L'indice pronominal est-il encore d'actualité ?

Margareta Kastberg Sjöblom
ILF-CNRS
Bases, Corpus et Langage (UMR 6039)
UFR Lettres, Arts et Sciences Humaines
98, bd É. Herriot. B.P. 209
06204 Nice Cedex 3
kastberg@unice.fr

ABSTRACT.

KEYWORDS :

RESUME.

MOTS-CLES :

1. Introduction

Désormais la quantification et la lemmatisation des corpus ouvrent la voie à la composante essentielle de l'écriture qu'est l'étude de la distribution des parties du discours et l'analyse des catégories grammaticales. L'intégration de lemmatiseurs et d'analyseurs morphosyntaxiques dans les logiciels lexicométriques permet le recensement automatique et impartial de ces choix inconscients faits par l'auteur lors de la création et de l'élaboration d'un texte, et autorise l'évaluation des choix grammaticaux caractéristiques et personnels.

Pourtant, l'interaction de la statistique et de la stylistique n'est pas nouvelle. En 1962 déjà Charles Muller avait adopté une démarche permettant de construire des indices stylistiques à partir des fréquences d'index. Son enquête avait révélé "assez clairement" que la distribution des pronoms personnels et possessifs pouvait fournir des indications intéressantes et c'est suite à ces observations qu'il a mis au point un "indice du style familier" [Muller 1979], qu'il a nommé l'indice pronominal. Il avait en effet constaté que le rapport entre le nombre des pronoms personnels "de dialogue" ($1^{ère}$ et 2^{e} personne du singulier et du pluriel) et le nombre des possessifs (adjectifs et pronoms) des mêmes personnes variait de façon significative suivant les textes considérés. Le quotient entre les

première et deuxième personnes des pronoms personnels et les mêmes personnes des possessifs dans les pièces de théâtre qu'il étudiait, pouvait en effet lui donner une mesure de la valeur stylistique du texte, indiquant son appartenance à un langage familier ou à une langue plus soutenue[1]. "D'une façon générale, écrit-il, l'indice pronominal est élevé quand le style est familier, bas quand le style tend à la noblesse ou au lyrisme."

Bien que la pertinence de cet indice s'explique par des alternances typiques du théâtre classique ("*je* suis déshonorée" *vs*. "*Mon* honneur est perdu"), son emploi dans d'autres types de corpus s'est avéré également pertinent[2]. Charles Muller avait observé des valeurs variant pour la plupart entre 1,6 et 3 dans le théâtre classique ; par exemple, l'indice moyen pour *Le Cid* était de 1,64 et celui de *Nicomède* était encore en-dessous (1,07). Étienne Brunet a en revanche relevé des valeurs de 5,72 chez Zola, 4,08 chez Giraudoux, 3,56 chez Hugo et 2,63 chez Chateaubriand, auteurs du XIX^{ème}, dont on sait par ailleurs que le style est soutenu. Chez Proust il a observé un indice de 3,30 dans son premier texte passant à 5,03 dans son dernier.

Depuis lors la pertinence et l'efficacité opérationnelle de cet indice ont souvent été mises en question [cf. Malrieu 2001], compte tenu notamment de son caractère spécifique lié au style théâtral de grande éloquence (comme les alternances citées ci-dessus) ainsi qu'à l'utilisation dans le théâtre classique des nombreuses périphrases "précieuses" qui substituent une expression nominale, du type "mon cœur", ou "vos yeux", à une expression pronominale : "je" ou "vous", caractéristiques notamment de Molière. On notera cependant dès à présent que dans *Frantext* l'indice pronominal est de 4,33 pour la prose littéraire et de 2,004 pour la poésie. On retrouve donc là, en dehors du contexte théâtral, une différence de niveau d'indice qui semble significative.

Nous nous sommes donc posé la question suivante : la transposition de cet indice à d'autres genres littéraires que le théâtre est-elle possible ? Quel est aujourd'hui l'intérêt de l'étude de cet indice, qui a si souvent été critiqué pour ne pouvoir s'appliquer qu'au théâtre classique ? Et cette analyse revêt-elle un intérêt quelconque dans un contexte contemporain qui depuis longtemps a abandonné les expressions précieuses, et où la grande éloquence n'est pas seulement considérée comme "démodée" mais souvent même perçue comme étant "assez" ridicule ?

Pour répondre à ces questions, nous nous proposons ici de soumettre un corpus contemporain et "multigénérique" à l'analyse de l'indice pronominal, qui - compte tenu du contexte que nous venons d'exposer - nous paraît devoir être faite avec beaucoup de prudence afin d'en tirer, avec toutes précautions possibles, les conclusions auxquelles conduira son application au corpus qui nous occupe ici : l'œuvre de Le Clézio.

L'œuvre de J.M.G Le Clézio s'adapte en effet parfaitement pour répondre aux questions qui nous intéressent. Il s'agit d'un corpus tout à fait contemporain (il s'étend du 1963 à l'an 2000) et qui se décline dans plusieurs genres.

---

[1] L'indice P/p (pronoms personnels de 1ère et 2ème personnes / adjectifs et pronoms possessifs de 1ère et 2ème personne).

[2] Cf. les différentes études d'É. Brunet.

2. Le corpus

Le vaste corpus de 31 ouvrages englobant l'œuvre de Le Clézio totalise 51.009 vocables différents, soit 2.257.931 occurrences, et regroupe divers genres littéraires : nous y trouvons des romans, des recueils de nouvelles, des récits poétiques, des essais littéraires, des ouvrages ethnologiques, des récits de voyages, une biographie ainsi que des livres pour les enfants et la jeunesse[3].

Le corpus est constitué tout d'abord des six premières œuvres[4], classées, par leur style particulier et innovant, comme appartenant à l'École du "nouveau roman" : *Le procès-verbal*, les nouvelles de *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre* et *Voyages de l'autre côté*. Les romans qui suivent cette période, considérés par les critiques comme plus "traditionnels", sont au nombre de neuf : *Désert*, *Le chercheur d'or* et *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. Lorsque certaines critiques les rapprochent de la poésie en prose, d'autres parlent de textes anecdotiques. Le corpus inclut ensuite les recueils de nouvelles : *Mondo et autres histoires*, *La ronde et autres faits divers* ainsi que *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. La culture amérindienne est 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* ; sont présents enfin *Diego et Frida*, unique biographie, et *Gens des nuages*, récit de voyage[5].

Le corpus "Le Clézio" a été traité et quantifié avec le logiciel Hyperbase (version 5.5) et lemmatisé selon le programme Cordial 7 qui aboutit au bout du traitement à quelque 200 codes grammaticaux différents, en utilisant toutes les combinaisons possibles. Ces procédés permettent, de façon exacte et efficace, d'extraire et d'isoler les différentes parties du discours d'un texte. Les codes grammaticaux fournis par l'étiqueteur morphosyntaxique au cours de l'opération de lemmatisation "automatique" constituent ici un outil indispensable pour l'analyse [cf. Kastberg Sjöblom 2002 : 80 – 88], qui demande l'accès à la forme canonique du mot, au lemme, et qui ne peut guère se fonder sur la distribution des effectifs d'un corpus s'appuyant sur les formes graphiques. C'est en effet 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, ce qui nous permet ici d'extraire les pronoms personnels et les pronoms possessifs du corpus.

La riche variation typologique des textes de notre corpus et l'importance du genre littéraire dans l'analyse lexicométrique, déjà bien documentée par ailleurs [Malrieu et

---

[3] Cette classification assez schématique pourrait être discutable, étant donné que Le Clézio lui-même aime brouiller les genres.

[4] Tous les textes de Le Clézio cités dans cet article ont été publiés aux éditions Gallimard, Paris.

[5] Cf. en annexe n°1 le tableau récapitulatif dans lequel les ouvrages sont classés selon les genres littéraires.

Rastier 2002, Kastberg Sjöblom 2002, 2003], nous a incitée à diviser ce corpus "multigénérique" en deux : le corpus A qui englobe l'ensemble des 31 ouvrages, et un sous-corpus "romanesque" qui comprend exclusivement les romans et les recueils de nouvelles (le corpus B) et qui permet ainsi d'affiner l'analyse.

Afin de bien saisir la situation et le contexte de l'indice pronominal dans notre corpus, nous nous intéressons à la catégorie des pronoms dans son ensemble avant de nous tourner vers les résultats des calculs de l'indice dans notre corpus et vers les corrélations grammaticales dont il dépend.

3. Les pronoms

Les pronoms constituent, avec leurs 243.150 occurrences, presque 10% du corpus global (A). L'histogramme ci-dessous rend compte de la distribution relative de la catégorie à l'intérieur du corpus :

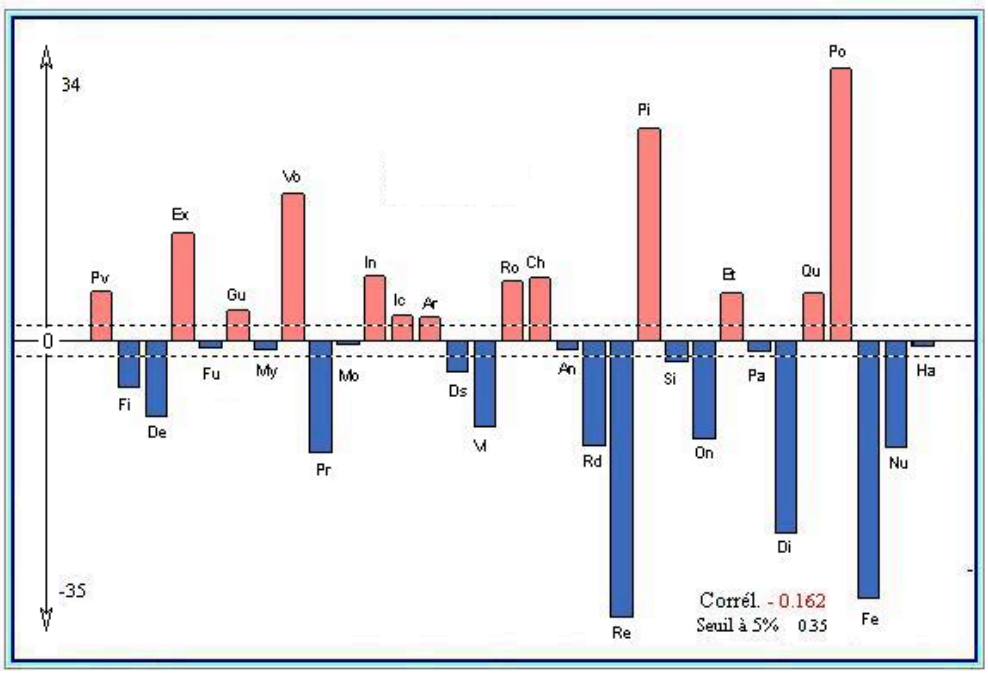


Figure n°1 : La distribution des pronoms dans le corpus A (écarts réduits).

Ce graphique permet de constater des écarts importants et alternés, donnant une ligne générale assez chaotique avec une distribution inversée par rapport à celle des substantifs [Kastberg Sjöblom 2002b, p. 310-314], notamment pour les écarts les plus grands (*L'extase matérielle*, *Voyages de l'autre côté*, *Prophéties de Chilam Balam*, *Voyage à Rodrigues*, *Le rêve mexicain*, *Printemps et autres saisons*, *Onitsha*, *Diego et Frida*, *Poisson d'or* et *La fête chantée*). Et les écarts autour de l'axe des abscisses, comme dans le graphique des substantifs, deviennent de plus en plus amples avec le temps, du côté négatif aussi bien que

du côté positif. Cette relation entre les deux courbes est intéressante, car s'il est vrai que le pronom est "mis pour" un nom, on voit ici qu'en réalité la distribution de ces deux catégories se fait de manière complémentaire et qu'elles n'apparaissent donc pas dans les mêmes contextes.

Les valeurs déficitaires sont à chercher dans les ouvrages ethnologiques, la biographie et les essais – à l'exception de *L'extase matérielle* et de *L'inconnu sur la terre* qui sont excédentaires en pronoms. Les déficits les plus importants se trouvent dans *Le rêve mexicain* et dans *La fête chantée*, avec des écarts réduits de -34,8 et de -32,3.

En revanche, la tendance générale des romans et des recueils de nouvelles est à la hausse avec pour point culminant *Poisson d'or* (écart réduit de +34,3). Toutefois, nous trouvons également des déficits importants – par exemple dans *Onitsha* – qui méritent d'être étudiés de près en prenant en compte uniquement l'œuvre romanesque de Le Clézio.

En effet, lorsqu'on regarde le même histogramme établi sur le corpus exclusivement romanesque, cette dynamique est plus aisément observable :

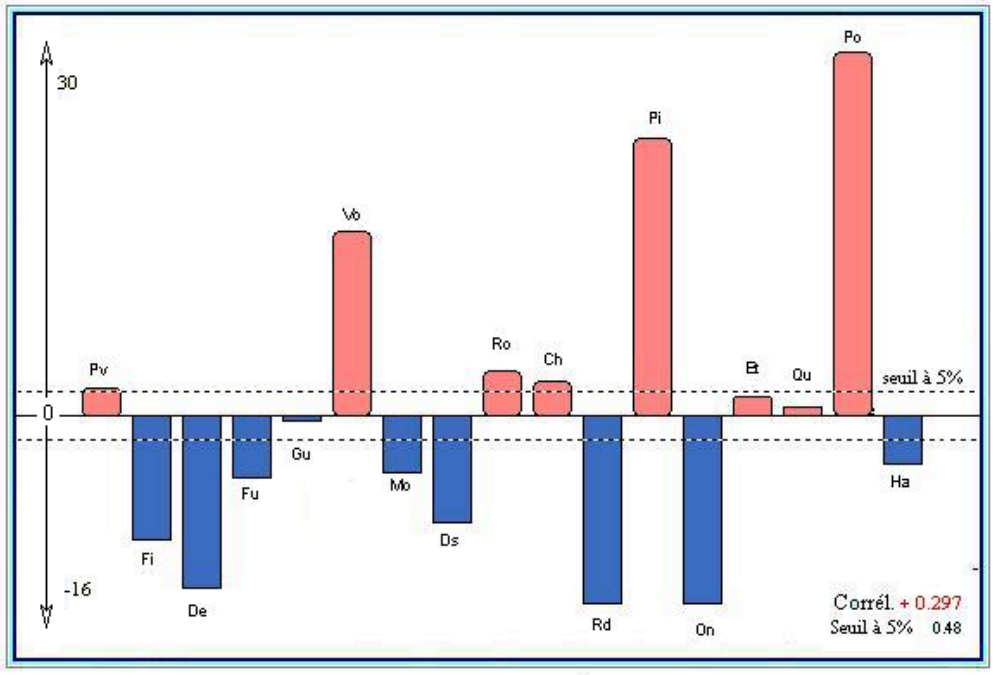


Figure n°2 : La distribution des pronoms dans le corpus B (écarts réduits).

La figure montre ici qu'il n'y a pas de tendance significative vers un usage plus prononcé du pronom chez Le Clézio. Bien que l'histogramme semble avoir une tendance à la hausse, les valeurs déficitaires des romans *Voyage à Rodrigues* et *Onitsha* sont trop importantes pour permettre des conclusions dans ce sens.

Pour ensuite diviser la catégorie des pronoms en sous-catégories, nous avons encore recours au logiciel Hyperbase qui, grâce à l'analyseur syntaxique Cordial, divise nos pronoms en sept sous-catégories : *pronoms personnels réfléchis*, *pronoms personnels non réfléchis*, *pronoms possessifs*, *pronoms démonstratifs*, *pronoms interrogatifs*, *pronoms indéfinis* et *pronoms relatifs*.

Cette classification permet de faire la même observation que dans presque toutes les autres études de même caractère : les plus grands nombres d'occurrences se trouvent dans les deux premières classes, celles des pronoms personnels. Pour mieux illustrer l'irrégularité de la distribution des différentes catégories, nous avons fusionné les deux premières sous-catégories en une seule : *pronoms personnels*. Le diagramme ci-dessous rend compte de la distribution des sous-catégories en ordre hiérarchique ; leur part dans la catégorie générale des pronoms est exprimée en pourcentage :

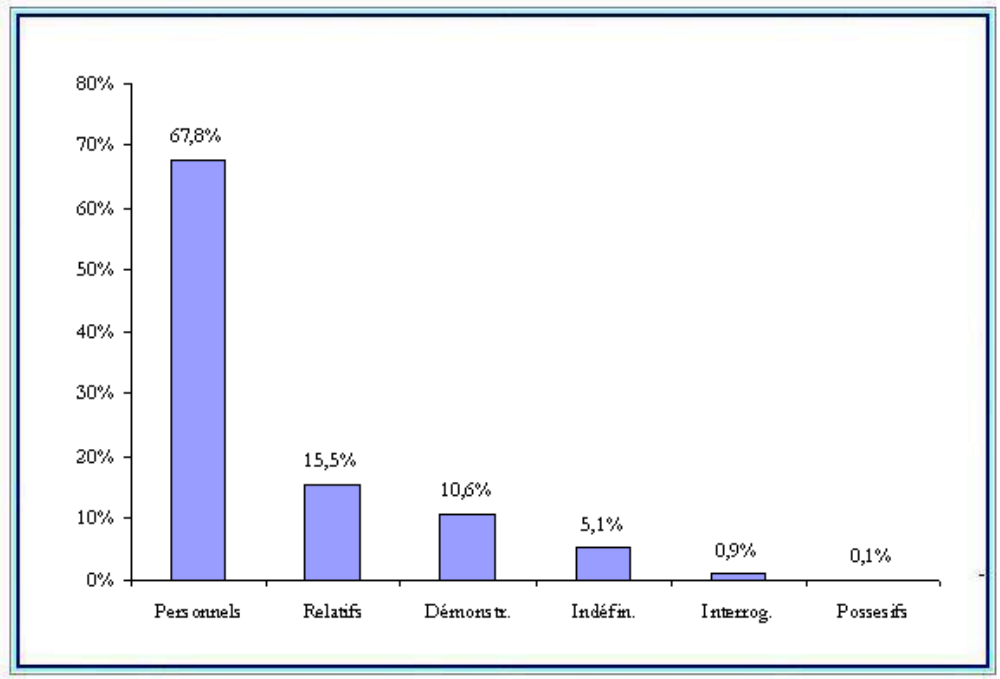


Figure n°3 : La proportion des différentes sous-catégories pronominales dans le corpus A (valeurs absolues).

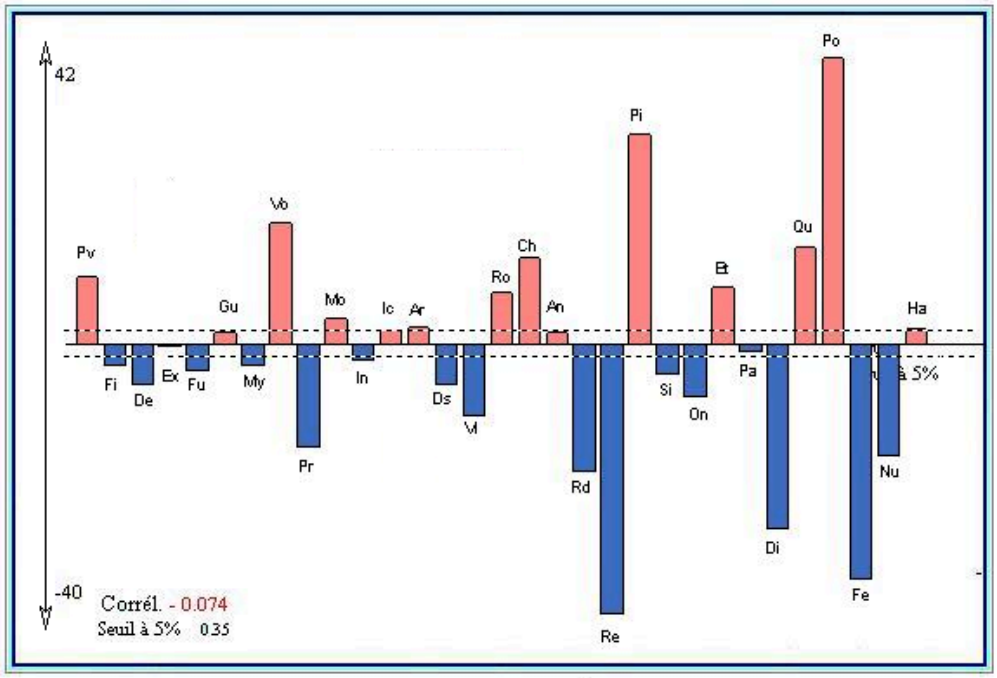
Nous constatons que la grande majorité des occurrences est fournie par les pronoms personnels qui, à eux seuls, occupent 67,8% du diagramme. Les pronoms relatifs en occupent 15,5%, les démonstratifs 10,6%, et les 6% restants sont partagés entre les autres pronoms (les pronoms possessifs n'occupant que 0,1% de l'effectif).

Il convient ici de rappeler que les pronoms personnels sont sujets à de nombreuses ambiguïtés. Une difficulté dans l'analyse statistique est leur homographie avec les articles :

le, *la*, *l'* et *les* sont facilement confondus avec les articles et les formes élidées sont souvent difficiles à désambiguïser.

Dans la littérature, étant donné qu'il s'agit souvent de fiction et de narration un autre phénomène rentre également en compte [Benveniste 1966, p. 251-257] dans la mesure où la situation énonciative est différente de celle du théâtre[6]. Certains phénomènes stylistiques notamment, dans la littérature de l'école "nouveau roman", mettent en jeu les conventions romanesques traditionnelles[7].

Néanmoins, malgré ces difficultés, les données numériques apportent quelques informations intéressantes. Si nous regardons la distribution des pronoms personnels dans notre grand corpus, nous pouvons observer une structure très semblable à celle de la distribution de l'ensemble des pronoms :



[6] Dans le genre romanesque en effet le *je* n'est pas le même qu'au théâtre - où le *je* désigne le locuteur qui s'adresse à *tu* allocutaire (ou parfois au public) ; dans le roman le *je* peut désigner l'instance énonciative du narrateur qui s'adresse à *tu* lecteur (une caractéristique de l'école du nouveau roman), ou bien il peut s'agir du *je* de chacun des personnages en situation de discours direct dans les dialogues [Malrieu, 2001, p. 5].

[7] Les auteurs de cette époque avaient l'habitude de "jouer" avec leurs lecteurs en les impliquant dans le récit. Maurice Nadeau (1992 : 184) écrit ceci à propos des innovations auxquelles se livre Le Clézio durant cette époque : "… une narration indirecte qui tourne brusquement à la confidence, substitution du "je" au "il" et au "nous", passages brusques chez le même individu de l'enfance à l'âge adulte et réciproquement, relations de situations invraisemblables (dans *Terra Amata*, Chancelade décrit sa propre mort et les sentiments qu'il éprouve en tant que cadavre)…"

Figure n°4 : La distribution des pronoms personnels dans le corpus A (écarts réduits).

Cette ressemblance n'a rien d'étonnant sachant que les pronoms personnels occupent pratiquement 70% de la catégorie. Nous trouvons les mêmes déficits et les mêmes excédents avec des ampleurs qui deviennent de plus en plus importantes au fur et à mesure que l'œuvre progresse. Il y toutefois quelques exceptions : les valeurs excédentaires de *L'extase matérielle* ont une autre source que celle des pronoms personnels. *Mondo*, déficitaire dans le premier histogramme, est ici légèrement excédentaire (écart réduit de +3,7), comme l'est aussi *Angoli Mala*.

Selon la formule de Ch. Muller, pour obtenir l'indice pronominal, l'effectif de cette importante catégorie de pronoms personnels doit ensuite être divisé par le nombre de pronoms possessifs. La distribution relative de la catégorie des possessifs dans notre corpus est la suivante :

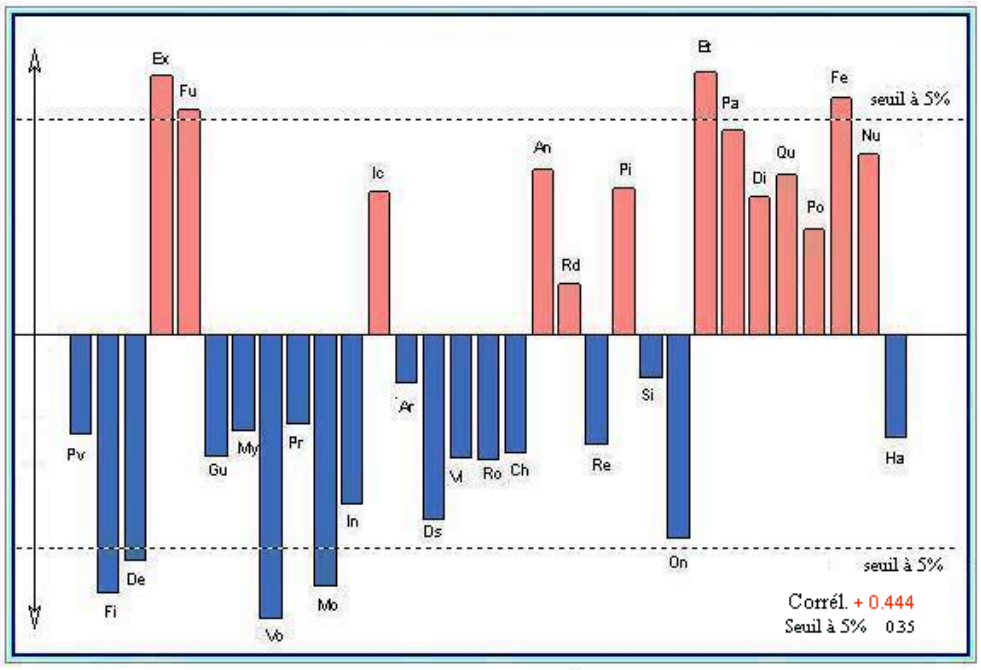


Figure n°5 : La distribution relative des pronoms possessifs dans le corpus A.

Cet histogramme permet de constater une tendance significative à la hausse avec une utilisation de plus en plus importante des possessifs au fur et à mesure que l'œuvre progresse (corrélation à 0,444 pour un seuil de significativité à 0,35). En dehors de cette tendance l'histogramme montre peu de valeurs significatives, un effet sans doute de la faible fréquence, et si l'on superpose les différents histogrammes, on ne trouve pas de relation - négative ou positive - observable entre les deux catégories de pronoms.

Quelle est donc la relation entre ces deux catégories et celle qui établit l'indice pronominal, indiquant un style familier ou soutenu ?

4. L'indice pronominal

Notre corpus A affiche un indice moyen de 5,81 qui indique un style relativement familier. Mais cette valeur moyenne mérite d'être détaillée et observée de plus près. Le tableau ci-dessous rend compte de cet indice pronominal dans les différents ouvrages du corpus :

	PERSONNELS			POSSESSIFS		Indice pronominal	
	1e pers.	2e pers.	Total	1e pers.	2e pers.	Total	Indice
Pv	1469	950	2419	135	103	238	10,16
Fi	1415	611	2026	162	52	214	9,47
De	1293	833	2126	154	86	240	8,86
Ex	2204	225	2429	610	54	664	3,66
Fu	1351	339	1690	198	40	238	7,10
Gu	1338	505	1843	122	83	205	8,99
My	4	10	14	0	6	6	2,33
Vo	758	1032	1790	222	117	339	5,28
Pr	23	1	24	5	1	6	4,00
Mo	412	399	811	59	59	118	6,87
In	1527	431	1958	267	91	358	5,47
Ic	29	16	45	5	0	5	9,00
Ar	4	15	19	1	0	1	19,00
Ds	438	256	694	208	47	256	2,71
Vi	53	32	85	32	9	41	2,07
Ro	571	181	752	72	10	82	9,17
Ch	5867	175	6042	1047	23	1070	5,65
An	109	116	225	15	11	26	8,65
Rd	489	17	506	269	4	273	1,85
Re	240	102	342	83	33	116	2,95
Pi	2666	239	2905	437	30	467	6,22
Si	27	1	28	9	0	9	3,11
On	239	199	438	34	44	78	5,62
Et	2301	216	2517	525	19	544	4,63
Pa	310	7	317	31	1	32	9,91
Di	369	112	481	113	24	137	3,51
Qu	4827	483	5310	816	58	874	6,08
Po	4439	393	4832	620	36	656	7,37
Fe	418	140	558	106	48	154	3,62
Nu	165	14	179	35	1	36	4,97
Ha	245	304	549	37	41	78	7,04
Tot.	**35600**	**8354**	**43954**	**6429**	**1131**	**7561**	**5,81**

Tableau n°1 : L'indice pronominal.

Le tableau permet en premier lieu d'observer l'importance du genre littéraire et de l'opposition typologique dans ce corpus. Les essais, les ouvrages ethnologiques et la biographie ont un indice très inférieur aux œuvres romanesques. *Mydriase*, le récit poétique, a une valeur très basse, comme l'essai *Trois villes saintes* où l'auteur insère, dans le récit très poétique, des extraits de textes sacrés maya ainsi que les mots des conquérants et des prêtres, comme ici Juan de la Cruz[8] :

"Ainsi donc, mes chers Chrétiens, fils des villages, gardez dans votre cœur mes commandements, parce que moi-même, mes enfants, je ne puis rester immobile, sans cesse je suis en route, ma gorge et mon ventre sont desséchés par une soif inextinguible, car sans cesse je suis en marche à travers le Yucatan pour vous défendre …"

Les autres ouvrages qui traitent du monde amérindien, par exemple *Le rêve mexicain* ou *La fête chantée*, sont aussi souvent riches en extraits de ce type, qui contribuent à donner une valeur très poétique, presque lyrique à ces textes bien qu'ils soient classés comme étant des ouvrages ethnologiques. Notons que le seul livre pour enfants du corpus, *Voyage au pays des arbres*, a un indice pronominal de 19,00 qui reflète bien son style particulier, "extrêmement familier" selon le barème de Muller.

É. Brunet avait déjà noté dans l'œuvre de Hugo d'amples variations selon les textes, avec des indices reflétant un style très relevé dans les recueils poétiques et une écriture beaucoup plus simple dans des romans comme les *Travailleurs de la mer* ; il en concluait que ces variations suivaient la loi du genre.

D. Malrieu [2001, p. 3] insiste aussi sur le fait "qu'il paraît plus intéressant de comparer des sous-corpus homogènes ou un texte à un corpus de son genre, qu'à un corpus, forcément aléatoire (ce qui n'est pas le cas de notre corpus) qui prétendrait être représentatif de la langue française et elle cite à ce propos Ch. Muller qui s'exprime à propos de son indice :

"On pourrait parler dans ce domaine de moyennes pour un genre bien délimité ou pour un groupe de textes très homogènes ; mais non d'une norme pour la langue, même bornée à une époque définie. Nous renoncerons donc d'emblée à comparer entre eux les chiffres obtenus dans différents textes."

Ch. Muller, tout comme D. Malrieu, se tient en effet à un raisonnement comparatif soit entre genres dans un même champ générique et une même période (comédie *vs.* tragédie classique), soit entre champs génériques (théâtre *vs.* poésie), etc. Cette importance de l'opposition générique nous incite à nous intéresser aussi au corpus B, relativement homogénéisé, qui regroupe l'œuvre romanesque de Le Clézio et ses nouvelles.

Les variations de l'indice pronominal à l'intérieur de ce corpus sont d'une très grande ampleur. La courbe ci-dessous en témoigne :

[8] *Trois villes saintes*, p. 55.

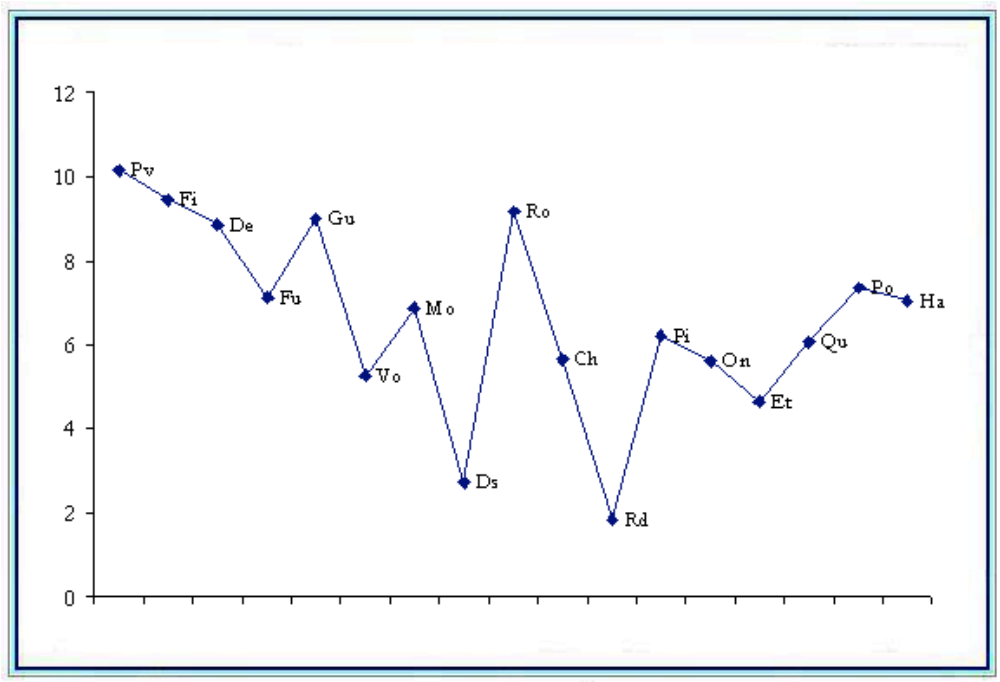


Figure n°6 : Courbe de l'indice pronominal du corpus romanesque (le corpus B).

Malgré quelques saccades de la courbe, nous pouvons observer un mouvement chronologique qui après des valeurs initiales élevées tend vers une baisse au milieu de l'œuvre, c'est-à-dire vers un style plus soutenu. Le Clézio semble laisser de côté le style très familier de ses débuts, puis – après cette période centrale au ton soutenu – revenir vers la fin de son œuvre à un style moins soutenu, sans pour autant atteindre les niveaux initiaux.

La courbe permet en effet d'observer un indice très élevé au début de l'œuvre dans les ouvrages "nouveau roman". C'est le premier livre du corpus *Le procès-verbal*, mais également *La guerre*, *La ronde* et *Poisson d'or*, qui ont les indices les plus élevés du corpus. Or c'est dans ces livres aussi – à l'exception de *Poisson d'or* – que l'usage de la deuxième personne est excédentaire. Nous avons montré [Kastberg Sjöblom 2002 ; 459-465] que ces récits ont en commun de donner la parole à des personnages adolescents qui s'expriment dans une langue de jeunes teintée de langue orale. L'extrait ci-dessous témoigne de ce style[9] :

"Puis d'un seul coup cela s'en va, et c'est elle maintenant qui parle, la voix un peu rauque, sans savoir très bien ce qu'elle dit. "Bon. Alors, on y va ? On y va maintenant ?"

Le garçon descend de son vélomoteur. Il embrasse Titi sur la bouche, puis il s'approche de Martine qui le repousse avec violence.

[9] *La ronde et autres faits divers*, p. 14.

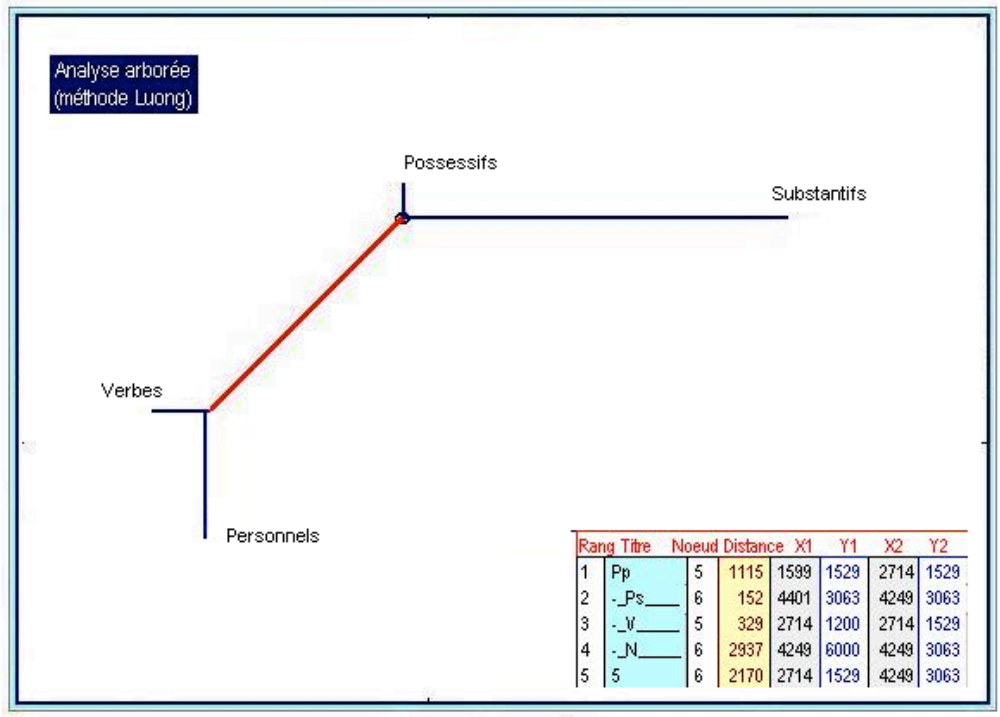
"Allez, laisse-la."

Les valeurs les plus basses de l'indice, témoignant d'un style soutenu ou lyrique (selon la qualification proposée par Ch. Muller), sont à trouver dans *Désert* et dans *Voyages à Rodrigues* mais pour des raisons diverses. L'indice de 2,71 est remarquable, bien qu'il n'étonnera guère le lecteur leclézien connaissant la veine poétique du roman *Désert* pour lequel Le Clézio a reçu de nombreuses récompenses. Quant à *Voyages à Rodrigues*, les raisons de l'indice très faible sont non seulement à chercher dans le style très particulier, déjà décrit auparavant, de ce livre, mais également dans le fait qu'il s'agit du récit à la première personne de l'aventure du grand-père de Le Clézio qu'il nomme "mon grand-père"[10] :

"La présence de mon grand-père dans ce lieu solitaire, c'est cela qui me trouble, me retient. C'est mon unique lien avec lui, car je ne sais rien de lui, hormis ces papiers et quelques photos jaunies."

La particularité grammaticale de ce livre est par ailleurs [Kastberg Sjöblom, 2002] celle de se montrer très excédentaire en substantifs et très déficitaire en verbes. Déjà Ch. Muller s'est posé, au moment de la présentation de l'indice, la question [1979 : 122] de savoir s'il existe une corrélation entre pronoms personnels et verbes d'une part, et pronoms possessifs et substantifs d'autre part. L'indice pronominal semble en effet avoir un rapport avec la proportion des substantifs et des verbes dans notre corpus : lorsqu'un texte est riche en substantifs il a un indice pronominal bas et *vice versa* comme le montre l'analyse arborée ci-dessous :

[10] *Voyage à Rodrigues*, p. 38.



Rang	Titre	Noeud	Distance	X1	Y1	X2	Y2
1	Pp	5	1115	1599	1529	2714	1529
2	-_Ps____	6	152	4401	3063	4249	3063
3	-_V_____	5	329	2714	1200	2714	1529
4	-_N_____	6	2937	4249	6000	4249	3063
5	5	6	2170	2714	1529	4249	3063

Figure n°7 Analyse arborée de l'opposition des catégories grammaticales.

La fréquence totale des pronoms d'un texte, qui va de paire avec celles des verbes, se reflète également sur les résultats de l'indice. Les textes riches en verbes, pleins d'action, sont aussi souvent riches en dialogues et en oralité, produisant un indice pronominal élevé. En revanche, il s'agit, pour les valeurs basses, aussi souvent d'un effet du genre littéraire que d'un signe de noblesse ou de lyrisme. Les possessifs (du moins les adjectifs qui sont très largement majoritaires dans cette catégorie) accompagnent nécessairement un substantif. L'exemple ci-dessous tiré de *L'inconnu sur la terre* illustre bien cette conjonction, cette affinité entre les diverses catégories morpho-syntaxiques[11] :

"Et la parole venue de l'espace, qui va vers l'espace, me traverse. Elle me donne mon vrai nom, ma vraie pensée, mon unique langage. […] et mon corps et mon esprit sont une mécanique dont tous les rouages s'entraînent, et j'avance, je respire, je vis."

5. La corrélation entre verbes et substantifs

On observe souvent dans un corpus clos, comme nous pouvons le faire dans le corpus Le Clézio, 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

[11] *L'inconnu sur la terre*, p. 138.

adverbes) s'opposent à la classe nominale qui réunit autour du substantif les adjectifs, les déterminants, les prépositions et souvent les coordinations.

Le graphique ci-dessous illustre comment les substantifs et les verbes varient les uns par rapport aux autres dans les différents livres du corpus. Les courbes s'appuient sur les écarts réduits des deux catégories grammaticales[12].

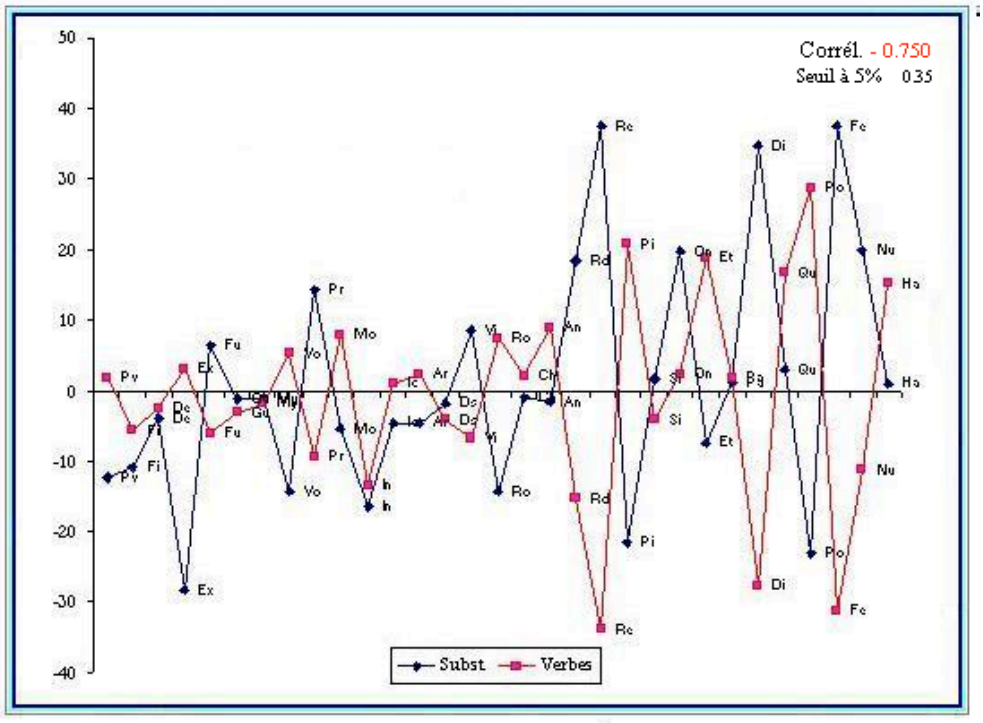


Figure n°8 : Comparaison de la distribution interne des substantifs et des verbes.

Le graphique rend compte 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. Ces courbes permettent en effet de voir de près ce que l'analyse factorielle nous a déjà indiqué par ailleurs [Kastberg Sjöblom, 2002] : lorsqu'une œuvre est riche en substantifs, elle est pauvre en verbes et *vice-versa*. Le coefficient de corrélation significativement négatif (-0,75) nous apporte la preuve formelle de ce

[12] Les écarts réduits sont reportés sur le graphique dont l'axe horizontal est formé par la moyenne sur tout le corpus, du moins comprise dans un intervalle de fluctuation "normale" de ± 2 écarts autour de l'axe. S'ils n'avaient de fluctuation stylistique significative (du point de vue probabiliste), tous les points seraient donc confondus avec l'axe horizontal.

phénomène que souligne encore, ci-dessous, l'histogramme du quotient entre les 459.957 substantifs et les 321.108 verbes, qui se révèle très sensible au genre[13] :

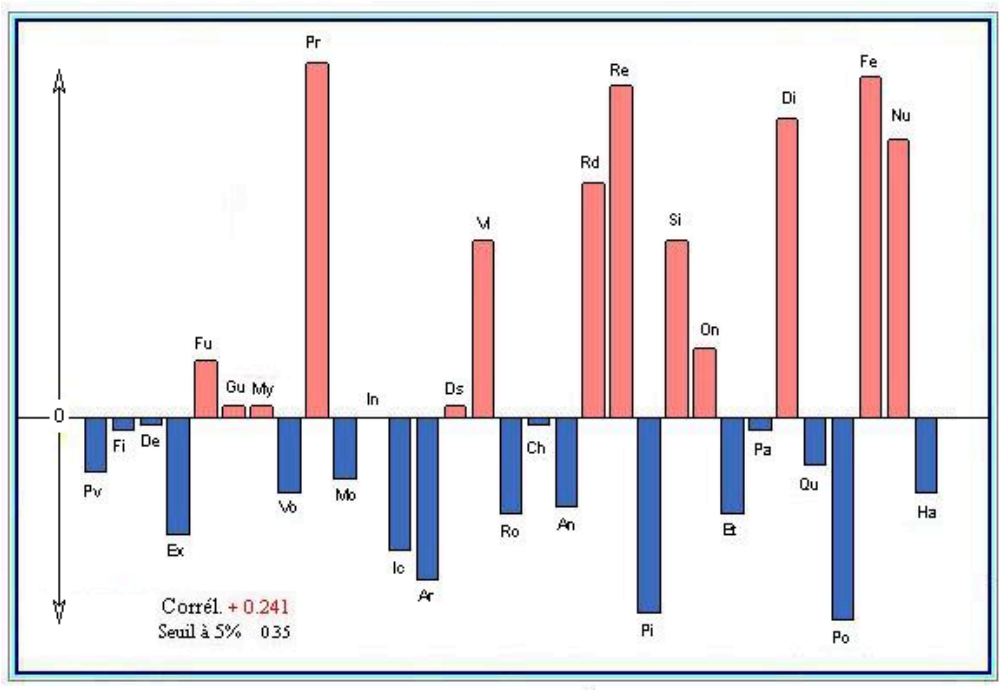


Figure n°9 : Histogramme du quotient substantifs/verbes (corpus B).

Toutefois, dans le graphique n°9 nous constatons qu'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 deux courbes deviennent parallèles dans *La guerre* et dans *Mydriase* pour se séparer de nouveau à partir de *Voyages de l'autre côté*. Dans les romans et dans les recueils de nouvelles qui paraissent entre 1975 et 1986, l'opposition des deux catégories est observable sans être très importante. 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

[13] 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.

excès de verbes. Ces deux dernières remarques confirment donc les observations de Pierre Guiraud [1954, p. 104].

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. Cette découverte capitale, Le Clézio veut en témoigner et il répète souvent : "Etre vivant c'est savoir regarder". Peut-être, à partir de ce moment, n'y a-t-il plus besoin du mouvement, des dialogues et 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.

Comme nous venons de le rappeler, la bipolarité que nous pouvons observer des catégories des substantifs et des verbes chez Le Clézio n'a rien d'originale : elle a été observée dans bien d'autres corpus : Étienne Brunet [1985, p. 155] l'a bien remarquée dans ses diverses études et il souligne également le rôle important de l'opposition des genres littéraires. De ce point de vue l'œuvre de Le Clézio s'inscrit tout à fait dans la dynamique générale de la littérature française.

Le graphique ci-dessous, qui isole l'œuvre romanesque, illustre encore plus nettement cette opposition qui s'amplifie vers la fin de l'œuvre, à l'exception du roman *La quarantaine* où les courbes tendent à converger :

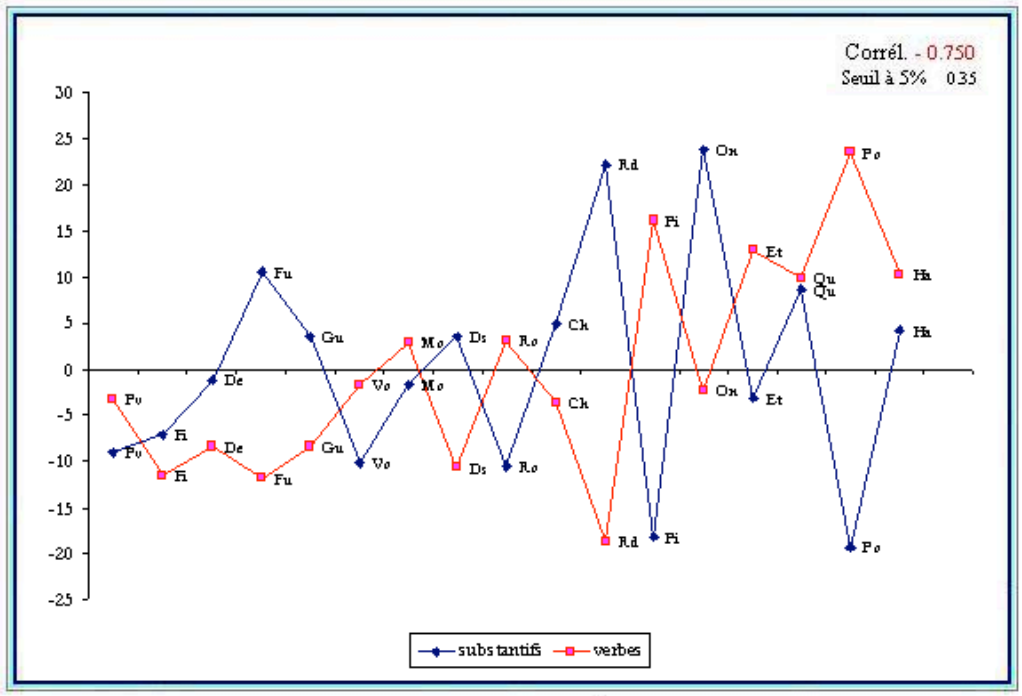


Figure n°10 : Comparaison de la distribution interne des substantifs et des verbes dans le corpus B.

Il s'avère que la courbe de verbes est pratiquement superposable à celle de l'indice pronominal et suit parfaitement le même mouvement. L'étroitesse de cette relation est remarquable et se pose inévitablement la question de l'effet de l'emploi du verbe sur cet indice stylistique.

6. Conclusion

Finalement, il s'est avéré que l'indice pronominal auquel nous croyions très peu après avoir pris connaissance des critiques émis sur son efficacité – sur d'autres genres que le théâtre classique – fonctionne parfaitement sur le corpus Le Clézio et fournit un complément intéressant à l'analyse stylistique.

Nous avons pu constater que la fréquence des pronoms, qui occupent environ 10% de notre corpus, est très variable. A l'intérieur de la catégorie des pronoms personnels nous avons aussi pu constater [Kastberg Sjöblom 2002, p. 353–356] la prédominance de la 3^{e} personne, qui occupe presque trois quarts des personnels. Le Clézio privilégie cette forme d'écriture, et l'usage de la 1^{e} et de la 2^{e} personnes est beaucoup plus restreint et bien limité à certains livres du corpus. C'est dans ces livres que nous avons relevé un indice pronominal élevé, témoignant d'un style encore plus familier que celui du corpus dans son ensemble.

Les pronoms étaient, nous l'avons vu dans les analyses factorielles de la distribution des parties du discours, étroitement liés à une autre catégorie grammaticale, celle des verbes :

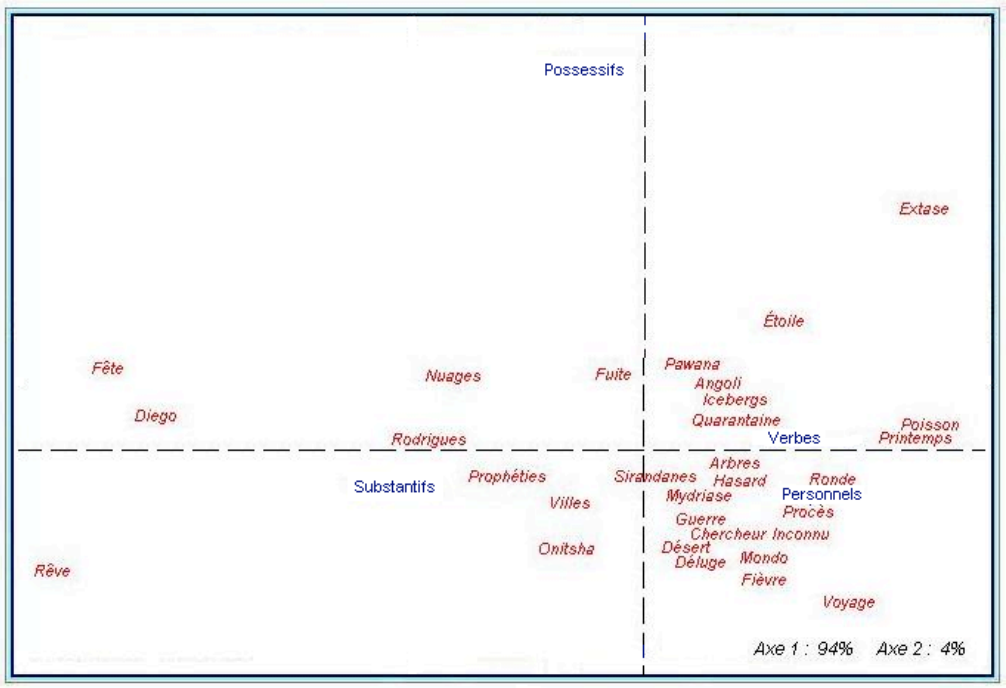


Figure n° 11 : Analyse factorielle de la distribution grammaticale dans le corpus A.

La mode n'est plus à la grande éloquence et par conséquent de chercher un indice servant à la mesurer dans des typologies textuelles différentes pourrait aujourd'hui sembler un peu hasardeux. Le remplacement du *je* et du *tu* ou du *vous* par des syntagmes nominaux à déterminants possessifs des personnes correspondantes serait, comme l'écrit D. Malrieu [2001, p.4], "à la fois l'expression d'un style précieux ou galant, qui s'exprime soit par des synecdoques corporelles euphémisantes soit dans les sphères des sentiments par une préférence pour l'expression indirecte ou voilée du vécu qui s'effectue à travers une nominalisation du subjectif et de la relation de l'interlocuteur".

Dans un contexte plus contemporain que celui du théâtre classique on pourrait néanmoins opposer des expressions telles que : "mon impression est que" *vs.* "j'ai l'impression que", "ton tort est de" *vs.* "tu as tort de", où la forme nominale sonne effectivement différemment de la forme verbale. Elle évoque en effet, comme le constate D. Malrieu, "un mode d'interlocution plus relevé, l'effacement de l'adresse personnelle 'inconvenante' au profit d'un discours pseudo-dégagé du *je* et du *tu* : elle évite en même temps la prise en charge des évaluations implicites et elle présente le locuteur comme capable d'expliciter en permanence son propre vécu, de l'objectiver, d'en faire une représentation pour l'interlocuteur, d'en faire un jeu social."

Il est vrai que l'indice pronominal a parfaitement fonctionné dans notre corpus où il souligne encore l'importance du genre et il nous semble que l'analyse trouve tout à fait sa

place dans une perspective typologique ; de plus l'interaction des prépondérances de certains catégories grammaticales vis-à-vis d'autres nous semble significative.

Ch. Muller insistait dans son premier article sur le fait que l'analyse stylistique reposait sur une analyse comparative des configurations des mots grammaticaux et non sur l'étude du lexique, comme elle l'avait été jusqu'alors. Le style s'analyse à travers les usages différentiels des catégories grammaticales, des structures syntaxiques, prosodiques et thématiques.

Il nous semble en effet que l'indice pronominal est bien plus lié au genre textuel qu'au niveau du style ; la corrélation pronoms personnels et verbes *vs.* pronoms possessifs et substantifs en est le fidèle témoin.

Les genres textuels tels que les ouvrages ethnologiques, les essais littéraires, les récits poétiques ont une très basse valeur d'indice pronominal, qui indique donc un niveau stylistique soutenu et élevé, mais par là même l'indice témoigne d'une richesse des substantifs au détriment de verbes, caractéristique d'un récit non favorable à l'action.

De la même façon, les valeurs élevées de l'indice que nous avons pu relever dans les romans témoignent aussi bien d'un récit de fiction, favorable à l'action, que d'un style familier.

Que mesure l'indice pronominal, en réalité ? Il nous semble bien qu'autant il est révélateur du niveau stylistique, autant il constitue une sorte de mesure de l'action dans une narration. L'indice pronominal serait donc aussi, indirectement, une mesure de l'action ou de la "non-action", étant des caractéristiques intimement liées au genre textuel et à leurs techniques narratives distinctes.

L'opposition de différentes typologies de textes se montre encore - comme elle le fait si souvent dans les analyses lexicométriques -, le facteur prépondérant et cet affrontement, au niveau stylistique aussi bien qu'au niveau thématique est régie par la finalité des textes.

Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement au niveau pour ainsi dire "atomique", ce qui transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

Références

ADAM Jean-Michel, *Les textes : Types et prototypes*, Paris, Nathan, 1992, coll. « fac. linguistique ».

BENVENISTE Emile, *Problèmes de Linguistique Générale*, Paris, Gallimard, 1966.

BRUNET Étienne, *Le vocabulaire de Zola*, Paris-Genève, Champion-Slatkine, 1985.

BRUNET Étienne, *Le vocabulaire de Victor Hugo*, Paris-Genève, Champion-Slatkine, 1988.

GUIRAUD Pierre, *Les caractères statistiques du vocabulaire*, Paris, puf, 1954.

KASTBERG SJÖBLOM Margareta, "Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus", in *JADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, Morin A., Sébillot P. (éds.), Saint-Malo, Irisa, Inria, 2002, p. 391-402.

KASTBERG SJÖBLOM Margareta, *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*, Nice, Université de Nice–Sophia Antipolis, 2002.

KASTBERG SJÖBLOM Margareta, "Analyse lexicométrique de l'opposition générique dans une perspective endogène" in *Actes des IIIèmes Journées de la linguistique de corpus*, septembre 2003, Lorient, Williams G. (éd.), Presses Universitaires de Rennes, en cours de publication.

MALRIEU Denise, "Stylistique et Statistique textuelle : A partir de C. Muller sur les « pronoms de dialogue » ", in *Texto !*, revue en ligne, revue électronique de sémantique des textes, F. Rastier (éd.), l'Institut Ferdinand de Saussure, Maison des Sciences de l'Homme, Paris, 2001, http://www.msh-paris.fr/texto/.

MALRIEU Denise et RASTIER François, "Genres et variations morphosyntaxiques", in *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*, Angel Martin Municio (éd.), Logroño, Fundación San Millán de la Cogolla, 2002.

MULLER Charles, *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977.

MULLER Charles, "Sur quelques scènes de Molière, essai d'un indice du style familier", in *Langue française et linguistiques quantitatives*, Genève, Slatkine, 1979, p. 107-124.

ONIMUS Jean, *Pour lire Le Clézio*, Paris, Collection Écrivains, puf, 1994.

RASTIER François, *Sens et textualité*, Paris, Hachette Supérieur, 1989

7. Annexe

Genres	Titre	No	Genres s	Titre	No
nouveau roman	Le procès-verbal (1963)	1			
	La fièvre (1965)	2	**récits poétiques**	Mydriase (1973)	7
	Le déluge (1966)	3		Vers les icebergs 1978)	12
	Le livre des fuites (1969)	5			
	La guerre (1970)	6	**essais littéraires**	L'extase matérielle (1967)	4
	Voyages de l'autre côté (1975)	8		L'inconnu sur la terre (1978)	11
				Trois villes saintes (1980)	15
romans traditionnels	Désert (1980)	14		Le rêve mexicain (1988)	20
	Le chercheur d'or (1985)	17			
	Angoli Mala (1985)	18	**ouvrages ethnologiques**	Les prophéties de Chilam Balam (1976)	9
	Voyage à Rodrigues (1986)	19		Sirandanes (1990)	22
	Onitsha (1991)	23		La fête chantée (1997)	29
	Etoile errante (1992)	24			
	La quarantaine (1995)	27	**enfant et jeunesse**	Voyages au pays des arbres (1978)	13
	Poisson d'or (1997)	28		Pawana (1992)	25
	Hasard (1999)	31			
			récits de voyages	Gens des nuages (1997)	30
nouvelles	Mondo et autres histoires (1978)	10			
	La ronde et autres histoires (1982)	16	**biographies**	Diego et Frida (1993)	26
	Printemps et autres histoires (1989)	21			