Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus

Margareta Kastberg Sjöblom

Bases, Corpus et Langage (CNRS, ILF) – UFR Lettres, Arts et Sciences Humaines – 98, bd. E. Herriot – B.P. 209 – 06204 Nice Cedex 3 – France

Abstract

The purpose of this paper is to outline some of the possibilities offered by using different tools and methods in stylo-statistical analysis. Different strategies are adopted in the statistical investigation of a corpus consisting of 990.000 tokens obtained from 12 novels of the French author J.M.G. Le Clézio. The investigation is made, on one hand, exploring not normalized data, on the other hand, by exploring normalized and tagged textual data, according to two different lemmatizing methods. These techniques make morphology and syntax investigations relevant and useful. Different technologies to study vocabulary variation and growth are also explored.

Résumé

Cet article s'intéresse à quelques résultats issus de différentes méthodes, en soumettant notre corpus Le Clézio à plusieurs traitements statistiques. Nous disposons d'une première version du corpus basé sur les formes graphiques et de deux autres versions de ce même corpus lemmatisées selon deux méthodes différentes. Nous tentons de mettre en parallèle la forme, le lemme et le code, pour pouvoir comparer les résultats. Nous nous intéressons à l'étude des parties du discours avec l'aide des codes grammaticaux, puis à l'accroissement lexical.

Mots clés : Lexicométrie, Lemmatisation, Lemmatiseurs, Accroissement lexical, Le Clézio, Le nouveau roman.

1. Introduction

Depuis les premiers travaux de lexicométrie et les progrès de l'informatique, les corpus devenant de plus en plus grands, le traitement statistique a cessé d'être manuel. Le débat de la lemmatisation, celle-ci difficilement standardisée par sa nature, s'est alors installé et aujourd'hui encore, le choix de méthode est controversé.

Nous nous proposons ici d'étudier quelques résultats issus de différentes méthodes (basées sur les formes graphiques ou sur les lemmes) en soumettant notre corpus Le Clézio à plusieurs traitements statistiques, et de mettre en parallèle la forme, le lemme et le code, pour pouvoir comparer les résultats. En premier lieu, nous nous intéressons à l'étude des parties du discours avec l'aide des codes grammaticaux, puis à l'accroissement lexical.

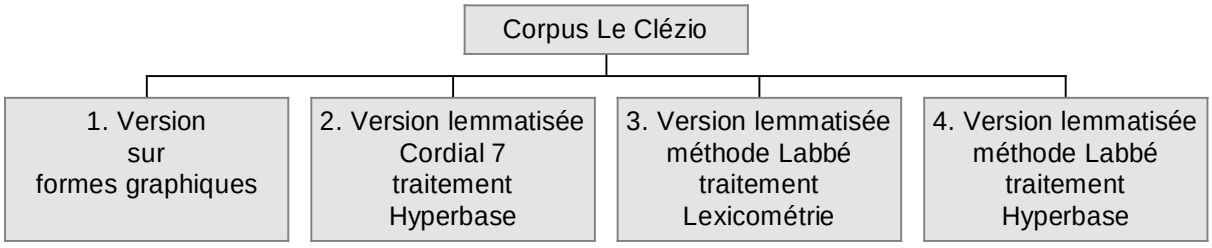
À cet effet, nous disposons de plusieurs versions du même corpus et de logiciels qui permettent la comparaison des résultats relatifs aux mêmes textes et aux mêmes traitements.

2. Le corpus

Le sous-corpus utilisé dans cet article est extrait de notre corpus exhaustif de l'œuvre de Jean-Marie Gustave Le Clézio (annexe 1 et Kastberg, Brunet 2000). Il est constitué de douze

oeuvres de l'auteur et il représente environ un tiers de sa production. Il s'étend sur 36 ans, de 1963 à 1999, et recouvre différents genres littéraires. Nous y trouvons son premier livre publié, *Le procès-verbal (1963)*, puis le recueil de nouvelles *La fièvre (1964*) puis *Mydriase (1973)*, que l'on pourrait qualifier de récit poétique. Le roman *Voyages de l'autre côté (1975*) est suivi par le recueil *Mondo et autres nouvelles (1978)* et le grand roman *Désert (1978).* Dans les années 80 paraît le recueil de nouvelles *La ronde et autres faits divers(1982)*, le roman écrit sous forme de journal de bord *Rodrigues (1986)*, ainsi que le recueil *Printemps et autres saisons (1989). Pawana (1992)* paraît dans la collection "enfant et jeunesse". Les derniers livres du corpus sont *La Quarantaine (1995)*, le long roman inspiré de l'histoire des ancêtres mauriciens de Le Clézio, et *Hasard (1999)* qui est le dernier roman de l'écrivain.

Les douze ouvrages du corpus totalisent 978.456 occurrences et 29.314 formes différentes. Nous disposons d'une part d'un dépouillement du corpus basé sur les formes graphiques, d'autre part de deux autres versions lemmatisées selon deux méthodes différentes. La première d'après les méthodes de Dominique Labbé (Labbé 1990b) et la seconde d'après le programme Cordial 7 Pour les traitements et l'exploitation statistique des données, nous avons eu recours à deux logiciels différents, le logiciel Hyperbase d'Étienne Brunet (Brunet 2000 et 2001) dans sa version 5.1 qui opère sur la forme graphique et dans sa version lemmatisée 5.3 qui s'appuie sur le programme Cordial, ainsi que le logiciel Lexicométrie de Dominique Labbé (Picard, Pibarot, Labbé 1995). Une version du corpus lemmatisé selon la méthode de Dominique Labbé a également été intégrée dans le logiciel Hyperbase par les soins d'Étienne Brunet.



Ces quatre versions ouvrent la voie à des recherches différentes et permettent de comparer les résultats. Les versions lemmatisées du corpus donnent notamment la possibilité de faire des études sur la morphologie et sur la syntaxe, ce qui n'est guère envisageable sur les formes graphiques. 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 aux codes grammaticaux et au lemme lui-même.

3. Les parties du discours

Au cours de l'opération de lemmatisation les catégories grammaticales sont fournies. Ces codes constituent un outil indispensable pour étudier les parties du discours et regrouper les différentes catégories grammaticales.

La version lemmatisée selon la méthode de Dominique Labbé trie les occurrences en une vingtaine de catégories fondamentales. Le logiciel Hyperbase regroupe, par la suite, les codes et fournit la liste des fréquences. Cette liste permet de voir la distribution des catégories grammaticales principales dans le corpus. Pour une vision plus synthétique, nous avons

recours à l'analyse factorielle :

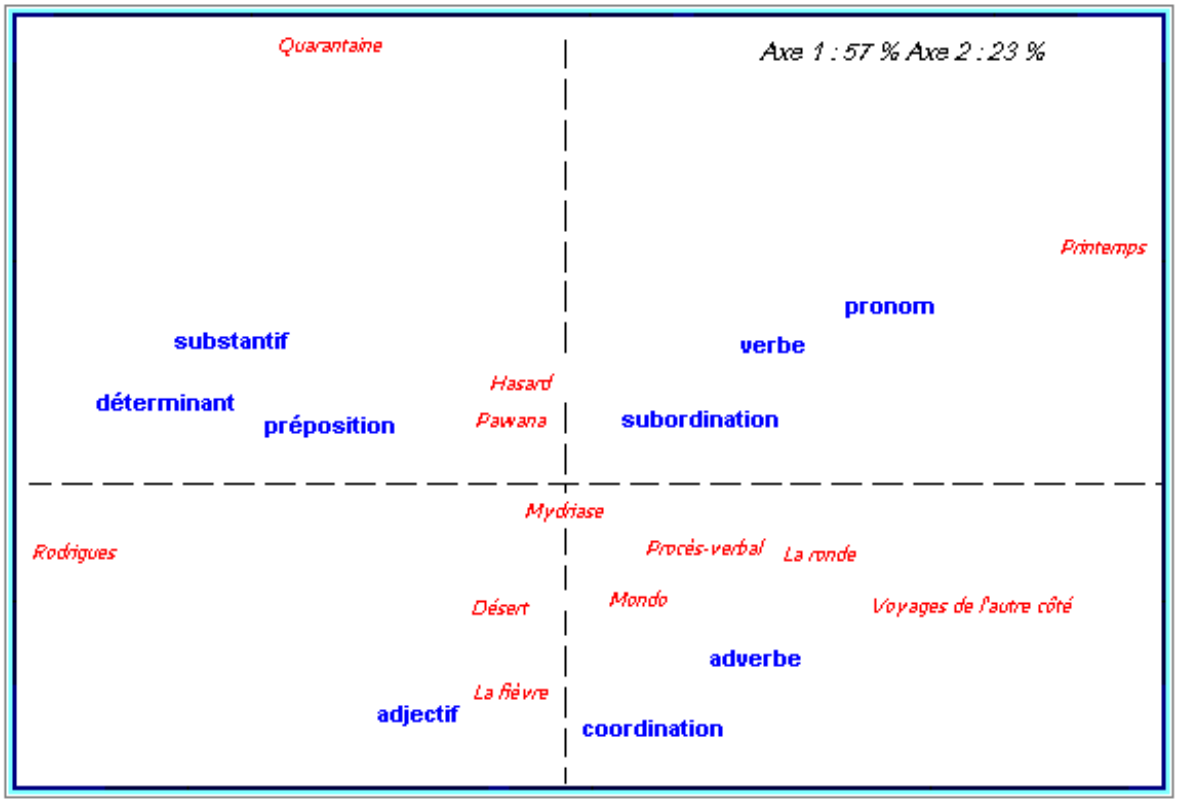


Figure 1. Analyse factorielle de la distribution grammaticale selon la lemmatisation Labbé

Le premier facteur oppose la catégorie des verbes à la catégorie nominale. Le substantif, en haut à gauche, attire les prépositions, les déterminants et les adjectifs, tandis que le verbe, en haut à droite, attire les pronoms, les subordonnées et les adverbes. Le second facteur parcourt la chronologie de l'écrivain, du bas vers le haut du graphique.

Le programme Cordial propose quelques 200 codes grammaticaux différents en utilisant toutes les combinaisons. Nous en avons extrait les 11 catégories fondamentales que propose le programme ; verbes, substantifs, adjectifs, déterminants, pronoms, numéraux, interjections, prépositions, adverbes, conjonctions et délimiteurs (signes de ponctuations), sans tenir compte du genre ou du nombre. L'analyse factorielle de la liste de fréquences de ces catégories permet de voir la distribution des catégories grammaticales dans le corpus.

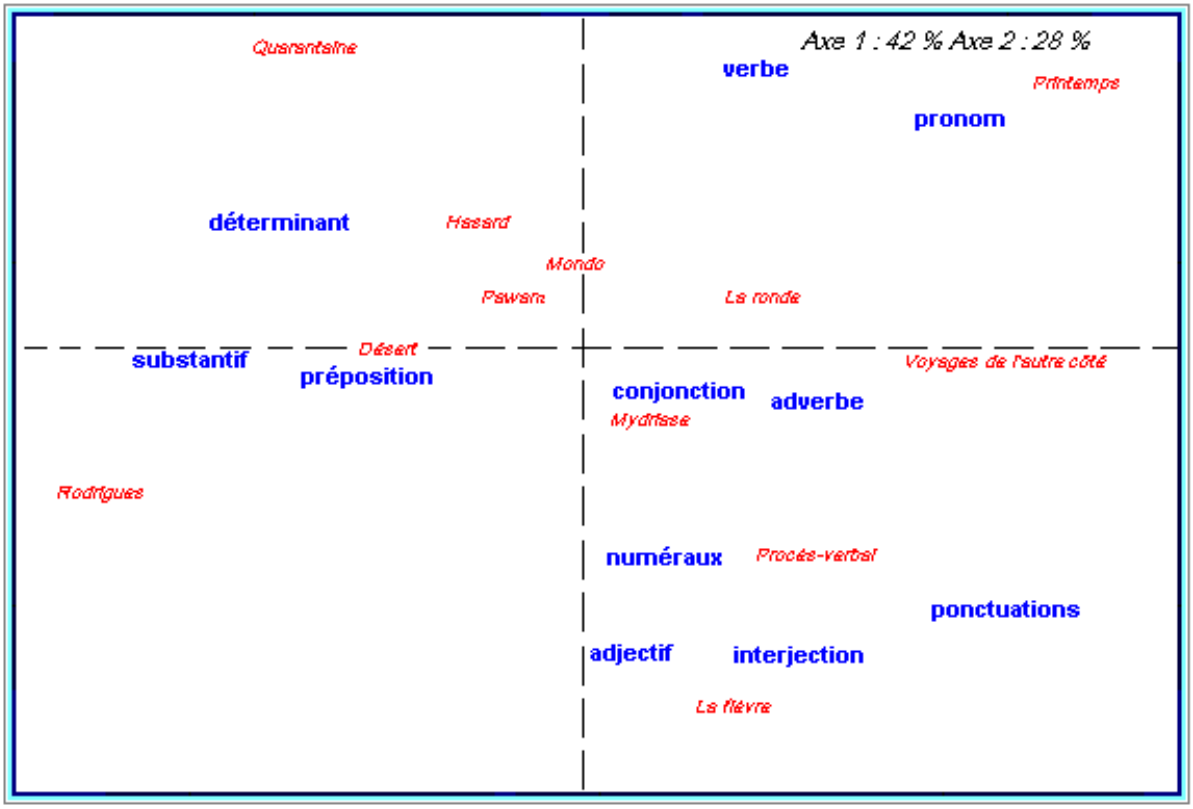


Figure 2. Analyse factorielle de la distribution grammaticale selon la lemmatisation de Cordial 7

Bien que les critères soient un peu différents, nous constatons la même opposition entre le verbe et le substantif avec les autres catégories qui gravitent entre ces deux pôles, ainsi que l'évolution dans le temps de l'auteur.

L'analyse factorielle permet également de constater le regroupement selon les genres littéraires. Les ouvrages "nouveau roman" se regroupent en bas à droite avec les adjectifs, les interjections, les numéraux et les signes de ponctuations. Le substantif sollicite les romans plus traditionnels tandis que le verbe attire les recueils de nouvelles. Le seul ouvrage sous forme de journal, *Voyage à Rodrigues*, se trouve isolé des autres.

Si nous regardons de près une catégorie grammaticale, comme par exemple celle des verbes, l'histogramme ci-dessous permet de mieux évaluer, ce que l'analyse factorielle avait déjà montré, la distribution qui n'est pas stable à travers le corpus.

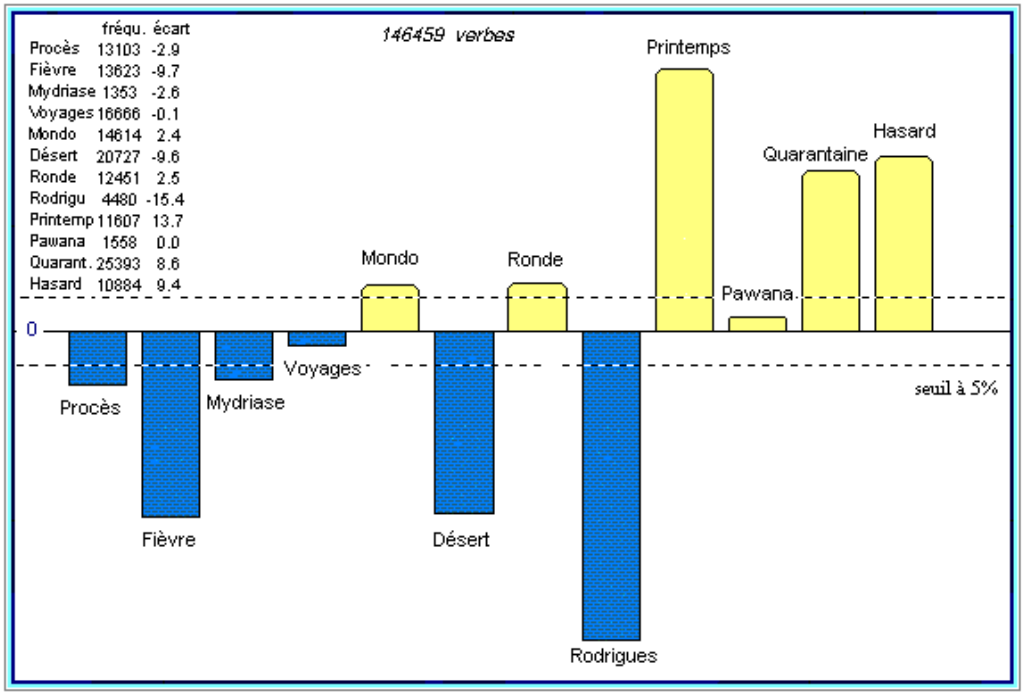


Figure 3. La distribution de la catégorie de verbes dans le corpus

La tendance générale, qui s'oriente de gauche à droite selon la chronologie, est croissante, avec des ruptures importantes qui correspondent aux nouvelles de *La fièvre*, dans lesquelles l'auteur expérimente des procédés pour mettre à jour les sensations, et aux romans *Désert* et *Rodrigues*. Ce dernier, écrit sous forme de journal de bord d'un marin, est constitué par des notes prises par le narrateur, ce qui explique son caractère particulier (ce livre est très riche en substantifs). La rupture au niveau de *Désert* et l'augmentation des verbes vers la fin du corpus rejoint la tendance générale du vocabulaire de Le Clézio.

En effet, la distribution de verbes dans notre corpus reflète souvent l'évolution du tout le vocabulaire. L'étude de l'accroissement lexical permet de mieux cerner l'évolution, avec ses variations et ses ruptures caractéristiques de l'écriture leclézienne.

4. L'accroissement du vocabulaire de Le Clézio

L'accroissement lexical est, pour un segment déterminé du texte, le nombre d'unités nouvelles qui apparaissent dans ce segment. Le calcul de l'accroissement pourrait être celui de Charles Muller qui a été le premier à proposer un modèle d'accroissement du vocabulaire (Muller 1979). Son calcul repose sur une "réduction" de la distribution des fréquences observée sur l'ensemble du corpus, à la taille de la tranche. C'est ce modèle que suit Dominique Labbé et que nous appliquerons aussi dans une deuxième approche.

Mais dans un premier temps, nous observons l'accroissement par un simple ajustement de courbe en choisissant pour jalons les césures naturelles du corpus. Ce sont les textes eux-mêmes qui constituent les tranches et les décomptes sont établis, par le logiciel Hyperbase, à partir des formes graphiques, chaque fois que l'on passe d'un texte à un autre. Le tableau ci-dessous rend compte de l'accroissement du vocabulaire dans l'ordre chronologique. Ici le calcul fait appel à un ajustement des deux séries parallèles (vocabulaire cumulé et étendue

cumulée) grâce à une fonction puissance. L'écart entre étendue théorique et étendue réelle est alors calculé pour chaque texte, puis pondéré :

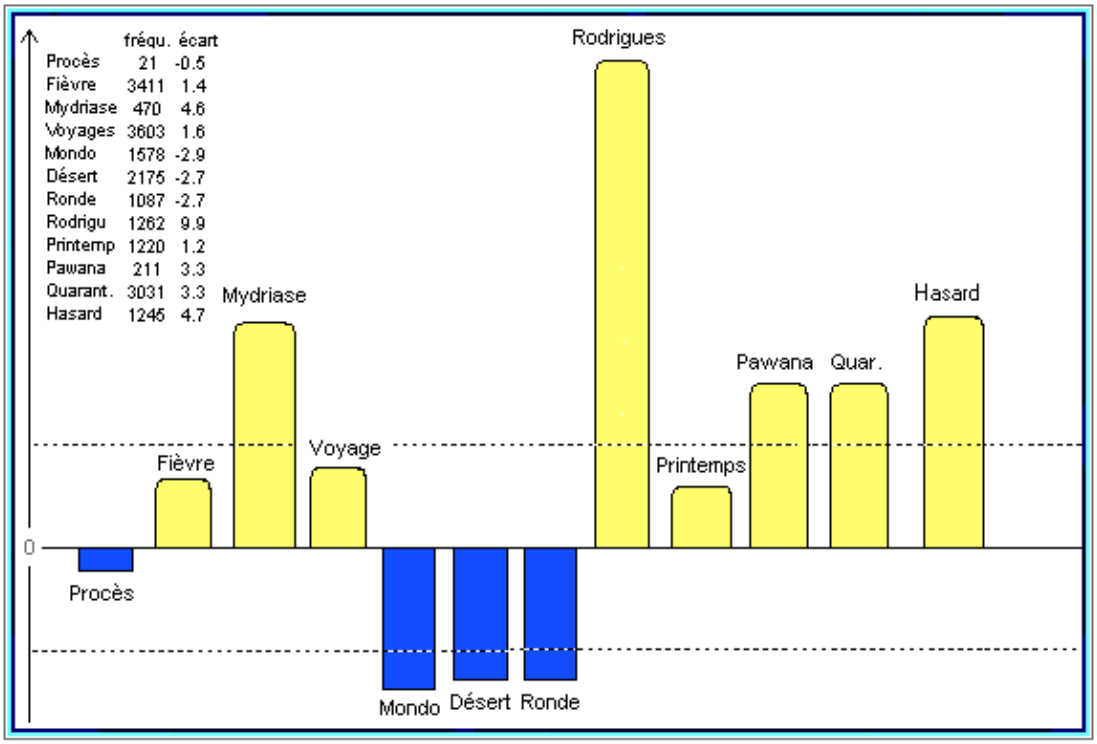


Figure 4. L'accroissement lexical calculé sur les formes

Les écarts autour de la moyenne, l'axe horizontal, sont de grande ampleur, avec des ruptures et des reprises comme l'indique le graphique qui, de gauche à droite, s'oriente selon la chronologie. Le vocabulaire s'accroît dès le début de la production, avec un apport assez important dans les premiers ouvrages, période classée "Nouveau Roman" par les critiques[1]. Par la suite, cet apport ralentit.

La chute la plus importante d'apport lexical arrive avec *Mondo* qui correspond à une rupture nette dans l'écriture et à un tournant chez Le Clézio. Ce recueil de nouvelles est en effet dépourvu de tout exotisme. La tendance littéraire de cette époque, la fin des années 70, vise le dépouillement de l'écriture avec le renoncement aux conventions romanesques et au pittoresque, et le discours exotique n'est plus en faveur.

"L'un des dangers de ce type d'écriture est de ne pas pouvoir se renouveler d'une œuvre à l'autre", écrit Michelle Labbé (Labbé M. 1999). "Ne risque-t-elle pas de s'enfermer dans cette alternative. Se taire ou de se répéter ?" Nous ne pouvons que confirmer ce propos, grâce aux statistiques.

Avec *Voyage à Rodrigues*, un apport lexical considérable est à noter. Nous découvrons un nouvel univers, celui du navigateur, du chercheur de trésor sur l'Océan Indien. À partir de ce

[1] Le Clézio lui même ne tient pas à voir son oeuvre s'inscrire dans une typologie trop précise et son attitude consiste à brouiller la division en genres.

moment, Le Clézio semble revenir dans ses derniers ouvrages, comme d'autres romanciers, à une conception plus conventionnelle du roman, avec la volonté de retrouver une certaine illusion réaliste, l'étude des verbes dans la partie précédente l'indiquait également.

En général, on observe chez les écrivains[2] un accroissement important du vocabulaire dans le début de la carrière littéraire qui diminue vers la fin de l'œuvre. Ce décroissement marque la tendance vers une écriture plus sobre à la fin d'une carrière littéraire. C'est, par exemple, le cas de l'œuvre de Zola où le renouvellement du vocabulaire s'amenuise vers la fin de sa production. Par contre, ce n'est pas le cas de Le Clézio, chez qui nous observons un vocabulaire croissant vers la fin[3], une tendance qui est assez exceptionnelle.

Mais le dépouillement sur les formes graphiques n'introduit-il pas un biais ? Pour le savoir nous faisons le même calcul sur les lemmes.

Les graphiques ci-dessous permet de comparer les deux versions lemmatisées.

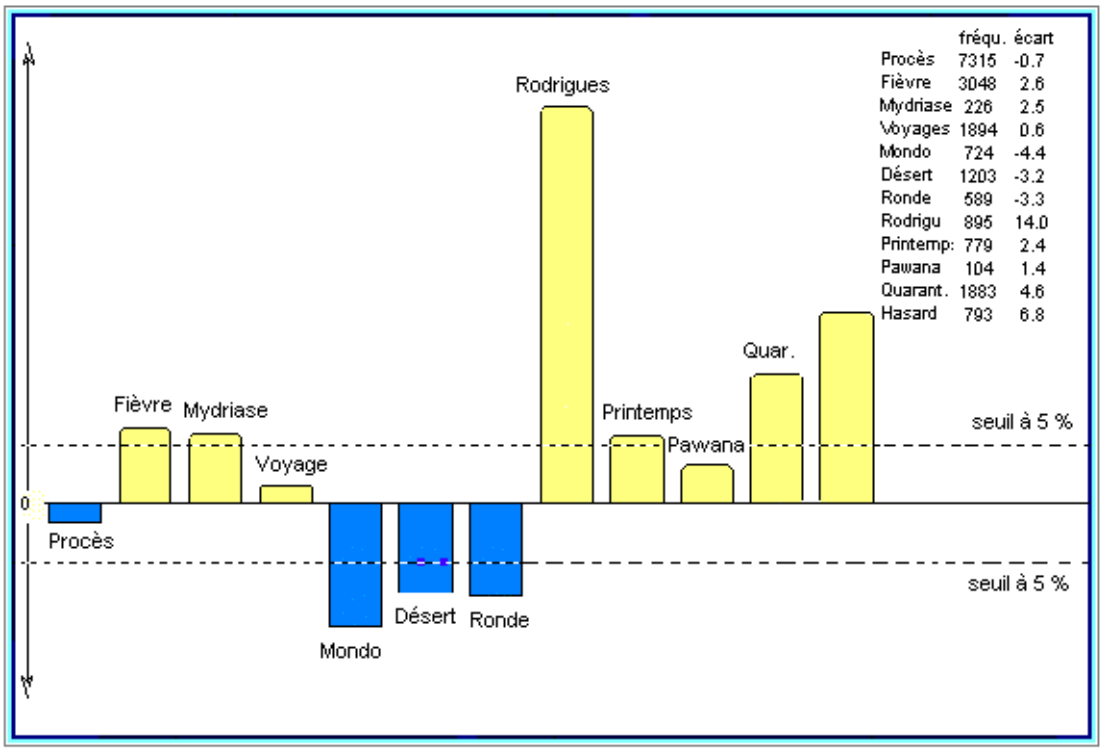


Figure 5. L'accroissement lexical calculé par Hyperbase sur les lemmes selon Cordial

[2] É. Brunet a montré ce phénomène dans ses ouvrages sur les écrivains du XIXe siècle.
[3] Dans notre cas, il ne s'agit pas d'une fin de carrière, puisque l'auteur est toujours vivant et très productif.

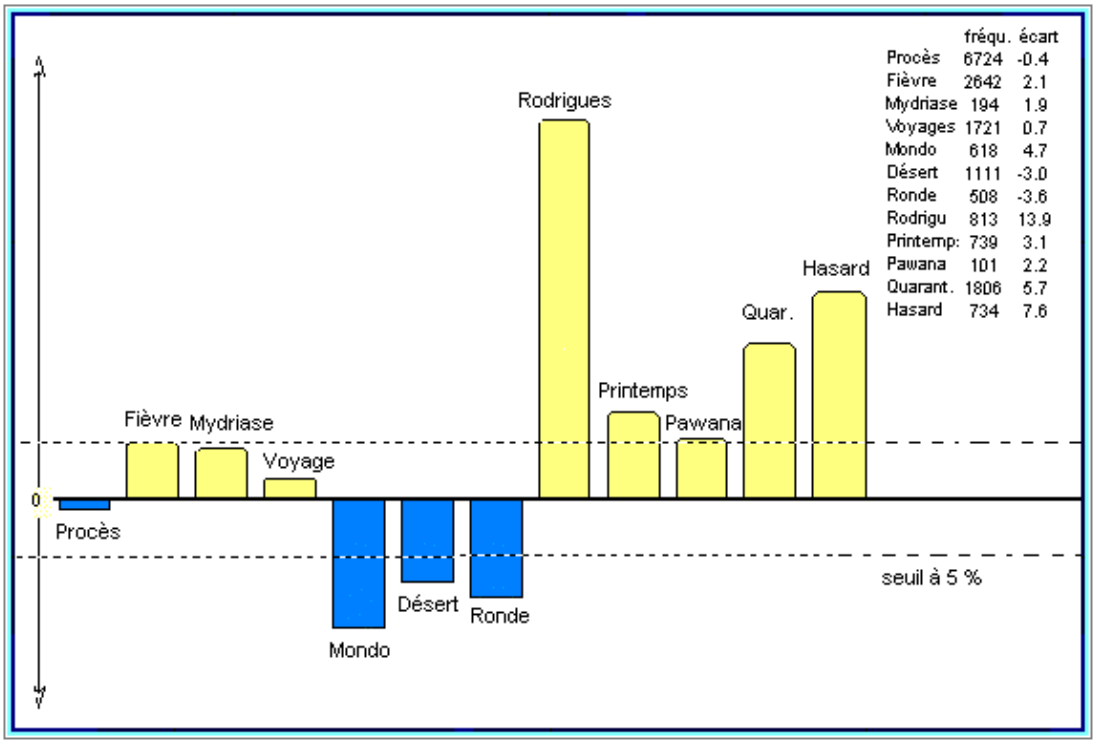


Figure 6. L'accroissement lexical calculé par Hyperbase sur les lemmes selon la méthode Labbé

Les deux graphiques sont très semblables. Dans la première moitié, les écarts sont légèrement plus grands lorsque les calculs sont basés sur le corpus lemmatisé par Cordial ; en revanche, d'après la méthode Labbé, c'est à partir de *Mondo* que les écarts sont légèrement plus grands qu'avec Cordial.

Si nous comparons ces deux graphiques avec celui que nous avons vu auparavant, nous constatons que les trois tableaux sont presque superposables : le même calcul aboutit, pour ces trois versions (la première basée sur les formes et les deux autres sur les lemmes), à des résultats quasi identiques.

Le logiciel "Lexicométrie" mesure l'accroissement lexical avec une technique un peu différente. Pour effectuer la mesure, cette fois-ci, on découpe le corpus en tranches de taille égale, c'est-à-dire en fragments comportant le même nombre de mots (1.000 mots). Cette technique permet de localiser les afflux de vocables nouveaux qui correspondent aux principales ruptures thématiques dans le corpus en dehors des césures naturelles, c'est-à-dire les différents livres. Ce modèle de partition permet de calculer le nombre attendu de vocables employés depuis le début du corpus en différents points de celui-ci dans l'hypothèse d'un accroissement régulier du vocabulaire, au fur et à mesure de l'augmentation de la taille, et en tenant compte de la spécialisation du vocabulaire (Hubert, Labbé. 1995).

La technique permet, comme avec Hyperbase, de localiser les ruptures thématiques dans le corpus, là où se produit un afflux de vocables nouveaux. En sens inverse, les fragments où l'accroissement est inférieur aux valeurs théoriques signalent l'épuisement d'un thème.

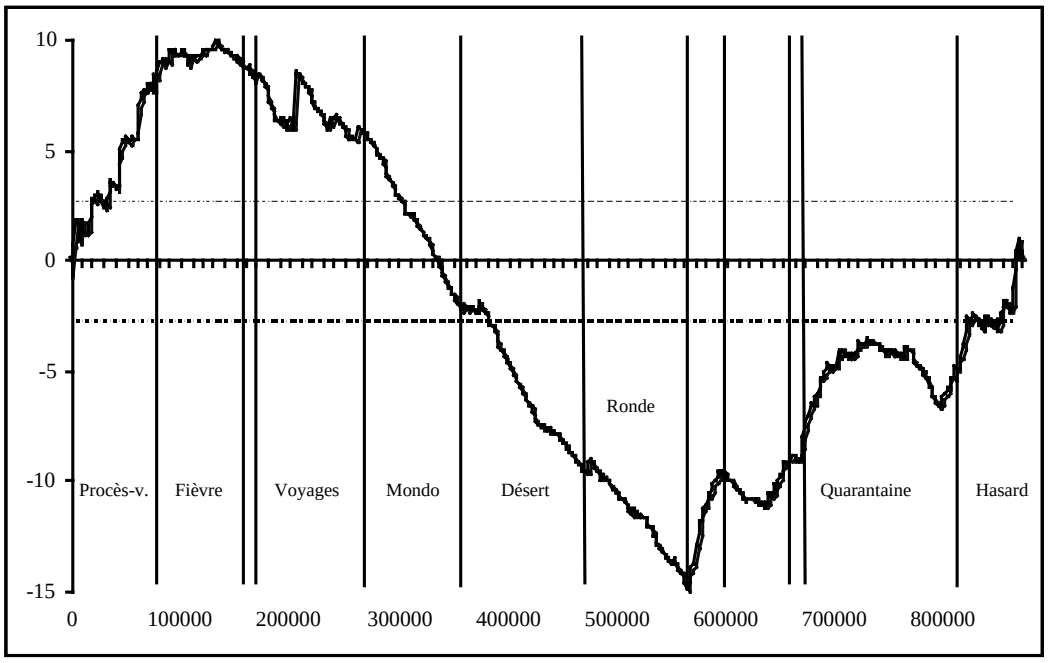


Figure 7. Accroissement lexical par le logiciel Lexicométrie sur les lemmes selon la méthode Labbé

La courbe qui s'inscrit autour de l'axe horizontal rend compte des très grands écarts autour de la tendance moyenne. Les fluctuations dépassent souvent les bornes de variations "normales" (±2 écarts types).

La courbe fait apparaître trois périodes principales dans la production leclézienne. Après un accroissement important de vocabulaire dans le *Procès-verbal*, le vocabulaire se maintient à un palier élevé, jusqu'à *Voyages de l'autre côté,* où commence la chute de la courbe. De *Voyages* à *Ronde* le déclin continue, excepté deux périodes très brèves correspondant aux premières pages de *Désert*, c'est-à-dire à la découverte de l'univers des hommes bleus du désert nord-africain, et à la première nouvelle de la *Ronde*. De *Rodrigues* à la fin, chaque œuvre représente un nouvel apport lexical. Chaque livre correspond à un nouvel univers exploré par l'auteur. Une exception peut, cependant, être constatée, ce sont les trois premières nouvelles de *Printemps* pour lesquelles le creux dans la courbe signale qu'elles appartiennent à la période précédente.

Cette technique permet également d'étudier le profil de chaque œuvre. On retrouve le profil très caractéristique de *La quarantaine* avec un fort apport, dans le premier quart, le *Hasard* présente, quand à lui, une césure vers le milieu. Dans *Voyages de l'autre côté,* l'accroissement est assez spectaculaire.

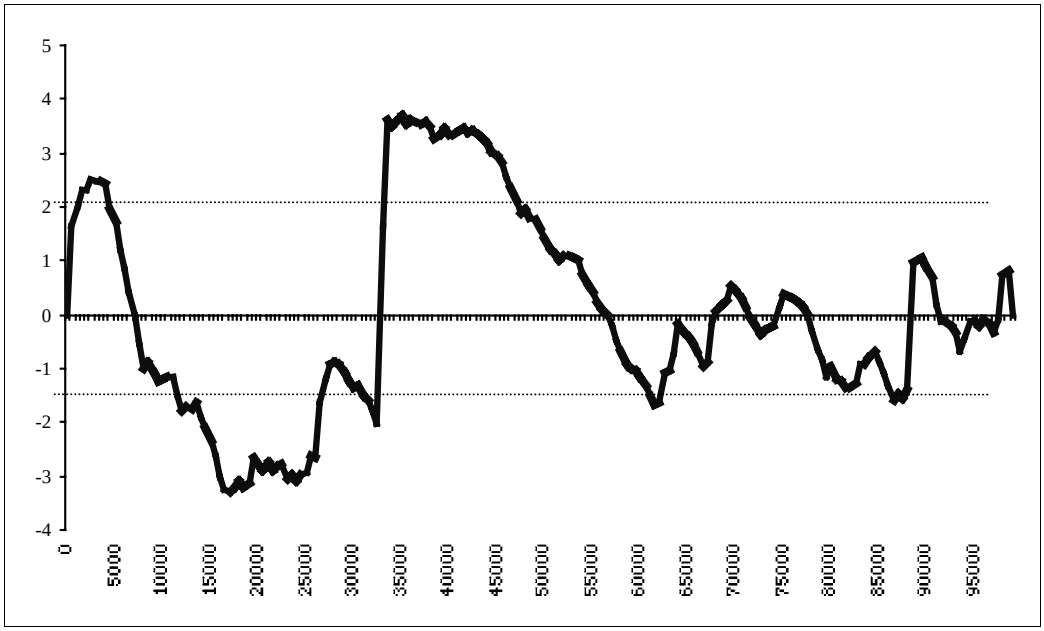


Figure 8. Profil de l'accroissement lexical dans Voyages de l'autre côté

On remarque un apport initial suivi par une chute importante, puis un pic correspondant sensiblement au milieu de *Voyages* (entre 30.000 et 35.000 mots). En moins de 1000 mots, la courbe s'élève de +4,5 écarts types. Le caractère abrupt est absolument unique. Il s'agit, en fait, d'un trait appartenant à l'écriture "Nouveau Roman" avec une longue énumération de noms propres ou d'objets (annexe 2).

Voyages de l'autre côté considéré comme l'œuvre qui annonce le tournant dans l'écriture de Le Clézio, montre bien jusqu'à quel point ces calculs sont sensibles au genre.

5. Conclusion

En conclusion, nous avons vu à travers cette étude que les différentes méthodes donnent des résultats souvent assez semblables, comme si l'étiquetage était relativement neutre et sans influence.

Toutefois, chacun des deux types de corpus, basé sur les formes graphiques ou sur les lemmes, permet une analyse différente et complémentaire. Les corpus lemmatisés offrent la possibilité d'analyses morphologiques et syntaxiques, impossibles avec des grands corpus non lemmatisés, ceux-ci en revanche permettent la comparaison des travaux de différents chercheurs sur un même ensemble d'œuvres.

De la même manière que les méthodes de travail sont complémentaires, l'exploitation des deux logiciels permet un travail complet à partir d'un corpus donné.

Références

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

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

Brunet É. (2000). *Hyperbase, Manuel de référence, version 5.0*, CNRS-INaLF, UPRESA "Bases,

corpus et langage" et sa mise à jour du janvier 2001.

Hubert P., Labbé D. (1995). Note sur l'approximation de la loi hypergéométrique par la formule de Muller et Un modèle de partition du vocabulaire. In Labbé D., Thoiron Ph., Serant D. éditeurs, *Etudes sur la richesse et la structure lexicale.* Slatkine-Champion, pages 77-114.

Kastberg, Brunet (2000). La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain", in Rajman M., Chappelier J.-C.éditeurs, *JADT 2000 Actes des 5es journées internationales d'Analyse statistique des Données Textuelles*, Lausanne.

Labbé D. (1990a). *Le vocabulaire de François Mitterand*, Presse de la Fondation Nationale des Sciences Politiques.

Labbé D. (1990b). *Normes de dépouillement et procédures d'analyse des textes politiques*, CERAT.

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

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Muller Ch. (1964). Calcul des probabilités et calcul d'un vocabulaire. Reproduit dans *Langue française et linguistique quantitative* (1979), pages 167-176.

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

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

Annexe 1 : Le corpus

N°	TITRE	OCCURRENCES	FORMES
1	Procès	93025	10021
2	Fièvre	75093	6612
3	Mydriase	9863	1875
4	voyage	113820	8368
5	Mondo	99131	6617
6	Désert	150894	8337
7	Ronde	83442	6425
8	Rodrigues	38340	4889
9	Printemps	70246	6136
10	Pawana	10616	1903
11	Quarantaine	165651	11173
12	Hasard	68335	7233
TOTAL		978456	29314

Annexe 2 : le passage le plus divers de tout le corpus

Voyages de l'autre côté p. 108

C'est Naja Naja qui nous donne tous nos noms. Sans elle peut-être que nous n'existerions pas, peut-être que nous ne saurions jamais qui nous sommes. Voici comment elle nous appelle : Longy Bay, Calopin, Tremblement de terre, Caoudal, Blancs-Manteaux, Gros Fumeur, Night Blue (ça c'est un joli nom), Splendid 66 (c'est Louise quand elle suit ses cours de dactylo). Quand elle est triste : Vizeacha. Une Seule Flèche, Dugong, Kit-e-cat, Baja California, Djakarta, Gavial, Paravent, Water Lily, Charron, Big Game, Mondo, Mansarde, Herbe Folle, Térébenthine, Juilen, Quarante-sept Touches, Doigts-de-fée, Carbone, Tséno, Gecko, Lars, Leroux, Écho, Kissang. Chamelon, Bolet Satan, Russule de Quélet, Ambroise, Fortune, Tournepointe, Palava, Don, Sépulcre, Marcomans, Comores, Passim, Sarigue, Koala, Carte Postale (si tu lui en envoies une). Racine, Bosse, Lipo, Caravane, Lieudit, hausse-col, hausse-pied, haut-de-forme, Indochine, Wagon-lit, Tangeers, Miel, Boule de Neige, Gaur. Patent, Cigarette-Bill. Python Molure (quand elle vous aime bien). Pygargue, Séquoia, Chaise longue, Chiltosic, Maureau, Heaume, Vincent Guerre, Glenn, Auvergne, Chanvre, Houilles-Forges, Nil, Bazooka, Citerne, Carotte, Petit Go, Lune, Apennine, Ulme, Énoch, Berg-op-Zoom. Petite Croix.
…..
Ce sont les noms que nous donne Naja Naja.