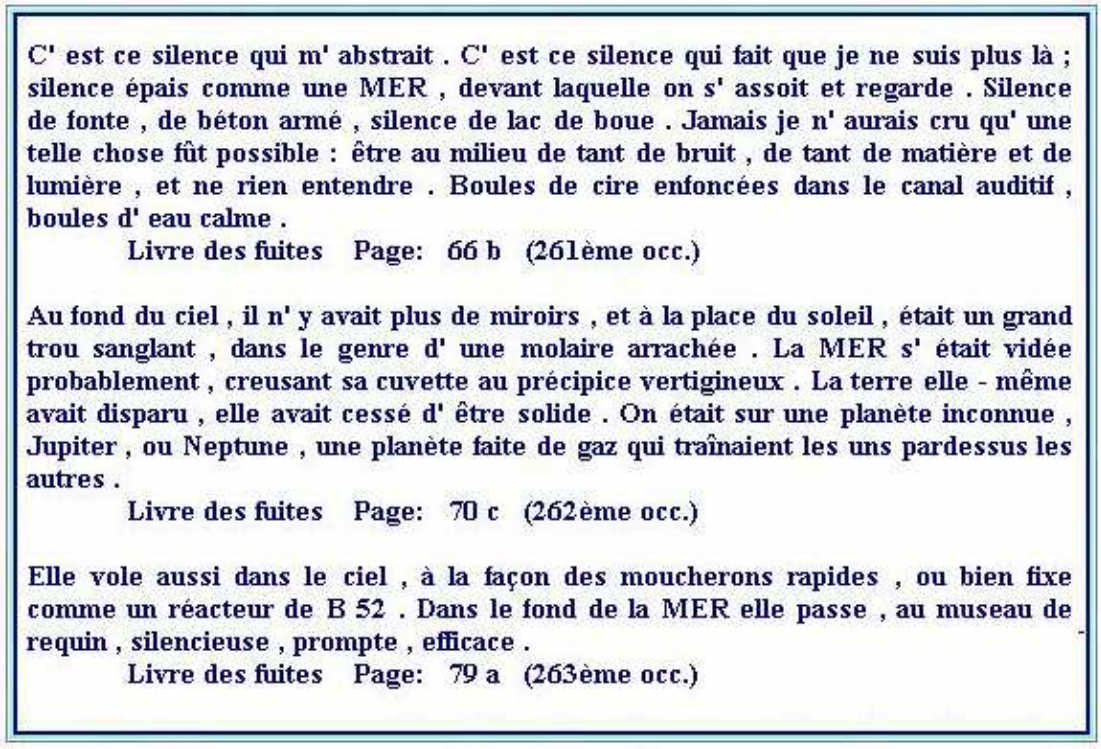


Figure n°9 : Quelques contextes du mot *mer*

Il est intéressant de regarder de près la structuration thématique autour du pôle *mer*. Le logiciel Hyperbase s'y emploie avec une fonction d'extraction de thèmes. Il s'agit d'un calcul de spécificité particulier, puisqu'on ne recherche plus une relation entre un mot et un texte, mais une relation privilégiée entre les mots eux-mêmes – ce que mesure aussi le calcul de corrélation, quand deux séries sont juxtaposées dans le même graphique. La procédure ne se réduit cependant pas ici à deux mots confrontés, mais à l'ensemble indéfini de tous les mots qui peuvent se trouver dans l'entourage d'un mot (ou d'un groupe de mots), qu'on définit comme étant le pôle (Kastberg Sjöblom & Brunet, 2000, pp. 457-465.).

En confrontant le mot *mer* à son entourage, ici le paragraphe, on obtient un fichier discontinu de 300 000 mots dont on fera un sous-corpus constitué par les mots qui gravitent autour du pôle. Reste à comparer ce sous-ensemble au corpus Le Clézio qui est 7 fois plus important. Ils produisent en fin de compte la liste hiérarchique ci-dessous.

écart	corpus	texte	mot	écart	corpus	texte	mot	écart	corpus	texte	mot
166.81	3367	3403	MER	12.91	105	53	DANIEL	9.68	776	170	BORD
34.86	2344	787	VENT	12.75	1961	393	BRUIT	9.63	88	38	PROUE
33.17	59489	9093	LA	12.66	506	145	BATEAU	9.46	293	83	ENTENDS
33.01	701	351	VAGUES	12.50	93	48	ÉTRAVE	9.30	871	182	NUAGES
23.25	598	244	HORIZON	12.30	1904	378	LOIN	9.28	53	27	PÊCHE
23.06	178	116	ÉCUME	12.11	222	81	DUNES	9.25	116	44	PLAGES
23.00	2936	714	CIEL	12.00	588	156	ÎLE	9.18	117	44	MARÉE
19.30	704	238	OISEAUX	11.94	70	39	ZETA	9.12	122	45	DENIS
19.16	782	254	PLAGE	11.91	2737	497	JUSQU'	9.01	622	139	SOMBRE
17.88	85	61	RÉCIFS	11.91	836	200	BLEU	8.99	220	66	HAUTE
17.54	462	169	ROCHERS	11.80	100	48	ALGUES	8.93	66	30	CORAIL
15.62	362	133	NAVIRE	11.67	734	180	MONTAGNES	8.88	4624	699	NOUS
14.80	2793	553	SOLEIL	10.82	73	37	TIMONIER	8.67	928	185	NAJA
14.67	1387	324	DESSUS	10.62	2851	492	EAU	8.67	151	50	CRÉPUSCULE
14.33	198	85	SEL	10.62	157	59	LAGON	8.55	153	50	LULLABY
14.14	44660	5860	LE	10.61	149	57	VOILES	8.52	246	69	ÉTENDUE
14.11	425	138	PONT	10.31	49	28	RIVAGES	8.47	284	76	PORT
13.66	225	89	BAIE	10.30	3874	628	VERS	8.47	63	28	VACOAS
13.65	198	82	RIVAGE	10.30	271	83	PLATE	8.42	663	141	SOUFFLE
13.52	102	54	MOUETTES	10.28	243	77	BLEUE	8.37	319	82	ROCHER
13.45	14736	2136	SUR	10.25	473	122	COLLINES	8.35	87	34	ESTUAIRE
13.37	255	95	VAGUE	10.20	139	53	BATEAUX	8.19	169	52	TEMPÊTE
13.35	215	85	ÎLES	10.10	12621	1747	AU	7.96	64	27	SILLAGE
13.17	889	221	SABLE	9.82	230	72	LISSE	7.93	810	160	COULEUR

Figure n°10 : L'environnement du mot *mer*

Au-delà des "corrélats" qui tiennent à la syntaxe (c'est le cas ici pour l'article féminin singulier *la* que la mer privilégie à l'exclusion des autres articles), à la phraséologie, aux expressions toutes faites, nous trouvons les véritables liens sémantiques, le partage de sèmes communs par lequels on peut définir un thème.

Le thème de la mer chez Le Clézio peut ainsi être défini à partir de l'usage de l'écrivain lui-même et non plus par une norme externe. En tête de liste se trouvent des mots que l'on associe souvent avec la mer : *vent*, *vagues*, *horizon*, *écume*, *ciel*, *oiseaux*, *plage*, etc. Il s'agit encore d'un univers où l'être humain et la civilisation moderne n'ont pas de place, cette mer, souvent de l'hémisphère sud (il n'y a point de glace ou d'icebergs dans cette liste) semble être chez Le Clézio bien plus un univers que l'on contemple qu'un espace de navigation et de transport.

Une fois le thème isolé – c'est-à-dire les mots qui gravitent autour du pôle *mer* – nous pouvons illustrer, par un graphique, l'évolution chronologique de la constellation lexicale qui entoure le pôle.

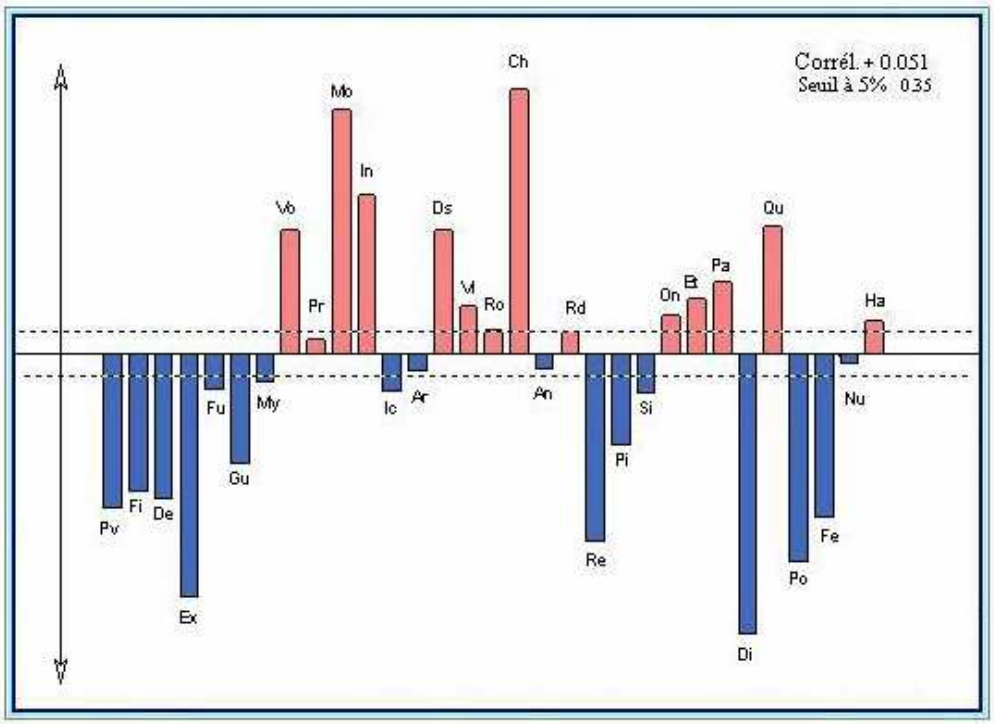


Figure n°11 : L'évolution d'ensemble du thème *mer*

La tendance du thème suit et complète celle du mot-pôle *mer* et elle est aussi en accord avec celle du vocabulaire spécifique. Les déficits de la période initiale font place à des excédents, du moins lorsque le genre romanesque est seul en cause. Les différents éléments qui forment cette structure ne sont évidemment pas distribués de la même façon dans les différents ouvrages. C'est peut-être l'analyse factorielle qui rend le mieux compte des multiples liens entre les différents termes constitutifs du thème et les diverses œuvres :

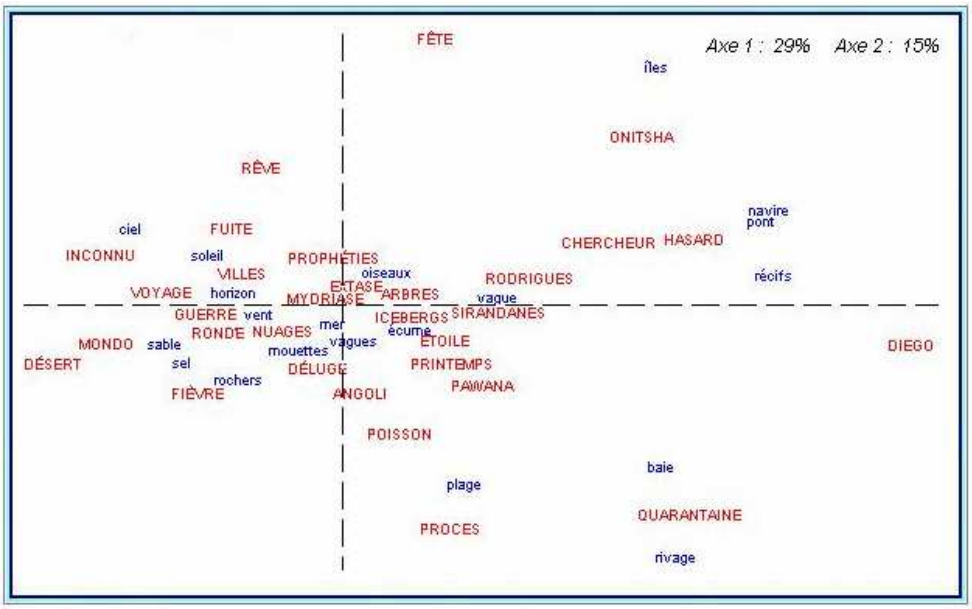


Figure n° 12 : Analyse factorielle de la structure thématique de *mer*

Dans cette analyse, nous ne retrouvons ni notre opposition générique habituelle ni la chronologie de l'œuvre. Il semble que d'autres facteurs prévalent dans ce cas bien que les axes ne soient pas très significatifs avec leurs 29% et 15%. Nous trouvons plutôt les décors et des éléments de l'action des différents livres. Sur le côte droit inférieur du tableau nous situons l'élément aquatique avec *Le procès-verbal*, qui se déroule en grande partie au bord de la Méditerranée. Dans *La quarantaine*, les personnages se trouvent dans la baie de l'île Plate en attendant de pouvoir poursuivre leur voyage à l'île Maurice ; dans *Le chercheur d'or*, *Hasard* et *Onitsha* il est question de traversées maritimes, d'où l'importance du *navire*. Sur le côté gauche du graphique nous trouvons dans la partie supérieure les romans qui ont plus affaire à l'élément terrestre, dans la partie inférieure on retrouve la bordure entre *mer* et *terre* avec le vent, le sable, les rochers et les mouettes, qui se trouvent également dans les ouvrages qui se déroulent dans le désert comme *Désert* et *Gens des nuages*.

De la même façon que dans les analyses structurales et morphosyntaxiques, l'analyse des corrélats sémantiques et thématiques révèle aussi des caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations des textes se retrouvent aussi bien au niveau lexical et syntaxique qu'au niveau thématique.

Conclusion

Ainsi, la numérisation et l'analyse lexicométrique de la quasi totalité des textes d'une monographie nous ont permis de mettre en exergue 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. Ces résultats contredisent d'une certaine manière ce qu'a souvent écrit Le Clézio à propos de son écriture et des genres littéraires, notamment dans l'ouvrage *La fièvre* (1965 : 143). "Tout et rien. Je prenais des feuilles de papier, les plus grandes possible, et je les couvrais d'écriture, presque sans y prendre garde, presque au hasard. Mais ça n'avait aucun genre littéraire, c'était simplement de l'écriture."

En effet, le refus de genres est souvent une position idéaliste ou sociopolitique. Aussi, bien que Le Clézio refuse toute appartenance à un genre littéraire et que les critiques aient souvent souligné le mélange des genres dans un même ouvrage, nos analyses ont montré que l'appartenance à un genre précis de chacun de ses livres est bien réelle.

Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

Bibliographie

Adam J.-M. (1992) : *Les textes : Types et prototypes*, Nathan, collection Fac. linguistique, Paris.

Adam J.-M. (2005) : *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*, 2005, Arman Colin, collection Fac. linguistique, Paris.

Brunet E. (1985) : *Le vocabulaire de Zola*, Champion-Slatkine, Paris-Genève.

Brunet E. (2000) : "Peut-on mesurer la distance entre deux textes ?", in Rastier F. (éd.) *Corpus littéraires – Recueil et numérisation, analyses assistées, didactique*,) Paris 20-21 octobre 2000.

Kastberg Sjöblom M. (2002) : *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*, Thèse de doctorat, Université de Nice-Sophia Antipolis, Nice.

Kastberg Sjöblom M.& Brunet E. (2000) : "La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain", in Rajman M. & Chappelier J.-C. (éds.), *JADT 2000, 5èmes Journées internationales d'Analyse statistique des Données Textuelles*, Ecole polytechnique fédérale de Lausanne, Lausanne, 2000, p. 457-465.

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

Labbé M. (1999) : *Le Clézio, l'écart romanesque*, L'Harmattan, Paris.

Le Clézio J.M.G. (1965) : *La fièvre*, Gallimard, L'Imaginaire n° 253, Paris.

Malrieu D. & Rastier F. (2002) : "Genres et variations morphosyntaxiques", in Angel Martin Municio (éd.), *Actas del segundo seminario de la escuela interlatina de altos estudios en lingüística aplicada, Matemáticas y tratamiento de corpus, San Millán de la Cogolla, 19-23 septiembre de 2000*, Logroño, Fundación San Millán de la Cogolla, p. 61-84.

Muller Ch. (1977) : *Principes et méthodes de statistique lexicale*, Hachette, Paris.