

Préambule.

“Écrire seulement sur les choses qu’on aime. Écrire pour lier ensemble, pour rassembler les morceaux de la beauté, et ensuite recomposer, reconstruire cette beauté. Alors les arbres qui sont dans les mots, les rochers, l’eau, les étincelles de lumière qui sont dans les mots, ils s’allument, ils brillent à nouveau, ils sont purs, ils s’élancent, ils dansent ! On part du feu, et on arrive dans le feu.”

L’inconnu sur la terre, page 10.

Comment cerner l’écriture de J.M.G. Le Clézio ? La place qu’occupe Le Clézio dans la littérature française, par l’importance de la diffusion de son œuvre, la forme et la thématique de ses récits au statut littéraire variable, le désignent aujourd’hui comme un des écrivains les plus connus et les plus aimés du grand public. La singularité de cette œuvre, souvent perçue comme originale et marginale, la rend difficile à classer. La production leclézienne, à travers les textes très divers qu’il a publiés, a évolué de façon tout à fait personnelle et indépendante : depuis ses débuts aux années 1960 elle a traversé romans, nouvelles, récits poétiques, traductions et adaptations, essais, etc., toujours en marge des courants littéraires.

L’œuvre de Le Clézio a déjà fait l’objet de diverses études universitaires - françaises aussi bien qu’étrangères -, et nombreux sont les signes d’une institutionnalisation en cours de la création littéraire de notre auteur. A en juger par ces travaux, l’intérêt de l’œuvre résidait jusqu’alors principalement dans l’étude de certains textes, de leurs aspects stylistiques, mais peut-être surtout des thématiques représentatives de l’œuvre ou d’une partie de l’œuvre ; la critique de la société moderne, l’exotisme géographique pour certains textes, le monde amérindien, la dimension mythique, l’aspect érotique, etc. en sont les exemples les plus prégnants. Ces analyses se sont inspirées de différentes sources, notamment

de la critique littéraire “classique”, bachelardienne, ou encore de l’école psychocritique.

Peut-être les contenus de ces analyses ne rendent-ils pas encore suffisamment compte de la spécificité de l’œuvre leclézienne, dont l’unité est au demeurant difficile à cerner ? C’est pourquoi aborder l’œuvre à partir de son vocabulaire, des mots eux-mêmes, dans une démarche endogène et non à partir d’une thématique artificiellement préétablie, peut se révéler fructueux. Pour ce faire, nous nous appuierons sur une technique bien précise, dans l’ambition de rapprocher les études littéraires des études linguistiques.

“Ce rapprochement paraît aujourd’hui assez naturel, écrit Roland Barthes¹. N’est-il pas naturel que la science du langage (et des langages) s’intéresse à ce qui est incontestablement langage, à savoir le texte littéraire ? N’est-il pas naturel que la littérature, technique de certaines formes de langage, se tourne vers la théorie du langage ? N’est-il pas naturel qu’au moment où le langage devient une préoccupation majeure des sciences humaines, de la réflexion philosophique et de l’expérience créative, la linguistique éclaire l’ethnologie, la psychanalyse, la sociologie des cultures ? Comment la littérature pourrait-elle rester à l’écart de ce rayonnement dont la linguistique est le centre ? N’aurait-elle pas dû, même, être la première à s’ouvrir à la linguistique ?”

Le Clézio lui-même se joint à cette idée. “Il n’y a, écrit-il², de réalité que celle que nous concevons par le service de la langue.” Notre auteur s’est en fait intéressé très tôt au rapprochement entre le langage et la littérature, et développe dans

¹ R. Barthes (1968) : p. 3.

² *L’extase matérielle*, p. 140.

différents ouvrages des théories dans ce sens³. C'est dans *L'extase matérielle* que nous trouvons la plupart de ses réflexions portant sur le langage⁴ :

“Une des erreurs de l'analyse est de distinguer forme et fond. Il est bien évident que forme et fond ne sont qu'une seule et même chose, et qu'il est tout à fait impossible de les dissocier. Parler de telle ou telle façon, employer tel ou tel mot sont des modalités qui engagent tout l'être. Le langage n'est pas une “expression” ni même un choix ; c'est être soi.”

De quelle façon aborder l'œuvre très abondante de Le Clézio, dont l'examen systématique et exhaustif manuel relève de l'impossible ? Nous avons choisi de recourir à l'outil informatique et aux méthodes lexicométriques, et plus largement, à celles de la statistique linguistique. Dans cette perspective, qui constitue l'unité de la présente entreprise, notre objectif est d'obtenir une vision d'ensemble de l'œuvre de Le Clézio et de l'évolution même de cette œuvre - évolution qui a d'ailleurs provoqué des variations importantes dans l'intérêt que lui a porté la critique.

Nous tenterons de suivre le parcours lexical et sémantique de J.M.G. Le Clézio, cet auteur que l'on qualifiait à ses débuts d'“écrivain de la ville” occidentale, “du monde moderne”, “de la civilisation mécanique”, “de l'électricité”, “du béton”, “de l'agression urbaine”, de la marginalité, et même d'une certaine forme de guerre. De cet homme qui, insensiblement, a évolué pour devenir l'écrivain du silence, du désert, de la transparence et de l'harmonie. C'est à travers l'étude du lexique, par des méthodes portant toutes les caractéristiques de l'univers automatisé et mécanique pourtant dénoncé dans l'œuvre leclézienne, que nous parviendrons à une étude plus objective de ses caractéristiques et de son évolution.

³ Nous émaillerons d'ailleurs notre texte de citations de Le Clézio qui n'auront pour autant pas la prétention d'illustrer parfaitement les chapitres qu'elles introduisent.

⁴ *L'extase matérielle*, p. 51.

La lexicométrie et le traitement informatique de textes permettent une description contrôlée et chiffrée du vocabulaire et de la syntaxe, ouvrant ainsi la voie à un traitement exact et impartial du corpus. Lorsque celui-ci est important, comme l'est celui de l'œuvre de Le Clézio avec plus de trente livres, l'aide de l'ordinateur devient indispensable pour saisir les nuances, les structures et les évolutions de la langue et de son usage. L'ordinateur et les logiciels appropriés permettent une description rigoureuse et objective du corpus, sur laquelle une réflexion et une analyse précises peuvent alors être faites.

En effet, bien que l'approche du texte soit avant tout quantitative, cette méthode d'analyse de texte exige un retour au texte et au contexte. Elle ouvre ainsi la voie à l'analyse qualitative, à une réflexion sur l'utilisation des mots, à un travail de recherche sur le vocabulaire et le langage⁵.

L'œuvre de Le Clézio n'a guère besoin de présentation ; la plupart de ses livres sont en effet connus de tous et l'écrivain jouit d'un succès indéniable. Nous ferons toutefois un très bref rappel du détail de son œuvre en relation avec notre corpus, avant de commencer cette étude qui met en œuvre plusieurs méthodes de travail.

En effet, notre thèse a un double but : elle ne s'intéresse pas seulement à l'œuvre leclézienne mais a également une mission méthodologique claire, consistant en l'exploitation et la comparaison, tout au long de ce travail, de plusieurs principes méthodologiques et par là même de différents logiciels lexicométriques. Cette ambition didactique nous a contrainte à limiter les longs développements théoriques, déjà bien documentés par ailleurs et qui ici auraient nuit à une appréhension globale du travail, au profit de l'évaluation des résultats de l'exploitation de techniques différentes, qu'il conviendra de définir au préalable. Ce sont ces quelques mots sur Le Clézio d'une part et sur les méthodes d'autre part qui constitueront notre première partie.

⁵ Il est évidemment impossible, faute de place, de commenter en détail, dans chaque analyse, chacun des 31 ouvrages qui constituent notre corpus.

Nous nous intéresserons ensuite à la structure du vocabulaire et à son évolution à travers l'œuvre de Le Clézio. C'est à partir de l'étude de la distribution des fréquences mais aussi de la richesse lexicale, de la diversité du vocabulaire, de l'accroissement lexical ainsi que des hapax que nous tirerons des conclusions sur la structure du vocabulaire leclézien.

Le chapitre suivant se penchera sur la phrase et le rythme du récit. L'analyse de la manière dont sont agencées les phrases, ainsi que l'examen de la distribution quantitative des mots à l'intérieur des segments, de la longueur des mots, tous deux rendus possibles par l'application de méthodes lexicométriques et des logiciels adaptés, donnent des informations intéressantes et permettent de révéler des caractéristiques stylistiques qui compléteront les études existantes sur le style de notre écrivain.

Dans le quatrième chapitre nous nous interrogerons sur les parties du discours et la syntaxe dans l'écriture leclézienne au travers d'une analyse "grammatico-métrique". Après avoir étudié la distribution des catégories grammaticales, possible grâce aux versions lemmatisées du corpus, nous nous attarderons sur certains aspects morphologiques et syntaxiques qui nous ont paru intéressants et révélateurs, dans la catégorie nominale aussi bien que dans la catégorie verbale.

Ce n'est que dans la cinquième partie que nous nous intéresserons réellement au contenu lexical et thématique de l'œuvre de Le Clézio. Après avoir étudié la distance lexicale entre les différents livres du corpus nous nous proposerons d'étudier les spécificités lexicales, dans une perspective exogène - prenant le corpus de *Frantext* comme référence – aussi bien que dans une perspective endogène. Cette étude mettra également en exergue les thématiques, ou isotopies, récurrentes dans l'œuvre. Nous en explorerons certaines plus en détail, notamment *la nature, la mer, la terre, le bestiaire* ainsi que *le regard* omniprésent dans l'œuvre leclézienne. Enfin, la dernière notion de cette cinquième partie à laquelle nous nous intéresserons est celle de l'intertextualité, très importante dans l'étude d'une œuvre comme celle de Le Clézio et qui mérite également une approche

technique ; nous nous attacherons notamment dans ce chapitre à une particularité de l'écriture leclézienne : les récits intercalés.

L'écriture leclézienne parcourt le monde entier et s'inspire de sources différentes. Dans la dernière partie nous nous intéresserons à quelques-unes de ses sources. Les méthodes lexicométriques s'adaptent en effet très bien aussi à des études comparatives. En exploitant ces techniques nous tâcherons de comparer l'œuvre de Le Clézio avec celle d'un de ses contemporains, Julien Gracq, susceptible d'offrir quelques affinités thématiques avec notre corpus. Par ailleurs, J.M.G. Le Clézio parle souvent des auteurs et des livres qui ont inspiré en partie son écriture, parmi lesquels on compte des romans d'aventures tels *L'île au trésor* et *Robinson Crusoé* et l'œuvre de Joseph Conrad. S'agissant de traductions, ces derniers seront étudiés uniquement dans leur aspect thématique.

Nous voulons ainsi, par l'exploitation de plusieurs méthodes de travail et par le choix d'une approche technique, dans un souci d'interdisciplinarité également, multiplier les points de vue sur la littérature leclézienne par la caractérisation systématique de son vocabulaire et de sa syntaxe, dans l'évolution d'une œuvre toujours en progression.

“Tout et rien. Je prenais des feuilles de papier, les plus grandes possible, et je les couvrais d'écriture, presque sans y prendre garde, presque au hasard. Mais ça n'avait aucun genre littéraire, c'était simplement de l'écriture.”

La fièvre, p. 143.

Première partie :

Introduction.

1.1. L'œuvre de J.M.G. Le Clézio.

“Il y a tant de choses à dire, de choses
belles, et stupides, et longues !
Je voudrais écrire pendant mille ans.”

Le livre des fuites, page 172.

“J’ai une conception sans doute morale de la littérature, car je crois, en effet, que la littérature est une fiction en vue d’autre chose. N’ayant pas acquis cette autre chose, n’étant sans doute pas assez mûr, j’attendrai la vieillesse sans être passé par la maturité puisque je vais continuer d’écrire en continuant de me demander quand vais-je enfin atteindre ce moment où je n’aurai plus besoin de mots.”

C’est par cette citation que commence la biographie de Gérard de Cortanze sur J.M.G. Le Clézio. L’auteur, qui écrit et publie depuis bientôt quarante ans, connaît désormais un grand succès et son lectorat voit paraître chaque année un ou parfois plusieurs livres, touchant à des domaines très divers. Sa large production, qui compte aujourd’hui plus d’une trentaine de livres, des articles et des traductions, s’accroît en effet chaque année, l’auteur étant toujours très actif et productif.

Le monde de Le Clézio est celui de la planète entière. Le planisphère ci-dessous témoigne de l’étendue géographique par son œuvre⁶ :

⁶ Reproduction de carte publiée dans Magazine littéraire, n° 362, février 1998 consacré à Le Clézio, p. 20.



Figure 1 : La planète de Le Clézio.

Le Clézio passe les vingt premières années de sa vie à Nice, “située en bordure de la mer, entre la plage et les collines et montagnes ; son soleil méditerranéen à la fois chaleur et lumière et sa population hétéroclite bronzée par le soleil”. Ces années lui ont fourni bien plus qu’un décor, écrit Germaine Brée⁷ : “un espace, une topographie où circulent, prennent forme et se déploient l’imagination et la pensée lecléziennes. Cet espace donne à ses premiers livres, si différents par ailleurs les uns des autres, une obsédante continuité.”

“Il y avait une petite fois...” débute l’œuvre et *Le procès-verbal*, son premier livre qui paraît en 1963. Dans ce roman erre Adam Pollo, jeune homme tourmenté qui refuse, pendant un été caniculaire, de s’intégrer dans le monde social plus ou moins accepté comme normal, jusqu’à ce que la police l’arrête et le mène à l’asile psychiatrique ; le roman se termine sur un “interrogatoire” d’Adam par sept étudiants-psychiatres, qui constitue sur plusieurs plans le vrai “procès-verbal”.

⁷ G. Brée (1993) : p. 15.

Lors de la parution de ce premier roman, pour lequel Le Clézio s'est vu attribuer le prix Théophraste Renaudot, la critique et la publicité font prévoir une carrière littéraire précoce et brillante à ce jeune auteur qui, à cette époque, n'a que 23 ans.

Le succès immédiat du roman, à structure expérimentale, valut tout de suite à Le Clézio d'être salué par certains critiques comme un nouveau "nouveau romancier".

Aujourd'hui, il est admis que l'école du "nouveau roman" ne fut qu'un groupement transitoire et hétérogène d'écrivains de tous âges. Les noms les plus cités sont : Nathalie Sarraute, Samuel Beckett, Claude Simon, Marguerite Duras, Robert Pinget, Alain Robbe-Grillet, Claude Ollier, Michel Butor et Jean Ricardou⁸. Ce mouvement était en effet relativement délimité, comme le souligne Hubert de Phalèse⁹ à propos de Claude Simon et *La route des Flandres*.

"Il serait illusoire de croire que durant une vingtaine d'années, la littérature française fut occupée uniquement de formes nouvelles, d'une transformation du modèle narratif. La plus grande forme, nul ne s'en étonnera, revient au roman de forme traditionnelle [...] Mais entre les époques réalistes de Troyat ou Jules Roy et la rêverie fantastique d'André Dhôtel ou de Julien Gracq, s'élève la voix sèche et courte de Françoise Sagan et celle plus ample, de Le Clézio..."

Pourtant, Le Clézio s'est démarqué assez vite de cette école, qui pour répondre aux attentes d'un public essentiellement universitaire aurait exigé de lui davantage de littérature théorisante. "J'ai voulu que ça fasse du bruit.", dit-il dans une interview télévisée, "et ça a fait du bruit, donc j'étais assez content. En même

⁸ Le noyau de cette école était, selon Jean Ricardou (1990 : p. 23), composé de Michel Butor, Claude Ollier, Robert Pinget, Jean Ricardou, André Robbe-Grillet, Nathalie Sarraute et Claude Simon. Il propose, toutefois que l'on étende la dénomination de "nouveau roman" "à tout ce qui en quelque manière contient dans le roman contemporain, et par rapport au roman académique, des traces de nouveautés". Parmi les titres divers dont il cite l'intérêt, se proposaient vers 1973, *Le Procès verbal*, *La Fièvre*, *Le Déluge* et *Terra amata* de Le Clézio.

⁹ Hubert de Phalèse (1997) : p. 16.

temps j'étais insatisfait, [...] j'ai eu une réaction de méfiance"¹⁰ Car malgré une certaine filiation avec le nouveau roman, au niveau du personnage et du montage romanesque, Le Clézio souhaitait proposer une nouvelle approche du genre romanesque, visant une critique à la fois du roman psychologique, de la narration classique et du nouveau roman.

“Pour lui, écrit Miriam Stendal Boulos¹¹, l'expérimentation formelle n'empêchait pas le lyrisme d'une contemplation amoureuse de la matière, la mise en valeur d'un cosmos vivant, l'expression de la soif de l'absolu dans la quête d'un ailleurs et dans une écriture exubérante.”

Cette écriture nouvelle et différente n'est pas seulement conçue par la critique comme appartenant à l'école du “nouveau roman” mais très vite, ce livre : *Le procès-verbal* et ceux qui suivront recevront l'étiquette ambiguë entre toutes d’“œuvre difficile”.

La fièvre, son deuxième livre, un recueil de nouvelles, paraît en 1965. Ce sont des histoires qui partent de l'expérience de la douleur, d'un peu de température, d'une rage de dents, d'un vertige passager.

“Ces neuf histoires de petite folie, écrit Le Clézio dans la préface¹², sont des fictions ; et pourtant elles n'ont pas été inventées. Leur matière est puisée dans une expérience familière [...] Tous les jours, nous perdons la tête à cause d'un peu de température, d'une rage de dents, d'un vertige passager. Nous nous mettons en colère. Nous jouissons. Nous sommes ivres. Cela ne dure pas longtemps, mais cela suffit. Nos peaux, nos yeux, nos oreilles, nos nez, nos langues emmagasinent tous les jours des millions de sensations dont pas une

¹⁰ Le Clézio dans *Un siècle d'écrivains*, émission diffusée sur FR3 le 8 mai 1996.

¹¹ M. Stendal Boulos (1999) : p. 15.

¹² La préface du roman.

est oubliée. Voilà le danger. Nous sommes de vrais volcans”.

Lorsque Le Clézio publie *Le déluge* en 1966, cet ouvrage est selon l’auteur la réalisation d’un rêve d’adolescent, le rêve de recréer le monde. Déjà dans la préface de *Le procès-verbal* Le Clézio avait annoncé ce nouveau départ : “un autre récit, beaucoup plus étendu, racontant avec le maximum de simplicité ce qui se passe le lendemain de la mort d’une jeune fille, description sommaire de *Le déluge*, continuation du premier volume.” Le roman qui déjà par son titre annonce le recours à un très vieux mythe judéo-chrétien, dénonce la confusion, l’angoisse et la peur de la grande ville occidentale, à travers 13 jours dans la vie de François Besson, qui, sous le choc déclenché par la mort d’une jeune fille, disparaîtra une fois la crise surmontée. C’est, pour la première fois, une histoire d’un événement qui comporte un début, des péripéties et une fin, tout comme le roman traditionnel ; mais la fin est aussi un commencement et un nouveau départ.

L’écriture de cette période, qui coïncide avec la découverte de l’Orient et du Mexique, est caractérisée par le refus et la critique de la société moderne occidentale : “Le monde gréco-romain, je ne suis plus son fils, je ne peux plus être de sa race, je ne sais plus rien de lui. [...] Sale monde latin, tu as voulu faire de moi un esclave, je ne suis plus ton fils, plus ton fils !” écrira Le Clézio un peu plus tard.¹³

Simultanément à ses premiers succès, Le Clézio présente un mémoire de maîtrise sur “Le thème de la solitude dans l’œuvre d’Henri Michaux” et le projet d’une thèse sur Lautréamont¹⁴.

L’année 1967 voit paraître deux livres, le roman *Terra amata* et l’essai *L’extase matérielle*. *Terra Amata* est une sorte de dialogue sur l’immortalité de l’âme, à travers

¹³ *Le livre des fuites*, p. 249.

¹⁴ Le Clézio ne finira pas sa thèse sur Lautréamont. Mais il ne reniera pas son goût pour les deux poètes. En témoignent *Vers les icebergs* (1978), essai sur le poème *Inji* de Michaux, et deux essais sur Lautréamont dans *Nouvelle Revue Française*, nrf, avril et mai, 1987.

l'histoire du personnage central, le petit garçon Chancelade. Ce récit d'une vie, qui se déroule chronologiquement de la naissance à la mort, est pourtant moins transparent qu'il ne semble. Le roman se transforme en une sorte de fable qui culmine par un drame métaphysique de la "mort de Dieu", que Chancelade remplace par une vision sommaire des agissements du grand capitalisme avant de terminer sur le cheminement vers la mort du personnage vieilli et solitaire.

L'extase matérielle, qui est présenté comme un "essai discursif", nous fait entrer directement dans le processus de pensée de ce jeune intellectuel poussé par le besoin de poser à nouveau, dans le cadre moderne, les problèmes fondamentaux de l'existence humaine. Le livre est "à l'opposé de tout système, composé de méditations écrites en marge de l'œuvre qu'on connaît, hors de toute tranquillité, destinées à remuer plutôt qu'à rassurer, oui, à faire bouger les idées reçues, les choses acquises ou apprises."¹⁵

"C'est un document tout personnel que cet essai, écrit Germaine Brée¹⁶ où le jeune intellectuel "privilegié" qu'est Le Clézio, tout imbu de littérature et de philosophie, éprouve le besoin de reposer dans un cadre "moderne" toutes les grandes questions concernant la vie."

A partir de 1967, tout en remplissant les fonctions d'enseignant à l'Université du Mexique et à l'Université du Nouveau Mexique à Albuquerque, Le Clézio voyagera dans le monde entier, parcours dont ses ouvrages gardent la marque.

Le livre des fuites paraît en 1969, un roman qui a pour thèmes les voyages, la déambulation et la fuite perpétuelle, à travers les aventures de Jeune Homme Hogan, un voyageur qui passe de pays en pays. Ce personnage, un et plusieurs à la fois, "se radiographie en radiographiant" l'univers des ses villes monstrueuses, ses autoroutes et ses déserts. Le mythe moderne pose à l'écrivain le problème de la

¹⁵ La quatrième de couverture.

¹⁶ G. Brée (1990) : p. 48.

conscience qui le pousse à des séquences d'autocritique que Jeune Homme Hogan décrit ainsi¹⁷ :

“Je veux fuir dans le temps, dans l'espace. Je veux fuir au fond de ma conscience, fuir dans la pensée, dans les mots. Je veux tracer ma route, puis la détruire, ainsi, sans repos. Je veux rompre ce que j'ai créé, pour créer d'autres choses, pour les rompre encore. C'est ce mouvement qui est le vrai mouvement de ma vie.”

Ensuite, en 1970, paraît *La guerre*, qui aborde l'expérience de la guerre, “Idéalement que voulez-vous exprimer ?” lui demande Pierre Lhoste à propos de ce roman¹⁸. A cette question Le Clézio répond :

“Peut-être essayer d'exprimer le constat qu'actuellement on vit dans une société en guerre permanente. Guerre de masse contre l'individu, guerre des objets contre l'être humain, guerre des êtres humains entre eux. C'est ce sentiment d'agression permanente qu'il y a autour de moi.”

C'est dans un milieu urbain que la jeune fille Bea B. circule clandestinement et le roman est composé en grande partie des pages du “semainier pratic”, journal où elle inscrit ses observations, ses sentiments, ses pensées ; et des lettres envoyées à un Monsieur X. Ces personnages qui ne se comportent pas logiquement, ne cherchent pas à donner un compte-rendu cohérent de leurs aventures et à la fin ils disparaissent, tout simplement.

Le Clézio reprend un procédé analogue dans *Les géants*, de 1973, où le monde d'aujourd'hui est dénoncé à travers la description d'un immense hypermarché,

¹⁷ *Le livre des fuites*, p. 108.

¹⁸ P. Lhoste (1971) p. 30.

Hyperpolis¹⁹. Dans ce supermarché apparaissent des forces anti-humaines, des “Maîtres” ou “Géants” invisibles qui manifestent leur pouvoir sur les êtres humains. Le livre est en effet une sorte de cri de révolte contre les mégapoles et leurs univers de bruits et de violence.

Entre ces deux romans, paraît en 1970 *Haiï*, un essai d’anthropologie sur la culture amérindienne où Le Clézio décrit en initié le rite magique dont usent les prêtres Embéras pour la guérison des maux, individuels et collectifs²⁰. Il compare, en analysant les différentes phases de rites Embéras, point par point, cette culture à celle de l’Occident, au détriment de l’Europe.

“Je ne sais pas trop comment cela est possible, commence le livre²¹, mais c’est ainsi : je suis un Indien. je ne le savais pas avant d’avoir rencontré les Indiens, au Mexique, au Panama. Maintenant, je le sais. [...] quand j’ai rencontré ces peuples indiens, c’est comme si j’avais connu des milliers de pères, de frères et d’épouses.”

Mydriase paraît en 1973 ; ce livre, parfois classé comme un texte “anecdotique”, parfois comme un récit poétique, décrit les modifications des perceptions de la personne sous l’effet de drogues hallucinogènes. Le mot “mydriase” désigne la dilatation des pupilles sous l’effet de certaines drogues, notamment par le datura qu’emploient les Amérindiens pour certaines cérémonies. Cette dilatation est sensée voiler la vision extérieure afin d’ouvrir une scène intérieure où le “moi” dédoublé est à la fois le spectateur et le lieu où déferlent d’étranges images, où les objets vivent avec intensité.

Grâce à sa rencontre avec le monde amérindien, Le Clézio s’éloignera des préoccupations théoriques et des débats critiques sur les rapports entre le langage

¹⁹ Il semble que Hyperpolis ait pris comme modèle le supermarché Cap 3000 à Saint Laurent du Var, près de Nice, qui venait d’ouvrir à cette époque.

²⁰ Les voyages de Le Clézio les plus marquants ont été, sans aucun doute, les séjours qu’il fit chez les indiens Embéras au Panama (1969-1973).

²¹ *Haiï*, p. 7.

et la réalité où s'attardaient les écrivains parisiens. On ne trouvera plus dans sa production littéraire les expériences de langage, de collages, de typographie insolite caractéristiques de l'école du "nouveau roman".

Les voyages de *Voyages de l'autre côté*, qui paraît en 1975, sont entrepris par une petite bande d'habitants de la ville, issus semble-t-il du folklore "pop", aux noms souvent répétés : Alligator Barks, Sursum Corda, Louise, Winston, Téclavé, Gin Fizz, Yamaha. Ils voyagent, entraînés par une étrange jeune femme-mythe, Naja Naja, qui initie ses amis à la vie inimaginable du monde qui les entoure en les conduisant au-delà de notre monde moderne et occidental, vers une autre rive, vers le silence et la liberté. L'histoire est encadrée par deux courtes sections intitulées respectivement "Watasenia" et "Pachacamac". "Watasenia" est l'évocation mythique d'une genèse, celle de la naissance de la terre, des choses et des êtres qui l'habitent, à partir des vastes étendues d'eau. C'est à "Pachacamac", ville "au cœur du monde, du désert" péruvien que s'annonce "le dernier moment de la métamorphose"²².

"Puis un jour, écrit Simone Domange²³, les nouvelles de *Mondo et autres histoires*, publiées en 1978, entre à l'école primaire." En effet, l'auteur dit "intellectuel" et dont l'œuvre est considérée comme difficile devient un auteur grand public. Mais qui a changé, le romancier ou les lecteurs ? Certes, la forme des ouvrages s'est modifiée : plus simple, plus "classique". Le Clézio lui-même en convient : "Jeune, on aime étonner, après, on va à l'essentiel."²⁴

Chacune des huit histoires de *Mondo* est animée par des enfants ou des jeunes adolescents et témoigne de la nostalgie de l'enfance et de l'innocence de la société préindustrielle. Les histoires racontent²⁵ "les étapes d'une initiation, d'une

²² *Voyages de l'autre côté*, p. 307.

²³ S. Domange (1993) : p. 7.

²⁴ Citation reprise de S. Domange (1993) : p. 8.

²⁵ G. Brée (1993) : p. 100.

rencontre lumineuse, magique, le bonheur et l'angoisse, l'enfant se trouve en présence d'une force lumineuse et vivante, qui l'appelle, l'attire et le dépasse."

La même année Le Clézio publie l'essai *L'inconnu sur la terre*, "pas tout à fait un essai"²⁶, qui, écrit en même temps que *Voyages de l'autre côté* et *Mondo et autres histoires*, constitue selon Germaine Brée²⁷ "le substratum subjectif, servant de toile de fond à ces trois livres, sorte de plaque tournante marquant une étape de la pensée et de l'écriture lecléziennes." Le livre est tourné vers la sensibilité primitive enfantine à travers la présence d'un "petit garçon inconnu" silencieux. Le texte de l'essai propose au lecteur le cheminement de la pensée vers un langage et un ton de rêverie et de méditation qui caractérisent cet essai.

Vers les icebergs, de 1978, est inspiré par la rencontre avec Henri Michaux. Ce texte est selon Jean Onimus²⁸ "une sorte de poème en prose où ses grandes aspirations spirituelles convergent dans une ambiance caractéristique inspirée d'Henri Michaux. Pris entre une introduction à la poésie de Michaux et une réédition de son poème *Inji*, ce texte est tout inspiré de l'art abrupt, dense, polysémique du poète qui a tant influencé Le Clézio."

Les débuts des années 1980 sont très productifs. En 1980 paraît *Désert* qui reprenant la confrontation de deux civilisations, raconte sur deux niveaux, d'une part le sort des hommes bleus du désert nord-africain au moment de la conquête du sud marocain par les troupes françaises entre 1909 et 1912, d'autre part l'histoire de Lalla, fille de nos jours, descendante de ces bédouins, qui est entraînée à l'immigration vers Marseille et la France.

Les onze nouvelles de *La ronde et autres faits divers*, de 1982, suivent le destin de quelques adolescents et s'intéressent aux conditions des immigrés et des exclus de

²⁶ La quatrième de couverture.

²⁷ G. Brée (1990) : p. 97.

²⁸ J. Onimus (1994) : p. 144-145.

notre hostile milieu urbain, tous angoissés et tourmentés par les difficultés du quotidien.

Voyage au pays des arbres, de 1978, seul ouvrage de Le Clézio édité pour les jeunes et écrit expressément pour ce public, n'est pourtant pas le seul texte que l'auteur ait composé pour les enfants. Des nouvelles apparues dans les différents recueils sont également parues dans des éditions jeunesse : il s'agit de *Lullaby*, *Peuple du ciel*, *La Montagne du dieu vivant* et *Celui qui n'avait jamais vu la mer* du recueil *Mondo et autres histoires*, ainsi que *La grande vie* et *Villa Aurore* de *La ronde et d'autres faits divers*. *Baalabilou*, une histoire extraite du roman *Désert* et illustré par Georges Lemoine qui raconte aux enfants le temps des djinns et le sacrifice d'un jeune homme transformé en oiseau musicien pour permettre à la belle princesse Leila d'échapper par la magie de son chant aux bêtes sauvages. Plus tard, en 1992, paraît le livre *Pavana*, qui à travers l'aventure d'un jeune garçon nous fait suivre la chasse cruelle aux baleines et la destruction d'une lagune ; d'abord édité dans une publication pour adultes, il sera repris dans la collection "Lecture Junior".

J.M.G. Le Clézio est né durant la deuxième guerre mondiale "par hasard" à Nice d'une mère bretonne et un père de nationalité anglaise, tous deux descendants des colons mauriciens. Les îles de l'océan Indien vers lesquelles l'ancêtre de Le Clézio s'est dirigé – fuyant la révolution française et les événements de Valmy – deviennent ses prochains grands buts ; le grand voyage en 1981 aboutit à deux livres : *Le chercheur d'or*, qui paraît en 1985, et *Voyage à Rodrigues* en 1986. *Le chercheur d'or* s'inspire des aventures de son grand-père paternel et du retour de l'écrivain à Rodrigues, île de l'archipel des Mascareignes, au large de l'île Maurice où le grand-père - ayant trouvé les papiers d'un corsaire indiquant en langage codé le plan de la cachette de son trésor - avait consacré de longues années à la recherche de ce trésor. *Voyage à Rodrigues*, de 1986, est un journal largement autobiographique, journal de bord qui éclaire la genèse du roman *Le chercheur d'or*. En effet, pour bien comprendre *Le Chercheur d'or*, il faudrait peut-être commencer par lire *Voyage à Rodrigues*, qui contient les clefs de l'histoire conjointe du double itinéraire des deux hommes qui parcourent la même piste à des époques différentes.

La fascination de Le Clézio pour la vie et la culture des Amérindiens est grande, il n'a jamais cessé de visiter et d'aimer cette région et y consacre des travaux importants qui dénoncent la destruction des civilisations amérindiennes du Mexique. L'essai *Trois villes saintes*, qui est publié en 1980, continue ce thème et parle des lieux sacrés.

Le Clézio traduit également des textes mythologiques indiens. "Ces livres ont été écrits aussi pour nous, comme un témoignage. Aujourd'hui, sachons les lire" dit-il. La première traduction jamais publiée en langue occidentale de *Les prophéties du Chilam Balam* paraît en 1976 et une autre traduction, *Relations de Michoacan* est publiée en 1984. Un essai, *Le rêve mexicain ou la pensée interrompue*, qui paraît en 1988, commente les textes mythologiques des amérindiens traduits auparavant.

Le recueil de cinq nouvelles, *Printemps et autres saisons*, est publié en 1989. A travers le destin de cinq femmes différentes, l'écrivain traite de "l'errance et de l'appartenance, du rêve et de la fragilité ainsi que du temps qui ne passe pas"²⁹. Ces figures de femme sont toutes exilées de leur terroir natal après "une enfance interrompue" et se trouvent dans des villes européennes confrontées à un univers hostile.

En 1990, notre auteur publie *Sirandanes*, un recueil de devinettes créoles, qu'il écrit en collaboration avec sa femme Jemia. "Le pouvoir des sirandanes, écrit-il dans la préface, comme celui de la langue créole, est celui de la jeunesse, qui survit au modernisme et aux bouleversements sociaux. Aujourd'hui, dans l'île Maurice du tourisme, de l'industrie et des crises, quel est l'avenir de cette langue créole et de son pouvoir imaginaire ? Combien de temps encore entendrons-nous les proverbes, les sirandanes, les *ségas* ?"

²⁹ La quatrième de couverture.

Durant les années 1990 paraissent de grands romans, Le Clézio confirme son succès auprès du grand public et ses livres connaissent une grande réussite commerciale. *Onitssha*, qui paraît en 1991, en partie autobiographique, est l'histoire d'un petit garçon qui embarque pour l'Afrique avec sa mère afin de rencontrer son père, médecin. Les parents de Le Clézio ont été séparés par les aléas de la guerre et le petit Le Clézio doit attendre la fin de la guerre avant de pouvoir embarquer sur le grand bateau pour le Nigeria pour rencontrer pour la première fois son père qui est médecin de bush.

Etoile errante, de 1992, qui forme un diptyque avec *Onitssha*, a pour sujet la souffrance de la deuxième guerre mondiale et l'émigration des Juifs vers l'Israël, le pays promis. L'action commence dans un petit village de l'arrière-pays niçois et se poursuit en Palestine où, à travers l'histoire d'Esther, la Juive, et celle de Nejma, la Palestinienne, le lecteur suit des destins marqués par la guerre et par l'exil. "Lorsque les Allemands ont remplacé les Italiens à Nice, en 1943, explique Le Clézio³⁰, il a fallu sortir très vite de la ville. Nous nous sommes réfugiés dans la montagne à Roquebillière, non loin du ghetto juif de Saint-Martin-Vésubie."

Diego et Frida, publié en 1993, est la biographie des deux artistes mexicains, Frida Kahlo et Diego Rivera, qui par leurs idées et leur art marquent un tournant dans la culture sud-américaine. Le Clézio partage avec le lecteur sa fascination pour leur étrange amour³¹ :

"S'il est vrai que Diego est à beaucoup d'égards un réactionnaire en amour - qui réduit le rôle des femmes à celui de mères fécondes ou de putains dispensatrices de jouissances -, l'extraordinaire force de Frida donne un sens à son appétit des plaisirs. Il est en quelque sorte l'intermédiaire entre le monde difficile qui se refuse à Frida et cet univers où tout s'harmonise, et dans lequel elle tient le rôle de la

³⁰ *Magazine Littéraire*, n° 362, février 1998, p. 33.

³¹ *Diego et Frida*, p. 183.

déesse-mère.”

1995 marque le retour de Le Clézio à l’île Maurice avec *La quarantaine*, un long roman, inspiré encore une fois des aventures de son grand-père, mis en quarantaine sur l’île Plate, au large de l’île Maurice. Le livre raconte l’histoire de son séjour forcé, au terme duquel il quitte pour toujours l’Océan indien pour vivre en France. Le point de départ du roman est la rencontre que fait le grand-père avec Rimbaud, dans un bar à Paris. “Si mon grand-père n’avait pas choisi de revenir en France, après cette expérience de quarantaine au large de l’île Maurice, il ne se serait pas marié, et je ne serais sans doute pas né...” dit-Le Clézio³².

Le Clézio publie *Poisson d’or* en 1997. Le roman raconte le sort vagabond et tragique d’une jeune fille d’Afrique du Nord, Laïla, petite fille noire volée à ses parents, battue et vendue, qui parcourt la planète à la recherche de ses racines, de son pays et réussit à aller jusqu’au bout du monde. En quinze années, de l’Afrique à Paris, en passant par Marseille, la jeune fille finit par revenir à son point de départ : dans le sable du désert... “C’est le récit, dit Le Clézio, de quelqu’un au terme de son errance”. Ce livre est sans doute le roman le plus politique de Le Clézio ; c’est en effet l’histoire des sans-papier, une métaphore de l’exil, de la recherche de soi, et de sa place dans le monde.

Quelques mois après, il publie *La fête chantée* qui est un recueil de textes ethnologiques qui continue le thème de la culture amérindienne ; il puise son expérience chez ces peuples et dans la compréhension de leurs rites et leurs coutumes.

“Il y a une vingtaine d’années, entre 1970 et 1974, écrit Le Clézio sur le rabat de la couverture, j’ai eu la chance de partager la vie d’un peuple amérindien, les Embéras, [...] dans la province du Darién au Panamá, expérience qui a changé toute ma vie, mes idées sur le

³² *Magazine Littéraire*, n° 362, février 1998, p. 33.

monde et sur l'art, ma façon d'être avec les autres, de marcher, de manger, de dormir, d'aimer, et jusqu'à mes rêves. J'appris la vanité des objets de notre monde de la consommation, vains parce que la chaleur, l'humidité, les insectes les rendent inutilisables."

En 1997 paraît également *Gens de nuages*, un livre de J.M.G. Le Clézio et de sa femme Jemia Le Clézio, qui est une sorte de journal de voyage d'un séjour dans le désert du sud du Maroc. A l'origine le livre est un double trajet : pour J.M.G. Le Clézio qui depuis ses séjours chez les Amérindiens est à la recherche d'une cohérence, d'un équilibre philosophique entre le cérébral et le physique, entre la culture occidentale et la culture "primitive" et pour Jemia c'est un retour aux sources, à la terre de ses ancêtres. "Les gens des nuages, qui habitent le désert, écrit Gérard de Cortanze³³, ne sont pas des curiosités. Ils ont quelque chose à nous dire sur ce que nous sommes, sur le monde que nous habitons, qui est le même que le leur. Parce qu'ils sont proches de la vérité universelle, ils sont, malgré les apparences, en avance sur nous."

1999 voit paraître un livre qui contient deux romans : *Hasard* suivi d'*Angoli Mala* (écrit en 1985), où les thèmes du grand voyage et de l'aventure sont confrontés au sentiment de déracinement, et au passage de l'enfance à l'adolescence.

Le premier, *Hasard* (le plus long), est l'histoire d'une adolescente, Nassima, hantée par le souvenir de son père qui a abandonné sa famille pour courir le monde. Nassima va parvenir à s'embarquer clandestinement sur le Azzar, yacht splendide ancré dans le port de Villefranche, qui appartient à un aventurier richissime et solitaire, à la fois cinéaste et forban, aussi abandonné dans la vie que la petite fille déguisée en garçon qui s'est réfugiée à son bord. Elle s'enfonce avec lui dans les profondeurs océanes, à la recherche d'un absolu qui se dérobe toujours à leur soif hallucinée.

³³ *Magazine Littéraire*, n° 362, février 1998, p. 43.

Angoli Mala, la seconde histoire, raconte le retour au pays d'un jeune Indien, Bravito, élevé à Panama. A la recherche de ses origines, des splendeurs et des légendes d'un paradis perdu, Bravito découvre un peuple dégradé par l'alcool qui se traîne dans les chaînes d'un nouvel esclavage, celui des petits tyrans locaux et des intérêts étrangers. Malgré l'amour d'une jeune noire, il deviendra presque fou de solitude et de désespoir, dans sa révolte sans issue. Il finira assassiné.

En l'an 2000, paraît *Cœur brûle et autres romances*. C'est un recueil de sept nouvelles qui racontent la vie des adolescents quittant l'incertitude de l'enfance et l'histoire d'une vieille femme qui épuise la fin de ses jours à se remémorer des souvenirs évanouis. L'appel à l'aventure suscite aussi de pitoyables désillusions où l'exil marque de façon inoubliable la destinée des jeunes aventuriers.

Le Clézio est très productif et sa production ne se limite pas aux livres que nous venons de citer ; il a notamment collaboré à l'élaboration d'*Enfances*, un livre de photos d'enfants dont les légendes sont de l'auteur.

En outre, dans un souci de sauvegarde des textes fondateurs de groupes ethniques, il dirige avec Jean Grosjean la collection *L'Aube des peuples* chez Gallimard, qu'il a lui-même enrichi de ses traductions préfacées des livres maya et porhépecha : *Les prophéties du Chilam Balam* et *La relation de Michoacan*.

Nous ne faisons pas ici mention de nombreux textes, principalement des nouvelles et des extraits de livres "à paraître" publiés au cours des années dans des revues et différents ouvrages, témoignant de la richesse de l'œuvre leclézienne. Le Clézio a écrit de nombreux articles, notamment de critique littéraire, et ses contributions aux œuvres collectives sont très abondantes.

J.M.G. Le Clézio a reçu de nombreux prix depuis le prix Renaudot qu'il a reçu pour son premier livre, *Le procès-verbal*. En 1980, il a été le premier à recevoir le prix Paul Morand, pour la totalité de son oeuvre. Plus tard, en 1994, il a été élu le plus grand écrivain vivant de langue française. En 1997, il s'est vu attribuer le

Grand Prix Jean Giono pour son roman *Poisson d'or* et, en 1998, il a reçu le prix littéraire de la Fondation Prince-Pierre de Monaco.

L'auteur et son œuvre ont été l'objet de nombreuses études universitaires et d'ouvrages de critique littéraire. Les travaux sont nombreux et il serait presque impossible de tous les citer, nous nous bornerons donc à mentionner quelques excellents ouvrages. L'ouvrage de Michelle Labbé, *Le Clézio, l'écart romanesque*³⁴ et celui de Germaine Brée, *Le monde fabuleux de J.M.G. Le Clézio*³⁵, s'intéressent à l'ensemble de l'œuvre, tandis que la thèse de Claude Cavallero, *J.M.G. Le Clézio : les marges du roman*³⁶ s'intéresse plus particulièrement aux débuts de l'écrivain. Certains ouvrages sont consacrés à des livres en particulier, notamment le commentaire de François Marotin de *Mondo et autres histoires*³⁷ et l'étude de Simone Domange du roman *Désert*³⁸. L'ouvrage *Chemins pour une approche poétique monde. Le roman selon J.M.G. Le Clézio* de Miriam Stendal Boulos³⁹ traite l'approche poétique du texte et Sandra L. Beckett s'intéresse à la littérature pour enfants dans son *De grands romanciers écrivent pour les enfants*⁴⁰. Plusieurs publications, notamment d'Amérique du Nord, font une analyse psychologique de l'œuvre parmi lesquelles nous trouvons l'ouvrage *L'œil du serpent, dialectique du silence dans l'œuvre de J.-M. G. Le Clézio*⁴¹ de Ruth Holzberg et celui de Jennifer Waelti-Walters ; *Icare ou l'évasion impossible, Étude psycho-mythique de l'œuvre de J.M.G. Le Clézio*⁴² qui ont suscité l'intérêt des spécialistes.

³⁴ M. Labbé, *Le Clézio, l'écart romanesque*, Paris, L'Harmattan, 1999.

³⁵ G. Brée, *Le monde fabuleux de J.M.G. Le Clézio*, Amsterdam – Atlanta, Rodopi, Collection Monographique Rodopi en Littérature Française Contemporaine, 1990.

³⁶ C. Cavallero, *J.M.G. Le Clézio : les marges du roman*, Thèse de doctorat, Université de Rennes, 1992.

³⁷ F. Marotin, *Mondo et autres histoires de J.-M. G. Le Clézio*, Paris, Éditions Gallimard, Collection Folio, Foliothèque, 1995.

³⁸ S. Domange, *Le Clézio ou la quête du désert*, Paris, Éditions Imago, 1993.

³⁹ M. Stendal Boulos, *Chemins pour une approche poétique monde. Le roman selon J.M.G. Le Clézio*, Copenhague, Etudes romanes n° 41, Museum Tusculanum Press, University of Copenhagen, 1999.

⁴⁰ S. L. Beckett, *De grands romanciers écrivent pour les enfants*, Montréal, Les Presses de l'Université de Montréal – Éditions littéraires et linguistiques de l'université de Grenoble, Collection espace littéraire, 1997.

⁴¹ R. Holzberg, *L'œil du serpent, dialectique du silence dans l'œuvre de J.-M. G. Le Clézio*, Sherbrooke, Québec, Éditions Naaman, 1981.

⁴² J. Waelti-Walters, *Icare ou l'évasion impossible, Étude psycho-mythique de l'œuvre de J.M.G. Le Clézio*, Sherbrooke, Québec, Éditions Naaman, 1981.

Les études de l'œuvre augmentent constamment et J.M.G. Le Clézio continue à écrire des livres. Un livre, parfois plusieurs, sont publiés chaque année. Il nous a fallu, dans notre étude, à un moment donné, délimiter notre corpus.

1.2. Le corpus.

“Écrire, si ça sert à quelque chose, ce doit être à ça : à témoigner. A laisser ses souvenirs inscrits, à déposer doucement, sans en avoir l’air, sa grappe d’œufs qui fermenteront. Non pas à expliquer, parce qu’il n’y a peut-être rien à expliquer ; mais à dérouler parallèlement.”

L’extase matérielle, page 105-106.

Le corpus est défini comme une collection finie et intangible de textes réunis pour l’étude d’un phénomène langagier ou littéraire comme dans notre étude.

Nous avons achevé la numérisation des textes lecléziens en 1999. Le roman *Hasard* est donc la dernière œuvre entrant dans le corpus. Etant donné l’extraordinaire productivité de J.M.G. Le Clézio nous savions que d’autres textes paraîtraient certainement avant l’achèvement de cette étude – ce fut bien le cas avec *Cœur brûlé* – qu’il n’a pas été possible d’intégrer à moins de recommencer au début l’ensemble des traitements.

1.2.1. Le choix d’un corpus : les œuvres écartées.

Notre but, dans cette étude quantitative, a été de créer un corpus le plus complet possible, parmi la quarantaine d’ouvrages de Le Clézio, pour travailler sur l’ensemble de l’œuvre, afin de répondre aux impératifs de la statistique lexicale et ainsi de donner l’image la plus juste du vocabulaire et du style de cet auteur.

Pourtant, quelquefois, le style de l’écrivain crée des difficultés quant au traitement informatique du texte.

Dans les années soixante à soixante-dix en effet, J.M.G. Le Clézio se livre à de nombreuses innovations : collages, textes raturés, citations d’articles de journaux

ou de prospectus. Il invente même une langue imaginaire. Il s'agit, selon Maurice Nadeau⁴³, de

“... descriptions qui sont au paysage ce que la carte postale est au tableau... passages d'une typographie à une autre... cette licence que se donne l'auteur, dans la façon dont il conçoit et mène son récit n'a d'autre but que de multiplier les points de vue de son objet : la peinture du vivant, comme si, de tous les horizons, convergeaient les rayons susceptibles d'illuminer ce point focal.”

Les livres qui contiennent de nombreuses originalités typographiques causent évidemment des difficultés pour la numérisation du texte au scanner et la reconnaissance optique des caractères ainsi que pour l'interprétation des données ; leur prise en compte fausserait les résultats des analyses statistiques.

Nous avons donc écarté quatre livres de notre corpus :

1. *Terra Amata* de 1967, où Chancelade décrit sa propre mort et les sentiments qu'il éprouve en tant que cadavre, occasionne de nombreuses difficultés de traitement. Parfois, J.M.G. Le Clézio utilise des signes au lieu de lettres de l'alphabet, par exemple des points ou des tirets pour illustrer que le personnage “parle en silence”. Ceci fausse évidemment les statistiques que nous les traitons ou que nous les excluons⁴⁴.

⁴³ M. Nadeau (1992) : p. 184.

⁴⁴ Cf. *Terra amata*, p. 110.

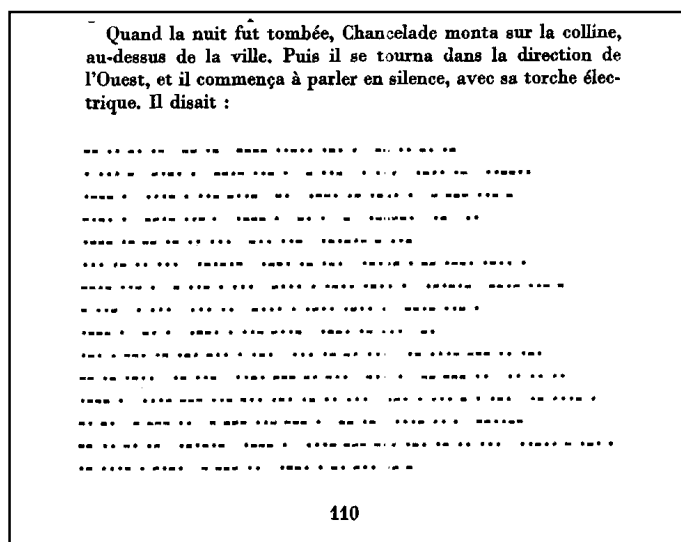


Figure n°2 : “Parler en silence” dans *Terra amata*.

D’autres parties du livre présentent des phrases saccadées, séparées par des barres obliques et des guillemets⁴⁵.

Si / « Non » / Si / « Non » / Les / « Les pommes de terre » / Ce n’est pas ça / « Les olives » / Non / « Les autos » / Non les / « Les » / Vite il faut le dire / « Je ne peux pas c’est trop dangereux » / Vite / « J’ai peur » / Vite vite / « Les » / Les / « Les loups sont noirs » /

Le Clézio invente en outre dans ce livre une langue. L’intégration de ces mots non français dans un dictionnaire de l’œuvre perturbe évidemment les analyses automatiques du vocabulaire⁴⁶.

⁴⁵ Cf. *Terra amata*, p. 66.

⁴⁶ Cf. *Terra amata*, p. 119.

EN DISANT DES MOTS INCOMPRÉHENSIBLES

Woolikanoc mana bori oclakokok. Zane prestil zani wang don bang. Geod de molladda apudax predongxi, ette, lalarus toolynk füranpelek gene, etti sali akka mundiu, chien roxudal woombyi waa nochli maabara sata klo kluoc. Fam rezon, fam loop griçetka sama super.

— Bojun vendery? banje Zaka.

— Kaz.

— Abele m'n poostu! Trix slamoc jdiokdong denyi, para munok fla monkx belu rezon mana jel silurgonge, staars hok.

— Kaz.

— Stunibo arhun chlajondyi.

— Ho. Bendri ravuda kok Trekei?

— Trekei.

— Ho.

— Kazniebolsk, pojnun handrei Trekei doulowo yan'de assa-mabulok.

Frjena degun paramanuk mizkanulunre delbo anuk delco, de Wü.

— Wü-o-kuchun-punku-rügani-yugwi-van-ntü-m(ü).

— Tumengao-ak.

Betsewdepaw dungsk selo-ubo-uti, mana bori oklako-oak, pendeba, m, astaro flemdiüs; grehonk flüstye merced, ozier ellgok vaarista, junje sapir trebik gonoktogo-ak duyž banje samovülk vuntu arboo dochlawatznyé. Ob dabok-ka-paok

Figure n°3 : Invention d'une langue dans Terra Amata.

2. Le roman *Les géants* de 1973, qui est la description de l'hypermarché Hyperpolis où les gens perdent toute existence, contient de nombreux collages et des publicités⁴⁷.

⁴⁷ Cf. *Les Géants*, p. 343.

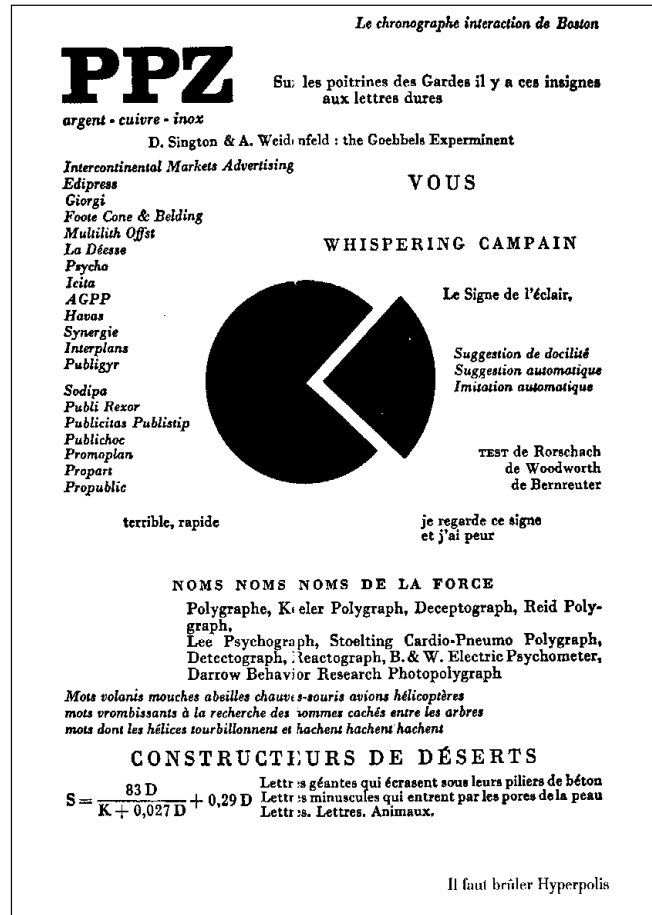


Figure n°4 : Collage publicitaire dans Les géants.

De même, il y a dans ce livre une volonté de jouer avec les lettres. Elles tracent ou elles composent des mots sous forme de dessins. L'exemple suivant illustre le paysage "plein de signes"⁴⁸.

⁴⁸ Cf. *Les Géants*, p. 93.

[illegible]

Figure n°5 : Le “paysage plein de signes” dans Les géants

Nous y trouvons encore des pages entières remplies de langage technique et de fichiers de texte scientifique. L'exemple ci-dessous montre une page de “langage machine” des années soixante (ALGOL)⁴⁹.

⁴⁹ Cf. *Les Géants*, p. 184.

```

real      a, b, c, D min (  $\uparrow^2 = c - b \uparrow^2 a$  at time
               $t = -b / a$ ), t;
a := (Vx [r] - Vx [s])  $\uparrow^2$  + (Vy [r] -
      Vy [s])  $\uparrow^2$ ;
b := Px [r] - Px [s] * (Vx [r] - Vx [s] +
      (Py [r] - Py [s]) * (Vy [r] - Vy [s]));
c := (Px [r] - Px [s])  $\uparrow^2$  + (Py [r] - Py [s])
       $\uparrow^2$ ;
D min := sqrt (c-b  $\uparrow^2$  2a);
t := -ba;
if
begin
  D min < gap and t > 0 then
    write text (' ) the $ ships $ now $ in $ positions (');
    write (Px [r]); write text (' ');
    write (Py [r]); write text (' $ and $ (');
    write (Px [s]); write text (' ');
    write (Py [s]); write text (' $ will $ close $ to $ a $
      distance $ '); write text (' $ in $ a $ time $ ');
    write (D min);
    write (t);

      end

    end

  end

  comment
  real      Ax, Ay, Bx, By, Cx, Cy, interval, gap;
  integer    N, r; Boolean first scan;
  on watch : set fixed data (Ax, Ay, Bx, By, Cx, Cy, interval, gap);
  restart :  first scan := true
              count (N);
  reset :    begin real array Bearing A, Bearing B, Bearing C,
              Px, Py, Old x, Old y, Vel x, Vel y (1 = N);
              repeat : scan (Bearing A, Bearing B, Bearing C,
                Bearing (, N);
                fix (Ax, Ay, Bx, By, Cx, Cy, N, Bearing A,
                  Bearing B, Bearing C, Px, Py, ALARM);

```

184

Figure n°6 : Exemple du langage machine dans *Les géants*.

3. Le livre *Haiï*, de 1971, est le très beau récit qui raconte la rencontre décisive de J.M.G. Le Clézio avec le monde amérindien ; il a également été écarté de notre corpus. Dans cet ouvrage qui contient de nombreuses photographies et illustrations, le texte est étroitement lié à l'image et commente souvent les illustrations. Il nous a semblé que l'exploitation simultanée de ces deux expressions artistiques différentes s'adapterait mal à la statistique lexicale et nous l'avons par conséquent écarté de notre étude.

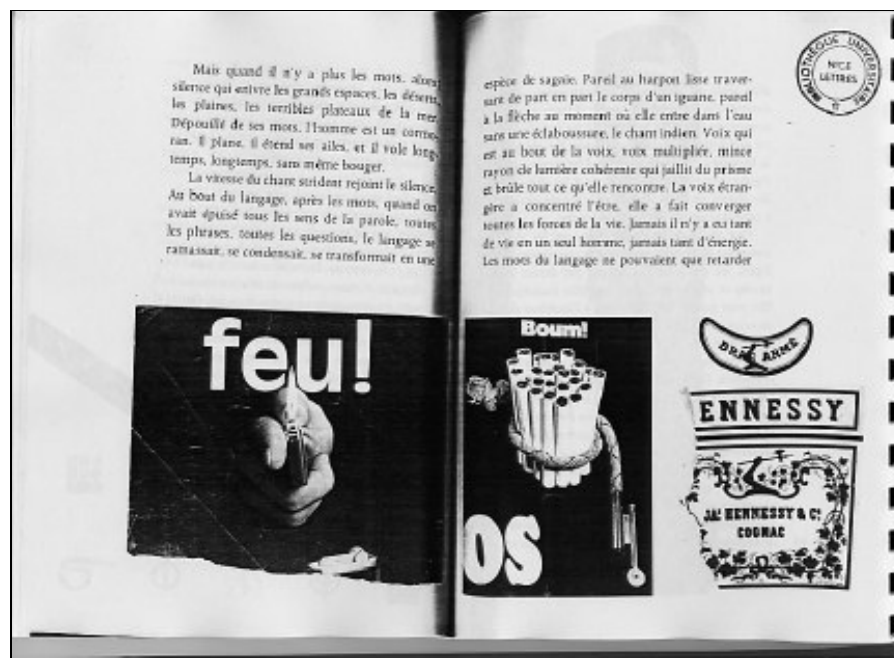


Figure n°7 : Une double page de Haï.

4. Nous n'avons pas inclus non plus dans notre corpus *La relation de Michoacán*, paru en 1984, car il s'agit d'une traduction, par Le Clézio, d'un livre sacré du peuple Maya. La petite présentation écrite par l'auteur-traducteur est trop courte pour motiver sa présence dans notre corpus⁵⁰.

Quant aux textes courts, dont nous avons dit au chapitre précédent que nous n'en ferions pas mention, ils ont été écartés en raison des impératifs du traitement statistique qui exige une certaine homogénéité de taille des sous-corpus.

Telles sont les raisons de nos choix, essentiellement techniques et imposés par la méthodologie utilisée. Nous pensons cependant qu'un tel corpus, par son ampleur, satisfait à la fois aux impératifs du traitement statistique et à notre désir de donner une image aussi complète que possible du vocabulaire - et, par là-même, de l'œuvre - de Le Clézio.

⁵⁰ L'introduction de l'autre traduction *Les prophéties de Chilam Balam*, incluse dans le corpus est plus importante.

1.2.2. Présentation des œuvres retenues.

“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. C’est la seule expérience à peu près probante. Après cela, on n’aurait plus qu’à se taire.”

L’extase matérielle, page 132.

Notre corpus principal s’étend sur presque quarante ans, de 1963 à 1999, et représente la quasi-totalité de la production de Jean-Marie Gustave Le Clézio, en englobant trente et un textes complets soit plus de deux millions de mots. On y trouve des romans, des recueils de nouvelles, des essais littéraires, des ouvrages de type ethnologique ainsi que des livres pour les enfants et pour la jeunesse.

Le tableau suivant regroupe, par ordre chronologique, les différents ouvrages présents dans le corpus qui constitue la base de notre travail, avec les abréviations utilisées dans l’exploitation informatique ainsi que les références des éditions employées.

No	Titre	Code	Parution	Éditeur
1	Le procès-verbal	PROCES, Pv	1963	©Gallimard, Folio n° 353
2	La fièvre	FIEVRE, Fi	1965	©Gallimard, L’imaginaire n°253
3	Le déluge	DELUGE, De	1966	©Gallimard, L’imaginaire n°309
4	L’extase matérielle	EXTASE, Ex	1967	©Gallimard, Folio n° 212
5	Le livre des fuites	FUITES, Fu	1969	©Gallimard, L’imaginaire n°225
6	La guerre	GUERRE, Gu	1970	©Gallimard, L’imaginaire n°271
7	Mydriase	MYDRIASE, My	1973	©Fata Morgana
8	Voyages de l’autre côté	VOYAGE, Vo	1975	©Gallimard nrf
9	Les prophéties de Chilam Balam	PROPHETIES, Pr	1976	©Gallimard nrf, le chemin

10	Mondo et autres histoires	MONDO, Mo	1978	©Gallimard, Folio n° 1365
11	L'inconnu sur la terre	INCONNU, In	1978	©Gallimard nrf, le chemin
12	Vers les icebergs	ICEBERGS, Ic	1978	©Fata Morgana
13	Voyages au pays des arbres	ARBRES, Ar	1978	©Gallimard, Folio Cadet
14	Désert	DESERT, Ds	1980	©Gallimard, Folio n° 1670
15	Trois villes saintes	VILLES, Vi	1980	©Gallimard, nrf
16	La ronde et autres histoires	RONDE, Ro	1982	©Gallimard, Folio n° 2148
17	Le chercheur d'or	CHERCHEUR, Ch	1985	©Gallimard, Folio n° 2000
18	Angoli Mala	ANGOLI, An	1985	©Gallimard nrf
19	Voyage à Rodrigues	RODRIGUES, Rd	1986	©Gallimard, Folio n° 2949
20	Le rêve mexicain	REVE, Re	1988	©Gallimard, Folio Essais n° 178
21	Printemps et autres histoires	PRINTEMPS, Pi	1989	©Gallimard, Folio n° 2264
22	Sirandanes	SIRANDANES, Si	1990	©Seghers
23	Onitsha	ONITSHA, On	1991	©Gallimard, Folio n° 2472
24	Etoile errante	ETOILE, Et	1992	©Gallimard, Folio n° 2592
25	Pawana	PAWANA, Pa	1992	©Gallimard nrf
26	Diego et Frida	DIEGO, Di	1993	©Stock
27	La quarantaine	QUARANTAINE, Qu	1995	©Gallimard, Folio n° 2974
28	Poisson d'or	POISSON, Po	1997	©Gallimard nrf
29	La fête chantée	FETE, Fe	1997	©Le Promeneur
30	Gens des nuages	NUAGE, Nu	1997	©Stock
31	Hasard	HASARD, Ha	1999	©Gallimard nrf

Tableau n°1 : La répartition des ouvrages constitutifs du corpus.

Comme nous l'avons vu dans le chapitre précédent, ces livres traitent de thèmes divers et appartiennent à des genres littéraires différents. Tout d'abord, on trouve les six premières œuvres, classées, par leur style particulier et innovant, comme appartenant à l'École du "nouveau roman" : *Le procès-verbal*, les nouvelles de *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre* et *Voyages de l'autre côté*.

Les romans qui suivent cette période, considérés par les critiques comme plus "traditionnels", sont au nombre de neuf : *Désert*, *Le chercheur d'or* et *Voyage à Rodrigues* écrit sous forme de journal personnel, *Angoli Mala*, *Onitsha*, *Etoile errante*, *La quarantaine*, *Poisson d'or*, et *Hasard*.

Mydriase et *Vers les icebergs* sont difficiles à classer dans un genre précis, ce sont plutôt des récits poétiques. Lorsque certaines critiques les rapprochent de la poésie en prose, d'autres parlent de textes anecdotiques⁵¹.

Nous avons ensuite inclus les recueils de nouvelles : *Mondo et autres histoires*, *La fièvre* (que nous avons également compté parmi les ouvrages "nouveau roman"), *La ronde et autres faits divers* ainsi que *Printemps et autres saisons*.

Les essais littéraires sont de différentes époques. *L'extase matérielle* et *L'inconnu sur la terre* traitent de thèmes généraux tandis que *Trois villes saintes* et *Le rêve mexicain ou la pensée interrompue* s'intéressent exclusivement à la culture amérindienne.

Cette culture est également le principal intérêt des ouvrages à vocation ethnologique, *Les prophéties du Chilam Balam* et *La fête chantée*, tandis que *Sirandanes* s'intéresse à la culture de l'île Maurice.

Sont inclus en outre dans le corpus deux livres pour enfants : *Voyage au pays des arbres* et *Pawana*, un livre qui a été écrit pour des adultes mais qui se trouve désormais uniquement dans des collections "Lecture jeunesse". Sont présents enfin *Diego et Frida*, la seule biographie, et *Gens des nuages*, le récit de voyage⁵².

Afin d'avoir une vue d'ensemble de ce vaste corpus et des différents genres qu'il regroupe⁵³, nous présentons ci-dessous un tableau récapitulatif des ouvrages du corpus classés selon les genres littéraires⁵⁴ :

⁵¹ Nous reviendrons plus tard dans cette étude (chapitre 2.3.) à la division générique de l'œuvre leclézienne.

⁵² D'autres ouvrages, notamment les essais littéraires et les ouvrages ethnologiques pourraient éventuellement être inclus dans ce genre.

⁵³ En annexe n° 1, nous avons inclus une courte présentation de chaque ouvrage représenté dans le corpus, nous permettant de nous situer rapidement parmi les trente et un livres tout au long de ce travail. Cette présentation est souvent celle donnée par l'éditeur correspondant à la quatrième de couverture.

⁵⁴ Cette classification schématique pourrait être discutable (cf. chapitre 2.3.2.) mais nous la gardons en première approximation.

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

Tableau n°2 : Classification du corpus selon les genres littéraires.

1.2.3. Le choix des éditions.

La majeure partie de la production littéraire de J.M.G. Le Clézio a été publiée par la maison d'édition Gallimard, dans différentes collections : “La nouvelle revue française”, “L’imaginaire” et dans la collection Folio en “livre de poche”. Nous avons essayé, dans la mesure du possible, de retenir la première version publiée, mais du fait des contraintes techniques et pratiques, nous avons souvent dû exploiter la collection de poche Folio⁵⁵.

⁵⁵ La collection Folio est la plus accessible et disponible en librairie et sa typographie s’adapte particulièrement bien à la lecture optique par scanner.

Deux ouvrages ont en outre été publiés par la maison d'éditions Fata Morgana : *Mydriase* et *Vers les Icebergs*. *Sirandanes* est paru chez Seghers et la maison d'édition Stock a publié *Diego et Frida* ainsi que *Gens des nuages*. Il va de soi que nous utiliserons ces textes aux fins exclusives de la recherche universitaire⁵⁶.

Dans l'œuvre leclézienne, il n'y a pas le problème d'éditions multiples comme chez certains écrivains. Nous n'y trouvons pas de transformations ultérieures des oeuvres après leur parution⁵⁷ et la pagination des différentes éditions est quasiment la même.

1.2.4. Le classement chronologique.

Le procès-verbal est le premier roman de Le Clézio à être publié, mais l'auteur écrit en réalité depuis l'âge de sept ans. Pour un écrivain comme lui qui écrit tout le temps et pour lequel "l'écriture est une question de vie ou de mort, conditionnée par un besoin intérieur", le moment de l'écriture et celui de la parution ne coïncident pas toujours et le décalage peut être important.

Or, la nature de notre recherche, par exemple dans l'étude de l'évolution du vocabulaire, demande une chronologie correcte. Non seulement l'incertitude attachée à différentes indications temporelles mais aussi le fait qu'un livre n'est pas forcément écrit d'un seul jet et qu'il peut y avoir des interruptions très longues entre le début de l'écriture d'un livre et son aboutissement, introduisent un facteur d'ambiguïté sur les datations. Nous avons donc dans ce travail choisi de respecter l'ordre chronologique de la première parution de chaque ouvrage sauf dans un cas, celui du dernier livre du corpus. Celui-ci contient en effet, comme nous l'avons déjà dit, deux récits : *Hasard* et *Angoli Mala*. A la fin de chaque récit, l'écrivain nous

⁵⁶ L'œuvre de Le Clézio est toujours sous copyright.

⁵⁷ Le Clézio a souvent exprimé l'envie de réécrire ses livres, ce qu'il n'a pourtant pas fait, à ce jour.

indique les lieux et dates de l'écriture, *Albuquerque 1999* pour le premier et *Jacona, septembre 1985* pour le second. Celui-ci a donc été écrit 14 ans avant sa parution. Le décalage chronologique nous a paru trop important pour être négligé et nous avons donc séparé les deux récits pour les insérer à leur juste place dans la chronologie.

Par ailleurs, nous avons contacté l'écrivain pour avoir plus de précisions sur la chronologie de l'œuvre ; J.M.G. Le Clézio nous a expliqué que, globalement, l'ordre chronologique de l'écriture et celui des différentes parutions sont les mêmes sauf dans le cas du roman *Etoile errante*, qui a été écrit longtemps avant de paraître, au milieu des années 1980. Comme le roman, en partie, se déroule en Israël et en Palestine et que le moment coïncidait avec les mouvements de l'Intifada, Le Clézio a préféré attendre avant de le publier, ne voulant pas que son livre ait une connotation politique. Toutefois, comme le livre forme un diptyque avec *Onitssha* et qu'il a été écrit sur une assez longue période, nous ne l'avons pas changé de rang dans l'ordre chronologique du corpus.

L'ordre chronologique complexe et plus encore l'étendue générique que nous venons d'évoquer sont des facteurs importants dans notre choix de créer plusieurs variantes de ce grand corpus.

1.2.5. Les différentes versions du corpus.

La taille importante et la relative complexité de ce corpus nous ont amenée, au fur et à mesure que notre travail a progressé, à créer différentes versions et sous-corpus permettant de répondre à des besoins différents. Notre corpus "Le Clézio" a en effet été décliné en plusieurs versions que nous utiliserons dans les diverses analyses de notre étude. L'alternance des versions que nous trouverons tout au long de cette étude permet de cibler des intérêts de recherche divers, soit concernant la méthode, soit concernant le contenu. Ces divisions et versions peuvent peut-être, à première vue, sembler assez compliquées. Pourtant, dans

notre ambition d'une comparaison méthodologique et technique tout au long de cette étude, il s'est avéré nécessaire de clairement délimiter les différentes bases sur lesquelles s'appuient nos diverses analyses, afin de pouvoir en tirer des conclusions adéquates.

Pour pouvoir comparer différentes méthodes statistiques⁵⁸, nous disposons, en plus de notre corpus de 31 œuvres s'appuyant sur les formes graphiques, du même corpus lemmatisé avec le programme Cordial. Sachant que l'inégalité de taille des sous-corpus pourrait être, dans certaines circonstances, un facteur d'erreur, nous avons un corpus écartant les textes de petite taille, regroupant ainsi des textes de taille relativement homogène. Nous avons également créé quatre versions d'un corpus réduit à 12 œuvres. Une première version basée sur les formes graphiques, une deuxième lemmatisée avec Cordial et exploitée par Hyperbase, une troisième lemmatisée selon la méthode Labbé exploitée par le logiciel Lexicométrie et une quatrième version lemmatisée selon la méthode Labbé et exploitée par Hyperbase⁵⁹.

Nous disposons également de différents sous-corpus qui permettent des analyses sémantiques ciblées. Ces sous-corpus présentent principalement les données sous formes graphiques. Un sous-corpus "fiction" a été ainsi créé qui comprend exclusivement les romans et les recueils de nouvelles ; la comparaison avec d'autres auteurs est également possible grâce à des corpus "comparés"⁶⁰. Enfin, l'étude sur l'intertextualité et l'alternance des voix s'appuie sur un corpus composé des romans *Désert*, *Onitsha*, *Etoile errante* et *La quarantaine*.

Le tableau n°3 rend compte du regroupement des différentes versions sur lesquelles s'appuieront les analyses dans les différents chapitres :

⁵⁸ Nous y reviendrons dans les chapitres 2.4. et 2.5.

⁵⁹ Ces techniques seront expliquées dans le chapitre 2.5.

⁶⁰ Les corpus N, O, et P seront exploités dans la partie V : "J.M.G. Le Clézio et les autres."

Corpus :	Composition :	Logiciels :
Corpus A	Formes graphiques 31 œuvres	Hyperbase
Corpus B	Lemmatisé Cordial 31 œuvres	Hyperbase
Corpus C	Corpus Homogénéisé 21 œuvres	Hyperbase
Corpus D	Corpus Romanesque formes graphiques 17 œuvres	Hyperbase
Corpus E	Corpus Romanesque lemmatisé 17 œuvres	Hyperbase
Corpus F	Formes graphiques 12 œuvres	Hyperbase
Corpus G	Lemmatisé Cordial 12 œuvres	Hyperbase
Corpus H	Lemmatisé Labbé 12 œuvres	Lexicométrie
Corpus I	Lemmatisé Labbé 12 œuvres	Hyperbase
Corpus K	Lemmatisé Winbrill, 31 œuvres	Hyperbase
Corpus L	Formes graphiques récits intercalés ¹	Hyperbase
Corpus M	Formes graphiques récits intercalés ²	Hyperbase
Corpus N	Formes graphiques (34 œuvres) Corpus Le Clézio-Gracq	Hyperbase
Corpus O	Formes graphiques (3 œuvres) Corpus romans d'aventures	Hyperbase
Corpus P	Formes graphiques (2 œuvres) Corpus Le Clézio - Conrad	Hyperbase

Tableau n°3 : Les différentes variantes du corpus.

Les corpus A, B et K sont les corpus qui englobent la totalité du corpus. Le corpus C (le corpus homogénéisé) comporte les 21 œuvres suivantes : *Le procès-verbal*, *La fièvre*, *Le déluge*, *L'extase matérielle*, *Le livre des fuites*, *La guerre*, *Voyages de l'autre côté*, *Mondo et autre histoires*, *L'inconnu sur la terre*, *Désert*, *La ronde et autres faits divers*, *Le chercheur d'or*, *Le rêve mexicain*, *Printemps et autres saisons*, *Onitssha*, *Etoile errante*, *Diego et Frida*, *La quarantaine*, *Poisson d'or*, *La fête chantée* et *Hasard*.

Les différentes versions du corpus romanesque sont composées de : *Le procès-verbal*, *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre*, *Voyages de l'autre côté*, *Mondo et autre histoires*, *Désert*, *La ronde et autres faits divers*, *Le chercheur d'or*, *Voyage à Rodrigues*, *Printemps et autres saisons*, *Onitssha*, *Etoile errante*, *La quarantaine*, *Poisson d'or* et *Hasard*.

Quant aux diverses variantes du corpus de 12 œuvres⁶¹, elles contiennent : *Le procès-verbal*, *La fièvre*, *Mydriase*, *Voyages de l'autre côté*, *Mondo et autre histoires*, *Désert*, *La*

⁶¹ Les critères du choix de ces variantes du corpus sont essentiellement d'ordre technique.

ronde et autres faits divers, Voyage à Rodrigues, Printemps et autres saisons, Pavana, La quarantaine et Hasard.

Le corpus romanesque, qui exclut les autres genres littéraires, s'est également avéré indispensable, permettant des comparaisons avec d'autres auteurs littéraires qui ne produisent pas dans tous les différents genres présents dans la production leclézienne. Or, les genres littéraires, nous le savons depuis le tout début des études lexicométriques⁶², sont un facteur prépondérant dans toute étude de statistique lexicale.

Il est vrai que la division générique selon le tableau que nous avons proposé ci-dessus n'est pas tout à fait catégorique et peut être sujette à discussion. Les différents niveaux de classification générique s'emboîtent souvent : *La fièvre* par exemple est un recueil de nouvelles que nous avons classé parmi les romans appartenant à l'école du "nouveau roman", les essais littéraires sont très proches des ouvrages ethnologiques et les deux types de textes abordent des problèmes philosophiques aussi bien que culturels ou anthropologiques, notamment lorsqu'il s'agit des ouvrages qui s'intéressent à la culture amérindienne.

En effet, la notion de genres littéraires ne va pas de soi en règle générale et chez un auteur comme J.M.G. Le Clézio, qui la conteste sciemment, elle est évidemment encore plus délicate à manier.

Nous nous proposons donc, avant de décrire plus en détail les méthodes de notre étude et les normes sur lesquelles elles s'appuient, d'approfondir la question des genres littéraires qui joue un rôle important dans toute étude lexicométrique et plus particulièrement celle de l'œuvre leclézienne.

⁶² Cf. Ch. Muller (1964) et (1966), É. Brunet (1978) et (1981).

1.3. Les genres littéraires.

“La poésie, les romans, les nouvelles sont des singulières antiquités qui ne trompent plus personne, ou presque. Des poèmes, des récits, pourquoi faire ? L’écriture, il ne reste plus que l’écriture, l’écriture qui tâtonne avec ses mots, qui cherche et décrit, avec minutie, avec profondeur, qui s’agrippe, qui travaille la réalité sans complaisance.”

La fièvre, page 8.

Les premières tentatives de classification des différents modes d’expression littéraire remontent, dans notre tradition culturelle, à la *Poétique* d’Aristote, sinon à *La République* de Platon.

Aujourd’hui encore la notion de genre, et notamment de genre littéraire, est utilisée pour classer les textes et en distinguer les différents types. Ainsi, dans les bibliothèques, regroupe-t-on les romans, les essais, la poésie, le théâtre ou encore les ouvrages d’histoire, de philosophie, etc. ; et au sein de chacun de ces groupes on peut distinguer encore divers types d’énonciation (poésie lyrique *vs.* poésie épique) ou divers sous-genres (comédie, tragédie, drame romantique, etc.).

Pourtant, cette institution première du code littéraire, formalisée, qu’est le code générique, a depuis souvent été discutée et mise en question. Les théoriciens la considèrent avec réserve, affirmant que chaque genre littéraire en englobe plusieurs : la nouvelle peut se présenter en effet sous forme de fable, de lettre, de poème en prose etc. Les hésitations terminologiques (nouvelle, conte, récit...) manifestent ce caractère “d’appartenance multiple et emboîtante” de tout écrit littéraire. Les genres ne sont pas imperméables, ils peuvent se partager des intertextes (dictons, métaphores, comparaisons, etc.) ou bien emprunter des éléments extra-littéraires, transférés de domaines extérieurs à la littérature (médecine, technique, administration publique, etc.) vers l’œuvre littéraire, comme

nous pouvons si souvent l'observer dans l'écriture leclézienne.

“Un des aspects souvent négligés de la problématique des genres littéraires, écrit Jean-Marie Schaeffer⁶³, réside dans le fait que ce que nous désignons d'un terme unique (genre) se réfère à des régularités textuelles qui peuvent avoir des statuts les plus divers... Enfin – et c'est sans doute le cas le plus fréquent – on peut se trouver face à un composé variable de plusieurs de ces facteurs.”

Gérard Genette s'intéresse au problème dans *Introduction à l'architexte*⁶⁴ où il montre que l'objet de la poétique n'est pas le texte mais l'architexte, c'est-à-dire l'ensemble des catégories générales ou transcendantes - types de discours, modes d'énonciation, genres littéraires - dont relève chaque texte singulier. Genette remet en cause la tradition générique occidentale et son système unifié qui depuis l'antiquité tente de recouvrir l'ensemble du champ littéraire. Son ouvrage *Palimpsestes*⁶⁵ cherche à fonder en théorie les différentes formes de la relation par laquelle l'œuvre littéraire peut se construire.

En effet, la codification des genres, on l'a déjà vu dans notre présentation du corpus, n'est pas chose aisée ni stabilisée. Le système traditionnel nous propose – ou nous impose – selon le code générique institutionnel, certaines classifications reconnues qui sont richement représentées dans notre corpus ; romans, notamment le sous-genre du “nouveau roman”, nouvelles, essais, etc.

Pourtant, quand il s'agit de repérer les facteurs génériques qui inscrivent le texte dans un code social de poétique, rentre également en compte la fonction de la réception du lecteur et d'autres facteurs de subjectivité qui rend cette codification discutable. Il convient ici, non seulement de laisser la parole aux spécialistes en

⁶³ J.-M. Schaeffer, “De deux facteurs institutionnels de la différenciation générique”, in F. Rastier (éd.) *Textes & Sens*, p.49.

⁶⁴ G. Genette (1979).

⁶⁵ G. Genette (1982).

linguistique qui se sont penchés sur cette question mais également d'entendre la voix de l'écrivain lui-même qui s'est beaucoup intéressé à ce problème.

1.3.1. La question du genre romanesque.

Le genre romanesque est le genre le mieux représenté dans notre corpus, et celui auquel nous prêterons une attention plus particulière.

“Le roman serait ainsi considéré, écrit Bernard Valette⁶⁶ comme l'aboutissement d'une (ou plusieurs) forme(s) narrative(s) en usage dès l'Antiquité. Cette conception s'accompagne d'une adhésion au moins implicite à une théorie évolutionniste : les genres naissent, se transforment et disparaissent – ou se perpétuent. Mais il est toujours loisible de s'affronter en vaines polémiques pour savoir si la perfection du genre se situe à son origine, du côté de sa création, ou lorsqu'il est à son apogée, au moment de sa maturité.”

En effet, les théories sur le roman et ses sous-catégories abondent : ce sont pour la plupart des romanciers ou des critiques qui réfléchissent sur leur art et justifient leur position idéologique ou leur technique littéraire selon différentes méthodes, en rappelant qu'il n'existe pas de frontière stricte entre les différents types de récits.

Pour citer un exemple qui touche l'œuvre de Le Clézio, un critique contemporain, Hans Boll-Johansen⁶⁷, s'est efforcé de mettre en parallèle les distinctions fondamentales entre le roman et la nouvelle. Il oppose les personnages, l'action, l'espace et le temps dans ces deux types de récits. Chacun des niveaux observés révèle selon lui une spécificité de la nouvelle : unicité de l'événement, rapidité de

⁶⁶ B. Valette (1992) : p. 14.

⁶⁷ Cf. D. Grojnowski (2000) : p. 11-12.

l'action, fixité des personnages, économie référentielle. Le tableau synoptique qu'il propose résume sa classification :

	Forces	Temps	Personnages	Espace
Roman	plusieurs	lent	- nombreux - développement psychologique	plusieurs lieux
Nouvelle	un seul couple	rapide	- peu nombreux - constance psychologique	un seul lieu

Tableau n°4 : Classification Roman – Nouvelle.

Traditionnellement, en effet, toute écriture est codifiée. Elle s'impose des formes fixes ou des contraintes techniques ; l'inspiration elle-même, le choix des sujets, la manière de les développer obéissent à des règles précises. Quant au roman, plus que la nouvelle – dont la taille réduite impose certaines contraintes - il semble jouir d'une totale liberté. Objet sémiotique complexe et hétérogène, il paraît insaisissable et se rebelle face à toute tentative de définition.

A propos de cette hétérogénéité Marguerite Yourcenar écrit dans *Carnets des notes des Mémoires d'Hadrien* : "Le roman dévore aujourd'hui toutes les formes ; on est à peu près forcé d'en passer par lui."

Ainsi ont effectivement été publiés sous le titre "roman", aussi bien des romans de gare que des œuvres où la notion d'intrigue devient imprécise, où les nœuds événementiels se font rares.

Un roman apparaît aujourd'hui comme une œuvre largement polysémique. Écrire un roman signifie peut-être simplement, aux yeux du lecteur moderne, "raconter des histoires", avec tout ce que cela implique de distanciation morale et de fascination esthétique. Bernard Valette emprunte le concept *d'œuvre ouverte* à

Umberto Eco “pour désigner un type d’énoncé susceptible de véhiculer des significations différentes en fonction des diverses situations pragmatiques”⁶⁸. L’œuvre ouverte est polyvalente et donne à l’écrivain une grande liberté d’expression. Elle transcende l’époque et le cadre de la société où elle a été produite et peut s’investir de significations nouvelles, parfois totalement opposées au dessein primitif de son créateur.

Cette ouverture introduit cependant une certaine ambiguïté. D’un côté l’écrivain est totalement libre de construire son ouvrage comme il l’entend, mais d’un autre côté, à moins de s’aliéner volontairement le public dans un défi avant-gardiste, l’écrivain est censé s’inscrire dans le code culturel de ses lecteurs potentiels.

Les débuts littéraires de Le Clézio ont lieu dans une époque où les traditions littéraires sont mises en question et où advient une véritable révolution, celle du “nouveau roman” dont les articles à la fois polémiques et fondateurs seront recueillis dans *L’Ère du soupçon*, de Nathalie Sarraute (1956), *Pour un Nouveau roman*, d’Alain Robbe-Grillet (1963), et *Essais sur le roman*, de Michel Butor (1964).

Les genres littéraires, qui sont classés dans les anthologies (avec des filiations, des sous-espèces), sont ici mis en cause. L’utilisation chez Butor par exemple des règles d’un genre littéraire pour faire travailler un autre genre, l’introduction par exemple des modes propres à la poésie dans le récit de voyage, a d’abord pour but de contester notre usage routinier de cadres de pensée. Elle nous fait percevoir ce qu’il y a d’hétérogène en un objet que nous réduisons à l’homogène. Cette façon de modifier les règles établies d’un genre – en l’occurrence le récit de voyage – apparaît très clairement déjà dans *Mobile*⁶⁹.

Les emprunts à une réalité contingente - des fragments de texte sans rapport avec le sujet du livre comme chez Le Clézio - ou à des genres voisins sont nombreux

⁶⁸ B. Valette (1992) : p. 28.

⁶⁹ M. Butor (1962).

dans le “nouveau roman” : théâtre, lettres, journal intime, articles de journaux, télégrammes et message radio etc., tout est possible. Pourtant, si le roman échappe à quelque contrainte génétique que ce soit, il ne relève pas pour autant du simple caprice des écrivains ni d’une inspiration aléatoire. Un ensemble d’habitudes culturelles, de règles tacites, de conventions littéraires conditionnent les auteurs de cette école.

Dans son ouvrage sur le “nouveau roman” Jean Ricardou fait une étude détaillée⁷⁰ de cet ensemble que l’on pourrait définir soit comme flou soit comme net.

“L’ensemble net est trop net, écrit-il⁷¹. Groupe ou école, il se prétend d’une certaine homogénéité et la conforte par des manifestes, des revues, de recueils collectifs, quelquefois une façon de chef.... L’ensemble flou est trop flou. Songeons au Nouveau Roman, ce n’est pas un groupe sûr, ni une école certaine. L’impression de ses contours suscite alors des oscillations prévisibles : maints critiques se sont sentis en effet autorisés par le vague des limites à considérer chaque fois l’ensemble qui convenait le mieux à leurs desseins. C’est à l’imprécision d’une foule de listes officieuses variées qu’on a bientôt affaire.”

Le point commun le plus important dans l’écriture qui caractérisait l’école du “nouveau roman”, malgré le caractère hétéroclite de ce groupement, semble être justement ce refus des genres littéraires préétablis. Pourtant, malgré cette attitude novatrice et ambitieuse qui consiste à brouiller les genres et à remettre en cause les conventions, les créations littéraires de ces auteurs sont presque toujours appelées “romans”. Ce sont ces derniers “romans” qui ont été responsables de ce qu’on a appelé la “mort du roman”.

⁷⁰ J. Ricardou *Le Nouveau Roman*, Paris, Seuil, Points essais, (1973) 1990.

⁷¹ J. Ricardou (1990) : p. 19-20.

“Au-delà des préoccupations scientifiques, écrit François Rastier⁷², la réflexion contemporaine sur les genres paraît influencée par l’esthétique littéraire moderne. Rien de surprenant à cela, puisque les théories des genres prennent traditionnellement la littérature comme objet primordial. Or, l’esthétique moderne, affadissant la pensée romantique, estime qu’un texte littéraire est avant tout l’expression d’une antériorité personnelle, ou, pour des avant-gardistes, d’une subjectivité désirante. ”

Les écrivains de cette époque ne sont pas les seuls à dénier droit de cité aux facteurs sociaux, que reflètent les genres, bien d’autres artistes estiment alors que la création authentique passe par la destruction des genres⁷³. Les bons textes doivent être inclassables et les textes exemplaires, “révolutionnaires” ou “de rupture”, ne renvoient à aucun genre, sinon pour le parodier.

Julia Kristeva adhère avec *La révolution du langage poétique* à ces réflexions et pousse encore plus loin l’idée sociale : puisque les genres reflètent l’incidence de facteurs sociaux, les “subvertir” révolutionnerait la société.

Il s’agit en fait, pour les contemporains, de se libérer du poids écrasant que le temps et la renommée ont fini par conférer à ce qui était, chez les romanciers du siècle précédant, leur recherche de la vérité. La littérature française participe à cette remise en cause de la tradition par la voix de ces auteurs, qui vont faire partie de ce que l’on a appelé le “nouveau roman”, mais qui ne s’est jamais voulu réellement école.

Le jeune Le Clézio est influencé par ces pensées et s’intéresse aux différentes théories sans pour autant adhérer à aucun mouvement ou groupement quel qu’il soit.

⁷² F. Rastier (1992) : p. 43.

⁷³ Déjà Maupassant, dans la préface de *Pierre et Jean*, était pour l’abolition des genres littéraires.

1.3.2. J.M.G. Le Clézio et les genres littéraires.

“Les formes que prend l’écriture, les genres qu’elle adopte ne sont pas tellement intéressants. une seule chose compte pour moi : c’est l’acte d’écrire.”

L’extase matérielle, page 106-107.

Le Clézio, héritier du “nouveau roman” pour certains, considéré par d’autres comme représentant à part entière de ce même “nouveau roman”, ne s’est, lui-même, jamais reconnu dans quelque groupe que ce soit.

“Tout ce qu’a écrit Le Clézio, écrit Michelle Labbé⁷⁴, du moins jusqu’à *Désert*, contient le roman de sa lutte contre le roman, sa quête de l’écriture, la grande histoire d’amour de l’œuvre. Il ne propose pas de théorie structurée sur la création romanesque ni de critique en forme sur le roman dit “traditionnel”, comme ont pu le faire ses contemporains N. Sarraute, A. Robbe-Grillet, J. Ricardou ou Ph. Sollers, mais plutôt des réflexions fréquentes, récurrentes et dispersées.”

On trouve en effet ses réflexions sur la littérature dans toute l’œuvre leclézienne, aussi bien dans les articles et les essais que dans les préfaces, dans les nouvelles, dans les romans et même dans les épigraphes aux chapitres de romans.

Teresa di Scanno rappelle la discussion qui suivit l’exposé de Gerda Zeltner au colloque de Strasbourg sur le roman contemporain en 1971⁷⁵, quand Jean Ricardou a fait la déclaration suivante :

“Il y a le romancier Le Clézio, il y a l’essayiste Le Clézio, et l’on a montré un certain irrationalisme chez lui. Cela conduit, dans *L’extase*

⁷⁴ M. Labbé (1999) : p. 17.

⁷⁵ T. di Scanno (1983) : p. 123.

matérielle, à une réussite beaucoup moins évidente que dans le roman. Cela conduit Le Clézio, j'ai le regret de le dire, mais je le dirai d'autant plus clairement que ses romans sont d'un haut intérêt, cela le conduit à une philosophie de... boy-scout, qui regarde vers... l'Orient.”

L'intérêt dans cette citation n'est point le jugement curieusement classique, formaliste, traditionaliste de Jean Ricardou qui surprend, mais la division qu'il fait entre le romancier et l'essayiste Le Clézio.

“En réalité, écrit Mme di Scanno, non seulement une telle distinction est insoutenable théoriquement dans l'état actuel de la littérature, où, s'il y a indistinction, c'est bien entre le roman et l'essai, mais encore il est apparu assez clairement au cours des présentes études que, chez Le Clézio, il est impossible de séparer le narrateur du penseur : jamais Le Clézio ne fait de “l'art pour l'art”, jamais il n'écrit un roman ou un récit pour le seul plaisir de raconter ou en vue de quelque recherche formelle, soit dans le procédé expressif, soit dans la structure narrative.”

Le jeune Le Clézio s'intéresse à tout le procédé de la création littéraire et ses idées se traduisent souvent par un refus de certaines normes littéraires et se présentent comme une contestation sociale⁷⁶ :

“Pourquoi est-ce que je ne donne jamais les noms des lieux, ou des hommes ? De quoi ai-je peur ? Le système, l'horrible système est là qui me guette. Il veut me faire agenouiller, ou lever le poing. Il veut m'apprendre à posséder les maisons et les voitures, sans parler des femmes.”

⁷⁶ *Le livre des fuites*, p. 172.

“Accepter les conventions du roman, écrit Michelle Labbé⁷⁷, ou tout autre type d’écriture, qui assoit son effet du réel sur les noms propres, présenterait donc le risque d’enfermer dans un système sociopolitique qui a érigé en principe et en finalité l’ascension à la propriété. Nommer n’est plus créer mais posséder.” Le cloisonnement conventionnel des genres dérange le jeune écrivain au plus profond et il écrit⁷⁸ :

“Romans psychologiques, romans d’amour, romans de cape et d’épée, romans réalistes, romans-fleuves, romans satiriques, romans policiers, romans d’anticipation, nouveaux romans, romans-poèmes, romans-essais, romans-romans ! Tous, faits pour les humains, connaissant leurs défauts, flattant leurs lâchetés, ronronnent tranquillement avec eux.”

Dans ses premiers livres, Le Clézio tentait des expériences en transgressant les catégories et les genres, ni essais, ni romans, ni poèmes, et pourtant tout cela à la fois. Dans ces textes des voix d’origine indéfinie s’entremêlent qui donnent aux textes un caractère souvent perçu comme déroutant.

“Échapper au règne du bégaiement de la société moderne, écrit Jean-Xavier Ridon⁷⁹ pour ne plus être l’“écho” d’autres paroles prononcées devant eux. On pourrait alors objecter que le style serait le moyen par lequel l’écrivain surpasserait cette présence de l’autre en construisant un mode d’écriture qui lui est propre. Là où il pourrait trouver ce qui est irréductible à aucun autre.”

Pourtant, le style est historique et social autant que “personnel” comme nous le montre Barthes dans son *Degré zéro de l’écriture*⁸⁰. C’est pourquoi le Clézio dénonce

⁷⁷ M. Labbé (1999) : p. 18.

⁷⁸ *Le livre des fuites*, p. 56.

⁷⁹ J.-X. Ridon (1995) : p. 30.

⁸⁰ R. Barthes (1972) : chapitre 1.

le style comme étant lui aussi un artifice. C'est une autre forme de mensonge qui n'est pas en mesure d'instaurer une différence. Le style est pour l'écrivain une façon de se piéger lui-même en lui donnant l'illusion d'une liberté de création dans un langage qu'il réinventerait :

“Le style est gouverné par des critères esthétiques, continue Jean-Xavier Ridon⁸¹, qui reflètent les conditions historiques et sociales dans lesquelles une écriture s'élabore et qui permet une classification des différents styles en genres littéraires. Mais si le style est un piège, toute considération de “genre littéraire” devient à son tour problématique car elle porte avec elle la même puissance de reproduction et de répétition. Toute définition d'un genre littéraire nous plonge encore dans l'espace du semblable. ”

Le Clézio s'interroge alors sur la possibilité d'écrire de sorte que son écriture échappe à toute classification. Est-il possible d'écrire de façon que cette écriture ne corresponde à aucun genre préétabli par l'héritage littéraire et critique ? Peut-on écrire dans un espace où il ne puisse plus y avoir de jugement de valeur issu d'une tradition littéraire ? Dans *Le livre des fuites* il écrit⁸² :

“Comment échapper au roman ?
Comment échapper au langage ?
Comment échapper, ne fût-ce qu'une fois, ne fût-ce qu'au
mot COUTEAU ?”

Sa rupture avec le contrat romanesque traditionnel s'annonce d'emblée dès les premières pages du premier roman *Le procès-verbal*⁸³ : “J'aimerais que mon récit soit pris dans le sens d'une fiction totale, dont le seul intérêt serait une certaine répercussion (même éphémère) dans l'esprit de celui qui le lit. Genre de

⁸¹ J.-X. Ridon (1995) : p. 31.

⁸² *Le livre des fuites*, p. 13.

⁸³ *Le procès-verbal*, p. 10.

phénomène familier aux amateurs de littérature policière, etc. C'est ce qu'on pourrait appeler à la rigueur le Roman-Jeu, ou le Roman-puzzle."

Dans la préface de *La fièvre*⁸⁴ il refuse également toutes les étiquettes de genre, parmi elles le terme "récit" ; ces neuf histoires sont les produits, annonce-t-il, de "l'écriture pure".

Ces premiers textes ne se conforment guère aux cloisonnements traditionnels, portés par un flot de langage qui passe de la prose au poème, se désarticulant par moments pour faire place à des mots en vrac, des dessins, des schémas, des chiffres, des panneaux publicitaires etc. "En ce qui concerne le genre, je continue de penser que ni le roman, ni l'essai n'existent vraiment" déclare Le Clézio dans un entretien en 1967⁸⁵.

Jean-Marie Gustave Le Clézio se révèle comme un écrivain de ruptures. A son écriture expérimentale initiale, qui avait permis à maints critiques de le rapprocher du "nouveau roman", succède, comme l'exprime Miriam Stendal Boulos⁸⁶, "une marginalité générique délibérément choisie pour dire que la rencontre entre l'être et le cosmos relève plus du discours poétique que du romanesque".

En effet, à partir de *Désert* (1980) l'écriture de Le Clézio s'apaise et s'assagit. Désormais elle se plie au récit et à une forme souple et traditionnelle vers une écriture plus conventionnelle. Néanmoins, le cloisonnement entre les catégories de genres conventionnels continue à éclater facilement, notamment dans les textes qui sont parus dans une édition pour enfants, et ce phénomène semble vraiment être une constante chez notre auteur.

⁸⁴ Cf. citation p. 42.

⁸⁵ R. Borderie, "Entretien avec J.M.G. Le Clézio", in *Lettres françaises* n°1180, 27 avril 1967, p.11-12.

⁸⁶ M. Stendal Boulos (1999) : p. 5.

Il se soucie toujours peu des genres littéraires, qui sans doute “existent”, mais qui n’ont selon lui “aucune importance⁸⁷” :

“Les formes que prend l’écriture, les genres qu’elle adopte ne sont pas tellement intéressants. [...] Les structures des genres sont faibles. Elles éclatent facilement.”

Le Clézio admet cependant que “le problème du genre” prend nécessairement une certaine importance, aussi “parce que, pour trop de gens, il y a un snobisme du genre”⁸⁸.

Pour Le Clézio ce que l’écriture recherche est cet “avant” des mots dans lequel le monde se donne en dehors du signe pour le dire et qui serait aussi un après des mots. L’origine n’est plus seulement le point de départ, c’est aussi le point d’arrivée.

Lorsque Gérard de Costanze, dans une interview récente⁸⁹, aborde de nouveau le problème des genres, il lui pose la question suivante : “Michaux avait brisé les genres littéraires. La nouvelle, la poésie, le roman. Ces notions ont un sens pour vous ? Pour vous, auteur ?”

“Il n’est pas, répond Le Clézio, d’une importance extrême de définir ce que c’est qu’un roman ni ce que c’est une nouvelle. Il s’agit simplement d’une question de rythme. Quand vous commencez certains livres, vous avez un rythme qui vous guide vers ce qui va être un roman, c’est-à-dire vers une œuvre qui est plus musicale peut-être que dans le cas de la nouvelle. Pour d’autres, vous vous rendez compte que cela s’apparente davantage au faits divers. Il pourrait presque s’agir d’une rubrique de journal, mais vous ne pouvez pas

⁸⁷ *L’extase matérielle*, p. 74.

⁸⁸ *Ibid.*, p. 74.

⁸⁹ *Magazine littéraire*, n° 362, février 1998, p. 34.

dire vraiment ce qui vous a conduit. Peut-être une certaine manière, votre disposition du moment... Je ne sais pas...”

L'écriture littéraire peut-elle vraiment tourner le dos aux genres ? Le procès d'écriture est générique, énonce Robert Scholes, en ce sens que tout écrivain conçoit sa tâche en fonction de sa tradition passant sous silence l'épineux problème du statut de la fiction. Un des objectifs de notre travail sera aussi d'évaluer le poids de cette tradition, de vérifier si, à travers les données objectives et quantifiables du corpus, transparait plutôt le maintien des frontières génériques ou plutôt leur subversion.

Toutefois, tout discours peut se définir, qu'on le veuille ou non, par des caractéristiques communes qui définissent des typologies diverses de textes. Ainsi, la linguistique d'aujourd'hui nous propose-t-elle de reprendre cette question sans pour autant forcément adhérer aux institutions traditionnelles des genres littéraires.

1.3.3. Les typologies de textes.

“Il faut reconnaître, écrit François Rastier⁹⁰, d’une part qu’il n’existe pas de texte sans *genre*, et en outre que tout genre relève d’un discours (ex. politique, juridique, religieux, etc.). [...] Même les échanges linguistiques qui paraissent les plus spontanés sont réglés par les pratiques sociales dans lesquelles ils prennent place, et relèvent donc d’un discours ou d’un genre. [...] Comme les textes ne sont pas gouvernés par des règles (au sens trop fort du terme qui est d’usage en linguistique), mais par des normes, la description des genres est un objectif pour la typologie des textes, et aussi un enjeu important pour la linguistique. Elle permet en effet d’articuler les connaissances sur la structure de la langue avec les observations sur la structure des textes particuliers. Les linguistiques textuelles dont nous disposons ont sous-estimé ces problèmes : la plupart sont universelles, et se révèlent trop puissantes en théorie pour être utiles en pratique. Il importe donc de discuter des critères de typologie.”

La plupart des typologies dont dispose la langue transcendent la division en discours et en genres. Chacune de leurs classes peut inclure plusieurs discours, ou du moins des genres relevant de plusieurs discours, la plupart étant du domaine de l’anthropologie philosophique, et énumèrent les différents types d’usage que l’homme peut faire du langage. Ces typologies relèvent d’une grande stabilité, et le caractère traditionnel des types est toujours présents.

“Il n’existe pas de texte (ni même d’énoncé), écrit François Rastier dans *Sens et textualité*⁹¹, qui puisse être produit par le seul système fonctionnel de la langue (un sens restreint de mise en linguistique).

⁹⁰ F. Rastier, “La macrosémantique”, in F. Rastier, M. Cavazza, A. Abeillé (1994) : p. 174-175.

⁹¹ F. Rastier (1992) : p. 37.

En d'autres termes, la langue n'est jamais le seul système sémiotique à l'œuvre dans une suite linguistique, car d'autres codifications sociales, le genre notamment, sont à l'œuvre dans toute communication verbale."

Georges Molinié ajoute à la notion de genres⁹² l'existence de trois types, trois classes de littérarité. Ces trois types ne correspondent pas à des types de discours littéraires, mais à trois types de questionnements sur quelque discours littéraire que ce soit, donc à trois niveaux d'approche de l'objet : la littérarité générale, la littérarité générique et la littérarité singulière.

La littérarité générale, c'est la littérarité comme telle : un discours est ou n'est pas littéraire. On oppose par exemple un discours littéraire à un discours technique, administratif ou scientifique.

La littérarité générique est celle des genres. On en connaît l'importance : une pièce de théâtre n'est pas un roman et en subdivisant les catégories, une tragédie n'est pas une comédie.

"Il semble d'ailleurs, écrit-il⁹³, curieusement, plus facile de démonter et de décrire les contraintes stylistiques caractéristiques de chacun de ces genres et sous-genres, contraintes dont il est possible de construire des systèmes, des systèmes de systèmes et des emboîtements systémiques partiels ! On arrivera justement ainsi à décrire des esthétiques de mélanges, non moins rigoureusement démontables et explicables."

L'étude des "esthétiques de mélanges" serait selon Molinié le seul moyen de dépasser l'impasse dans laquelle ce type d'étude risque de buter, et d'envisager

⁹² G. Molinié, A. Viala (1993) : p. 13-15.

⁹³ *Ibid.*, p. 14.

paradoxalement sous cet angle les pratiques littéraires qui justement ne paraissent pas justiciables de la stylistique des genres, c'est-à-dire une bonne partie de la littérature actuelle. "L'indéfinition générique, voire le refus affiché de la catégorisation générique, relève encore de l'examen de la littérarité générique" dit-il⁹⁴.

La littérarité singulière enfin est l'analyse de la manière individuelle de chaque production littéraire singulière : c'est le domaine des études de style consacrées à tel ou tel auteur. C'est la littérarité qui paraît avoir été de tout temps la mieux étudiée.

D'autres classifications sont également possibles, Greimas propose par exemple une répartition entre texte "pratique" et texte "mythique", d'origine lévi-straussienne⁹⁵, qui justifie la distinction entre les textes scientifiques et les textes axiologiques, comme par exemple les descriptions littéraires. Cette opposition serait intéressante à observer dans l'écriture leclézienne où ces deux classifications emboîteraient les classifications plus traditionnelles. Toutefois, les textes qui décrivent des pratiques mythiques (les rituels et les coutumes des cultures amérindiennes dans les ouvrages ethnologiques de Le Clézio, par exemple), sont par leur précision et leur organisation séquentielle bien plus proches des textes techniques. Dans bien des sociétés, notamment la société amérindienne, la distinction entre la pratique et le mythique semble aussi difficile à établir. Peut-être une telle distinction ne reflète-t-elle qu'une exigence de notre société occidentale ?

"Il m'apparaît, écrit J.-M. Schaeffer⁹⁶, que la problématique des genres nous force à prendre en compte la dimension institutionnelle de la littérature : s'il est vrai que le destin des traditions génériques est *inscrit dans* les textes, on ne saurait guère nier qu'en revanche il se

⁹⁴ G. Molinié, A. Viala (1993) : p. 14.

⁹⁵ Cf. F. Rastier, "La macrosémantique", in F. Rastier, M. Cavazza, A. Abbeilé (1994) : p. 176.

⁹⁶ J.-M. Schaeffer "De deux facteurs institutionnels de la différenciation générique", in F. Rastier (éd.) *Textes & Sens*, p. 50.

décide entre les textes, dans le passage d'une œuvre à l'autre – donc dans la tête des écrivains, et par-là même [...] dans l'espace public de la littérature comme pratique créative socialisée.”

La connaissance des genres resterait donc indispensable pour interpréter les textes, même littéraires et d'avant-garde comme le sont souvent les textes de Le Clézio. Elle permet de formuler des critères de plausibilité des lectures, de contribuer à fixer la référence, fictionnelle ou non, des textes et en même temps d'accepter l'hétérogénéité comme un facteur irréductible, pour tenter ensuite de la comprendre.

“Or, écrit encore J.-M. Schaeffer⁹⁷, l'importance changeante accordée à tel ou tel des mécanismes de régulation générique par les diverses traditions littéraires dessine des types de généricité et d'évolution générique fort différents les uns des autres, qui à leur tour nous mettent en prise directe avec la diversité des traditions littéraires selon les époques, les pays ou les aires culturelles.”

La typologie des genres textuels paraît donc essentielle pour toute activité d'analyse textuelle car la connaissance des genres reste indispensable pour interpréter les textes, même littéraires et d'avant-garde.

Enfin, les critères d'une typologie des textes se divisent, selon François Rastier, en trois groupes : les critères anthropologiques, les critères philosophiques, et les critères sémantiques. La technique de la lexicométrie et l'analyse des corpus que nous allons mettre en œuvre nous permettront d'aborder ces trois critères et elles ouvriront aussi la voie à une approche de classification purement endogène, fondée sur des traits linguistiques propres aux textes étudiés, aisément objectivables et dénombrables par l'outil informatique.

⁹⁷ J.-M. Schaeffer “De deux facteurs institutionnels de la différenciation générique”, in F. Rastier (éd.) *Textes & Sens*, p. 49.

1.3.4. La lexicométrie et les genres.

Les études lexicométriques et l'analyse du corpus en situation montre en effet que le lexique, la morphosyntaxe, la structure et la longueur des phrases, entre autres, varient avec les genres.

L'opposition entre les différents genres est toujours présente et souvent même prépondérante dans les différentes analyses statistiques. Étienne Brunet aborde ainsi le phénomène à propos de son étude sur le vocabulaire de Victor Hugo⁹⁸ :

“..... le relief des genres littéraires vient sans cesse contrarier la pente chronologique qui conduit les textes de l'amont à l'aval. Dans les nombreuses analyses factorielles que nous avons réalisées, relativement aux catégories grammaticales, à l'emploi des mots outils, au rythme de la phrase, à la structure lexicale, nous avons toujours rencontré les massifs des genres, avant de découvrir, serpentant parmi les monts, le cours du temps. [...] C'est que le genre préexiste à l'œuvre et à l'écrivain, avec ses codes et ses lois qui s'opposent aux règles des autres genres. Et l'opposition entre les genres l'emporte généralement sur toutes les autres, écrasant la perspective chronologique et même effaçant les différences des tempéraments, au point que deux textes varient davantage s'ils sont écrits par le même écrivain dans deux genres différents que s'ils partagent le même genre sans partager le même auteur.”

Il a noté chez Giraudoux que le vocabulaire, la syntaxe et le rythme de la phrase différaient selon qu'il s'agit d'un roman ou d'une pièce de théâtre, que chez Hugo la poésie s'opposait à la prose, le roman au théâtre et la fiction à la correspondance. Gunnel Engwall a même pu constater dans son étude des

⁹⁸ É. Brunet (1988) : p. 398.

romans contemporains⁹⁹ que le genre du texte peut jouer un plus grand rôle pour la proportion des classes de mots dans un texte que la langue elle-même.

Cette opposition générique dans les analyses de statistique lexicale est si forte que, selon Étienne Brunet, elle empêcherait de fonder de grands espoirs sur les méthodes quantitatives pour attribuer un texte à un écrivain plutôt qu'à un autre¹⁰⁰ :

“La démarche se justifierait si les écarts entre les genres étaient moindres que les différences qui séparent l'écriture des deux écrivains. Or c'est le contraire qu'on observe. Les distances *intra* s'avère souvent plus importantes que les distances *inter*.”

Dans notre corpus aussi cette opposition s'est avérée bien souvent un facteur prépondérant dans les différentes analyses. Le tableau suivant, représentant l'analyse factorielle du dictionnaire entier de notre corpus A – en donnant en même temps un avant-goût des méthodes que nous allons aborder dans le prochain chapitre - montre bien l'opposition générique qui opère à tout niveau :

⁹⁹ G. Engwall (1983) : p. 101.

¹⁰⁰ É. Brunet (1988) : p. 398.

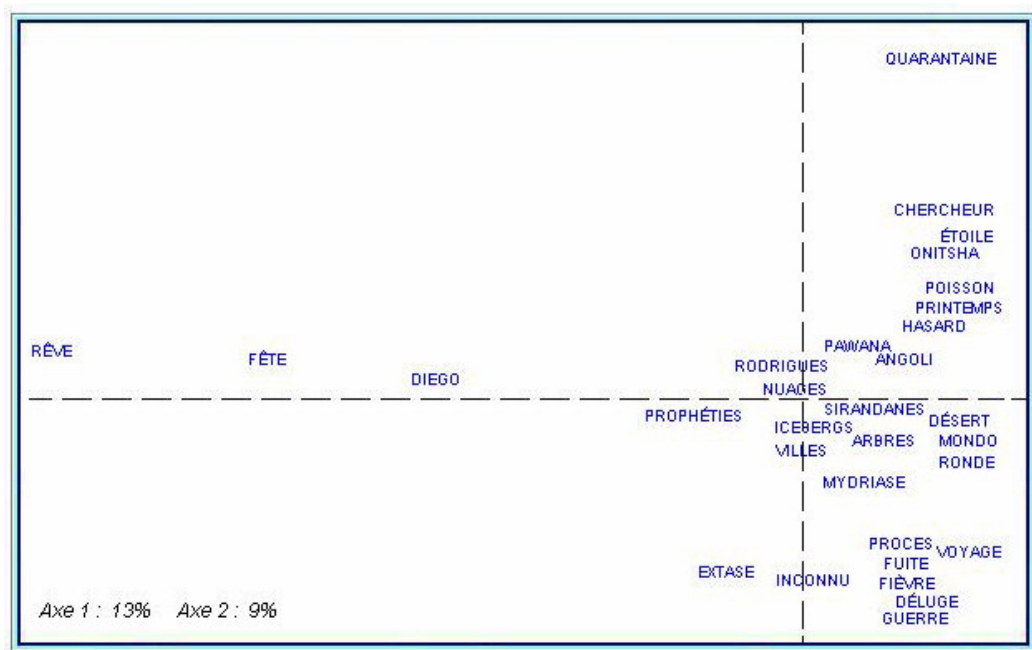


Figure n°8 : Analyse factorielle du dictionnaire du corpus A.

Nous trouvons effectivement des oppositions très fortes dans ce tableau. Les ouvrages de vocation ethnologique *La fête chantée*, *Le rêve mexicain*, *Les prophéties de Chilam Balam* (traitant du monde amérindien) et la biographie sur les peintres mexicains *Diego et Frida* s’opposent de façon significative des autres. Les autres essais, *Trois villes saintes* et *L’inconnu sur la terre*, se trouvent au milieu sur l’axe vertical près du récit de voyage *Gens des nuages*, et le seul ouvrage du corpus écrit sous forme de journal personnel : *Voyage à Rodrigues* ainsi que les récits poétiques *Vers les icebergs* et *Mydriase*. Sur la partie droite du graphique est rassemblé l’ensemble du genre romanesque. Il est à noter qu’il y a chez notre écrivain une division nette entre la première période de la production, caractérisée par une écriture expérimentale et influencée par l’école du “nouveau roman” (*Le procès-verbal*, *La fièvre*, *Le déluge*, *La guerre*, *Le livre des fuites* et *Voyages de l’autre côté* regroupés en bas du graphique) et les autres romans et recueils de nouvelles. Cette opposition prépondérante nous a amenée à traiter le groupe comme une entité à part et nous l’appelons dans cette étude, par commodité, le genre du “nouveau roman”, bien que l’appartenance de l’écrivain à cette école soit contestable.

Ces divisions génériques - ou typologiques - des textes, sont toujours sur le plan théorique, discutables, comme nous avons voulu le montrer dans ce chapitre. D'une part des théoriciens ont critiqué le caractère institutionnel et traditionnel qui cloisonne la création littéraire en genres différents, d'autre part le refus de se plier à des règles souvent classées comme "petit-bourgeois" caractérise l'époque des débuts littéraires de notre écrivain. Il est évidemment impossible de ne pas prendre en considération ces critiques lors de l'étude d'une écriture comme celle de J.M.G. Le Clézio qui lui-même a été, est peut-être encore jusqu'à un certain degré, un des plus fervents opposants à ce système préétabli et conventionnel.

Pourtant, d'autres théoriciens, notamment celles de François Rastier, ont montré que les genres existent, qu'on le veuille ou non, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence des différentes typologies de textes. Par ailleurs, dans l'étude lexicométrique l'opposition générique est extrêmement claire et permet de définir des caractéristiques génériques en s'appuyant, non sur des valeurs anthropologiques ou sociales, mais sur les propriétés même des textes.

C'est dans cet esprit que nous tâcherons de mener notre étude de l'écriture leclézienne et d'employer la terminologie des genres, sans pour autant ajouter un jugement de valeur ou d'adhérer à une idéologie sociale quelconque. C'est dans une perspective linguistique et scientifique que l'on doit aborder les genres littéraires, et les techniques et méthodes de la lexicométrie – auxquelles nous consacrerons le chapitre suivant – nous semblent être, par leur caractère impartial, une façon adéquate d'aborder le problème.

1.4. La méthode.

“Rien d’autre, rien d’autre pour moi
que le langage, c’est le seul problème,
ou plutôt la seule réalité.”

L’extase matérielle, page 33.

Nous suivrons principalement dans ce travail consacré à l’analyse du vocabulaire de Le Clézio les méthodes lexicométriques de la statistique lexicale.

La statistique lexicale, que l’on peut appeler lexicologie quantitative ou lexicométrie, traite de façon exhaustive et systématique le vocabulaire d’un corpus. L’analyse est automatisée et porte sur des critères quantifiés. Grâce à l’indexation et aux procédés informatisés, on associe à chacun des mots (occurrence ou vocable) du corpus plusieurs valeurs chiffrées qui permettent l’analyse statistique.

C’est à Charles Muller, de l’université de Strasbourg, que revient le mérite d’avoir, avec Pierre Guiraud, implanté la statistique lexicale en France. Deux volumes présentent en toute rigueur et clarté les techniques devenues classiques qu’il préconise : *Initiation aux méthodes de la statistique linguistique*¹⁰¹ et *Principes et méthodes de statistique lexicale*¹⁰².

Faut-il aujourd’hui expliquer les méthodes et les principes de la lexicométrie ? “Ces méthodes sont désormais connues de tous, a déclaré Charles Muller lors d’une récente soutenance de thèse¹⁰³, et il ne faut pas informer sur ces méthodes, cela serait comparable à expliquer le fonctionnement du sèche-cheveux ou de la montre à quartz”.

¹⁰¹ Ch. Muller, *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette, 1973.

¹⁰² Ch. Muller, *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977.

¹⁰³ décembre 2001.

Suivant son conseil, nous ne nous attarderons que très brièvement sur ces méthodes qui sont, aujourd'hui, depuis les travaux pionniers des années 60, bien acceptées et établies. Nous renvoyons le lecteur, tout au long de cette étude, aux textes des fondateurs de cette branche de recherche qui en expliquent de façon détaillée les principes et les méthodes¹⁰⁴.

Les méthodes de la lexicométrie se fondent pour l'essentiel sur la loi binomiale (tirages aléatoires non exhaustifs) et utilisent les approximations les mieux ajustées aux corpus et à leur partition, aux fréquences et sous-fréquences des vocables (loi normale et écart réduit, distance de Khi2, loi de Poisson, etc.). Jugée sur tests, la comparaison d'effectifs "théoriques" calculés et des observations faites sur les listes de vocabulaires ou de classes obtenues après dépouillement permet de caractériser le style d'un écrivain, la particularité d'une œuvre ou les constantes d'un genre, comme nous l'avons indiqué dans le chapitre précédent.

Pour les traitements et l'exploitation statistique des données de notre corpus nous utiliserons différents logiciels de traitement statistique et hypertextuel, ainsi que différentes techniques de traitement statistique.

1.4.1. Les principes statistiques.

La statistique dans son aspect descriptif est un recueil de renseignements chiffrés sur des domaines variés, comme des populations, des groupes humains ou non, des faits physiques. Dans notre cas il s'agit d'une collection de renseignements chiffrés concernant des mots ou des groupes de mots.

¹⁰⁴ Pour plus de détails sur l'histoire et sur les fondements de cette méthode de recherche, voir en outre P. Guiraud (1954) et É. Brunet (1978) et (1981).

L'analyse lexicométrique repose en grande partie sur le calcul de probabilité et l'écart-type. A son tour l'écart-type permet de passer à l'écart réduit, noté z . Il se calcule à partir de l'écart absolu, qui est la différence entre l'effectif observé (ou la fréquence absolue) et l'effectif théorique (ou la fréquence théorique)¹⁰⁵ :

$$Z = \frac{\text{fréquence absolue} - \text{fréquence théorique}}{\sqrt{\text{fréquence théorique}}}$$

En employant cette technique, on ne considère donc plus les effectifs absolus mais les relatifs. Ainsi la neutralisation des différences de taille entre les œuvres permet-elle de savoir si un phénomène ou un sous-corpus s'écarte significativement de l'ensemble du corpus. Un écart réduit supérieur à 2 (ou inférieur à -2) indique que l'on a moins de 5 chances sur 100 de se tromper en affirmant que deux valeurs sont significativement différentes et qu'il y a (ou qu'il n'y pas) un écart distinctif entre par exemple un roman et l'ensemble de l'œuvre.

Jusqu'ici l'usage a été d'avoir recours à la loi normale, notamment dans les calculs de spécificités et chaque fois que la distribution d'un mot, d'une classe ou d'une catégorie devait être pondérée pour tenir compte de l'inégale étendue des textes comparés. Toute fréquence (ou effectif) observée était rapprochée de la fréquence attendue et convertie en écart réduit selon la formule ci-dessus.

Toutefois, la loi normale est une approximation d'un calcul plus exact qui repose sur la distribution hypergéométrique¹⁰⁶. Pierre Lafon a montré que ce modèle est celui qui convient le mieux à la lexicométrie parce qu'il s'applique à des données discrètes (alors que la loi normale traite de valeurs continues) et qu'il propose un

¹⁰⁵ formule simplifiée, la formule exacte fait intervenir la probabilité q .

¹⁰⁶ Pour des informations plus détaillées sur ce calcul, voir É. Brunet (1978) et (1981).

tirage exhaustif (alors que la loi binomiale s'accommode d'un tirage non exhaustif)¹⁰⁷ :

“...les instruments statistiques habituellement utilisés pour traiter la variabilité de la fréquence des formes, Khi 2, loi normale avec calcul d'écart réduit pour les formes les plus fréquentes, loi de Poisson pour celles qui sont rares, ne constituent jamais que des approximations du calcul adéquat, certes légitimes dans certaines conditions d'emploi, mais qui deviennent scabreuses, voire fausses, dans d'autres. Aussi, pour une application de cette nature, il nous semble préférable d'emprunter à la statistique les formules de la distribution hypergéométrique, qui, bien que lourdes à manier – mais l'ordinateur n'est-il pas là pour soulager ? – sont exactement adaptés à la population discrète des occurrences de vocabulaire, et valides pour toute la gamme des fréquences rencontrées.”

Néanmoins Étienne Brunet¹⁰⁸, ayant reconnu la supériorité du modèle hypergéométrique dans *Le vocabulaire français de 1789 à nos jours*, précise, après divers tests, que le gain en précision et en exactitude n'est pas sensible dans les corpus de grande ampleur comme l'est le nôtre. “Dans les très grands corpus non seulement la loi hypergéométrique est inutile, écrit-il encore en 1980¹⁰⁹, mais elle est ruineuse et impraticable.”

La puissance multipliée des ordinateurs actuels qui permet désormais les lourds calculs exigés par la loi hypergéométrique fait aujourd'hui de l'utilisation de la loi normale ou de la loi hypergéométrique un choix d'école souvent sujet à polémique.

¹⁰⁷ P. Lafon (1984) : p. 46.

¹⁰⁸ É. Brunet (1981) : p. 31-50.

¹⁰⁹ É. Brunet, “Loi hypergéométrique et loi normale. Comparaison dans les grands corpus”, *Actes du 2^{ème} Colloque de lexicologie politique*, Paris, Klincksieck, 1982, vol 3, p. 699-717.

Étant donné les résultats quasi identiques qui ressortent des analyses diverses effectuées sur notre corpus Le Clézio avec l'une et l'autre méthodes de calcul, nous ne rentrerons pas ici dans cette polémique. Nous aurons recours dans la plupart des cas à des calculs s'appuyant sur la loi hypergéométrique, fournis par les logiciels, et lorsque nous aurons recours à la loi normale et aux écarts réduits, par exemple dans les histogrammes, cela sera indiqué dans la légende.

Les résultats de ces calculs seront présentés par une illustration graphique de la distribution, le plus souvent sous forme d'histogramme¹¹⁰ dans lesquels les "bâtons" se répartissent de part et d'autre de la ligne médiane qui représente la valeur 0 de l'écart réduit.

Nous ferons référence pour les comparaisons externes dans cette étude à la base *Frantext* issue du *Trésor de la langue française* et son dictionnaire de fréquence, le plus important de ces dictionnaires – concernant la langue française – (réalisé à Nancy sous la direction de P. Imbs et R. Martin). Dans ce cas, on n'a pas jugé utile¹¹¹ d'avoir recours au modèle hypergéométrique. En effet ce corpus de référence qui dépasse 100 millions d'occurrences permet de bénéficier de la loi des grands nombres. Comme le calcul n'est entrepris que dans la zone sûre où la fréquence théorique est suffisamment élevée, la loi normale offre ici les mêmes garanties et la même précision que le modèle hypergéométrique.

La méthode de χ^2 (khi2) analyse des fréquences absolues dans une table de fréquences. On procède à la comparaison des fréquences vraies avec les fréquences on devrait s'attendre selon l'hypothèse nulle¹¹². Le test de χ^2 sert selon Charles Muller¹¹³ à apprécier en probabilité l'écart constaté entre une observation et un modèle théorique, quel que soit le nombre de variables. "Il est d'une

¹¹⁰ Il s'agit d'histogrammes directement issus du logiciel Hyperbase que nous avons retouchés dans le logiciel Paintshop Pro.

¹¹¹ Le choix s'impose en fait par le logiciel utilisé. Pour plus de détails, cf. É. Brunet (1995) : *Hyperbase*, Manuel de référence, version 5.0, CNRS-INaLF, UPRES "Bases, corpus et langage", et sa mise à jour de janvier 2001.

¹¹² S. Körner, L. Wahlgren (1998) : p. 147.

¹¹³ Ch. Muller (1973) : p.116.

application constante en statistique linguistique aussi bien que dans les autres domaines et son maniement est simple” écrit-il. Si l’on représente par e une valeur calculée ou théorique (ou attendue) et par o une valeur observée ou réelle, la formule générale du test est :

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Dans une étude statistique il s’agit souvent de traiter un grand nombre de variables corrélatives et tout naturellement on se pose, comme le fait Philippe Cibois¹¹⁴, les questions suivantes :

- “Y-a-t-il une possibilité de comprimer toute cette information ? Un nombre inférieur de variables pourrait-il remplacer l’information originale ?
- Quelles sont les facteurs sous-jacents que nous mesurons réellement avec ces grands nombres de données statistiques ?”

Une méthode efficace pour traiter des complexes de variables et des applications multilinéaires est l’analyse factorielle, qui sans réduire l’information traite de grands matériaux statistiques. Le point de départ est que les valeurs des différentes variables étudiées sont en réalité l’expression d’un nombre sous-jacent réduit de dimensions. En créant des combinaisons diverses des variables étudiées on essaie d’identifier ces facteurs.

L’analyse factorielle des correspondances (A.F.C.) est en effet une manière de représenter la réalité sous une forme plus intelligible et lisible. Philippe Cibois explique ainsi le principe¹¹⁵ :

“L’analyse factorielle traite des tableaux de nombres et elle remplace

¹¹⁴ Ph. Cibois (1983) : p. 5.

¹¹⁵ *Ibid.*, p. 5.

un tableau difficile à lire par un tableau plus simple à lire qui soit une bonne approximation de celui-ci.”

L’analyse factorielle permet de soumettre au calcul une série de formes, qui seront traitées ensemble selon les méthodes multidimensionnelles, et conduit à une cartographie signifiante, dont les éléments peuvent être comparés à la cartographie planaire courante, aux fins de détection des principales distorsions induites par l’introduction des facteurs non spatiaux. La cartographie prend, à chaque fois, une valeur significative. Philippe Cibois fournit une explication détaillée dans son ouvrage *L’analyse factorielle*¹¹⁶

Toutefois, il ne saurait être question, dans les limites de ce travail, de fournir une présentation complète de l’analyse factorielle. En tout état de cause, on peut constater à travers la bibliographie que l’A.F.C. est une procédure d’investigation scientifique très représentative des méthodes contemporaines : on ne la présente jamais indépendamment de l’application que l’on propose.

Il convient toutefois de rappeler lors de l’exploitation de cet outil, comme le fait Jean-Marie Viprey¹¹⁷ :

“...une fois de plus, que l’on écarte ici toute illusion positiviste sur la validité ‘en soi’ de cette cartographie. Elle devrait être un outil d’exploration de l’œuvre, reposant sur l’objectivation provisoire de rapports spatiaux paramétrés, et nécessitant (et ne se justifiant que par) le retour au texte.”

Le programme utilisé dans cette étude¹¹⁸ a été fourni par l’association ADDAD, qui distribue un logiciel complet pour l’analyse des données. Le module mis en

¹¹⁶ Ph. Cibois, *L’analyse factorielle*, Paris, PUF, Collection que-sais-je, 1983.

¹¹⁷ J.-M. Viprey (1997) : p. 109.

¹¹⁸ Cf. É. Brunet (1995) : *Hyperbase*, Manuel de référence, version 5.0, CNRS-INaLF, UPRES “Bases, corpus et langage” et sa mise à jour de janvier 2001.

œuvre est celui de l'analyse de correspondance, qui suit l'algorithme proposé par Jean-Paul Benzécri et dont l'adaptation à Windows a été réalisée par André Salem et celle au logiciel Hyperbase par Étienne Brunet.

Certains de nos calculs¹¹⁹ seront fondés sur une autre méthode : l'analyse arborée élaborée par Xuan Luong¹²⁰, que nous expliquerons plus en détail dans le chapitre consacré à la distance lexicale. Il s'agit là aussi d'un mode de représentation des distances extrêmement visuel et dont chaque élément graphique (les groupements, les oppositions, les nœuds et les distances) est signifiant.

Dans l'étude des tableaux lexicaux, l'analyse des correspondances permet avant tout de dégager de grands traits structuraux qui portent sur plusieurs ensembles mis en correspondance. Cependant, lorsque le nombre d'éléments représentés est important, il peut devenir délicat, dans la pratique, d'apprécier leurs positions réciproques au seul vu des résultats graphiques. Nous emploierons donc, dans certaines analyses, conjointement aux méthodes factorielles, la méthode de la classification.

Les regroupements obtenus se font en effet à partir des distances calculées *dans tout l'espace*, et non seulement dans le ou les premiers plans factoriels. "Ils seront donc, selon Ludovic Lebart et André Salem¹²¹, de nature à corriger certaines des déformations inhérentes à la projection sur un espace de faible dimension".

La classification ainsi obtenue peut être représentée de plusieurs manières différentes. La représentation sous forme d'arbre hiérarchique ou *dendrogramme*, à laquelle nous aurons également recours, constitue sans doute la représentation la plus parlante.

¹¹⁹ Plus précisément ceux de la distance lexicale.

¹²⁰ Cette analyse est décrite en détail dans *Analyse arborée des données textuelles*, CUMFID 16, juin 1989.

¹²¹ L. Lebart, A. Salem (1994) : p. 112.

La lexicologie quantitative, qui permet avant tout une description contrôlée et chiffrée du vocabulaire, ouvre la voie à un traitement exact et impartial. Lorsque le corpus est important, comme le nôtre, l'ordinateur, qui est à la source d'une évolution de la lexicométrie en direction d'une sophistication des traitements et d'une macro-statistique de corpus, devient un outil indispensable pour saisir les nuances, les structures et les évolutions d'un ensemble linguistique que l'œil et la mémoire humaine ne sauraient embrasser. L'ordinateur et les logiciels appropriés permettent une description rigoureuse et objective du corpus sur lequel une réflexion et une analyse précise peuvent être faites.

Pour les traitements et l'exploitation statistique des données de notre corpus nous utiliserons différents logiciels de traitement statistique et hypertextuel, ainsi que différentes techniques de calcul. Nous espérons ainsi à la fois enrichir l'étude sur le vocabulaire de Le Clézio et développer une réflexion méthodologique à travers la confrontation des résultats fournis par les différents outils.

1.4.2. Les logiciels.

Notre étude exploite principalement le logiciel Hyperbase, conçu par Étienne Brunet à l'Université de Nice-Sophia Antipolis dans le cadre du laboratoire "Statistique linguistique", puis "Bases, Corpus et Langage". Plusieurs versions sont utilisées, notamment la version 5.1 et la version 5.3 qui s'appuie sur le lemmatiseur Cordial. Ce logiciel permet des traitements exploratoires et statistiques très étendus à partir d'une base de données. Il offre un grand choix de tris et de calculs, souvent accompagnés de graphiques. Nous aurons également recours, pour certaines analyses, au logiciel Lexicométrie¹²² de Dominique Labbé. Ce logiciel, créé au laboratoire de C.E.R.A.T. à Grenoble, comprend un ensemble de programmes qui s'enchaînent et des tables incrémentielles.

¹²² Pour des détails sur le logiciel Lexicométrie voir Picard J., Pibarot A. et Labbé D. (1995) : p. 395-396.

Le logiciel Hyperbase ouvre la voie à des applications diverses. Il permet l'exploitation documentaire et l'obtention des contextes, des concordances et de listes de mots. L'exploration statistique du logiciel donne la possibilité de nombreuses analyses, notamment sur la richesse lexicale, l'étude des hapax, l'accroissement lexical, la distance lexicale, la corrélation chronologique, les spécificités internes et externes, ainsi que la recherche thématique que nous aborderons longuement dans notre étude.

Le programme fournit de nombreux graphiques que nous exploiterons tout au long de cette étude afin d'offrir une visualisation plus aisée des résultats des différentes analyses.

Pour les nombreux traitements que fournit Hyperbase nous renvoyons au manuel du logiciel¹²³ qui décrit de façon précise le fonctionnement ainsi que le large éventail d'analyses disponibles. Nous nous bornons ici à n'en citer que quelques-unes, que le logiciel fournit sous forme de listes :

- l'index alphabétique de chaque sous-ensemble du corpus ;
- le dictionnaire de fréquence alphabétique de tout le corpus ;
- la liste de fréquences par ordre décroissant, qui nous présente tous les vocables ;
- du plus fréquent au moins fréquent ;
- les listes de corrélation chronologique dans le sens de la progression aussi bien que dans le sens de la régression.

Le logiciel Lexicométrie permet une préparation très sophistiquée des données et met l'accent sur la technique de la lemmatisation. Nous reviendrons donc plus en détail sur ses fonctionnements après avoir traité les principes de cette technique dans le prochain chapitre.

¹²³ É. Brunet (1995) : *Hyperbase*, Manuel de référence, version 5.0, CNRS-INaLF, UPRES "Bases, corpus et langage" et sa mise à jour de janvier 2001.

Le recours à ces logiciels demande tout naturellement beaucoup de rigueur dans le traitement des textes et la préparation des données au préalable. L'ordinateur nous offre en fait la possibilité de traiter de façon impartiale et efficace de grandes quantités de données, mais il ne saurait produire de jugements ou de rectifications sur des données préparées de façon inadéquate ; il convient par conséquent de définir préalablement les normes linguistiques employées.

1.4.3. Les normes linguistiques.

“C'est qu'il s'agit bien d'un jeu, le plus cruel, le plus vain des jeux. Concevoir selon les normes de la conception, écrire selon l'écriture, être selon l'être.”

Le livre des fuites, page 237.

Toute étude quantitative du vocabulaire commence par l'indexation des textes que l'on va étudier. Après avoir délimité notre corpus, nous avons procédé à la saisie optique, au moyen du scanner, qui constitue la seule méthode lorsque les textes ne sont pas déjà disponibles et accessibles sous forme numérique, ce qui n'a pas été le cas de notre corpus. Après ce long travail, il a fallu, par la suite, corriger les fautes inévitables dues à la lecture et à la reconnaissance optique, ainsi que certaines imperfections typographiques.

Enfin, après un travail de préparation de données selon la codification requise, le dépouillement automatique des textes et l'indexation du corpus ont été effectués d'un côté de façon automatique par le logiciel Hyperbase, et de l'autre par le logiciel Lexicométrie qui travaille selon d'autres paramètres et demande une contribution manuelle partielle.

C'est avec l'aide du logiciel Hyperbase que nous avons, après avoir préparé les

différents sous-corpus, créé les multiples variantes de notre corpus, notamment celles de l'œuvre romanesque créée en différentes versions qui ont également requis une préparation des données minutieuse.

1.4.4. Le codage.

“L'un des préceptes élémentaires de toute statistique, écrit Charles Muller¹²⁴, est que l'on n'opère valablement que sur des *populations*, des *individus* et des *caractères* bien définis.”

Il en déduit qu'il peut sembler hasardeux de faire des statistiques sur les *mots* car la définition de *mot* en linguistique est ambiguë et n'a pas de délimitation satisfaisante. Selon sa terminologie¹²⁵, les *mots* sont les “unités dont la suite constitue un énoncé ou un texte ; il s'agit essentiellement d'une unité graphique séparée des unités voisines par un blanc ou un signe de ponctuation.” Le *mot* est donc simplement une unité graphique de texte. Dans notre étude, nous employons, comme le propose Étienne Brunet¹²⁶, le terme d'*occurrence* ; l'ensemble des occurrences d'un texte est symbolisé par *N*.

“Si l'appréciation du nombre de mots (*N*), écrit Charles Muller¹²⁷, crée relativement peu de difficultés, c'est que la tradition orthographique et typographique apporte au problème une solution qui est acceptable pour le linguiste dans la très grande majorité des cas. Si on appelle *unité graphique* un groupe de signes alphabétiques séparé des autres signes de ponctuations, on pourra presque toujours admettre l'équivalence : unité graphique = mot.”

¹²⁴ Ch. Muller (1979) : p. 125.

¹²⁵ Ch. Muller (1977) : p. 4.

¹²⁶ É. Brunet (1978) : p. 23.

¹²⁷ Ch. Muller (1979) : p. 128.

Toutefois, dans la réalité de la langue française cet idéal ne pourrait s'admettre de façon générale et ce phénomène cause de nombreuses difficultés à tout travail de traitement statistique et automatique de la langue. Il faut tenir compte des cas où l'équivalence de l'unité graphique et de l'unité linguistique peut être mise en cause. Il s'agit souvent des cas où :

- L'unité graphique peut correspondre à plusieurs mots ; il y a par exemple les formes contractées *au*, *aux*, *du* et *des*. Ensuite il y a le cas de deux signes graphiques qui ne sont ni des signes alphabétiques, ni des signes de ponctuation et dont la présence à l'intérieur d'une séquence peut créer un doute : l'apostrophe et le trait d'union. Il faut considérer l'apostrophe comme ambiguë. "Faut-il compter différemment le loup et l'agneau ? demande Charles Muller¹²⁸. Le considérer comme deux mots c'est briser des unités comme *aujourd'hui* et *presqu'île*." Le trait d'union est également ambigu, d'une part il sert à réunir des groupes lexicalisés : *chauve-souris*, *après-midi*, *va-et-vient* etc., d'autre part, il unit certaines combinaisons libres formées d'un mot et d'un second élément autonome comme *celui-ci*, *lui-même*, *vient-il*, etc.
- Parfois plusieurs unités graphiques peuvent former un seul mot, soit pour des raisons morphosyntaxiques, soit pour des raisons lexicales. Au niveau morphosyntaxique on pourrait par exemple compter (vous) *avez été surpris* comme un seul mot, c'est-à-dire comme une occurrence d'une forme du verbe *suspendre* ; selon l'analyse et au niveau lexical on cite généralement des lexies telles que *pomme de terre* et *chemin de fer*.

Les *vocables*, désignés par (V), sont les unités lexicales de signification qui, invariables ou constituées par plusieurs formes flexionnelles, sont représentées dans les dictionnaires par une entrée, le *lemme* (le mot sous sa forme canonique : l'infinitif pour le verbe, la forme masculine au singulier pour l'adjectif, etc.) Le

¹²⁸ Ch. Muller (1979) : p. 129.

lemme représente donc l'ensemble des formes flexionnelles relevant de la même entrée de dictionnaire.

L'étude quantitative peut se fonder sur la fréquence d'occurrences ou de vocables d'un texte, mais également sur celles des *formes* (appelées aussi *types*), qui sont les différentes formes graphiques d'un texte qui regroupent les *occurrences* (*tokens*) qui sont à leur tour le nombre d'unités total d'un corpus. Par exemple : soit une phrase comportant 3 fois la forme *faisait* et 1 fois la forme *fit*, on comptera donc 1 vocable (*faire*), 2 formes (*faisait*, *fit*) et 4 occurrences. Dans l'étude du nombre de *formes*, il n'y a ni séparation d'homographes ni regroupement des formes fléchies, c'est-à-dire qu'il n'y a pas de lemmatisation.

Dans des études lexicales quantitatives assistées par ordinateur, la complexité de la sémantique lexicale pose évidemment des difficultés. L'ordinateur, considérant comme unité lexicale celle qui est comprise entre deux caractères délimiteurs (les "blancs" graphiques, la ponctuation et les apostrophes), ne peut tenir compte de la polysémie d'un mot, des mots composés, des flexions, etc.¹²⁹.

"La statistique lexicale, écrit Charles Muller¹³⁰, à un moment de son œuvre, doit exclure ces nuances. Après avoir dressé l'inventaire des formes, il arrive un instant où elle doit fournir des chiffres en face des symboles N et V, et opter dans chacun des cas qui ont été examinés ici. C'est l'ensemble de ces options qui constituera la norme. [...] On admettra qu'il n'y a pas de norme parfaite, parce que le caractère complexe et mouvant du langage n'obéit jamais parfaitement à la quantification. On admettra même qu'il puisse y avoir plusieurs normes, non seulement parce que les avis des linguistes diffèrent sur certains faits, mais parce que les recherches vont dans des directions bien déterminées et divergentes. Une

¹²⁹ Ch. Muller (1963) : "Le mot, unité de texte et unité de lexique", in *Langue française et linguistique quantitative – recueil d'articles*, Genève, Slatkine, 1979, p. 125-143, et (1977) : *Principes et méthodes de statistique lexicale*, ch. 3 et 4.

¹³⁰ Ch. Muller (1979) : p. 142.

statistique dont le but est pratique... est forcément amenée à considérer les faits en fonction de ce but.”

Néanmoins, depuis ce constat de Charles Muller, la technique a fait quelques progrès, notamment en matière de lemmatisation et de levée d’ambiguïté. Nous effectuerons donc dans notre étude, afin d’enrichir l’étude sur le vocabulaire de Le Clézio, une comparaison de plusieurs méthodes : premièrement sur les formes graphiques du texte, c’est-à-dire les unités séparées par des blancs ou des signes de ponctuation dans le texte courant, où les occurrences identiques sont réunies en types, définies uniquement par leur représentation graphique (sans tenir compte de leur fonction dans la phrase ou de leur flexion) ; deuxièmement, sur un corpus lemmatisé, dans lequel les formes flexionnelles sont réunies sous leur lemme. Toutefois, la lemmatisation s’appuie sur plusieurs techniques et sur des normes différentes qu’il convient, avant de procéder à l’étude même, de définir plus en détail.

1.5. La lemmatisation.

“Seul le langage est rationnel :
les mots sont faits de la même
substance que la réalité”

L'extase matérielle, page 52.

La lemmatisation consiste à donner des règles d'identification qui permettent de regrouper les unités graphiques qui correspondent aux différentes flexions d'une même entrée de dictionnaire, et à associer à chaque mot d'un texte, dont la graphie a été normalisée, un lemme composé d'une forme canonique et d'une catégorie grammaticale.

Pour lemmatiser le vocabulaire d'un texte français, le regroupement en lemmes s'effectue en général¹³¹ selon les principes suivants, en ramenant :

- les formes verbales à l'infinitif ;
- les substantifs selon leur genre ;
- les adjectifs au masculin singulier ;
- les formes élidées à la forme sans élision ;
- Les formes contractées aux deux éléments constitutifs de base.

Ce procédé vise à permettre des décomptes sur des unités beaucoup plus soigneusement définies du point de vue de la langue. Toutefois, comme l'écrivent Ludovic Lebart et André Salem¹³²:

“...s'il est relativement aisé de reconstituer, “en langue”, l'infinitif d'un verbe à partir de sa forme conjuguée, le substantif singulier à partir du pluriel etc., la détermination systématique du lemme de

¹³¹ L. Lebart, A. Salem (1994) : p. 36.

¹³² *Ibid.*, p. 37.

rattachement pour chacune des formes graphiques qui composent un texte suppose, dans de nombreux cas, que soient levées préalablement certaines ambiguïtés.”

La langue française en effet, présente de nombreuses ambiguïtés. Elles peuvent résulter d’une homographie entre deux formes flexionnelles de lemmes différents ou d’une homographie entre deux lemmes ; l’ambiguïté peut ainsi se situer à divers niveaux. Citons l’exemple que donne Charles Muller¹³³ de la forme graphique *marche* où il distingue quatre niveaux :

- “un niveau *lexical* : la forme graphique *marche* peut être une occurrence du verbe *marcher* ou un nom féminin ;
- un niveau *sémantique* : le substantif féminin *marche* peut désigner soit un objet (la partie d’un escalier), soit un mouvement (“pratiquer la marche”), soit encore une province frontière (cette acceptation étant en général distinguée au niveau lexical) ;
- un niveau *grammatical* : la forme verbale *marche* peut être indicatif ou subjonctif ou impératif 1^e ou 3^e personne ou impératif 2^e personne ;
- un niveau *syntactique* : le nom *marche*, ou tout autre nom, peut être sujet, objet, complément prépositionnel, etc.”

En outre, il y a d’autres formes d’ambiguïté : la délimitation du mot que nous avons évoquée dans le chapitre précédent où un même caractère, par exemple le point, l’apostrophe ou le trait d’union, peut ou non jouer le rôle de séparateur, pose de nombreux problèmes.

¹³³ Ch. Muller (1993) : p. 61.

Lorsqu'il s'agit du lemme, il est encore plus difficile à délimiter. La référence de l'entrée dans le dictionnaire ne suffit pas pour régler le problème. Les lemmes à plusieurs constituants sont souvent traités à l'intérieur des articles des dictionnaires sous les lemmes "simples" et les lexicographes accordent rarement une entrée indépendante aux syntagmes ou aux expressions lexicalisées. Il faut donc les relever de façon manuelle et les ajouter à la nomenclature, ce qui amène à des choix souvent délicats. La langue française est riche de ces constructions et il est en fait difficile de savoir s'il s'agit d'une suite de mots intangible, lexicalisée ou seulement d'une suite de mots fréquente mais modifiable. Le résultat de ce travail de tri et de choix varie évidemment selon les principes et les normes choisis.

Ces ambiguïtés et bien d'autres posent problème lors du regroupement, qui impose souvent des choix arbitraires. Par conséquent, il est très important d'avoir recours à une norme bien définie lors de la lemmatisation. La difficulté d'établir une norme de dépouillement cohérente pour la lemmatisation est encore aujourd'hui d'actualité.

En fait, le choix entre les deux méthodes, soit de travailler sur un corpus s'appuyant sur la forme graphique, soit d'avoir recours à un corpus qui s'appuie sur le lemme, n'est pas aisé et ce choix de méthode a de nombreuses implications. Anthony A. Lyne l'exprime ainsi¹³⁴ :

"The decision to lemmatize is a very weighty one. It implies that we commit ourselves to deciding, with regard to every word-token (segment) of the corpus without exception, (a) whether it is a canonical form or lemma, (b) if not, to which lemma it is to be assigned. Quantification will not allow us to leave anything in doubt."

¹³⁴ A. A. Lyne (1985) : p. 49.

Il est intéressant, avant de parler des différentes méthodes et de présenter la nôtre, d'exposer les arguments des uns et des autres, ceux qui rejettent comme ceux qui recommandent la lemmatisation des données.

La lemmatisation a longtemps été l'objet de discussions parmi les experts en lexicométrie. Le débat a commencé il y a trente ans, quand certains chercheurs¹³⁵ ont commencé à contester les recommandations de Charles Muller pour la lemmatisation des données.

Selon la tradition française, les travaux de lexicométrie s'étaient, jusqu'alors, principalement intéressés aux vocables et le traitement des textes dépouillés était, en grande partie, manuel. Avec le développement de l'informatique, l'évolution des machines et des logiciels, l'accès aux textes est devenu relativement facile et le travail a cessé d'être manuel. Au fur et à mesure que les corpus devenaient de plus en plus grands, l'effort pour effectuer les dépouillements manuellement devenait surhumain. En ce qui concerne les corpus en langue française, il a pendant longtemps manqué les logiciels d'étiquetage qui auraient diminué la part manuelle du travail. L'effort de désambiguïsation semblait alors démesuré par rapport aux bénéfices que l'on pouvait en tirer. De nombreux chercheurs ont alors choisi de travailler sur les formes, c'est-à-dire sur les unités graphiques.

L'effort et le coût de la désambiguïsation n'est pas le seul argument des adversaires de la lemmatisation. Il y a en effet d'autres motivations et raisons en faveur de la méthode de travail qui consiste à s'appuyer sur les formes graphiques, outre celui de la relative facilité de préparation et de traitement.

En effet, certains chercheurs opposent toujours aux dénombrements de mots lemmatisés, désambiguïsés et classifiés, une quantification qui porte directement sur les formes graphiques et leur organisation à la surface du texte. Regroupés dans le laboratoire de lexicologie politique (E.N.S. de Saint-Cloud), ils affirment, à

¹³⁵ Le laboratoire de lexicologie politique E.N.S. de Saint-Cloud.

la suite de Maurice Tournier¹³⁶, “leur scepticisme quant aux études hors contexte, à la possibilité même d’une statistique du “lexique” et aux interventions qui transforment un corpus en témoin de langue”.

“Cette position critique, écrit-il, les conduit à éviter des interprétations prématurées au niveau des signes observés. Pour eux, le “texte est roi”, son vocabulaire ne constitue pas l’échantillon représentatif d’un lexique virtuel mais la réalité première. S’en tenir ainsi aux données matérielles, c’est proposer une description de séries et de réseaux de “formes” indicées, dont l’interprétation s’opère non par inférence sur une langue en soi mais par retour aux conditions de la production des textes.”

Il évoque aussi un second principe ; “refuser pour les mots de parler de ‘fréquence en langue’, c’est considérer les variables psycho-sociales, repérées et choisies lors de la constitution du corpus, comme explicatives des distorsions constatées dans les comptes des sorties-machine”.

Les critiques évoquent également les polysémies propres à la langue comme une argumentation contre la lemmatisation. Parfois les mots d’une même souche étymologique acquièrent des acceptions différentes et les variations du genre, du nombre, du temps, du mode et de la personne peuvent également introduire une polysémie qu’il convient de conserver et repérer. P. Lafon¹³⁷ cite des exemples de variations de sens qui vont au-delà de la simple nuance entre, par exemple, *peuple* et *peuples*, *histoire* et *histoires*, *lumière* et *Lumières*. Dans le cas de *Liberté* et *Libertés* la lemmatisation fait perdre une opposition de sens distinctive de deux discours politiques opposés. En effet, dans les études des textes politiques, les chercheurs ont souvent pu constater que le singulier et le pluriel d’un même substantif renvoient à des notions différentes, partant à des idéologies diverses. André Salem

¹³⁶ M. Tournier, “Lexicométrie”, in *Encyclopadia Universalis*, DVD version 6.0.

¹³⁷ P. Lafon (1984) : p. 104.

et Ludovic Lebart¹³⁸ signalent en exemple le cas de *défense de la liberté* et *défense des libertés*.

La standardisation dans l'opération de décanation est une difficulté souvent invoquée, elle porte inévitablement sur des critères divers lors de l'établissement d'une norme et des inévitables choix qui s'ensuivent. Il est en effet nécessaire, lors de la lemmatisation, de prendre position sur les cas ambigus, que nous venons de souligner, et cette prise de position n'est jamais neutre. Benoît Habert et André Salem évoquent "l'inévitable éparpillement des étiquetages"¹³⁹. Selon Étienne Brunet¹⁴⁰ "on court toujours le risque de la "babélisation".

Un argument important de ceux qui ne considèrent que les formes brutes est celui de la comparaison, les résultats d'une étude sur un corpus non lemmatisé la permettant plus facilement : les données lemmatisées offrent en effet peu de prises à la comparaison, non seulement du fait de leur rareté, mais aussi parce que les exigences en matière de lemmatisation sont variables selon les corpus et les chercheurs, de sorte qu'il est difficile de trouver un standard pour les corpus lemmatisés, chacun proposant ses propres règles. La réduction au traitement des formes graphiques a en effet l'avantage d'offrir la possibilité de rapprochement entre des corpus comparables, quoique issus de sources diverses. "Les rencontres et les comparaisons sont plus faciles quand les règles sont minimales, écrit Étienne Brunet¹⁴¹. Le degré zéro et le niveau le plus bas correspondent à l'absence complète de toute manipulation préalable des données".

La principale difficulté de la lemmatisation, selon Étienne Brunet¹⁴², réside dans le fait que "c'est le contexte qui fixe la nature et les limites de la chaîne graphique. Le

¹³⁸ L. Lebart, A. Salem (1994) : p. 34.

¹³⁹ B. Habert, A. Nazarenko, A. Salem (1997) : p. 23.

¹⁴⁰ É. Brunet (2002) : "Le lemme comme on l'aime", in *J-ADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, A. Morin, P. Sébillot (éds.), Saint-Malo, Irisa, Inria, 13-15 mars, p. 221-233.

¹⁴¹ É. Brunet (2000) : "Qui lemmatise dilemme attise", in *Scolia, 11èmes rencontres linguistiques en Pays Rhénan*, L. José et A. Theissen (éds.), Strasbourg, Publications de l'Université Marc Bloch n°13, 2000, p. 3.

¹⁴² É. Brunet (2002) : "Le lemme comme on l'aime", in *J-ADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, A. Morin, P. Sébillot (éds.), Saint-Malo, Irisa, Inria, 13-15 mars, p. 222.

recours aux dictionnaires ne résout pas le problème de la nomenclature. Encore moins celui des homographes.”

Pourtant, d’après Étienne Brunet¹⁴³

“Dans le procès que font à la lemmatisation certains défenseurs de la forme brute, on omet souvent de considérer que la lemmatisation n’est pas une facilité, une réduction, un appauvrissement. Certes le nombre d’unités diminue et les marques de flexion disparaissent mais le matériau a été raffiné et le contenu a gagné bien plus que la forme n’a perdu. Le respect de la forme devrait conduire au respect du sens et si l’on refuse de confondre des formes différentes, combien plus forte encore devrait être l’exigence de ne pas mêler les contenus.”

Charles Muller n’a quant à lui jamais cessé de recommander de soumettre les données à la lemmatisation. Il expose ses arguments en faveur de cette contrainte préalable, et contredit ainsi Pierre Lafon dans la préface qu’il a donnée à son ouvrage¹⁴⁴.

“Les recherches linguistiques, stylistiques ou lexicométriques ne se réduisent pas à la thématique, écrit-il. Si l’on s’intéresse à la syntaxe ou à plus forte raison à la morphologie, il faut avoir accès à la forme canonique du mot.”

L’étude grammaticale demande en effet l’accès au lemme et ne peut s’effectuer de façon satisfaisante sur un corpus qui s’appuie sur la forme graphique. Elle permet l’étude des fonctions grammaticales, des parties du discours et l’étude à l’intérieur d’une catégorie grammaticale, par exemple l’étude des temps, des modes et des personnes du verbe.

¹⁴³ É. Brunet (1981) : p. 26.

¹⁴⁴ P. Lafon (1984) : p. 2.

Selon Étienne Brunet¹⁴⁵ :

“Les homographes engendrent une pollution diffuse, qui, de loin, peut échapper à l’observation. Mais pour le détail des mots individuels, la différence est sensible entre formes brutes et lemmes. La sûreté de ces derniers vient certes de leur pureté mais aussi d’une assise plus large livrée à l’observation. L’émiettement des formes diminue en effet l’effectif de chacune. Elle affaiblit aussi leur témoignage que d’autres formes, appartenant au même lemme, peuvent contester ou contredire, et le jugement reste en suspens quand un même vocable distribue ses représentants parmi les excédents et les déficits. Quand les formes verbales, par exemple, sont rassemblées derrière un même lemme, il est plus aisé de tirer des conclusions de leur fréquence cumulée.”

Dominique Labbé n’hésite pas à souligner jusqu’à quel point l’analyse s’appuyant sur des données brutes peut “brouiller”, jusqu’aux moindres éléments morphosyntaxiques, les résultats obtenus¹⁴⁶.

“Comment mettre de l’ordre dans cette émeute des formes graphiques ?” écrit-il. Tous ceux qui ont été un jour confrontés aux listes “segments répétés” le savent : le “bruit” provoqué par les règles de la graphie et de la syntaxe française est si fort qu’il couvre presque complètement le sens.”

La lemmatisation et l’étiquetage des corpus apporte en effet beaucoup de bénéfices. En particulier, on retrouve enfin les substantifs *devoir*, *pouvoir*, etc., le point cardinal *est* ou la saison estivale perdus dans les participes passés de *être*, etc.

¹⁴⁵ É. Brunet (2000) : “Qui lemmatise dilemme attise”, in *Scolia, 11èmes rencontres linguistiques en Pays Rhénan*, L. José et A. Theissen (éds.), Strasbourg, Publications de l’Université Marc Bloch n°13, 2000, p. 17.

¹⁴⁶ D. Labbé (2001) : “Analyse des données textuelles et Statistique lexicale”, in *Lexicometrica*, n°3, 2001, p. 3.

Il convient ici de rappeler à cet égard, comme le fait Dominique Labbé, qu'en moyenne plus du tiers des mots d'un texte sont "homographes" et que certains sont extrêmement usuels.

"Imaginons maintenant, écrit-il, une vaste base de données textuelles normalisées et étiquetées contenant plusieurs millions de mots judicieusement choisis dans les grandes œuvres du présent et du passé mais aussi dans la presse, sur le net ou dans les conversations quotidiennes. Grâce à cet étalon de référence, nous pourrions donner instantanément les différents sens d'un mot dans la langue, fournir les citations les plus illustratives et mesurer la singularité de tel ou tel locuteur."

Les meilleurs travaux, selon Étienne Brunet¹⁴⁷ sont bien sûr ceux qui s'appuient sur des données rigoureusement lemmatisées. Toutefois, les logiciels les plus répandus s'appuient de préférence sur des données brutes, ce qui fait que les corpus lemmatisés sont les plus rares.

Aujourd'hui, osons le dire, la normalisation et l'étiquetage des données textuelles non seulement sont devenus des nécessités impérieuses qui doivent être effectuées selon des standards explicites, mais ces procédés et la réflexion qui s'ensuit font partie intégrante de tout travail lexicométrique.

Faut-il lemmatiser ? "La décision, conclut André Salem¹⁴⁸, est d'ordre économique. Il est dans l'absolu toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes). Une lemmatisation complète, sur un corpus important, reste une opération coûteuse. Indispensable dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations et des typologies [...]"

¹⁴⁷ É. Brunet (2000) : "Qui lemmatise dilemme attise", in *Scolia, 11èmes rencontres linguistiques en Pays Rhénan*, L. José et A. Theissen (éds.), Strasbourg, Publications de l'Université Marc Bloch n°13, 2000, p. 3.

¹⁴⁸ L. Lebart, A. Salem (1994) : p. 226.

Nous avons voulu, dans notre étude, afin de comparer ces deux méthodes de travail, soumettre notre corpus non seulement à la quantification en formes graphiques mais également à la lemmatisation assistée par ordinateur selon différentes techniques.

1.5.1. Les lemmatiseurs.

Les différents logiciels procèdent à l'étiquetage du corpus selon des règles établies au préalable. Ils associent à des segments de texte, c'est-à-dire aux mots, une ou plusieurs étiquettes, le plus souvent leur catégorie grammaticale ou leur lemme. L'étiquetage, qui jadis était manuel, est désormais effectué par un programme informatique que l'on appelle "étiqueteur". Les étiqueteurs travaillent soit de façon automatique, soit en combinant le traitement informatique avec une annotation manuelle. En effet, étant donné qu'un corpus étiqueté n'est pas forcément totalement désambiguïsé et qu'il reste certains problèmes non résolus pendant l'opération d'étiquetage automatique, certains logiciels aboutissent à une phase interactive qui permet le dialogue entre homme et machine pour lever ces ambiguïtés.

L'avantage de cette méthode est qu'elle permet une désambiguïsation plus fine et complète, l'inconvénient est que les réponses aux questions posées par le logiciel ne sont jamais neutres, et des réponses différentes peuvent avoir des conséquences lors de la comparaison avec d'autres corpus étiquetés par une autre personne ou à une époque plus ou moins éloignée.

Toutefois, il existe de nombreuses variantes d'analyseurs lexicaux. Beaucoup se limitent au repérage et à l'étiquetage des mots simples, sans restitution de leur forme canonique et sans prise en compte des mots composés. La plupart des

analyseurs syntaxiques indiquent la fonction des mots de la phrase ou construisent l'arbre syntaxique de celle-ci, selon le type de grammaire sur lequel ils s'appuient, mais ceux qui accomplissent les deux tâches à la fois sont rares. Il est encore plus difficile de trouver un lemmatiseur qui lève toute ambiguïté des prédicats. Pour les analyseurs sémantiques, la variété des sorties est encore plus grande : traduction en formules logiques, ou en graphes conceptuels, décomposition en phrases simples (minimales), etc. Pour répondre à la question qui se pose tout naturellement de connaître l'avantage d'une base étiquetée sur celle qui ne l'est pas et des différents modes d'étiquetage, nous avons soumis la base "Le Clézio", successivement, à plusieurs traitements, avec ou sans étiquetage.

1.5.2. Les méthodes de quantification de notre corpus.

Après avoir effectué de nombreux traitements informatiques, nous disposons désormais de plusieurs versions de notre corpus, s'appuyant sur les formes graphiques ou sur les lemmes.

Les différents outils utilisés sont :

- l'étiqueteur intégré dans Hyperbase
- le lemmatiseur Winbrill
- le programme Cordial
- L'application de la méthode de Dominique Labbé :
 - intégré dans son logiciel Lexicométrie
 - intégré dans Hyperbase par Étienne Brunet

Ces logiciels opèrent selon des techniques et des méthodes différentes et s'appuient sur des normes qui varient d'un logiciel à l'autre. Il est par conséquent utile, avant de présenter les résultats issus des différents traitements, de donner une présentation de chacune.

1.5.3. Hyperbase.

Hyperbase, dans sa version originale¹⁴⁹ qui s'appuie sur les formes graphiques, propose des fonctions simples de regroupement.



Figure n° 9 : Fonctions de regroupement des mots d'Hyperbase.

Le logiciel permet en premier lieu le regroupement des formes graphiques. Mais d'autres regroupements sont également possibles. Le bouton "lemme" permet le regroupement des formes derrière la forme canonique, et d'obtenir par exemple toutes les formes conjuguées d'un verbe. Il ne s'agit pas, dans ce cas, d'une vraie lemmatisation. Le regroupement des formes est fait de façon rudimentaire, ce qui oblige les homographes à se ranger tous ensemble sous une catégorie ou sous une autre. Le logiciel permet des recherches sur des expressions, c'est-à-dire sur une suite de mots ou de signes, par exemple *pomme de terre*. La recherche se fait également sur les débuts de mots, ce qui permet, en fournissant par exemple le préfixe *aim*, d'atteindre indirectement toutes les formes conjuguées du verbe *aimer*. Le bouton "final" permet un regroupement à partir des fins de mots. Les suffixes permettent, "par déduction", d'effectuer de nombreuses analyses. Nous trouvons par exemple de nombreuses études sur les adverbes en *-ment*, qui sont souvent regroupés et étudiés grâce à ces fonctions de base. La fonction "chaîne" permet la recherche d'une chaîne de caractères où qu'elle se trouve dans un mot. Le bouton "liste" permet de faire des recherches sur une liste constituée au préalable.

Ce type de regroupement lexical s'exerce uniquement sur le dictionnaire, et aucun examen du contexte n'est effectué. Par exemple, nous y trouvons toutes les

¹⁴⁹ Nous travaillons principalement avec la version 5.1.

formes de *marche* ou de *marches* rattachées ensemble au verbe ou au substantif, ce qui limite évidemment la fiabilité et l'utilité du regroupement.

1.5.4. Winbrill.

Durant la phase initiale de notre travail, nous avons exploité le lemmatiseur Winbrill qu'Étienne Brunet a intégré dans une version d'Hyperbase¹⁵⁰, adapté à l'environnement Windows. Dérivé des travaux de Brill et du MIT, il a été adapté au français par deux chercheurs de l'INaLF à Nancy, J. Lecomte et G. Souvay. Le logiciel Winbrill n'opère qu'un étiquetage grammatical ; les classes qu'on y distingue ne sont pas les classes traditionnelles et la division manque parfois de clarté. La figure ci-dessous montre les codes utilisés par le programme.

étiquette d'origine	étiquette <u>simplifiée</u>	signification	étiquette d'origine	étiquette <u>simplifiée</u>	signification
ABR	<u>ABR</u>	abréviation	-----	<u>S</u>	substantif -----
-----	<u>AV</u>	avoir-----	SBC		nom commun
ACJ		AVOIR conjugué	SBP		nom propre
ANCF		AVOIR infinitif	SBP?		nom propre probable
ANCNT		AVOIR forme en -ant	-----	<u>DTN</u>	déterminant-----
APAR		AVOIR p.passé après AVOIR	DTN		déterminant
-----	<u>E</u>	être -----	DTC		déterminant contracté
ECJ		ETRE conjugué	-----	<u>PR</u>	pronom_adj.-----
ENCF		ETRE infinitif	PRV		pronom supporté par le verbe
ENCNT		ETRE forme en -ant	PRO		autre pronom (ou adj)
EPAR		ETRE p.passé après AVOIR	PRV_\$\$		pronom indéterminé (en, y, se)
-----	<u>V</u>	verbe -----	-----	<u>REL</u>	-----
VCJ		verbe conjugué	REL		relatif (adj., pron. ou adv.)
VNCF		verbe infinitif	-----	<u>PREP</u>	-----
VNCT		verbe forme en -ant	PREP		préposition
VPAR		verbe p. passé après AVOIR	-----	<u>SUB</u>	-----
-----	<u>ADV</u>	-----	SUB		subordonnant
ADV		adverbe	SUB\$		code par défaut pour QUE
-----	<u>A</u>	adjectif-----	-----		-----
ADJ		adjectif (participe passé exclu)	INJ		interjection ou onomatopée
ADJ1PAR		part.passé adjectif derrière ETRE	PUL		particule
ADJ2PAR		par.passé adjectif NON derrière ÊTRE	SYM		symbole ou signe mathém.
		CAR			cardinal (chiffres ou lettres)
			FGV		mot étranger
			sg		singulier
			pl		pluriel

CLIQUER DANS LE CHAMP POUR
LE FAIRE DISPARAITRE

CLIQUER...dans le champ pour le faire disparaître

Figure 10 : Les codes grammaticaux de Winbrill.

¹⁵⁰ Hyperbase 3.0, version Windows.

L'opération d'étiquetage est entièrement automatique, ce qui est, selon Étienne Brunet¹⁵¹ "à la fois la force et le handicap de Winbrill". Comme il n'y a pas de choix subjectif, les comparaisons sont possibles d'un corpus à l'autre, le traitement étant identique. Malheureusement, les erreurs d'analyse sont nombreuses et il y a souvent des confusions entre classes grammaticales. La figure ci-dessous montre une page de recherche de concordance du verbe *marcher* où nous voyons que le logiciel n'a pas su séparer le substantif homographe de la forme fléchie du verbe.

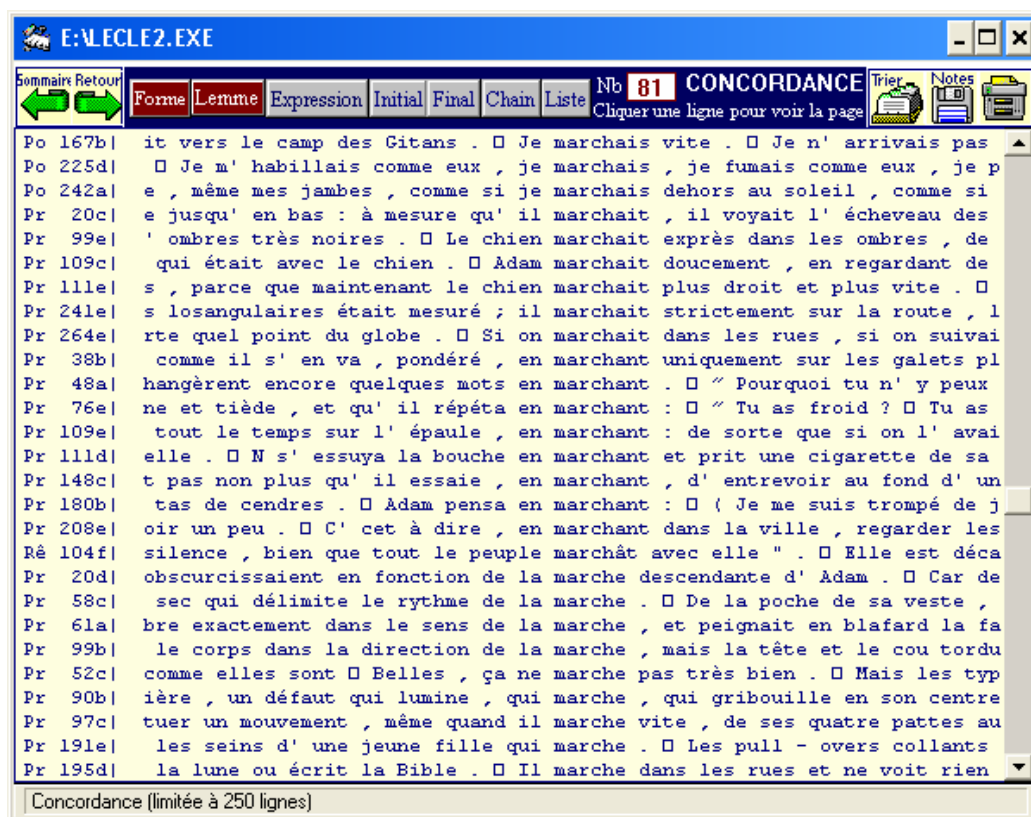


Figure n°11 : Une page de concordance analysé par Winbrill.

Une autre difficulté liée à l'exploitation du corpus lemmatisé avec Winbrill est que la lemmatisation n'est pas complète car il manque la mention du lemme correspondant, le code, qu'exige l'exploitation adéquate du logiciel.

¹⁵¹ É. Brunet (2002) : "Le lemme comme on l'aime", in *JADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, A. Morin, P. Sébillot (éds.), Saint-Malo, Inria, 13-15 mars, 2002, p. 221-222.

Vu ces difficultés, nous nous sommes intéressée à d'autres lemmatiseurs, notamment celui élaboré par Dominique Labbé où la lemmatisation respecte les normes de lemmatisation définies par Charles Muller.

1.5.5. La méthode Labbé et le programme Lexicométrie.

Le logiciel Lexicométrie est une suite de programmes permettant la mise à la norme des textes et le codage grammatical des mots¹⁵². La normalisation et l'étiquetage des textes s'appuient sur les théories et les normes de Charles Muller et celles du Laboratoire de lexicologie politique de Saint Cloud. Ce logiciel est composé d'un ensemble de programmes qui s'enchaînent de la façon suivante¹⁵³ :

Le premier programme appelé "Saint Cloud" vise à standardiser les formes graphiques et les délimiteurs de forme ou de séquences (blancs et ponctuations) selon des règles préétablies. Il repère la structure du texte, la séparation en phrases, les ponctuations et les mots. Chaque élément séparé fait l'objet d'un enregistrement dans le fichier de sortie (extension SC3). Cette segmentation met en minuscule la première lettre du premier mot de la phrase, sauf si c'est un nom propre retrouvé dans une table des noms propres. Les mots composés, avec ou sans tirets, ne sont pas séparés s'ils sont retrouvés dans la table des mots composés, il en est de même pour les locutions figurant dans la table. Enfin le "Saint Cloud" transforme les nombres cardinaux en lettres. A la fin de ce programme le fichier se trouve sous forme segmentée et normalisée.

Le deuxième traitement de codage CM¹⁵⁴ se poursuit en trois phases. La première phase constitue la confrontation avec la table générale des formes. Si la forme est

¹⁵² Les normes de saisie et de dépouillement sont décrites en détail dans : Labbé D. (1990) : *Normes de saisie et de dépouillement des textes politiques*, CERAT, cahier n°7, avril 1990.

¹⁵³ Pour plus de précisions sur les différents programmes, voir J. Picard, A. Pibarot et D. Labbé (1995), *Un outil de statistique textuelle : le lemmatiseur*, p. 305-306.

¹⁵⁴ Ce programme évoque les initiales de Charles Muller qui a énoncé (comme nous l'avons déjà évoqué) les principes fondamentaux de lemmatisation.

trouvée, alors le codage a lieu en précisant la catégorie grammaticale, le genre et le nombre. Si la table indique un homographe possible avec un verbe pour lequel la forme n'est pas trouvée dans la table, alors la procédure de lemmatisation des verbes est lancée. Si aucun verbe ne correspond à la forme, alors selon la présence ou non dans la table générale, on considère que l'homographie est levée ou que c'est une forme inconnue.

Dans la deuxième phase (CM2) la procédure de résolution automatique des homographies est lancée. Le principe réside en une analyse de contexte étroit limité à quelques mots. Par un certain nombre de règles, on arrive à lever un grand nombre d'ambiguïtés.

La dernière phase est celle de la correction interactive pour traiter les mots inconnus et lever les homographies non résolues. Le programme pose des questions en fonction de l'expertise, des règles et du contexte et il propose un choix de solutions.

Le fichier final (CM3) est constitué d'enregistrements contenant le numéro d'ordre de la forme, le lemme associé et le code de la catégorie grammaticale. Les fichiers issus de cette procédure peuvent être soumis à des études statistiques textuelles directes comme l'indexation, l'index alphabétique, l'index hiérarchique et les concordances.

Toutefois, même si le pourcentage de recours à l'utilisateur dans la dernière phase est faible et ne dépasse pas 2 % (alors que dans un texte quelconque environ 30 % des mots présentent des difficultés de codification) il représente, lorsque le corpus est de taille, un travail manuel considérable. Nous avons, pour cette raison, choisi un corpus plus restreint de 12 œuvres, le corpus H, pour cette expérience.

D'autres lemmatiseurs sont également proposés sur le marché, dont certains sont intégrés à des logiciels d'analyse textuelle. Nous exploiterons dans cette étude celui

qu'Étienne Brunet vient d'intégrer à son logiciel Hyperbase, le programme Cordial.

1.5.6. L'outil Cordial.

Le programme Cordial¹⁵⁵ (dans sa version 7), dont la compétence linguistique a été reconnue par de nombreux prix, a été intégré dans Hyperbase.

Le logiciel fournit quelque 200 codes grammaticaux différents, et quelques classes sémantiques. Ces codes associés aux formes sont d'un grand secours lorsque l'on veut isoler un homographe et par exemple ne plus confondre un article et un pronom. Ils sont aussi indispensables pour étudier les parties du discours. En effet, quand on s'intéresse à la syntaxe ou, à plus forte raison, à la morphologie, il faut avoir accès non seulement à la graphie mais également aux codes grammaticaux et au lemme lui-même.

Nous empruntons au projet Weblex¹⁵⁶ les tables de décodage des étiquettes qui ont été conçues pour aider (ou apprendre) l'interprétation d'étiquettes morpho-syntaxiques du type Cordial.

Ces étiquettes ont été conçues en faisant correspondre systématiquement à deux zones dans l'étiquette, d'une part la partie du discours et d'autre part la sous-catégorisation spécifique à cette partie du discours. La zone partie du discours est construite en associant à une partie du discours un seul caractère réservé. Par exemple, le caractère "V" est réservé à l'étiquetage des verbes du texte. La zone de sous-catégorisation est construite en associant à chaque trait de sous-catégorisation un seul caractère à une position réservée. Par exemple, la partie du

¹⁵⁵ Le correcteur que Microsoft Word a intégré dans sa version française est emprunté au programme Cordial.

¹⁵⁶ "Weblex", S. Heiden, Septembre 2000, UMR8503 Lyon, V 2.3 Juillet 1999, UMR8503, St Cloud, V. 1.0, septembre 1998, UMR9952, St Cloud.

discours Verbe est sous-catégorisée en traits type, mode, temps, personne, nombre et genre.

Voici les 11 tables de décodage de chaque partie du discours :

<u>N</u>	Nom ;	<u>D</u>	Déterminant ;
<u>V</u>	Verbe ;	<u>P</u>	Pronom ;
<u>R</u>	Adverbe ;	<u>I</u>	Interjection ;
<u>A</u>	Adjectif ;	<u>M</u>	Numéral ;
<u>S</u>	Préposition ;	<u>Y</u>	Ponctuation
<u>C</u>	Conjonction ;		

N	Nom						
	PdD	Type	Genre	Nombre			
N c m s	Nom	commun	masculin	singulier	tigre	menuisier	table
p f p		propre	féminin	pluriel	Jean	Liège	Québécois

V Verbe							
V m i p 1 s m	PdD	Type	Mode	Temps	Personne	Nombre	Genre
a s i 2 p f	Verbe	principal	indicatif	présent	première	singulier	masculin
f f 3		auxiliaire	subjunctif	imparfait	deuxième	pluriel	féminin
c s			impératif	futur	troisième		
n r			conditionnel	passé			

p m			infinitif	subj. Prés.			
c			participe	subj. Imp.			
é			conditionnel				
a			impératif				
			part. passé				

R Adverbe				
	PdD	Type	Degré	
R g p	Adverbe	général	non comparatif	et non négatif
p c				
n			négatif (ne ou n')	

A Adjectif					
	PdD	Type	Degré	Genre	Nombre
A f p m s	Adjectif	qualificatif	non comparatif	masculin	singulier
o c f p		ordinal	comparatif	féminin	pluriel
i		indéfini			
s		possesif			

S Préposition			
	PdD	Type	
			à, après, attendu, avant, avec, chez, concernant, contre, dans, de, depuis, derrière, dès,
S p	Adposition	Préposition	devant, durant, en, entre, envers, excepté, hormis, hors, jusque(s), malgré, moyennant,
			outre, par, parmi, passé, pendant, plein, pour, près, proche, sans, sauf, selon, sous,
			suivant, supposé, sur, touchant, vers, vu

C Conjonction			
	PdD	Type	
			rapports:
			addition, liaison: et, ni, puis, ensuite, alors, aussi, bien plus, jusqu'à, comme,
			ainsi que, aussi bien que, de même que, non moins que, avec,
			alternative, disjonction: ou, soit..., soit, soit...ou, tantôt, ou bien,
			cause : car, en effet, effectivement
	coordination		conséquence : donc, aussi, partant, alors, ainsi, par conséquent, en conséquence
			conséquemment, par suite, c'est pourquoi
			explication : savoir, à savoir, c'est-à-dire, soit
C c			opposition, restriction : mais, au contraire, cependant, toutefois, néanmoins,
s			pourtant, d'ailleurs, aussi bien, au moins, du moins, au reste, du reste,
			en revanche, par contre, sinon
Conjonction			
			Rapports :
			but : afin que, pour que, de peur que, ...
			cause : comme, parce que, puisque, attendu que, vu que, étant donné que
			comparaison : comme, de même que, ainsi que, autant que, plus que, moins que
			non moins que, selon que, suivant que, comme si, ...
	subordination		concession, opposition : bien que, quoique, alors que, tandis que,
			condition, supposition : si, au cas où, à condition que, pourvu que, à moins que,
			conséquence : que, de sorte que, en sorte que, de façon que, de manière que,
			temps : quand, lorsque, comme, avant que, alors que, dès lors que,
			tandis que, depuis que, ...

D Déterminant							
	PdD	Type	Personne	Genre	Nombre	Possesseur	Type d'article
	Déterminant	article	première	masculin	singulier	singulier	défini
D a 1 m s s d		adjectif démonstratif	deuxième	féminin	pluriel	pluriel	indéfini
d 2 f p p i		adjectif indéfini	troisième				
i 3		(adjectif interrogatif ?)					
s		adjectif possessif					
t		adjectif					
		interrogatif/exclamatif					
		(adjectif indéfin ?)					

P Pronom						
	PdD	Type	Personne	genre	Nombre	Cat
P p 1 m s n	Pronom	personnel non réfléchi	première	masculin	singulier	sujet (nominatif)
x 2 f p a		personnel réfléchi	deuxième	féminin	pluriel	COD (accusatif)
d 3 d		démonstratif	troisième			COI (datif)
i		indéfini				
s	possessif					
t	interrogatif					
r	relatif					

I Interjection		
	pDd	
		adieu, ah, ah!, aïe, allo (allô), bah, baste, barnique, bravo, ça, chiche, chut, crac, dame,
	Interjection	dia, eh, euh, fi, fichtre, foin, gare, ha, haïe, hardi, hé, hein, hélas, hem, ho, holà, hon,
I		, hourra, hue, hum, là, las, mince, motus, ô, oh, ohé, ouais, ouf, ouiche, ouste, paf,
		pan, patatras, pif, pouah, pst, quoi, sacristi, saperlipopette, saperlotte,
		sapristi, st, sus, vivat, zest, zut, ah ! ça, à la bonne heure, bonté

M Numéral				
	PdD	Type	Genre	Nombre
M c m s	Numéral	commun	masculin	singulier
f p			féminin	pluriel

Y Punctuation			
	PdD	Type	Genre
Y p w	Punctuation	ponctuation	finale
s s		autre	pause
o			ouvrante
c			fermante
s			autre

Dans le logiciel Hyperbase le système a été intégré par le concepteur avec l'ambition d'une convivialité d'utilisation et d'une interface synthétisée. La figure ci-dessous rend compte des choix proposés par le logiciel du point de vue morphologique aussi bien que syntaxique¹⁵⁷.

Catégorie 1	4Sous-cat.	Mode 3	Temps 4	Personne 5	Code choisi	1 2 3 4 5 6 7	Retour	Sommaire
Verbe V	principal m auxiliaire a	Infinitif n Indicatif i Subjonctif s Conditionnel c Impératif f Participe p	Présent p Imparfait i Passé s Futur f Subjonctif présent r Subjonctif imparfait m Participe passé a	1re pers. 1 2e pers. 2 3e pers. 3	N			
Substantif N	nom commun c nom propre p		Genre 4 Masculin m Féminin f	Numéral Mo	Effacer	Fonction 7		
Adjectif A	qualificatif f ordinal o	Positif p Comparatif c		Interjection Ij	Continuer			
Déterminant D	article a démonstratif d interrogatif i indéfini t		Nombre 5-6 Singulier s Pluriel p	Préposition Sp				
Pronom P	personnel réfléchi x pers. non réfléchi p possessif s démonstratif d interrogatif t indéfini i relatif r	1re personne 1 2e personne 2 3e personne 3	Fonction 6 sujet n objet direct a objet indirect d	Adverbe R Conjonction C Ponctuation Y				

Substantif,

Cliquer sur les critères souhaités puis sur le bouton CONTINUER

A - attribut du sujet
B - groupe attribut du sujet
C - objet direct
D - groupe objet direct
E - objet indirect
F - groupe objet indirect
G - complément d'agent
H - circonstanciel
K - circ. de temps
L - circ. de lieu
M - apposition
N - groupe apposition
O - apostrophe
P - groupe apostrophe
Q - complément de négatio
S - sujet
T - groupe sujet
U - pronominalisation
V - verbe base de propositi
Y - sujet réel
Z - groupe sujet réel
1 - ajout à l'adjectif
2 - reprise du COD
3 - reprise du COI
4 - reprise du circonstanciel
5 - ajout au nom
6 - ajout au pronom
7 - reprise du sujet
8 - ajout au verbe

Figure n°12 : Les différentes possibilités d'exploitation des codes grammaticaux.

Ce programme permet de nombreuses combinaisons de recherches différentes. Il offre également la possibilité de comparer, dans les analyses statistiques, des structures grammaticales et de choisir de travailler non seulement sur les formes et les lemmes mais également sur les codes. La figure ci-dessous rend compte d'une des potentialités qu'offre ce programme.

¹⁵⁷ Hyperbase modifie parfois le code de Cordial. Par exemple les conjonctions de coordination sont distinguées des subordonnantes, et les pronoms relatifs sont distingués des autres pronoms.

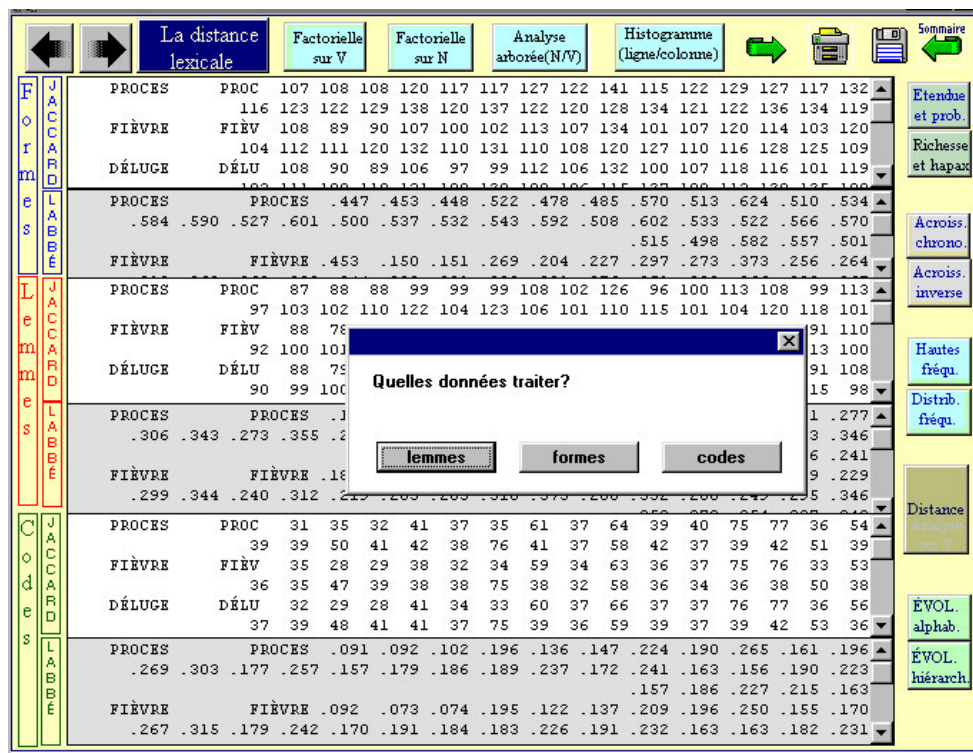


Figure n°13 : Le choix de différentes méthodes d'analyse d'Hyperbase avec Cordial.

Le logiciel permet de choisir quelles données traiter, c'est-à-dire de s'appuyer sur les lemmes, les formes ou bien les codes grammaticaux. Il offre également le choix entre plusieurs méthodes de calcul dans les traitements statistiques, comme nous allons le voir dans notre étude ; par exemple dans le calcul de la distance lexicale¹⁵⁸ le choix est donné entre le calcul sur la présence V (méthode Jaccard) et sur la fréquence N (méthode Labbé).

La possibilité de s'appuyer sur les différentes méthodes offre en fait au chercheur de nombreuses ouvertures de recherche et permet également de comparer les résultats issus de différents traitements. La comparaison des résultats provenant des différentes méthodes permet aussi d'avoir une indication sur l'utilité et la validité de la lemmatisation dans les analyses statistiques.

¹⁵⁸ Cf. chapitre 5.2. "La distance lexicale".

“Plutôt que choisir définitivement, écrit André Salem¹⁵⁹, au début de l’expérience le type d’unités qui présidera à la segmentation du texte (formes graphiques, segments répétés, lemmes etc.) il est préférable d’observer simultanément le comportement statistique de plusieurs types d’unités rendues disponibles par l’apparition des catégoriseurs automatiques.”

Il nous a également semblé intéressant d’adopter une telle attitude devant l’étude de ce vaste corpus qu’est l’œuvre de J.M.G. Le Clézio, et de ne pas trancher entre une méthode ou une autre, en ce qui concerne la lemmatisation et l’étiquetage des données, mais au contraire de comparer les différentes techniques offertes par des logiciels de plus en plus performants, afin d’enrichir les résultats d’une analyse qui portera en premier lieu, comme souvent dans les études lexicométriques, sur la structure du vocabulaire.

Nous avons choisi d’appliquer ces méthodes et ces divers programmes informatiques dans les différentes analyses qui constituent notre travail, afin d’éviter une certaine “lourdeur” de lecture. La comparaison des résultats issus des différentes techniques sera donc présentée dans les différentes parties de notre étude au fur et à mesure de sa progression.

¹⁵⁹ A. Salem (1995) : “Les unités lexicométriques.”, in *JADT 1995, III Giornate internazionali di Analisi Statistica dei Dati Testuali*, S. Bolasco, L. Lebart, A. Salem (éds). Rome, Università degli Studi di Roma, La Sapienza3 – Università degli Studi di Salerno, CISU, p. 25.