Analyse grammatico-métrique d'une monographie « multi-générique » ; le substantif

Margareta Kastberg Sjöblom

ILF-CNRS, Bases, Corpus et Langage (UMR 6039)
UFR Lettres, Arts et Sciences humaines, 98, Bd Edouard Herriot, 06204 NICE Cedex 3
kastberg@unice.fr

Abstract

The purpose of this paper is to explore an analysis of the parts-of-speech of a literary corpus. Today's techniques allow, by the integration of lemmatizers into the statistical tool, direct access to normalized and tagged textual data, as well as to the indispensable grammatical codes. This automatic sorting of the grammatical divisions paves the way to an impartial overview of the specific and personal style of an author as well as to the genre-related variations. The grammatical distribution, by its unconscious character, seems to show more subtle nuances than the traditional thematically orientated study and provides new valuable elements to stylistic studies.

Résumé

Cet article s'intéresse à l'étude des parties du discours d'un corpus littéraire contenant des textes de genres différents. Désormais l'intégration de lemmatiseurs dans les logiciels lexicométriques permet, grâce à l'accès aux codes grammaticaux, le recensement automatique et impartial des divisions grammaticales, personnelles et caractéristiques des typologies de textes diverses. Cette distribution, qui est bien un critère de distinction des œuvres, manifeste peut-être des choix plus subtils que celui du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, ce qui peut apporter à l'analyse des éléments nouveaux.

Mots-clés : lexicométrie, lemmatisation, stylométrie, stylistique de corpus, genres littéraires, typologie textuelle, Le Clézio.

1. Introduction

Les parties du discours, ou les catégories grammaticales, sont une classification des mots qui nous a été transmise par la syntaxe traditionnelle, basée sur trois critères qui sont la nature, la fonction et la position. Quel est l'intérêt d'analyser la distribution des parties du discours dans un corpus littéraire « multi-générique » informatisé, et de quelle façon cette analyse peut-elle contribuer à la compréhension de l'utilisation et de la fonction des catégories grammaticales ?

« Investigating this question, écrit Douglas Biber (1998 : 57), can help to understand how different varieties exploit the grammatical categories of words available to them… Using the corpus, we can analyze the distribution and function of these different categories of words and study the part that they play in fulfilling the communicative function of different registers."

En effet, la distribution des parties du discours n'est pas constante, elle varie selon les époques, les auteurs et les genres. Déjà l'étude de Guiraud (1954) montre que la distribution des catégories grammaticales dans les ouvrages littéraires est fortement influencée par le genre et l'époque où ils s'inscrivent. Cette distribution, qui est bien un critère de distinction des œuvres, manifeste peut-être des choix plus subtils que celle du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, ce qui peut apporter à l'analyse des éléments très révélateurs. Peut-on à partir de l'étude des parties du discours mieux comprendre l'écriture et

les différentes typologies de textes ? Quels sont les particularités grammaticales qui contribuent au style d'un écrivain et quel est le rôle du genre littéraire dans leur distribution ? Dans cette étude qui s'appuie sur une monographie contemporaine « multi-générique » – l'œuvre de J.M.G. Le Clézio (annexe 1) – nous nous attarderons, après avoir étudié la distribution des catégories grammaticales, sur quelques aspects morphologiques et syntaxiques qui nous ont paru intéressants et révélateurs dans la catégorie nominale.

Malheureusement les parties du discours ne sont pas des catégories entièrement stables et étanches les unes aux autres. Si dans chaque contexte l'appartenance d'un mot à une catégorie précise s'impose sans grande difficulté, en revanche le même mot rentrera dans une autre catégorie pour peu que change le contexte, même si le contenu sémantique reste constant. Les cas de *Venir pour dîner* et *Venir pour le dîner* ou bien *Tout est bon* et *Le tout est bon* sont des exemples de ce phénomène relativement courant dans la langue française. En effet, les fondements et les méthodes sur lesquels s'appuient les études des catégories grammaticales varient d'un ouvrage à l'autre. Il convient donc, avant de se lancer dans l'étude de catégories grammaticales de s'attarder un peu sur les principes et les techniques qui guident cette étude.

2. Les méthodes d'analyse

Désormais la quantification et la lemmatisation des corpus ouvrent la voie à cette composante essentielle de l'écriture qu'est l'étude de la distribution des parties du discours. Cette analyse, qui demande l'accès à la forme canonique du mot, au lemme, ne peut 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.

Toutefois, la lemmatisation automatique repose sur des critères préétablis selon des méthodes d'analyse différentes et les résultats peuvent varier selon le choix de l'étiqueteur et de ses principes. Il faut également savoir que la lemmatisation automatique n'est pas sans erreurs : erreurs d'analyse contextuelles, mauvaise définition de certaines classes grammaticales, etc. Il est donc recommandé de tenir compte de ces imperfections lors de l'exploitation des résultats. Les parties du discours distinguées par les logiciels d'analyse morpho-syntaxique sont toutefois nettement plus fiables que les fonctions grammaticales car leur reconnaissance est plus facile. Pour beaucoup de mots qui ne sont pas sujets à l'homographie, le codage est automatique et indépendant du contexte : la proposition du dictionnaire, unique, est immédiatement acceptée. Et quand il s'agit de diviser les homographes en catégories différentes (par exemple les cas très nombreux, du type *la marche/il marche*, où le substantif peut se confondre avec un verbe), une analyse syntaxique simple, désormais quasi-automatique (selon les différents logiciels), emporte la décision. D'autres cas peuvent être plus difficiles, comme la reconnaissance de l'adverbe *fort* dans le syntagme *un homme fort bon*. Face à ces ambiguïtés d'analyse, il y a différents types de lemmatiseurs : ceux qui prennent une décision dans tous les cas de figure (Cordial) et ceux qui rendent la main au linguiste (Logiciel Labbé). Le choix d'un type de lemmatiseur ou l'autre se fait en fonction de la taille de corpus (s'il est grand l'analyseur automatique offre plus d'avantages que d'inconvénients, s'il est de taille plus modeste l'autre solution est peut-être préférable). Nous avons pratiqué les deux solutions dans notre thèse, ce qui nous a permis de constater d'ailleurs que les résultats restaient assez stables de l'un à l'autre. L'étiquetage morphologique est généralement plus fiable que le « parsage » syntaxique qui recourt à des modèles et des procédures complexes.

Dans cette étude nous exploitons un corpus qui englobe la quasi-totalité de l'œuvre de Le Clézio (Cf. Kastberg Sjöblom 2002 : 100-110) lemmatisé avec l'analyseur Cordial 7, désormais intégré au logiciel Hyperbase, le logiciel lexicométrique sur lequel s'appuie cette analyse.

3. La distribution des catégories grammaticales du corpus

L'analyse morpho-syntaxique Cordial aboutit à quelque 200 codes grammaticaux différents en utilisant toutes les combinaisons possibles. Le logiciel Hyperbase regroupe par la suite de façon « automatique » les codes et fournit la liste des fréquences.

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). Cette extraction met en évidence la richesse en substantifs (20,5 %) de notre corpus, une des caractéristiques des corpus de langue française. Les catégories des déterminants, des signes de ponctuation et des verbes occupent chacune entre 14 et 15 % du corpus. Quant aux pronoms et aux prépositions, ils se regroupent autour de 10 % chacun tandis que les adverbes, les adjectifs et les conjonctions se rassemblent autour de 5 % chacun. Les catégories de numéraux et d'interjections sont dérisoires par leur nombre. Pierre Guiraud (1954 :104) avait déjà observé que dans la littérature le nombre de substantifs et celui des verbes varient en proportion inverse, le substantif étant dominant dans la prose abstraite et le verbe dans les récits. La richesse du substantif dans l'écriture leclézienne, caractéristique partagée avec les grands écrivains du XIX^{ème} siècle, est étonnante car on aurait pu croire que les romans et les nouvelles de Le Clézio étaient plus proches des récits et donc devaient favoriser le verbe.

Si nous regardons d'un peu plus près les catégories grammaticales dans notre corpus, nous constatons que leur distribution n'est pas régulière. L'analyse factorielle de la liste de fréquences des différentes codes grammaticaux illustre des variations importantes :

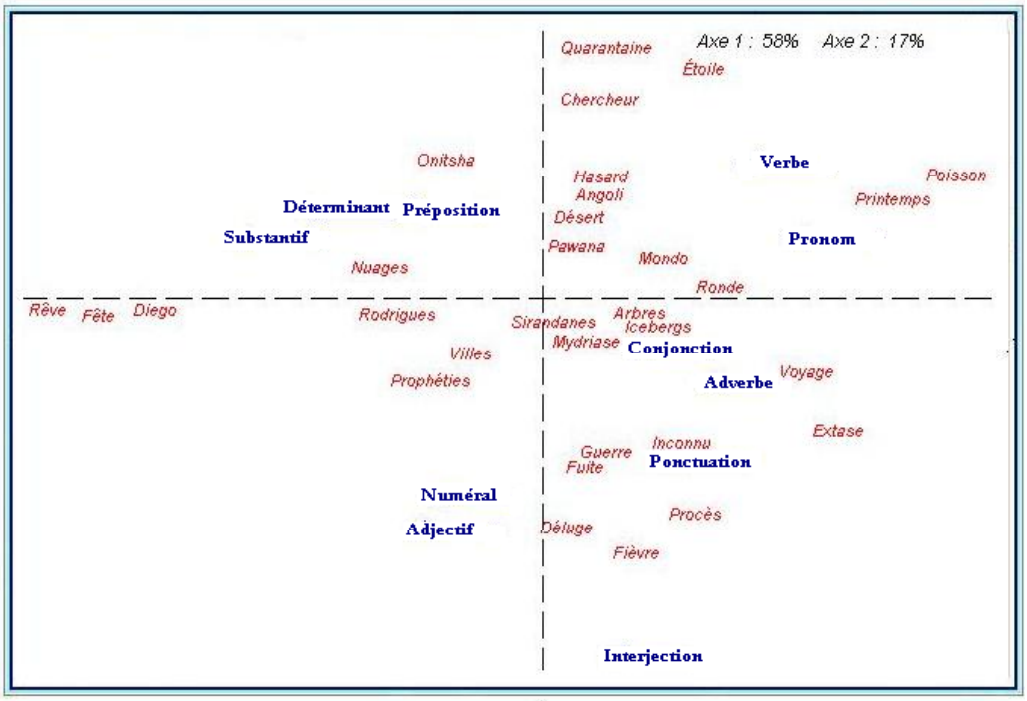

Figure 1.
Analyse factorielle de la distribution grammaticale du corpus selon la lemmatisation par Cordial 7.

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. Le second facteur parcourt la chronologie de l'écrivain du bas vers le haut du graphique. Les premiers ouvrages se trouvent en bas du graphique autour des catégories secondaires qui témoignent d'une écriture foisonnante (adjectifs, adverbes et interjections). Les derniers romans 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 également compte de l'importante 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. Examinons pour l'instant le phénomène dont témoigne le premier facteur de l'analyse factorielle, l'opposition du groupe verbal et du groupe nominal.

4. L'opposition des catégories grammaticales

On observe souvent dans un corpus clos, comme nous venons de le faire, que deux camps, la catégorie nominale et la catégorie verbale, s'affrontent : la classe du verbe et les catégories qui lui sont proches (subordonnants, relatifs, pronoms et adverbes) s'opposent à la classe nominale qui réunit autour du substantif les adjectifs, les déterminants, les prépositions et souvent les coordinations.

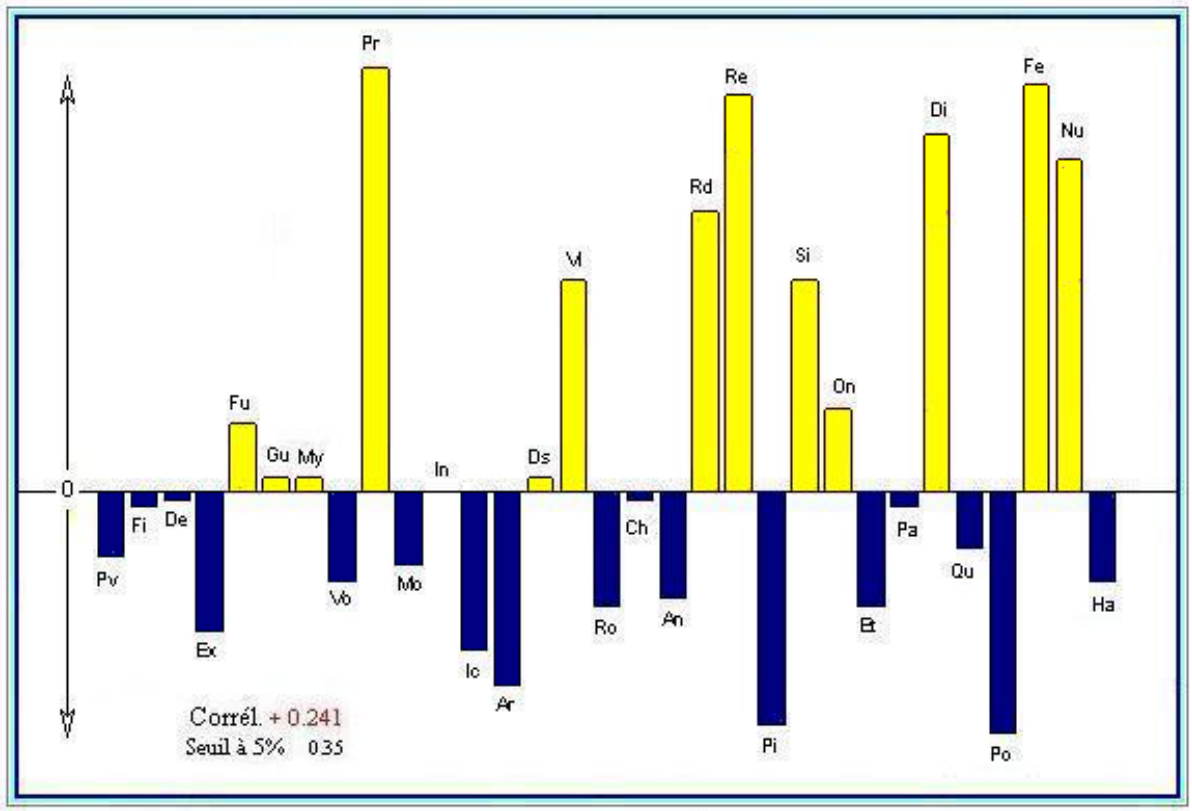

Figure 2. Histogramme du quotient substantifs/verbes.

L'histogramme du quotient entre les 459.957 substantifs et les 321108 verbes permet de voir de près ce que l'analyse factorielle nous a déjà indiqué : lorsqu'une œuvre est riche en subs-

tantifs, elle est pauvre en verbes et vice-versa. Il nous apporte la preuve formelle de l'opposition des deux catégories qui semble s'accentuer, avec des écarts de plus en plus importants et des mouvements de plus en plus amples, au fur et à mesure que l'œuvre progresse et des écarts qui se révèlent très sensibles au genre littéraire[1].

Au début de la production de l'écrivain, dans sa période « nouveau roman », les deux courbes ne s'écartent point, elles se suivent au contraire, les deux catégories étant déficitaires dans cette partie de l'œuvre. C'est à partir de l'essai *L'extase matérielle* que l'opposition se déclare. Les écarts les plus importants – avec un déficit important de verbes et un grand excédent de substantifs – sont à trouver dans les ouvrages d'ethnologie et dans les essais qui traitent du nouveau monde, comme *Le rêve mexicain* ainsi que dans la biographie *Diego et Frida*. *Poisson d'or* est le seul roman de cette époque qui présente un écart d'une grande amplitude, mais l'écart cette fois-ci témoigne d'un déficit important de substantifs et d'un excès de verbes.

Dans les œuvres non fictionnelles – les ouvrages ethnologiques, les essais, les récits de voyage et la biographie – l'évolution de l'opposition entre la catégorie du substantif et celle des verbes est en effet assez spectaculaire. Au début, les substantifs sont déficitaires et les verbes excédentaires, mais assez vite les rôles s'inversent et l'écart s'amplifie de façon importante. Il est difficile de fournir une explication précise, mais à un moment qui correspond à la découverte de la culture amérindienne et mexicaine, capitale pour notre écrivain, les substantifs commencent à abonder tandis que les verbes diminuent de façon considérable. Peut-être, à partir de ce moment, n'y a-t-il plus besoin du mouvement, des dialogues ni des verbes (d'action ou de parole) ; il suffit de regarder et Le Clézio observe, décrit et partage ce qu'il voit avec ses lecteurs en recourant à de nombreux substantifs.

5. Le syntagme nominal

Nous avons déjà pu constater l'ampleur de la catégorie nominale dans notre corpus, un phénomène qui est constaté dans pratiquement toutes les études statistiques sur le vocabulaire français[2]. L'histogramme ci-dessous rend compte de sa distribution relative à travers l'œuvre :

Nous pouvons observer que la fréquence des substantifs à l'intérieur du corpus n'est pas constante. Le graphique rend compte de deux dynamiques différentes : la différenciation générique et l'évolution dans le temps. Concernant l'opposition des genres littéraires, la tendance générale est celle d'un substantif déficitaire dans les romans tandis qu'il est excédentaire dans les ouvrages d'ethnologie, dans la biographie et dans les récits de voyages. En revanche, dans les livres qui favorisent le dialogue, et par conséquent une langue plus orale, le substantif est déficitaire. Des livres comme *Printemps et autres saisons* et *Poisson d'or* ne sont pas seulement riches en dialogues mais sont également des livres où l'action prime sur la description.

[1] Le quotient est le rapport entre les deux séries. Il permet de voir comment se séparent les parallèles quand deux séries sont liées et parallèles. Comme les deux séries peuvent avoir un poids très inégal, la seconde est d'abord ramenée à la dimension de la première, proportionnellement, pour que le total des deux séries soit le même. Le quotient est calculé ensuite terme à terme, et s'équilibre nécessairement autour de la valeur 1.

[2] Selon Marc Hug (1989 : 5) : « Le syntagme nominal est l'unité qui s'impose d'emblée quand on commence l'étude statistique : non seulement c'est l'unité la plus abondamment représentée dans tout énoncé, et une unité très multiforme, mais c'est en même temps une unité fragmentaire, qui peut être étudiée indépendamment de la structure d'ensemble de la phrase… »

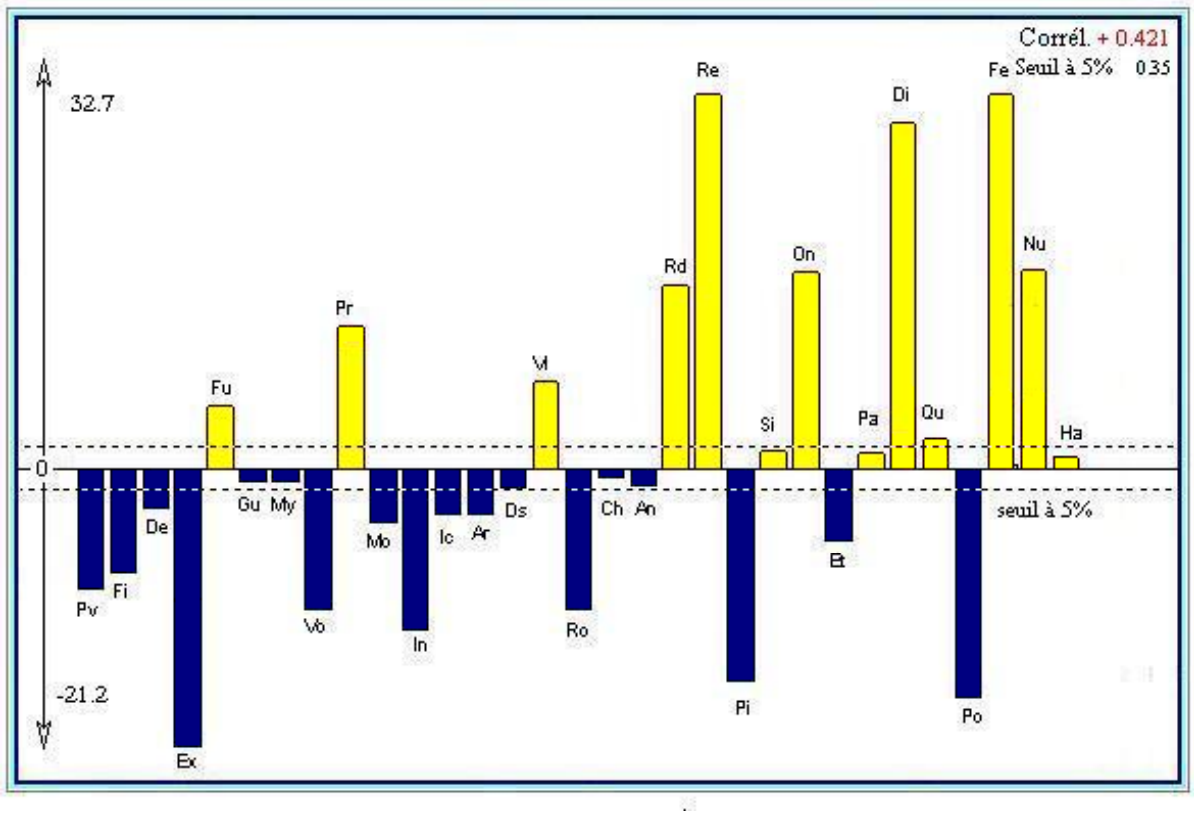

Figure 3. La distribution des substantifs dans le corpus(écarts réduits).

Ces résultats sont confirmés par les constats qu'a faits Étienne Brunet à propos du vocabulaire français de 1789 à nos jours (1981 : 304) concernant « le style substantif ». Le langage technique et le discours intellectuel (que l'on trouve dans les essais et dans les ouvrages ethnologiques de Le Clézio) utilisent plus volontiers les substantifs car ils sont plus adaptés aux définitions abstraites et « s'enchaînent avec une grande rigueur logique dans des phrases reflétant sous une forme très ramassée une pensée complexe. »

Il est intéressant de comparer cette analyse avec celle de la longueur de la phrase que nous avons effectuée sur ce même corpus (Cf. Kastberg Sjöblom, 2002a : 230-240). Les deux histogrammes sont pratiquement superposables : celui de la distribution des substantifs et celui de la longueur de la phrase. Cela confirme ce que nous avons pu constater auparavant, c'est-à-dire que la construction de la phrase longue chez Le Clézio repose souvent sur l'énumération et l'accumulation, et ce sont, bien évidemment, des substantifs qui prolifèrent dans ce système. Nous notons également chez Le Clézio la même corrélation entre la fréquence du substantif et l'accroissement lexical ainsi que la richesse du vocabulaire qu'a déjà trouvée Charles Muller (1967 : 111-115) dans ses études sur le vocabulaire de Corneille, où les pièces les plus riches en substantifs avaient en général aussi un vocabulaire étendu, bien qu'il y eût des exceptions à cette règle.

La division de la catégorie de substantifs en genres, c'est-à-dire en masculin et en féminin, pourrait nous apporter plus de précisions sur l'usage du nom dans l'œuvre de Le Clézio[3].

[3] Dans les études statistiques s'appuyant sur des corpus non lemmatisés, l'analyse du genre masculin et féminin se faisait surtout à partir de l'étude des articles, c'est-à-dire *la*, *le*, ou *une*, les formes *l'* et *les* ne disant rien sur le genre. Or ces formes sont ambiguës et des confusions avec les numéraux aussi bien qu'avec des pronoms personnels rendaient la statistique moins fiable. Les formes contractées au et du étaient délicates à traiter et le phénomène de la neutralisation de l'opposition du genre masculin-féminin dans les formes *l'*, *les*, *aux* et *des* ne facilitait certes pas la tâche.

Désormais, le logiciel Hyperbase dans sa version lemmatisée permet d'extraire les noms féminins et les noms masculins distinctement grâce à l'analyseur incorporé. Si nous divisons la catégorie des substantifs en genres, nous pouvons constater que le masculin domine avec 56,5 % des effectifs, le féminin n'en comportant que 43,5 %. L'analyse de la distribution relative permet de constater que la distribution des deux genres est relativement parallèle et qu'il n'y a pas d'opposition franche entre le masculin et le féminin (coefficient de corrélation significativement positif (+0.60)).

À la comparaison avec les valeurs obtenues lors de l'analyse des deux articles *le* et *la*[4], nous pouvons constater que l'article *le* emporte sur l'article *la* avec 57 % contre 43 %, c'est-à-dire un résultat quasi identique à celui du masculin et du féminin. Les courbes obtenues font preuve d'un parallélisme encore plus fidèle que celles des substantifs, avec un coefficient de corrélation de +0,749. Dans des études antérieures on avait presque toujours pu observer que l'article féminin l'emportait sur le masculin avec des proportions autour de 55 % contre 45 % (Brunet, 1988 ; Engwall, 1982)[5]. La préférence pour l'article féminin est encore plus marquée dans *Frantext* avec 56 % contre 44 % pour le masculin. La prépondérance de l'article *le* par rapport à *la* semble vraiment être une caractéristique de notre écrivain.

Comment interpréter ces résultats ? Nous savons qu'en français l'opposition des genres n'a pas de motivation précise. Toutefois dans certains substantifs on peut observer une régularité en genre due à la présence de suffixes. Les suffixes *-tion*, *-té* ainsi que *-ade* marquant une action ou une qualité imposent le féminin, tandis que *-ment* et *-isme* exigent l'article masculin. Étienne Brunet écrit ceci à propos de la prépondérance habituelle de l'article féminin (1988 : 193) : « L'abstraction intéresse plus l'article défini que l'indéfini (on dit moins souvent une justice que la justice) et concerne davantage le féminin que le masculin. […] Le rapport *la*/*le* serait donc une mesure indirecte du degré d'abstraction d'un texte ou d'un corpus. »

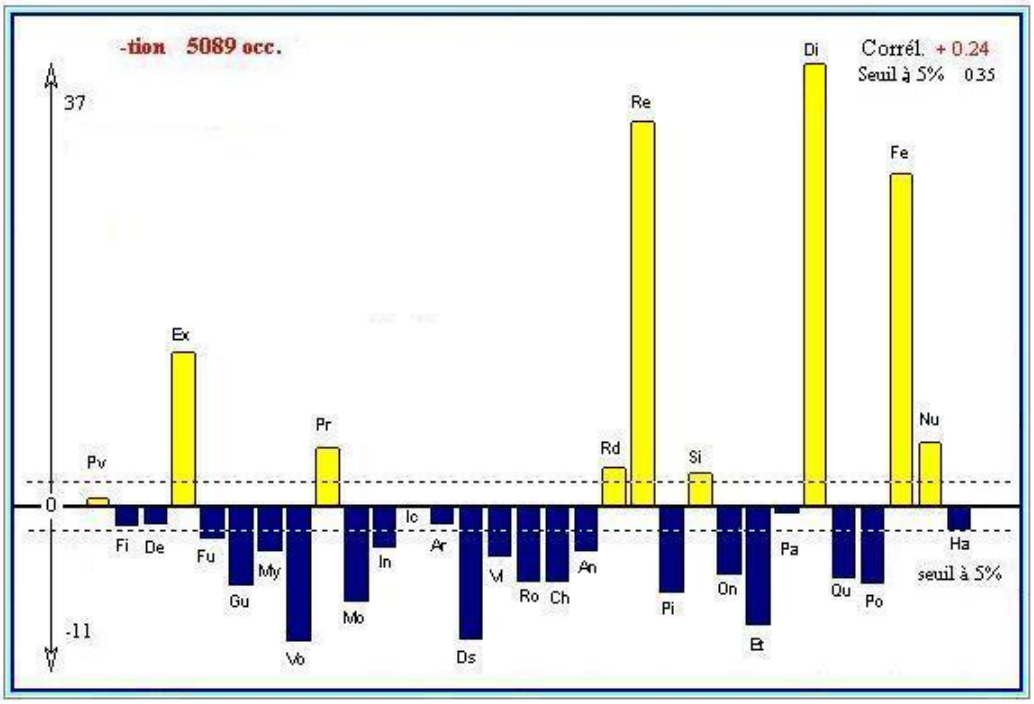

Figure 4. La distribution relative du suffixe -tion *(écarts réduits).*

[4] La forme unique du pluriel est exclue de l'analyse, ne permettant pas la distinction du genre en français.

[5] Rappelons qu'il s'agit dans son étude de formes graphiques et qu'il faut en tenir compte lors de la comparaison.

Les valeurs excédentaires du suffixe *-tion* dans notre corpus se trouvent dans les ouvrages ethnologiques et dans les essais où effectivement plus de place semble être donnée à l'abstraction et au discours intellectuel.

Dire à partir de ces analyses que Le Clézio fuit l'abstraction dans son œuvre romanesque serait sans doute une surinterprétation. Mais il y a chez notre auteur une volonté manifeste de rester dans le concret et de garder un langage simple et « terre à terre » ; il y revient souvent dans les divers entretiens et même dans ses textes et l'analyse statistique le confirme. Toutefois, l'histogramme de l'article *la* seul témoigne d'une tendance croissante vers l'abstraction, avec un coefficient de corrélation significativement positif :

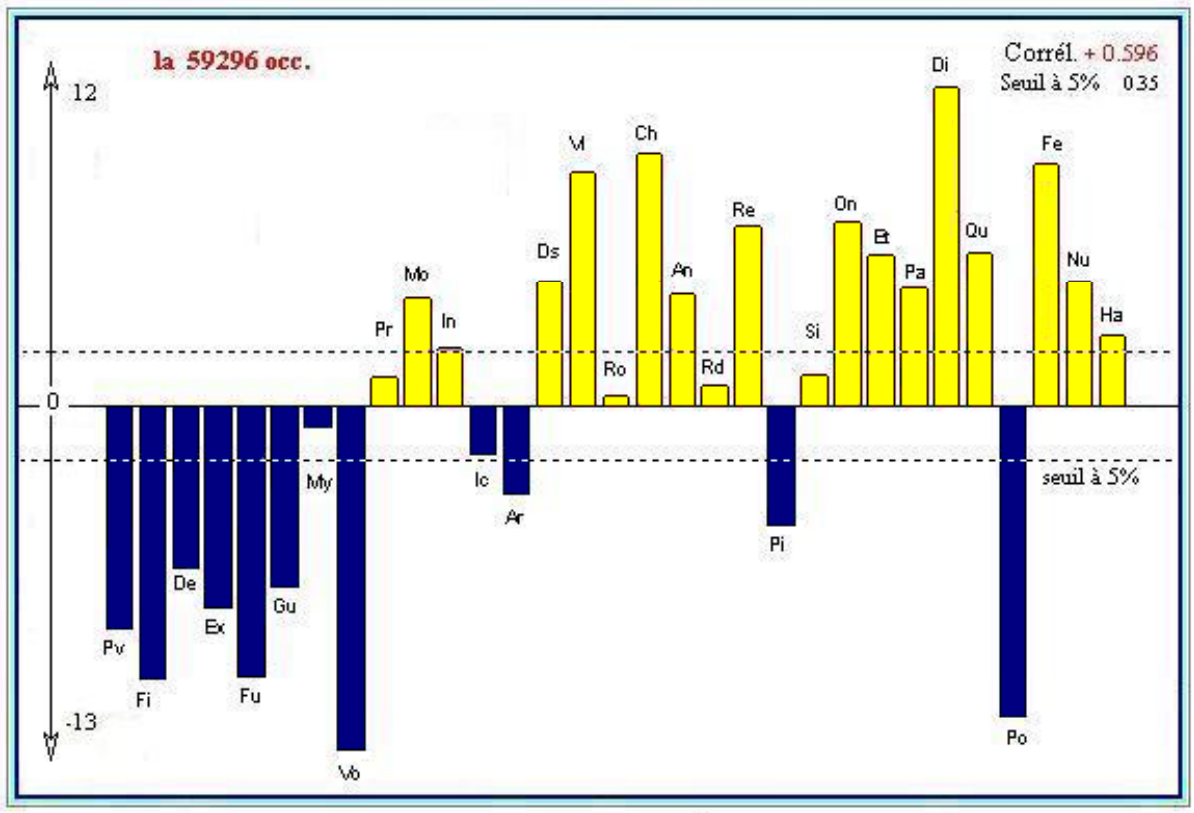

Figure 5. La distribution relative de l'article la dans le corpus (écarts réduits).

Que la première période « nouveau roman » soit largement déficitaire n'est pas étonnant avec cette écriture particulière qui ne favorise pas l'abstraction. Bien au contraire, elle est favorable au concret, car c'est une réalité bien concrète, la société de consommation avec ses attributs et l'envahissement d'objets, que Le Clézio dénonce durant cette période. Par la suite, l'écriture semble tendre de plus en plus vers l'abstraction dans tous les genres littéraires (accentuée dans les ouvrages ethnologiques, les essais et la biographie), à l'exception des livres riches en dialogues et en oralités comme *Voyage au pays des arbres*, *Printemps et autres saisons* et *Poisson d'or*.

6. Conclusion

Cette étude qui ne prétend qu'à donner quelques exemples d'utilisation des codes grammaticaux dans l'étude stylistique et typologique de textes et ne peut pas – faute de place – traiter tous les aspects sous lesquels on pourrait étudier les parties du discours et les éléments syntaxiques en exploitant cette technique. Mais elle aura montré qu'un excédent de l'une ou

l'autre catégorie dans un texte donné peut avoir plusieurs causes et produire de multiples effets stylistiques.

Malgré les difficultés rencontrées dues à l'opacité relative de l'analyseur Cordial, notamment à la façon dont il tranche entre le singulier et le pluriel du substantif et le fait qu'il faut admettre une certaine marge d'erreur ou d'incertitude sur les chiffres, ces doutes ne mettent pas en cause les grandes tendances qu'au contraire seule l'analyse statistique d'un grand corpus lemmatisé peut mettre en évidence.

En effet, le profil grammatical de J.M.G. Le Clézio qui émerge de nos différentes analyses est celui d'une écriture globalement marquée par une forte présence de la catégorie nominale mais qui change, qui évolue dans le temps et dont la richesse s'exprime dans plusieurs genres littéraires ; les variations génériques se reflètent en effet dans presque toutes les analyses que nous avons effectuées.

L'analyse lexicométrique et « grammatico-métrique » permet de mettre à jour ces procédés peu conscients que l'on a comparés aux empreintes digitales d'un écrivain. Les empreintes digitales pourtant ne changent pas avec le temps, au contraire, elles demeurent à jamais les mêmes. Notre étude montre que les procédés langagiers peuvent changer, même de façon significative, à l'intérieur d'une œuvre. Le jeune écrivain débutant ne fait pas forcement le même emploi des catégories grammaticales qu'il fera au sommet de sa production ou bien vers la fin de sa carrière et ce phénomène semble très accentué dans le cas de Le Clézio.

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. Dans l'œuvre de Le Clézio les procédés morphosyntaxiques ne sont pas statiques, les techniques d'expression évoluent et sont constamment mises en question.

La distinction de typologies de textes n'opère, nous semble-t-il, ni à un niveau conscient lors de la production, ni à un niveau interprétatif. L'opération de classification par laquelle un lecteur donne une certaine cohésion à une suite textuelle est une opération de lecture-interprétation qui confère au discours une certaine structure compositionnelle opèrant au niveau le plus profond, régi par la finalité des textes. Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement au niveau pour ainsi dire « atomique » et cela transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

Références

Biber D., Conrad S. et Reppen R. (1998). *Corpus linguistics, Investigating Language, Structure and Use*. Cambridge Approaches to Linguistics.

Brunet Ét. (1981). *Le vocabulaire français de 1789 à nos jours*. Champion – Slatkine.

Brunet Ét. (1988). *Le vocabulaire de Victor Hugo*. Champion-Slatkine.

Engwall G. (1982). Le vocabulaire dans les best-sellers des années 1960. In *Actes du 2° colloque de lexicologie politique*, Saint-Cloud, 15-20 septembre 1980, vol (1). Klincksieck.

Guiraud P. (1954). *Les caractères statistiques du vocabulaire : essai de méthodologie*. PUF.

Hug M. (1989). *Structures du syntagme nominal français, Étude statistique*. Champion-Slatkine.

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

Kastberg Sjöblom M. (2002b). Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus. In *Actes des JADT 2002 :* 391-402.

Kastberg Sjöblom M. (2003). Analyse lexicométrique de l'opposition générique dans une perspective endogène. In Williams G. (Ed.), *Actes des IIIèmes Journées de la linguistique de corpus*. Presses Universitaires de Rennes.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. CERAT. Cahier n° 7.

Malrieu D. et Rastier F. (2002). Genres et variations morphosyntaxiques. In Angel Martin Municio (Ed.), *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.

Muller Ch. (1967). *Le vocabulaire du théâtre de Pierre Corneille, Étude de statistique lexicale*. (réédition Slatkine 1993).

Picard J., Pibarot A. et Labbé D. (1995). Un outil de statistique textuelle : le lemmatiseur. In *Travaux scientifiques C.R.S.S.A*., vol. (16) : 395-396.

Annexe 1

Genres	Titre	No	Genres	Titre	No
nouveau roman	Le procès-verbal	1			
	La fièvre	2	**récits poétiques**	Mydriase	7
	Le déluge	3		Vers les icebergs	12
	Le livre des fuites	5			
	La guerre	6	**essais littéraires**	L'extase matérielle	4
	Voyages de l'autre côté	8		L'inconnu sur la terre	11
				Trois villes saintes	15
romans traditionnels	Désert	14		Le rêve mexicain	20
	Le chercheur d'or	17			
	Angoli Mala	18	**ouvrages ethnologiques**	Les prophéties de Chilam Balam	9
	Voyage à Rodrigues	19		Sirandanes	22
	Onitsha	23		La fête chantée	29
	Etoile errante	24			
	La quarantaine	27	**enfant et jeunesse**	Voyages au pays des arbres	13
	Poisson d'or	28		Pawana	25
	Hasard	31			
			récits de voyages	Gens des nuages	30
nouvelles	Mondo et autres histoires	10			
	La ronde et autres histoires	16	**biographies**	Diego et Frida	26
	Printemps et autres histoires	21			