

Deuxième partie :

La structure du vocabulaire.

2.1. Introduction.

“Il y a devant moi un océan de mots libres, une plaine grise de langage qui avance, recule, avance, qui danse sur place. Il y a autour de ma bulle une montagne de gélatine violette qui tressaille tout au long de sa chair sans nerfs. Moi, je suis dedans.”

Livre des fuites, page 212.

La structure du vocabulaire d'un corpus comporte l'observation des fréquences et l'analyse des rapports que ces fréquences tissent entre elles. Elle est en principe indépendante des éléments qu'elle contient et ignore le contenu sémantique du discours. La lexicométrie fournit des méthodes qui permettent de procéder à des analyses statistiques portant sur le vocabulaire à partir d'une ou diverses segmentations du corpus.

L'étude la plus traditionnelle en lexicométrie est 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.

Ces recherches offrent la possibilité, indépendamment du contenu lexical, de situer, de distinguer et de comprendre la structure formelle des textes afin de pouvoir comparer, au niveau endogène, les parties de l'œuvre d'un écrivain ou de tout autre producteur de texte ou de parole aussi bien qu'au niveau exogène différents discours, genres, époques ou auteurs différents. Car, comme l'écrit J.M.G. Le Clézio¹⁵⁸ :

“Chacun a son aire, son langage, sa structure, sa vue. Le monde des araignées n'est pas le monde des mouches, le monde des oiseaux n'est pas celui des chenilles. Le monde des crustacés n'est pas le monde des

¹⁵⁸ *L'inconnu sur la terre*, p. 36.

mollusques. Le monde des éléphants n'est pas le monde des tigres, ni celui de l'herbe des savanes.”

Afin de mieux connaître la structure du vocabulaire de l'œuvre de Le Clézio, nous nous intéresserons dans cette partie à la distribution des fréquences, aux basses et aux hautes fréquences et à leurs rapports, à l'étude des hapax, à la richesse lexicale, à la diversité du vocabulaire ainsi qu'à l'accroissement lexical.

Il convient toutefois, avant de présenter les résultats de ces études, de bien définir la taille des différents sous-corpus et leur poids dans le corpus total afin de procéder aux pondérations nécessaires.

2.2. L'étendue du corpus et de ses variantes.

“Il entra dans l'écriture comme dans un paysage, sans chercher, sans s'attarder. Il vit les mots pleuvoir par phrases entières, filant rapidement vers la droite, pareils à de minuscules animaux.”

Le déluge, page 125

La fréquence absolue d'un vocable est le nombre de fois où il apparaît dans un texte ou un corpus, c'est-à-dire le nombre de ses attestations, son effectif. La fréquence relative est le rapport entre l'effectif absolu des attestations d'un mot et la taille du corpus.

Dans les études statistiques, pour effectuer des analyses quantitatives différentes notamment de la fréquence et des hapax, les fréquences absolues ne suffisent pas, il est important de connaître l'étendue de son corpus et de ses parties. Selon Charles Muller¹⁵⁹ :

“La fréquence d'un fait de langue quelconque (phonétique, morphosyntaxique, lexical, stylistique,...) n'a de signification que si elle est rapportée à l'étendue du texte considéré. Dire qu'un fait est plus fréquent dans le texte A que dans le texte B ne saurait s'appliquer à la fréquence absolue, mais à une fréquence relative qui ne peut être établie que si l'on connaît avec une précision suffisante le rapport de longueur des deux textes.”

En effet, les valeurs de N et de V ne sont pas liées par une relation fixe. Certes V augmente quand N augmente, mais la relation qui les unit n'est pas une loi simple :

¹⁵⁹ Ch. Muller (1979) : “Comment mesurer l'étendue d'un texte”, in *Langue française et linguistique quantitative*, p. 243.

l'étendue du vocabulaire est, comme l'exprime Charles Muller¹⁶⁰, “un fait de style limité par des contraintes linguistiques”.

Or, les calculs effectués par le logiciel Hyperbase permettent de mesurer l'étendue des textes dans le corpus en prenant en compte ces contraintes. 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 binomiale¹⁶¹, et elles servent aux calculs de pondération dans les différents traitements statistiques de notre étude.

Notre corpus principal, dans la version qui s'appuie sur les formes graphiques, le corpus A, contient 2.281.659 occurrences et 51.009 formes réparties sur les trente et une œuvres du corpus. Le tableau ci-dessous rend compte de la distribution des occurrences et des formes ainsi que de l'étendue.

¹⁶⁰ Ch. Muller (1979) “Calcul des probabilités et calcul d'un vocabulaire”, in *Langue française et linguistique quantitative*, p. 167.

¹⁶¹ Cf. Ch. Muller (1977) : p. 159-169.

N°	TITRE	OCCURRENCES	FORMES	P
1	PROCES	92955	10019	0.0407
2	FIÈVRE	101666	10124	0.0446
3	DÉLUGE	122269	12070	0.0536
4	EXTASE	90616	8924	0.0397
5	FUITE	104451	9962	0.0458
6	GUERRE	110602	9581	0.0485
7	MYDRIASE	9771	1859	0.0043
8	VOYAGE	113790	8367	0.0499
9	PROPHÉTIES	8690	1894	0.0038
10	MONDO	99114	6616	0.0434
11	INCONNU	124433	8455	0.0545
12	ICEBERGS	2637	767	0.0012
13	ARBRES	2910	646	0.0013
14	DÉSERT	150843	8335	0.0661
15	VILLES	15266	2448	0.0067
16	RONDE	83380	6424	0.0365
17	CHERCHEUR	134435	9219	0.0589
18	ANGOLI	24313	3121	0.0107
19	RODRIGUES	38238	4881	0.0168
20	RÊVE	87414	9259	0.0383
21	PRINTEMPS	70230	6136	0.0308
22	SIRANDANES	2463	749	0.0011
23	ONITSHA	76046	7187	0.0333
24	ÉTOILE	118756	7737	0.052
25	PAMANA	10593	1899	0.0046
26	DIEGO	73266	8888	0.0321
27	QUARANTAINE	165612	11170	0.0726
28	POISSON	86051	7578	0.0377
29	FÊTE	71025	8955	0.0311
30	NUAGES	21563	3841	0.0095
31	HASARD	68261	7230	0.0299
	TOTAL	2281659	50119	

Tableau n°5 : Étendue, occurrences et formes dans le corpus A.

Ce tableau regroupe les occurrences de chaque œuvre constitutive dans le corpus, classée par ordre chronologique. La deuxième colonne regroupe le nombre de formes distinctes et la proportion p est présentée dans la troisième colonne.

Il est aisé de constater que la taille des différents sous-corpus varie beaucoup. Certains ouvrages sont d'une étendue importante tandis que d'autres sont très courts. *La quarantaine* et *Désert* sont les plus longs avec plus de 150.000 occurrences. *Voyage au pays des arbres* et *Vers les icebergs* en comptent moins de 3000. Le graphique ci-dessous permet de visualiser ces variations.

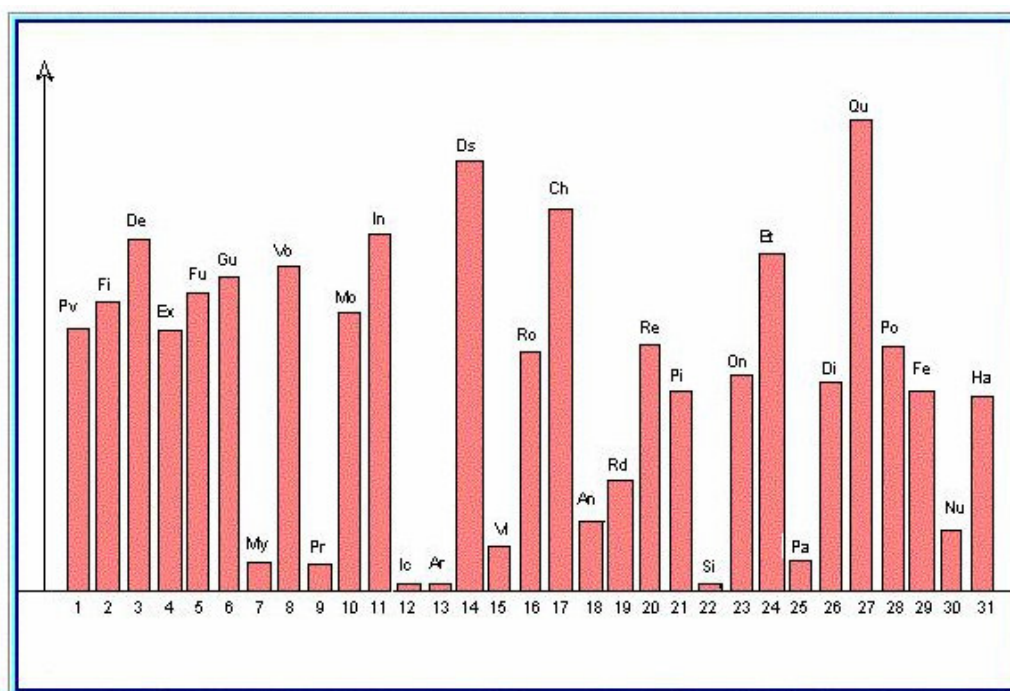


Figure n°14 : Étendue relative du corpus A.

Les premières œuvres, *Le procès-verbal*, *La fièvre*, *Le déluge*, *L'extase matérielle*, *Le livre des fuites* et *La guerre* sont assez homogènes au niveau de la taille, tandis que les œuvres romanesques qui suivent varient plus. Les derniers romans sont un peu plus courts que les premiers, à l'exception de *La quarantaine* – le plus long livre du corpus –, sans toutefois présenter de très grands écarts. Les ouvrages ethnologiques et la biographie *Diego et Frida* ont à peu près la même taille. En revanche, les récits *Mydriase* et *Vers les icebergs* et les livres pour enfants *Voyage au pays des arbres* et *Pawana* sont très courts. Le récit de voyage *Gens des nuages*, l'essai *Trois villes saintes*, *Les prophéties du Chilam Balam* et *Sirandanes* sont également très peu étendus.

Il est vrai que ces différences de taille pourraient nuire dans certains traitements statistiques et occasionner des erreurs d'interprétation. Charles Muller n'a jamais cessé de conseiller une certaine homogénéité de taille des différents sous-corpus. Dominique Labbé recommande, notamment dans le calcul de la distance lexicale, d'éviter la comparaison des corpus où la taille du texte la plus restreinte se trouve

inférieure au dixième du plus grand texte. Au-delà, tout calcul risque d'être influencé par la différence de taille.

Pour éviter des erreurs dues à une trop grande hétérogénéité de taille des textes, sans céder à notre ambition d'un corpus le plus complet possible, nous avons créé une version du corpus qui exclut les ouvrages très courts. Les ouvrages écartés sont *Mydriase*, *Les prophéties de Chilam Balam*, *Vers les icebergs*, *Voyage au pays des arbres*, *Trois villes saintes*, *Angoli Mala*, *Voyage à Rodrigues*, *Sirandanes*, *Pawana* et *Gens des nuages*.

Ce corpus, que nous appelons corpus C – le corpus “homogène”, est composé de 21 œuvres, il contient 2.146.241 occurrences regroupées en 48.643 formes différentes :

N°	TITRE	OCCURRENCES	FORMES	P
1	Procès	93025	10021	0.0433
2	Fièvre	101735	10128	0.0474
3	Déluge	122346	12074	0.0570
4	Extase	90633	8927	0.0422
5	Fuites	104456	9964	0.0487
6	Guerre	110609	9582	0.0515
7	voyage	113820	8368	0.0530
8	Mondo	99131	6617	0.0462
9	Inconnu	124464	8456	0.0580
10	Désert	150894	8337	0.0703
11	Ronde	83442	6425	0.0389
12	Chercheur	134526	9222	0.0627
13	Rêve	87458	9261	0.0407
14	Printemps	70246	6136	0.0327
15	Onitsha	76104	7192	0.0355
16	Étoile	118822	7739	0.0554
17	Diego	73341	8897	0.0342
18	Quarantaine	165651	11173	0.0772
19	Poisson	86164	7582	0.0401
20	Fête	71039	8959	0.0331
21	Hasard	68335	7233	0.0318
	TOTAL	2146241	48643	

Tableau n°6 : Étendue du corpus C, occurrences et formes.

La figure n°15 permet de constater la relative homogénéité des tailles des différents sous-corpus :

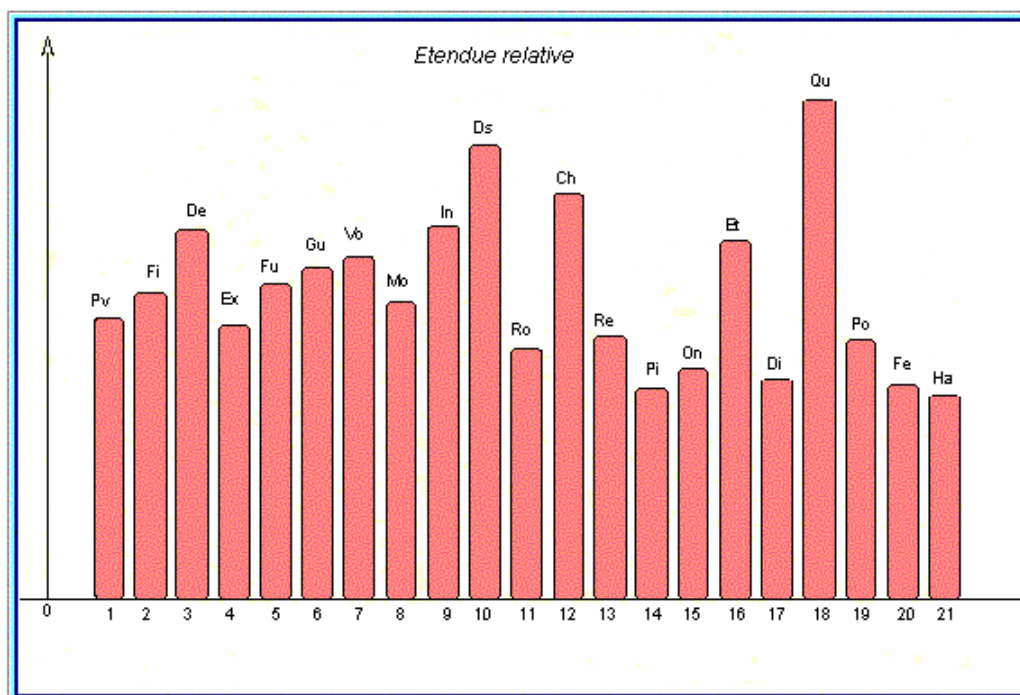


Figure n°15 : Étendue relative du corpus C.

La taille des ouvrages croît au début de la production pour culminer avec *Désert* de 1980. Par la suite, à l'exception de *La quarantaine*, elle diminue légèrement au fur et à mesure que l'œuvre progresse. La taille des ouvrages dépend également du genre littéraire dans lequel ils s'inscrivent. De manière générale, les ouvrages ethnologiques et les essais sont un peu plus courts que les œuvres de fiction.

Nous avons également jugé utile de constituer un corpus qui ne s'intéresse qu'à l'œuvre romanesque de Le Clézio et qui écarte les ouvrages à vocation ethnologique, la biographie et les essais. Ce corpus permet de cibler les livres de fiction, qui sont au cœur de notre recherche, et de faire abstraction des oppositions des différents genres littéraires.

Dans ce corpus nous avons écarté les romans les plus courts, *Voyage au pays des arbres*, *Angoli Mala* et *Pawana*, afin de respecter une certaine homogénéité de taille.

Cette variante du corpus, le corpus “roman” (D), englobe 17 livres et compte 1.717.096 occurrences et 41.317 formes.

N°	TITRE	OCCURRENCES	FORMES	P
1	Procès	91243	10018	0.0531
2	Fièvre	100374	10110	0.0585
3	Déluge	122453	12078	0.0713
4	Fuites	102769	9926	0.0599
5	Guerre	109196	9563	0.0636
6	Voyage	111852	8450	0.0651
7	Mondo	97332	6614	0.0567
8	Désert	149595	8337	0.0871
9	Ronde	82670	6433	0.0481
10	Chercheur	133326	9229	0.0776
11	Rodrigues	37999	4885	0.0221
12	Printemps	69622	6135	0.0405
13	Onitsha	75271	7189	0.0438
14	Étoile	117674	7743	0.0685
15	Quarantaine	163938	11182	0.0955
16	Poisson	85083	7577	0.0496
17	Hasard	66699	7307	0.0388
TOTAL		1.717.096	41.317	

Tableau n°7 : Étendue du corpus D.

L’histogramme de la figure n°16 permet de mieux apprécier l’étendue relative des différents ouvrages du corpus.

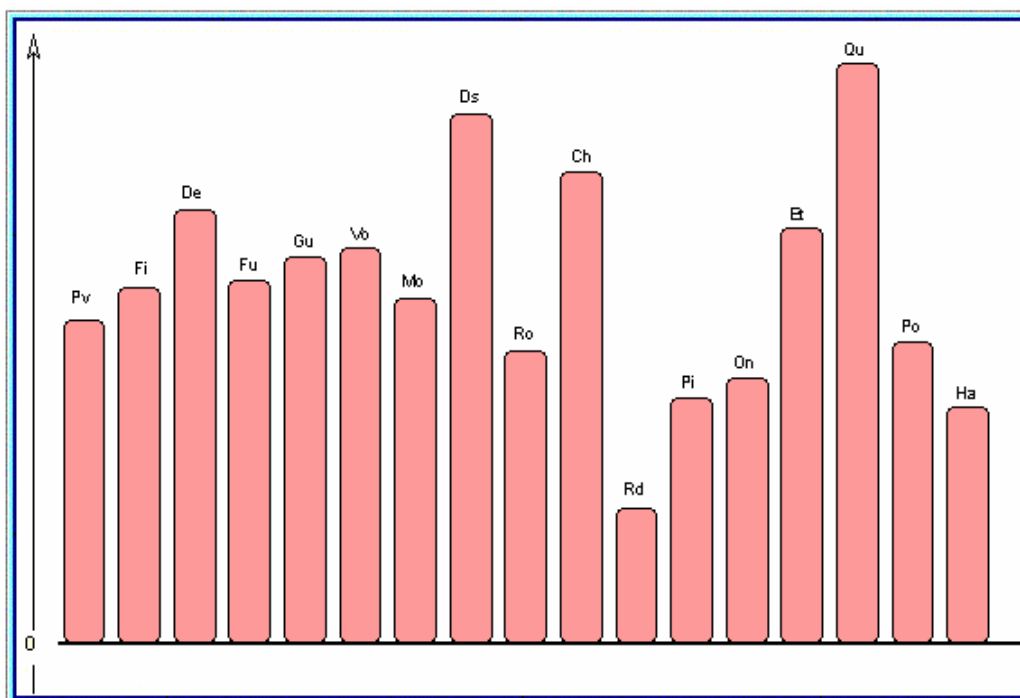


Figure n°16 : Étendue relative du corpus D.

Nous constatons que la taille des divers livres est relativement égale. *Voyage à Voyage à Rodrigues* est plus court que les autres et, de manière générale, les recueils de nouvelles sont un peu plus courts que les romans.

Enfin les multiples variantes du même corpus¹⁶² qui servent à comparer les différentes méthodes de lemmatisation, décrites auparavant¹⁶³, comprennent douze œuvres et ce corpus, dans sa version F, contient 993.054 occurrences et 29.938 formes différentes.

¹⁶² Les corpus F, G, H et I.

¹⁶³ Cf. le chapitre 2.5. "La lemmatisation."

N°	TITRE	OCCURRENCES	FORMES	P
1	Procès	91001	10025	0.0916
2	Fièvre	99966	10113	0.1007
3	Mydriase	9836	1871	0.0099
4	Voyages	113099	8359	0.1139
5	Mondo	97249	6612	0.0979
6	Désert	149477	8337	0.1505
7	Ronde	82659	6430	0.0832
8	Rodrigues	38032	4885	0.0383
9	Printemps	69602	6135	0.0701
10	Pawana	10562	1904	0.0106
11	Quarantaine	163918	11182	0.1651
12	Hasard	67652	7233	0.0681
	TOTAL	993053	29998	

Tableau n°8 : Étendue du corpus F, occurrences et formes.

Ce corpus, réduit pour les raisons techniques expliquées dans le chapitre 1.5. *La lemmatisation*, contient, outre le récit poétique *Mydriase*, exclusivement des romans et des recueils de nouvelles.

Ces multiples variantes de notre corpus serviront de référence dans les différentes études afin de permettre des comparaisons à partir de critères divers. Nous espérons pouvoir, grâce à ce “jeu” de corpus, donner une image plus fine et diversifiée de la structure du vocabulaire de J.M.G. Le Clézio.

Pour obtenir une première vue générale de la structure quantitative de notre corpus, nous étudierons pour commencer la distribution des fréquences et le rapport entre les hautes et les basses fréquences.

2.3. La distribution des fréquences.

“Ce qu’il faudrait faire, pour vraiment percer le mystère de l’écriture, c’est écrire jusqu’aux limites de ses forces. Penser, et déterminer cette pensée avec des signes, sans arrêt, jusqu’à ce que l’on tombe endormi, évanoui, ou mort.”

L’extase matérielle, page 132.

Une classe de fréquence réunit tous les mots de même fréquence dans un texte ou dans un corpus. L’ensemble des classes de fréquences forme la distribution des fréquences ou la gamme des fréquences.

“Classer les mots d’une langue par leur fréquence, ou tout au moins par grandes classes de fréquence, écrit Charles Muller¹⁶⁴, est une opération qui a intéressé très tôt la pédagogie des langues ; depuis que la linguistique appliquée a étendu son domaine, d’autres utilisations d’une bonne liste de fréquences se dessinent : programmation de la traduction automatique, micro-glossaires, codage, etc. A mesure que ces besoins se précisent et se multiplient, il est normal que les méthodes qui conduisent à un classement quantitatif se perfectionnent.”

La connaissance de la distribution des fréquences permet d’étudier les proportions relatives des basses, moyennes et hautes fréquences. Elle est également nécessaire pour l’application de la méthode fondée sur la loi binomiale qu’exploite, entre autres, le logiciel Hyperbase. De nombreux calculs ont été proposés sur cette distribution des fréquences, notamment l’application de la loi de Zipf¹⁶⁵ qui stipule très grossièrement que “le produit (rang x fréquence) est à peu près constant”.

¹⁶⁴ Ch. Muller (1979) : “Fréquence, dispersion et usage, à propos des dictionnaires de fréquence.”, in *Langue française et linguistique quantitative*, p.197.

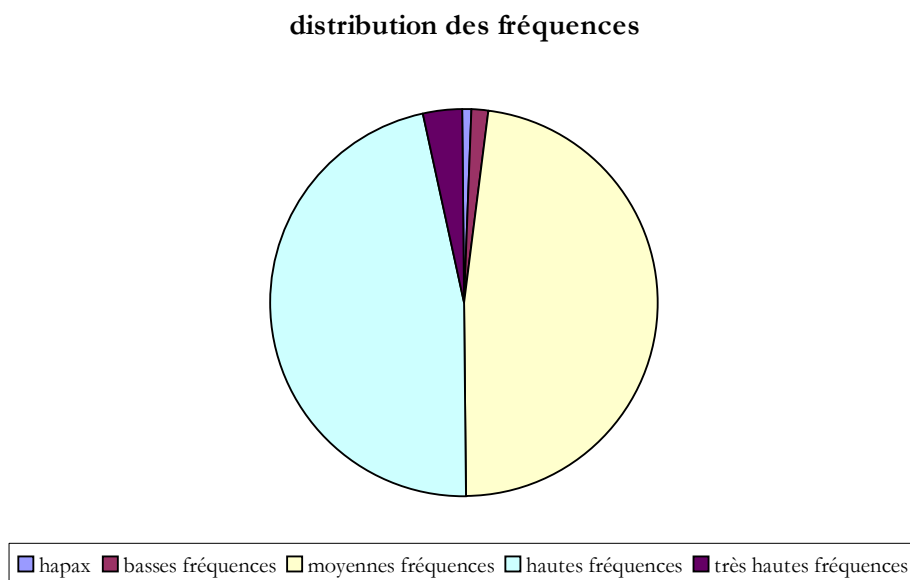
¹⁶⁵ Pour plus de détails sur ce calcul voir L. Lebart, A. Salem (1994) : p. 47-48 et É. Brunet (1978) : p. 95-109.

Dans ce chapitre, nous nous intéresserons à la distribution des fréquences dans notre corpus principal et au rapport entre les basses et les hautes fréquences ainsi qu'à l'étude des hapax dans les différentes versions de notre corpus.

Nous avons, en recourant à l'index des formes et des lemmes¹⁶⁶ fourni par le logiciel Hyperbase¹⁶⁷, divisé les mots en groupes de fréquence. Ces groupes de fréquences ont ensuite été regroupés en cinq classes :

1. Hapax
2. Basses fréquences (2-100)
3. Moyennes fréquences (101-1999)
4. Hautes fréquences (1000-2000)
5. Très hautes fréquences (>2000)

La distribution dans notre corpus des différentes classes de fréquences par rapport à N (le nombre d'occurrences) est la suivante :



¹⁶⁶ Lemmatisé avec Cordial 7.

¹⁶⁷ Le dictionnaire du corpus entier est fourni en annexe.

Hapax : 20668 occurrences

Basses fréquences : 28.614 occurrences

Moyennes fréquences : 1.011.778 occurrences

Hautes fréquences : 1.006.406 occurrences

Très hautes fréquences : 71.193 occurrences

Les groupes de fréquences se comportent d'une manière tout à fait différente selon qu'ils sont calculés sur V (les vocables) ou sur N (les occurrences). Dans notre corpus les hapax occupent 40 % du vocabulaire mais seulement 1 % du texte. Britt-Marie Kylander constate dans son étude sur le vocabulaire de Molière¹⁶⁸ que les classes de faibles fréquences occupent 65 % du vocabulaire, mais seulement 4 % du texte, tandis que les fréquences élevées ne constituent que 6 % du vocabulaire et 80 % du texte. Elle écrit :

“Les vocables à fréquence faible sont donc très importants dans la composition du vocabulaire, mais ne jouent qu'un rôle modeste dans celle du texte, alors que le contraire vaut pour les vocables à haute fréquence.”

Le tableau suivant présente la distribution des fréquences jusqu'à la fréquence 100 dans le corpus, en tenant compte du calcul sur les formes et sur les lemmes. Le premier élément de chaque classe précise la classe de fréquence (de 1, 2, n mots), le second l'effectif de la classe correspondante (combien de mots employés 1, 2, n fois). Par exemple, les mots de fréquence 1 (hapax) sont 20668 etc.

¹⁶⁸ B.-M. Kylander (1995) : p. 16-17.

Formes					Lemmes				
1 20668	21 214	41 53	61 33	81 17	1 1910	21 6	41 1	61 0	81 0
2 7197	22 173	42 59	62 36	82 18	2 436	22 4	42 1	62 0	82 0
3 3980	23 176	43 60	63 35	83 14	3 188	23 0	43 3	63 0	83 0
4 2579	24 162	44 60	64 39	84 17	4 104	24 4	44 1	64 2	84 0
5 1909	25 164	45 53	65 31	85 22	5 66	25 6	45 0	65 0	85 0
6 1475	26 132	46 50	66 25	86 19	6 61	26 4	46 1	66 0	86 0
7 1099	27 134	47 60	67 24	87 19	7 31	27 1	47 1	67 0	87 0
8 919	28 121	48 50	68 35	88 20	8 20	28 5	48 1	68 2	88 1
9 828	29 114	49 62	69 35	89 21	9 23	29 1	49 0	69 0	89 0
10 678	30 105	50 48	70 24	90 22	10 15	30 0	50 0	70 0	90 0
11 605	31 120	51 52	71 29	91 18	11 14	31 3	51 0	71 0	91 0
12 509	32 119	52 37	72 19	92 15	12 11	32 3	52 0	72 0	92 0
13 440	33 98	53 53	73 21	93 19	13 9	33 2	53 1	73 0	93 1
14 386	34 101	54 40	74 24	94 12	14 5	34 5	54 0	74 0	94 0
15 355	35 85	55 39	75 32	95 19	15 9	35 1	55 1	75 0	95 0
16 287	36 99	56 32	76 25	96 9	16 2	36 1	56 0	76 0	96 0
17 341	37 75	57 46	77 32	97 10	17 4	37 1	57 0	77 1	97 0
18 258	38 73	58 42	78 23	98 17	18 6	38 1	58 0	78 1	98 0
19 239	39 78	59 45	79 21	99 20	19 6	39 0	59 0	79 0	99 1
20 231	40 72	60 32	80 31	100 10	20 4	40 0	60 0	80 0	100 0

Tableau n°9 : Distribution des fréquences, forme et lemmes dans le corpus B.

Nous constatons que, dans cette gamme de fréquence, le nombre des vocables de fréquence 1 est supérieur à celui des vocables de fréquence 2, qui à son tour est supérieur à celui des vocables de fréquence 3, et ainsi de suite. Les effectifs diminuent au fur et à mesure que la fréquence croît, même si quelques irrégularités se produisent à partir d'une certaine limite. Selon Charles Muller¹⁶⁹ “cette inégalité reste vraie malgré les accidents aléatoires qui peuvent la mettre en défaut, surtout quand les effectifs deviennent petits.”

Notre corpus ne fait pas exception à cette règle, les classes de fréquences se succèdent de la manière décrite par tous les chercheurs en lexicométrie.

“Visiblement, écrit Etienne Brunet¹⁷⁰, les classes de fréquences, comme les couleurs du spectre, forment un continuum et les mouvements s'inversent progressivement en passant de la première image à la

¹⁶⁹ Ch. Muller (1977) : p. 81-82.

¹⁷⁰ É. Brunet (1988) : p. 44.

dernière”.

Arrêtons-nous un instant pour regarder de près “la première et la dernière image de ce spectre” afin d’en mesurer les contrastes.

2.3.1. Les hautes et les basses fréquences

“C’est ce mystère plus que tout autre que j’aimerais délayer. Car il porte en lui la clé du langage, et peut-être même la raison originelle. Chaque mot, comme un clou, devrait me permettre de fixer un peu plus durablement cette toile. Mais il faut choisir ces clous, ni trop faibles, ni trop blessants.”

L’extase matérielle, page 41.

Les graphiques¹⁷¹ suivants permettent de constater une autre caractéristique de notre corpus¹⁷². Le premier histogramme regroupe les 100 plus hautes fréquences et rend compte de leur distribution¹⁷³. Le deuxième illustre la distribution des hapax¹⁷⁴ dans les différentes œuvres du corpus:

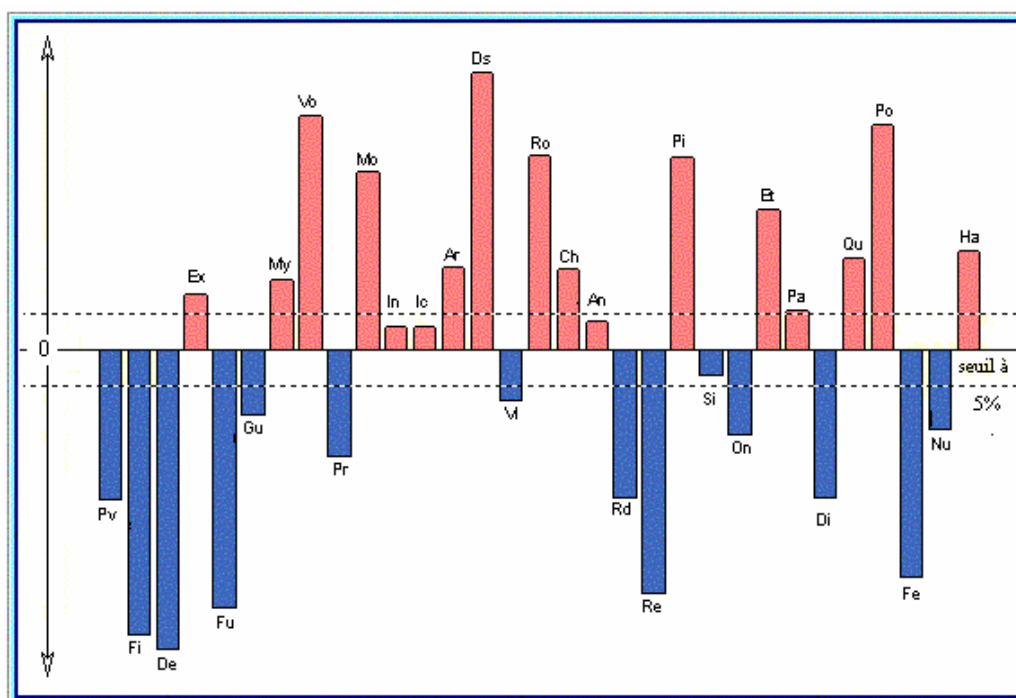


Figure n°17 : La distribution des plus hautes fréquences à travers le corpus (A).

¹⁷¹ Pour plus de détails sur l’élaboration des histogrammes, cf. p. 68.

¹⁷² Pour l’interprétation du graphique la zone délimitée par les deux traits pointillés indique l’intervalle de fluctuation “normale”, celle pour laquelle on ne peut pas écarter l’hypothèse que les fluctuations sont dues au hasard. Plus on s’écarte de cette zone, en dessous et en dessus, plus les écarts sont significatifs (du point de vue probabiliste).

¹⁷³ La surface occupée par les 100 mots les plus fréquents/divisée par la surface totale de l’œuvre.

¹⁷⁴ Cf. chapitre 2.3.2. “Les hapax”.

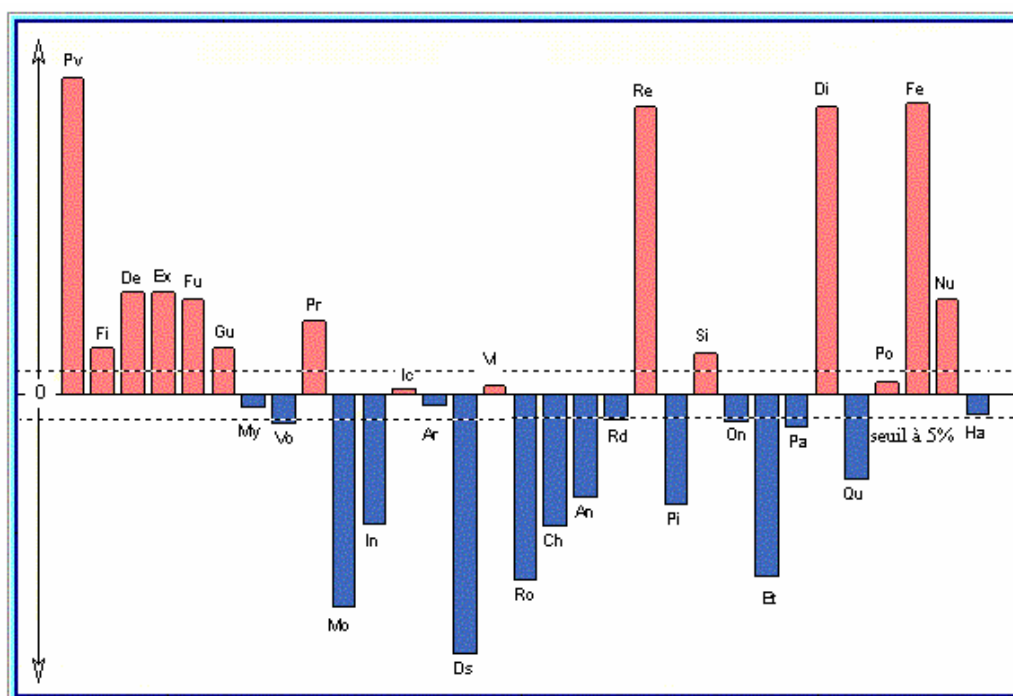


Figure n°18 : La distribution des hapax à travers le corpus (A).

Les histogrammes illustrent l'inversion du mouvement. Il est aisé de constater que les livres qui contiennent le plus d'hapax sont les plus "pauvres" en hautes fréquences.

Etienne Brunet a constaté la richesse extraordinaire de hautes fréquences dans le vocabulaire de Proust par rapport aux autres auteurs. Il écrit à ce propos¹⁷⁵ :

“Car les autres auteurs font un choix dans le lexique alors que Proust affirme son originalité dans la syntaxe... En fait, 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.”

¹⁷⁵ É. Brunet (1983) : p. 28.

Nous sommes tentée de faire la même interprétation en ce qui concerne les œuvres qui sont riches en hautes fréquences dans notre corpus.

En effet, 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'effectif des différentes classes de fréquence dépend aussi de l'étendue du texte. Plus un texte est long, plus les répétitions semblent inévitables et les hapax deviennent relativement rares (puisque $F_1 > F_2$). Néanmoins, les contrastes évidents entre les nombres des hapax sont très intéressants parce qu'ils donnent des indications sur la richesse lexicale.

Toutefois, avant d'étudier d'autres aspects fondamentaux dans l'étude de la structure d'un vocabulaire comme celui de l'analyse de la richesse lexicale, il convient d'effectuer une étude détaillée des hapax dans notre corpus.

2.3.2. Les hapax.

“Ou bien il doit y avoir un mot, un vrai mot, qui à lui seul ferait éclater toutes ces gangues. Pas un mot intelligent, ni un mot d’amour, mais un mot quelconque, qui exploserait dans la chair à la manière d’une balle-obus dans le crâne d’un rhinocéros. Un mot, un seul mot. Mais j’ai beau le chercher, je ne le trouve pas.”

La guerre, page 90.

Lorsque l’on parle des mots qui ne figurent qu’une fois dans un corpus, il convient de faire la distinction entre la notion d’hapax et celle de sous-fréquence 1. En effet, par le terme hapax, on désigne les vocables de fréquence 1 qui ont été rencontrés une seule fois dans un corpus, et conséquemment dans un seul texte. Les mots de ce corpus qui n’ont qu’une seule occurrence dans un texte donné sont désignés comme des vocables de sous-fréquence 1. Les vocables de fréquence 1 ont une forte influence sur la richesse lexicale. Selon Etienne Brunet¹⁷⁶ :

“Le phénomène qu’on appelle richesse lexicale est davantage soumis aux variations des basses fréquences et particulièrement de la fréquence 1”.

Nous tenterons de donner une image plus juste de la richesse lexicale en prenant en compte l’étude des hapax. Les fluctuations de la fréquence d’hapax sont-elles d’une influence considérable quant aux études sur la richesse lexicale de notre corpus qui semble être soumise à ces variations ?

Mais l’analyse de ces vocables à fréquence 1 présente également d’autres aspects. Les critiques littéraires ont souvent évoqué l’intérêt de l’étude des hapax car ils y

¹⁷⁶ É. Brunet (1981) : p. 65.

trouvent de nombreux substantifs et des noms propres qu'ils considèrent comme les unités les plus parlantes et les plus explicites dans les études stylistiques. Michael P. Oakes s'exprime ainsi à ce propos¹⁷⁷ :

“These words that are hapax legomena tend to be interesting, reflecting such factors as the background, experience and powers of mind of an author, and conveying delicate shades of meaning. Many qualitative studies of literature have focused their attention on rare words, since the commonest words tend to be connectives, pronouns and articles, words whose meaning and use are much more prescribed and conventional than the case of rare words.”

Toutefois, Michael Oakes souligne la fragilité de l'analyse des hapax du fait justement de leur faible fréquence, notamment dans les distributions de textes. “Any stylistic trait which is to be used in the determination of authorship must be sufficiently frequent”, écrit-il¹⁷⁸. Pour avoir recours à l'hypothèse statistique de χ^2 par exemple, il faut un minimum de 5 fréquences, et de préférence 10. Il convient donc selon lui de faire preuve de prudence dans les conclusions des études stylistiques basées sur les hapax.

La proportion des vocables de fréquence 1 ne dépend pas seulement de certaines caractéristiques stylistiques ou linguistiques, mais aussi de la longueur des textes. Pour pouvoir comparer nos différents sous-corpus et pondérer correctement les valeurs, nous avons encore une fois recours à la loi normale.

Le tableau n°10 rend compte de leur distribution dans le corpus ainsi que des valeurs des écarts réduits calculés par le logiciel Hyperbase.

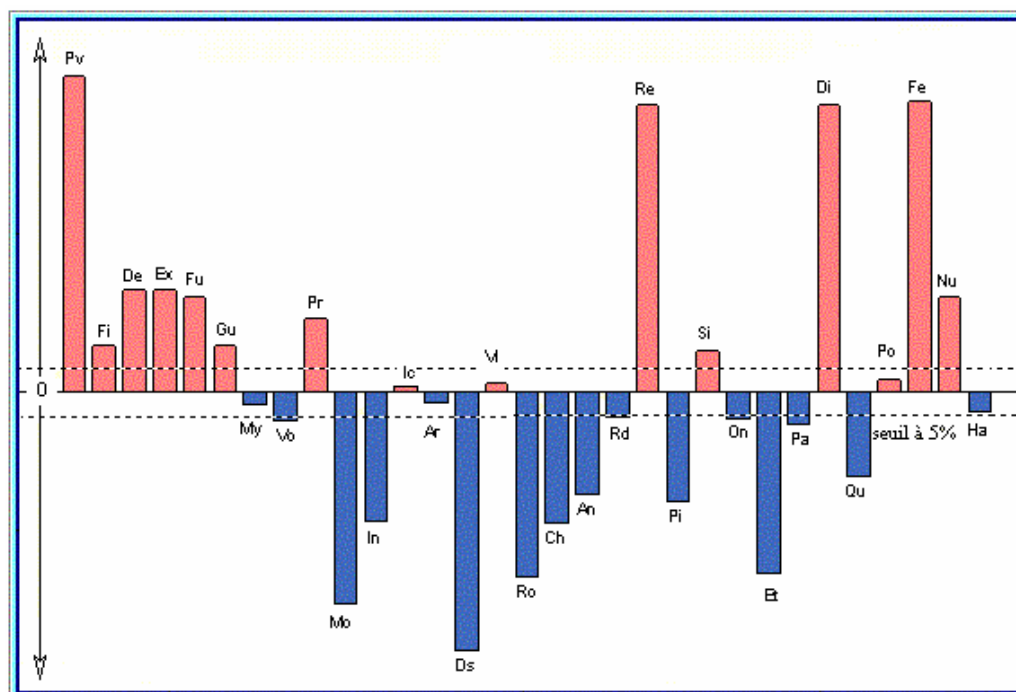
¹⁷⁷ M.P. Oakes (1998) : p. 201.

¹⁷⁸ *Ibid.*, p. 201-202.

n°	Titre	Hapax	éc.réduit	n°	Titre	Hapax	éc.réduit
1	PROCES	1263	15.90	17	CHERCHEUR	815	-11.02
2	FIÈVRE	1027	4.54	18	ANGOLI	82	-9.06
3	DÉLUGE	1399	10.14	19	RODRIGUE	304	-1.78
4	EXTASE	1080	10.23	20	RÊVE	1473	25.89
5	FUITE	1166	8.35	21	PRINTEMPS	394	-9.14
6	GUERRE	1128	5.10	22	SIRANDANES	40	3.94
7	MYDRIASE	77	-0.97	23	ONITSHA	621	-1.88
8	VOYAGE	951	-1.62	24	ÉTOILE	551	-15.68
9	PROPHÉTIE	136	6.83	25	PAWANA	67	-2.72
10	MONDO	334	-18.62	26	DIEGO	1304	26.41
11	INCONNU	666	-13.32	27	QUARANTA	1206	-6.82
12	ICEBERGS	26	0.58	28	POISSON	803	1.71
13	ARBRES	22	-0.71	29	FÊTE	1276	26.47
14	DÉSERT	531	-22.61	30	NUAGES	314	9.07
15	VILLES	143	0.76	31	HASARD	566	-1.43
16	RONDE	304	-16.15				
					TOTAL :	20069	

Tableau n°10 : La distribution des hapax.

L'histogramme ci-dessous illustre la distribution des hapax à travers le corpus :



L'histogramme rend compte des valeurs positives au début de l'œuvre, avec un écart réduit très élevé au niveau du premier livre *Le procès-verbal*. Les ouvrages qui suivent, *La fièvre*, *Le déluge*, *L'extase matérielle* et *La guerre* sont également riches en hapax. Il s'agit vraisemblablement d'une caractéristique du genre "nouveau roman". L'auteur fait de nombreuses énumérations, souvent en détaillant des mots représentant "l'envahissement commercial" caractéristique de notre temps et dont nous sommes les victimes. L'auteur énumère des sigles commerciaux, des noms de produits chimiques ou, comme dans l'exemple suivant¹⁷⁹, des noms de matériaux synthétiques :

“Matières ! Matières ! Luisantes, douces, fragiles, inflammables, pareilles à des nuages de fumée. Couleurs rouges, noires, profondes. Ce sont elles qui pensent maintenant. Ce sont elles qui inventent les histoires, les religions, les sciences. Elles bougent, elles enlacent. Chlorofibre Rhovyl, Polyamide, Rhodia 100 % polyester, acétate, Gama, Kreon de Skaï, fibre acrylique ACSA, Leacril, Dralon acrylique, polyuréthane synthétique, Brinyl polyamide texturé, Masulyne, Mérinovyl, Clévyl, Flanyl, Dropnvl Polyamide 66, Terital, Tercryl, Viscose, Fibranne, Crylor, expansé Vinyle.”

Les ouvrages de type ethnologique et biographique sont également riches en hapax. *Le rêve mexicain*, *Diego et Frida* ainsi que *La fête chantée* sont les œuvres regorgeant le plus d'hapax. De la même façon que dans les études précédentes, la découverte d'un nouvel univers et la rencontre culturelle se reflète par l'abondance de termes spécifiques et des passages en langue étrangère.

Mydriase et *Voyages de l'autre côté* semblent annoncer un changement dans l'écriture leclézienne, avec des valeurs négatives après la richesse de la période "nouveau roman". Ceci se confirme à partir de *Mondo*, dans les romans et les nouvelles où nous trouvons des valeurs très négatives. Les écarts négatifs de ce livre et du

¹⁷⁹ *La Guerre*, p. 232.

lemmatisation. En dehors de ces écarts, il semble globalement que la lemmatisation ne modifie pas, dans notre corpus, de façon significative, les résultats de l'étude des hapax.

Toutefois, bien que la distribution soit essentiellement la même, la lemmatisation diminue nettement le taux d'hapax dans un corpus, notre corpus B ne compte que 12507 hapax pour ses 31520 lemmes :

n°	Titre	Hapax	éc.réduit	n°	Titre	Hapax	éc.réduit
1	PROCES	727	10.12	17	CHERCHEUR	482	-9.71
2	FIÈVRE	507	-2.04	18	ANGOLI	35	-8.59
3	DÉLUGE	763	3.44	19	RODRIGUE	192	-1.28
4	EXTASE	636	6.18	20	RÊVE	1028	25.89
5	FUITE	830	11.21	21	PRINTEMPS	248	-7.11
6	GUERRE	745	5.87	22	SIRANDANES	21	1.98
7	MYDRIASE	53	-0.20	23	ONITSHA	427	0.52
8	VOYAGE	646	0.81	24	ÉTOILE	307	-13.86
9	PROPHÉTI	102	7.72	25	PAWANA	19	-5.18
10	MONDO	148	-17.19	26	DIEGO	917	26.22
11	INCONNU	438	-9.76	27	QUARANTA	695	-7.34
12	ICEBERGS	16	0.25	28	POISSON	480	0.41
13	ARBRES	16	-0.11	29	FÊTE	928	27.85
14	DÉSERT	327	-18.01	30	NUAGES	215	8.81
15	VILLES	115	3.30	31	HASARD	292	-4.34
16	RONDE	152	-14.55				
					TOTAL :	12507	

Tableau n°11 : La distribution des hapax dans le corpus lemmatisé, le corpus B.

La richesse du système verbal français avec ses nombreuses formes a pour conséquence que la lemmatisation intervient essentiellement dans les formes verbales. Il n'y a en effet pas beaucoup d'"hapax-verbes" lorsque le corpus est lemmatisé et les différences constatées semblent émaner de ce phénomène.

Le graphique ci-dessous illustrant la distribution des hapax dans notre corpus "homogène" (Corpus C) permet une analyse plus synthétique du mouvement général, les ouvrages très courts étant écartés :

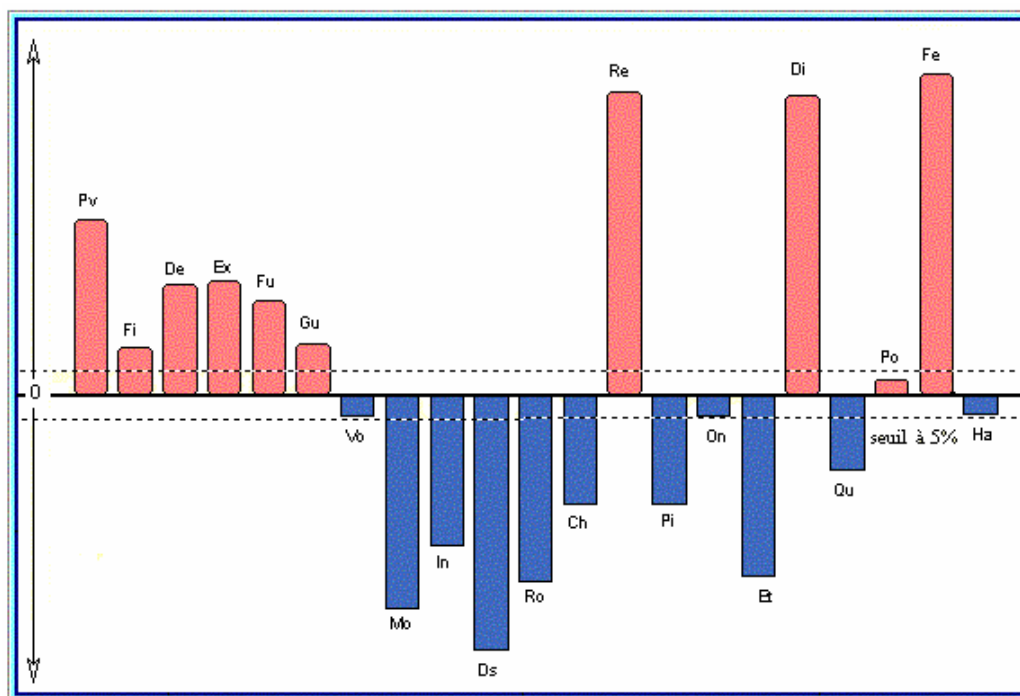


Figure n°21 : Histogramme de la distribution des hapax dans le corpus C.

Ce graphique confirme une fois encore les observations précédentes ; il met en exergue les trois périodes distinctes que nous avons pu constater auparavant : celle du début, riche en hapax ; celle du milieu, la période des grands romans, déficitaire ; une troisième période où les hapax sont nombreux dans les ouvrages ethnologiques et manifestent une tendance à la hausse dans l'œuvre romanesque, sans jamais atteindre des valeurs positives, à l'exception de *Poisson d'or*.

Pour regarder de près les œuvres de fiction, nous faisons la même étude basée sur notre corpus romanesque, le corpus D. Ce corpus contient 16.646 hapax distribués selon le tableau suivant :

n°	Titre	occurences	Hapax	écart réd.
1	Procès	10021	1522	22.48
2	Fièvre	10128	1258	10.03
3	Déluge	12074	1729	17.68
4	Fuites	9964	1478	16.30
5	Guerre	9582	1337	9.50
6	Voyage	8368	1139	2.16
7	Mondo	6617	409	-17.65
8	Désert	8337	654	-21.26
9	Ronde	6425	379	-14.85
10	Chercheur	9222	1007	-7.56
11	Rodrigues	4889	450	4.76
12	Printemps	6136	465	-7.77
13	Onitsha	7192	740	0.92
14	Étoile	7739	656	-14.31
15	Quarantaine	11173	1485	-1.95
16	Poisson	7582	930	4.31
17	Hasard	7233	707	2.58
Tot		40350	16345	

Tableau n°12 : La distribution des hapax dans le corpus "roman" (C).

L'histogramme illustre le même mouvement général que nous avons observé dans l'étude du grand corpus, en permettant toutefois de dégager des traits particuliers.

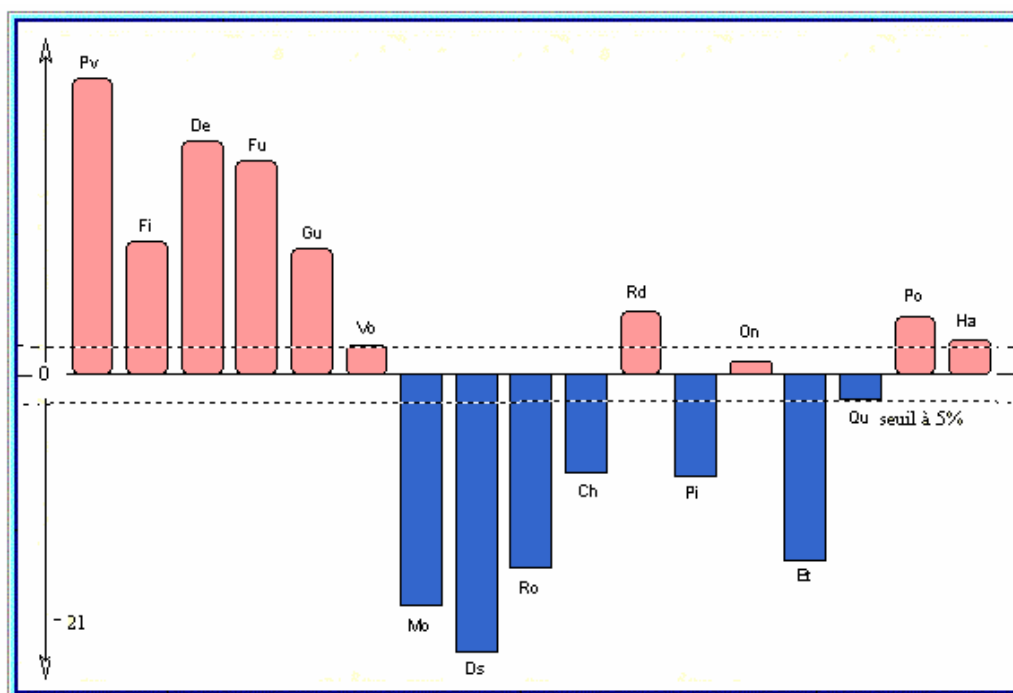


Figure n°22 : Distribution des hapax dans le corpus "roman" (D).

Le graphique distingue également les trois périodes que nous venons de commenter, mais il permet cependant de noter des fluctuations plus importantes

dans la dernière période. Nous remarquons que *Le chercheur d'or* est pauvre en hapax alors que le livre suivant, *Voyage à Rodrigues*, qui traite du même univers que le précédent roman, est quant à lui riche en hapax. Il convient de tenir compte de la richesse de termes techniques dans le dernier et du fait qu'il est écrit, en partie, sous la forme d'un journal personnel. Le Clézio écrit ceci à propos du langage du personnage principal, le grand-père¹⁸⁰ :

“Il invente aussi une langue, une véritable langue avec ses mots, ses règles de grammaire, son alphabet, sa symbolique, une langue pour rêver plus que pour parler, une langue pour s'adresser au monde étrange dans lequel il a choisi de vivre... C'est une langue pour parler au temps passé, pour s'adresser aux ombres, au monde à jamais disparu, du temps où la lumière brillait si fort sur la mer des Indes... C'est avec ces traces qu'il élabore, année après année, le langage du Corsaire inconnu...”

Nous avons indiqué auparavant que le corpus Le Clézio dans sa version lemmatisée (le corpus B) contient 12507 hapax, c'est-à-dire 39.7% du vocabulaire (31520 vocables).

Ce taux est-il exceptionnel ? Chaque lecteur de Le Clézio connaît l'abondance, dans ses livres, des mots de langues étrangères - il y a des mots en langues amérindiennes, en anglais, en italien et en espagnol - des noms propres ainsi que des termes techniques, qui auraient une forte influence sur cette analyse.

Il est intéressant de le comparer avec d'autres bases littéraires. L'accès aux bases de données “Hyperbase” traitant d'autres auteurs français permet de constater des valeurs semblables. Parmi les écrivains du XIX^{ème} siècle, le corpus Balzac compte 19970 hapax, c'est-à-dire 33% du vocabulaire, le corpus Chateaubriand 40 %, le corpus Flaubert 38 %, le corpus Maupassant 34 % et le corpus Jules Verne 37 %.

¹⁸⁰ *Voyage à Rodrigues*, p. 103-107

Cela permet de constater que le taux d'hapax chez Le Clézio n'a rien d'exceptionnel. Etienne Brunet écrit ceci à propos du vocabulaire de Proust et de Zola¹⁸¹ :

“Il est vrai que les enquêtes que Zola mène dans des milieux fort divers favorisent l'extension du vocabulaire et que l'enquête de Proust s'exerce plus en profondeur qu'en surface. Il semble toutefois que l'explication soit ailleurs : ni Chateaubriand, ni Giraudoux ne prétendent faire œuvre documentaire. Mais ils ont un certain goût des mots rares ou recherchés, que Proust ne partage pas.”

Nous avons pu constater à travers cette étude que les hapax se trouvent surtout, à l'exception des ouvrages “nouveau roman”, dans les livres qui traitent des milieux et des cultures éloignées ou exotiques. En revanche, le vocabulaire des livres de fiction n'est pas riche en hapax.

Le mouvement général émanant de l'analyse des hapax est-il vraiment représentatif de la structure lexicale de Le Clézio ? Il convient, avant d'en tirer des conclusions précises, d'étudier d'autres aspects fondamentaux de l'étude de la structure d'un vocabulaire, comme celui de l'analyse de la richesse lexicale.

¹⁸¹ É. Brunet (1985) : p. 25.

2.4. La richesse du vocabulaire

“Faire des pots, tisser des trames, vanner ou simplement vivre selon les rythmes agraires, c’est-à-dire produire et non pas représenter, c’est ce que peut faire l’écrivain avec les mots, s’il possède ces mots.”

*Ailleurs*¹⁸², page 29.

On n’a pas attendu les études quantitatives pour qualifier de “riche” ou de “pauvre” le vocabulaire d’un écrivain ou d’une œuvre littéraire. “Appréciation toute subjective d’ailleurs, écrit Charles Muller¹⁸³, qui ne se fonde généralement sur aucune donnée sûre, et qui traduit plutôt la présence dans le texte de quelques vocables jugés rares, ou au contraire l’absence de tels éléments du lexique.”

En fait, un nombre relativement élevé de vocables rares dans un corpus n’indique pas nécessairement un vocabulaire riche, même s’il y a souvent une bonne corrélation entre les deux phénomènes. Ainsi, Étienne Brunet¹⁸⁴ a constaté que tous les termes techniques rares et spectaculaires chez Zola, qui pouvaient donner l’impression d’un vocabulaire riche, ne compensaient pas, en nombre, le grand déficit des vocables abstraits, notamment créés à l’aide d’affixes. Une abondance de ces vocables pouvait par contre, chez d’autres auteurs, contribuer à un vocabulaire nuancé et varié.

“Ce qui justifie un traitement mathématique de la richesse lexicale, écrit Charles Bernet¹⁸⁵, c’est la volonté d’apporter une solution que l’on voudrait objective à un problème qui n’a longtemps reçu que des réponses impressionnistes et subjectives.”

¹⁸² *Ailleurs*, Entretiens avec Jean-Louis Ezine, Paris, Arléa (Entretiens France-Culture), 1995.

¹⁸³ Ch. Muller (1977) : p. 115.

¹⁸⁴ É. Brunet (1985) : p. 25-27.

¹⁸⁵ Ch. Bernet (1988) : “Faits lexicaux. Richesse du vocabulaire. Résultats.”, in Ph Thoiron., D. Labbé, D. Serant (éds). Paris – Genève, Champion – Slatkine, p. 2.

La richesse du vocabulaire d'un texte ne dépend pas seulement de l'étendue supposée du lexique de son auteur, de sa "culture". Elle est également influencée par le genre dans lequel il s'inscrit et par la spécialisation du vocabulaire en fonction du thème traité dans tel ou tel passage de ce texte. Naturellement, il n'y a aucune péjoration dans "pauvre" ni valorisation dans "riche", mais ces termes se sont imposés. Il ne s'agit que de mesurer des styles et des manières d'utiliser ou de ne pas utiliser l'étendue considérable du lexique de la langue.

Lorsque l'on veut comparer la richesse lexicale de deux textes qui sont d'égale étendue, ou à peu près, le calcul est simple, on s'interroge sur le rapport V/N ¹⁸⁶ (nombre de mots différents ou vocables dans un nombre donné d'occurrences).

Par contre, quand on est en face de deux corpus de taille différente, la situation est plus difficile. Il est compliqué et délicat d'établir, de manière précise, comment la richesse lexicale varie avec N et V lorsque les textes sont de longueurs différentes¹⁸⁷. La relation entre N et V n'est pas constante dans des textes d'étendue inégale. Quand N croît, V croît aussi, mais de plus en plus lentement.

Ainsi, pour comparer la richesse lexicale de deux textes d'étendue inégale¹⁸⁸, on dispose de plusieurs calculs, notamment celui du calcul d'un vocabulaire théorique fondé sur la loi binomiale¹⁸⁹.

La méthode binomiale résout¹⁹⁰ le problème de la longueur inégale des textes en raccourcissant le plus long aux dimensions du plus court. Connaissant la distribution des fréquences du texte le plus long, on peut calculer le vocabulaire d'un fragment théorique de ce texte qui aurait la même longueur que le texte le plus court tout en conservant sa propre structure lexicale. Le vocabulaire

¹⁸⁶ Type/token ratio

¹⁸⁷ Pour ces notions, cf. chapitre 1.4.4. "Le codage".

¹⁸⁸ Ch. Muller (1970) : "Sur la mesure de la richesse lexicale" in *Langue française et linguistique quantitative*, Genève, Slatkine, 1979, p. 281-307.

¹⁸⁹ Ch. Muller (1977) : ch. 16.

¹⁹⁰ Les chercheurs émettent, toutefois, des réserves sur son efficacité lors de la comparaison des textes d'étendue très différente. Cf. P. Hubert, D. Labbé (1994) : "La richesse du vocabulaire", *Communication au congrès de l'ALLC-ACH*, Paris, La Sorbonne.

théorique de ce fragment peut ensuite être directement comparé au vocabulaire du texte le plus court et ainsi permettre de déterminer lequel des deux textes a le vocabulaire le plus riche.

Le vocabulaire de J.M.G. Le Clézio est-il “riche” ou “pauvre” ? Nous tentons, par cette étude, de répondre à cette question et d’arriver à un classement objectif sur la richesse lexicale, au-delà des impressions subjectives.

Parallèlement à l’analyse lexicale à proprement parler, nous nous proposons de comparer différentes techniques de travail. Dans un premier temps, nous aurons recours au logiciel Hyperbase et nous comparerons les résultats de deux techniques de travail différentes, sur un corpus lemmatisé ou non.

Dans un deuxième temps nous comparerons ces résultats avec ceux obtenus par l’application du modèle de partition, qui permet aussi d’analyser la diversité et la spécialisation du vocabulaire.

Enfin, nous testerons la pertinence de l’indice W élaboré par Etienne Brunet.

2.4.1. Mesures de la richesse lexicale.

Pour la première étude sur la richesse lexicale, nous avons recours à nos deux corpus de 31 œuvres, le corpus qui s’appuie sur les formes (corpus A) et celui qui est lemmatisé avec Cordial (corpus B). Afin de détecter des anomalies éventuelles dues à la diversité de tailles des sous-corpus, nous effectuerons la même analyse du corpus homogénéisé, basé sur les formes (corpus C).

Le nombre d’occurrences du corpus A est de 2.257.931 et le nombre de vocables est de 51.009. Le taux de répétition est de 0,0226 ($51.009/2.257.931$). Nous trouvons pratiquement le même taux de répétition dans le corpus C : 0.0227

(48.643/2.146.241). En revanche, si nous nous appuyons sur les lemmes du corpus B, le taux de répétition est de 0,0149 (33.652/2.257.931).

Nous mesurons, en premier lieu, la richesse lexicale des sous-ensembles des trois corpus “Le Clézio” avec l’aide du logiciel Hyperbase. Le calcul est exécuté par le programme, suivant la loi binomiale, en s’appuyant sur le tableau de distribution des fréquences et sur l’étendue relative des textes.

<u>Richesse lexicale, classement chronologique.</u>					
n°	réel	théo	écart	réduit	Titre
1	10019	10668	-649	-6.28	PROCÈS
2	10124	11229	-1105	-10.43	FIÈVRE
3	12070	12464	-394	-3.53	DÉLUGE
4	8924	10512	-1588	-15.49	EXTASE
5	9962	11403	-1441	-13.49	FUITE
6	9581	11780	-2199	-20.26	GUERRE
7	1859	3098	-1239	-22.26	MYDRIASE
8	8367	11970	-3603	-32.93	VOYAGE
9	1894	2954	-1060	-19.50	PROPHÉTI
10	6616	11067	-4451	-42.31	MONDO
11	8455	12587	-4132	-36.83	INCONNU
12	767	2107	-1340	-29.19	ICEBERGS
13	646	2147	-1501	-32.39	ARBRES
14	8335	14001	-5666	-47.88	DÉSERT
15	2448	3803	-1355	-21.97	VILLES
16	6424	10018	-3594	-35.91	RONDE
17	9219	13140	-3921	-34.21	CHERCHEU
18	3121	4869	-1748	-25.05	ANGOLI
19	4881	6324	-1443	-18.15	RODRIGUE
20	9259	10296	-1037	-10.22	REVE
21	6136	9063	-2927	-30.75	PRINTEMP
22	749	2082	-1333	-29.21	SIRANDAN
23	7187	9495	-2308	-23.69	ONITSHA
24	7737	12262	-4525	-40.86	ETOILE
25	1899	3207	-1308	-23.10	PAWANA
26	8888	9291	-403	-4.18	DIEGO
27	11170	14732	-3562	-29.35	QUARANTA
28	7578	10203	-2625	-25.99	POISSON
29	8955	9123	-168	-1.76	FETE
30	3841	4557	-716	-10.61	NUAGES
31	7230	8913	-1683	-17.83	HASARD
Tot 50119					

Tableau n°13 : La richesse lexicale du corpus A.

Nous voyons ici que les écarts de richesse sont assez importants. Entre *La fête chantée* n°29 et *Diego et Frida* n°26, les livres qui mobilisent le vocabulaire le plus

“riche”, et *Désert* n° 14 qui, avec *Mondo* n°10 mobilise le vocabulaire le plus “pauvre”, la différence atteint presque quarante écarts-types.

Le tableau ci-dessous, qui s’appuie sur les lemmes, permet de constater à peu près les mêmes écarts entre les livres. Les valeurs des écarts réduits sont un peu moins importantes, ce qui s’explique par la lemmatisation et le regroupement des effectifs sur un nombre d’unités moins important.

<u>Richesse lexicale, classement chronologique.</u>					
n°	réel	théo	écart	réduit	Titre
1	7307	7807	-500	-5.66	PROCÈS
2	7053	8203	-1150	-12.70	FIÈVRE
3	8270	9090	-820	-8.60	DÉLUGE
4	6509	7776	-1267	-14.37	EXTASE
5	7226	8339	-1113	-12.19	FUITE
6	6722	8584	-1862	-20.10	GUERRE
7	1537	2613	-1067	-21.05	MYDRIASE
8	5850	8745	-2895	-30.96	VOYAGE
9	1630	2508	-878	-17.53	PROPHÉTI
10	4285	8081	-3796	-42.23	MONDO
11	6084	9169	-3085	-32.22	INCONNU
12	674	1896	-1222	-28.06	ICEBERGS
13	539	1922	-1383	-31.55	ARBRES
14	5578	10083	-4505	-44.86	DÉSERT
15	1956	3117	-1161	-20.80	VILLES
16	4441	7417	-2976	-34.56	RONDE
17	6451	9513	-3062	-31.39	CHERCHEU
18	2199	3861	-1662	-26.75	ANGOLI
19	3862	4884	-1022	-14.62	RODRIGUE
20	6938	7564	-625	-7.20	RÊVE
21	4429	6767	-2338	-28.42	PRINTEMP
22	687	1871	-1184	-27.37	SIRANDAN
23	5218	7055	-1837	-21.87	ONITSHA
24	5172	8925	-3753	-39.73	ETOILE
25	1457	2683	-1226	-23.67	PAWANA
26	6993	6910	83	1.00	DIEGO
27	7687	10560	-2873	-27.96	QUARANTA
28	5546	7532	-1986	-22.88	POISSON
29	6801	6792	9	0.11	FÊTE
30	3133	3651	-518	-8.57	NUAGES
31	5312	6664	-1352	-16.56	HASARD
Tot 33652					

Tableau n°14 : La richesse lexicale du corpus B.

Au regard du premier tableau, nous pouvons constater que tous les écarts sont négatifs.

“Une fois de plus, écrit Etienne Brunet à propos du vocabulaire de Victor Hugo¹⁹¹, on constate que tous les écarts sont négatifs, ce qui traduit le phénomène partout observé de la spécialisation lexicale : les mêmes mots se trouvent dans les mêmes textes, en vertu des contraintes, thématiques ou stylistiques, de la situation de discours.”

Le deuxième tableau, qui s’appuie sur le corpus lemmatisé, permet toutefois de relever deux exceptions, *Diego et Frida* ainsi que *La fête chantée*. Ces deux ouvrages sont-ils exceptionnellement riches en vocabulaire ? Nous y reviendrons plus tard¹⁹².

Le tableau ci-dessous rend compte de la distribution de la richesse lexicale du corpus C, où les ouvrages courts ont été écartés.

<u>Richesse lexicale</u>					
n°	réel	théo	écart	réduit	Titre
1	10021	10610	-589	-5.72	Procès
2	10128	11168	-1040	-9.84	Fièvre
3	12074	12395	-321	-2.88	Déluge
4	8927	10452	-1525	-14.92	Extase
5	9964	11337	-1373	-12.90	Fuites
6	9582	11711	-2129	-19.67	Guerre
7	8368	11902	-3534	-32.39	Voyage
8	6617	11004	-4387	-41.82	Mondo
9	8456	12514	-4058	-36.28	Inconnu
10	8337	13920	-5583	-47.32	Désert
11	6425	9964	-3539	-35.45	Ronde
12	9222	13067	-3845	-33.64	Chercheu
13	9261	10239	-978	-9.67	Rêve
14	6136	9010	-2874	-30.28	Printemp
15	7192	9444	-2252	-23.17	Onitsha
16	7739	12194	-4455	-40.34	Étoile
17	8897	9242	-345	-3.59	Diego
18	11173	14645	-3472	-28.69	Quaranta
19	7582	10152	-2570	-25.51	Poisson
20	8959	9070	-111	-1.17	Fête
21	7233	8865	-1632	-17.33	Hasard
Tot	48643				

Tableau n°15 : La richesse lexicale du corpus C.

¹⁹¹ É. Brunet (1988) : p. 27

¹⁹² Cf. p. 145.

Les valeurs négatives des écarts réduits sont globalement plus importantes que dans le tableau précédent (sauf pour les textes n°2 et n°3), mais les ordres de grandeur sont conservés et, surtout, le classement des œuvres et leur ordre respectif dans cette échelle de la richesse lexicale restent les mêmes, comme on va le voir sur les graphes suivants.

Le premier histogramme ci-dessous mesure la richesse lexicale du corpus A, celui qui s'appuie sur les formes, le deuxième celle du corpus B, qui s'appuie sur les lemmes ; le troisième, enfin, mesure la richesse lexicale du corpus homogénéisé (le corpus C).

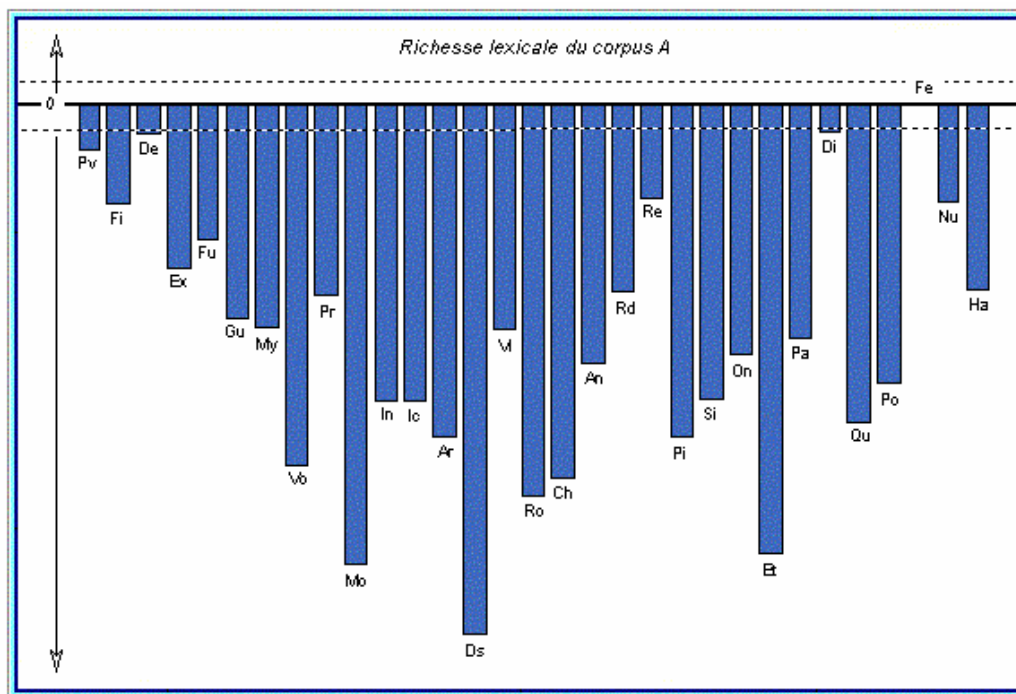


Figure n°23 : La richesse lexicale du corpus A (formes) calculée sur l'étendue relative des textes, suivant la loi binomiale.

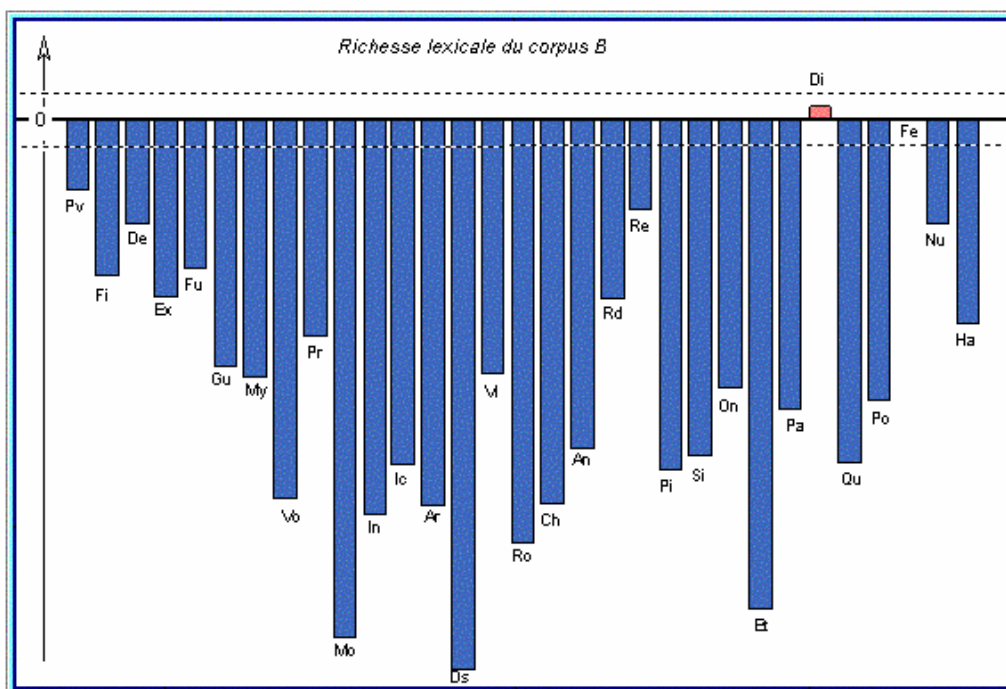


Figure n°24 : La richesse lexicale du corpus B (lemmes).

Il est aisé de constater que les deux histogrammes concernant les corpus A et B sont pratiquement superposables. L'influence de la lemmatisation sur ce calcul est en fait insignifiante dans le cas de notre corpus.

La superposition des deux courbes, celle de la richesse lexicale du corpus A et celle du corpus B permet de constater la parfaite parallélisme des deux courbes :

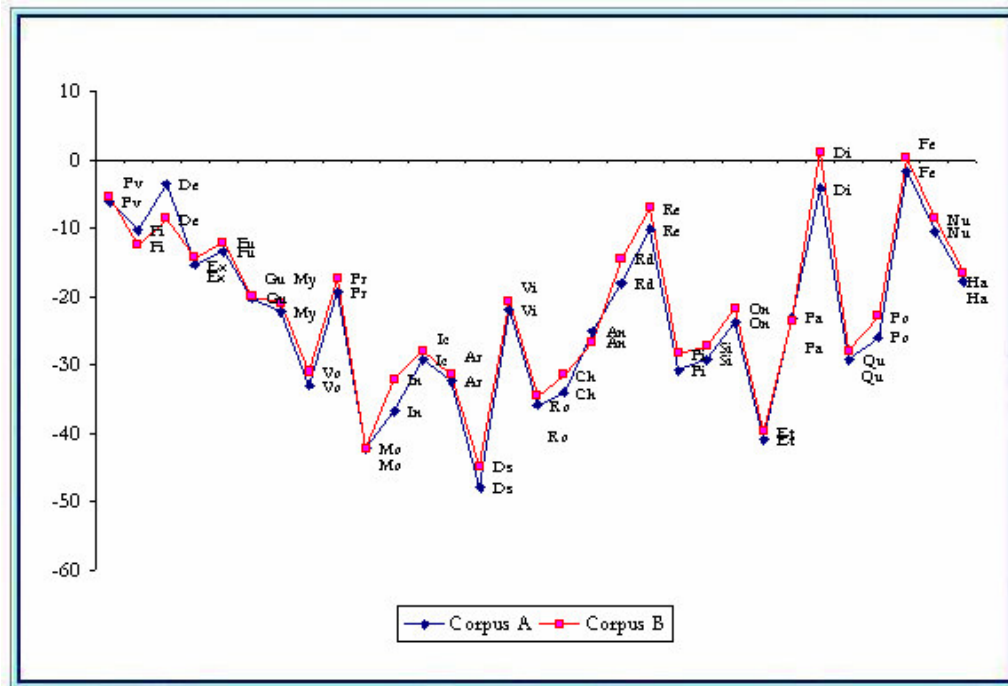


Figure n°25 : Comparaison de la richesse lexicale du corpus A et du corpus B, l'influence de la lemmatisation.

Nous faisons la même observation, en examinant le corpus C. Les histogrammes sont très semblables et l'analyse permet de constater que les différences de taille n'influencent pas l'analyse de la richesse lexicale de façon considérable.

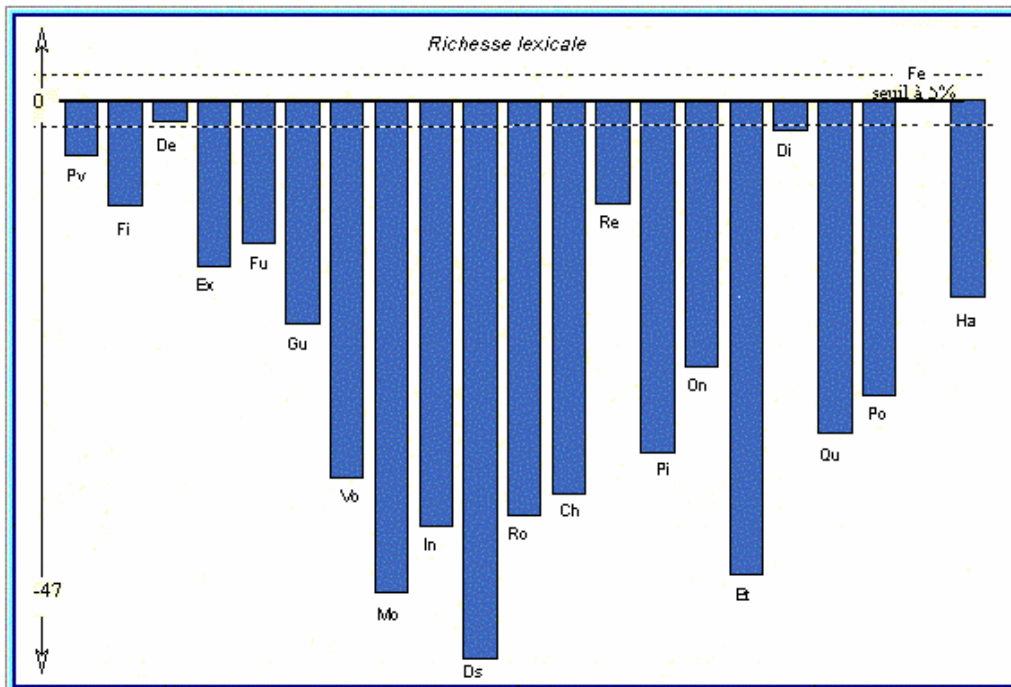


Figure n°26 : La richesse lexicale du corpus C (homogénéisé).

L'analyse est la même dans les trois cas. Nous notons un vocabulaire "riche" au début de l'œuvre qui s'appauvrit pour "s'enrichir" à nouveau vers la fin de l'œuvre, avec des fluctuations toutefois remarquables.

Au début de la carrière littéraire de l'écrivain, dans la période "nouveau roman", nous nous trouvons devant un vocabulaire abondant. L'écriture expérimentale des années 1960 de l'auteur est en fait très "riche" en vocabulaire. Il y a chez lui une volonté manifeste de sortir des formes conventionnelles de la littérature.

"Le propos de l'écrivain, dit J.M.G. Le Clézio à cette époque¹⁹³, est peut-être de joindre le langage et le monde, que le langage de l'écrivain rejoigne le langage des mathématiques... Imaginez ce que les mathématiques ont découvert, imaginez ce que la physique moléculaire a découvert. Imaginez ce qu'est le langage à côté. Qu'a-t-il découvert ? Imaginez ce que la psychologie a découvert et ce que le langage courant lui a découvert. A quoi sert d'avoir un langage de brouette alors qu'autour de lui tout est voiture ? A quoi sert d'avoir un langage de calcul élémentaire alors que tout est calcul infinitésimal."

Au milieu des années 1970, l'écriture de J.M.G. Le Clézio change. Il abandonne les nombreuses expérimentations langagières et le goût des excentricités lexicales, pour adopter un style d'écriture beaucoup plus "simple" et épuré. C'est *Voyages de l'autre côté* (1975) qui ouvre la voie à son changement d'écriture. *Mondo* (1978), avec un écart réduit de -42.3, est le premier livre de l'auteur qui a pour thèmes les enfants et leur univers. Il y développe un style et un langage proche de celui de l'enfant. Le critique Alain Clerval s'exprime ainsi concernant le langage particulier de *Mondo*¹⁹⁴ :

¹⁹³ P. Lhoste (1971) : p. 101.

¹⁹⁴ A. Clerval (1978) : *nrf*, n° 305, juin 1978, in F. Marotin (1995) : p. 159.

“La pensée et l’esprit de système qui se développent parallèlement au langage articulé n’ont pas encore, chez l’enfant, produit le monstrueux accouplement de l’intelligence et de l’appétit de domination, où l’auteur aperçoit, à juste titre, l’origine de la technologie dont nous sommes devenus les esclaves. A l’inverse, les mots utilisés par les enfants-fées de *Mondo et autres histoires* ont un pouvoir magique : ils ouvrent, sans peine, l’accès d’un univers qui n’est plus hostile mais merveilleux. En état de grâce, l’enfant sauvage utilise les mots comme le talisman qui lui permet de s’aventurer jusqu’au cœur du paysage.”

Le livre que l’on peut considérer comme étant le moins diversifié avec un écart réduit de -47,88 est le roman *Désert*. On retrouve ici la même volonté de s’exprimer par un lexique d’une extrême simplicité, presque enfantine.

Dans *Trois villes saintes* nous retrouvons un vocabulaire plus abondant, bien que sans approcher la richesse du début de l’œuvre romanesque. Le vocabulaire des derniers romans du corpus témoigne d’une tendance à la hausse de la richesse lexicale.

La richesse lexicale des différents ouvrages reflète aussi un autre phénomène connu, celui de l’influence du genre dans lequel il s’inscrit. Notre corpus ne fait pas exception à cette règle. 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, comme nous l’avons constaté à propos des œuvres de fiction.

Les deux livres qui se rapprochent de la poésie, que l’on pourrait éventuellement classer dans le genre “poésie en prose” : *Mydriase* et *Vers les icebergs* (écarts réduits

de -22.26 et de -26.19), se trouvent entre les deux groupes de genres cités auparavant.

Les ouvrages de type ethnologique ont un vocabulaire “riche” qui s’explique par la multitude des phénomènes culturels et géographiques qu’ils décrivent. *Diego et Frida*, la biographie qui s’intéresse aux deux peintres mexicains, est riche en thèmes. Elle traite à la fois de la peinture, de l’histoire, de la vie sud-américaine et de la culture indienne, des voyages en Europe et des séjours aux États Unis. De plus J.M.G. Le Clézio utilise beaucoup de termes en espagnol, en anglais aussi bien que des termes en langue amérindienne. *La fête chantée* et *Le rêve mexicain* reflètent le même phénomène, nous sommes toujours dans des récits qui témoignent de deux cultures différentes et de leur impossible rencontre.

Dans ces cas-là, il est évident que la richesse des noms propres, des toponymes et des patronymes joue un rôle important sur les résultats et reflète très nettement ce phénomène d’une richesse lexicale importante dans les ouvrages qui traitent de différentes cultures.

Il ne faut pas se contenter de cette analyse qui “met en avant” ce type d’ouvrages (les ouvrages ethnologiques et les essais qui traitent des cultures différentes). La question que nous nous posons est de savoir quelle serait la richesse lexicale des textes sans ces termes spécifiques.

“On se souviendra, écrit Charles Muller¹⁹⁵, que les calculs théoriques sont fondés sur une hypothèse qui rejette la spécialisation lexicale des fragments et qui, postulant une stabilité parfaite du lexique virtuel, admet que chaque vocable a autant de chance d’apparaître dans un fragment que dans un autre, à longueur égale de ceux-ci. Or il est évident que la réalité s’écarte d’autant plus de cette hypothèse stylistiquement nulle que des causes diverses, thématiques ou

¹⁹⁵ Ch. Muller (1970) : “Sur la mesure de la richesse lexicale”, in *Langue française et linguistique quantitative*, Genève, Slatkine, 1979, p. 302.

stylistiques, modifient cette probabilité.”

En fait, le vocabulaire utilisé dans un corpus est partagé en deux groupes : un vocabulaire général employé quelles que soient les circonstances, et différents vocabulaires “spécialisés” dont chacun est mobilisé dans une partie seulement du corpus¹⁹⁶.

Pour mieux comprendre ces grandes variations dans le vocabulaire leclézien, il est intéressant d’écarter, dans un premier temps, les ouvrages ethnologiques, les biographies et les essais, pour viser plus précisément les livres de fiction et le genre romanesque.

¹⁹⁶ Cf. P. Hubert, D. Labbé “Un modèle de partition de vocabulaire”, in *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, p. 94-114.

2.4.2. La richesse lexicale dans les œuvres romanesques.

Le corpus D, qui regroupe uniquement les romans et les recueils de nouvelles, contient 1.737.646 occurrences et le nombre de vocables (formes graphiques) est de 40.350. Le taux de répétition est de 0,0232 (40.350/1.737.646). Ce taux diffère très peu de ceux que nous avons calculés pour le corpus A et le corpus C.

Le tableau ci-dessous rend compte de la distribution de la richesse lexicale, encore une fois calculée par le logiciel Hyperbase :

Richesse lexicale					
n°	réel	théo	écart	réduit	Titre
1	10021	9893	128	1.29	Procès
2	10128	10402	-274	-2.69	Fièvre
3	12074	11517	557	5.19	Déluge
4	9964	10556	-592	-5.76	Fuites
5	9582	10896	-1314	-12.59	Guerre
6	8368	11069	-2701	-25.67	Voyages
7	6617	10252	-3635	-35.90	Mondo
8	8337	12899	-4562	-40.17	Désert
9	6425	9303	-2878	-29.84	Ronde
10	9222	12127	-2905	-26.38	Chercheur
11	4889	5897	-1008	-13.13	Rodrigues
12	6136	8429	-2293	-24.98	Printemps
13	7192	8827	-1635	-17.40	Onitsha
14	7739	11334	-3595	-33.77	Étoile
15	11173	13555	-2382	-20.46	Quarantaine
16	7582	9474	-1892	-19.44	Poisson
17	7233	8295	-1062	-11.66	Hasard
Tot 40350					

Tableau n°16 : La richesse lexicale du corpus D.

L'illustration graphique à la page suivante (figure n°27) permet la visualisation du mouvement chronologique et rend ainsi l'interprétation des données plus aisée :

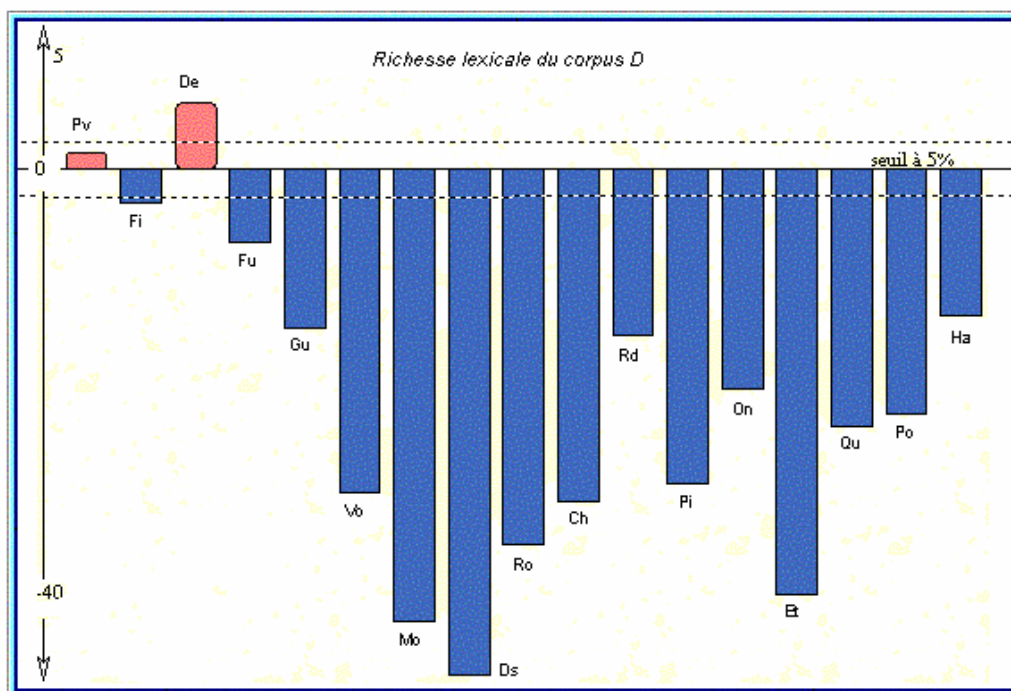


Figure n°27 : La richesse lexicale du corpus D.

La tendance générale est la même que dans les analyses précédentes, un vocabulaire “riche” au début de la production qui tend vers “l’appauvrissement” vers le milieu de l’œuvre pour s’amplifier de nouveau au fur et à mesure que l’œuvre progresse. Cette fois-ci nous relevons deux valeurs positives, il s’agit des romans *Le procès-verbal* et *Le déluge*. Ces valeurs exceptionnelles relèvent sans doute du style particulier du “nouveau roman”. Quand Le Clézio abandonne cette écriture et que son style s’épure, nous constatons les mêmes valeurs négatives importantes qu’auparavant dans les livres de la fin des années 1970 et du début des années 1980. Par la suite nous constatons, avec l’exception du roman *Etoile errante*, la tendance à un vocabulaire plus “riche”. Le roman *Voyage à Rodrigues*, plus “riche” que les autres livres de cette période, abonde en termes marins et techniques et il est le seul livre de Le Clézio écrit sous forme de journal personnel.

Le tableau n°17 rend compte de la richesse lexicale calculée cette fois-ci sur les lemmes dans le corpus romanesque dans sa version lemmatisée, le corpus E :

Richesse lexicale					
n°	réel	théo	écart	réduit	Titre
1	6959	6762	197	2.40	Procès
2	6652	7096	-444	-5.27	Fièvre
3	7798	7833	-35	-0.40	Déluge
4	6782	7179	-397	-4.69	Fuites
5	6324	7399	-1075	-12.50	Guerre
6	5582	7481	-1899	-21.96	Voyages
7	4028	6985	-2957	-35.38	Mondo
8	5277	8637	-3360	-36.15	Désert
9	4225	6431	-2206	-27.51	Ronde
10	6110	8166	-2056	-22.75	Chercheur
11	3704	4281	-577	-8.82	Rodrigues
12	4246	5886	-1640	-21.38	Printemps
13	4995	6128	-1133	-14.47	Onitsha
14	4907	7679	-2772	-31.63	Étoile
15	6588	8311	-1723	-18.90	Quarantaine
16	5305	6525	-1220	-15.10	Poisson
17	5144	5755	-611	-8.05	Hasard
Tot 24767					

Tableau n°17 : La richesse lexicale du corpus E.

Les fréquences de la première et de la deuxième colonnes sont sensiblement réduites par rapport au tableau s'appuyant sur les formes sans pour autant modifier les écarts réduits de façon significative. L'histogramme à la page suivante (figure n°28) permet une comparaison plus aisée :

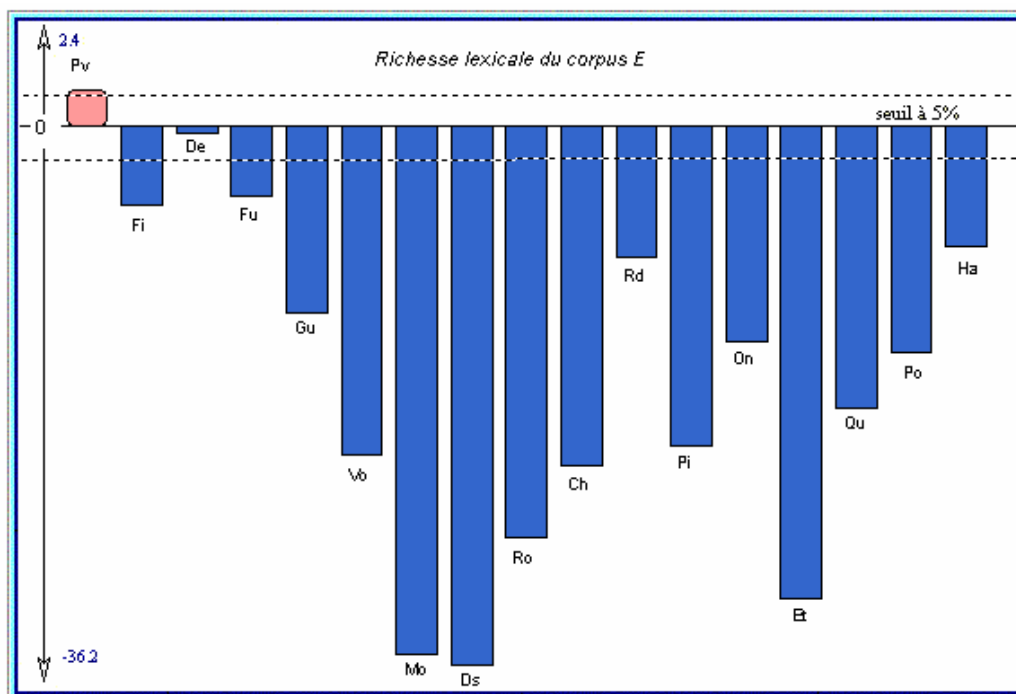


Figure n°28 : La richesse lexicale du corpus E.

Nous pouvons constater le même mouvement que dans l'histogramme de la richesse lexicale du corpus D. Dans un seul cas une différence importante est à noter, il s'agit du roman *Le déluge* qui n'est plus excédentaire, comme il l'était dans l'étude du corpus D, lorsque l'analyse est basée sur les lemmes.

Nous avons pu constater que, même à l'intérieur de ce corpus romanesque, les variations sont importantes et que les mots spécialisés, souvent colorés d'exotisme et des voyages de Le Clézio, jouent un rôle important dans les résultats de nos analyses.

Une autre technique pour analyser la richesse d'un vocabulaire, en tenant compte du vocabulaire général et de différents vocabulaires "spécialisés", est celui du modèle de partition¹⁹⁷.

¹⁹⁷ Cf. P. Hubert, D. Labbé "Un modèle de partition de vocabulaire", in *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, p. 94-114.

Dans la partie suivante nous essayerons d'appliquer cette méthode sur un sous-ensemble du corpus, celui de 12 œuvres (corpus H).

2.4.3. Le modèle de partition

Cette méthode applique un modèle qui s'appuie aussi sur la méthode présentée par Muller en 1964 et 1970. Il compare le nombre de vocables différents dans chacun des textes en raccourcissant les plus longs à la dimension des plus courts, en s'appuyant sur la loi binomiale. Dans le modèle de partition cependant, un paramètre supplémentaire (p) estime le poids relatif des deux vocabulaires (vocabulaire général et vocabulaire spécialisé) : la valeur de ce paramètre donne ensuite une estimation de la spécialisation lexicale d'une œuvre dans le corpus. Ce procédé a pour ambition d'obtenir un chiffre plus exact car la comparaison des richesses lexicales des différentes œuvres du corpus ne serait alors pas influencée par la spécialisation du vocabulaire.

Pour cette analyse, nous avons recours à notre corpus H, c'est-à-dire le corpus basé sur 12 œuvres, lemmatisé selon la méthode Labbé, avec exploitation par le logiciel Lexicométrie. Il est important de tenir compte du fait que, pour des raisons techniques, ce corpus ne couvre pas la totalité du corpus Le Clézio et que ce sont les ouvrages ethnologiques et les essais, contenant une multitude de termes rares, qui ont été écartés de l'étude. Cette analyse, néanmoins, complète celle que nous avons effectuée dans le chapitre précédent.

La comparaison des textes ne peut pas s'appliquer sur des textes de taille trop inégale¹⁹⁸. Elle n'est vraiment sûre que lorsque les différences ne vont pas au-delà de 1/10ème. Comme *La quarantaine* compte plus de 140.000 mots, nous avons

¹⁹⁸ Cf. Ch. Muller (1970) : p. 302 et (1977) : p. 142-144, ainsi que P. Hubert, D. Labbé "Un modèle de partition de vocabulaire", in *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, p. 94-114.

écarté les nouvelles de moins de 15.000 mots. C'est pourquoi la comparaison ne prend en compte que 3 nouvelles, les plus longues : *Printemps*, *Mondo* et *Les bergers*.

Le tableau de synthèse ci-dessous permet la comparaison entre les textes, tenant compte de leur taille respective, et il se lit de la façon suivante :

La première colonne correspond au nombre d'occurrences (N) par texte. La deuxième (V) au nombre de vocables différents dans le texte entier. Les autres colonnes permettent de faire la comparaison des textes, deux à deux, en réduisant le plus grand à la taille du plus petit selon la méthode de Charles Muller.

Par exemple si *La quarantaine* avait la taille de la nouvelle *Les bergers*, soit 19.340 mots, elle compterait 2.774 vocables différents. Le vocabulaire de la nouvelle *Les bergers* n'étant que de 1725 vocables, nous en déduisons que le vocabulaire employé dans *La quarantaine* est nettement plus "riche" que celui de la nouvelle *Les bergers*.

	Taille	V	Quarantaine	Désert	Voyages	Mondo	Fièvre	Procès-verbal	Ronde	Printemps	Hasard	Rodriguez	Printemps	Mondo
Bergers	19340	1725	2774	1798	2106	1841	3006	2901	2151	1887	2729	2723	1912	1824
Mondo (seul)	19626	1837	2796	1812	2124	1855	3030	2927	2167	1906	2752	2745	1928	
Printemps (seul)	31237	2508	3575	2341	2773	2357	3899	3898	2744	2623	3579	3491		
Rodriguez	33494	3616	3704	2432	2885	2442	4045	4066	2841	2753	3719			
Hasard	59602	5061	4912	3328	4011	3260	5413	5722	3750	4122				
Printemps	60342	4158	4941	3351	4039	3280	5446	5763	3772					
Ronde	72502	4110	5381	3704	4488	3594	5950	6409						
Procès-verbal	78783	6726	5589	3877	4708	3746	6189							
Fièvre	85901	6447	5812	4066	4949	3912								
Mondo	85940	3913	5813	4067	4951									
Voyages	98862	5372	6187	4395										
Désert	132595	5183	7027											
Quarantaine	144479	7287												

Tableau n° 18 : Tableau de synthèse du modèle de partition.

Le tableau permet d'obtenir deux classements différents :

- 1. en adoptant une base de comparaison unique - le texte le plus court (*Les bergers*, soit 19340 mots) - nous obtenons le classement du plus “pauvre” au plus “riche” :

Les bergers, *Désert*, *Mondo*, *Printemps*, *Voyages de l'autre côté*, *La ronde*, *Voyage à Rodrigues*, *La quarantaine*, *Hasard*, *Le procès-verbal* et *La fièvre*.

- 2. en suivant la diagonale :

La quarantaine > *Désert* ; *Désert* < *Voyages* ; *Voyages* < *Mondo* ; *Mondo* > *La fièvre* ; *La fièvre* < *Le procès-Verbal* ; *Le procès-verbal* < *La ronde* (total) ; *La ronde* (total) > *Printemps* (total) ; *Printemps* (total) < *Hasard* ; *Hasard* > *Voyage à Rodrigues* ; *Voyage à Rodrigues* > *Printemps 1* ; *Printemps1* > *Mondo1* ; *Mondo1* > *Les bergers*.

La deuxième opération, par transitivité, donne le même résultat que la première, sauf pour les deux premières œuvres qui sont interverties et pour *La ronde* qui change de place. La raison de cet écart de résultat est la spécialisation du vocabulaire, assez forte dans la première, *Le procès-verbal*, et beaucoup plus faible dans la seconde, *La fièvre*. Ces résultats sont à comparer avec ceux obtenus dans le premier classement où le premier livre, *Le procès-verbal*, “emportait” sur le deuxième, *La fièvre*.

Nous pouvons encore une fois constater que l'écriture de Le Clézio opère une sorte de cycle et qu'après la presque “préciosité” de ses premières œuvres, il adopte une grande sobriété qu'il abandonne ensuite progressivement. Entre le texte le plus “riche” et le texte le plus “pauvre”, la différence est supérieure à 50%, ce qui est tout à fait considérable alors même que les ouvrages les plus “riches” ont été écartés de cette étude.

Il convient toutefois de tenir compte de l'inconvénient que présente cette technique et qui a été signalé par Etienne Brunet¹⁹⁹ ainsi que par Dominique Labbé²⁰⁰. Ce dernier écrit que “(...) la technique de “raccourcissement” selon le procédé Muller - du texte le plus long au texte le plus court – “avantage” le premier et cet avantage sera d'autant plus grand que leur différence de longueur est importante ou que la spécialisation lexicale est intense.²⁰¹” Notre corpus semble pouvoir témoigner de cet effet.

A la comparaison avec le même corpus traité par le logiciel Hyperbase et les méthodes décrites auparavant le classement est effectivement assez différent. La différence est-elle due au fait que la spécialisation est prise en compte ? Il semble que le calcul qu'effectue le logiciel Lexicométrie neutralise l'effet de la spécialisation lexicale et permet donc d'avoir une mesure exacte de la propension de l'auteur à diversifier son vocabulaire (ou à l'inverse à simplifier l'expression), autrement dit le choix stylistique. L'introduction de mots rares, de noms propres ou exotiques peut donner l'illusion de la “richesse” mais ne résiste pas à l'examen du logiciel. Toutefois, il est délicat d'en tirer des conclusions trop fermes concernant les différences assez considérables auxquelles conduisent les différents calculs.

A travers l'analyse suivante nous chercherons à situer la richesse lexicale de l'œuvre de Le Clézio dans la littérature française et à faire une comparaison avec d'autres auteurs en recourant à une différente méthode, l'indice W.

¹⁹⁹ É. Brunet (1978) : p. 35-41.

²⁰⁰ P. Hubert, D. Labbé “Un modèle de partition de vocabulaire”, in *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, p.112-113.

²⁰¹ *Ibid.* p. 112.

2.4.4. L'indice W et la comparaison externe.

Une troisième manière de mesurer la richesse lexicale est de calculer l'indice W²⁰². Élaboré par Etienne Brunet, cet indice a été appliqué à un grand nombre d'œuvres, de Chateaubriand à Zola en passant par Giraudoux, Gracq, Hugo et Proust.

Depuis, cet indice W a été un peu abandonné car, selon Etienne Brunet lui-même, “c’est une approximation qui n’est pas indépendante, malgré tout, de l’étendue et surtout parce que c’est un indice fondé sur des données lemmatisées.” Or, souvent les données des corpus traités ne sont pas lemmatisées.

Étant donné que nous disposons d’un corpus lemmatisé et que sa taille correspond à peu près aux corpus auxquels Etienne Brunet a appliqué l’indice, nous trouvons l’expérience intéressante car elle permet la comparaison avec d’autres écrivains.

Notre corpus B contient 33.652 vocables différents pour une étendue de 2.281659 occurrences. L’indice W est donc, selon la formule de Brunet, de 8,40.

Auteur	Indice W
Le Clézio	8,40
Giraudoux	9,43
Hugo	10,13
Chateaubriand	9,71
Proust	9,84
Zola	10,97
Corpus du XX ^e siècle du T.L.F.	13,90

²⁰² Cf. É. Brunet (1978) : p. 45-67. La formule de l'indice W est $W = NV^{-a}$. Il propose plusieurs valeurs, fruits de maints essais, pour “a”. Nous utilisons 0,185, la valeur qu'il utilise dans *Le vocabulaire de Victor Hugo*. Rappelons que la richesse du vocabulaire est d'autant plus forte que W est faible.

La valeur de l'indice W sur le corpus Le Clézio témoigne d'un vocabulaire plus riche que celui des auteurs du XIX^{ème} siècle. Il est aussi plus "riche" que le vocabulaire du Giraudoux et que celui du corpus T.L.F. du XX^{ème} siècle.

Un facteur qui contribue à l'importante variété du vocabulaire leclézien est celui du style "nouveau roman" qui attribue une richesse lexicale quelque peu artificielle. L'exemple ci-dessous tiré de *La guerre*²⁰³ illustre ce phénomène de propagation de vocabulaire :

"C'est simple, tu vois, il suffisait d'y penser. Il y a sûrement un schéma. Peut-être qu'on le trouvera dans les livres, dans la Bible, le Coran, ou bien dans l'Atlas Bartholomew. Ou bien dans les salles de cinéma, quand la lumière s'éteint et qu'au même moment l'écran s'allume dans dix villes différentes, avec, Ivan le Terrible, Le Baron Vampire, Ngu, Viridiana, Le Cri, Caballo Prieto Azabache, Coups de feu dans la sierra, La Mère, Pickpocket, Feux dans la plaine.

Je suis sûre que je finirai par trouver le plan. Il doit avoir la forme d'une ville gigantesque, avec des rues et des boulevards, des ponts, des immeubles, des lignes de chemin de fer ; la carte de Berlin, de Londres, de Tokyo. Le plan d'Helsinki, ou de Bogota. Des avenues longues de quinze kilomètres, comme à Mexico. Des tranchées pleines de boue et d'ordures, comme à Tegucigalpa."

Dans son ouvrage sur le vocabulaire français de 1789 à nos jours, Etienne Brunet a constaté une augmentation constante de la "richesse lexicale" avec le temps²⁰⁴, augmentation qui tiendrait à un phénomène "démographique" propre à toute langue vivante : l'excédent des "naissances de mots" sur les "décès" engendre une "inflation lexicale". Logiquement, nous déduisons que, plus un auteur est contemporain, plus ses textes ont de chances de contenir un vocabulaire étendu

²⁰³ *La guerre*, p. 83.

²⁰⁴ Cf. É. Brunet (1981) : p. 63-71.

(puisque le stock de mots à sa disposition est supérieur à celui des générations précédentes).

L'œuvre de J.M.G. Le Clézio s'inscrit dans cette lignée d'évolution de la langue française.

Il convient donc de tenir compte de cette évolution lors de la comparaison de différents auteurs séparés par un trop grand nombre d'années, car on risquerait d'attribuer un indice, témoignant d'un vocabulaire "riche", à la culture de l'écrivain alors qu'il peut tenir à la différence d'étendue du lexique disponible aux deux époques.

Autrement dit, la comparaison des richesses lexicales ne peut s'opérer que sur les textes des auteurs plus ou moins contemporains, tout au moins lorsque la richesse lexicale du plus contemporain l'emporte sur le plus ancien. "En effet, écrit Dominique Labbé²⁰⁵, l'époque contemporaine offre un vocabulaire d'une variété et d'une étendue sans égale par le passé. Dès lors, il ne faut pas comparer des auteurs d'époques différentes au risque d'attribuer aux plus anciens une moindre richesse qui ne serait que la marque de leur époque."

Dans un cas comme le nôtre la conclusion n'est peut-être pas très significative. L'indice W témoigne d'un vocabulaire très "riche" s'appuyant sur la totalité du corpus.

Les travaux d'Etienne Brunet permettent également de constater des différences assez nettes en fonction du genre auquel appartient l'œuvre analysée. Il a pu relever que la richesse lexicale était toujours supérieure dans la poésie, et plus particulièrement dans la poésie en prose. De même, la prose littéraire l'emportait toujours sur les écrits techniques ou la correspondance. Ces constatations amènent

²⁰⁵ M. Brigidou, Dominique Labbé (2000) : p. 72.

à reformuler, comme le fait Dominique Labbé²⁰⁶, la question de la richesse lexicale :

“...avant de mesurer celles-ci, il faut se demander si les œuvres comparées sont vraiment comparables, c’est-à-dire si elles appartiennent à la même époque et au même genre littéraire.”

En effet, quand nous écartons les ouvrages d’ethnologie, les récits de voyage, la biographie et les essais pour n’étudier que le genre romanesque, le résultat qui en découle est différent. En effet, lorsque l’on renouvelle l’analyse sur notre corpus romanesque, le corpus E, l’indice W est alors de 9,11.

Nous constatons que le vocabulaire est moins diversifié lorsque nous écartons les genres non romanesques. Toutefois, la différence avec le calcul précédent n’est pas significative et l’écart par rapport aux autres auteurs reste d’une telle importance qu’il nous semble imprudent de tirer des conclusions précises de cette analyse de la richesse lexicale ; nous retenons simplement l’idée que les résultats obtenus sur la totalité du corpus s’inscrivent dans la lignée de l’évolution de la langue littéraire française.

Il convient donc, avant d’apporter des conclusions sur la richesse du vocabulaire de notre corpus, d’analyser d’autres notions qui sont importantes pour comprendre la richesse lexicale. Un autre facteur non négligeable est celui de la diversité du vocabulaire.

²⁰⁶ D. Labbé, “La richesse du vocabulaire politique.”, in *Mots chiffrés et déchiffrés*, S. Mellet, M. Vuillaume (éds.), Paris, Honoré Champion, 1998, p. 175.

2.4.5. La diversité du vocabulaire.

“C’est peut-être la plus grande souffrance de l’homme de ne pouvoir être dans la forme sans appartenir à la diversité, et de ne pouvoir s’enfourer dans le grouillement sans regretter les systèmes.”

L’extase matérielle, page 59.

Nous ajoutons donc à la mesure de la richesse lexicale une autre notion, celle de la propension à diversifier l’expression, c’est-à-dire à éviter ou au contraire à chercher la répétition des mêmes mots sur de courtes périodes.

Pour cette étude, nous abandonnons un instant la méthode qui s’appuie sur la loi binomiale (comparaison des textes deux à deux en réduisant le plus grand à la taille du plus petit) pour une autre technique.

Il s’agit de l’indice global de diversité du vocabulaire qui, pour un texte donné, mesure l’étendue moyenne du lexique mobilisé pour la réalisation de ce texte. Pour neutraliser l’effet de la longueur et utiliser un étalon commun entre les textes comparés, on simule le prélèvement d’échantillons de même longueur dans ces textes et l’on observe le nombre de vocables différents présents dans ces extraits²⁰⁷. Cet indice de diversité du vocabulaire est le nombre moyen de vocables différents observés dans tous les échantillons possibles de 1000 mots contigus que l’on peut découper dans le texte étudié. Dans une perspective probabiliste — pour intégrer les fluctuations normales d’échantillonnage — on associe à cet indice un intervalle de confiance (centrée sur l’axe horizontal, la moyenne) égal à deux écarts-type qui correspond à la fluctuation “normale” du vocabulaire.

²⁰⁷ Pour plus de détails sur ce calcul, cf. P. Hubert, D. Labbé (1990) et (1994).

Nous avons, cette fois-ci, recours à notre corpus D (corpus de 12 oeuvres, lemmatisé selon la méthode Labbé) et le calcul est effectué par le logiciel Lexicométrie²⁰⁸, avec le modèle de partition.

Le tableau ci-dessous permet la visualisation de la diversité du lexique dans notre corpus. Sur l'axe des ordonnées sont représentées les valeurs de l'indice et sur l'axe des abscisses sont indiquées les tranches de mots :

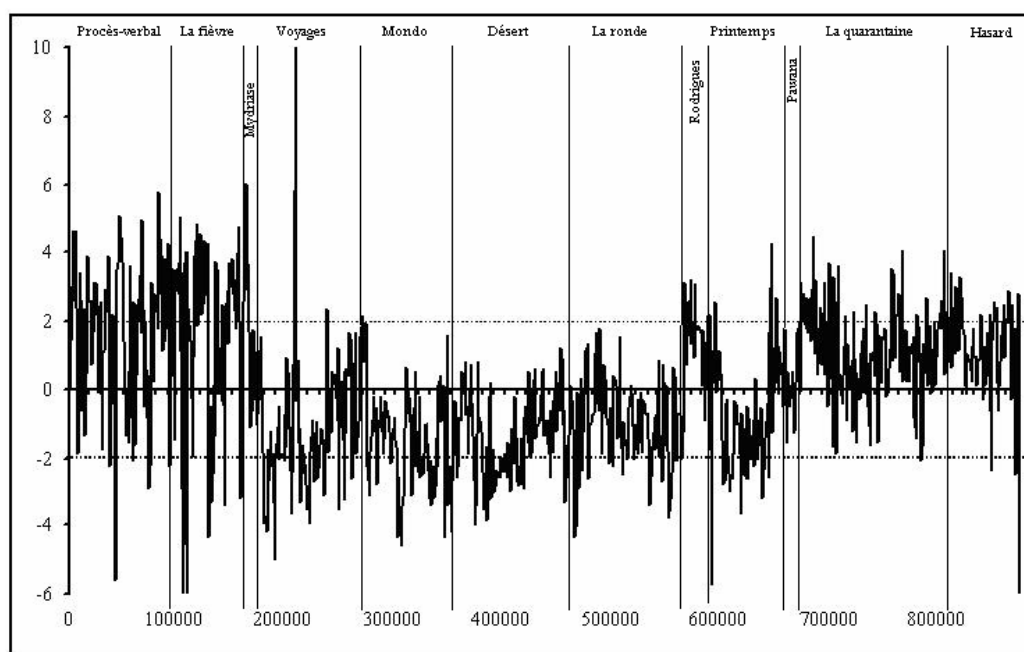


Figure n°29 : La diversité du vocabulaire du corpus H.

Chez Le Clézio, nous pouvons constater que la diversité moyenne n'est pas très élevée par rapport à celle d'autres textes littéraires, pour lesquels cette courbe prend normalement un aspect plus heurté et saccadé²⁰⁹. Les "pics" anormaux sur la courbe correspondent aux passages où l'auteur s'écarte significativement de son style habituel. Pour les "pics" vers le haut il peut s'agir de l'introduction d'un

²⁰⁸ Cf. R. Alvarez, M. Bécue, J. J. Lanero, "Vocabulary diversity and its variability: A tool for the analysis of discursive strategies. Application to the investiture speeches of the Spanish democracy" in *JADT 2000, 5^{es} journées internationales d'Analyse statistique des Données Textuelles*, 9-11 mars 2000, M. Rajman, J.-J. Chappelier: (éds.), EPFL, Lausanne, p. 111-118.

²⁰⁹ Cf. P. Hubert et D. Labbé (1994)

nouveau thème ou tout simplement d'un passage en langue étrangère et les "pics" vers le bas de l'épuisement d'un thème²¹⁰.

De manière générale, une faible diversité du vocabulaire est l'indice d'un texte à dominante scientifique ou pédagogique. A l'inverse, un indice élevé peut signaler par exemple une visée polémique ou poétique.

Nous retrouvons ici le même mouvement, de façon toutefois moins nette que dans l'étude de la richesse lexicale du chapitre précédent, qui nous permet de découper l'œuvre en trois grandes périodes, cette fois-ci du point de vue de la diversité comme indice synthétique du style.

La première période correspond aux deux premières œuvres, *Le procès-verbal* et *La fièvre*. Elle est caractérisée par une moyenne plus élevée, signe d'une recherche dans l'écriture, voire d'un "forçage" par rapport à la pente "naturelle" de l'auteur qui n'est pas à la recherche d'une grande diversification dans l'ensemble. Nous observons des fluctuations brutales et de grande ampleur autour de l'axe moyen. Il s'agit encore une fois de cette écriture particulière qui s'approche du style "nouveau roman". Adam Pollo, le protagoniste du *Le procès-verbal*, raconte des pans de sa vie en récits successifs ("... il continue à raconter son histoire pour des autres, parodiant à bon marché des souvenirs troubles"²¹¹). *La fièvre* prolonge cette technique, mettant l'accent sur l'artifice de la structure. La condamnation du style et du récit conventionnels par Le Clézio durant cette période avec une écriture riche en fragments disparates est bien reflétée par la courbe. Le Clézio écrit ceci en 1967²¹² à propos de son style et de sa volonté d'expression :

"En écrivant, aussi : en écrivant pour accomplir cette destinée cachée,
en écrivant pour couvrir le temps, pour couvrir l'espace, pour couvrir
tout ce qui existe avec des signes délicats et puissants...en écrivant à

²¹⁰ En annexe n°2 ; les passages correspondant aux pics les plus "hauts" et les plus "bas" des trois périodes.

²¹¹ *Le procès-verbal*, p. 39.

²¹² *L'extase matérielle*, p. 279-280.

nouveau ce que les autres, ce que nous-mêmes avions mille fois écrit. Ces lignes voulaient (...) recopier matériellement le présent, rendre tangible l'intouchable. Elles voulaient que tout ce qui existe soit arraché dans l'instant de son existence au présent qui s'enfuit et le dénommer à jamais.”

Pourquoi alors ne pas recopier des horaires avec le prix des billets, des tableaux synoptiques, des textes publicitaires ? “Cette main qui écrivait voulait accumuler les mots comme des coups, pour voiler la face de la vérité, pour dissimuler l’abîme de joie et de malheur...”²¹³

Dans la deuxième période, de *Voyages de l'autre côté* à *Voyage à Rodrigues*, nous remarquons une diversité basse avec des variations de faible amplitude. Cela correspond au changement de l’écriture leclézienne, à la rupture dans l’œuvre. L’auteur renonce à l’écriture expérimentale de la première période, aux sortes de “collages” que le logiciel repère dans la première partie. De même, sans renoncer aux dialogues “pauvres”, il les rend moins caricaturaux. Cette nouvelle écriture est caractérisée par une très grande simplicité dans l’expression. Simone Domange²¹⁴ fait la remarque suivante à propos de la diversité dans *Désert*, autre ouvrage de cette période :

“...le registre familier, phrases à vocabulaire et constructions simples, frappe le lecteur - d’autant plus que, malgré ces traits facilement repérables, l’écriture de *Désert* ne se calque pas, c’est le moins que l’on puisse dire, sur le langage de tous les jours.”

Dans la troisième période, à partir de *Voyage à Rodrigues*, la diversité augmente, à l’exclusion des trois premières nouvelles de *Printemps et autres saisons* qui semblent appartenir à la deuxième période (diversité tout à fait en bas de la plage de variation normale). La diversité continue à monter légèrement à partir de *Pawana*.

²¹³ *L’extase matérielle*, p. 280.

²¹⁴ S. Domange (1993) : p. 122.

S'agit-il d'un retour à une certaine recherche dans l'écriture ou bien de l'introduction de thèmes nouveaux ? Les variations toutefois de faible ampleur montrent que le style semble plus apaisé et relativement régulier vers la fin du corpus, en terminant par *Hasard*.

Il est intéressant de regarder de près la diversité à l'intérieur du corpus et ainsi d'isoler le profil des différents livres²¹⁵. Le calcul est toujours le même, à savoir le nombre de mots par échantillons de 1000 mots. Il est à noter que, comme cette méthode ne tient compte que d'un échantillon à la fois, il s'agit de mesurer la variété à l'intérieur de chaque sous-corpus et non plus de constater un renouvellement dans l'œuvre.

Le procès-verbal, le premier roman, celui que l'étude de la richesse lexicale selon la méthode binomiale a classé comme le roman le plus "riche" du corpus, fait ici apparaître une grande unité du vocabulaire lorsqu'il est pris en compte isolément.

Nous pouvons toutefois remarquer quelques ruptures. Notons par exemple la chute suivie par un pic entre le 42000e et le 45000e mot. Elle correspond à une rupture dans le récit, où la déambulation du jeune homme tourmenté s'arrête un instant pour suivre la pensée et le rêve. Aussitôt après l'action reprend avec l'ouverture du nouveau chapitre (N.)²¹⁶ comme on peut le constater sur le graphique suivant.

²¹⁵ En annexe n°3 : le profil de la diversité du vocabulaire des 12 livres du corpus H.

²¹⁶ Les chapitres du romans sont nommés par des simples lettres. Il s'agit ici du 14ème chapitre.

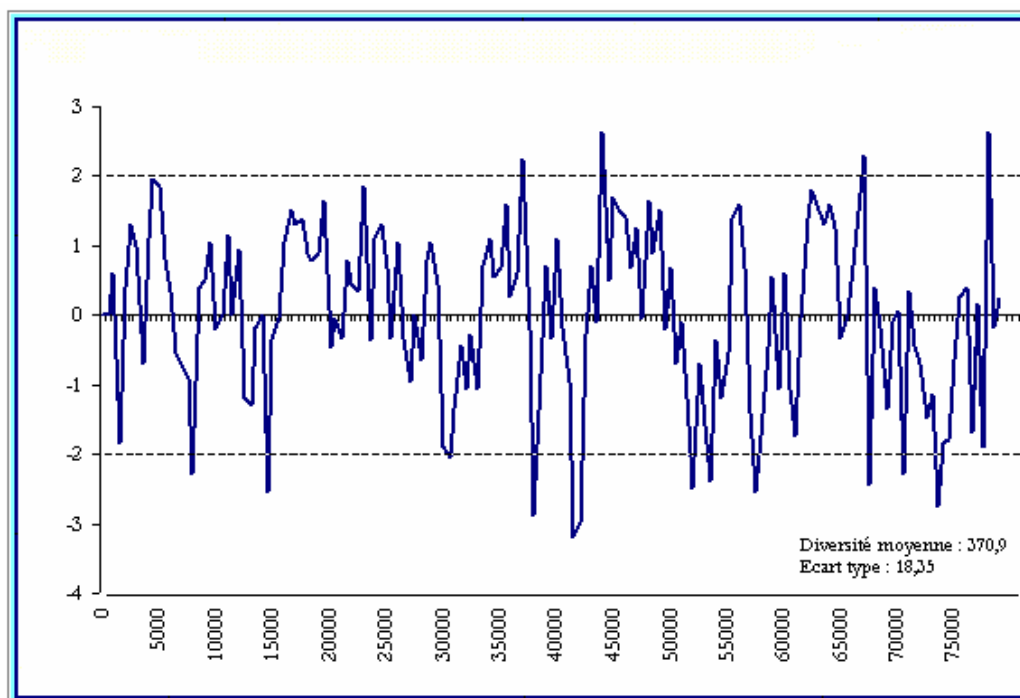


Figure n°30 : La diversité dans le roman *Le procès-verbal*.

Le roman *Désert*, caractéristique de la deuxième période, fait aussi preuve d'une grande unité. A un seul moment du récit, nous relevons des valeurs qui dépassent les deux écarts-types. Il s'agit du passage où les hommes bleus sont écrasés par les soldats occidentaux, riche en noms propres et en toponymes.

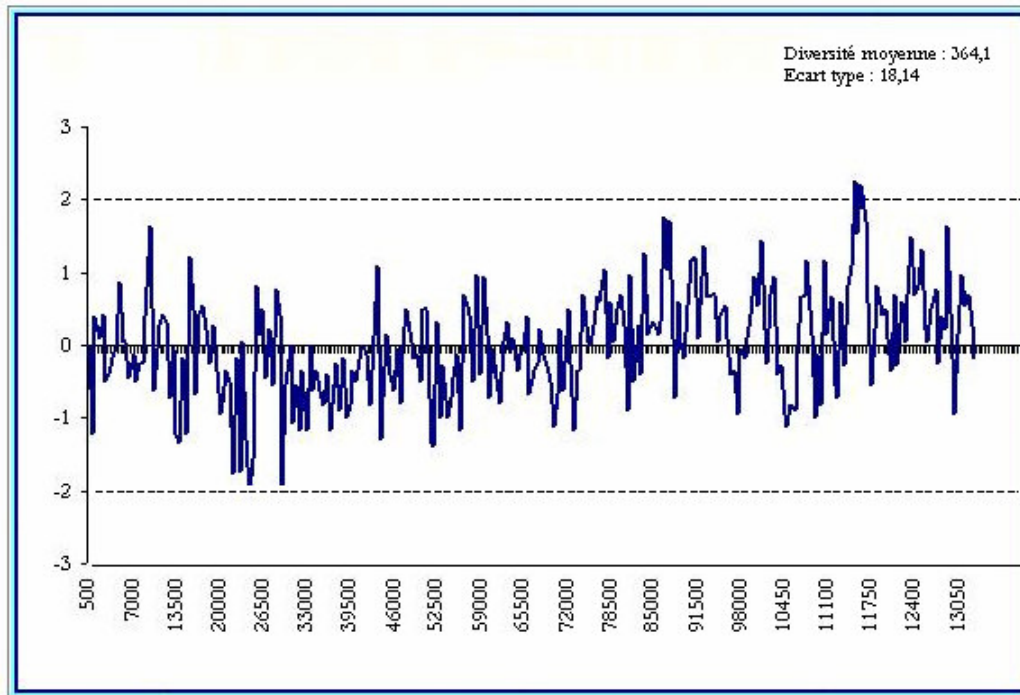


Figure n°31 : La diversité du roman *Désert*.

Le recueil de nouvelles *Printemps et autres histoires* montre également des valeurs très unies lorsque nous l'observons individuellement. Contrairement à ce que l'étude de la totalité du corpus indiquait, nous sommes ici en face d'un récit sans ruptures spectaculaires. La division du recueil en cinq parties correspond aux cinq nouvelles du livre. Cette division s'accorde avec les fluctuations du graphique. A l'intérieur de chaque nouvelle, la diversité semble maintenue à un palier relativement homogène. Le seul dépassement du seuil de signification des deux écarts-types a lieu dans la nouvelle *La saison des pluies*, la cinquième du recueil, et correspond à la découverte de la France par la jeune Gaby.

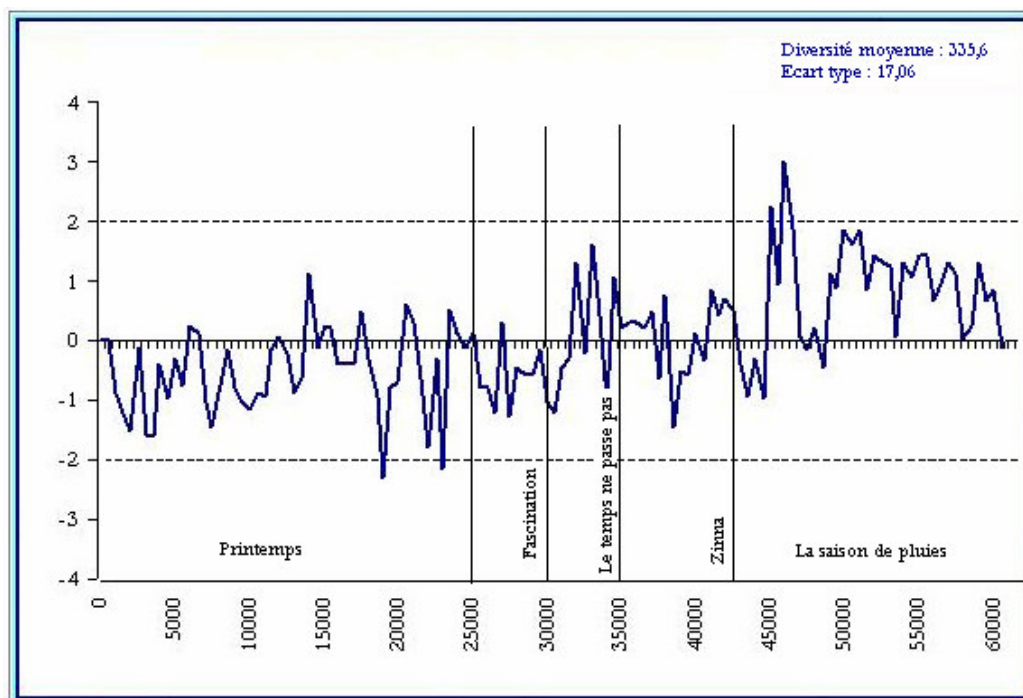


Figure n°32 : La diversité du recueil Printemps et autres saisons.

Cette deuxième période dans l'écriture leclézienne, caractérisée par le dépouillement et un style très sobre, est intéressante à étudier au point de vue de la richesse du vocabulaire car elle reflète l'image de la langue de l'écrivain dépourvue d'innovations et de toute fioriture.

Nous nous proposons d'étudier maintenant plus en détail le recueil de nouvelles *Mondo et autres nouvelles* qui, avec *Désert*, était l'ouvrage qui présentait le vocabulaire le plus "pauvre" dans notre étude effectuée selon la méthode binomiale.

Les huit nouvelles du recueil sont, cette fois-ci, prises en compte, afin de pouvoir comparer tous les petits textes entre eux.

Le tableau à la page suivante (tableau n°19) rend compte des résultats obtenus avec le même calcul sur le vocabulaire des nouvelles composant le recueil *Mondo* :

	Taille	V	Mondo seul	Bergers	Lullaby	Hazaran	Montagne	Celui qui...	Peuple	Roue
V (1000)	-	-	336	329	362	302	338	339	309	340
Roue	5 108	865	912	868	989	770	915	927	794	
Peuple	7 057	943	1 089	1 031	1 180	913	1 089	1 109		
Celui qui...	7 285	1 128	1 108	1 048	1 200	928	1 107			
Montagne	7 495	1 124	1 125	1 064	1 219	942				
Hazaran	8 440	1 001	1 198	1 132	1 298					
Lullaby	11 589	1 527	1 413	1 332						
Bergers	19 340	1 725	1 837							
Mondo (seul)	19 626	1 837								

Tableau n°19 : Comparaison du vocabulaire des nouvelles du recueil Mondo et autres histoires.

Le nombre d'occurrences est donné en première colonne. La seconde colonne présente le nombre de vocables différents dans le texte entier. La diversité du vocabulaire est donnée dans la première ligne (nombre de vocables différents dans n'importe quelle tranche de 1000 mots). Les autres colonnes permettent la comparaison des nouvelles deux à deux, en réduisant la plus grande à la taille de la plus petite.

Pour interpréter les résultats, nous avons recours à l'échelle suivante, établie par Dominique Labbé à partir du dépouillement de textes politiques et de romans contemporains pour l'écrit et, pour l'oral, d'entretiens avec des hommes politiques et des responsables syndicaux (radio-télévisés ou enregistrés...) ²¹⁷ :

²¹⁷ Cf. D. Labbé (1990) : p. 114-115.

	Oral	Ecrit
Très pauvre	inférieur à 280	inférieur à 340
Pauvre	280 - 300	340-360
Moyen	300 - 320	360 - 380
Riche	320- 340	380 - 400
Très riche	supérieur à 340	supérieur à 400

Selon ce tableau, toutes les nouvelles de *Mondo* ont donc un vocabulaire très “pauvre”, sauf *Lullaby* qui présente une valeur plus élevée mais dont la diversité du vocabulaire n’en reste pas moins “juste moyenne”. Cette caractéristique de pauvreté du vocabulaire semble assez constante chez Le Clézio durant la deuxième période de l’œuvre. A partir de ce moment de sa production littéraire, elle reflète sa recherche d’un style dépouillé dans le vocabulaire et l’expression. Nous constatons, en fait, les mêmes signes de dépouillement, la même recherche manifeste par l’auteur d’un vocabulaire “simple et restreint”, que nous avons obtenus lors de l’étude de la richesse du vocabulaire selon la loi binomiale. L’étude de ce recueil de nouvelles illustre bien la rupture stylistique de Le Clézio et son abandon définitif de la forme “nouveau roman”.

Cette analyse nous permet de conclure que la propension à diversifier le vocabulaire est assez faible. Il est en fait caractérisé par des répétitions et des redondances dans les dialogues aussi bien que dans les monologues, ainsi que des séquences très répétitives tout au long de l’œuvre. Nous relevons des passages, surtout au début de l’œuvre, où l’auteur “colle” de nombreux mots étrangers, des noms propres et des mots rares. La deuxième période est caractérisée par un vocabulaire “pauvre”. La narration, durant cette période, se réduit souvent à une suite d’événements banals ou de petites anecdotes, comme dans *Mondo* ou dans *La ronde*. En revanche, la découverte de l’univers, liée à l’exercice du regard et à la contemplation, donne de l’importance aux descriptions. Notons, par exemple, l’énumération des trésors de la mer lors des explorations de Daniel dans *Celui qui*

*n'avait jamais vu la mer*²¹⁸, ou dans *Les bergers* les longues contemplations du ciel étoilé²¹⁹. Vers la fin du corpus, nous pouvons observer une légère “hausse” de l'indice de la diversité du vocabulaire.

Cependant, il convient, avant de tirer des conclusions définitives sur la richesse lexicale de Le Clézio, de prendre en compte un autre facteur important pour caractériser l'évolution dans le temps du vocabulaire : la mesure de l'accroissement lexical.

²¹⁸ *Mondo et autres histoires*, p. 178-179.

²¹⁹ *Ibid.*, p. 281-284.

2.5. L'accroissement lexical.

“Ce langage n’est plus discursif mais répétitif. Il opère la grande métamorphose que l’homme, au fond, ne cesse d’espérer, et qui le libère de sa prison anthropomorphique (...). Il approche de cet instant miraculeux, mythique, où l’ancêtre des hommes a créé le monde : étant l’instigateur de toutes les formes d’existences, et l’inventeur de toutes les formes du langage.”

Les cris des crapauds,²²⁰ page 25.

Selon Charles Muller²²¹, “on admet qu’à tout moment de l’acte de discours, le locuteur a un certain lexique “en jeu”, dans lequel il puise et dont chaque unité a une probabilité d’emploi non nulle. L’étendue et la structure de ce lexique déterminent l’apport lexical, qui est plus grand si le lexique est plus étendu, mais qui est plus faible si les probabilités d’emploi sont inégalement réparties.”

L’étude de l’accroissement lexical détermine donc 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.

Le calcul d’accroissement du vocabulaire peut être fondé sur le même modèle que celui de la richesse lexicale, grâce à la formule binomiale²²². La représentation des données permet ici aussi de localiser les ruptures thématiques dans le corpus, là où se produit un afflux de vocables nouveaux. A l’inverse, les fragments où l’accroissement est inférieur aux valeurs théoriques signalent l’épuisement d’un thème.

²²⁰ J.M.G. Le Clézio in *Les Cahiers du Chemin* n° 12. 15.4.1971, p. 23-38.

²²¹ Ch. Muller (1977) : p. 130.

²²² Cf. le chapitre précédent, et Ch. Muller (1964) : “Calcul des probabilités et calcul d’un vocabulaire”, in *Langue française et linguistique quantitative*, Genève, Slatkine, p 167-176.

Les événements thématiques interviennent toujours dans la structure du vocabulaire en modifiant l'apport lexical. Nous essaierons donc de situer les ruptures thématiques qui ont provoqué soit un afflux de nouveaux vocables, soit l'épuisement d'un thème.

Dans cette étude de l'accroissement lexical, nous aurons recours aux différentes versions de notre corpus. Pour une étude sur la totalité de l'œuvre, nous utiliserons notre corpus de 31 œuvres basé sur les formes graphiques. Ensuite, nous nous intéresserons au corpus exclusivement romanesque et enfin, pour comparer différentes méthodes et techniques, nous exploiterons les différentes versions du corpus de douze œuvres²²³ s'appuyant sur des données lemmatisées.

2.5.1. L'accroissement du vocabulaire de Le Clézio à partir des formes graphiques.

Observons dans un premier temps l'accroissement chronologique de notre corpus de 31 œuvres s'appuyant sur les formes graphiques.

L'accroissement est, dans ce premier cas, calculé par un simple ajustement de courbe en choisissant pour jalons les césures naturelles du corpus. Ce sont les textes eux-mêmes qui constituent les tranches. Les décomptes sont établis, par le logiciel Hyperbase, chaque fois que l'on passe d'un texte à un autre. Le programme recherche la première apparition de chaque mot et, selon qu'il l'a trouvé dans tel ou tel texte, il augmente l'effectif d'une unité.

Le tableau ci-dessous rend compte de l'accroissement du vocabulaire dans l'ordre chronologique. Ici le calcul fait appel à un ajustement des deux séries parallèles (vocabulaire cumulé et étendue cumulée) grâce à une fonction-puissance de type : $y = ax^b$ pour $x = \text{vocabulaire cumulé}$ et $y = \text{étendue cumulée théorique}$. L'écart

²²³ Cf. chapitre 2.2.5. "Les différentes versions du corpus".

entre étendue théorique et étendue réelle est alors calculé pour chaque texte, puis pondéré par l'étendue de chaque texte²²⁴ :

<u>Accroissement du vocabulaire</u>							
ACCROISS. CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
PROCES	0018	10018	10018	90985	90985	-13689	-1.50
FIÈVRE	5529	10111	15547	99938	190923	17805	1.78
DÉLUGE	4750	12079	20297	121853	312776	25060	2.06
EXTASE	2896	8920	23193	90315	403091	20584	2.28
FUITE	2778	9982	25971	103175	506266	18656	1.81
GUERRE	2301	9566	28272	109093	615359	3407	0.31
MYDRIASE	186	1870	28458	9835	625194	-279	-0.28
VOYAGE	1844	8360	30302	113033	738227	-14538	-1.29
PROPHÉTIES	438	1916	30740	8734	746961	15667	17.94
MONDO	869	6615	31609	97170	844131	-47608	-4.90
INCONNU	1448	8499	33057	123952	968083	-37967	-3.06
ICEBERGS	50	781	33107	2707	970790	338	1.25
ARBRES	38	661	33145	2946	973736	-628	-2.13
DÉSERT	1229	8338	34374	149412	1123148	-72863	-4.88
VILLES	251	2462	34625	15294	1138442	718	0.47
RONDE	605	6433	35230	82593	1221035	-43465	-5.26
CHERCHEUR	1810	9230	37040	133199	1354234	-11655	-0.88
ANGOLI	208	3135	37248	24133	1378367	-9733	-4.03
RODRIGUES	576	4886	37824	37987	1416354	2353	0.62
REVE	2860	9273	40684	85723	1502077	124747	14.55
PRINTEMPS	752	6135	41436	69563	1571640	-11401	-1.64
SIRANDANES	64	761	41500	2465	1574105	2539	10.30
ONITSHA	1053	7190	42553	75201	1649306	8364	1.11
ETOILE	890	7744	43443	117607	1766913	-45169	-3.84
PAWANA	115	1903	43558	10558	1777471	-1077	-1.02
DIEGO	1996	8903	45554	72348	1849819	96631	13.36
QUARANTAINE	1707	11183	47261	163846	2013665	-12677	-0.77
POISSON	1050	7578	48311	85039	2098704	11006	1.29
FETE	1629	8966	49940	70064	2168768	83570	11.93
NUAGES	390	3846	50330	21538	2190306	16080	7.47
HASARD	679	7235	51009	67625	2257931	-1357	-0.20

Fonction $y=a(x \text{ exposant } b)$: $a=2.90048898388014e-004$ $b=2.1060101417775$
 $r^2=0.989405109759551$ $r=0.994688448590588$

Tableau n°20 : L'accroissement lexical du corpus A.

La représentation graphique ci-dessous permet une observation plus aisée de l'accroissement.

²²⁴ Le programme ne donne pas des valeurs théoriques et l'écart est isolé pour chaque terme de la série.

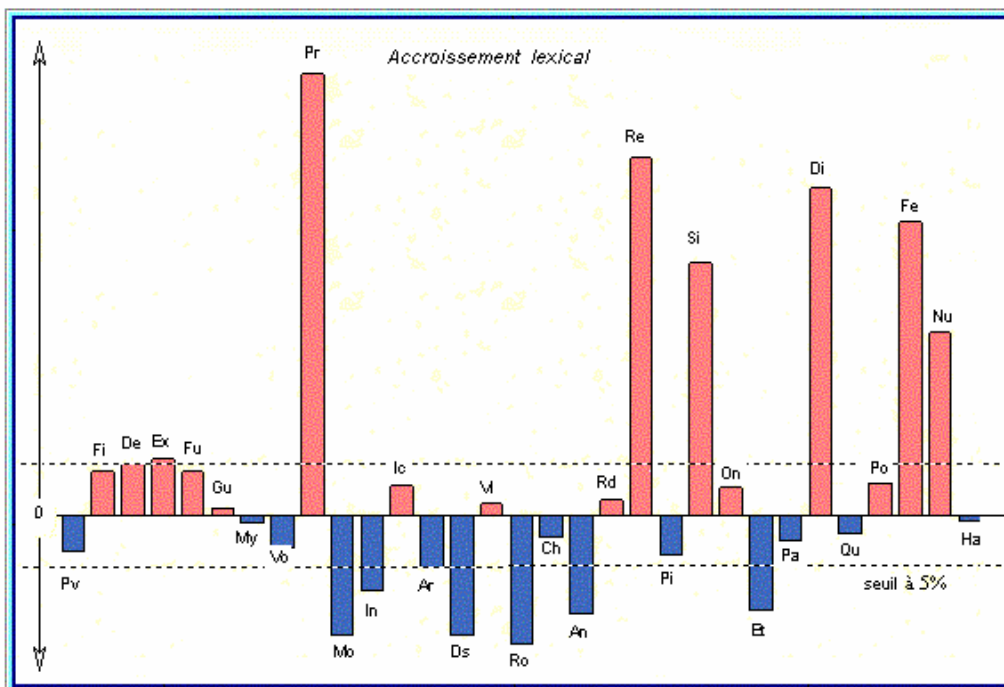


Figure n°33 : Accroissement lexical du corpus A.

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 dans le sens négatif.

Le vocabulaire s'accroît dès le début de la production. Nous constatons, en effet, un apport important au passage du roman *Le procès-verbal* à *La fièvre*. Cet apport est maintenu dans les romans qui suivent, les romans de l'école du “nouveau roman”, ainsi que dans l'essai *L'extase matérielle*. Les œuvres se succèdent, par la suite, sans significativement apporter de vocabulaire nouveau jusqu'à *Mydriase* et *Voyages de l'autre côté*. Là, l'apport lexical ralentit et nous relevons des valeurs négatives.

Ensuite, au neuvième livre nous pouvons observer un grand “pic” correspondant au livre *Les Prophéties de Chilam Balam*, avec un apport lexical très important (valeur de l'écart réduit de +17,94). Ceci s'explique par la grande spécialisation du texte. En effet, ce livre marque un événement décisif dans la vie et dans la production

littéraire de Le Clézio : la découverte du monde amérindien qui influencera, par la suite, sa conception de la vie autant que son écriture. Dans ce cas, l'écrivain commente sa traduction d'un des grands livres sacrés du peuple Maya. Nous nous trouvons devant une véritable explosion lexicale. J.M.G. Le Clézio explique cette culture ancienne avec des termes jusqu'alors inconnus dans le corpus. L'extrait ci-dessous rend compte de sa particularité lexicale :

“Et, lorsque les jours Kan, Muluc, Hix, Cauac étaient revenus chacun pour la quatrième fois, une nouvelle roue tournait, celle des cycles de 52 ans. Enfin, la roue comprenant treize Katun était celle des Ahau Katun (260 ans)”²²⁵

La chute la plus notable d'apport lexical arrive aussitôt après avec *Mondo et autres histoires*, ce qui correspond à une rupture nette dans l'écriture et à un tournant dans l'œuvre de Le Clézio. Cette rupture, selon notre analyse, semble être de caractère stylistique étant donné que l'écrivain utilise son “ancien” vocabulaire. L'apport très faible en mots nouveaux suggère que les thèmes de *Mondo* étaient déjà présents dans certaines œuvres précédentes. Ce recueil de nouvelles est en effet dépourvu de tout exotisme. Nous retrouvons l'espace des premières œuvres. Le cadre est, de nouveau, la ville occidentale, peut-être Nice. La tendance littéraire de cette époque, la fin des années 70, vise au dépouillement de l'écriture avec le renoncement aux conventions romanesques et au pittoresque ; le discours exotique n'est plus en faveur.

L'écriture des œuvres où les personnages sont souvent des enfants ou des êtres énigmatiques et symboliques semble être souvent répétitive et témoigne d'une recherche “musicale” qui donne au récit un caractère incantatoire.

Le roman *Désert* de la même période rend compte de la même retenue lexicale.

²²⁵ *Les prophéties de Chilam Balam*, p. 17.

Simone Domange²²⁶ s'exprime ainsi à propos de la "certaine pauvreté" qu'elle constate dans ce roman :

"Une certaine pauvreté, justifiable, croit-on d'abord, par la jeunesse de l'héroïne, le fait que ce n'est pas une intellectuelle, naît en apparence de ce choix lexical et syntaxique ; mais très vite, ces reprises de mots, de constructions, deviennent incantatoires."

Les livres de cette période sont caractérisés par les répétitions et des reprises fréquentes. A propos du nom de *Lullaby*, une des nouvelles de *Mondo*, qui veut dire "berceuse" en anglais, une remarque faite par l'auteur lors de la publication du recueil semble être révélatrice : "Il y a (...) le langage qui a un caractère musical ou d'incantation, j'aime assez l'idée selon laquelle le langage ne sert qu'à soi-même pour se bercer"²²⁷.

Relevons quelques phrases du roman *Désert* comme exemple de ce style incantatoire : "Mais Lalla aime être dehors ces jours-là." (p. 109), "La lumière est belle, ici, sur la Cité, tous les jours." (p. 118), "Lalla aime bien marcher entre les collines." (p. 128), "Lalla aime le feu." (p. 133) et "C'est l'eau qui est belle, aussi." (p. 150).

"L'un des dangers de ce type d'écriture est de ne pas pouvoir se renouveler d'une œuvre à l'autre", écrit Michelle Labbé²²⁸. "Ne risque-t-elle pas de s'enfermer dans cette alternative. Se taire ou de se répéter ?" Nous ne pouvons que confirmer ce propos, grâce aux statistiques.

Les romans et les recueils qui suivent ont les mêmes caractéristiques, les valeurs négatives des écarts réduits témoignent de la reprise de thèmes antérieurs. Toutefois, nous constatons deux exceptions, *Vers les icebergs* et *Trois villes saintes*,

²²⁶ S. Domange (1993) : p. 123.

²²⁷ P. Boncenne (1978) : p. 44.

²²⁸ M. Labbé (1999) : p. 15.

relevant d'univers différents d'où l'afflux d'un vocabulaire nouveau. Le premier est une sorte de poème et le deuxième, un essai littéraire qui continue le thème des peuples amérindiens avec la quête et le mythe des villes sacrées. De même, avec *Voyage à Rodrigues*, roman écrit sous forme de journal personnel, un léger apport lexical est à noter. Le Clézio nous amène vers un nouvel univers, celui du navigateur, du chercheur de trésor sur l'Océan Indien.

A partir de ce moment, Le Clézio semble revenir, dans ses derniers ouvrages, comme d'autres romanciers, à une conception plus conventionnelle du roman, avec la volonté de retrouver une certaine illusion réaliste. Dans cette troisième partie de l'œuvre, on voit nettement s'opposer les genres littéraires.

Les essais *Le rêve mexicain* et *La fête chantée* traitent des cultures et mythologies amérindiennes. Comme nous l'avons noté auparavant, les apports lexicaux sont considérables dans ces œuvres. Le même phénomène se constate dans *Sirandanes*, qui traite du monde créole, ainsi que dans *Les gens des nuages*, le récit d'un voyage initiatique dans le désert marocain. *Diego et Frida* présente un apport lexical nouveau. Il est riche en termes espagnols, anglais et amérindiens et le livre énumère une multitude de personnages et de lieux différents.

Le goût de Le Clézio pour l'énumération influence évidemment ce type de statistique. Ce dernier livre en témoigne. Nous y trouvons de longues sections d'énumérations²²⁹ :

“Pour Diego Rivera, et pour la plupart des créateurs de sa génération, le “docteur” Atl, Rufino Tamayo, Carlos Mérida, Orozco, Jean Charlot, Vasconcelos, Pellicer, comme plus tard pour la génération de Justino Fernández et d'Octavio Paz, le monde préhispanique est hanté par cette déesse qui dort son sommeil magique sous la turbulence du monde mexicain moderne .”

²²⁹ *Diego et Frida*, p. 90.

Les livres de fiction se succèdent et le lexique s'accroît légèrement. De nouveaux thèmes apparaissent sans présenter d'apports lexicaux spectaculaires. Quelques fluctuations sont toutefois intéressantes à noter. Regardons cette catégorie de près.

2.5.2. L'accroissement lexical de l'œuvre romanesque.

Notre corpus D, le corpus "Roman", permet de faire le classement suivant :

Accroissement du vocabulaire								
ACCROISS.	CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
Procès		10018	10018	10018	91243	91243	-15530	-1.70
Fièvre		5527	10110	15545	100374	191617	21646	2.16
Déluge		4749	12078	20294	122453	314070	33848	2.76
Fuite		3115	9926	23409	102769	416839	26862	2.61
Guerre		2508	9563	25917	109196	526035	11259	1.03
Voyage		2181	8450	28098	111852	637887	4794	0.43
Mondo		955	6614	29053	97332	735219	-42725	-4.39
Désert		1439	8337	30492	149595	884814	-63213	-4.23
Ronde		737	6433	31229	82670	967484	-36507	-4.42
Chercheur		2082	9229	33311	133326	1100810	4173	0.31
Rodrigues		690	4885	34001	37999	1138809	9896	2.60
Printemps		871	6135	34872	69622	1208431	-7497	-1.08
Onitsha		1201	7189	36073	75271	1283702	13453	1.79
Étoile		1028	7743	37101	117674	1401376	-38894	-3.31
Quarantaine		2078	11182	39179	163938	1565314	3350	0.20
Poisson		1199	7577	40378	85083	1650397	16373	1.92
Hasard		939	7307	41317	66699	1717096	15292	2.29
Fonction $y=a(x \text{ exposant } b)$: $a=1.37302202765734e-004$ $b=2.18494658975019$ $r^2=0.987975980987316$ $r=0.993969808891254$								

Tableau n°21 : Accroissement lexical du corpus D.

Le graphique ci-dessous rend compte de la tendance générale de l'accroissement lexical dans les romans et les recueils de nouvelles, les autres genres littéraires écartés :

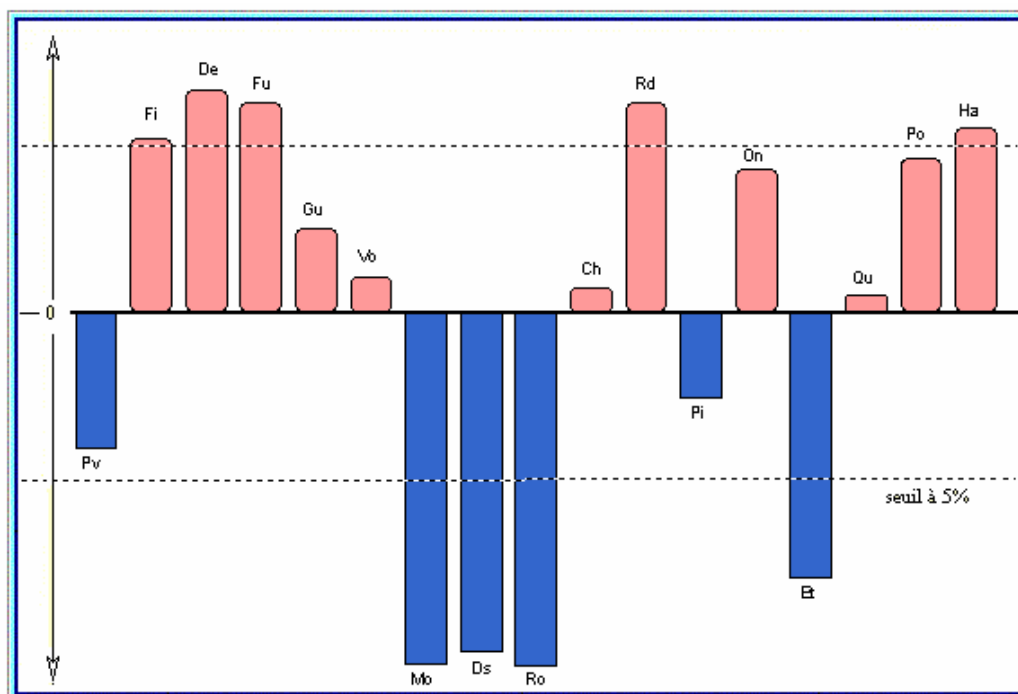


Figure n°34 : Accroissement lexical du corpus D.

Après un apport important de vocabulaire au début de l'œuvre, dans les ouvrages "nouveau roman", nous voyons l'apport lexical décliner à partir de *La guerre* pour être pratiquement nul (écart réduit de 1,45) dans *Voyages de l'autre côté*. Les valeurs négatives arrivent avec *Mondo* et les deux œuvres qui suivent. C'est à partir de *Le chercheur d'or* que nous pouvons constater à nouveau des apports de vocabulaire. Cette tendance continue jusqu'à la fin de notre corpus. *Voyage à Rodrigues* apporte du nouveau vocabulaire avec la découverte d'un nouvel univers, tandis que *Printemps et autres saisons* constitue un retour au monde occidental, aussi bien par les thèmes traités que par l'action et le décor. Il convient également de noter la chute importante d'apport lexical entre *Onitsha* et *Etoile errante*. *Onitsha*, c'est la découverte de l'Afrique noire et de ses mythes. Ce thème apparaît en effet pour la première fois ici. *Etoile errante*, par contre, semble appartenir à la période précédente.

Cette analyse permet de cibler la tripartition dans l'œuvre qui, de façon plus distincte, rejoint notre conclusion tirée de l'étude de l'accroissement lexical sur le corpus entier.

L'étude de l'accroissement fait très clairement apparaître, comme dans l'étude de la richesse lexicale et des hapax, les trois périodes principales dans l'œuvre de Le Clézio. Premièrement, le maintien d'un apport lexical élevé, de 1963 jusqu'au milieu des années 70. Puis une chute importante d'apport lexical qui dure jusqu'en 1986, et ne présente pas de vocabulaire supplémentaire. Enfin, la troisième période, qui 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.

Notons, au passage, que seuls *Les Prophéties de Chilam Balam*, *Le rêve mexicain* et *La fête chantée* s'opposent au classement de la richesse lexicale. Ils contribuent à un apport important de mots nouveaux, certes, mais sans présenter une richesse lexicale spectaculaire.

En général, on observe chez les écrivains²³⁰ un accroissement important du vocabulaire dans le début de la carrière littéraire qui diminue vers la fin de l'œuvre. Ce décroissement marque la tendance vers une écriture plus sobre à la fin d'une carrière littéraire. C'est, par exemple, le cas de l'œuvre de Zola où le renouvellement du vocabulaire s'amenuise vers la fin.

En revanche, il est intéressant de remarquer, dans l'œuvre de Le Clézio, la présence d'un vocabulaire croissant vers la fin²³¹, une tendance qui est assez exceptionnelle. Etant donné que l'œuvre de notre auteur est en progression, ce phénomène reflèterait-il plutôt un retour vers un niveau "normal" après une période d'apport lexical extrêmement réduit ?

2.5.3. L'accroissement du vocabulaire à partir des lemmes.

²³⁰ É. Brunet a montré ce phénomène dans ses ouvrages sur les écrivains du XIX^{ème} siècle.

²³¹ Dans notre cas, il faut se souvenir qu'il ne s'agit pas d'une fin de carrière, puisque l'auteur est toujours vivant et très productif.

Dans un deuxième temps, le même calcul de l'accroissement lexical de notre corpus est fait à partir des lemmes. C'est le corpus lemmatisé de douze œuvres qui est exploité dans cette analyse. Les tableaux ci-dessous rendent compte des résultats.

ACCROISS. CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
Procès	7315	7315	7315	91020	91020	-6174	-0.68
Fièvre	3048	7051	10363	100028	191048	26297	2.63
Mydriase	226	1535	10589	9838	200886	2431	2.47
Voyages	1894	5845	12483	113141	314027	7169	0.63
Mondo	724	4275	13207	97255	411282	-42584	-4.38
Désert	1203	5563	14410	149486	560768	-47349	-3.17
Ronde	589	4433	14999	82667	643435	-27320	-3.30
Rodrigues	895	3850	15894	38029	681464	53050	13.95
Printemps	779	4423	16673	69613	751077	16723	2.40
Pawana	104	1454	16777	10562	761639	1473	1.39
Quarantaine	1883	7676	18660	163923	925562	75415	4.60
Hasard	793	5305	19453	67652	993214	45695	6.75

Fonction $y=a(x \text{ exposant } b)$: $a=6.49706103392554e-006$ $b=2.61784501473123$
 $r^2=0.985692651226386$ $r=0.992820553386354$

Tableau n°22 : Accroissement lexical calculé sur les lemmes selon Cordial.

ACCROISS. CHRONO	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
Procès	6724	6724	6724	93184	93184	-3630	-0.39
Fièvre	2642	6451	9366	102942	196126	21976	2.13
Mydriase	194	1447	9560	10050	206176	1856	1.85
Voyages	1721	5367	11281	115749	321925	8048	0.70
Mondo	618	3918	11899	99599	421524	-46754	-4.69
Désert	1111	5177	13010	153639	575163	-46754	-3.04
Ronde	508	4114	13518	84526	659689	-30367	-3.59
Rodrigues	813	3606	14331	39334	699023	54517	13.86
Printemps	739	4170	15070	70992	770015	22224	3.13
Pawana	101	1416	15171	10924	780939	2415	2.21
Quarantaine	1806	7255	16977	168180	949119	95569	5.68
Hasard	734	5080	17711	69070	1018189	52261	7.57

Fonction $y=a(x \text{ exposant } b)$: $a=7.33097927304565e-006$ $b=2.6352931967328$
 $r^2=0.984892265151098$ $r=0.992417384546995$

Tableau n°23 : Accroissement lexical calculé par Hyperbase sur les lemmes selon la méthode Labbé.

La représentation graphique permet de mieux comparer les valeurs des deux versions lemmatisées.

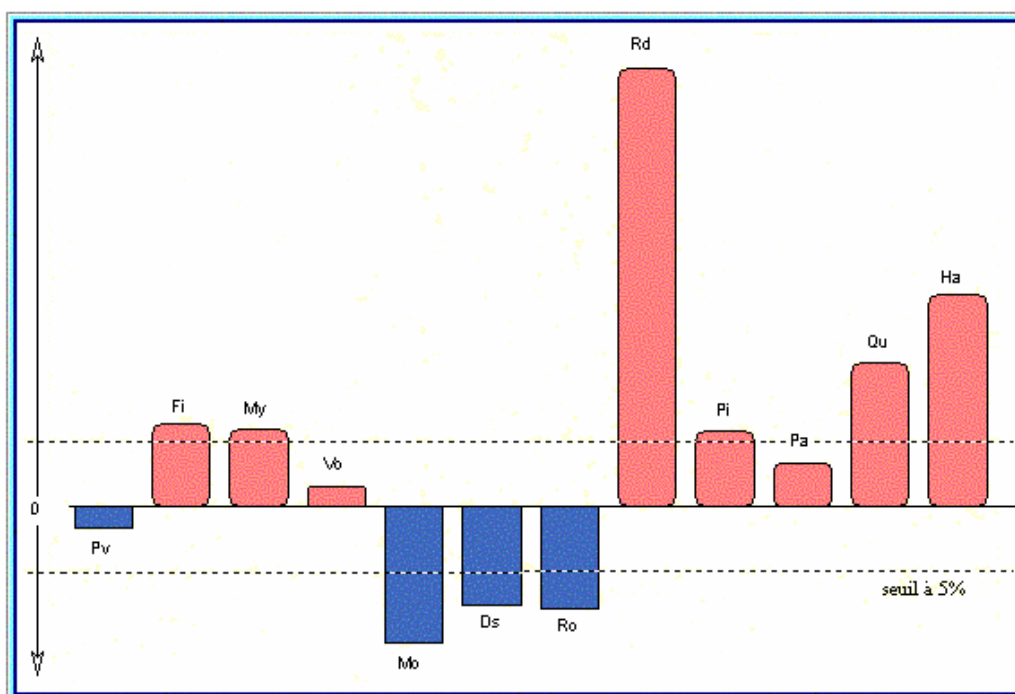


Figure n°35 : Accroissement lexical calculé par Hyperbase sur les lemmes selon la méthode Cordial.

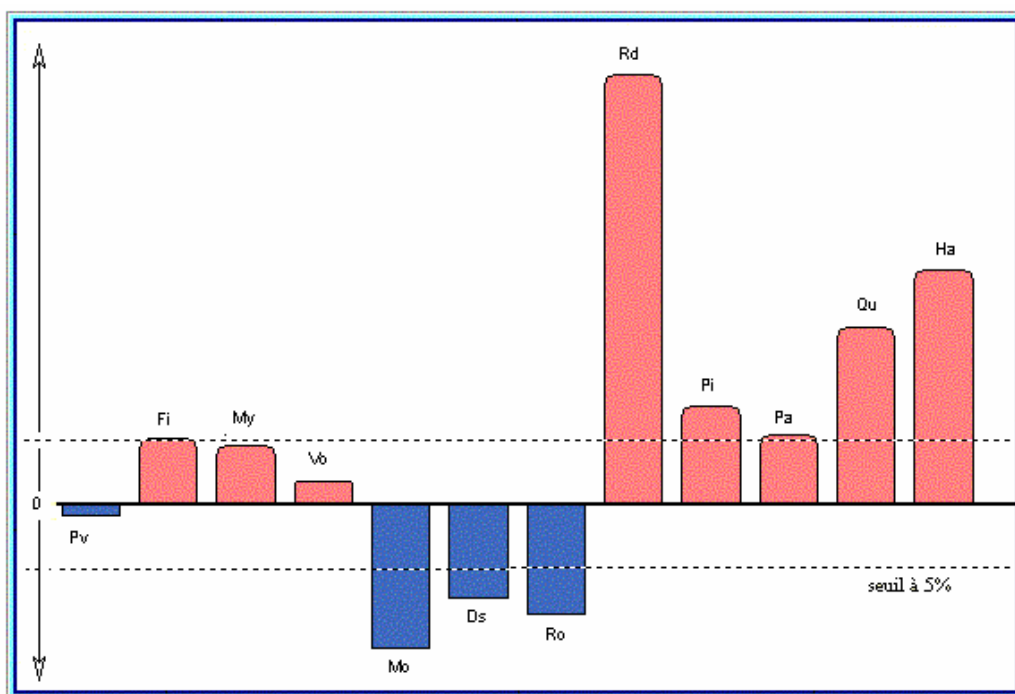


Figure n°36 : L'accroissement lexical calculé par Hyperbase sur les lemmes selon la méthode Labbé.

Les deux graphiques sont très semblables et les différences dans les valeurs des écarts réduits sont minimales. Dans la première moitié du corpus, les écarts sont légèrement plus grands lorsque les calculs sont basés sur le corpus lemmatisé par

Cordial ; en revanche, d'après la méthode Labbé, c'est à partir de *Mondo* que les écarts sont légèrement plus grands.

Afin de comparer ces résultats avec ceux de l'accroissement lexical à partir des formes graphiques du même corpus, nous insérons l'histogramme suivant :

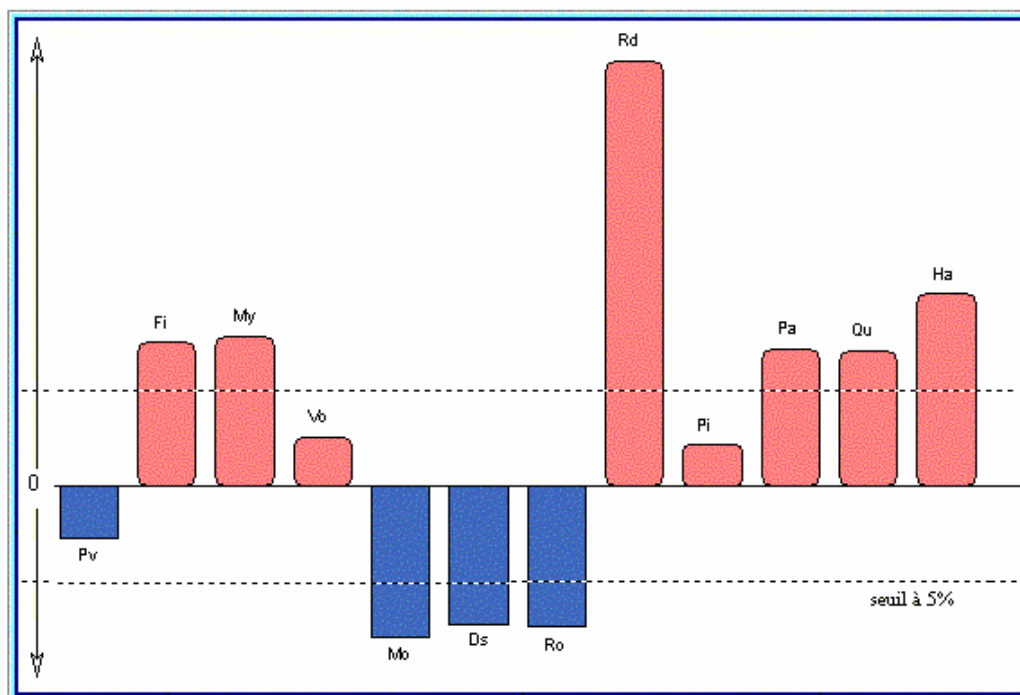


Figure n°37 : *Accroissement lexical calculé sur les formes.*

La comparaison des deux graphiques réalisée auparavant avec celui qui s'appuie sur les formes graphiques permet de constater que les trois tableaux sont en effet presque superposables : le même calcul aboutit, pour ces trois versions (la première basée sur les formes et les deux autres sur les lemmes), à des résultats quasi identiques.

Le logiciel Hyperbase fournit également une autre technique qui permet encore d'étudier l'accroissement lexical : il s'agit du classement inverse.

2.5.4. Le classement inverse.

En règle générale, l'analyse de l'accroissement lexical se porte sur l'œuvre d'un écrivain et sur ses sous-ensembles dans l'ordre naturel, l'ordre chronologique.

Cependant, le logiciel Hyperbase permet d'effectuer cette étude en choisissant l'ordre inverse, en mesurant l'apport lexical à partir de la dernière partie. Selon Étienne Brunet²³² :

“Les deux visées sont en fait complémentaires, la première désigne les thèmes qui apparaissent pour la première fois, la seconde ceux qui n'apparaîtront pas. La première marque le jaillissement d'une inspiration, la seconde son épuisement.”

Dans cette analyse, nous visons, en remontant le temps à rebours, la mise en évidence de l'épuisement de thèmes. Notre corpus sur les 31 œuvres s'appuyant, cette fois-ci, sur les formes graphiques, permet d'avoir une vue d'ensemble de l'œuvre leclézienne.

De la même façon que dans notre analyse de l'accroissement chronologique de l'ensemble de l'œuvre, Hyperbase calcule l'accroissement par un simple ajustement de courbe, en choisissant pour jalons les césures des textes mêmes.

Le tableau à la page suivante (tableau n°24) rend compte de l'accroissement du vocabulaire dans l'ordre inverse²³³ :

²³² É. Brunet (1978) : p.81.

²³³ Pour le calcul, cf. chapitre 2.5.1. “L'accroissement du vocabulaire de Le Clézio à partir des formes graphiques”.

L'accroissement à l'ordre inverse							
ACCROISS. INVERSE	Acc	Vocab	VocCum	Occur	OccCum	Ecart	Pondéré
HASARD	7235	7235	7235	67625	67625	-18898	-2.79
NUAGES	2003	3846	9238	21538	89163	8552	3.97
FÊTE	5476	8966	14714	70064	159227	48092	6.86
POISSON	3221	7578	17935	85039	244266	8776	1.03
QUARANTAINE	4329	11183	22264	163846	408112	-9641	-0.59
DIEGO	3124	8903	25388	72348	480460	58848	8.13
PAWANA	240	1903	25628	10558	491018	209	0.20
ETOILE	1675	7744	27303	117607	608625	-39733	-3.38
ONITSHA	1475	7190	28778	75201	683826	-2680	-0.36
SIRANDANES	76	761	28854	2465	686291	1371	5.56
PRINTEMPS	894	6135	29748	69563	755854	-23697	-3.41
RÊVE	2574	9273	32322	85723	841577	53888	6.29
RODRIGUES	750	4886	33072	37987	879564	4799	1.26
ANGOLI	227	3135	33299	24133	903697	-10995	-4.56
CHERCHEUR	1633	9230	34932	133199	1036896	-36130	-2.71
RONDO	776	6433	35708	82593	1119489	-34890	-4.22
VILLES	241	2462	35949	15294	1134783	-272	-0.18
DÉSERT	1072	8338	37021	149412	1284195	-81412	-5.45
ARBRES	42	661	37063	2946	1287141	-242	-0.82
ICEBERGS	50	781	37113	2707	1289848	515	1.90
INCONNU	1400	8499	38513	123952	1413800	-32019	-2.58
MONDO	669	6615	39182	97170	1510970	-52076	-5.36
PROPHÉTIES	176	1916	39358	8734	1519704	3253	3.73
VOYAGE	1424	8360	40782	113033	1632737	-14128	-1.25
MYDRIASE	113	1870	40895	9835	1642572	-1840	-1.87
GUERRE	1685	9566	42580	109093	1751665	12649	1.16
FUITE	1638	9982	44218	103175	1854840	19727	1.91
EXTASE	1454	8920	45672	90315	1945155	22540	2.50
DÉLUGE	1914	12079	47586	121853	2067008	32090	2.63
FIÈVRE	1236	10111	48822	99938	2166946	2721	0.27
PROCES	2187	10018	51009	90985	2257931	96900	10.65

Fonction $y=a(x \text{ exposant } b)$: $a=1.23981596647702e-003$ $b=1.96775022971229$
 $r^2=0.983252905581372$ $r=0.991591097974045$

Tableau n°24 : Accroissement lexical en ordre inverse du corpus A.

Le graphique suivant (figure n°38) illustre la dynamique de l'accroissement à l'inverse sous forme d'histogramme.

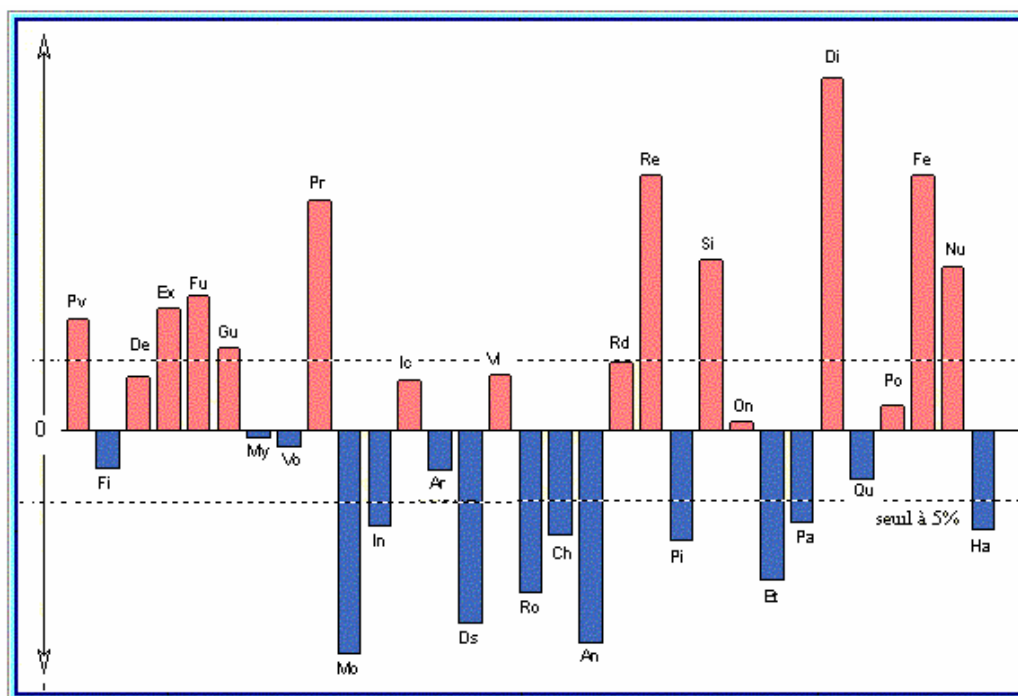


Figure n°38 : Accroissement lexical dans l'ordre inverse du corpus A.

Les données exposées ci-dessus nous permettent de confirmer ce que nous avons pu constater précédemment lors de l'analyse de l'accroissement chronologique. Nous retrouvons des valeurs positives considérables dans les ouvrages d'ethnologie, la biographie ainsi que dans les essais. Les valeurs négatives réapparaissent dans les romans de la deuxième période, avec des signes d'épuisement de thèmes très forts au niveau de *Mondo*, *Désert*, *La ronde* et *Angoli Mala*. Les romans de la première période "nouveau roman" présentent des apports lexicaux importants. Ce qui est exceptionnel, ce sont les valeurs du premier livre, *Le procès-verbal*.

Il faut se souvenir, en effet, lorsque l'on regarde des graphiques illustrant l'accroissement chronologique d'un corpus que, dans le cas du premier livre de la série, l'effectif coïncide nécessairement avec celui des vocables, puisqu'il s'agit d'un accroissement à partir de zéro. Par conséquent, il ne peut pas être réellement pris en compte. L'analyse semble défavoriser en effet non seulement le premier texte de la série - qui apparaît toujours plus pauvre que ne le serait un texte aléatoire exempt de spécialisation lexicale - mais aussi tous les textes qui

s'éloignent du centre de gravité, aux deux bouts de l'histogramme, comme dans notre cas : *La fièvre* et *Hasard*. Cette pénalisation du texte initial profite aux autres, et la comparaison s'en trouve faussée²³⁴.

Cette étude de l'accroissement à l'inverse permet néanmoins, dans le cas de notre corpus, de constater ce que l'étude de l'accroissement chronologique n'a pas pu relever, *Le procès-verbal* étant le premier sous-corpus de la série. Cette analyse fait apparaître une petite "explosion lexicale" au niveau de ce livre. Par ailleurs, elle confirme les résultats de l'étude de l'accroissement chronologique.

Il est à remarquer que cet histogramme est presque superposable à celui de la distribution des hapax. Nous trouvons le même phénomène d'un écart réduit très élevé au niveau du premier livre *Le procès-verbal*. Les hapax semblent en effet jouer un rôle important dans l'explosion lexicale que nous avons pu constater auparavant.

Regardons maintenant les résultats du calcul de l'accroissement lexical, non plus basés sur les livres différents mais sur des tranches égales.

2.5.5. L'accroissement lexical mesuré à partir des tranches égales.

Le logiciel "Lexicométrie" effectue le calcul de l'accroissement lexical d'une autre manière. Pour effectuer la mesure, cette fois-ci, on découpe le corpus en tranches de taille égale, c'est-à-dire en fragments comportant le même nombre de mots (1.000 mots) et non par les césures naturelles des différents livres.

Le modèle de partition permet de calculer le nombre attendu de vocables employés depuis le début du corpus, en différents points de celui-ci, dans l'hypothèse d'un accroissement régulier du vocabulaire, au fur et à mesure de l'augmentation de la taille, et en tenant compte de la spécialisation du

²³⁴ Cf. L'histogramme de l'accroissement lexical chronologique, figure n° 33, p. 174.

vocabulaire²³⁵. La technique est donc différente de celle d’Hyperbase qui, connaissant la taille d’une œuvre et sa place dans le corpus, calcule le nombre théorique de mots nouveaux qu’elle devrait apporter et compare avec le nombre effectivement constaté.

Lexicométrie observe en un point du corpus la position de la variable “accroissement du vocabulaire” depuis le début du corpus. C’est donc toute l’histoire de l’œuvre qui est prise en compte ainsi, les fluctuations marginales telles que mesurées par Hyperbase peuvent être retrouvées observant la pente de chaque segment correspondant à l’œuvre considérée.

La technique permet ainsi de localiser les ruptures thématiques dans le corpus là où se produit un afflux de vocables nouveaux (une pente ascendante du segment) ; les fragments où l’accroissement est inférieur aux valeurs théoriques signalent l’épuisement d’un thème (une pente descendante du segment) .

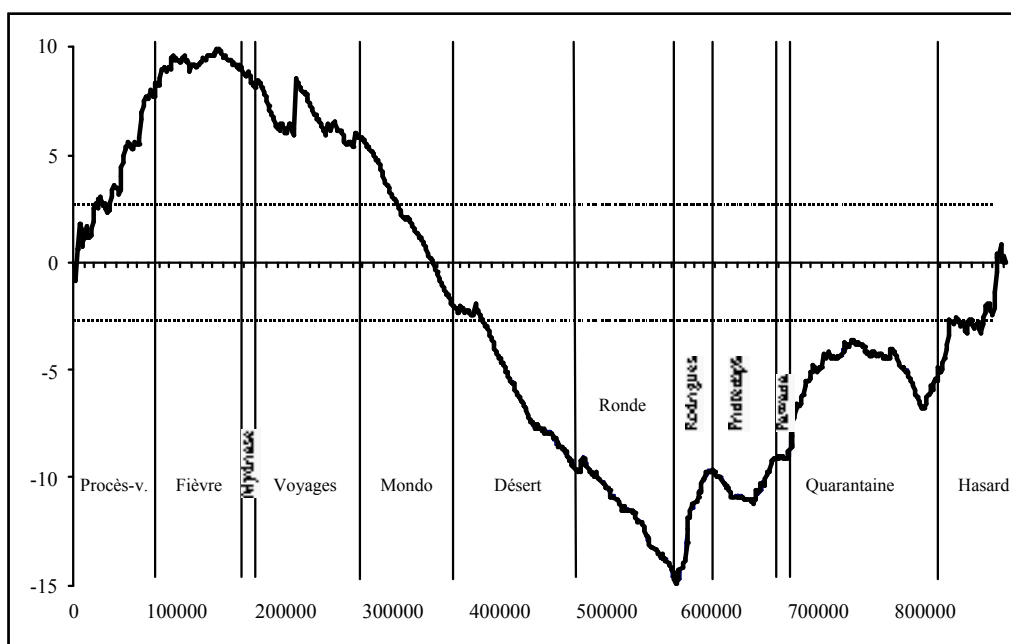


Figure n°39 : Accroissement lexical du corpus H calculé par le logiciel Lexicométrie sur les lemmes.

²³⁵ Hubert P., Labbé D. (1995) : p. 77-114.

Nous pouvons constater un mouvement semblable à celui observé dans les histogrammes précédents. La courbe qui s'inscrit autour de l'axe horizontal rend compte des très grands écarts autour de la tendance moyenne. Les fluctuations dépassent souvent les bornes de variations "normales" et se trouvent en dehors des limites de l'intervalle de fluctuation attendue (± 2 écarts types).

La courbe fait encore apparaître le même phénomène que dans l'analyse précédente, c'est-à-dire une évolution où se dégagent trois périodes principales dans la production leclézienne. Nous notons qu'après l'apport important au début de l'œuvre commence la chute de la courbe. De *Voyages à l'autre côté* à *La ronde* le déclin continue, excepté pendant deux périodes très brèves correspondant aux premières pages de *Désert*, c'est-à-dire à la découverte de l'univers des hommes bleus du désert nord-africain, et à la première nouvelle du recueil *La ronde*. De *Voyage à Rodrigues*, où nous observons la césure majeure, jusqu'à la fin, chaque œuvre représente un nouvel apport lexical. La courbe illustre, comme dans l'analyse précédente, que chaque livre correspond à un nouvel univers exploré par l'auteur. Une exception peut cependant être constatée, ce sont les trois premières nouvelles de *Printemps*, pour lesquelles le creux dans la courbe signale que du point de vue du vocabulaire elles appartiennent à la période précédente.

En effet, l'apport lexical dans l'œuvre, en dehors de quelques "accidents" constatés, étant pratiquement nul, nous amène à deux interprétations : premièrement, à partir du début avec *Le procès-verbal* et *La fièvre*, il ne semble pas s'ajouter de nouveaux thèmes dans l'œuvre – l'interprétation la plus probable –, deuxièmement, il pourrait s'agir d'un parti-pris de dépouillement qui amène l'auteur à traiter de nouveaux thèmes avec le même vocabulaire que ses anciens.

La même technique permet également d'étudier le profil de chaque œuvre de façon individuelle. Regardons de près quelques livres de notre corpus.

2.5.6. Le profil d'accroissement lexical des différentes œuvres du corpus.

Certains ouvrages du corpus ont, en effet, un profil qui sort de l'ordinaire²³⁶. On retrouve le profil très caractéristique de *La quarantaine* avec un fort apport, dans le premier quart du livre ; *Hasard* présente quant à lui une césure vers le milieu. Dans *Voyages de l'autre côté*, l'accroissement est assez spectaculaire. La courbe de la figure suivante met en évidence un mouvement tripartite :

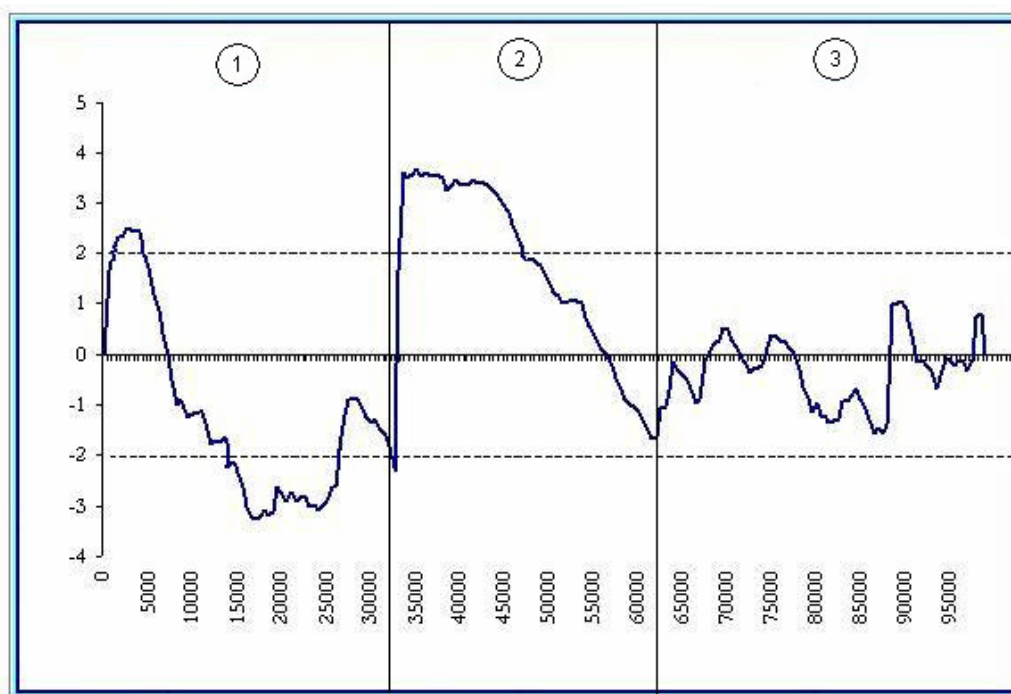


Figure n°40 : Profil de l'accroissement lexical dans *Voyages de l'autre côté*.

Nous remarquons un apport lexical initial suivi par une chute importante, puis un pic correspondant sensiblement au milieu du livre (vers 34.000 mots). En moins de 1000 mots, la courbe s'élève de +4,5 écarts types. Ce caractère abrupt est absolument unique. C'est un trait appartenant à l'écriture "nouveau roman" avec une longue énumération de noms propres ou d'objets.

²³⁶ En annexe n°4 : le profil de l'accroissement lexicale des individuel des œuvres du corpus.

Voyages de l'autre côté, considéré comme l'œuvre qui annonce le tournant dans l'écriture de Le Clézio, montre jusqu'à quel point ces calculs sont sensibles au genre. Les genres différents dans lesquels s'exprime l'écrivain entraînent en effet des fluctuations importantes et de très grandes variations.

La courbe de l'accroissement du corpus présente une chute importante au niveau du roman *Désert*. L'observation de son profil ne montre pourtant pas la même pente :

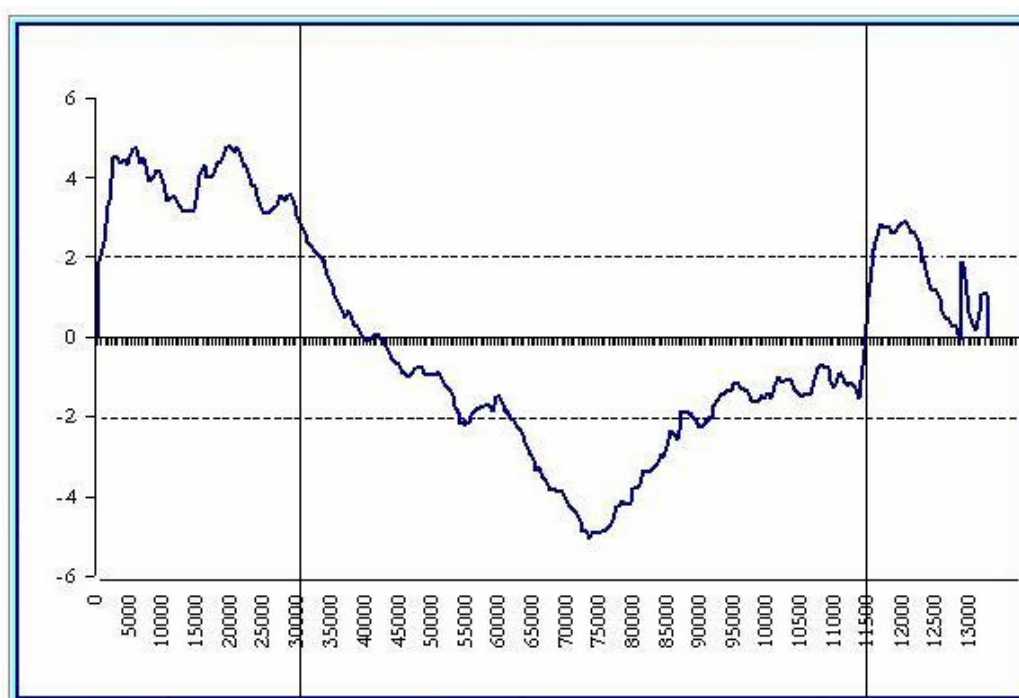


Figure n°41 : Accroissement lexical dans *Désert*.

La courbe permet ici aussi de constater une division de ce livre en trois périodes : la première période, historique, correspond à l'introduction d'un nouvel univers dans le corpus, celui du désert et des hommes bleus. La deuxième période, contemporaine, narre principalement le destin de Lalla et la troisième, par laquelle se termine l'œuvre, met l'accent sur la guerre coloniale du Maroc²³⁷.

²³⁷ Nous consacrerons le chapitre à cette technique de récits entrelacés : 5.5.2. "Les récits entrelacés".

La comparaison de cette courbe avec celle du corpus entier fait apparaître le maintien d'un apport lexical au début de l'œuvre qui précède la chute. Cette chute semble beaucoup moins importante dans la courbe consacrée à l'œuvre isolée que celle que nous avons constatée auparavant.

L'explication peut être fournie par le fait que la courbe du corpus entier enregistre tous les événements depuis le premier mot, de manière cumulative. A la hauteur de *Désert*, elle indique que ce roman n'apporte aucun renouvellement par rapport aux œuvres antérieures, sinon les premiers mots. En revanche, la même courbe, pour le roman *Désert* seul, déplace le point de départ du calcul au premier mot de ce récit et "oublie" toutes les œuvres antérieures. Elle indique qu'il y a un apport initial, une chute au milieu de l'œuvre et, à la fin, un nouvel apport de vocabulaire. Ces césures correspondent bien à la technique de récits intercalés²³⁸ qui est caractéristique de l'écriture leclézienne. Il semble que les récits où nous suivons le destin des "hommes bleus", guerriers du désert saharien, apportent plus de vocabulaire que les parties qui sont consacrées à Lalla, la jeune Maghrébine, et à son périple dans l'envers du monde occidental.

Nous notons que la réconciliation des deux graphiques se fait assez aisément : il suffit d'imaginer que l'axe horizontal du second bascule et suit une pente orientée vers le bas. La portion de courbe du graphique sur l'œuvre entière correspond alors à la courbe du roman *Désert* et nous pouvons constater que le mouvement général est relativement semblable.

De la même façon, le dernier ouvrage du corpus, le roman *Hasard*, présente un apport de vocabulaire assez spectaculaire, surtout vers la fin de la courbe qui couvre le corpus entier. Cet apport, significatif lorsque l'on regarde le corpus dans son ensemble, semble à l'étude individuelle beaucoup plus modéré :

²³⁸ Cf. chapitre 5.5.2. "Les récits entrelacés".



Figure n°42 : *Accroissement lexical du roman Hasard.*

L'apport de vocabulaire dans la courbe qui couvre l'ensemble du corpus se retrouve au début du profil individuel. Il s'agit, encore une fois, de thèmes nouveaux : la traversée de l'Atlantique sur le yacht *Hazard* représente en effet un nouvel univers. Par la suite, le livre ne présente pas d'apport significatif de vocabulaire. Au contraire, la courbe dépasse le seuil de deux écarts-types dans le sens négatif au milieu de l'œuvre, pour ensuite monter légèrement vers la fin. Ce mouvement correspond bien à la dynamique du récit.

Après la traversée maritime, les protagonistes vont retrouver la terre qu'il ont quittée quelques mois auparavant, les hommes et les soucis²³⁹ :

“Nassima s'est rhabillée, et lui n'a pas fait un geste, n'a pas dit un mot pour la retenir, d'ailleurs il sait que ç'aurait été inutile. Elle part avec une fausse hâte, comme si elle allait revenir plus tard, le retrouver sur

²³⁹ *Hasard*, p. 172-173.

le quai d'honneur, et lui serait en train d'étendre une nouvelle couche de vernis sur un épars, en train de faire une épissure ou Dieu sait quoi. Mais il a compris que tout était terminé, qu'elle ne reviendrait plus."

Comme avec *Désert*, nous pouvons réconcilier les deux graphiques, en basculant l'axe horizontal de *Hasard*, mais cette fois-ci vers le haut. Nous retrouvons la portion de courbe du graphique sur l'œuvre entière correspondant au roman *Hasard* et nous pouvons constater qu'elle est relativement linéaire. L'examen attentif de la première courbe montre aussi les incidents isolés qui dépassent l'intervalle normal de fluctuation.

De la même façon, l'étude de chaque œuvre de notre corpus permet un regard "à la loupe" sur les livres individuels qui, ensemble, de façon cumulative, contribuent à l'apport de vocabulaire au fur et à mesure que l'œuvre avance dans le temps.

2.7. Conclusion générale de la partie structure lexicale.

Ces différentes analyses nous amènent à des conclusions non seulement sur la structure lexicale de l'œuvre de Le Clézio, en tenant compte des différentes notions exploitées dans notre corpus, mais aussi sur la méthode de travail.

En effet, chacun des types de corpus permet une analyse différente et complémentaire. De la même manière que les méthodes de travail se complètent, l'exploitation des deux logiciels permet un travail complet à partir d'un corpus donné.

L'application des différentes méthodes basées sur les formes graphiques ou sur les lemmes permet de constater que les résultats sont souvent assez semblables : la superposition des différents graphiques est, dans la plupart des analyses, presque parfaite. Dans les différentes études, qu'il s'agisse de la richesse, de la diversité ou de l'accroissement du lexique, le parallélisme est respecté, comme si l'étiquetage était neutre et sans influence.

L'étude sur la structure lexicale de l'œuvre de J.M.G. Le Clézio a permis de constater une tendance générale que confirment les différentes analyses. En premier lieu, nous avons pu noter, à travers ces études, 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. Cette forte opposition générique nous a motivée à cibler notre recherche et à créer un corpus uniquement basé sur les œuvres de fiction.

Concernant l'œuvre romanesque de l'auteur, nous avons constaté 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.

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 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, qui divise l'œuvre en deux périodes principales.

La première période, caractérisée par un vocabulaire riche avec des apports lexicaux importants, témoigne non seulement d'une époque de la littérature française influencée par le style "nouveau roman", mais elle reflète surtout l'intention manifeste du jeune Le Clézio de tout exprimer et de laisser jaillir la spontanéité créatrice librement. Selon Jean Onimus²⁴⁰ :

“Une ivresse verbale s’empare de celui qui accumule avec rage et volupté, comme le peintre Arman accumule les cuillères et les réveille-matin.”

Nos diverses études sont nettement influencées par cette “accumulation lexicale” caractéristique de l'écriture leclézienne des années 1960 à 1970.

Les différentes analyses mettent ensuite en exergue la rupture dans l'œuvre, au niveau de *Mondo*, amorcée dès *Voyages de l'autre côté*, et le renouvellement complet qui suit. L'écriture est désormais caractérisée par un vocabulaire très réduit et une expression concentrée, souvent incantatoire, qui tend vers le dépouillement. Michelle Labbé²⁴¹ décrit ainsi ce langage, pauvre en paroles, qui est devenu très caractéristique de Le Clézio :

²⁴⁰ J. Onimus (1994) : p. 167.

²⁴¹ M. Labbé (1999) : p.71.

“Ce qui est recherché, c’est un rythme, c’est-à-dire un langage affectif, universel, primordial qui rejoint le fameux langage des oiseaux, l’absence de toute parole. A la pointe extrême de l’écriture surgit le rythme et à la pointe extrême du rythme se trouve le silence.”

Vers la fin de l’œuvre, nous avons constaté un vocabulaire croissant reflétant une écriture plus conventionnelle, en relevant toutefois quelques irrégularités. *Etoile errante*, nous l’avons vu, ne s’inscrivait pas dans la chronologie : il était, avec son vocabulaire “pauvre”, plus proche des romans des années 1980, *Mondo*, *Désert* et *La ronde*, que des livres de l’époque de sa parution. *Le Poisson d’or*, en revanche, se distinguait par son vocabulaire “riche”. C’est l’écrivain lui-même qui nous livre l’explication sur ces soupçons éveillés par l’analyse statistique²⁴² : le roman *Etoile errante* comme nous l’avons déjà indiqué dans l’introduction²⁴³ a bien été écrit au milieu des années 1980, mais comme le moment coïncidait avec les mouvements de l’Intifada et le roman, en partie, se déroule en Israël et en Palestine Le Clézio, ne voulant pas donner au livre une connotation politique, a préféré attendre avant de le publier. Quant à *Poisson d’or*, il a été pensé et écrit pour être un scénario de film plus qu’un roman.

Ces explications nous permettent d’insérer *Etoile errante* à “sa place” dans la chronologie et de considérer *Poisson d’or*, au moins en partie, comme un autre genre littéraire ; elles nous livrent de plus une image cohérente du mouvement général du vocabulaire. 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.

²⁴² A notre demande, J.M.G. Le Clézio a gentiment fourni ces précisions dans un entretien.

²⁴³ Cf. p. 38.