Université de Toulouse

THESE

en vue de l'obtention du
DOCTORAT DE L'UNIVERSITE DE TOULOUSE

Délivré par : Université de Toulouse Le Mirail
Discipline : lettres modernes
Spécialité : stylistique statistique

La musique des lettres

Variations sur Yourcenar, Tournier et Le Clézio

Thèse soutenue par Stephan Vonfelt

le 15 décembre 2008

JURY
Etienne Brunet, Professeur émérite à l'Université de Nice Sophia Antipolis
Louis Ferré, Professeur à l'Université de Toulouse Le Mirail
François-Charles Gaudard, Professeur à l'Université de Toulouse Le Mirail
Sylvie Mellet, Directeur de recherche au CNRS de Nice Sophia Antipolis
Georges Molinié, Professeur à l'Université de Paris Sorbonne

Ecole doctorale : ALLPH@
Unité de recherche : LLA
Directeur de thèse : François-Charles Gaudard

VOLUME 1

Edition d'origine : octobre 2008
Edition revue : janvier 2009

A mon père, ma mère, mes frères-chiens

Remerciements

A Damon Mayaffre : son livre, *Paroles de président*, m'a incité à approfondir la stylométrie.

Au professeur François-Charles Gaudard, qui a su accueillir un éléphant parmi la porcelaine. Son ouverture d'esprit, sa confiance et son intérêt ont été les carburants essentiels de cette aventure.

A Didier Bourrigault et Patrick Séguéla : leurs logiciels *Syntex* et *Cordial* ont enrichi les données linguistiques.

Aux professeurs Henri Caussinus et Louis Ferré, qui ont vérifié les orientations statistiques de ce travail.

A Pascal Daniel, pour ses lumières informatiques.

« Toute chose est nombre »
Pythagore

Lecteur ami
Ne crains pas cette formule
Elle n'est faite que de lettres
Si tu me fais confiance
Je t'invite à un voyage
Je t'emmène vers des rivages étranges
Au pays des chimères romantiques
Tu verras le peuple du symbole
Métaphores et équations
S'y mêlent en silence
Derrière un voile ambigu
Et quand ton corps sera las
Ferme les yeux écoute ton cœur
Il est plus grave que les mots avenirs

Préambule

Après cet exergue éthéré, je vais préciser ce qui symboliquement et pragmatiquement m'a amené à la stylistique.

Initialement ingénieur, j'ai fait une incursion dans le monde de l'information. En deux mots, je suis passé de l'objet à sa représentation. La lecture du triangle sémiotique m'a été fatale : il manquait une pierre à mon édifice. Au sommet de la pyramide, dans une représentation au carré, se trouvait l'œil de l'information, la linguistique. Un degré plus haut vers la subjectivité, la stylistique littéraire m'est apparue.

Pour le salut de mon âme, j'ai su refouler jusqu'ici les assauts du Carré et du Pentagone. Je prie tous les jours Sainte Amnésie de me maintenir en cette heureuse innocence.

Résumé

L'objectif de ce travail est d'analyser et de synthétiser un corpus littéraire à l'aide de statistiques.

Classiquement, les fréquences des unités linguistiques indiquent la composition d'un texte ou son « thème ». D'inspiration stylistique et musicale, cette thèse propose de mesurer la rareté à la place de l'abondance, et de prendre en compte l'organisation des unités par leur rythme.

Au sein d'un texte, les temps de retour d'une unité sont quasiment décorrélés. Ils se caractérisent par leur distribution en forme de cloche asymétrique, linéarisable avec un conditionnement par le passé. La répartition qui lisse ce spectre se mue alors en pierre de touche.

Comparant deux textes, la distance généralisée mesure les écarts entre les répartitions. Dans l'ensemble, elle suit les évolutions de sa version classique fondée sur les fréquences, mais des divergences significatives apparaissent localement selon l'intensité de l'arythmie.

Le corpus comprend trois romans du 20e siècle écrits par Yourcenar, Tournier et Le Clézio : *Mémoires d'Hadrien*, *Vendredi ou les limbes du Pacifique* et *Désert*. Les mesures linguistiques portent parallèlement sur les plans graphémologiques, syntaxiques et sémantiques.

Globalement, ces plans se répondent et semblent obéir profondément aux mêmes lois linguistiques. Les graphèmes peuvent être privilégiés pour leur objectivité et leur abondance.

Stylistiquement, l'intuition littéraire est confirmée par les mesures, qui montrent une gradation entre les œuvres en suivant leur chronologie. Leurs divisions forment des ensembles homogènes au sein du corpus, si bien qu'un style se dégage et permet de simuler avec succès une attribution d'auteur.

Mots-clés : attribution d'auteur - classement - linguistique - littérature - rythme - statistique - style - stylistique - stylométrie.

Abstract

The purpose of this work is to analyse and synthetise a literary corpus with the use of statistics.

Traditionally, the frequencies of linguistic units indicate the composition of a text or its “theme”. This thesis, inspired by stylistics and music, proposes to measure rarity instead of abundance, and to consider the organisation of the units on the basis of their rhythm.

Within a text, the recurrence times of a unit are virtually decorrelated and characterised by their asymmetrical bell-shaped distribution, linearisable with a conditioning by the past. The cumulative distribution that smoothes out this spectrum thus becomes a touchstone.

Comparing two texts, the generalised distance measures the differences between the cumulative distributions. Taken overall, it follows the developments of its traditional version based on the frequencies, but significant discrepancies appear locally depending on the intensity of the arrhythmia.

The corpus consists of three 20th Century novels by Yourcenar, Tournier and Le Clézio : *Mémoires d’Hadrien*, *Vendredi ou les limbes du Pacifique* and *Désert*. The linguistic measurements are carried out simultaneously on the graphemological, syntactic and semantic planes.

Globally, these planes correlate and seem to be deeply in accordance with the same linguistic laws. The graphemes may be favoured because they are objective and abundant.

Stylistically, the literary intuition is confirmed by the measurements, which show a grading between the works following their chronology. Their divisions form homogeneous assemblies within the corpus, in such a way that a style appears and permits to succesfully simulate the attribution of an author.

Key words : author attribution - classification - linguistics - literature - rhythm - statistics - style - stylistics - stylometry.

Sommaire

VOLUME 1

VOLUME 2

Annexes

Introduction

On reconnaît qu'une œuvre a du style à ceci qu'elle donne la sensation du fermé ; on reconnaît qu'elle est située au petit choc qu'on en reçoit, ou encore à la marge qui l'entoure, à l'atmosphère spéciale où elle se meut[1].

Après une présentation générale, le contexte du travail est donné par un bref historique des disciplines en jeu. Puis sont précisées les orientations de la thèse et l'organisation du mémoire.

1 Présentation générale

1.1 Problématique

Le besoin d'ordonner semble la marque de l'esprit humain : citons Linné[2] pour sa classification des espèces, Harvard[3] pour celle des étoiles. Plus proche de notre domaine, la classification décimale de Dewey[4] organise le savoir au sein des bibliothèques. En matière artistique, le rôle échoit à la stylistique qui rapproche ou discerne les œuvres.

Si la subjectivité intervient toujours dans cette opération délicate, il est sans doute raisonnable de la pousser dans ses retranchements. L'approche suivie se fonde donc sur des unités tangibles et dénombrables, objets de mesures et de statistiques finalement

---

[1] Jacob, *Le cornet à dés*, préface de 1916.
[2] Linné, *Systema Naturae*.
[3] Draper, *The Draper Catalogue*.
[4] Dewey, *Classification and subject index for a library*.

interprétées par l'analyste[5].

Selon les œuvres et les unités, les volumes des données peuvent devenir considérables. La feuille de papier, la règle de calcul sont alors dépassées et rendent nécessaires l'intervention d'un nouvel acteur : l'ordinateur entre en scène et l'informatique donne la puissance aux principes de la stylométrie.

En d'autres termes, la stylistique est le compositeur, la statistique est le chef d'orchestre, et l'informatique est l'instrumentiste. Discipline mêlée, la stylométrie moderne s'inscrit dans un triangle cerné par ces éléments.

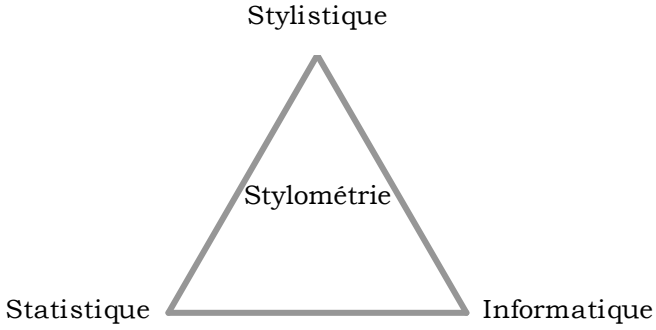


Figure 0.1 : contexte disciplinaire

1.2 Motivation de la recherche

Traditionnellement, la stylométrie effectue des comptages globaux sur les unités de différents plans linguistiques : graphèmes, catégories grammaticales, concepts… Ce faisant, on efface le souffle et le rythme de la littérature et l'on manque peut-être l'essentiel. Ce travail vise à compléter une approche classique et statique pour aborder les

[5] La subjectivité reste donc présente dans le choix des unités, de leur traitement et de leur interprétation.

composantes dynamiques du style[6].

2 Contexte historique

Les lignes qui suivent retracent les principaux développements des disciplines en jeu, la stylistique, la statistique et l'informatique. Pour chacune d'entre elles, un bref rappel des origines précède un état de l'art schématique.

2.1 La stylistique

2.1.1 De la rhétorique à la littérature[7]

Les fondements voient le jour avec la rhétorique antique : Aristote distingue le style périodique du style continu[8], puis Cicéron perfectionne la typologie en séparant les genres simples, tempérés ou sublimes d'un discours[9]. L'enjeu principal est pragmatique : il s'agit pour l'orateur de convaincre son auditoire par le jeu de trois registres : logique, esthétique et pathétique :

> Il n'y a que trois genres de style, et dans chacun des trois il s'est fait d'assez belles réputations [...) Dans le sublime, on voit des orateurs soutenir, par la majesté de l'expression, l'élévation de la pensée. Véhémence, variété, abondance, force, pouvoir de remuer les âmes, et de les pousser en tous sens, tels sont les caractères essentiels de ce genre [...) Le genre simple, au contraire, n'est que fin, et ne veut qu'instruire; il ne grossit pas les objets, mais il en éclaire toutes les faces [...] Entre ces deux genres, un troisième tient le milieu. Tempéré, par excellence, il amortit les foudres du premier et

[6] A une échelle de temps beaucoup plus large, Labov a étudié les variations historiques dans *Principles of Linguistic Change*.
[7] Karabetian, *Histoire des stylistiques*, chapitre 1.
[8] Aristote, *La Rhétorique*, Livre 3, chapitre 9.
[9] Cicéron, *Œuvres complètes,* « L'orateur » , VI..

> les traits du second [...] Toujours doux et coulant, ce style n'a, dit-on, d'autre caractère qu'une égalité soutenue[10].

Dans ce contexte, un discours fortement déterminé s'élabore en obéissant à des techniques éprouvées : l'*inventio* ou la recherche des idées ; la *dispositio* ou le plan en parties ; l'*elocutio* ou le choix des mots et de leur syntaxe ; la *memoria* ou l'apprentissage par cœur ; l'*actio* ou la performance oratoire[11].

Peu à peu, ce dispositif se démantèle et s'affaisse : avec le développement de l'écrit, *memoria* et *actio* passent au second plan. D'autre part, une argumentation devenue autonome se traite séparément, si bien que l'*elocutio* prédomine au sein de la rhétorique médiévale : cette composante est l'ancêtre de la stylistique.

Après la traduction latine de la *Poétique* d'Aristote en France, l'ornement prend pied au 17e siècle. Il ne s'agit plus uniquement d'argumenter, mais aussi de plaire par le déploiement d'une langue fleurie et colorée : le fondement logique du verbe glisse vers l'esthétique. Sur ce terreau propice, les « Belles Lettres » s'épanouissent.

Un siècle plus tard, l'esprit change : la littérature se fait le miroir de l'âme et des sentiments humains. Ce bouleversement romantique amène Condillac à intégrer l'affect dans son *Art d'écrire*[12] : les structures figées et universelles s'effacent pour laisser libre cours à l'expression individuelle. Fontanier[13] définit alors le style comme un écart vis-à-vis de « l'expression simple et commune ».

---

[10] Cicéron, *Œuvres complètes,* « L'orateur » , VI.
[11] Dans certains cas, les tâches se partagent : le scripteur prend en charge l'*inventio*, la *dispositio* et l'*elocutio*, tandis que l'orateur se concentre sur la *memoria* et l'*actio*.
[12] Condillac, *Œuvres complètes*, tome VII.
[13] Fontanier, *Les figures du discours*, p. 279.

2.1.2 Une discipline incertaine

Le terme « stylistique » est attribué à Novalis. Il entre en 1800 dans le *Deutsche Wörterbuch*, mais il faut attendre 1877 pour qu'il soit repris par le *Littré* français. Cette nouvelle science tente notamment d'établir un lien entre le style d'une œuvre et son contexte. A la frontière du monde intérieur et extérieur, elle fait communiquer la psychologie, la philosophie, l'histoire et la sociologie. Dans la sphère verbale, elle se tient entre l'art littéraire et la science linguistique. Insaisissable et centrale, voilà peut-être son âme.

2.1.2.1 Entre linguistique et littérature

Le partage des eaux entre linguistique et littérature[14] se révèle tumultueux au cours de l'histoire, dans un mouvement global du large vers l'étroit : surgi au 19e siècle avec Humboldt[15], le courant de la stylistique externe embrasse la diversité des langues nationales. A l'entrée du 20e siècle, la stylistique interne de Bally[16] étudie l'expression de l'affectivité dans une langue commune et spontanée, loin de toute intention esthétique et de tout choix conscient. Après la seconde guerre mondiale, la stylistique fonctionnelle de Larthomas[17] prend en compte le genre littéraire. Un pas de plus vers la subjectivité, la stylistique génétique de Spitzer[18] recherche l'âme du poète ou un étymon spirituel.

---

[14] Guiraud et Kuentz, *La Stylistique*, chapitre I, section A.
[15] Humboldt, *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts.*
[16] Bally, *Traité de stylistique française,* p. 16-21.
[17] Larthomas, *Le Français moderne*, p. 185-193.
[18] Spitzer, *Etudes de style*, p. 54.

2.1.2.2 Entre objet et sujet

Au sein de cette discipline, deux versants méthodologiques apparaissent. D'un côté, la stylistique descriptive de l'Ecole française se fonde prudemment sur les unités palpables du texte.

De l'autre, une stylistique subjective et subtile se subdivise elle-même en deux, selon que l'on porte l'attention sur l'émetteur ou le récepteur. Dans le premier cas, la psychologie linguistique allemande fait résonner la pensée de l'auteur et son langage[19]. Dans le second, Riffaterre[20] ou Jauss[21] mettent en relation le texte et le lecteur.

2.2 La statistique

2.2.1 Du recensement au sondage[22]

Les premiers dénombrements remontent à la haute antiquité : des listes de personnes et de biens apparaissent à Sumer 5000 ans avant notre ère, puis en Egypte 2900 ans avant J.C. Dès l'origine, des sentiments ambivalents accompagnent cette démarche : menace sur le secret de la vie et de la création, le recensement est sacrilège dans la tradition d'Israël et ne se justifie que si Dieu l'ordonne. En Grèce, il répond de façon plus neutre au précepte humaniste « connais-toi toi-même ». A Rome, le cens est d'abord l'instrument du pouvoir et de la domination de l'Empire.

Le Moyen-Age met en sommeil des dénombrements qui supposent une administration structurée. Néanmoins, un « état des feux » français

---

[19] Wundt, *Essays*, chapitre 8 : « Die Sprache und das Denken ».
[20] Riffaterre, *Essais de stylistique structurale*.
[21] Jauss, *Pour une esthétique de la réception*.
[22] Droesbeke & Tassi, *Histoire de la statistique*.

paraît en 1328. Ce relevé permet d'estimer indirectement la masse de la population à l'aide d'un coefficient multiplicateur sujet à caution. On voit poindre ici une approche théorique qui s'oppose à un comptage rigoureux mais coûteux. Cette pratique acquiert ses lettres de noblesse en Angleterre vers 1660, par l'arithmétique politique de Graunt[23] et Petty[24].

Le terme « statistique » apparaît au même moment en France sous l'administration de Colbert. Son étymologie évoque premièrement l'Etat, mais plus profondément une stabilité[25], une régularité qui se dégage de la nébuleuse des chiffres. La discipline naît formellement au sein de la « Staatenkunde » allemande : en 1746, Achenwall donne les premiers cours de statistique à l'Université de Göttingen.

Les décennies qui suivent voient l'émergence de grands recensements périodiques, propres à fournir des données indiscutables aux différents Etats : en Suède dès 1749, ailleurs en Europe à partir de 1801.

Le champ de la statistique s'élargit avec les sondages d'opinion. Initiés lors des élections présidentielles américaines en 1824, ils se développent rapidement. Un tournant est opéré au cours de l'année 1936 : la méthode des quotas est désormais préférée à un échantillonnage arbitraire.

Progressivement, la statistique quitte le berceau politique pour essaimer sur d'autres domaines : l'économie, les sciences humaines, la médecine, la physique. Elle gagne le terrain linguistique avec les

---

[23] Graunt, *Natural and Political Observations upon the Bills of Mortality*.
[24] Petty, *Political arithmetic*.
[25] Du grec σταω, "je me tiens debout".

travaux précurseurs de Morgan[26] et de Markov[27]. En France, les premiers acteurs sont Müller[28] et Brunet[29] pour leurs études sur le vocabulaire théâtral et littéraire.

2.2.2 Description et inférence

Comment cerner la statistique contemporaine ? Le *Trésor de la Langue Française Informatisé* donne la définition suivante : « branche des mathématiques ayant pour objet l'analyse (…) et l'interprétation de données quantifiables ».

Deux voies apparaissent : la première, objective et descriptive, met l'accent sur la collecte des données et leur restitution fidèle sous une forme condensée ; la seconde, subjective et inférentielle, se fait forte d'élaborer un modèle qui explique ces valeurs expérimentales.

2.2.2.1 Statistique descriptive

Selon cette approche, le recueil des informations doit couvrir l'intégralité de la population supposée finie.

La réduction et la représentation des données s'opèrent à l'aide de techniques variées. Les plus élémentaires consistent à calculer des valeurs typiques[30] ou à tracer la distribution des chiffre par des

---

[26] Ses idée est d'analyser la longueur des mots employés par un auteur. Elle est explorée en 1887 par Mendenhall, « The characteristic curves of composition ».
[27] Cf. son étude de 1913 sur les séries de lettres dans *Eugène Onéguine*.
[28] Müller, *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*.
[29] Brunet, *Le vocabulaire français de 1789 à nos jours*.
[30] La moyenne arithmétique est utilisée par Tycho Brahe à la fin du 16e siècle. L'écart-type est formalisé par Pearson en 1893 dans « Contributions to the Mathematical Theory of Evolution ».

diagrammes[31]. D'autres plus complexes permettent de classer les individus[32] ou de schématiser des nuages de points dans des espaces multidimensionnels[33].

2.2.2.2 Statistique inférentielle

Quand les populations sont trop nombreuses pour être traitées intégralement, le choix d'un échantillon s'impose : la subjectivité imprime une première marque.

Les mondes de l'échantillon et de la population communiquent : aux fréquences expérimentales correspondent les probabilités théoriques. Quand la taille de l'échantillon croît vers l'infini, l'écart entre ces grandeurs tend vers zéro : cette loi des grands nombres[34] est à l'origine de tous les sondages. Le théorème limite central[35] précise que la distribution des écarts converge vers une loi normale, qui acquiert ainsi une portée universelle.

L'inférence consiste à extrapoler un échantillon pour imaginer une population idéale susceptible de l'engendrer. Techniquement, elle tente d'aligner la distribution d'une loi théorique sur celle des données. Des

---

[31] Les diagrammes en barres sont employés par Playfair dans *The Commercial and Political Atlas* en 1786. Il fait apparaître les diagrammes en secteurs dans *The Statistical Breviary* en 1801.
[32] Les premiers dendogrammes remontent à Adanson dans son *Histoire Naturelle du Sénégal* (1757). La méthode de partition avec des centres mobiles est dûe à Mac Queen dans « Some Methods for classification and Analysis of Multivariate Observations » (1967).
[33] L'analyse en composantes principales d'un nuage est initiée en 1901 par Pearson dans « On Lines and Planes of Closest Fit to Systems of Points in Space ». L'analyse factorielle des correspondances entre deux nuages est développée par Benzécri à partir de 1963.
[34] La loi est énoncé par Bernoulli dans *Ars conjectandi*
[35] Si les fondements remontent à Laplace avec la *Théorie analytique des probabilités* en 1812, le choix du nom est attribué à Polya en 1920.

tests statistiques[36] dégrossissent le problème en cherchant une loi adaptée, puis l'estimation[37] opère un réglage plus fin en ajustant les paramètres de cette loi. A chaque étape, la validité des choix se mesure au moyen d'intervalles de confiance[38]. Ainsi, la taille des Français se distribue selon une courbe en cloche : avec un certain pourcentage de confiance, elle se modélise à l'aide d'une loi normale, définie par sa moyenne et son écart-type.

2.3 L'informatique

2.3.1 De l'argile au silicium[39]

Les plus anciennes traces de calcul[40] se situent entre le 10e et le 4e millénaire avant J.C : les Sumériens utilisent des jetons placés dans des boules d'argiles pour faire leurs comptes. Vers 3400 avant J.C., l'octogone de l'Empereur chinois Fou-Hi représente les huit premiers nombres au moyen de trigrammes : à travers le principe du Ying et du Yang, la logique binaire est née. Le boulier semble issu du Moyen-Orient aux alentours de 500 avant J.C.

Vers 820, le mathématicien Al-Khowarizmi et sa « science de l'élimination et de la réduction » initient les algorithmes. Puis en l'an 1000, le pape Sylvestre II adopte la numération arabe et le zéro.

---

[36] Le plus connu et le plus ancien est celui du X^2, défini par Pearson en 1900 avec son article « On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling ». D'autres apparaissent plus tardivement en 1933, comme celui Kolmogorov paru dans *Grundbegriffe der Wahrscheinlichkeitsrechnung.*
[37] Fisher utilise le principe du maximum de vraisemblance dès 1912. L'article décisif paraît en 1925 : « Theory of Statistical Estimation ».
[38] Développé en 1937 par Neyman dans « Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability ».
[39] Birrien, *Histoire de l'informatique.*
[40] « Calcul » à la même étymologie que « caillou », du latin « calculus ».

Dès 1500, de Vinci évoque dans ses notes le principe du calcul mécanique. En 1614, les logarithmes de Napier[41] fondent la conception de la règle à calcul, encore récemment employée pour multiplier et diviser. La première machine est attribuée à Schickard vers 1624 : destinée aux additions et aux soustractions, elle reste peu opérationnelle. La « Pascaline » mise au point en 1642 par le mathématicien éponyme connaît un sort plus heureux, et voit l'embryon de sa commercialisation. Il faut attendre 1694 pour que Leibniz automatise les quatre opérations de base.

En 1833, la machine analytique de Babbage « tisse des structures algébriques de la façon que le métier de Jacquard tissait des fleurs et des feuilles[42] ». Puisant ses données et ses instructions de cartes perforées, son architecture préfigure les ordinateurs modernes : un lieu de stockage, une unité de calcul et un centre de commande. Ce projet ambitieux trouve un écueil dans la technologie de l'époque, mais il inspire le premier programme informatique fait d'itérations.

Conçue à partir de 1884, la machine d'Hollerith sert à dépouiller le recensement de 1890 aux Etats-Unis[43]. Son principe électro-mécanique utilise le passage du courant à travers les trous d'une carte perforée. La « Tabulating Machine Company » créée à cette occasion devient plus tard la célèbre « International Business Machine ».

Les recherches quêtent un automate binaire, supposé à l'image de la pensée humaine. En 1886, Peirce[44] fait intuitivement correspondre

---

[41] Napier, *Mirifici logarithmorum canonis descriptio.*
[42] La métaphore est celle de Lovelace, collaboratrice de Babbage.
[43] Hollerith, *In connection with the electric tabulation system which has been adopted by U.S. government for the work of the census bureau.*
[44] Peirce, *Writings of Charles S. Peirce,* vol. 5, p. 421-424. L'idée est développée dans la thèse de Shannon en 1938.

l'algèbre booléenne[45] et la commutation électrique. Le puzzle se complète en 1904 : alternant deux états en fonction du sens du courant, la diode de Fleming est le premier tube à vide.

En 1937, la machine virtuelle de Turing devient la mère de tous les calculateurs algorithmiques. Constituée d'un curseur qui lit et écrit sur les cases d'un tableau infini, elle prend ses états dans un ensemble fini. Ce dispositif vise à évaluer la puissance des processus algorithmiques, mais aussi à montrer leurs limites à travers les problèmes indécidables[46].

L'ordinateur[47] naît peut-être en 1949 avec l'EDVAC d'Eckert et Mauchly. Electronique, binaire et programmable, sa mémoire stocke les données mais aussi les programmes : la nouvelle architecture due à von Neuman[48] donne l'autonomie à ce cerveau artificiel, qui reste à l'état de prototype.

Egalement conçu par Eckert et Mauchly, l'UNIVAC est le premier produit commercialisé. Le transistor au silicium en 1954 puis le circuit intégré en 1959 donnent le jour à des appareils plus petits et moins chers : l'informatique se démocratise, la voie est ouverte vers l'ordinateur individuel à la fin des années 1980[49] et vers le réseau Internet durant la décennie suivante.

---

[45] Boole, *Les lois de la pensée*.
[46] Turing, « On Computable Numbers, With an Application to the Entscheidungsproblem ».
[47] Le terme apparaît dans son acception technique en 1955. Au Moyen-Age, il désignait Dieu, grand ordonnateur de la Création.
[48] Neumann, « First Draft of a Report on the EDVAC».
[49] L'Apple II est commercialisé en 1977 suivi par le PC d'IBM en 1981.

2.3.2 La modernité

2.3.2.1 Principes et définition

Le terme « informatique » remonte à 1962 : il est l'invention de Dreyfus[50] qui contracte « information » et « automatique ».

Au sens technique comme étymologique, l'information met l'accent sur la forme du message, non sur son contenu : il s'agit de coder simplement pour transmettre efficacement, d'où le recours au binaire[51]. L'automatique[52] se fonde quant à elle sur la boucle de l'action et de la rétroaction : à travers l'itération, le rythme réapparaît sous des traits inattendus.

Plus généralement, le *Trésor de la Langue Française Informatisé* définit ainsi l'informatique : « science du traitement rationnel, notamment par machines automatiques, de l'information ».

2.3.2.2 Extensions et limites

Si l'informatique se préoccupe historiquement du calcul, elle gagne vite d'autres rivages : la forme binaire de l'information englobe tant les chiffres, les textes, les sons que les images. Par ce cheval de Troie, une informatique initialement scientifique pénètre successivement maints aspects de notre quotidien : loisirs, gestion, administration, renseignement… La numérisation n'épargne pas le domaine des lettres : la base Frantext[53] du CNRS couvre cinq siècles de littérature française jusqu'à nos jours. Appelé à grandir, ce corpus électronique rassemble

---

[50] Breton, *Une histoire de l'informatique*, p. 38.
[51] Shannon, *The Mathematical Theory of Communication*.
[52] Wiener, *Cybernetics, or control and communication in the animal and the machine*.
[53] www.frantext.fr

actuellement mille auteurs et quatre mille œuvres.

Ce déploiement insidieux fait naître des sentiments ambivalents au sein de la population : espoir de se libérer du travail, interrogations sur la froideur de la technique, voire crainte d'être asservi par ses propres créatures.

3 Orientations

Le contexte cerné, il s'agit à présent de définir les principaux choix qui sous-tendent la thèse. D'un objet linguistique à un objet littéraire et d'une méthode descriptive à une méthode subjective, de nombreuses combinaisons stylistiques s'offrent à l'exploration. Parmi elles, l'option hybride de Jakobson[54], qui défend l'étude rationnelle de formes poétiques. C'est la voie étroite et abrupte que nous tentons de suivre.

3.1 Corpus et unités

Avant tout traitement, le premier choix est celui des données : en premier lieu le corpus, ou d'un point de vue statistique, la population étudiée ; en second lieu les unités linguistiques, ou les caractères recensés. Les paragraphes à venir expliquent les options retenues.

3.1.1 Légitimation du corpus

Précisons d'emblée que l'objectif de cette thèse est d'abord méthodologique : il ne s'agit pas d'étudier une œuvre, un auteur, un genre ou un courant particulier, mais plutôt de définir des mesures

[54] Jakobson, *Essais de Linguistique générale*, p. 209-248.

adaptées à notre problématique principale : situer objectivement un texte par rapport à un autre.

Le corpus est d'abord un terrain de jeu, et cette liberté laisse une certaine place à la subjectivité. Le choix des œuvres répond en partie à un goût personnel, sans exclure des facteurs plus rationnels de différents ordres.

3.1.1.1 Considérations stylistiques

A contre-pied de Bally et suivant Merleau-Ponty, nous prenons le parti littéraire :

> Il y a lieu, bien entendu, de distinguer une parole authentique, qui formule pour la première fois, et une expression seconde, une parole sur des paroles, qui fait l'ordinaire du langage. Seule la première est identique à la pensée[55].

A cet argument intuitionniste, on pourrait ajouter que la création originale — objet stylistique par excellence — n'est sans doute pas uniquement un don de la nature ou du Ciel, mais aussi une construction faite de choix successifs et individuels. Ces deux aspects semblent plus développés dans le domaine littéraire qu'ailleurs, d'où notre préférence.

Précisément, le corpus comprend trois romans du 20e siècle :

- *Mémoires d'Hadrien,* de Yourcenar ;
- *Vendredi ou les limbes du Pacifique,* de Tournier ;
- *Désert,* de Le Clézio.

Au-delà de leur popularité, ces livres ne sont pas pris au hasard parmi les étagères de notre bibliothèque. Si leurs auteurs décident tous de vivre en marge de l'agitation du monde, leurs personnages diffèrent

[55] Merleau-Ponty, *Phénoménologie de la perception*, p. 207.

radicalement : entre l'Empereur tout-puissant, le gouverneur ambigu de l'île Speranza et la nomade guidée par le vent, la palette du théâtre humain défile, de la présence à l'effacement. De la sorte, une gradation semble apparaître entre ces trois oeuvres : la question est de savoir si elle se traduit dans les chiffres.

3.1.1.2 Considérations statistiques

Avec un point de vue plus logique et métrique, deux œuvres sont nécessaires pour amorcer une comparaison. Associons à chacune un point de l'espace : sans échelle, la carte qui se dessine reste muette sur la distance entre ces individus. Mais l'arrivée d'un troisième compagnon produit des échanges enrichissants par la mise en regard des écarts deux à deux. Plus nombreux, ce groupe ne se laisse généralement pas enfermer dans un plan et menace la perspective : les cartes projetées déforment la réalité spatiale[56]. Finalement, le nombre trois semble être un choix raisonnable.

Figure 0.2 : Trois, nombre d'or de la stylométrie plane ?

3.1.1.3 Considérations informatiques

Dernier critère, mais non des moindres d'un point de vue pratique : le texte intégral de ces œuvres est disponible sous forme numérique dans la base Frantext.

---

[56] Cet aspect est détaillés dans le chapitre 2, section 3.2.

3.1.2 Choix des unités

L'idée est de tracer un panorama linguistique en balayant le spectre des unités de l'étroit vers le large et du simple vers le complexe, soit schématiquement :

– à l'échelle du mot, le phonème ou le graphème ;
– à l'échelle du syntagme, la catégorie grammaticale ;
– à l'échelle de la phrase ou du texte, le concept et le sens.

3.1.2.1 Phonèmes ou graphèmes ?

Paradoxalement, le choix le plus épineux concerne le niveau élémentaire. A l'origine de toute communication[57], l'oral se voit concurrencé au fil des siècles par l'écrit, au point que les graphèmes semblent acquérir une autonomie vis-à-vis des phonèmes :

> La société humaine, le monde, l'homme tout entier est dans l'alphabet. La maçonnerie, l'astronomie, la philosophie, toutes les sciences ont là leur point de départ, imperceptible, mais réel ; et cela doit être. L'alphabet est une source[58].

D'un point de vue informatique, l'examen du plan phonologique nécessite un étiqueteur spécifique. De tels outils sont disponibles, mais ils sont payants ou demandent des développements hors du cadre de cette étude. En revanche, les graphèmes sont directement dénombrables et offrent l'avantage d'une mesure objective, lavée de tout soupçon.

La voie graphémologique est ouverte par Markov[59] en 1913. Dans le contexte de la théorie sur l'information, Shannon[60] suit ses traces en

---

[57] Castarède, Konopczynski & alii, *Au commencement était la voix.*
[58] Hugo, *Alpes et Pyrénées*.
[59] Markov, « Primer statisticeskogo issledovanija nad tekstom "Evgenija Onegina", illjustrirujuscii svaz ispytanii v cep ».
[60] Shannon, « Prediction and entropy of printed english».

1951 avec les n-grammes[61], couramment exploités pour la reconnaissance de formes. Brunet[62] reprend cette approche en 1981 dans son étude du vocabulaire français. Au cours de la dernière décennie, Khmelev et Tweedie[63] puis Jardino[64] utilisent les graphèmes dans le but d'identifier des auteurs. En dépit de ces travaux, la piste reste relativement blanche et poudreuse, d'où le choix de l'approfondir.

3.1.2.2 Syntaxe

Le recours à un étiqueteur automatique est ici inévitable : s'il est clair que rien ne remplace l'œil minutieux de l'expert pour attribuer un terme à une catégorie grammaticale, l'humain est rapidement dépassé quand la taille du corpus grossit. On parie alors que le myopie locale est compensée par la largeur de vue. Le risque reste limité : le taux d'échec est inférieur à 5 % selon les concepteurs du logiciel mis à contribution[65].

3.1.2.3 Sémantique

Le débat entre les partisans de la forme d'un mot et de son lemme fait couler beaucoup d'encre[66]. Pour les uns, la forme immédiate et indiscutable doit fonder l'analyse. Pour les autres, la profusion des visages est rédhibitoire à la clarté du jugement : il faut quitter la jungle touffue des graphies et faire intervenir le lemme unificateur et simplificateur. Un nouveau pas est franchi par une génération de

---

[61] Une séquence de n graphèmes.
[62] Brunet, *Le vocabulaire Français de 1789 à nos jours.*
[63] Khmelev & Tweedie, « Using Markov Chains for Identification of Writers ».
[64] Jardino, « Identification des auteurs de textes courts avec des n-grammes de caractères ».
[65] Le logiciel est présenté dans le chapitre 3, section 4.2.
[66] Un article de Brunet fait le point sur la question : « Qui lemmatise dilemme attise ».

logiciels qui étiquettent un terme et lui associent un sens ou un concept[67]. L'opération qui inclut une analyse syntaxique est complexe, mais le taux d'erreur est supposé inférieur à 10 %[68]. Un pari analogue au précédent permet d'appréhender une unité d'un genre nouveau et d'une essence plus synthétique.

3.2 Méthode comparative

La description de la méthode porte successivement sur ses composantes stylistiques, statistiques et informatiques.

3.2.1 Aspects stylistiques

La notion de comparaison stylistique ne va pas de soi. Idéaliste et romantique, Croce considère que l'intuition de l'artiste échappe par nature à toute pensée logique[69]. La pertinence de cet argument esthétique semble cependant moins concerner la pente poétique de l'art : la structure de la création répond sans doute à la fusion de l'esprit. Pour éviter un dualisme trop strict, nous passons outre cette remarque de fond et postulons la légitimité de la démarche stylistique.

Dès lors, comment comparer des œuvres susceptibles de diverger par leurs auteurs, leurs genres ou leurs chronologies ?

Paradoxalement, la variété et le nombre sont ici des appuis, plus que des ennemis. D'essence, le « style » est singulier, mais cet esprit malicieux ne se laisse appréhender que par la comparaison avec l'Autre.

---

[67] Par exemple, « homme » et « femme » sont rangés dans la classe « être humain ».
[68] Le logiciel est présenté dans le chapitre 3, section 4.3.
[69] Antoine, *Revue d'Enseignement supérieur*, p. 49-60.

Pour reprendre une terminologie classique, ce travail s'inscrit dans le cadre de la stylistique des écarts. Des différences absolues entre une œuvre et une norme sont délicates à établir : se pose la question de la référence qui la définit. On considère donc les divergences relatives aux textes du corpus.

La focalisation sur l'œuvre n'est pas anodine. Il ne s'agit pas d'exclure les approches subjectives, mais de les réserver à un second mouvement : le message ou l'information se placent au centre de toute communication. Chaque création semble d'ailleurs acquérir une unité organique au fur et à mesure qu'elle se rapproche du beau, en relation avec le choix littéraire. Le symbolisme romantique y est sensible, Novlalis l'exprime par ces propos aux accents hermétiques :

> L'artiste primitif n'attache aucune valeur à la beauté intrinsèque de la forme, à sa cohérence et à son équilibre. Il ne vise et ne veut seulement que l'expression bien assurée de son intention : son but est l'intelligibilité du message. […] Le langage à la deuxième puissance, la fable par exemple, est l'expression d'une pensée entière et appartient au hiéroglyphisme mis au carré, à la *langue chiffrée des figures et des sons*[70].

Progressivement, un ordre interne se substitue à la contrainte externe. Condillac l'exprime par une belle analogie :

> Il en est ici du discours comme de la *marche*. La marche habituelle a son but *en dehors d'elle-même*, elle est un pur *moyen* pour parvenir à un but, et elle tend incessamment vers ce but, sans tenir compte de la régularité ou de l'irrégularité des pas séparés. Mais la passion, par exemple la joie sautillante *renvoie la marche en elle-même*, et les pas séparés ne se distinguent plus entre eux par ceci que chacun rapproche davantage du but ; ils sont tous égaux, car la marche n'est plus dirigée vers un but, mais a lieu *plutôt pour elle-même*. Comme de la sorte les pas séparés ont acquis une *importance* égale, l'envie devient *irrésistible de mesurer et de subdiviser ce qui est devenu identique de nature* ; de la sorte est née la danse[71].

Mesure, le mot est lâché dans ces propos. D'aucuns diront que l'intuition est reine dans le jugement artistique et la comparaison littéraire. Mais en-deçà des fulgurances incertaines, le nombre

---

[70] Novalis, *Œuvres complètes*, tome II, p. 89 et 103.
[71] Condillac, *Œuvres complètes*, tome VII, p. 185-186.

apprivoisé peut devenir un support fidèle : c'est l'objet de la stylométrie. Géométrisant le message initial, sa radioscopie fournit une *Gestalt* finalement interprétée par l'œil expert.

Encore faut-il s'entendre sur le point de vue de cette radioscopie et son appréhension de l'environnement : chassée dans les coulisses, la subjectivité réapparaît. Les méthodes traditionnelles de la stylométrie privilégient l'espace aux dépens du temps, à l'encontre des théories physiques modernes qui jouent sur l'équivalence entre ces deux composantes[72]. Il s'agit donc de rétablir un équilibre à travers un facteur essentiel de l'art en général et de la littérature en particulier : le rythme.

Ce dernier, tant par son concept que ses réalisations, menace de se révéler complexe, et l'on risque fort de se perdre dans ce maelström. La dualité entre ordre et désordre musical est d'ailleurs fort ancienne :

> Or, ce sont les Dieux, disions-nous, qui, prenant en pitié notre sort, nous ont donné pour nous accompagner dans nos chœurs comme pour être les chefs de ces chœurs Apollon et les Muses : divinités auxquelles, comme il est naturel, nous en avons, si vous vous en souvenez, ajouté une troisième, Dionysos[73].

Si les unités peuvent être nombreuses, leurs permutations ouvrent des mondes quasi infinis :

> Deux pierres bâtissent deux maisons, trois pierres bâtissent six maisons, quatre pierres bâtissent vingt-quatre maisons, cinq pierres bâtissent six cent vingt maisons, sept pierres bâtissent cinq mille quarante maisons. A partir

---

[72] Cette analogie formelle est à la source de la théorie de la relativité d'Einstein. La référence à la science physique peut sembler incongrue : or, loin de voir dans l'Homme et ses créations de merveilleuses exceptions de la nature suspendues dans l'Ether, nous défendons l'idée d'un continuum entre des univers isomorphes régis par des lois sinon identiques, du moins semblables.

[73] Platon, *Lois*, II, 664e-665a. La partition est reprise par Nietzsche dans *Die Geburt der Tragödie*.

> d'ici continue, calcule ce que la bouche ne peut exprimer et ce que l'oreille ne peut entendre[74].

Plus tard et dans un autre domaine, Newton renonce à caractériser chaque trajectoire et se contente de la comparer à une référence rectiligne uniforme[75]. Notre étude fait ainsi intervenir des écarts relatifs par rapport à un mouvement régulier dans le temps, en d'autres termes des arythmies.

3.2.2 Aspects statistiques

> Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (...) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques[76].

Les propos tranchés de Benzécri ont le mérite de la clarté : si la stylistique est une typologie, la voie désignée est celle de la statistique descriptive. De la sorte, pour comparer deux formes, le plus simple est de se tenir au fait plutôt que de faire intervenir un modèle tiers et hypothétique.

Ce parti général n'empêche pas des incursions locales dans le monde théorique. Ainsi, le recours à des modèles fiabilistes ou à des processus stochastiques pour éclairer la distribution et la succession des temps de retour d'une unité. Mais la perspective est alors linguistique : il s'agit de dégager des traits communs et non de préciser des points singuliers.

---

[74] Papus, *Le Sepher Jesirah*, chapitre 4, section 16. Saisi de vertige, on n'ose imaginer ce nombre pour un livre entier : un nouvel ouvrage serait nécessaire pour écrire ses chiffres...
[75] Weyl, *Symétrie et mathématique moderne*, p. 129.
[76] Benzécri, *Analyse des données,* tome II, p. 3.

3.2.3 Aspects informatiques

Ce travail peut s'inscrire dans le champ du « Traitement automatique des langues ». Alors que les développements de la discipline touchent la reconnaissance vocale et la synthèse de la parole, il ne s'agit ici que de traiter du texte. A côté des applications historiques — traduction automatique, correction orthographique et grammaticale, l'analyse textuelle facilite l'appréhension des corpus volumineux par le pourvoi d'une information précise ou l'élaboration d'une vue d'ensemble. Notre démarche s'inspire principalement de ce dernier axe et fait appel à des techniques de classification et de cartographie.

D'un point de vue matériel, la recherche se satisfait d'un simple ordinateur individuel : ce n'est pas le moindre des charmes d'une discipline qui permet un travail autonome et léger. Nul besoin ici d'accélérateur de particules ou de télescope spatial…

D'un point de vue logiciel, les outils du marché sont mis à contribution autant que possible, notamment les étiqueteurs linguistiques en amont des traitements. Le développement de programmes spécifiques est néanmoins nécessaire pour récupérer ces résultats et réaliser les statistiques en aval.

4 Organisation

4.1 Une vue d'ensemble

Le corps du mémoire commence par une partie qui présente le corpus et les unités, décrit les méthodes de mesure et spécifie les outils informatiques. L'étude proprement dite et ses résultats expérimentaux

sont exposés dans une seconde partie[77].

Les annexes, faces immergées à l'image de ce corps visible, suivent le même parcours, des données aux résultats.

4.2 Deux mouvements

Plus précisément, l'étude comprend cinq scopies agencées en deux mouvements :

4.2.1 « Le chemin mystérieux va vers l'intérieur[78] »

Du regard de Sirius à l'oreille humaine, l'analyse se déploie d'abord selon quatre âges :

- dans la macroscopie, le temps est gelé, un élément du corpus est perçu comme un bloc.
- dans la mésoscopie, la vie s'éveille, une œuvre est entendue comme la succession de ses parties ;
- dans la microscopie, la dynamique prend son envol, cadencée par le retour d'une même unité linguistique[79] ;
- dans la nanoscopie, le tempo s'emballe et se multiplie par les échos de cette unité, l'ensemble de ses occurrences voisines.

4.2.2 Ascension

Dans un second mouvement, la télescopie suit une logique inverse et

---

[77] La vue générale du corpus venant éclairer l'analyse des mesures, on parcourt une moitié du cercle herméneutique de Schleiermacher. Le processus complet intègre le chemin inverse, les faits venant alimenter la perception globale. Une nouvelle boucle peut alors s'engager.

[78] Novalis, *Œuvres complètes*, tome I, p. 357-358.

[79] Pour fixer les idées, l'intervalle entre deux « a » successifs.

vise plus haut : tandis que les premières scopies se contentent d'analyser un corpus étiqueté, cette dernière tente d'identifier des fragments inconnus et lointains. A partir de mesures adaptées, il s'agit de classer les divisions de chaque œuvre, voire de les attribuer à un auteur.

4.3 Trois plans d'expériences

Dans le détail enfin, chaque scopie analyse les trois niveaux linguistiques retenus : la graphémologie[80], la syntaxe et la sémantique. De façon générale, les mesures sont d'autant plus significatives que les échantillons sont grands : les synthèses de ces éléments diffus ponctuent ainsi les principales sections et chapitres de l'étude.

[80] En musique, ce terme désigne la transcription d'un son.

Première partie : principes

Chapitre 1 : le corpus et les unités

Le chapitre débute par une vue générale du corpus, suivie par la revue de détail de chaque œuvre. Le choix des unités linguistiques est précisé à la fin.

1 Aperçu du corpus

Les trois œuvres ont des points communs : ce sont des romans de la deuxième moitié du 20e siècle, de longueurs sensiblement égales.

Plus précisément, trente ans s'écoulent entre la publication du premier et du dernier livre. *Mémoires d'Hadrien* est résolument classique, tant par son thème que sa forme. A l'opposé, le récit exotique de *Désert* est influencé par le Nouveau Roman. Au milieu, *Vendredi ou les limbes du Pacifique* apporte sous des habits traditionnels les idées novatrices d'un Tournier philosophe.

Les œuvres divergent sur d'autres points : tandis que le thème progresse de l'humain vers la nature, la forme prosaïque tend vers la poésie, et l'ethos semble évoluer du masculin au féminin[81] (fig. 1.1) :

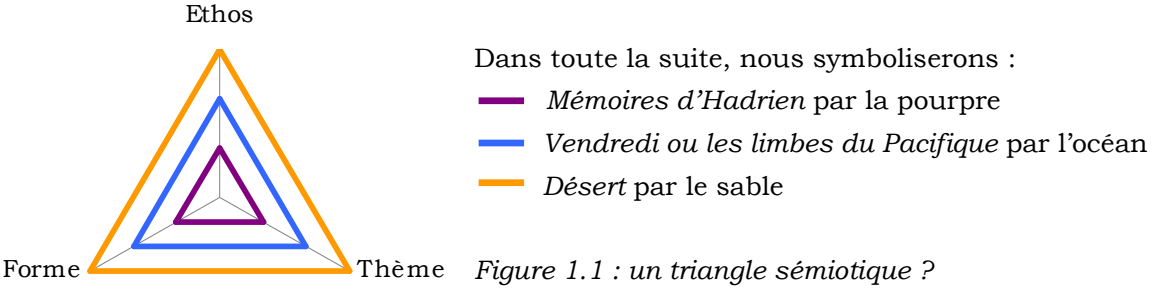


Figure 1.1 : un triangle sémiotique ?

[81] Cette vision est évidemment intuitive : il sera intéressant de voir si elle se traduit dans les chiffres.

✯

Venons à l'étude détaillée du corpus, plus précisément des auteurs et de leurs œuvres. Pour chaque écrivain, les traits essentiels de sa vie, de son écriture et de sa pensée sont esquissés sans chercher l'exhaustivité. Nous n'entrons pas plus dans le commentaire thématique, plus attachés à la genèse d'une œuvre et à sa forme. L'étude stylistique commence donc ici, même si ses aspects numériques n'apparaissent que dans la seconde partie du mémoire.

2 Le pouvoir lucide

2.1 Marguerite de Crayencour

2.1.1 La vie choisie[82]

Née en 1903 d'une vieille famille de l'aristocratie belge, Marguerite perd précocement sa mère. Son père cultivé et original assume la plus grande partie de son éducation, lui donnant les bases d'une culture classique mêlée au goût du voyage. Partie à Paris, la famille se réfugie au début de la première guerre mondiale en Angleterre, puis s'établit dans le sud de la France. Au hasard d'une anagramme sur son patronyme, le nom d'auteur apparaît un soir de 1920[83].

Son père mort en 1929, Yourcenar poursuit ses voyages à travers

[82] Guslevic, *Mémoires d'Hadrien* .
[83] Changeant de nom, Yourcenar reste attachée à son prénom et à sa consonance mystique. Désignant une fleur en français et en grec, il est issu de l'iranien « perle ». Cf. Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 55.

l'Europe, notamment en Italie et en Grèce. Ce sont aussi les années de la bohême sentimentale. Après un amour malheureux et impossible pour son éditeur André Fraigneau, homosexuel, elle exorcise cette passion en écrivant *Feux*, un recueil de poèmes.

Un an plus tard, Yourcenar rencontre Grace Frick, une universitaire américaine qui partagera sa vie pendant quarante ans. Avec le bonheur amoureux arrivent les premiers succès littéraires, *Nouvelles Orientales* et *Coup de Grâce.* Mais cette plénitude est de courte durée : chassée par la guerre et immigrée aux Etats-Unis aux côtés de Grace, contrainte de gagner sa vie en enseignant et réduite à l'immobilité, Yourcenar traverse sa nuit littéraire.

L'éclaircie arrive par la lecture d'anciens manuscrits consacrés à Hadrien, et une nouvelle vie sur l'île des Monts-Déserts où les deux amies s'installent. La publication de *Mémoires d'Hadrien* donne à Yourcenar la célébrité et l'occasion de prendre les airs à travers l'Europe. Dix ans plus tard, elle plonge à nouveau dans le retrait insulaire pour écrire un second livre majeur : *L'Œuvre au Noir*.

Rongée par le cancer de Grace, la décennie suivante sera douloureuse, mais Yourcenar accompagne son amie fidèlement dans la maladie. Poursuivant son parcours littéraire, elle s'éloigne du roman et se tourne vers l'autobiographie avec la publication de *Souvenirs pieux* et *Archives du Nord*. En novembre 1979, Grace meurt.

Un an plus tard, Yourcenar est la première femme élue à l'Académie Française. Entrée dans l'histoire, elle commence une nouvelle vie et entreprend un tour du monde avec le photographe Jerry Wilson. Son nouveau compagnon victime du sida, la mort frappe encore. Abattue par cette perte, Yourcenar décède à son tour sur l'île des Monts-Déserts en 1987.

2.1.2 Faits et visions[84]

Paradoxalement, « l'essentiel, ce n'est pas l'écriture, c'est la vision ». De longues années peuvent ainsi s'écouler entre l'une et l'autre, comme avec *Mémoires d'Hadrien*. Ecrire répond individuellement à un « mystérieux » besoin d'exprimer, qui trouve son utilité dans une société quelquefois en difficulté pour formuler ce qu'elle ressent[85].

Le thème influe évidemment sur l'œuvre, si bien qu'il semble problématique de définir le style d'un auteur :

> Dans une écriture, dans un style, il existe une sorte de soubassement qui appartient en propre à la nature de l'auteur. ; et encore, je n'en suis pas sûre : la mise en œuvre finale dépend du sujet et du moment[86].

Son ambivalence apparaît dans le processus de ses créations : soucieuse de la réalité et du détail[87], mais aussi médium abandonné à la « visitation » de ses personnages. Une communion spirituelle s'établit alors, le poète est « en contact », « traversé par un courant »[88].

> Chaque livre naît avec sa forme tout à fait particulière, un petit peu comme un arbre[89].

A l'image de la littérature japonaise qu'elle admire, Yourcenar aime le dépouillement et l'épure. Arrivé à maturité, l'auteur se méfie ainsi des adjectifs, mais aussi du « je » anecdotique et particulier[90]. Avec la patience d'un artisan, Yourcenar remet cent fois l'ouvrage sur son métier. Allégé des scories et des ornements maladroits, le style retrouve la clarté essentielle d'une esthétique classique[91]. Puis vient le sentiment du devoir accompli :

---

[84] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 231-242.
[85] Ibid., p. 304.
[86] Ibid., p. 236.
[87] Ibid., p. 60-61.
[88] Ibid., p. 209-211.
[89] Ibid., p. 85.
[90] Ibid., p. 97.
[91] Ibid., p. 46-47.

> C'est comme le pain : il y a un moment où l'on sent qu'il ne faut plus pétrir[92].

Par une symétrie étrange, ses deux romans phares — *Mémoires d'Hadrien* et *L'Œuvre au Noir* — sont les pôles solaires et lunaires de sa création[93]. L'un dirigé vers l'extérieur et le pouvoir, l'autre vers l'intérieur et la connaissance.

2.1.3 Bouddhisme

La pensée de Yourcenar est marquée par la philosophie et la spiritualité, notamment à travers Nietzsche et Bouddha[94]. Etablie sur l'île des Monts-Déserts, elle cultive une vie réduite à l'essentiel au contact de la nature, des animaux et de ses voisins : Yourcenar prône l'écologie avant l'heure et tend au végétarisme :

> Tout comme Zénon, il me déplaît de digérer des agonies[95].

Ses ambitions, loin de l'argent et du pouvoir[96], se placent sur le terrain de la conscience et de la culture. Ancrée dans la réalité, elle aspire à l'universel. Elle cite ainsi le poème d'un moine chinois :

> Quelle merveille !
> Je balaie la cour et je vais chercher de l'eau au puits[97] !

---

[92] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 235.
[93] Ibid., p. 191.
[94] Ibid., p. 333-334.
[95] Ibid., p. 307.
[96] Ibid., p. 186-188.
[97] Ibid., p. 303.

2.2 L'univers d'Hadrien

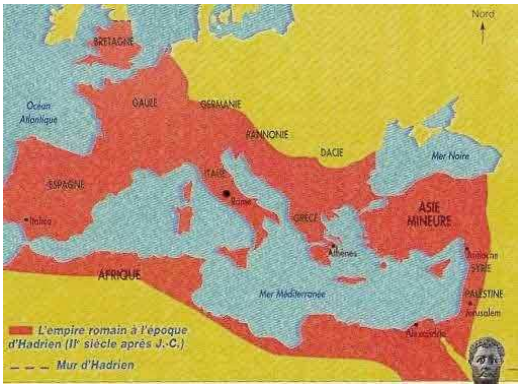


Figure 1.2 : l'empire romain sous Hadrien[98]

2.2.1 Maturation

C'est sans doute au British Museum de Londres que Yourcenar encore enfant rencontre pour la première fois l'empereur, « le viril et presque brutal *Hadrien* de bronze vers la quarantième année[99] ». Une décennie plus tard à Rome, la visite de la villa Adriana a été l' « étincelle » de sa lente passion[100]. Son premier fruit, une version appelée *Antinoos* est refusée en 1926.

Le travail de documentation et d'imprégnation se poursuit jusqu'en 1929, mais le projet tombe peu à peu dans l'abandon tout en mûrissant souterrainement, au fil de l'acquisition d'œuvres d'arts, de lectures et de rencontres. Ce n'est qu'en 1941 que les événements prennent réellement tournure[101] : ouvrant une malle rapatriée de Suisse, Yourcenar retrouve des feuillets oubliés et l'adresse initiale « Mon cher Marc ». Le cadre de son livre est cerné et elle cite Flaubert :

> Les dieux n'étant plus, et le Christ n'étant pas encore, il y a eu, de Cicéron à

[98] Source : corumcle.edres74.ac-grenoble.fr
[99] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 31.
[100] Ibid., p. 146-166.
[101] Yourcenar, *Mémoires d'Hadrien*, « carnet de notes », p. 327.

> Marc Aurèle, un moment unique où l'homme seul a été[102].

A l'image de cet équilibre incertain, elle trouve enfin le point de vue de son personnage central, Hadrien : « je commence à apercevoir le profil de ma mort ».

Yourcenar entreprend alors de « refaire du dedans ce que les archéologues du XIX^e^ siècle ont fait du dehors », reconstituant la bibliothèque d'Hadrien[103] ou recourant à la méditation[104]. « Un pied dans l'érudition, l'autre dans la magie[105] », elle utilise aussi une voie intermédiaire en faisant parler Hadrien dans sa propre langue :

> Immédiatement, je me suis aperçue qu'un certain nombre de phrases écrites en français passaient en grec, et qu'il y en avait une ou deux qui ne passaient pas, parce qu'elles étaient de moi et non d'Hadrien[106].

2.2.2 Forme classique

Pour Yourcenar, le choix du genre est influencé par le contexte de l'écriture. Il s'agit de choisir à un instant donné ce qui saura toucher le lecteur :

> Le roman dévore aujourd'hui toutes les formes. On est à peu près forcé d'en passer par lui. Cette étude sur la destinée d'un homme qui s'est nommé Hadrien eût été une tragédie au XVII^e^ siècle ; c'eût été un essai à l'époque de la Renaissance[107].

Au moment où apparaît le Nouveau Roman et à l'ère du soupçon[108], les *Mémoires d'Hadrien* se situent dans la tradition classique, décrivant avec réalisme un passé lointain. La forme se doit donc d'être cohérente avec l'histoire :

---

[102] Yourcenar, *Mémoires d'Hadrien*, « carnet de notes », p. 321.
[103] Ibid., p. 327.
[104] Humblement ou ironiquement, elle parle de « méthodes de délire » : ibid. p. 330.
[105] Ibid., p. 330.
[106] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 108.
[107] Yourcenar, *Mémoires d'Hadrien*, « carnet de notes », p. 340.
[108] Sarraute, *L'ère du soupçon*.

> En réalité, mon brouillon ne contenait qu'un début de lettre, beaucoup plus près du ton du journal intime, chose impossible pour un Romain, je m'en suis rendu compte tout de suite. […] si c'était un Romain qui parlait, il devait s'agir d'un discours organisé. […] le monologue était la seule forme possible, et je n'ai pas introduit dans le texte de conversation, parce que nous ignorons comment ces gens se parlaient[109].

Si *Mémoires d'Hadrien* est un roman historique, de quel côté penche la balance ? Il semble impossible de dénouer l'intrication des faits et des interprétations :

> Ceux qui mettent le roman historique dans une catégorie à part oublient que le romancier ne fait jamais qu'interpréter, à l'aide des procédés de son temps, un certain nombre de faits passés, de souvenirs conscients ou non, personnels ou non, tissus de la même matière que l'Histoire[110].

La notion même de fiction est remise en cause par Yourcenar, et elle évoque un contact intime avec ses personnages :

> Vous avouerais-je que je n'ai jamais eu le sentiment d'écrire « de la fiction » ? J'ai toujours attendu que ce que j'écrivais fût assez incorporé à moi pour n'être pas différent de ce que seraient mes propres souvenirs (…) la maladie d'Hadrien me paraît aussi authentique que mes maladies[111].

Rapprochons-nous du livre : une « Note » en recense minutieusement les sources. Ajouté plus tardivement[112], un « Carnet » déjà évoqué est une fausse note et une vraie création qui permet au lecteur d'entrer dans le secret de la création littéraire. Associés dans la quatrième page de couverture, le « Carnet » et les *Mémoires d'Hadrien* se rejoignent dans leur ambiguïté entre réalité et fiction[113].

Formellement, le roman est une lettre de l'empereur Hadrien adressée à celui qui doit lui succéder, son petit-fils adoptif Marc-Aurèle. Deux voix s'élèvent : le vieillard qui fait le bilan de son existence et envisage sa mort ; l'empereur qui fait revivre son quotidien passé. Un prologue et un épilogue encadrent ainsi le cœur du roman — les

---

[109] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 148-149.
[110] Yourcenar, *Mémoires d'Hadrien*, « carnet de notes », p. 330.
[111] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 326-327.
[112] A partir de la seconde édition de 1953.
[113] Notre étude stylométrique se restreint au corps principal.

mémoires à proprement parler. Finalement, le texte comporte six chapitres[114] titrés en latin.

Comme le fait remarquer Yourcenar, les mémoires suivent un double mouvement :

> J'ai toujours vu — mes lecteurs le voient rarement — l'histoire d'Hadrien comme une espèce de construction pyramidale. : la lente montée vers la possession de soi et celle du pouvoir ; les années d'équilibre suivies de l'enivrement, qui est aussi le grand moment, si vous voulez ; puis l'effondrement, la descente rapide ; et de nouveau, la reconstruction à ras de terre des dernières années, les usages, les rites religieux romains acceptés, après les expériences exotiques d'autrefois, les travaux poursuivis coûte que coûte, la maladie supportée[115].

Cette progression des événements s'inscrit dans un temps cyclique du récit : l'épilogue rejoint le prologue.

Les principaux passages de ces mouvements sont recensés en annexe[116].

Dans le chapitre central « Sæculum aureum », la vie d'Hadrien semble passer par deux sommets : la rencontre d'Antinoüs sur un plan intime, l'ascension de l'Etna sur un plan extime[117]. A l'opposé, dans « Patientia », Hadrien rencontre la mort, hypogée du corps mais peut-être apogée de l'esprit. Pendant ces phases sublimes, le ton se fait volontiers lyrique et poétique.

En revanche, la maturité d'Hadrien s'exprime par un style mesuré et contenu que Yourcenar appelle « oratio togata[118] » :

> […] cette catégorie du style soutenu, mi-narratif, mi-méditatif, mais toujours essentiellement écrit, d'où l'impression et la sensation immédiates sont à peu

---

[114] « Animula vagula blandula », « Varius multiplex multiformis », « Tellus stabilita », « Sæculum aureum », « Disciplina augusta », « Patientia ».
[115] Galey, *Marguerite Yourcenar, Les yeux ouverts*, p. 101.
[116] Cf. annexe 1.1.
[117] Néologisme emprunté au livre de Tournier, *Journal extime*.
[118] Yourcenar, *Le temps, ce grand sculpteur* , « Ton et temps dans le roman historique », p. 37.

près exclues, et d'où tout échange verbal est ipso facto banni.

Les extraits associés, « Trajan » et « Un successeur », sont ainsi marqués par une logique minutieuse.

3 Une île duale

3.1 Michel Tournier

3.1.1 Marges et honneurs[119]

Le père spirituel de Tournier est peut-être du côté maternel : son grand-oncle, l'abbé Gustave Fournier consacre sa vie à la religion, à la musique et à la « Germanistik ». Dans la même branche de l'arbre familial, sa grand-mère est sans doute d'origine juive[120].

Tournier naît en 1924 de parents eux aussi germanistes, dans un milieu plutôt aisé de la région parisienne. A cinq ans, il subit « l'Agression, l'Attentat, un crime qui a ensanglanté mon enfance et dont je n'ai pas encore surmonté l'horreur » : l'ablation sans anesthésie des amygdales. Il connaît « son premier exil » pendant l'hiver de ses sept ans : envoyé en Suisse pour soigner sa santé fragile, il vit douloureusement cette séparation de la chaleur maternelle.

Ecolier indiscipliné, sa vraie vie est ailleurs, il découvre l'alchimie de « mots mystérieux et d'une extrême précision » dans l'officine de son grand-père apothicaire. Ce dernier l'initie aussi à l'art du portrait photographique. A l'affût des enregistrements de son père, fondateur

[119] Pour une biographie détaillée : Merllié, *Michel Tournier*, p. 215-237.
[120] Plus tard, Tournier songera à emprunter son nom — Anus — pour signer dérisoirement ses œuvres.

lointain de la Sacem, Tournier s'identifie au clown Grock, et écoute avec une « fascination horrifiée » la *Voix humaine* de Cocteau. Son aversion pour une certaine forme de féminité commence peut-être ici : « On ne choisit pas son initiation sentimentale ».

Ballotté entre plusieurs collèges religieux, puis rebuté par les matières scientifiques au lycée, « le vilain petit canard » prépare seul son baccalauréat et l'obtient avec une mention : Tournier prend conscience de ses capacités personnelles.

Un moment tenté par la médecine, fasciné par la figure du monstre et du fou, il découvre sa vocation profonde en lisant Bachelard :

> Il m'avait donné la soudaine révélation que la philosophie était un instrument apéritif, une clé multiple, un ouvre-boîtes universel permettant une effraction incomparable de tout ce qui passe aux yeux du vulgaire pour clos, irrémédiablement obscur, secret et inentamable[121].

Après son D.E.S.[122] de philosophie et sa licence en droit, il part quatre ans en Allemagne pour préparer l'agrégation. Parallèlement, il suit les cours d'ethnographie de Lévi-Strauss. Echouant à son concours en 1949, Tournier n'enseignera pas la philosophie, et sa vie prend une nouvelle direction :

> Ainsi donc s'il fallait dater la naissance de ma vocation littéraire, on pourrait choisir ce mois de juillet 1949 où dans la cour de la Sorbonne Jean Beaufret m'apprit que mon nom ne figurait pas sur la liste des admissibles du concours d'agrégation[123].

De retour à Paris, Tournier entre dans la vie active, il réalise des émissions de radio et de télévision consacrées à la photographie. Le jeune homme occupe aussi plusieurs fonctions dans l'édition, de la lecture à la traduction.

Mais sa vie littéraire ne commence vraiment qu'avec l'installation au

---

[121] Tournier, *Le vent Paraclet*, p. 152.
[122] Diplôme d'Etudes Supérieures.
[123] Tournier, *Le vent Paraclet*, p. 163.

presbytère de Choisel en 1962, acheté par sa famille quelques années auparavant. Retiré du monde, il écrit en secret les premières pages de *Vendredi ou les limbes du Pacifique*. Cinq ans plus tard, le livre est publié et obtient le Grand Prix du roman de l'Académie française. Puis, Tournier débute la rédaction du *Roi des Aulnes*, récompensé cette fois avec le Prix Goncourt en 1970.

Consacré par l'élection à l'Académie Goncourt — un paradoxe pour celui qui se voit en marge — il entreprend une série de voyages et se documente pour son nouveau projet *Les Météores*, qui paraissent en 1975. Puis dans *Le Vent Paraclet*, Tournier fait son autobiographie intellectuelle et explicite son cheminement philosophique et littéraire.

Avec les années, Tournier aspire à la simplicité et écrit des contes, renouant avec les émois littéraires de son enfance, Lagerlöf et Andersen. Parallèlement, il continue ses voyages, qui prennent vers l'ancienne RDA une dimension officielle et politique. Il reçoit d'ailleurs François Mitterrand à plusieurs reprises dans son presbytère de la vallée de Chevreuse.

3.1.2 Copie sinistre[124]

Même anticonformiste, Tournier obéit aux lois de l'intertextualité et se nourrit de l'ancien pour créer le neuf. Il accorde d'ailleurs une place privilégiée au livre des origines :

> Le livre le plus important de ma bibliothèque, c'est une Bible en vingt volumes[125].

Mais de façon plus systématique, il se caractérise par la dualité de son processus de création :

---

[124] Dans l'esprit de Tournier et les écrits de Tiffauges, ce mot n'est pas péjoratif.
[125] Luk, *Michel Tournier et le détournement de l'autobiographie*, p. 207.

> L'usage de la dualité est constant chez moi, mais c'est qu'elle constitue aussi le ressort de toute pensée[126].

Il est ainsi aimanté entre deux pôles, copie et inversion[127] :

> Il ne faut pas trop mépriser la répétition. L'uniforme a sa beauté. Une famille où tout le monde se ressemble avec des variantes dues seulement aux sexes et aux âges, c'est passionnant. Le clonage engendre une sorte de vertige[128].

> L'inversion maligne, c'est une idée si simple et si fondamentale que je ne saurais vraiment pas vous dire d'où je la tiens. Lucifer, image inverse de Dieu, les messes noires, la Reine des neiges d'Andersen, etc.[129]

Toujours dans une logique dualiste et une quête de l'extrême, l'esthétique[130] de Tournier accompagne la réalité pour mieux la subvertir :

> Pour arriver à une effraction des choses, ils [Ernst, Picabia, Magritte, Delvaux, Dali] font confiance plus à la sûreté infaillible du trait qu'au tremblé atmosphérique du rêve. Plus platement on copiera le réel, plus intimement on le bouleversera[131].

Situé dans le genre fantastique et le courant du réalisme magique, Tournier n'a pas de véritable ambition stylistique, stricto sensu :

> Mon propos n'est pas d'innover dans la forme, mais de faire passer au contraire dans une forme aussi traditionnelle, préservée et rassurante que possible une matière ne possédant aucune de ces qualités[132].

Loin des expérimentations du Nouveau Roman, Tournier s'inspire des romanciers classiques du 19e siècle comme Stendhal, Balzac, Flaubert, Zola et Maupassant[133].

Venons-en au processus d'écriture[134] proprement dit. Tournier avoue n'avoir aucune imagination. Il est une « pie voleuse » qui accumule les

---

[126] Bouloumié, *Michel Tournier*, Questions, p. 256.
[127] Tournier utilise aussi d'autres déformations, comme le changement d'échelle.
[128] Bouloumié, *Michel Tournier*, Questions, p. 258.
[129] Ibid., p. 252.
[130] Pezechkian-Weinberg, *Michel Tournier, Marginalité et création*, p. 8-13.
[131] Tournier, *Le Vent Paraclet*, p. 115.
[132] Ibid., p. 195.
[133] Cités dans Tournier, *Le Vol du vampire*.
[134] Rambures, *Comment travaillent les écrivains*, p. 163-167.

« cadeaux du destin ». Un long travail de documentation[135] précède les premiers mots.

Ceux-ci s'inscrivent méthodiquement dans un plan en miroir :

> L'un des secrets consiste à écrire la fin du roman avant le début. [...] je procède ensuite à un découpage rigoureux. Le livre se compose toujours de deux versants séparés au milieu par une crise. [...] Pour obtenir les correspondances, il suffit de travailler simultanément à chacun de ces versants. je n'hésite pas, s'il le faut, à écrire à reculons[136].

Le procédé de l'inversion se manifeste sur plusieurs échelles, à l'image d'une fractale : au sein de l'intertexte, dans le plan du livre, enfin au cœur même des personnages, qui voient leurs vies bouleversées par une initiation de nature ésotérique, une Transfiguration.

Le texte écrit dans un « climat d'obsession », Tournier réécrit[137]. Il applique cette fois le principe de la copie, mais réflexivement à sa propre création. Le résultat n'est cependant pas une reproduction fidèle, il s'agit de tendre vers la simplicité :

> Ce n'est pas une écriture enfantine, c'est une écriture idéale qui est simple, qui est limpide. [...] Et je suis à la recherche de cela. Pour moi, ça, c'est le fin du fin, c'est le summum : des choses simplissimes qui sont à la porté de tout le monde et que je suis le premier à avoir[138].

3.1.3 Philosophe-pédagogue

Etudiant, Tournier consacre son mémoire de D.E.S. à « L'intuition intellectuelle dans la philosophie de Platon ». Plus tard, il dira à propos de Spinoza et de sa partition de la connaissance — sensible, rationnelle et intuitive :

> L'*Ethique* est à mes yeux le livre le plus important qui existe après les

---

[135] Merllié, *Michel Tournier*, p. 254-256.
[136] Rambures, *Comment travaillent les écrivains*, p. 166.
[137] Merllié, *Michel Tournier,* p. 271-274.
[138] Luk, *Michel Tournier et le détournement de l'autobiographie*, p. 210.

> *Evangiles*[139].

Philosophe, Tournier suit d'abord une idée[140]. Pédagogue, il s'attache à la restituer clairement par des mots. Littéraire enfin, il entend rendre ces derniers vivants :

> Il ne fallait pas renoncer aux armes admirables que mes maîtres métaphysiciens[141] avaient mises entre mes mains. je prétendais bien sûr devenir un vrai romancier, écrire des histoires qui auraient l'odeur du feu, des champignons d'automne ou du poil mouillé des bêtes, mais ces histoires devraient être secrètement mues par les ressorts de l'ontologie et de la logique matérielle[142].

La métaphysique se « transmute » ainsi de façon voilée dans cette littérature « venue d'ailleurs ». Le mythe y joue un rôle primordial :

> La fonction sociale — on pourrait même dire biologique — des écrivains et de tous les artistes créateurs est facile à définir. Leur ambition vise à enrichir ou au moins à modifier ce « bruissement » mythologique, ce bain d'images dans lequel vivent leurs contemporains et qui est l'oxygène de l'âme[143].

Ambigu, le mythe est universel, mais il est à l'origine création, récit de héros légendaires. Derrière la connivence de surface, il rappelle le chaos initial[144]. D'un « rire blanc », Tournier veut insuffler une nouvelle vie dans les mythes et retrouver l'esprit dans la lettre morte.

> Lorsque les lattes disjointes de la passerelle où chemine l'humanité s'entrouvrent sur le vide sans fond, la plupart des hommes ne voient rien mais certains autres voient le rien. Ceux-ci regardent sans trembler à leurs pieds et chantent gaiement que le roi est nu. Le rire blanc est leur cri de ralliement[145].

Il s'agit finalement, par ce geste hygiénique, de secouer les idées reçues pour entrevoir l'essence des choses :

> Le véritable sens de la nature morte, c'est plutôt, me semble-t-il, de considérer des objets d'usage — normalement oblitérés à nos yeux par leur

---

139 Tournier, *Le vent Paraclet*, p. 235.
140 Schématiquement, Tournier prend le chemin inverse de Yourcenar, qui part plus volontiers du fait.
141 Dont le « cryptométaphysicien » Sartre.
142 Tournier, *Le vent Paraclet*, p. 179.
143 Ibid., p. 192.
144 L'enfance est un « chaos brûlant » : ibid., p. 19.
145 Ibid., p. 199.

> utilité — hors de tout usage non seulement actuel, mais possible. Leur présence, habituellement très effacée dans notre vie, devient tout-à-coup exorbitante. Le dessin les fait passer du relatif à l'absolu. La cafetière et le pot à tabac se refusent désormais à contenir du café ou du tabac. Ce sont des archétypes, des idées platoniciennes[146].

Plus ambitieusement, un Tournier « naturaliste mystique » aspire sans doute à dépasser la dualité évoquée et les oppositions terrestres. Par un jeu de double inversion — la bénigne et individuelle rachète la diabolique et universelle — il entend se rapprocher de l'harmonie céleste, voire muter vers l'Androgyne au corps glorieux. Dans cette tentative d'unir les contraires, il rejoint le romantisme et l'ésotérisme de son courant allemand.

3.2 La terre de Vendredi

Figure 1.3 : l'archipel de Juan Fernandez[147]

3.2.1 Généalogie

De façon lointaine, *Vendredi ou les limbes du Pacifique* est peut-être né vers 1948 au Musée de l'Homme, alors que Tournier suit les cours de Lévi-Strauss :

---

[146] Tournier, *Le Vagabond immobile*, p. 69.
[147] Source : www.periodistadigital.com

> L'idée que Robinson eût de son côté quelque chose à apprendre de Vendredi ne pouvait effleurer personne avant l'ère de l'ethnographie[148].

Pendant les quinze ans qui suivent, Tournier écrit plusieurs manuscrits qui restent dans ses tiroirs[149]. Le récit de la solitude et de la communion élémentaire s'enracine véritablement en 1962 avec le retrait et l'isolement : son presbytère insulaire constitue une sorte de microcosme, qui unit l' « horizontalité de l'étang » et la « verticalité de l'arbre » :

> Il faut réussir à se créer un terrier, comme le blaireau qui va mettre bas. J'ai la chance d'habiter seul, une grande maison à la campagne. [...] Au bout de quatre ans, au moment où le livre touche à sa fin, je ne quitte plus mon grenier que quelques heures par jour pour bêcher, scier, ou développer quelques photographies[150].

D'un point de vue intertextuel, son roman réécrit a contrario celui de Defoe[151] et s'inspire parallèlement des robinsonnades de Giraudoux et Verne. S'y ajoutent des influences romantiques : ainsi Rousseau pour le thème du bon sauvage et Novalis pour celui de l'alchimie.

Enfin, Tournier écrira lui-même d'autres versions de son roman initial : si *Vendredi ou la vie sauvage* évolue vers la simplicité, *La fin de Robinson* conclut le cycle sur le registre de la dérision.

3.2.2 Echos structurés[152]

Le titre du livre mérite une certaine attention : marquée par une alternative, la dualité apparaît d'emblée. Celle-ci se fait aussi sentir dans le jour de la mort du Christ avant sa résurrection, et dans des limbes ambigus suspendus entre Terre et Ciel. A l'image du roman,

---

[148] Tournier, *Le Vent Paraclet*, p. 227.
[149] Ibid., p. 163.
[150] Rambures, *Comment travaillent les écrivains*, p. 164.
[151] Defoe, *Robinson Crusoé.*
[152] Epinette-Brengues, *Vendredi ou les limbes du Pacifique*, p. 29-34.

Robinson reste ici dans l'ombre du « sauvage » et de son île. Paradoxalement, cette dernière est peut-être première, si l'on en croit les longueurs des termes articulés par « ou ».

Le prologue fait plus qu'annoncer le récit, il est un abyme étrange : son image réduite et déformée semble être vue dans le cristal d'un globe. Ce procédé, utilisé dans certains opéras, suggère la conception musicale[153] de cette œuvre littéraire :

> Mais pour loi, le modèle des modèles, c'est ce paroxysme d'alchimie verbale que constitue l'*Art de la fugue* de Jean-Sébastien Bach. Lorsqu'on sait que c'est sur la dernière fugue de cette dernière œuvre que l'on a retrouvé mort le compositeur, et que, par un raffinement suprême, celle-ci était construite sur les notes allemandes B.A.C.H., il faut bien reconnaître qu'il n'y a rien de plus romantique, de plus intolérable, dans toute l'histoire de la musique[154].

Comme dans une fugue, sujet et réponse semblent se faire écho : dans un temps, le rythme quotidien d'un « log-book » écrit le plus souvent au présent et à la première personne ; dans un autre, la perception mythique[155] d'un récit traduit au passé et à la troisième personne.

L'histoire est divisible en deux parties, avant et après l'arrivée de Vendredi[156]. Mais derrière cette apparence, un centre caché apparaît au sein de la caverne platonicienne : là-bas, « l'obscurité changea de signe[157] » et l'œuvre alchimique passe du noir au blanc. De part et d'autre de ce sommet, deux versants s'abaissent : le stade prénatal et l'apprentissage de la vie.

Après la gestation suivent les initiations telluriques, éoliennes et solaires[158], selon les trois stades spinoziens de la connaissance : la

---

[153] Bouloumié, *Michel Tournier, le roman mythologique*, p. 73-81.
[154] Rambures, *Comment travaillent les écrivains*, p 167.
[155] Sur la structure feuilletée du mythe et de la musique : Lévi-Strauss, *L'homme nu.*
[156] Tournier, *Le Vent Paraclet*, p. 232.
[157] Tournier, *Vendredi ou les limbes du Pacifique*, chapitre V, p. 124.
[158] Vierne, *Rite, roman, initiation*, p. 119-123.

sensibilité, la raison et l'intuition. Associant la gestation à l'eau, le tellurique à la terre, l'éolien à l'air et le solaire au feu, l'alchimie fondatrice de la religion[159] du Pacifique se fait jour.

Plus finement, le livre est divisé en douze chapitres, sans doute en référence à la Bible — les tribus d'Israël et les apôtres du Christ — ou à l'astrologie et l'astronomie — les signes du zodiaque et les mois de l'année[160].

Avec l'arrivée du *Whitebird*, un nouveau cycle commence. Les rôles de Robinson et Vendredi s'inversent : le premier choisit de vivre dans l'île, le second part découvrir la civilisation. Initié et parvenu à la phase du rouge alchimique, Robinson devient l'initiateur de Jaan, jeune mousse réfugié à Speranza. L'histoire se poursuit avec ce nouveau « naufragé ».

Les extraits retenus en annexe correspondent aux quatre éléments évoqués. Le style y est résolument classique.

4 Le silence du désert

4.1 Jean-Marie Gustave Le Clézio

4.1.1 Voyage et contemplation

Rebelle, son aïeul François refuse de couper ses cheveux pour entrer

[159] Du latin *religio*, « ce qui relie ».
[160] En incluant le prologue, on obtient 13 parties : insidieusement, Tournier subvertit peut-être la tradition.

dans l'armée révolutionnaire. Faisant fi de son patronyme breton[161] et ouvert au monde, il immigre à l'île Maurice. Autre figure de cette famille fantasque, son grand-père paternel abandonne sa charge de magistrat et part à la recherche d'un trésor sur une île voisine, Rodrigues. De cet archipel métissé naissent le père anglais et la mère française de l'écrivain.

Le Clézio naît à Nice en 1940 et passe les premières années de son enfance chez ses grands-parents. Dans l'arrière-pays, il découvre la nature :

> Les mots dont vous parlez [silence, fourmi, rocher, mouche, poussière, terre] sont des points de repère : je veux dire que tout le monde a une mythologie personnelle et il se trouve que ces mots-là représentent ce à quoi je me suis d'abord intéressé quand j'étais enfant[162].

A sept ans, il fait son premier voyage : avec sa mère, il part à la rencontre de son père, médecin de brousse en Afrique. C'est aussi le moment du premier livre, *Un long voyage* :

> A bord — je me souviens très bien de ça — ma mère me disait : « Viens, on voit la côte ». Mais je restais dans la cabine et j'écrivais. J'écrivais ce que je ne voyais pas[163].

Enfant rêveur et sensible[164], Le Clézio se sent étranger à la société des hommes. Il prend « le parti de ne pas parler[165] » et trouve refuge dans l'écriture de romans d'aventures ou de poèmes :

> Au lieu de m'amuser, au lieu de faire l'effort d'être comme tout le monde, je préférais rester chez moi à écrire[166].

Plus tard au lycée, la réalisation de bandes dessinées lui permet de trouver une place parmi ses camarades, mais aussi de dépasser

---

[161] Clézio signifierait « clos » dans cette langue.
[162] Lhoste, *Conversations avec J.M.G. Le Clézio,* p. 39.
[163] Ezine, *Ailleurs*, p. 26.
[164] Adulte, il dira ne vivre « que d'émotions » : Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 118.
[165] Ibid., p. 17.
[166] Ibid., p. 49.

certains clivages :

> J'aurais aimé être dessinateur de bandes dessinées. [...] je crois que les arts qui réalisent une fusion entre deux ou trois éléments sont particulièrement accomplis[167].

A la fin de ses études de lettres, son mémoire de maîtrise porte sur la *Solitude dans l'œuvre d'Henri Michaux*, et il prépare une thèse consacrée à Lautréamont[168]. Son premier livre *Le Procès-Verbal* est récompensé par le prix Renaudot en 1963. Cette entrée précoce dans le monde littéraire lui permet « d' ignorer les nécessités de la vie[169] ».

Il part ensuite en Thaïlande faire sa coopération militaire. Ce séjour en Orient et la lecture des *Veda* lui inspire en 1967 un essai, *L'Extase matérielle*. Conjointement, il dénonce l'Occident dans *Le Déluge*. Inadapté à la société, le héros de ce roman qui finira par se suicider déclare :

> Le monde s'agite trop pour moi. On fait trop de choses à la fois. C'est ça que je ne peux pas supporter[170].

La bataille contre « Hyperpolis » culminera avec *La Guerre* en 1970 et plus tardivement avec *Les Géants* en 1973.

Sa vie prend un tournant décisif lors d'un voyage au Mexique. Invité comme professeur, il découvre la civilisation sauvage amérindienne :

> Il y a une vingtaine d'années, entre 1970 et 1974, j'ai eu la chance de partager la vie d'un peuple amérindien, les Emberas, et leurs cousins germains, les Waunanas, dans la province du Darien au Panama, expérience qui a changé toute ma vie, mes idées sur le monde et l'art, ma façon d'être avec les autres, de marcher, de manger, d'aimer, de dormir, et jusqu'à mes rêves[171].

Plus tard en 1975, Le Clézio fait une autre rencontre : sa femme Jemia, d'origine saharienne. Tous deux rêvent secrètement de connaître

---

[167] Ezine, *Ailleurs*, p. 19.
[168] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 35-36.
[169] Ezine, *Ailleurs*, p. 22.
[170] Le Clézio, *Le Déluge*, p. 259.
[171] Le Clézio, *La fête chantée*, p. 9.

ce pays mystérieux, dans le sud du Maroc. Pour préparer un voyage qui aura lieu quelques années plus tard, Le Clézio écrit le roman *Désert*. C'est le livre de la maturité, consacrée en 1980 par le prix Paul Morand de l'Académie française.

En 1981, un voyage teinté de nostalgie le ramène à l'archipel des Mascareignes. Du retour aux sources dans l'Océan Indien naissent quelques années plus tard un roman et un journal : *Le Chercheur d'or* et *Voyage à Rodrigues*.

Le Clézio partage actuellement sa vie entre le Mexique et la France :

> J'ai besoin de ce déséquilibre. J'ai besoin d'avoir deux portes. [...] C'est quelque chose qui m'impressionne énormément, ce moment où deux êtres, deux sociétés ou deux cultures se rencontrent. Parce que je me sens moi-même dans cette situation[172].

Le Rêve mexicain ou la pensée interrompue en 1988 est un autre essai sur le passé de son pays adoptif. Depuis, les livres de Le Clézio poursuivent leur voyage entre le pays niçois, l'océan, l'île Maurice et l'Afrique.

4.1.2 Ecrits du silence

Ecrire est sans doute un paradoxe, pour l'enfant taciturne et l'adulte épris du silence indien :

> Justement parce que le silence n'y est pas perçu comme une absence de paroles, mais comme une autre manière de s'exprimer. On pourrait comparer le silence des Amérindiens à la non-violence des Indiens à l'époque de Gandhi. Quand les Mexicains se taisent, c'est qu'ils ont quelque chose d'important à dire[173].

Néanmoins, le langage joue un rôle central pour cet homme en exil sur terre :

---

[172] Ezine, *Ailleurs*, p. 93.
[173] Ibid., p. 71.

> La langue française est peut-être mon seul véritable pays[174].

Ecrire en français est d'ailleurs plus qu'un choix esthétique, c'est aussi exercer un contre-pouvoir face à un anglais dominant[175].

Hormis ce choix, Le Clézio nourrit des sentiments ambivalents à l'égard des mots qui le font vivre et qui paradoxalement expriment ses réserves. Incapable de tout dire[176], elle n'est pas un système clos, suffisant et arbitraire, mais une voie d'accès à la vie et au rêve[177] :

> Si le langage n'est fait que de mots, il n'est rien du tout. Quelques bruits avec la bouche, quelques gestes, quelques silences : ce n'est pas une musique. Mais quand dans les mots viennent la danse, le rythme, les mouvements, les pulsations du corps, les regards, les odeurs, les traces tactiles, les appels; quand les mots jaillissent non seulement de la bouche mais du ventre, des jambes, des mains, quand tout l'air vibre et qu'il y a comme une auréole de lumière autour du visage; quand surtout les yeux parlent, et le regard est une route sans fin qui traverse le cosmos; alors on est dans le langage, dans sa beauté, et il n'y a plus rien de muet, ou d'insensé[178].

Tous les mots ne sont donc pas mis sur le même plan. Certains sont même explicitement bannis :

> Bien sûr, on les croyait importants, ces mots du langage, ces mots courants ; Dressés comme des meutes, utiles à chasser, chercher, aboyer, tuer. Mais il y a une autre langue, qu'on parlait avant sa naissance. Une langue très ancienne, qui ne servait à rien, qui n'était pas le langage du commerce avec les hommes. Pas une langue de séduction pour suborner, ou pour asservir[179].

Face à l'hydre centrale et lacunaire de la langue, la fonction de l'écrivain est de « produire et non pas représenter[180] ». Loin du mime, le poète est un individu libre :

> [Ecrire] est bien une folie parce que c'est contraire à toutes les règles de la bienséance et de l'efficacité, et de la vie de tout le monde. Ecrire, ça implique qu'on ne vit pas comme tout le monde. En même temps, — et c'est peut-être

---

[174] Chanda, *Entretien avec Jean-Marie Le Clézio*.
[175] Le Clézio, *La Quinzaine littéraire* 436, p. 6.
[176] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 19.
[177] Face à Hermogène, Le Clézio se situe plutôt du côté de Cratyle.
[178] Le Clézio, *L'inconnu sur la terre*.
[179] Le Clézio, *Vers les Icebergs*, Inigi, p. 62.
[180] Ezine, *Ailleurs*, p. 29.

> une part de ma folie —, c'est croire en la liberté. Je suis persuadé qu'on est libre[181].

Par un acte de magie, l'écrivain-sorcier modifie la perception du temps et de l'espace :

> Quand on écrit des livres, on rêve un peu de ça — pas d'arrêter le temps, bien sûr, mais de le faire durer le plus longtemps possible[182].
> […]
> Ce serait bien d'écrire comme on vole. […] Quand on vole, en effet, on a cette impression plus vaste et plus large, et on respire mieux[183].

Avec ce regard neuf, écrivain et lecteur communient au monde : un art total trouve ainsi sa force[184].

Plus concrètement, comment Le Clézio procède-t-il pour écrire ? De longs mois peuvent passer avant de prendre la plume :

> C'est comme un orage qui s'accumule en moi. […] J'ai une telle tension nerveuse que je ne peux plus faire autrement que d'écrire[185].

Un « sentiment inexplicable de l'œuvre future[186] » l'habite alors. Arrivé dans cette phase, il travaille plus volontiers dans le silence de la nuit solitaire pour se placer dans un état réceptif :

> J'ai besoin que la vie s'estompe pour pouvoir l'imaginer à nouveau[187].

Si l'inspiration se fonde sur des réminiscences de plus en plus nombreuses avec l'âge, l'imagination pure est le « paradis des écrivains[188] ». Chez Le Clézio, cette dernière est d'abord visuelle, sans doute en relation avec son amour de jeunesse, la bande dessinée : il crayonne ainsi ses personnages pour les voir vivre[189]. Puis viennent

---

[181] Ezine, *Ailleurs*, p. 121.
[182] Ibid., p. 100.
[183] Ibid., p. 114-116.
[184] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 32.
[185] Ibid., p. 105-106.
[186] Rambures, *Comment travaillent les écrivains*, p. 96.
[187] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 53.
[188] Ezine, *Ailleurs*, p. 28-29.
[189] Ibid., p. 22.

« des séquences, des mouvements, comme un film qui se déroule[190] ». Les premiers mots surgissent alors, spontanés et imprévisibles :

> Ecrire sans savoir où l'on va, en laissant les choses se faire d'elles-mêmes, sans aucun plan. [...] Ca, c'est bien ; c'est laisser dériver le fil[191].

De fil en aiguille, les chapitres s'enchaînent non logiquement, mais organiquement, comme « une marée qui monte et qui descend[192] ».

Dans un second temps, Le Clézio relit et tend à supprimer les adjectifs inutiles[193]. Il s'agit non seulement d'épurer le style, mais plus profondément d'effacer le jugement de son ego pour être mieux compris des autres.

> Ce qui est important, ce n'est pas ce qu'un individu ressent, c'est pourquoi il le ressent et dans quelle mesure les autres le ressentent[194].

Le livre achevé, les critiques les plus importantes viennent ainsi du lecteur[195], et le temps de l'écriture décrit un nouveau cercle, dans une alternance de souffrances et de libérations[196].

Sur le fond, Le Clézio entend engendrer son œuvre dans la continuité[197]. Cependant, au fil des années, les formes évoluent autour de deux extrêmes :

– les écrits initiaux situés dans la ville occidentale : marqués par la rébellion tant thématique que stylistique, ils clament le Nouveau Roman ; les protagonistes y sont le plus souvent masculins ;
– les écrits de la maturité, qui ont pour cadre le désert, de glace ou de sable : l'attitude apaisée et le ton assagi, ces œuvres traversent les

---

190 Rambures, *Comment travaillent les écrivains*, p. 98.
191 Ezine, *Ailleurs*, p. 51.
192 Lhoste, *Conversations avec J.M.G. Le Clézio,* p. 105.
193 Ibid., p. 71.
194 Ibid., p. 76.
195 Ibid.., p. 31.
196 Ibid., p. 92-93.
197 Ibid., p. 61.

courants littéraires, et leurs figures sont essentiellement féminines.

A partir de 1980, Le Clézio paraît entamer un nouvelle phase et revisite des mondes plus humains. Sur une orbite plus large que celui du livre, le temps de l'écrivain est également circulaire[198].

4.1.3 Exil chamanique

Face à l'Occident, Le Clézio vit un triple exil, il fuit la ville tentaculaire, le verbe mercantile et l'ego dévorant. Ce rejet épidermique, tenté par le nihilisme, gravite autour de l'œil d'une raison impérialiste :

> L'intelligence est un mur qui se dresse à des kilomètres d'altitude et qui écrase[199].

Rebuté par le monde moderne, Le Clézio trouve des sources d'inspiration dans l'histoire grecque, notamment dans le mythe d'Icare et la pensée présocratique de Parménide[200].

Traversant les pays, Le Clézio s'enrichit de traditions et de spiritualités variées, animistes ou orientales[201]. Mais c'est la culture amérindienne qui oriente durablement sa vision de la vie. Avec un chamanisme enraciné dans la terre et ouvert au rêve, il trouve ce qu'il cherche depuis son enfance :

> Ce que j'entrevoyais dans cette civilisation, c'était un monde beaucoup plus passionné que celui de la Renaissance européenne. Un monde qui n'était pas fondé sur la raison, ni sur les grandes idées humanistes — sans cesse contredites —, mais sur d'autres choses. Un monde animé par cette danse, cet élan vers la magie, le surnaturel ; fondé sur une perception plus intuitive du monde[202].

Débarrassé de ses oripeaux mentaux, un Le Clézio initié entre en

---

198 Ezine, *Ailleurs*, p. 41.
199 Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 76.
200 Jollin-Bertocchi & Thibault, *J.M.G. Le Clézio*, p. 9.
201 Par son titre, *L'extase matérielle* peut évoquer Sri Aurobindo, Mère et Satprem, notamment l'essai *Le mental des cellules*.
202 Ezine, *Ailleurs*, p. 43-44. Nous retrouvons les trois connaissances de Spinoza.

osmose avec la nature et communie avec les éléments :

> Ici, il n'y a plus d'hommes. Il n'y a plus de maisons, plus de barrières. On y est, on est arrivés, enfin. On est au sommet de la terre, sous l'étoile. On ne cherche plus rien, on ne désire plus rien. On est là, seulement, totalement, dans la substance de l'air, comme on est légers ! On a perdu l'épaisseur, on n'est plus opaques. La lumière froide et qui ne scintille pas vous traverse, le vent passe comme par une fenêtre ouverte[203].

Parmi ces éléments, certains tiennent une place particulière :

> Pour les Amérindiens, l'or était une substance extraterrestre, comme tombée du soleil ; des gouttes de soleil sur la terre[204].

Paradoxalement, cette visite intime des éléments, cet accès à « l'autre versant de la réalité[205] » sont associés au rêve métaphysique :

> Je crois que toutes les sociétés amérindiennes sont marquées par cette possibilité de recours au rêve. Elles ne considèrent pas le réel comme la solution définitive à tous les problèmes[206].

Au-delà des éléments, il s'agit de trouver un nouvel équilibre entre les différentes formes de vie :

> Il y a la certitude que l'être humain ne doit pas être séparé de son milieu naturel. La ville n'est pas son milieu naturel. Son milieu naturel, c'est l'équilibre entre toutes les forces — y compris animales et végétales. On pourrait d'ailleurs voir là une sorte de panthéisme[207]…

Les relations humaines sont elles aussi appelées à évoluer vers moins d'agressivité et plus d'égalité[208]. La terre appartient ainsi à tout le monde et les générations du futur y ont déjà leur place :

> Nous, notre volonté, c'est de laisser à nos enfants la terre telle qu'on l'a trouvée. Il y a des forêts, on leur laissera les forêts. Il y a une rivière, on leur laissera la rivière[209].

Le cœur de chacun est touché par cet art de vivre : de la

---

[203] Le Clézio, *Vers les Icebergs*, p. 30.
[204] Ezine, *Ailleurs*, p. 67-68.
[205] Ibid., p. 116.
[206] Ibid., p. 50.
[207] Ibid., p. 78.
[208] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 111-112.
[209] Ezine, *Ailleurs*, p. 59. Propos attribués à un chef indien.

« plénitude[210] » individuelle naît alors une culture collective, qui vise essentiellement à laisser des traces :

> Je trouve fabuleux d'imaginer des sociétés où tout le savoir — c'est-à-dire la science astronomique, le calcul mathématique, etc. — avait comme fin de laisser une trace pour des survivants actuels[211].

Dans la prémonition, voire l'attente de la catastrophe, l'Amérindien regarde la mort en face : là est précisément sa survie[212].

Rêvant de dépasser les frontières de la terre et de l'esprit, du temps et de l'espace, Le Clézio est fondamentalement en quête de l'unité.

4.2 Le vent et Lalla

Figure 1.4 : la vallée de Saguia el Hamra

4.2.1 Sources et débouchés

A l'origine de l'écriture *Désert*, il y a d'abord l'attirance physique de Le Clézio pour ce lieu :

> C'est le désert, la mer minérale, la pierre devenue mer et c'est fascinant[213].

---

[210] Ezine, *Ailleurs*, p. 37.
[211] Ibid., p. 40.
[212] Ibid., p. 39.
[213] Lhoste, *Conversations avec J.M.G. Le Clézio*, p. 42

Il s'agit aussi pour sa femme de retrouver la terre de ses ancêtres, à travers un roman qui prépare le voyage réel :

> Jemia connaît depuis toujours son identité. Sa mère faisait à la fois référence à son ethnie saharienne et à sa couleur en lui disant qu'elle était une Hamraniya, une Peau-Rouge en quelque sorte[214].

Plus généralement, *Désert* est la rencontre et le choc entre deux civilisations :

> C'est dans ce désert qu'était née la première grande insurrection quand les marabouts lançaient leurs appels à la guerre sainte et que le cheikh Ma el Aïnine, enveloppé dans son immense *khount* (voile) bleu de mer, exhortait ses fils Mohammed Laghdaf et Ahmed el Dehiba, la « Parcelle d'or », à combattre avec leurs cavaliers et leurs méharis l'une des plus puissantes armées du monde équipée de mitraillettes et de canons, et promettait à ses guerriers l'invulnérabilité en soufflant sur ses ennemis du sable chargé de sortilèges[215].

Dans cet affrontement, la préférence de Le Clézio ne fait aucun doute :

> Leur temps est plus vrai, plus réel, il se calcule sur le mouvement des astres et les phases de la lune, non suivant des plans établis à l'avance. Leur espace n'a pas de limites, il loge dans leurs yeux, dans leur volonté d'aller au gré de leurs routes[216].

Plus fondamentalement encore, le désert est un symbole. Fait de vide, de silence et de lumière, ce lieu est une source mystique :

> […] il n'y a rien qui vienne troubler les sens, l'homme peut se sentir plus près de Dieu. […] dans cette vallée où les étrangers disent qu'il n'y a rien, mais où il y a, au contraire, la plénitude de la pensée et l'infinité de l'amour[217].

Le motif du désert trouve plusieurs débouchés dans l'œuvre de Le Clézio. A treize ans, il écrit un premier roman dont le héros, un cheikh vêtu de blanc, préfigure étrangement Ma el Aïnine. Deux ans avant *Désert*, un hommage à Michaux, *Vers les Icebergs*, rejoint d'autres espaces vides. Plus tard, le rêve de sable devenu réalité, il publie avec sa femme le journal de cette aventure : *Gens des Nuages*.

---

[214] Le Clézio, *Gens des nuages*, p. 11.
[215] Ibid., p. 19.
[216] Ibid., p. 117.
[217] Ibid., p. 110.

4.2.2 Arabesques

Derrière la polyphonie de surface et l'écheveau des identités, un ordre profond mène le récit. Caché, l'inspirateur du mouvement romanesque prend deux formes, l'une historique avec Al Azraq, l'autre spirituelle avec Es Ser. De façon analogue, une même entité semble animer plusieurs personnages, comme le chef des nomades Ma el Aïnine et le berger Hartani, ainsi que les héros adolescents Nour et Lalla.

Le livre entrelace ainsi deux récits : l'histoire collective des nomades et les aventures de Lalla. Le premier est le plus souvent cadré dans le temps et l'espace au niveau de l'en-tête, tandis que son corps est marqué typographiquement par une marge. En revanche, le second reste sans repère précis. Paradoxalement, la fiction écrite au présent a les accents de la réalité, quand les faits d'armes sont relatés au passé.

Pour chacun de ces récits, quatre temps se dessinent :

- la naissance à la réalité et la rencontre de l'initié avec son initiateur, respectivement Nour et Ma el Aïnine, Lalla et Hartani ;
- la maturation et l'éloignement lors de la marche de la caravane vers le nord et du départ de Lalla vers la France ;
- la mort des nomades et de leur chef après la bataille, ainsi que celle de l'ami de Lalla à Marseille ;
- la renaissance et le retour au désert des Hommes Bleus et de Lalla.

Ces quatre temps ne sont cependant pas de durées égales : peut-être contaminés par la modernité, les événements se précipitent dans la seconde moitié du cycle.

Le livre s'achève de la même façon qu'il commence avec un chiasme :

la fin du récit historique, « comme dans un rêve, ils disparaissaient » répond à l'incipit « Ils sont apparus comme dans un rêve » ; de même, Lalla accouche au pied d'un arbre à l'image de sa mère une génération plus tôt.

A bien regarder, il ne s'agit pas d'un retour à l'état initial : débarrassés des rêves de conquête et des mirages de la ville, les nomades et Lalla entament une nouvelle phase de leur vie. Selon une spirale ascendante, les êtres s'allègent de l'illusion et rejoignent leur essence.

Par ce jeu de répétitions, la prose de Le Clézio a les accents du récit poétique :

> Si nous reconnaissons, avec Jakobson, que la poésie commence aux parallélismes, nous trouverons, dans le récit poétique, un système d'échos, de reprises, de contrastes qui sont l'équivalent, à grande échelle, des assonances, des allitérations, des rimes[218].

Présent à toutes les échelles dans l'ensemble du livre, le rythme est particulièrement sensible dans la psalmodie du *dzikr*[219]. Egrenant les mots berbères, Le Clézio fait entendre les sons de la langue arabe[220]. Les références fréquentes au chant et à la danse dans le récit suggèrent aussi une conception musicale ou chorégraphique.

A un autre niveau, des métaphores fréquentes traduisent la vision chamanique de Le Clézio. Les frontières dépassées, les quatre éléments entrent en fusion :

> La lumière des étoiles tombe doucement comme une pluie. Elle ne fait pas de bruit, elle ne fait pas de poussière, elle ne creuse aucun vent[221].

Les extraits sélectionnés en annexe correspondent aux quatre temps

---

[218] Tadie, *Le Récit Poétique*, p. 8.
[219] Le Clézio, *Désert*, p. 56-72.
[220] Le Clézio, *L'inconnu sur la Terre*, p. 97-98.
[221] Le Clézio, *Désert*, p. 220.

évoqués. Le dernier passage semble se référer aux trois phases alchimiques noires, blanches et rouges : Lalla est devenue un être solaire, un symbole vivant qui transmet la connaissance.

5 Unités linguistiques

Le corpus présenté, il reste à définir les unites linguistiques destinées à l'analyse.

Dans une perspective stylistique, il s'agit moins de connaître avec précision la composition d'une œuvre que l'organisation de ses unités. Par ailleurs, des classes générales et larges rassemblent des populations nombreuses, gages de statistiques assurées. La prime est donc le plus souvent donnée à la simplicité pour représenter chaque plan d'analyse[222].

5.1 Graphémologie

Le plan retenu comporte trente-cinq éléments :

- l'espace ;
- les signes de ponctuation : le point final, d'interrogation et d'exclamation ; la virgule, le point-virgule et les deux-points ; les guillemets et les parenthèses ;
- les lettres de l'alphabet.

[222] Hormis dans la macroscopie : d'inspiration thématique, celle-ci détaille parfois l'analyse des unités en utilisant un grain plus fin.

5.2 Syntaxe

L'analyse porte sur les huit parties du discours, soit d'après la taxinomie du logiciel *Syntex*[223] : adjectifs, adverbes, conjonctions, déterminants, noms, prépositions, pronoms et verbes.

5.3 Sémantique

L'univers sémantique se voit décomposé en vingt-huit concepts principaux selon la nomenclature du logiciel *Cordial*[224] : fondamental, ordre et mesure, esprit, être humain, temps, action, rapport à l'autre, vie collective, vie sociale, mouvement et forces, espace, perception, volonté, communication, affectivité, art, information, vie spirituelle, morale, hiérarchie, guerre et paix, quotidien, droit, économie, matière, vie, corps et vie, santé.

[223] Cf. chapitre 3, section 4.2.
[224] Cf. chapitre 3, section 4.3.

Chapitre 2 : la mesure

Musica est exercitium arithmeticae occultum nescientis se numerare animi[225].

1 Introduction

Ce chapitre voué à la méthode commence par un aperçu général de la stylométrie. La voie choisie pour notre étude est précisée à partir de la section 3, en fonction des scopies d'analyse et des plans linguistiques.

Si cet ensemble théorique concentre l'essence de ce travail, il est susceptible de se révéler indigeste aux âmes légitimement assoiffées de concret. Le lecteur indisposé se ressourcera dans la seconde partie pour toucher et goûter les fruits de ces principes.

2 La stylométrie

Ce continent est abordé par ses fondements et ses développements, avant d'élargir le point de vue par des considérations sur le temps et l'espace.

Suit un panorama des mesures envisageables et des processus qui œuvrent en arrière-plan.

Enfin sont dessinées les ouvertures et les perspectives de cette

[225] Leibniz, *Opera omnia*, vol. 3, p. 437 : « La musique est une pratique occulte de l'arithmétique, l'âme comptant inconsciemment ».

discipline.

2.1 Principe et histoire

2.1.1 Etymologie et terminologie

L'étymologie de style est sans doute le mot grec στυλος[226], la colonne. Ce support fait de pierre exprime la force, affirme une direction, un caractère qui s'oppose à la réserve, au cycle et à la neutralité. D'un côté le I et le 1, de l'autre le O et le 0.

Ces remarques apparement banales prendront un sens nouveau à la lumière des analogies physiques faites à la fin de cette section.

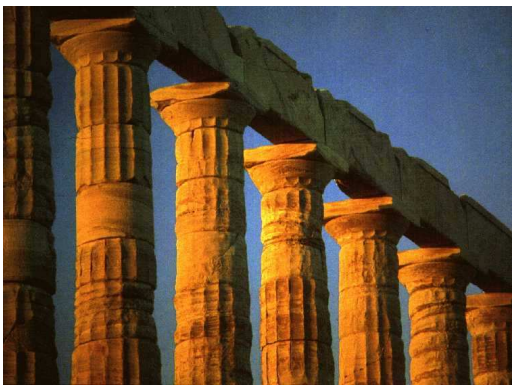


Figure 2.1 : colonnes du Cap Sounion[227]

Risquons une définition : la stylométrie consiste à faire des statistiques sur les unités d'une œuvre, puis à interpréter les spécificités de la forme ainsi découverte.

Cette approche n'est pas spécifique à la littérature. Formellement du

[226] Source : *Trésor de la Langue Française Informatisé.*
[227] Source : www.ae.gatech.edu.

moins, elle s'applique aussi à la musique, aux arts plastiques, et à toutes les manifestations du signe, à visées esthétiques ou non : c'est le bénéfice d'une mathématique abstraite et hautaine, qui appréhende conjointement des domaines parallèles.

2.1.2 Polémique et synthèse

Déjà, peut-être avec quelque raison, des voix s'élèvent : l'art relèverait du qualitatif, et la science du quantitatif. Mêler ces deux éléments serait une alchimie douteuse, une chimère romantique.

A réfléchir, la stylistique semble pourtant au milieu de ce fleuve insaisissable. La lecture d'un mot blanc provoque manifestement un effet. Mais le papillon doit se reproduire plus ou moins périodiquement pour se métamorphoser en style.

2.1.3 Limites et sauts

D'emblée, nous touchons une limite sérieuse d'une méthode qui dans l'état de l'art ne sait compter que des marques simples : l'implicite et les figures complexes de la pensée passent au travers des mailles du dispositif.

En revanche, cette approche entend dépasser la subjectivité de la pensée et la limitation de la mémoire humaine. Il s'agit en somme de dépasser le sensible, la surface par une radioscopie des profondeurs inconscientes.

2.1.4 Un peu d'histoire

Pour tenter de situer cette discipline étrange, voici quelques repères parmi ses manifestations dans l'histoire.

Pythagore (6è s. av. J-C) : peut-être le père lointain de la stylométrie. Mathématicien mais aussi philosophe, au tempérament mystique, il défend l'unité profonde du monde, derrière la profusion des formes. Unité entre l'homme et l'animal, mais aussi entre différents domaines comme l'art et la science, qu'il traduit par l'harmonie du nombre d'or.

Morgan : logicien britannique, il suggère en 1851 que la longueur des mots d'un texte pourrait révéler son auteur. L'étude est finalement réalisée en 1887 par Mendenhall[228], mais elle ne confirme pas vraiment cette intuition, et les résultats décevants dissuadent les chercheurs de poursuivre cette voie.

Markov : statisticien russe amoureux de poésie, il s'essaie aux études stylistiques. En 1913, il utilise ses processus stochastiques pour analyser les séries de lettres d'un roman en vers de Pouchkine, *Eugène Onéguine*[229].

Yule : statisticien écossais, il s'intéresse aux séries temporelles dès 1926. Associé à Georges Zipf[230], il élabore en 1932 une loi qui distribue les fréquences lexicales dans un texte. Quelques années plus tard en 1938, il propose de caractériser un auteur par la longueur de ses phrases[231] et produit en 1944 une étude statistique sur le vocabulaire

---

[228] Mendenhall., « The characteristic curves of composition »
[229] Markov, « Primer statisticeskogo issledovanija nad tekstom "Evgenija Onegina", illjustrirujuscii svaz ispytanii v cep.
[230] Zipf, « Selected studies of the Principle of Relative Frequency in Language ».
[231] Yule, « On sentence-length as a statistical analysis of style in prose, with application to two cases of disputed authorship.

littéraire[232].

Herdan : cet Anglais entend faire de la linguistique statistique une discipline autonome. Il introduit les travaux de Markov et propose une intéressante distinction entre les statistiques en masse qui prennent le texte en bloc, et celles en ligne qui intègrent la dimension temporelle[233].

Müller : soutenu par l'informatisation des pièces de Corneille, le pionnier de la statistique textuelle en France publie en 1967 une étude sur le vocabulaire théâtral[234]. Le principe consiste à mesurer les écarts d'une œuvre par rapport à la base Frantext qui sert de référence.

Benzécri : statisticien, il s'oriente en 1973 vers la représentation de données multi-dimensionnelles dans le but d'établir des correspondances entre plusieurs variables[235]. Il applique en 1981 cette méthode à la linguistique et à la lexicologie[236]. Mais la véritable ambition de cet homme complexe est philosophique *:* l'analyse des données est « un outil pour dégager de la gangue des données le pur diamant de la véridique nature[237] ».

Brunet : inspiré par les travaux de Müller et mettant à profit la ressource Frantext, il publie en 1981 une étude du vocabulaire français depuis la Révolution[238]. Lors de la célébration du Bicentenaire, il donne naissance au logiciel Hyperbase, qu'utilisera notamment Kastberg[239] pour sa thèse sur Le Clézio en 2002.

---

[232] Yule, *The statistical study of literary vocabulary.*
[233] Herdan, *Language as Choice and Chance.*
[234] Müller, *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille.*
[235] Benzécri, *L'analyse des données.*
[236] Benzécri, *Pratique de l'analyse des données, Linguistique et lexicologie.*
[237] Benzécri, *Histoire et préhistoire de l'analyse des données*, p. 144.
[238] Brunet, *Le vocabulaire français de 1789 à nos jours.*
[239] Kastberg, *L'écriture de J.M.G. Le Clézio, une approche lexicométrique.*

2.2 Temps et espace

2.2.1 Temps et fréquence

Le style semble intimement lié à la récurrence, d'où une digression sur la dualité entre le temps et la fréquence. Celle-ci traverse différents domaines.

En mathématiques, un signal est représenté par son évolution en fonction du temps ou par son spectre de fréquences : les deux points de vue sont équivalents et complémentaires, le passage de l'un à l'autre s'opérant par une transformée de Fourier[240].

En physique, cette dualité fonde la distinction entre les théories macroscopiques et microscopiques. D'un côté, l'univers lumineux et serein de Newton et Einstein fait de corps en mouvement. De l'autre, la jungle incertaine des quanta défrichée par De Broglie et Schrödinger composée d'ondes en interférence.

Entre ces deux pôles, le domaine qui nous intéresse. Au sein du langage médiateur, l'ambiguïté apparaît dans l'alternative entre les consonnes percutantes et les voyelles vibrantes. Mais aussi entre la lecture séquentielle d'un livre, de la première page à la dernière, et le parcours d'un index fréquentiel[241].

2.2.2 Texte dual

Ainsi, un texte est classiquement considéré comme une succession

---

[240] Samuelidès & Touzillier, *Analyse harmonique*.
[241] Dans ce jeu catastrophé, les termes sont rangés non par ordre alphabétique, mais par nombre d'occurrences.

d'unités qui évoluent sur différents plans linguistiques.

Mais de façon plus originale, il peut être vu comme un réseau d'ondes qui se propagent selon la transparence de ces surfaces et dessinent des interférences colorées.

2.2.3 Un espace, trois plans

Les unités usuellement observées sont schématiquement groupées en trois plans.

Au niveau phonologique, les sons s'analysent en distinguant les voyelles (antérieures ou postérieures, orales ou nasales, ouvertes ou fermées,…) et les consonnes (antérieures ou postérieures, constrictives ou occlusives, sonores ou sourdes,…). Parallèlement se placent les graphèmes : espaces, lettres et ponctuation.

Au niveau syntaxique se trouvent les parties du discours : déterminants, noms, adjectifs, pronoms, verbes, adverbes, prépositions, conjonctions, ainsi que les temps et modes verbaux.

Au niveau sémantique se déploient les dimensions, les domaines et les taxèmes. De façon plus structurelle, l'étude comporte aussi la richesse du vocabulaire (la variété des termes employés) et son niveau (simple, moyen ou élevé).

2.2.4 Vers l'inconscient ?

Qualitativement, la conscience privilégie souvent la composante sémantique, et relègue au second plan la syntaxe ou la phonologie. Cependant, l'esprit reste sensible à chacun de ces niveaux, d'où l'intérêt de les appréhender d'une humeur égale, et de mettre en évidence sinon

ce qui est inconscient, du moins ce qui est voilé.

Quantitativement, la mémoire humaine est pauvre : en fonction des individus, elle ne dépasse guère sept éléments à court terme. Si l'ordinateur connaît des pathologies malheureuses, son cerveau ignore ce handicap génétique et contribue à donner la vision globale d'un corpus.

2.2.5 Vers le continuum

Formellement, le temps se lit comme un espace. Dans les deux cas se pose la question de sa nature, discrète ou continue.

Les espaces des états et du temps sont le plus souvent discrets dans le cadre de cette étude. Mais ce propos mérite d'être modulé : l'entropie — cette grandeur d'origine thermodynamique[242] est reprise plus loin — est susceptible de se muer en « unité » et d'introduire sournoisement la continuité dans cette bergerie aux arêtes pointues. De même, le temps linguistique cadencé par des occurrences successives est la trace d'un temps physique fluide et ininterrompu.

2.3 Des mesures multiples

La première idée qui vient à l'esprit est de compter les unités et d'estimer leur fréquence relative dans le texte, réalisant ce qu'Herdan appelle des statistiques « en masse » . C'est l'objet de la stylométrie classique, et cette voie élémentaire fournit des informations significatives dans le chapitre 4.

La méthode traditionnelle trouve cependant ses limites : un texte est

[242] Cf. Clausius, *Abhandlungen über die mechanische Wärmetheorie*.

écrit et lu de gauche à droite, puis de bas en haut. Il semble donc difficile de s'abstraire de la dimension du temps, ou de statistiques « en ligne ».

2.3.1 Séquences

Pour fixer les idées, voici un exemple simple d'ailleurs étudié par Markov. Il s'agit de l'alternance des consonnes et des voyelles dans une séquence de lettres.

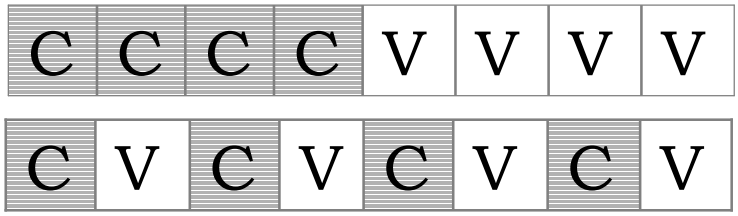


Figure 2.2 : séquences solides et alternées

Une certaine redondance se fait voir dans la représentation des séquences. Elle n'est pas gratuite : le principe s'applique également à la succession des ombres et des lumières d'un tableau, ou à la mélodie d'une partition.

La première séquence est solide et correspond à un vers classique figé. La seconde est rythmique et se rapproche d'une seconde génération de vers alternés.

2.3.2 Statistiques d'ordre n

Le simple comptage des voyelles et des consonnes ne permet pas de séparer les exemples de la figure 2.2.

D'où l'idée de recenser les bi-grammes : consonne-consonne, consonne-voyelle, voyelle-consonne, voyelle-voyelle. Cette statistique d'ordre deux permet de différencier les séquences.

Pour distinguer les faux jumeaux des vrais, rien n'empêche de faire des pas supplémentaires vers l'intimité de nos créatures, et de produire des statistiques monstrueuses sur des n-grammes. Cette fuite en avant se borne à la réalité de notre conscience : la charge cognitive de ces statistiques décroît sûrement.

Dans une conjecture fameuse, le neurologue Julesz[243] avance en 1962 que l'être humain discrimine des textures jusqu'à l'ordre deux. En 1973, il revient lui-même sur son hypothèse et apporte des contre-exemples. Plus récentes et dans le domaine linguistique, les études de Clément et Sharp[244] placent la frontière du sensible autour de sept.

Cette limite reste pertinente dans le cadre humain. Imaginons cependant qu'un ordinateur crée une œuvre étrange et complexe : sa spécificité restera invisible à l'homme, et seul un être de métal en appréciera le style.

2.4 Derrière la mesure

Ce qui précède se contente de décrire la surface. Dissimulé sous des formes profuses, l'artisan reste dans l'ombre. Essayons d'entrer en contact avec ce grand ordonnateur.

Pour illustrer le propos, voici une petite métaphore biologique : à

---

[243] Julesz, « Visual pattern discrimination ».
[244] Clement & Sharp, « Ngram and Bayesian Classification of Documents for Topic and Authorship ».

gauche du schéma (figure 2.3) se trouvent la genèse et la poétique ; au centre, le monde objectif de la forme ; en face, l'esthétique et la réception.

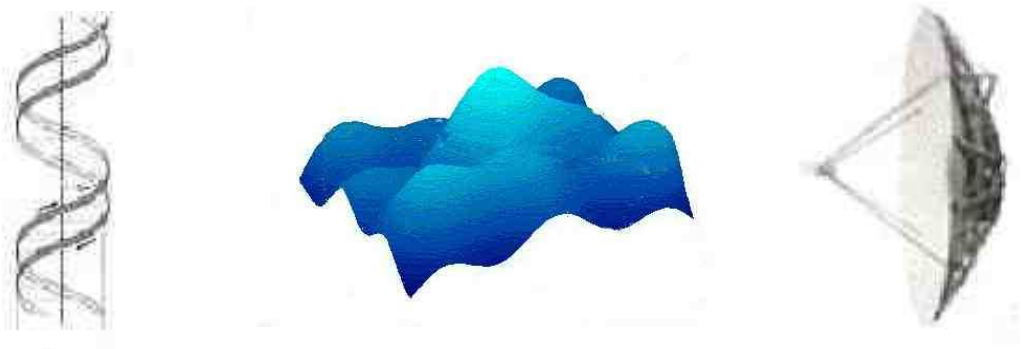


Figure 2.3 : l'ADN, la forme et le radioscope

Cette réception peut être sensible, intuitive, humaine, à l'image d'une auscultation immédiate. Mais dans certains cas, l'instrument est utile pour explorer la profondeur. Du stéthoscope au radioscope, ces média se heurtent aussi à leurs limites, si bien qu'un saut est nécessaire pour interpréter les mesures et remonter à la genèse.

Hasard et nécessité

Nos vies comme les signes semblent mus par des principes lointains et antagonistes : le futur reste incertain, mais il est orienté par le passé[245] : en liberté surveillée, pour ainsi dire.

Ces principes s'incarnent sous la forme de processus aléatoires[246] ou stochastiques qui recensent plusieurs espèces, selon la nature des

[245] Des esprits rebelles ont pu envisager d'autres schémas et entreprendre d'éclairer le passé par le présent. Ainsi, l'historien sensé expliquer le monde contemporain par ses origines, est souvent amené à utiliser les modèles actuels pour reconstituer les données perdues. Voir notamment Labov, *Principles of Linguistic Change*, p. 9-27.
[246] Pac, *Processus aléatoires*.

espaces où ils croissent et se multiplient.

Les processus considérés sont stationnaires : le modèle n'évolue pas en fonction du temps. Cette hypothèse est discutable lorsque l'écriture s'étale dans la durée — et c'est sans doute le cas des œuvres de notre corpus. Cet aspect est cependant oblitéré pour ne pas compliquer exagérément une modélisation fatalement schématique.

Parmi ces processus, les plus élémentaires vivent dans des espaces entièrement discrets. Ces êtres bondissent d'état en état par saccades. Ce sont les Markoviens, détaillés dans la section suivante.

Plus complexes sont les processus qui passent d'un état continu à un autre et ne se révélent qu'à des instants discrets. Parmi eux, les Armas[247] font aussi l'objet d'une observation attentive, plus loin dans le chapitre.

En revanche, certaines souches étranges sont quasiment absentes du continent linguistique : ainsi les ondins Poissonniens et les pollens Browniens, aux vies pleines et continues. Cette riche famille promet d'engendrer des créatures nombreuses et exotiques, cette perpective est précisée plus loin.

Etat \ Temps	Discret	Continu
Discret	*Markov*	*Poisson*
Continu	*Arma*	*Brown*

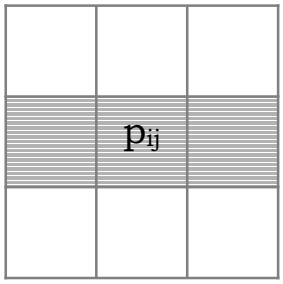
Figure 2.5 : taxonomie de processus stochastiques

[247] Ces processus sont explicités en 2.4.2.

2.4.1 Chaînes de Markov

Ce sont les processus les plus simples, mais aussi les plus utilisés en linguistique[248] : les espaces de temps et d'état sont tous les deux discrets.

La probabilité d'un état à un instant n dépend d'un nombre fini d'états passés. Une chaîne d'ordre deux est donc définie par son état initial, et une matrice de transition entre deux états successifs. Si m désigne le nombre d'états possibles, la matrice est de dimension m^2.



p_{ij} = probabilité d'atteindre l'état j si le précédent est i

$\Sigma\ p_{ij} = 1$ sur une ligne.

Figure 2.6 : matrice de transition

Apollon et Dionysos

Derrière son apparente simplicité, cette matrice renferme des éléments précieux pour notre interprétation. Balayons du regard chacune de ses lignes :

S'il existe un p_{ij} égal à 1, le présent est déterminé par le passé : les séquences solides et rythmiques de la figure 3.2 correspondent aux matrices : $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Si tous les p_{ij} sont égaux, le présent est indépendant du passé. Inspirée par le chaos, la séquence de la figure 2.7 se rapproche du vers libre ou de la musique atonale.

---

[248] Petruszewycz, *Les chaînes de Markov dans le domaine linguistique.*

Figure 2.7 : séquence chaotique

2.4.2 Processus Arma

Contexte

Ces processus sont développés par les mathématiciens Box et Jenkins[249] pour étudier des phénomènes économiques. Comme les Markoviens, les Armas vivent dans un espace de temps discret. En revanche, leurs états prennent des valeurs continues : la palette des signaux s'enrichit.

Pawlowski[250] les utilise dans un cadre linguistique et stylistique. Le Graal poursuivi ici est l'attribution d'auteur parmi les textes de Romain Gary et d'Emile Ajar. En l'occurrence, ces processus modélisent la distance entre les mots grammaticaux, la longueur des phrases, et l'entropie.

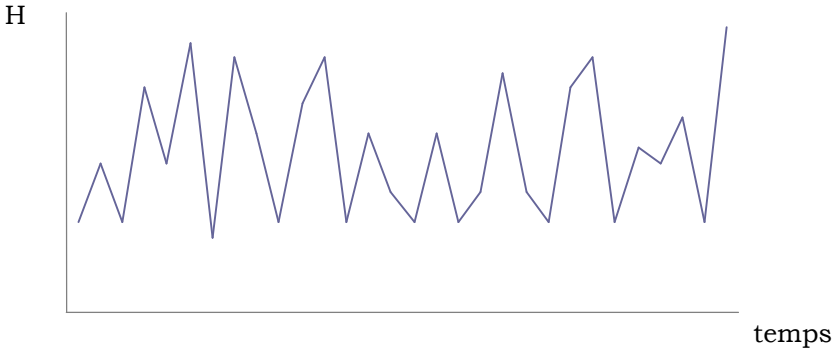


Figure 2.8 : oscillations entropiques

A titre d'illustration, la figure 2.8 montre les « oscillations » typiques de cette dernière grandeur, qui traduit ici l'alternance entre les unités

[249] Box, Jenkins & Reinsel, *Time series analysis : forecasting and control.*
[250] Pawlowski, *Séries temporelles en linguistique* .

rares et fréquentes[251].

Prolifiques, les Armas se retrouvent dans bon nombre de logiciels de statistiques, ce qui ne retire rien à leur charme. Il sont détaillés plus loin dans ce chapitre, dans la section consacrée à la nanoscopie.

2.4.3 Processus à temps continu

Ce sont sans doute les moins utilisés en linguistique. Certains voient des états discrets comme celui de Poisson, mais d'autres sont entièrement continus comme celui de Brown : un nouveau cap vers la richesse et la fluidité est franchi.

Même si les données textuelles sont discrètes, rien n'empêche d'inférer un modèle continu à l'image des phénomènes humains sous-jacents, et des ondes cérébrales qui rythment l'activité de l'écrivain ou du lecteur.

2.5 Ouvertures

Au-delà de l'analyse stylistique stricto sensu, les processus stochastiques trouvent des débouchés dans différents domaines : formellement, seule la dimension des espaces considérés varie : un pour le texte et la musique, deux voire trois pour les arts plastiques.

Un premier ensemble d'applications, de nature esthétique, comprend l'attribution d'auteur déjà évoquée, mais aussi la reconnaissance de la voix ou de formes géométriques. Parmi d'autres, les travaux de Khmelev

[251] Une définition formelle de l'entropie est donnée dans la section 5 consacrée à la microscopie.

et Tweedie[252].

Inversement, dans une fonction poétique, les processus peuvent s'employer pour générer du texte, de la musique ou une image. Par exemple, les textures de la figure 2.9 ont été crées par un processus Markovien[253]. Leur aspect rythmique ou chaotique dépend du réglage de ses paramètres.

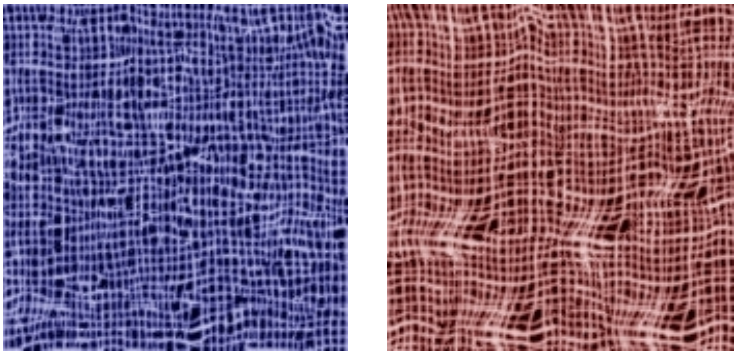


Figure 2.9 : textures markoviennes

2.6 Perspectives

Plus loin dans l'avenir, il reste à inventer une *sémiodynamique*, une « physique » du signe qui modélise son évolution. Voici quelques pistes de réflexions, sans aller jusqu'au bout.

2.6.1 Newton

Revenons quelques siècles en arrière, avec la relation fondamentale de la dynamique de Newton[254] : $F = d^2x/dt^2$, le terme de droite désignant l'accélération de x :

– à gauche de l'égalité s'exprime l'origine du mouvement ou du texte,

[252] Khmelev & Tweedie, « Using Markov Chains for Identification of Writers ».
[253] Paget & Longstaff, « *Texture synthesis via a Non-parametric Markov Random Field* ».
[254] Newton, *Philosophiæ naturalis principia mathematica*.

la force physique ou l'auteur[255] ;

– à droite se trouve sous une forme dérivée la description du mouvement ou du texte, issue de la cinématique ou du structuralisme.

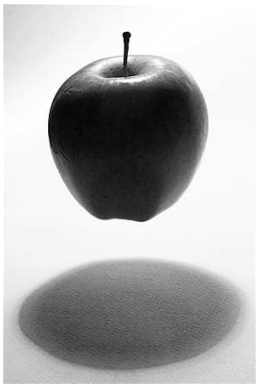


Figure 2.10 : force[256]

Apparaît alors une analogie troublante, qui entre en résonance avec l'intuition d'Henri Meschonnic[257] : *Le sujet est le rythme.*

Le propos mérite cependant d'être nuancé, à la lumière de ce qui précède : le « sujet » est naturel ou artificiel, tandis que la dynamique est faite de rythme et de chaos.

2.6.2 Einstein

L'espace de Newton est primitif, fait de droites et de plans. Face à la complexité des lois qui régissent les corps, Einstein repousse dans sa relativité générale[258] la difficulté vers la géométrie : quitte à déformer l'espace, la force de gravité s'évanouit, et le mouvement de chaque objet devient, dans une simplicité biblique, rectiligne uniforme.

---

[255] Cf. la remarque initiale du chapitre et l'étymologie du syle.
[256] Source : chiaroscuro.baltiblogs.com
[257] Meschonnic, *Critique du rythme.*
[258] Einstein, *La théorie de la relativité restreinte et générale.*

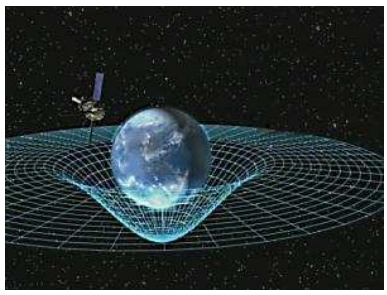


Figure 2.11 : déformation de l'espace-temps[259]

La quantification de la physique va en sens inverse : dans un espace simplifié et discrétisé, des lois complexes renoncent au déterminisme : il n'est plus question de prévoir l'avenir, mais au mieux d'estimer la probablilité d'un événement. C'est peut-être la situation de la linguistique qui recourt abondamment aux processus aléatoires.

D'où l'idée de suivre la voie d'Einstein, et de revenir à la continuité pour cesser de « jouer aux dés » dans le langage, ou du moins limiter la consommation de cette drogue. Dans cette analogie, le corps en mouvement joue le rôle du signe, la masse astrale ou stellaire celui de la quantité d'information contextuelle, et l'espace celui de la structure générale qui accueille ces éléments. A travers cette trame déformée, le flux de la quantité d'information par unité de temps devient localement constant.

La poésie va quelquefois plus vite que la science. Voici une phrase troublante de Le Clézio qui émerge dans *Icebergs* :

> Oui, on est arrivés au lieu de la naissance du langage, là où il n'y a plus qu'un seul mot, un mot intense et bref, un mot fixe qui brille comme cette étoile[260].

Ce mot immuable ou répété à l'infini ressemble étrangement à la course tranquille du corps isolé dans l'espace…

---

[259] Source : www.sciencedaily.com
[260] Le Clézio, *Vers les Icebergs*, p. 36.

2.6.3 Guillaume, Merleau et les autres

Dans l'ombre de Saussure, la psychomécanique de Guillaume gagne à être connue. Inspirée par la phénoménologie, elle s'attache à créer des liens entre la pensée et le langage. La conscience et son expression progressent ainsi du large vers l'étroit, puis de l'étroit vers le large : ce double mouvement se révéle notamment dans l'emploi alterné des articles, entre le « un » universel et un « le » particulier[261].

Cette oscillation en évoque une autre, mise en lumière par Merleau-Ponty :

> Pensée et parole s'escomptent l'une l'autre. Elles se substituent continuellement l'une à l'autre. Elles sont relais, stimulus l'une pour l'autre. Toute pensée vient des paroles et y retourne, toute parole est née dans les pensées et finit en elle[262].

Dans une synthèse romantique, ce ne sont d'ailleurs pas seulement la pensée et la parole qui sont réunies, mais aussi la parole et le corps :

> Les mots, les traits, les couleurs qui m'expriment sortent de moi comme mes gestes par ce que je veux faire[263].

Il s'agit d'une stylistique holistique, par opposition à ses versions analytiques. Combinaison des idées de Guillaume et Merleau-Ponty, le schéma de la figure 2.12 apparaît.

De part et d'autre du plan du langage se déploient ceux du corps et de l'esprit. L' énergie circule entre les trois pour inscrire un signe à chaque passage : dans la montée, elle s'incarne sous les traits d'une consonne qui involue du large vers l'étroit ; dans la descente, elle se

---

[261] *Le problème de l'article et sa slotion dans la langue française.*
[262] Merleau-Ponty, *Signes*, p. 25.
[263] Merleau-Ponty, *La prose du monde*, p. 122.

sublime en voyelle qui évolue de l'étroit vers le large[264]. A un niveau supérieur se place l'alternance de noms et de verbes, ou celle de concepts opposés.

Figure 2.12 : cycle de naissances et de morts du signe

Dans ce va-et-vient, l'aiguille de l'énergie tisse une toile qui unit organiquement ces trois plans : le texte et son étymologie apparaissent sous un jour nouveau.

2.6.4 Tentative de synthèse

Essayons de mettre en relation ces conjectures. Une onde fondamentale circule entre les plans du corps, du langage et de l'esprit. Coupant celui du langage, elle laisse une trace discrète de son essence continue. Déformée par la masse d'information contextuelle, notre perception détachée observe du chaos tandis que de son propre point de vue, un signe original se répète à l'infini.

Après ce panorama de la stylométrie, quels sont les chemins empruntés par cette étude ?

---

[264] A ce mouvement vertical, on pourrait associer une dynamique horizontale entre nos hémisphères cérébraux.

Dans chaque phase, de la macroscopie à la nanoscopie, la mesure individuelle propice à l'analyse est d'abord définie. Une mesure de synthèse est ensuite proposée pour l'intelligence des phénomènes en jeu.

3 Macroscopie

Dans cette première phase, chaque œuvre est vue comme un bloc, une boîte grise laissant filtrer son contenu, c'est-à-dire sa composition selon les unités d'un plan linguistique. Cependant, l'organisation ou la structure restent dans l'ombre.

3.1 Mesure unitaire

En général, les mesures évoluent linéairement avec la taille des ensembles considérés. De façon à comparer des textes de longueurs différentes, des fréquences relatives sont mises à contribution, soit le nombre d'occurrences d'une unité ramené au nombre d'occurrences toutes unités confondues.

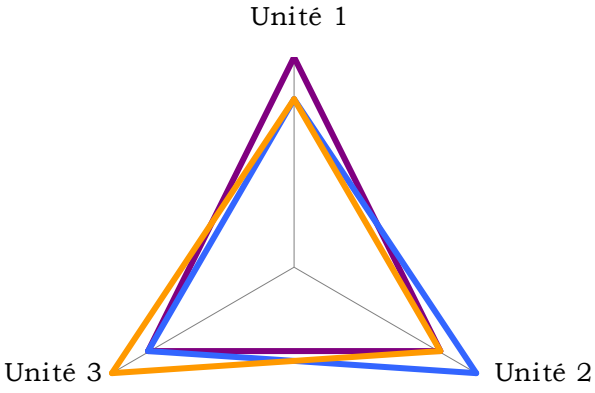


Figure 2.13 : stylogramme

Le « stylogramme[265] » de la figure 2.13 permet d' illustrer la variété des fréquences en fonction des œuvres. Il représente des excentricités ou des écarts par rapport à une référence qui fixe l'échelle de chaque axe.

Comment choisir cette référence ? Une première idée est d'adopter celle du corpus, vu comme la juxtaposition de ses trois composantes. Cela revient à ajouter un quatrième acteur dans notre jeu, donnant chair à un être artificiel dépourvu de toute unité organique et littéraire. En outre, d'un point de vue mathématique, ce procédé favorise les œuvres volumineuses au détriment des brèves dans le calcul de la référence. Or il sagit de mettre chaque oeuvre sur un pied d'égalité pour amorcer la comparaison. On lui préfére donc une référence équanime qui respecte chaque entité textuelle, d'où le recours à une moyenne tripartite[266].

A ce stade, les écarts pourraient s'évaluer en calculant les différences par rapport à cette référence. Or la physique et la psychologie semblent souvent plus sensibles aux proportions[267]. Chaque fréquence est donc divisée par la référence, de sorte que les points de mesure sont centrés autour de la valeur 1 sur chaque axe du stylogramme.

3.2 Synthèse des mesures

Les mesures qui précèdent restent fragmentées, d'où l'intérêt d'algorithmes qui les rassemblent dans un espace de dimension adaptée.

---

[265] Guiraud & Kuentz, *La stylistique*, p. 217-222.
[266] Pour une unité donnée, la moyenne des fréquences sur chacune des trois oeuvres.
[267] Citons ici les expériences de Pythagore sur sa monocorde : divisant la longueur par deux, on multiplie la fréquence du son par ce même rapport, et l'on passe à l'octave supérieure.

Inspirée des travaux de Benzécri sur l'analyse des données[268], la méthode a été introduite en stylistique par François-Charles Gaudard lors de sa thèse[269].

Soit le corpus formé de I œuvres, analysées selon J axes de mesures[270]. Chaque élément i est vu comme un point M_i d'un espace de dimension J, muni d'un repère et d'une origine O.

Désignons par x_{ij} la proportion de l'unité j dans l'œuvre i. L'ensemble des points est inclus dans l'hyperplan d'équation $\Sigma\, x_{ij} = 1$: les vecteurs OM_i sont donc mécaniquement ramenés à une échelle commune.

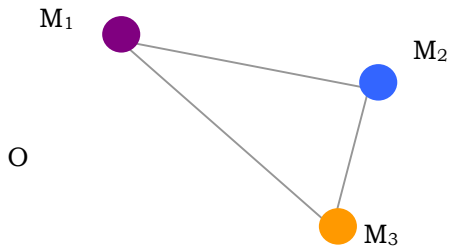


Figure 2.14 : distances

Les distances mutuelles se calculent simplement par :
$|M_{i1}M_{i2}| = (\Sigma\, (x_{i1j}-x_{i2j})^2)^{1/2} \quad (j = 1, J)$.

Dans le cas général, le nuage de points M_i est projeté sur son plan d'inertie, ce qui permet de le cartographier au prix des erreurs de perspective[271]. Ici, les trois points du corpus définissent un plan, support d'une carte fidèle.

---

[268] Benzécri, *Pratique de l'analyse des données, Linguistique et lexicologie.*
[269] Gaudard, *Contribution à l'analyse des discours littéraires : exploration stylistique de l'espace poétique Baudelairien.*
[270] Pour fixer les idées dans le cas des lettres, J est égal à 26.
[271] L'analyse arborée de Luong est une alternative intéressante, mais elle produit des graphes et non des cartes : les distances entre deux points ne sont pas représentées directement mais interprétées par le nombre de segments à parcourir.

Abitrairement, M_1 est placé à l'origine et M_1M_3 est choisi comme premier vecteur de base. Une triangulation donne la position de M_2 : $\cos \theta = (|M_1M_2|^2 + |M_1M_3|^2 - |M_2M_3|^2) / (2\ |M_1M_2|.|M_1M_3|)$.

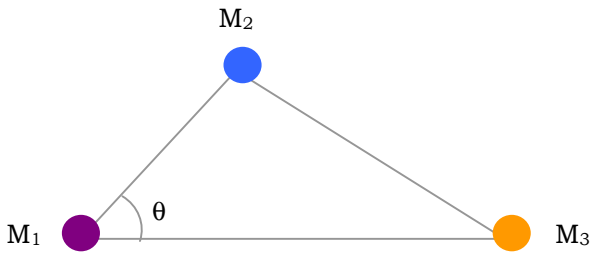


Figure 2.15 : carte

4 Mésoscopie

Si la phase précédente fournit une information intéressante sur le corpus, elle privilégie les variations dans l'espace selon les plans et les unités, mais néglige celles liées au temps. Implicitement, elle opère une dissymétrie dans l'appréhension de l'univers linguistique. Ce qui suit tente de rétablir l'équilibre par l'intégration de cette dimension supplémentaire : des premières indications sur la dynamique du corpus sont ainsi données.

4.1 Mesure unitaire

4.1.1 Mélodie textuelle[272]

Une œuvre est considérée comme la succession de ses divisions naturelles. Chacune d'entre elles est vue comme un bloc homogène,

[272] D'intéressants croisements sont rencensés dans un article de Molino, *La musique et les nombres* (in Chouvel & Lévy, « *Observation, analyse, modèle : peut-on parler d'art avec les outils de la science ?* »).

traité avec la méthode macroscopique. L'évolution de la fréquence[273] d'une unité dessine de la sorte une « portée musicale » (fig. 2.16) :

Figure 2.16 : portée musicale

Cette portée donne une autre information, la durée de chaque note, blanche ou noire : en l'occurrence, elle correspond à la taille de chaque division.

L'axe vertical des fréquences musicales se gradue traditionnellement par une échelle logarithmique : d'une octave à l'autre, les valeurs doublent. Ce modèle, partagé par de multiples cultures à travers les âges, trouve son origine tant dans la physique — la longueur des cordes de Pythagore — que dans la physiologie — la perception de l'oreille. Il s'agit d'un nouvel avatar du logarithme, indirectement évoqué avec l'entropie : ce point est précisé dans la section 5.1.1.

4.1.2 Variabilité

A partir de cette suite de fréquences, les écarts-types sont calculés, les moyennes restant bien entendu identiques à celles de la phase précédente[274]. La variabilité se définit alors comme le rapport de l'écart-type à sa moyenne.

Cette mesure quelque peu délaissée fournit pourtant de précieux

[273] Analyse temps-fréquence en langage mathématique.
[274] On passe ainsi de statistiques d'ordres 1 à 2.

renseignements sur le texte, voire son auteur : un caractère instable est susceptible de produire inconsciemment des œuvres plus variables. Le phénomène peut d'ailleurs aussi être le fruit de la volonté, un choix esthétique. Quoiqu'il en soit, il s'agit sans doute ici d'un fait stylistique majeur.

4.2 Synthèse des mesures

Il s'agit à présent de mesurer la progression des divisions dans l'œuvre, et d'estimer la variabilité globale.

Par rapport au formalisme de la macroscopie, la composante temporelle ajoute une dimension aux espaces considérés : les données d'une œuvre i, précédemment stockées dans des vecteurs de dimension J au chapitre précédent, sont maintenant rassemblées dans des matrices de dimension J x K_i, où K_i désigne le nombre de ses divisions.

Géométriquement, dans un espace de dimension J, les points M_{ik} forment un nuage de K_i éléments autour de leur centre de gravité M_i.

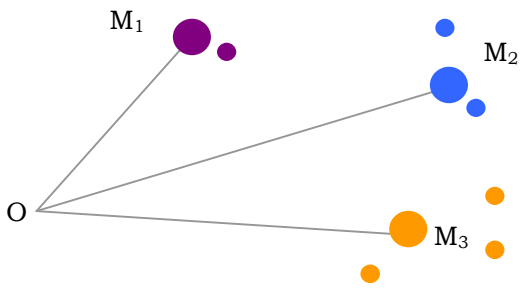


Figure 2.17 : nuées de divisions

4.2.1 Dynamique

Comment trouver une dynamique interne pour chaque roman, en

selon ses divisions ? La macroscopie permet de situer deux œuvres entre elles. Il suffit alors d'appliquer la méthode à une échelle plus petite, en évaluant les distances de chaque division par rapport à une référence, ici l'œuvre complète.

Soit $d_{ik} = |M_{ik}M_i|$ la distance de la division k à sa référence i. Pour faciliter la comparaison, cette grandeur est normalisée par une valeur moyenne, soit $d_i = (\Sigma\ n_{ik}/n_i\ d_{ik}^2)^{1/2}\ \ (k = 1, K_i)$, où n_{ik} est la taille de la division considérée et n_i celle de l'œuvre.

Cette seconde référence n'est autre que l'écart-type global de l'œuvre i, défini dans le paragraphe suivant[275].

4.2.2 Variabilité

Soit x_{ijk} la proportion de l'unité j dans la division k de l'œuvre i. La moyenne m_{ij} et l'écart-type s_{ij} de l'unité j sont estimés par :

$m_{ij} = \Sigma\ n_{ik}/n_i\ x_{ijk}\ \ (k = 1, K_i)$[276]

$s_{ij} = (\Sigma\ n_{ik}/n_i\ (x_{ijk} - m_{ij})^2)^{1/2}\ \ (k = 1, K_i)$

d'où la variabilité $v_{ij} = s_{ij}/m_{ij}$.

Sur l'axe j, M_{ij} est vu comme le centre de gravité ou barycentre des points M_{ijk} lestés des masse n_{ik}, et s_{ij} comme la distance moyenne de ces points par rapport à cette référence.

Dans l'espace de dimension J, les valeurs globales pour chaque œuvre sont alors :

---

[275] Il suffit d'inverser les sommations par rapport aux divisions et aux unités. Nous retrouvons donc avec bonheur la symétrie entre le temps et l'espace déjà évoquée.
[276] Cette valeur se confond avec la fréquence relative de la macroscopie x_{ij}.

$m_i = (\Sigma\ m_{ij}^2)^{1/2}$ (j = 1, J)

$s_i = (\Sigma\ s_{ij}^2)^{1/2}$ (j = 1, J)

$v_i = s_i/m_i$.

Géométriquement, l'écart-type s'interprète comme la distance moyenne des points M_{ik} autour de leur centre M_i, tandis que la variabilité compare cette dispersion à la référence $|OM_i|$.

Finalement, la variabilité d'une œuvre est normalisée par sa moyenne sur le corpus pour être représentée sur les stylogrammes.

5 Microscopie

Ce chapitre plonge dans la matière littéraire pour rejoindre le temps de l'écrivain et du lecteur : un nouveau pas vers la fluidité et la continuité numérique est fait.

5.1 Mesure unitaire

Isolons par la pensée une unité, par exemple un « a », un nom ou un concept.

5.1.1 Abondance ou rareté ?

La considération de la fréquence implique un choix souvent délicat, les intervalles sur lesquels elle est calculée. En quête de finesse, un découpage plus fidèle suit les occurrences successives de chaque unité (fig. 2.18) :

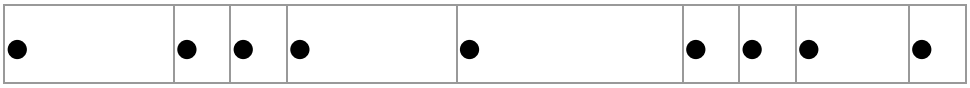


Figure 2.18 : découpage

La première idée est d'appliquer la méthode mésoscopique et de calculer une fréquence sur chaque intervalle microscopique. Un autre paramètre intervient alors.

Dans le champ de la figure 2.19, les points verts sont évidemment beaucoup plus nombreux que les rouges, tandis que les distances entre voisins de même couleur suivent une loi inverse. Néanmoins, l'observateur pourra oublier la prairie et se focaliser sur les coquelicots, non en raison de leur charme intrinsèque, mais plus prosaïquement de leur rareté et de la quantité d'information qu'ils offrent : c'est la théorie développée par Shannon[277] à partir de la notion d'entropie : I = - log P, où I désigne la quantité d'information et P la probabilité d'un événement.

Figure 2.19 : paradoxe du coquelicot

Pour en revenir au domaine de cette étude, l'abondance verse sans doute plus dans la linguistique, tandis que la rareté est d'une essence

[277] Shannon., « Prediction and Entropy of Printed English »

plus stylistique. D'où le choix de considérer non plus des fréquences, mais des distances ou des temps de retour[278], supposés ici indépendants les uns des autres.

Cette variable posée, comment l'interpréter, la quantifier, et la modéliser ? C'est l'objet des lignes qui suivent.

5.1.2 Interprétation

5.1.2.1 Structure géométrique

Levons les yeux : dans le ciel noir, les constellation apparaissent. Certaines présentent de belles régularités, d'autres sont plus chaotiques, et l'astronome les classe en amas de styles voisins. Plus proche du monde humain, le géomètre groupe des cercles ou des triangles semblables. A partir des réseaux atomiques, le chimiste assemble les cristaux. De façon générale, comment réaliser une telle taxinomie ?

Faisons une digression mathématique. Gallois résout les équations polynomiales par des considérations de symétrie :

> Ce qu'il importe de connaître, c'est par quelles substitutions peuvent être invariables les relations entre les racines[279].

En d'autres termes, ce sont moins les éléments qui comptent que leurs relations, et les transformations qui conservent ces relations. Gallois fonde ainsi la théorie des groupes en 1831[280].

---

[278] Précisément, le temps de premier retour, car il est aussi envisageable d'analyser le temps du second, ou plus généralement du n^{ième} retour.
[279] Gallois, *Œuvres Mathématiques*.
[280] Le structuralisme de Saussure en est peut-être l'écho.

Klein transpose ces idées à la géométrie et écrit en 1872 le programme d'Erlangen[281] : dans cette nouvelle approche, les figures sont mises en arrière plan au profit des opérations qui leur sont appliquées, d'où la ligne de partage entre la géométrie euclidienne et fractale[282] : dans la première, les distances sont conservées par des transformations « rigides » ou isométries telle que les translations, rotations ou symétries. Dans la seconde, la forme se maintient à travers des dilatations ou des homothéties. Ces deux types de transformations sont réunis par le groupe des similitudes.

Repartons vers le ciel pour illustrer ce propos : une constellation observée à des heures différentes est transformée par une rotation, tandis que deux amas identiques situés à des distances distinctes sont homothétiques.

Pour appréhender l'ensemble de ces transformations d'un regard, la figure 2.20 dessine des triangles similaires, communément transformés par une translation, et selon le cas par une rotation, une symétrie ou une homothétie.

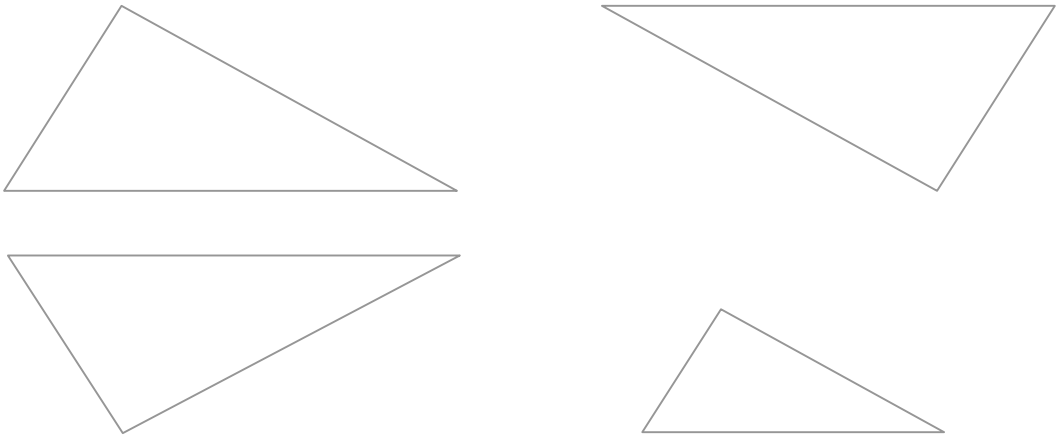


Figure 2.20 : triangles similaires

[281] Klein, *Vergleichende Betrachtungen über neuere geometrische Forschungen.*
[282] Développée par Mandelbrot, *Les objets fractals : formes, hasard et dimension.*

D'où deux concepts linguistiques, qui reprennent la dichotomie de Saussure entre les axes paradigmatiques et syntagmatiques :

– le thème est la composition d'une oeuvre en fonction de ses unités[283] ;
– le style est l'organisation spatio-temporelle de ces unités.

En d'autres termes, le thème définit les éléments d'une œuvre, et le style est la structure qui les accueille. Fondamentalement, cette dernière reste invariante par le groupe des similitudes correspondantes.

Dans le cas de la linguistique ou de la musique, l'espace-temps est de dimension 1 et les similitudes se réduisent aux translations, symétries et homothéties. Mais pour les arts plastiques, les dimensions considérées sont supérieures — deux pour la peinture, trois pour la sculpture — et il faut alors inclure les rotations.

Thème et style

Ce thème généralisé ne se restreint pas au domaine sémantique, comme le sens commun pourrait l'entendre. Ainsi, la langue allemande utilise volontiers les lettre K, W ou Z. Cet emploi fréquent, purement contextuel voire accidentel, ne relève pas du style selon la définition précédente. De même, imaginons un démiurge chargé en haut lieu d'apporter la vie sur une planète désolée. Dans son lointain exil, il forme ses créatures avec la glaise qui est à la portée de sa main. Or cette ressource est contingente, et l'on cherche des analogies entre les formes qu'il organise sur différentes planètes, en d'autres termes le style.

---

[283]On pourra rapprocher le θεμα grec — ce qu'on pose — de la composition. Le thème est donné avec les billes jetées sur le sol, alors que le style est formé par le joeur qui les dispose.

La définition donnée à ce dernier concept est fort classique et cohérente avec celle du *Trésor de la Langue Française Informatisé* : « catégorie de l'esthétique permettant de caractériser l'organisation de formes verbales, plastiques, musicales ». Elle précise cependant la notion souvent floue d'organisation par le critère d'invariance et de similitude.

Cette approche semble diviser en deux caissons étanches le thème et le style. C'est effectivement le cas pour un niveau donné, mais la perspective change quand le point de vue s'élargit.

A partir des éléments et de leur organisation, le pont supérieur est bâti, des lettres aux mots puis des mots au sens. En d'autres termes, le style apparaît comme un catalyseur, un ferment qui associé au thème fait lever celui du niveau suivant (fig. 2.21).

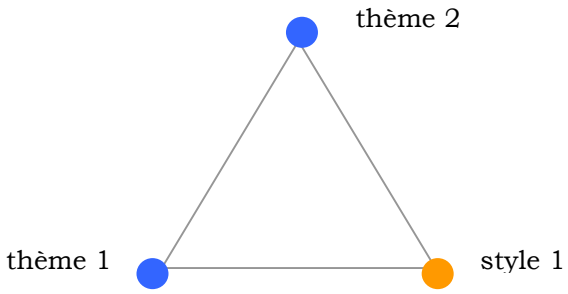


Figure 2.21 Abyme génétique

Si ces deux acteurs se complètent, ils n'ont pas la même complexité : soit un livre soit formé de n éléments, chacun pouvant prendre j valeurs. Il y a n^j possiblités de les choisir, mais n ![284] façons de les organiser : le premier terme cède rapidement le pas au second[285].

[284] n ! = 1.2...(n-1).n ou factorielle n.
[285] L'approxmation de Stirling pour la factorielle donne : $n ! \sim (2\Pi n)^{1/2} n^n e^{-n}$.

Une portée limitée

Cette définition formelle évacue avec un certain soulagement des notions problématiques : elle n'exclut pas des interactions entre la structure et son auteur, ni d'ailleurs son lecteur, mais celles-ci restent dans l'ombre : ainsi, une œuvre est vue comme filiation d'un créateur dans un contexte irrémédiablement mouvant ; de même, sa réception reste sujette à des variations impondérables. La comparaison entre deux auteurs se restreint donc à leurs écrits.

Dans sa marche vers l'étroit, la stylistique semble franchir un nouveau seuil. Historiquement, le style a successivement caractérisé une langue nationale, un genre, un auteur. Le voici congru à une œuvre, ou plus exactement à la tribu des œuvres qui lui sont similaires. Sa portée dépasse donc une création particulière, comme notre intuition le souffle, mais dans un sens précis.

5.1.2.2 Relativité

Le principe structural acquis, faisons un détour vers la physique et la cinématique.

Figure 2.22 : mouvement uniforme

La figure 2.22 marque la position d'un objet à des instants successifs. Les points sont alignés et les distances voisines sont identiques : le corps parcourt librement l'espace à vitesse constante.

Indépendamment de la valeur de cette vitesse, la théorie mécanique considère que tous ces mouvements sont équivalents en raison de leur accélération nulle : c'est un nouvel avatar de l'invariance par une dilatation du temps.

D'où l'idée de Newton[286] : définir la dynamique à partir de ce mouvement une référence.

Figure 2.23 : mouvement accéléré

Sur la figure 2.23, les distances entre deux points voisins sont en moyenne identiques à celles de la figure 2.22. Mais d'une logique fort différente, elles évoluent et décrivent une succession d'accélérations et de décélérations.

Classiquement, la mécanique explique ce mouvement par l'action de forces. Dans sa théorie de la relativité, Einstein géométrise ces forces et interprète le phénomène comme les déformations de la structure spatio-temporelle :

> C'est pourquoi on utilise des corps de référence non rigides, qui non seulement se meuvent dans leur ensemble d'une façon quelconque, mais qui subissent aussi pendant leur mouvement des changements de forme quelconques [...] Chaque point du mollusque est traité comme un point de l'espace, et chaque point matériel qui est immobile par rapport à lui est tout simplement traité comme immobile, tant que le mollusque est traité comme corps de référence. Le principe de relativité générale exige que tous ces mollusques puissent être employés, avec un égal droit et un égal succès, comme corps de référence pour la formulation des lois générales de la nature ; les lois elles-mêmes doivent être tout à fait indépendantes du choix du mollusque[287].

Revenons à la linguistique : les points des figures précédentes

---

[286] Newton, *Philosophiæ naturalis principia mathematica*
[287] Einstein, *La théorie de la relativité restreinte et générale*, p. 110.

symbolisent alors les apparitions d'une unité. Par analogie, un signe imaginaire livré à lui-même suit une trajectoire uniforme : c'est le *degré zéro du style*. A partir de cette vitesse de référence, un style particulier se traduit par une accélération née d'une force ou d'une déformation de la structure spatio-temporelle[288].

Ce modèle se généralise à des espaces de dimensions supérieures. Avec deux unités linguistiques (fig. 2.24), la structure est comme une nappe posée sur une table. Au repos, sa trame est symétrique et régulière, tandis qu'un style propre se caractérise par des contractions ou des dilatations locales.

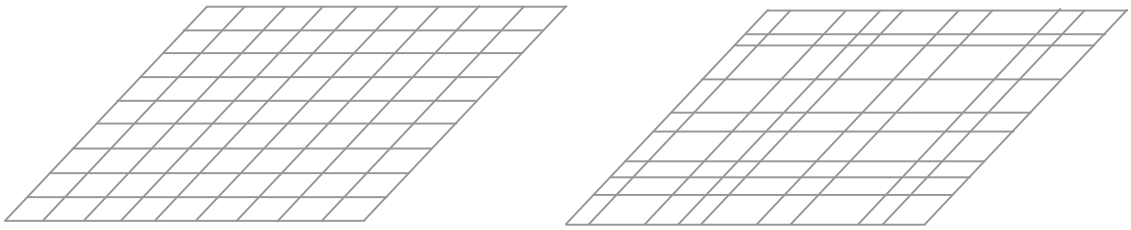


Figure 2.24 : trames

Si le thème est une vitesse globale, le style est une accélération locale.

Avec une autre image, le premier est le mouvement d'ensemble d'un solide, et le second la rotation autour de son centre.

5.1.3 Quantification

Soit X le temps de retour d'une unité, et m sa moyenne sur l'œuvre.

[288] Evidemment, le parallèle s'arrête ici : dans un cas, la physique établit un lien précis entre le mouvement et son origine ; dans l'autre, la relation entre le verbe et son auteur reste plus lâche.

La variable centrée X-m est invariante par une translation ou une symétrie dans le temps, mais elle reste sensible à une homothétie. Elle est donc normalisée, d'où l'écart-relatif : x = (X-m)/m.

Sur de petits intervalles de temps et un nombre limité d'unités, la représentation de cette variable donne une image fidèle du style précédemment défini. Il n'est évidemment pas question de procéder ainsi sur l'ensemble d'une œuvre et de ses unités, d'où le recours à des statistiques qui synthétisent ces données.

5.1.3.1 Moment statistiques

Le moment d'ordre k de la variable x est la moyenne de la variable x^k, soit m_k. Sa version réduite permet d'éliminer l'effet de puissances croissantes sur x, soit $\mu_k = m_k^{1/k}$.

Par définition de l'écart relatif x :

- μ_1 est nul ;
- μ_2 évalue la variabilité de X ;
- μ_3 estime l'asymétrie de X, grandeur expliquée plus loin dans le paragraphe consacré aux spectres.
- les moments d'ordres supérieurs comme l'aplatissement sont moins palpables physiquement et ne sont pas abordés.

Comment faire parler ces notions ? En musique, la moyenne fait entendre le tempo, et les moments relatifs d'ordres supérieurs le rythme. Si la première statistique semble avoir pâli d'un point de vue stylistique, cette couronne jaunie n'en reste pas moins fondamentale, elle acquiert même le statut de *référence interne* : caractéristique d'une œuvre, elle devient son métronome, un « la » régulier autour duquel s'organise le mouvement.

Fondamentalement, le style est une asymétrie, une arythmie par rapport au tempo imprimé par le thème.

Quantité d'information

Ces considérations s'appliquent à la quantité d'information, formulée classiquement par I = - log P, où P est la probabilité d'un événement. Or la distance X mesure la rareté et reste supérieure ou égale à 1 : elle est donc assimilée à l'inverse de P, d'où la relation : I = log X.

Cette valeur brute se compare alors à sa référence, I_0 = log m, d'où la quantité d'information relative I_r = log X/m. Lorsque X est voisin de m, I_r tend vers l'écart relatif x, c'est à dire l'arythmie.

I_r est reste nulle quand les occurrences d'une unité linguistique sont parfaitement réglées et rythmées, mais se fait sentir quand les apparitions deviennent plus chaotiques : l'horizon d'attente du lecteur est troublé et l'effet de surprise de Jauss[289] se manifeste sous un nouveau jour.

Le style est lié à la quantité d'information relativement à la référence du thème.

5.1.3.2 Spectres

Les moments qui précèdent restent parfois flous. Faisons un pas de plus pour interroger les signes littéraires.

Par définition, la distribution d'une variable représente la probabilité d'obtenir une valeur, en fonction de cette valeur. En l'occurrence, la

[289] Jauss, *Pour une esthétique de la réception.*

grandeur est un temps de retour, une période, si bien que la courbe s'interprète comme un spectre. Elle représente la fréquence à laquelle vibre cette unité, sa « musique ».

La figure 2.25 donne quelques exemples de spectres calculés sur l'écart relatif x[290].

- Le premier voit tous ses moments nuls : sa raie concentre l'énergie d'un mouvement parfaitement périodique.
- Le second a la même moyenne nulle, mais sa variabilité disperse l'énergie et élargit la courbe, qui reste symétrique autour de l'axe des ordonnées ;
- Le troisième conserve la moyenne et la variabilité du précédent, mais présente une asymétrie, une cambrure : le sommet de la vague est soufflé par un vent orienté à l'ouest, tandis qu'un courant profond entraîne sa base vers l'est. Le moment qui résulte de ces flux contraires est positif, les valeurs extrêmes de x prenant l'ascendant sur les médianes. Sur le fond, cette asymétrie trouve souvent son origine dans une contrainte physique : un temps de retour est toujours positif, et son écart relatif toujours supérieur à -1.

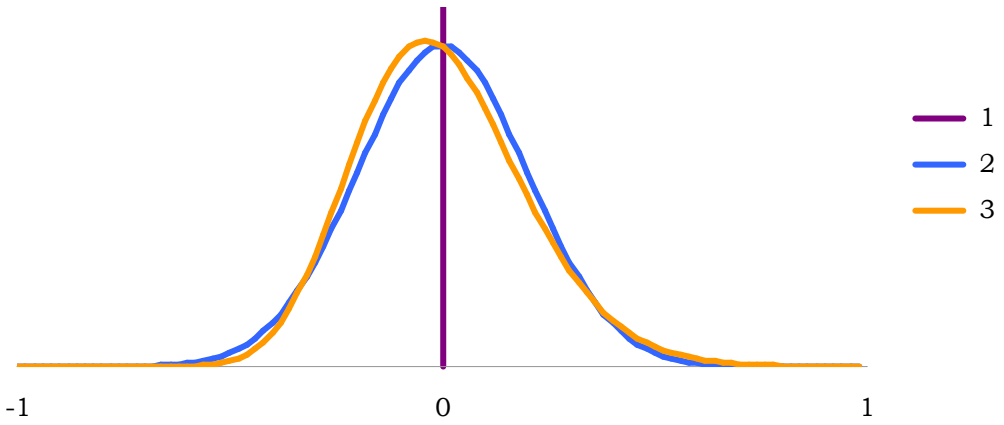


Figure 2.25 : spectres

[290] Les spectres de la linguistique sont discrets, mais ils sont illustrés par des courbes continues, plus lisibles.

Le style d'une œuvre se manifeste dans la forme des spectres. Le thème reste transparent, les moyennes étant effacées par le centrage et la normalisation des variables.

Cette forme est cernée en première approche par la variabilité. L'asymétrie affine le croquis, tandis que les moments supérieurs sont négligés ici.

Précautions et limites

Les statistiques qui précèdent supposent que les temps de retour se forment à l'identique tandis que le temps s'écoule, en d'autres termes que le processus est stationnaire. Les moments sont alors indépendants du temps : c'est l'hypothèse de stationnarité faible généralement utilisée.

Pour que les moments temporels et empiriques tendent vers leurs homologues ensemblistes et probabilistes, il faut un élément supplémentaire. Si les observations successives sont indépendantes, la convergence des moyennes est établie par la simple application de la loi des grands nombres. Cette condition est cependant restrictive, et l'on se contente en général de la décroissance rapide des corrélations. Ces phénomènes à la mémoire courte sont dits ergodiques[291]. La prise en compte de ces interactions temporelles est l'objet de la nanoscopie.

Par ailleurs, ce modèle isole le mouvement de chaque unité et ne tient pas compte des interactions entre ces éléments. Formellement, si T_{jk} désigne le temps de retour k de l'unité j, les statistiques sont invariantes par une permutation sur les indices spatio-temporels : à défaut d'être exhaustif, notre modèle est cohérent et équilibré dans son

[291] Pac, *Processus alétaoires*, p. 20.

appréhension de l'univers.

5.1.4 Modélisation

Les spectres évoqués précédemment illustrent les moments statistiques et caractérisent un style. D'un point de vue linguistique, cherchons un gabarit qui les unisse.

Plusieurs voies ont été suivies pour modéliser ces courbes. Les plus simples dérivent d'une loi normale déformée par un changement de variable : un temps de retour toujours positif invite une loi log-normale malheureusement peu fidèle à la réalité.

D'autres composent plusieurs variables aléatoires positives, et créent des modèles complexes réglés par de multiples paramètres : des lois Gamma[292] ou de Poisson[293] ont ainsi été proposées pour expliquer les longueurs de phrases.

Une théorie plus robuste, qui reflète mieux la physique littéraire, est donc appelée.

5.1.4.1 Distribution relative

Le spectre de la figure 2.25 reste quelque peu éthéré. Calculé a priori, il ignore une information pourtant essentielle dans un processus réel : la connaisance du passé.

Dans le domaine de la fiabilité, l'incidence d'une panne à un instant

[292] Woronczac, « Statistische Methoden in der Verslehre ».
[293] Hjort, « And Quiet Does Not Flow the Don : Statistical Analysis of a Quarrel between Nobel Laureates ».

donné dépend généralement du comportement antérieur du matériel considéré. Précisément, le lien avec le passé ne se fait sentir que durant la vie d'un même élément. A contrario, le comportement d'un nouveau composant est décorrélé du précédent[294].

De façon analogue, les occurrences d'une unité sont vues comme des accidents successifs et indépendants.

Mathématiquement, ce principe se traduit par l'emploi de probabilités conditionnelles formalisées par Bayes[295] : la probabilité d'un événement A est alors remplacée par celle de l'événement A sachant B.

Supposons qu'une unité arrive à l'instant 0, et soit T le temps de premier retour. A partir de la distribution absolue P(T), la distribution relative estime la probabilité que cette unité réapparaisse à l'instant t, sachant qu'elle n'est pas revenue depuis l'instant 0. Cette grandeur est appelée *taux d'apparition instantané* : $\lambda(t) = P(T = t) / P(T \geq t)$ [296].

Sur l'aire hachurée de la figure 2.26, le premier terme de ce rapport est la longueur de sa borne verticale, et le second sa surface.

---

[294] Bessis, *La probabilité et l'évaluation des risques*, « Relativité de la notion de probabilité », p . 31-34.
[295] Bayes, *An essay towards solving a Problem in the Doctrine of Chances.*
[296] Dans le cas d'une distribution continue, P(T=t) désigne la densité de probabilité d'un temps de retour t.

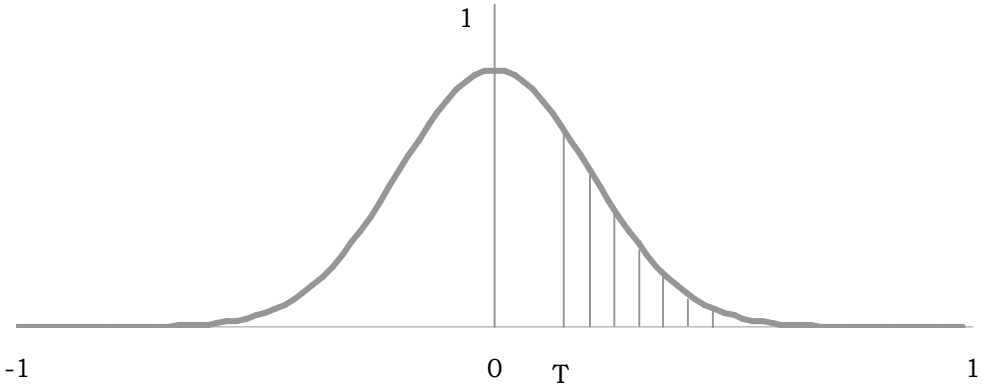


Figure 2.26 : longueur et surface

L'évolution de λ trace la distribution relative recherchée. Deux grandes familles font alors leur entrée (fig. 2.27) :

– si λ est constant, la distribution absolue est exponentielle ;

– si λ croît linéairement, la distribution absolue est de Rayleigh[297].

A l'origine, les courbes absolues et relatives se confondent, puis divergent rapidement.

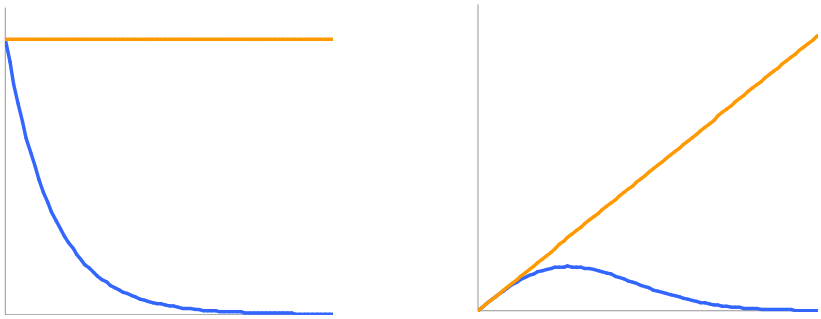


Figure 2.27 : distributions exponentielles et de Rayleigh

Dans le domaine de la fiabilité, les exponentielles modélisent des matériaux et des composants qui ne vieillissent pas, ou dont l'usure est stabilisée : le taux de défaillance reste identique, indépendamment du passé et de pannes éventuelles.

[297] Rayleigh (1842-1919) : physicien anglais, spécialiste des ondes optiques et acoustiques. Avec Ramsay, il découvre l'argon et obtient le prix Nobel en 1904.

En revanche, les distributions de Rayleigh sont associées à des éléments neufs enclins à l'usure : à mesure que le temps passe et que le fonctionnement reste normal, la menace d'un incident se fait plus pressante.

D'un point de vue linguistique, une distribution exponentielle traduit l'indifférence d'une unité à la présence de celle qui la précède.

La distribution de Rayleigh reflète justement cette sensibilité. En particulier, la probabilité de présence est nulle à l'origine, marquant l'impossibilité de la cohabitation de deux unités, puis cette barrière se fait moins sentir au fur et à mesure de l'éloignement. A leur image, deux ions chargés identiquement se repoussent violemment dans leurs sphères d'influences intimes, oublient progressivement leur aversion avec la séparation, et finissent par trouver un équilibre plus serein.

Les principales caractéristiques de ces lois sont résumées dans le tableau suivant. Les probabilités, par définition positives, supposent des paramètres λ et σ eux aussi positifs. Dans la variance et l'asymétrie, la notation m° désigne les moments de la variable centrée (fig. 2.28).

	Exponentielle ($\lambda > 0$)	Rayleigh ($\sigma > 0$)
Distribution absolue	$\lambda\, e^{-\lambda t}$	$\sigma t\, e^{-\sigma t^2/2}$
Distribution relative	λ	$\sigma\, t$
Mode (maximum)	0	$\sigma^{-1/2}$
Moyenne (m_1)	$1/\lambda$	$(\pi/2)^{1/2}\, \sigma^{-1/2}$
Variance ($m°_2$)	$1/\lambda^2$	$(4-\pi)/2\ \sigma^{-1}$
Asymétrie ($m°_3$)	$2/\lambda^3$	$(\pi-3)\ (\pi/2)^{1/2}\ \sigma^{-3/2}$

Figure 2.28 : lois exponentielles et de Rayleigh

Rien n'empêche alors de mêler le sang de ces deux familles pour

engendrer une distribution mixte, qui généralise l'évolution de λ sous la forme d'une droite quelconque. Ses caractéristiques sont données dans le tableau 2.29. Les formules complexes des moments n'apportent rien ici et ne sont pas précisées.

	Linéaire (λ, σ)
Distribution absolue	$(\lambda + \sigma t)\, e^{-\lambda t - \sigma t^2/2}$
Distribution relative	$\lambda + \sigma t$
Mode (maximum)	$(\sigma^{1/2} - \lambda) / \sigma$

Figure 2.29 : loi linéaire

Dans cet univers élargi, les expressions restent cependant soumises à la contrainte, et gardent un sens mathématique pour des valeurs positives, soit $\lambda + \sigma t > 0$. Physiquement, seuls les taux d'apparition positifs à l'origine sont retenus, soit $\lambda > 0$.

L'alternative est la suivante :

– $\sigma < 0$: la distribution est définie sur l'intervalle de temps $[0, -\lambda/\sigma]$
– $\sigma > 0$: la distribution est définie sur $[0, +\infty[$, mais deux cas se présentent à leur tour : tant que σ reste modéré, la distribution absolue décroît de façon monotone, tandis que si σ dépasse le seuil λ^2, un mode pointe près de l'origine.

Ces différents scénarios déploient leurs courbes sur la figure 2.30.

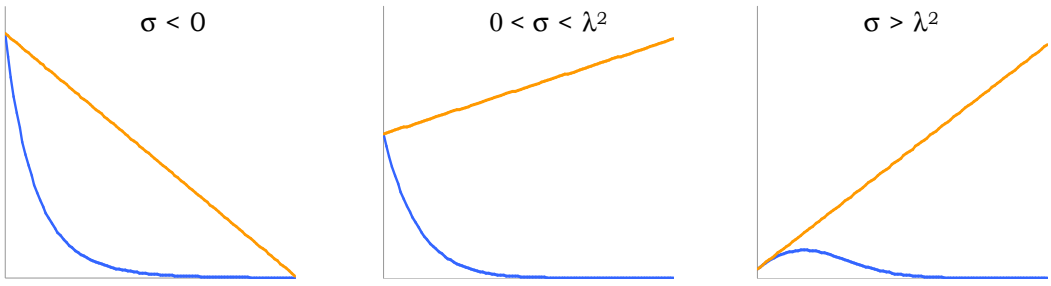


Figure 2.30 : distributions linéaires

La composition de lois linéaires permet d'élargir une dernière fois le modèle (fig. 2.31) :

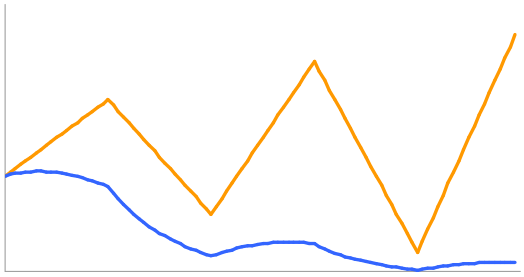


Figure 2.31 : distribution linéarisée

Le taux d'apparition instantané λ(t) donné, il est en général difficile d'exprimer analytiquement la probabilité absolue P(t). Celle-ci se calcule néanmoins à l'aide d'une intégration numérique :

$P(t)= \lambda(t)\ e^{-\int\lambda(u)du}$, $u \in [0, t]$[298].

5.1.4.2 Simulation

Inversement, le modèle qui précède est susceptible de générer des séquences aléatoires, par exemple le tirage d'un « a ». Formellement, il s'agit de bâtir à partir du taux d'apparition instantané λ une variable aléatoire X qui prenne la valeur 1 en cas de succès, 0 en cas d'échec.

La linguistique se satisfait de temps discrets, désignons donc par λ_i la suite des valeurs du taux et par P_i celle des probabilités associées. Soit X_i une variable aléatoire prenant la valeur 1 avec la probabilité P_i et la valeur 0 avec la probabilité $1-P_i$. La simulation est aisément réalisée selon l'algorithme qui suit :

---

[298] Cette formule résulte de la relation entre P(t) et λ(t). Elle reste donc valable pour des évolutions non linéaires de λ.

$P_0 = 0$;

$P_1 = \lambda_1$; si $X_1 = 1$, le processus repart à 0 ;

…

$P_i = \lambda_i (1-\Sigma P_j)$ $(j < i)$; si $X_i = 1$, le processus repart à 0 ;

5.2 Synthèse des mesures

L'idée générale consiste à considérer non les distances physiques pour chaque unité, mais une arythmie abstraite qui plisse la structure spatio-temporelle. Ce seuil franchi, l'espace se réduit à un axe et l'intégration devient naturelle.

5.2.1 Moments

Soit X_{ijk} le $k^{ième}$ temps de retour de l'unité j dans l'œuvre i , k variant de 1 à $n_{ij.}$

La moyenne de l'unité j s'évalue par $m_{ij} = \Sigma\, x_{ijk} / n_{ij}$ $(k = 1, n_{ij})$, d'où l'arythmie qui servira de base aux statistiques :

$x_{ijk} = (X_{ijk} – m_{ij}) / m_{ij}.$

Come nouvelle variable est centrée, son moment d'ordre 1 est nul, et ceux d'ordre 2 et 3 sont calculés par :

$m_{2ij} = \Sigma\, x_{ijk}^2 / n_{ij}$ $(k = 1, n_{ij})$

$m_{3ij} = \Sigma\, x_{ijk}^3 / n_{ij}$ $(k = 1, n_{ij})$.

On en déduit la variabilité et l'asymétrie de l'unité j :

$v_{ij} = m_{2ij}^{1/2}$

$a_{ij} = m_{3ij}^{1/3}$.

Pour s'abstraire des unités et se hisser au niveau de l'œuvre, il suffit de cumuler l'ensemble des déformations :

$m_{2i} = \Sigma\, x_{ijk}^2 \,/\, n_i \quad (j = 1, J), (k = 1, n_{ij})$

$m_{3i} = \Sigma\, x_{ijk}^3 \,/\, n_i \quad (j = 1, J), (k = 1, n_{ij})$.

Pratiquement, ces valeurs globales se déduisent des valeurs locales par :

$m_{2i} = \Sigma\, n_{ij}\, m_{2ij} \,/\, n_i \quad (j = 1, J)$

$m_{3i} = \Sigma\, n_{ij}\, m_{3ij} \,/\, n_i \quad (j = 1, J)$.

La variabilité et l'asymétrie d'une œuvre s'expriment finalement ainsi :

$v_i = m_{2i}^{1/2}$

$a_i = m_{3i}^{1/3}$.

5.2.2 Spectres

La logique précédente pourrait conduire à étudier la distribution de l'arythmie X/m –1, où X est un temps de retour et m sa moyenne. Or X est toujours positive, la variable relative X/m est donc préférée afin de conserver cette propriété physique fondamentale. Le spectre de la première se déduit de la seconde par une translation horizontale.

6 Nanoscopie

La microscopie observe les temps de retour d'une unité sans tenir compte de leur ordre d'apparition. La phase présente vise à combler

cette lacune et à affiner l'étude du rythme à l'aide des corrélations temporelles[299].

Les analyses peuvent se classer en deux familles.

– Les premières, d'inspirations empiriques, abandonnent l'espace du temps pour rejoindre celui des fréquences. La transformée de Fourier de la fonction de corrélation fournit le spectre du signal aléatoire considéré. Citons ici les travaux de Dreher, Young, Norton & MA[300] sur les segments du chinois moderne, ainsi que ceux d'Azar & Kedem[301] sur les phonèmes de l'hébreu biblique. Ces méthodes sont surtout adaptées à des phénomènes rythmiques, qui forment des pics aisément repérables sur les spectres. Mais la voie devient une impasse pour des distributions souvent chaotiques.
– Face à ces signaux bruités, des approches récentes apportent une réponse différente. A partir des fonctions de corrélation, elles infèrent un modèle stochastique qui opère dans le domaine temporel. Y figurent notamment les travaux de Corduas[302] sur les longueurs de mots italiens et de Pawlowski[303] sur l'attribution entre Romain Gary et Emile Ajar. C'est la voie choisie.

La section qui suit est organisée comme les précédentes, et décrit dans un premier temps la méthode appliquée à une unité, pour aborder ensuite les principes de la synthèse.

---

[299] Paradoxalement, l'horizon de temps s'élargit, d'une microscopie qui se borne à un intervalle à une nanoscopie qui embrasse les séquences voisines. Les titres se réfèrent à la finesse de l'analyse opérée.
[300] Dreher, Young, Norton & MA, « Power Spectral Densities of Literary Rythms (Chinese) ».
[301] Azar & Kedem, « Some Time Series in the Phonetics of Biblical Hebrew ».
[302] Corduas, « La struttura dinamica dei dati testuali ».
[303] Pawlowski, *Séries temporelles en linguistique.*

6.1 Mesure unitaire

Avant de décrire les modèles envisagés, la notion de corrélation est précisée ici.

6.1.1 Corrélations

6.1.1.1 Corrélation simple

Elle permet d'analyser la succession des distances entre occurrences d'une même unité. Schématiquement, une distance est considérée comme grande lorsqu'elle dépasse sa valeur moyenne, et comme petite dans le cas contraire.

Si X(t) désigne la distance courante, X(t-1) est sa voisine, et X(t-k) est son homologue décalé à k reprises.

La covariance[304] empirique de X d'ordre k est donnée par la formule : $R_k = < (X(t)-m) (X(t-k)-m) >$, où <> symbolise la moyenne temporelle.

Comme dans la microscopie, le processus sous-jacent est supposé stationnaire et ergodique. La statistique converge alors vers la covariance théorique $E(X(t)-m)(X(t-k)-m)$, où E désigne l'espérance stochastique.

Afin de s'affranchir des effets d'échelle, on utilise généralement la corrélation : $r_k = R_k/R_0$[305].

---

[304] Traditionnellement, on parle plutôt d'autocovariance, pour préciser que la statistique porte sur une seule variable. C'est toujours le cas ici, d'où une notation simplifiée.

[305] R_0 n'est autre que la variance empirique de X, non nulle en général.

6.1.1.2 Corrélation partielle

La corrélation apparente entre deux variables masque souvent l'intervention d'une troisième, qui est le véritable agent de l'interaction.

En surface, des coups de soleil apparaissent souvent au moment où l'on porte des lunettes protectrices, mais il faut surtout y voir l'effet d'une cause centrale, le rayonnement. De même ans, la corrélation des observations aux instants t et t+k dépend des contributions entre ces deux bornes de temps. La corrélation partielle soustrait les éléments intermédaires pour garder la partie congrue.

Mathématiquement, les influences mutuelles sont supposées linéaires, d'où la régression[306] :

$X(t) = \phi_{k1} X(t-1) + \ldots + \phi_{kk} X(t-k) + w(t)$, où w(t) est le résidu.

Le dernier coefficient ϕ_{kk} n'est autre que la corrélation partielle d'ordre k, notée r_k*.

Dans le cas particulier du premier ordre, les corrélations totales et partielles se confondent naturellement : aucune valeur intermédiaire ne vient influencer les observations initiales et finales.

6.1.2 Modèles

Ils sont issus en grande part de la sphère économique, fruits de nombreuses recherches pour prévoir ou du moins modéliser les cours de la Bourse.

---

[306] La méthode la plus employée est celle des moindres carrés.

En 1927, Slutzky[307] et Yule[308] font simultanément un constat paradoxal : des facteurs aléatoires sont capables d'engendrer des signaux corrélés.

Dix ans plus tard, Wold[309] formalise cette approche et énonce un théorème fondateur : toute série temporelle stationnaire se décompose en deux termes, l'un déterministe et l'autre aléatoire. Le premier s'explique par ses valeurs passés, tandis que le second résulte d'une série de chocs hasardeux.

Sous cette forme théorique, chaque terme comprend un nombre infini d'éléments qui rend la décomposition peu applicable. Une formulation plus pratique fait naître deux familles, les processus Auto Regressive (AR) et Moving Average (MA), qui divergent par la forme de leurs corrélations totales et partielles. Leurs principaux aspects sont présentés ci-dessous sans entrer dans la démonstration[310].

6.1.2.1 Processus AR

L'observation à l'instant t dépend des p précédentes, ainsi que d'un bruit w, d'où la définition d'un AR(p) :

$X(t) = \Sigma\ \phi_{pi}\, X(t-i) + w(t),\ i \in [1, p]$[311].

On reconnaît la régression linéaire évoquée précédement. L'identification des coefficients ϕ_{pi} en résulte, et les corrélations

---

[307] Slutzky, « The summation of random causes as the source of cyclic process ».
[308] Yule, « On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers ».
[309] Wold, *A study in the analysis of stationary time series.*
[310] Une description plus détaillée se trouve dans Pawlowski, *Séries temporelles en linguistique*.
[311] Le terme constant est omis pour la simplicité de la formulation : il ne change rien sur le fond.

partielles s'annulent donc au-delà de l'ordre p.

Les corrélations totales sont données par l'équation de Yule Walker, qui découle de la définition du processus AR :

$r_j = \Sigma\ \phi_{pi}\, r_{j-I},\ i \in [1, p]$.

Cette formule est valable quel que soit j[312]. Son application aux valeurs comprises entre 1 et p fournit un système de p équations, qui détermine les inconnues r_i à partir des estimations de ϕ_{pi}.

$$\begin{pmatrix} r_1 \\ r_2 \\ . \\ r_p \end{pmatrix} = \begin{pmatrix} 1 & r_1 & \dots & r_{p-1} \\ r_1 & 1 & \dots & r_{p-2} \\ & . & \dots & . \\ r_{p-1} & r_{p-2} & \dots & 1 \end{pmatrix} \begin{pmatrix} \phi_{p1} \\ \phi_{p2} \\ . \\ \phi_{pp} \end{pmatrix}$$

Par extrapolation, les corrélations totales n'ont aucune raison de s'annuler au-delà de l'ordre p. Un calcul précis montre qu'elles ont la forme d'une exponentielle ou d'une sinusoïde amortie, en fonction des coefficients ϕ_{pi}.

Les figures 2.32 et 2.33 représentent les corrélations totales partielles d'un processus AR d'ordre 1 pour ϕ_1 négatif puis positif.

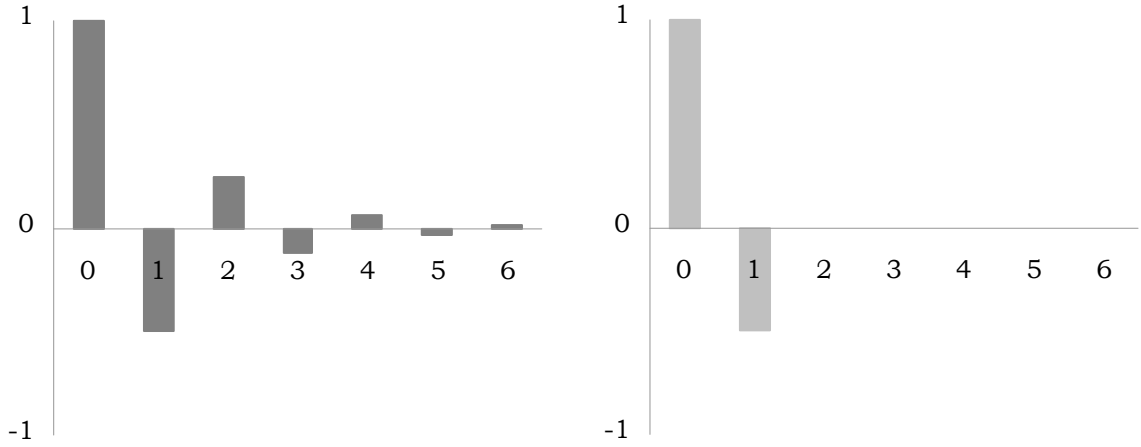


Figure 2.33 : corrélations totales et partielles, $\phi_1 < 0$

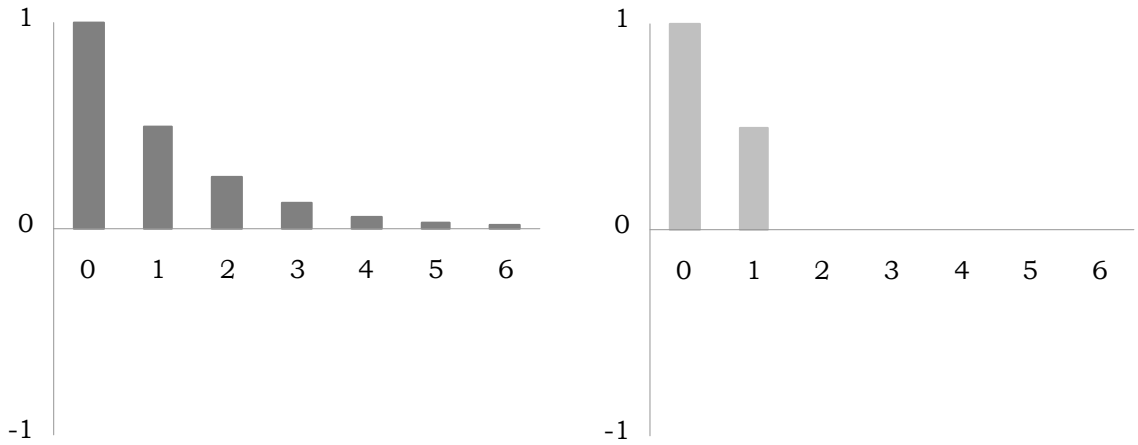


Figure 2.32 : corrélations totales et partielles, $\phi_1 > 0$

6.1.2.2 Processus MA

L'observation à l'instant t dépend de chocs aléatoires survenus au temps courant et aux q précédents, d'où la définition d'un MA(q) :

$X(t) = \Sigma\ \theta_{qi}\, w(t-i) + w(t)$, $i \in [1, q]$.

Les corrélations totales sont nulles au-delà de l'ordre q, car les chocs qui interviennent sont indépendants. La définition du processus induit les valeurs intermédiaires :

$r_k = (\theta_k + \Sigma\ \theta_{k+i}\theta_i) / \Sigma\ \theta_j^2$, $i \in [1, q-k]$ et $j \in [1, q]$.

Sans entrer dans la démonstration, les corrélations partielles décroissent progressivement selon une exponentielle ou une sinusoïde amortie.

Les courbes ont donc la même allure que celles des processus AR : il suffit d'échanger les versions totales et partielles.

L'estimation des paramètres est plus complexe dans ce cas. La

méthode la plus employée a été développée par Durbin[313] : elle consiste à faire dériver le processus MA d'un processus AR par un passage à la limite[314], puis à appliquer à ce dernier l'identification déjà décrite.

6.1.2.3 Processus ARMA et autres

Les familles AR et MA sont intimement liées. De ce constat, Box et Jenkins[315] créent en 1970 un processus généralisé ARMA, engendré par ces deux lignées :

$X(t) = \Sigma\, \phi_{pi}\, X(t-i) + \Sigma\, \theta_{qj}\, w(t-j) + w(t)$, $i \in [1, p]$ et $j \in [1, q]$.

Pour tenir compte de phénomènes saisonniers ou non linéaires, d'autres modèles ont été envisagés, non détaillés ici.

6.2 Synthèse des mesures

De façon analogue à la microscopie, une mesure globale est donnée à partir des temps relatifs ou de l'arythmie.

Formellement, au lieu de considèrer isolément la variable X_j d'une unité, on analyse la succession des valeurs de X_j/m_j par la méthode qui précède.

---

[313] Durbin, « Efficient Estimation of Parameters of Moving Average Mcdels ».
[314] Les deux familles sont duales : on peut montrer qu'un MA fini correspond à un AR infini, et inversement.
[315] Box, Jenkins & Reinsel, *Time series analysis : forecasting and control.*

7 Télescopie

La perspective de cette phase est l'inverse des précédentes : le corpus auparavant familier est désormais composé d'éléments inconnus que l'on cherche à classer ou étiqueter. En point de mire de cette approche figure l'attribution d'auteur.

Cette dernière notion ne va pas de soi, et suppose celle de style d'auteur : dans l'arbre de la création d'un individu, chaque branche est alors plus proche de son axe d'origine que des plants voisins ; en termes mathématiques, les distances internes sont plus faibles que les externes, et le rapport entre ces deux termes reflète la qualité de la partition opérée. L'hypothèse de style d'auteur demande à être confirmée : si l'auteur transmet ses gènes tandis que l'environnement impulse ses soubresauts, il est difficile de trancher a priori entre la part de l'inné et de l'acquis. En revanche, l'expérimentation et la mise en œuvre d'une mesure pertinente contribuent à lever le doute.

Quelle distance adopter ? On cherche une mesure représentative, qui conserve l'information et taise le bruit ; une mesure fondée sur le thème et la composition, mais tenant compte du style et de l'organisation.

Revenons sur le cheminement de cette thèse :

- la macroscopie établit le thème par l'analyse statique des fréquences ;
- la mésoscopie ébauche le mouvement et l'étude du style avec les variations des fréquences selon les divisions ;
- la microscopie va au cœur de la dynamique par la distribution des temps de retours ; sur les spectres, la valeur moyenne indique le thème, tandis que les écarts autour de cette référence traduisent le style ;

– la nanoscopie affine la microscopie et constate la quasi-indépendance des temps de retour successifs[316].

Si l'on écarte les interactions entre les unités pour se concentrer sur les aspects rythmiques, il semble logique de caractériser un texte par la collection des spectres de la microscopie. Evidemment, ce rythme a des accents singuliers : des intervalles indépendants se succèdent, mais ils obéissent à la même distribution.

L'approche suivie est modestement et résolument descriptive : il s'agit uniquement de mesurer la distance entre deux textes à l'aide de statistiques, et non d'inférer et d'estimer la probabilité qu'ils aient le même auteur. En somme, la comparaison porte sur le bois touché du doigt, sans référence à des arbres imaginaires.

A partir de ces principes, comment bâtir une mesure spécifique, puis synthétique ?

7.1 Mesure unitaire

L'idée est de fonder la distance sur les écarts entre les spectres de chaque unité. Pratiquement, des distributions bruitées sont difficiles à comparer. Un lissage naturel est obtenu à l'aide des répartitions qui les intègrent.

Mathématiquement, si P est la probabilité que le temps de retour T vaille t, alors la fonction de répartition F est définie par :

$F(t) = P(T \leq t)$[317].

[316] Cf. chapitre 7 dans la partie expérimentale.
[317] D'autres définitions utilisent $P(T < t)$. Les deux options sont équivalentes, seule compte la cohérence de leu emploi.

F croît donc monotonement de 0 à 1 quand t varie de 0 à l'infini.

La figure 2.34 compare l'agitation d'une distribution à la sérénité de sa répartition.

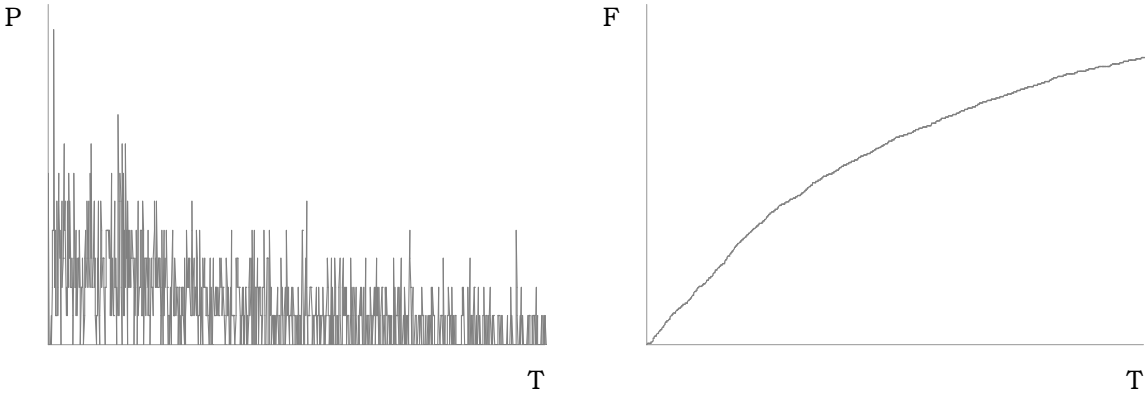


Figure 2.34 : distribution et répartition

Dans le détail, les répartitions se présentent comme des courbes en escalier (fig. 2.35) :

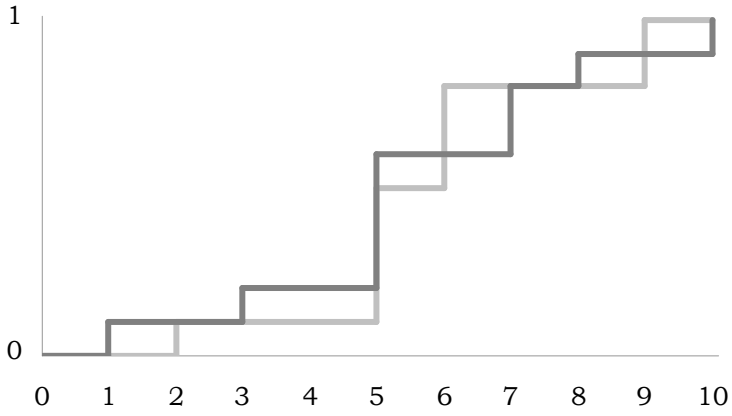


Figure 2.35 : comparaison des répartitions

- une marche se forme quand le temps de retour prend une nouvelle valeur ; la hauteur du front est la proportion des observations qui prennent cette valeur.
- entre deux marches, la fonction reste constante.

Plus précisément :

- les abscisses des marches sont celles de la série T^*_i des temps de retour, classés par valeurs croissantes pour i variant de 1 à I.
- l'ampleur du saut est donnée par le ratio $k_i / \Sigma k_i$, où k_i désigne le nombre d'observations égales à T^*_i..
- la valeur constante entre deux marches, qui intègre les sauts précédents, correspond au cumul de la fonction de répartition.

Pour comparer les deux courbes, il suffit de moyenner les écarts à chaque front montant avec la pondération des effectifs concernés.

Soient k_{1i} et k_{2i} les nombres d'observations égales à T^*_i pour les populations 1 et 2. La distance entre les deux répartitions est alors donnée par :

$$D_{12} = (\Sigma\ (k_{1i}+k_{2i})\ (F_1(T^*_i)-F_2(T^*_i))^2\ /\ \Sigma\ (k_{1i}+k_{2i}))^{1/2}.$$

Cette expression complexe n'est autre que l'espérance de la variable aléatoire $(F_1-F_2)^2$ sur l'ensemble des deux populations, d'où une écriture plus dense : $D_{12} = (E\ (F_1-F_2)^2)^{1/2}$.

On retrouve une statistique connue et utilisée par le test d'Anderson[318]. Mais il s'agit ici uniquement de mesurer une distance, tandis que la perspective de l'inférence est d'estimer une probabilité que deux échantillons appartiennent à une même population théorique. Cette dernière question, plus délicate, et n'est pas abordée dans le cadre de ce travail[319].

Des variantes de cette distance existent, fondées non sur L_2, mais L_p où p est un entier : $D_{12} = (E\ (F_1-F_2)^p)^{1/p}$.

[318] Anderson, « On the Distribution of the Two-Sample Cramer-Von-Mises Criterion ».
[319] En particulier, le test d'Anderson considère des distributions continues, et ne peut donc pas être appliqué directement aux échantillons discrets de cette étude.

L_1 est théoriquement la plus réaliste, puisqu'elle traduit linéairement les écarts observés entre les fonctions de répartition[320]. A l'autre bout du spectre, L_∞ est la plus schématique et ne retient que l'écart extrême, soit $\sup|F_1-F_2|$. Cette dernière fonde le test de Kolmogorov-Smirnov, couramment utilisé en statistique inférentielle.

Le plus souvent, L_2 est privilégiée : elle déforme peu les écarts tandis que sa forme euclidienne permet de précieuses interprétations géométriques.

7.2 Synthèse des mesures

Appliquée à chaque unité, la méthode précédente permet de calculer une distance locale : $D_{12j} = (E\ (F_{1j}-F_{2j})^2)^{1/2}$.

Pour obtenir une distance globale, il suffit de moyenner cette grandeur avec la pondération des effectifs de l'unité j sur l'ensemble des deux populations :

$$D_{12} = (\Sigma\ (k_{1j}+k_{2j})d_{12j}^2\ /\ \Sigma\ (k_{1j}+k_{2j}))^{1/2}$$
$$= (\Sigma\ \Sigma\ (k_{1ij}+k_{2ij})\ (F_{1j}(T^*_{ij})-F_{2j}(T^*_{ij}))^2\ /\ \Sigma\ \Sigma\ (k_{1ij}+k_{2ij}))^{1/2}$$

en faisant varier i de 1 à I, et j de 1 à J.

La dernière expression est une moyenne analogue à celle de la mesure locale, étendue à l'ensemble des unités, d'où la notation simplifiée : $D_{12} = (E\ (F_1-F_2)^2)^{1/2}$.

---

[320] Xiao, Gordon & Yakovlev, « The L1-Version of the Cramér-von Mises Test for Two-Sample Comparisons in Microarray Data Analysis »

Propriétés mathématiques

Une première caractéristique intéressante de cette distance est la suivante : chaque unité et chacune de ses occurences est prise en compte une fois et une seule par les temps de retour successifs[321], si bien que la statistique réalise un pavage du texte (fig. 2.36).

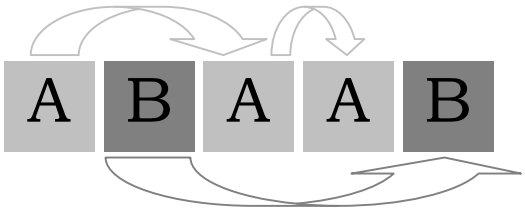


Figure 2.36 : pavage textuel

L'expression de la distance suit ce principe d'équanimité : chaque intervalle observé se retrouve une fois et une seule dans le cumul des écarts $(F_1\text{-}F_2)^2$.

Le mot « distance » a été employé. Vérifions que cette appellation est justifiée :

- la séparation est claire : D_{12} est nulle si et seulement si les fonctions de répartition F_1 et F_2 sont identiques ;
- la symétrie est triviale : $D_{12} = D_{21}$;
- l'inégalité triangulaire se démontre en identifiant D_{13} à $(E\ (F_1\text{-}F_3)^2)^{1/2}$: il suffit d'intercaler F_2 et d'appliquer l'inégalité de Cauchy-Schwarz[322] pour se convaincre que D_{13} est inférieure ou égale à la somme de D_{12} et D_{23} ; en d'autres termes, le chemin direct reste le plus court (fig. 2.37) :

[321] A l'exception de la première occurrence de chaque unité. Cet effet de bord est le plus souvent négligeable au regard de la taille de la population.
[322] Rényi, *Calcul des probabilités*, p. 97.

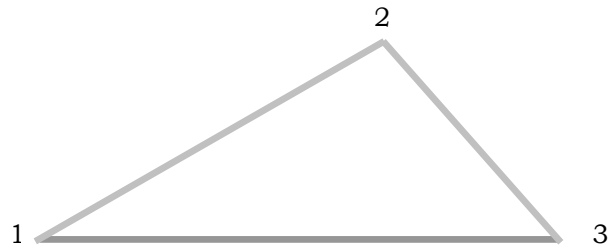


Figure 2.37 : inégalité triangulaire

Quels sont les ordres de grandeur de cette distance ?

F_1 et F_2 sont comprises entre 0 et 1, il en va donc de même pour $E\,(F_1\text{-}F_2)^2$ et la distance :

- la valeur 0 correspond à des répartitions identiques d'après ce qui précède ;
- la valeur 1 est atteinte lorsqu'une des répartitions est constamment nulle et que l'autre se concentre sur une seule valeur ; dans la première configuration, l'unité considérée est absente, tandis que la seconde structure est parfaitement symétrique et rythmique.

En présence de deux structures symétriques, précisons le comportement de cette distance et sa relation avec les fréquences.

La population de chaque unité se concentre sur une raie, associée à un temps de retour et à une fréquence qui lui est inversement proportionnelle, notés respectivement T_{1j}, T_{2j}, f_{1j} et f_{2j}.

Les formules se simplifient alors considérablement :

$$D_{12j}^2 = \begin{cases} 0 \text{ si } T_{1j} = T_{2j} \\ k_{1j}/(k_{1j}+k_{2j}) \text{ si } T_{1j} < T_{2j} \\ k_{2j}/(k_{1j}+k_{2j}) \text{ si } T_{2j} < T_{1j} \end{cases}$$

D_{12j} se fait sentir quand f_{1j} et f_{2j} diffèrent ; D reflète donc la moyenne des écarts de fréquences.

D généralise donc la distance classique d fondée sur les fréquences.

La correspondance entre les différentes distances est donnée par le schéma de la figure 2.38 :

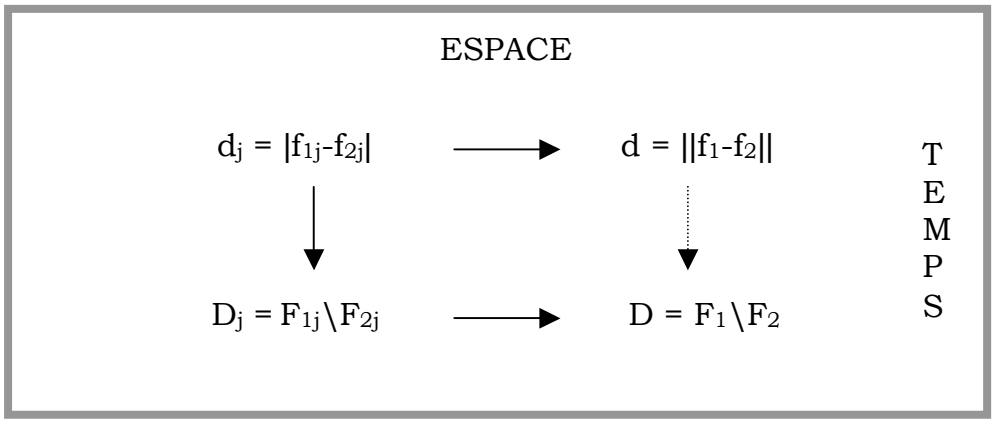


Figure 2.38 : analogies formelles

D résulte d'une double intégration, l'une dans le temps à travers les répartitions, l'autre dans l'espace à travers les unités, qui lui donne son assise et son équilibre.

Chapitre 3 : les instruments

1 Introduction

Après ce chapitre méthodique, il s'agit de mettre en œuvre les concepts. Devant le volume des données, des outils informatiques sont évidemment nécessaires.

Dans un premier temps sont recensés les logiciels de référence, à la croisée de la linguistique et de la statistique. Suit alors la description du chemin emprunté par cette étude.

2 Le marché des outils

Cet inventaire présente sommairement les principaux logiciels disponibles. Le cas échéant, leur utilisation dans ce travail est signalée par le symbole *.

Alceste est un logiciel issu d'une coopération entre le CNRS et l'ANVAR, actuellement commercialisé par la société Image[323]. Parmi ses fonctions, l'analyse du vocabulaire, la classification descendante hiérachique, l'analyse factorielle des correspondances et la cartographie. Il traite de nombreuses langues européennes, dont le français, l'anglais et l'allemand.

BDLex est un lexique phonologique constitué par l'IRIT, à l'Université

[323] www.image.cict.fr/index_alceste.htm.

de Toulouse[324]. Il regroupe environ 450 000 formes fléchies générées à partir de 50 000 formes canoniques.

*Cordial-Analyseur** est développé par d'anciens étudiants de l'Université de Toulouse Le Mirail, au sein de la société Synapse[325]. A partir d'un texte, ce logiciel distribue des étiquettes syntaxiques et sémantiques, puis évalue le style par une analyse statistique.

Hyperbase est la création de Brunet à l'Université de Nice. Né à l'occasion du Bicentenaire de la Révolution, ce logiciel rassemble initialement des fonctions documentaires (navigation fréquentielle, contexte et concordance) et statistiques (spécificités, graphiques, analyses factorielles)[326]. Une nouvelle version quitte la surface graphique du texte par l'exploitation des données sémantiques de *Cordial Analyseur*.

Neuronav est conçu par Lelu, statisticien de l'Université de Franche-Comté, et développé dans le cadre de la société Diatopie[327]. Ce logiciel classe et cartographie un corpus, mais il propose aussi une navigation triangulaire entre mots, documents et thèmes.

*Nomino** est le fruit du Département de Linguistique de l'Université du Québec[328]. Il construit dans un premier temps une base de connaissances à partir de textes français ou anglais. Chaque élément textuel est alors représenté par des lemmes, des catégories syntaxiques et des unités plus complexes. Enfin un moteur de recherche permet à l'utilisateur d'interroger cette base en langage naturel.

---

[324] www.irit.fr/PERSONNEL/SAMOVA/decalmes/IHMPT/ress_ling.v1/rbdlex.php.
[325] www.synapse-fr.com/Cordial_ Analyseur/Presention_Cordial_Analyseur.htm
[326] ancilla.unice.fr/~brunet/pub/ hyperbase.html
[327] www.diatopie. com/NN decription.htm
[328] www.ling.uqam. ca/nomino. Nomino n'est cependant n'est plus maintenu, remplacé par Semato.

*S-Plus** est un logiciel spécialisé en statistiques produit par la société Insightful[329] : sa palette très riche complète *Excel* pour les analyses pointues.

*Syntex** est l'œuvre de Bourigault et Fabre, au sein de l'équipe ERSS à Université de Toulouse[330]. Son objectif est de faciliter la constitution de terminologies ou d'ontologies spécialisées. En aval de *TreeTagger*, cet outil étudie les dépendances syntaxiques (nominales, verbales, …) présentes dans un corpus. Le fichier de sortie de *Syntex* est une liste d'étiquettes grammaticales associées à chaque terme du corpus.

TreeTagger est produit par l'Institut de Linguistique de l'Université de Stuttgart331. Européen dans l'âme, il lemmatise et étiquette syntaxiquement des textes allemands, français, anglais, italiens, espagnols, portugais, grec et bulgares.

Tropes-Zoom est un logiciel de la Société Acetic332. Bilingue (français et anglais), Tropes analyse un texte à partir de ses classes sémantiques et de leurs relations d'antériorité ou de postériorité. Il est associé à un moteur de recherche, Zoom.

3 Procédure générale

Il s'agit maintement de décrire les processus employés dans ce travail. Selon les phases de l'étude et les plans linguistiques, diverses

[329] www.insightful.com/products/splus
[330] www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN05-bourigault-Syntex.pdf.
[331] www.ims.uni -stuttgart.de/projekte/corplex/TreeTagger/Decision TreeTagger.html
[332] www.acetic.fr/tropes.htm et www.acetic.fr/zoom.htm.

voies sont empruntées. La cohérence d'une mesure est évidemment maintenue en utilisant la même méthode sur l'ensemble du corpus.

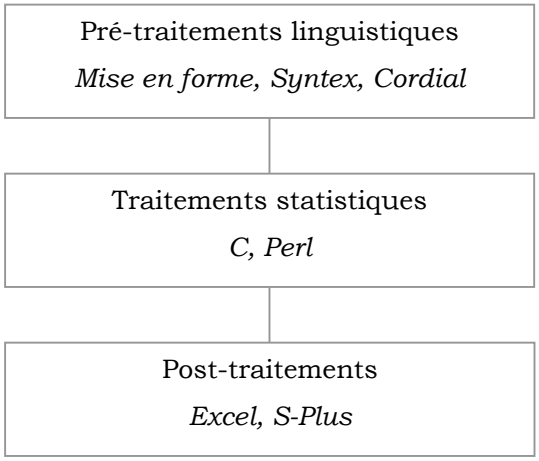


Figure 3.1 : orchestration

De façon générale, les opérations linguistiques sont réalisées à l'aide de logiciels trouvés sur le marché, principalement *Cordial* pour la sémantique, *Syntex* pour la syntaxe ; les graphèmes sont quant à eux appréhendés par une simple mise en forme du texte.

Les développements les plus importants concernent les analyses statistiques, en aval des sorties linguistiques. Ces programmes sur mesure sont surtout nécessaires à partir de la microscopie, tandis que les phases initiales et traditionnelles sont prises en charge par les outils standard.

Le langage *C*, adapté aux éléments de bas niveaux, est principalement utilisé pour la graphémologie. Conçu a contrario pour manier des chaînes de caractères, *Perl* est mis à contribution pour la syntaxe ou la sémantique.

Les programmes sont présentés dans les sections suivantes, à l'aide

d'organigrammes qui suivent le formalisme de la norme ISO 5807[333].

Enfin, les synthèses sont le plus souvent réalisés avec *Excel*, tandis que les statistiques complexes le sont avec *S-Plus*.

4 Macroscopie

4.1 Graphémologie

4.1.1 Mise en forme

Au préalable et afin d'éliminer les effets de la mise en page tributaire de l'édition, chaque texte est concaténé en un seul paragraphe, les retours à la ligne étant remplacés par des espaces[334].

Cette manipulation quelque peu brutale serait hors de propos pour certaines formes[335], mais elle semble plus anodine à l'égard des œuvres de ce corpus, et reste dans tous les cas neutre pour les niveaux syntaxiques ou sémantiques. Elle ramène ainsi chaque création à une référence sinon universelle, du moins partagée, qui fonde sainement les statistiques à venir.

Concrètement, l'opération est simplement réalisée avec la fonction « Rechercher… Remplacer » de *Word*.

Les lettres minuscules et les majuscules sont généralement regroupées afin de constituer des populations plus vastes. En outre, les caractères accentués sont ignorés, comme ils le sont par un codage

---

[333] Les symbologie et les listings sont donnés en annexe 3.
[334] Dans le cas de retours successifs, un seul espace a été inséré.
[335] On pense aux *Calligrammes* d'Apollinaire.

ASCII inspiré de la langue anglaise[336].

4.1.2 Comptages

Les statistiques de *Word* sont sommaires et parfois fluctuantes, d'où le recours à des programmes spécialisés et sécurisés sécrétés en *C* :

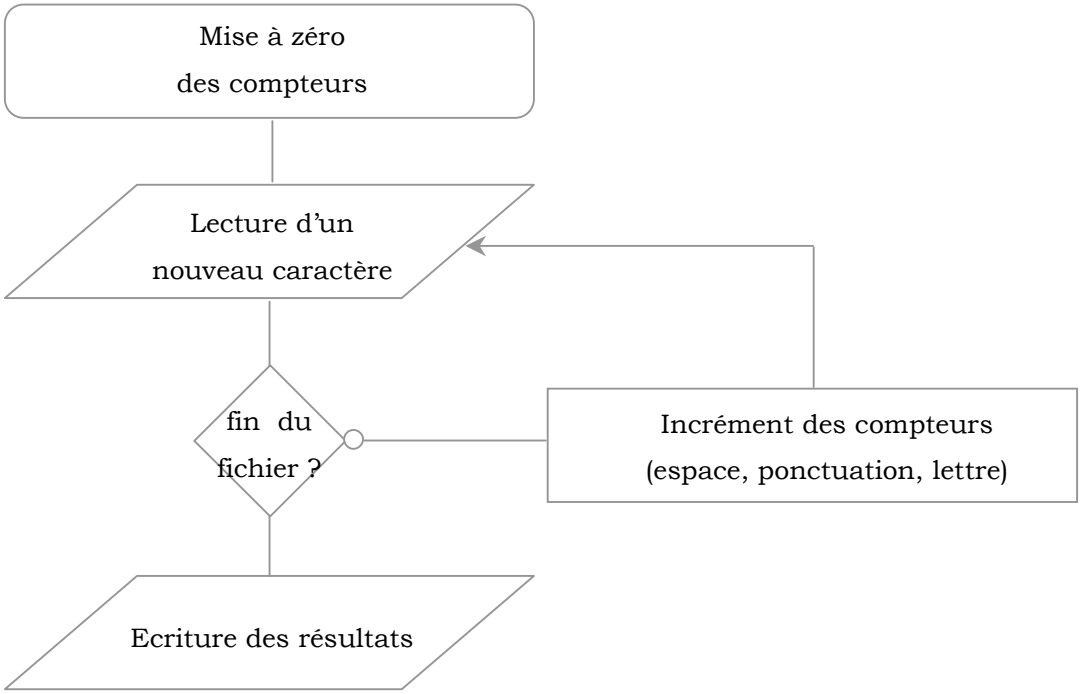


Figure 3.2 : comptage des graphèmes

- la fonction *fgetc* lit un nouveau caractère et donne son code *ASCII* ;
- les valeurs obtenues permettent de trier les graphèmes : EOF si le fichier est fini, 32 pour l'espace, entre 33 et 63 pour la ponctuation, entre 65 et 122 pour les lettres ;

[336] Dans sa version étendue, l'*American Standard Code for Infomation Interchange* prend en compte ces différences. Afin de simplier l'analyse, le codage basique a été retenu.

– les comptages finaux sont écrits dans un fichier séparé.

4.2 Syntaxe

4.2.1 TreeTagger

En amont de *Syntex*, ce logiciel étiquette chaque terme en fonction de son contexte à l'intérieur d'une phrase.

Son principe s'inspire d'un processus de Markov : l'étiquette la plus probable est choisie à partir des attributions voisines. Classiquement, les algorithmes travaillent sur des séquences fixes de n éléments. Pour certaines d'entre elles, les occurrences se font rares et l'identification du modèle devient hasardeuse. Afin de pallier cette difficulté, *TreeTagger* utilise des arbres de décision dont les branches s'allongent ou se raccourcissent selon le contexte : les effectifs sont ainsi régularisés.

La précision de l'étiquetage dépasse 96 % d'après les concepteurs du logiciel[337].

4.2.2 Syntex

Théoriquement, il est possible de tirer les informations syntaxiques de ce premier programme. Pratiquement, elles sont fournies par un logiciel de l'Université du Mirail qui analyse en aval les dépendances syntaxiques.

La figure 3.3 montre une sortie simplifiée de *Syntex,* le format complet incluant le recteur et le régi de chaque terme. L'exemple

[337] Schmid, « Probabilistic Part-of-Speech Tagging Using Decision Trees ».

correspond à la première phrase de *Mémoires d'Hadrien* :

Etiquette	**Lemme**	**Forme**
Pro	je	Je
VCONJS	être	suis
PpaMS	descendre	descendu
DetMS	ce	ce
NomXXDate	matin	matin
Prep	chez	chez
DetMS	mon	mon
NomMS	médecin	médecin
NomPrXXInc	Hermogène	Hermogène
Typo	,	,
ProRel	qui	qui
VCONJS	venir	vient
Prep	de	de
VINF	rentrer	rentrer
Prep	à	à
DetFS	le	la
NomFS	villa	Villa
Prep	après	après
DetMS	un	un
Adv	assez	assez
AdjMS	long	long
NomMS	voyage	voyage
Prep	en	en
NomPr	Asie	Asie
Typo	.	.

Figure 3.3 : sorties de Syntex

La nomenclature de ces étiquettes, précisée en annexe 3, s'organise autour de quatorze catégories grammaticales : adjectif, adverbe, conjonction de coordination, conjonction de subordination,

déterminant, nom, participe passé, participe présent, préposition, pronom, pronom relatif, typographie, verbe conjugué, verbe à l'infinitif, ainsi qu'une catégorie éliminatoire.

La première colonne est extraite du tableau avec des opérations standard d'*Excel*. Le fichier qui en résulte forme une suite d'étiquettes, soit avec l'exemple précédent : Pro VCONJS PpaMS …

4.2.3 Comptages

Des scripts en *Perl* comptent ces étiquettes suivant un algorithme analogue à celui de la figure 3.2, les mots jouant le rôle des graphèmes. Les motifs recherchés sont spécifiés à l'aide d'expressions régulières : par exemple, ^Adj recense toutes les étiquettes commençant par « Adj ».

Finalement, les individus sont groupés en huit parties du discours : adjectif, adverbe, conjonction, déterminant, nom, préposition, pronom, verbe. La typographie déjà analysée en tant que graphème est ignorée, de même que la catégorie éliminatoire[338].

4.3 Sémantique

4.3.1 Cordial

Parmi d'autres fonctions, ce logiciel donne le sens d'un terme, compte tenu de son contexte dans la phrase. Il se réfère à des dictionnaires et des règles de grammaire établis empiriquement, à partir de corpus représentatifs de la langue française.

Dans l'ordre, *Cordial* commence par étiqueter syntaxiquement

[338] Sauf une analyse particulière dans la macroscopie.

chaque terme, ce qui permet de filtrer les significations envisageables.

Le programme tire alors ses concepts d'une taxinomie sémantique, qui découpe l'univers en 28 éléments[339]. Leur liste est donnée en annexe 3.

La précision de l'étiquetage est estimée à 90 % par la société Synapse.

Le figure 3.4 illustre ce traitement, à partir de la même phrase de *Mémoires d'Hadrien.* Certains termes n'ont pas d'étiquette tandis que d'autres en ont plusieurs[340].

Mot	**Lemme**	**Concept**
Je	je	
suis	être	fondamental, temps, esprit
descendu	descendre	mouvement et forces, quotidien, économie
ce	ce	
matin	matin	temps
chez	chez	
mon	mon	
médecin	médecin	santé
Hermogène	Hermogène	
,	,	
qui	qui	
vient	venir	mouvement et forces
de	de	
rentrer	rentrer	mouvement et forces, temps, action
à	à	
la	le	

[339] Ce nombre d'éléments varie avec la finesse du feuilletage choisi : 3 au niveau le plus général, 3387 au niveau le plus précis.
[340] Il ne faut pas voir dans ce dernier cas l'hésitation du logiciel entre plusieurs concepts : le mot touche effectivement plusieurs domaines.

Villa	Villa	
après	après	
un	un	
assez	assez	ordre et mesure, affectivité
long	long	temps
voyage	voyage	quotidien
en	en	
Asie	Asie	
.	.	

Figure 3.4 : sorties de Cordial

Cette sortie spécifique réalisée par le concepteur de *Cordial* est exploitée dans les scopies ultérieures, qui nécessitent une information sur la dynamique du texte. A ce stade, les indications sémantiques globales fournies par le logiciel suffisent, et la figure précédente est d'abord explicative.

4.3.2 Comptages

Les concepts sont directement comptés par *Cordial*, à partir de la fonction « statistiques et sémantique » du menu « outils ». Aucun développement n'est donc nécessaire ici.

5 Mésoscopie

Les traitements de ce niveau font appel aux ressources de la macroscopie, employées non sur l'œuvre entière, mais sur ses divisions successives.

Le titre de chaque œuvre, inclus dans les chiffres macroscopiques, n'est affecté à aucune division particulière : il est donc exclu à ce niveau. Evidemment, ceci est sans incidence statistique au regard des

volumes considérés.

6 Microscopie

Le principe général de cette phase est de refléter au mieux la topologie du texte.

Avec ce pas supplémentaire vers la finesse, une frontière linguistique importante est traversée : si la phrase reste largement englobée par les divisions de la mésoscopie, les intervalles entres deux occurrences successives sont le plus souvent internes. La notion n'est pas aussi élémentaire qu'elle y paraît : son cadre de référence fluctue selon les sensibilités, de la période rhétorique fondée sur l'argument au simple syntagme nominal ou verbal. D'autre part, scinder le texte en phrases réduit la tailles des échantillons et fragilise les statistiques. Finalement, le parti est pris d'ignorer cette limite incertaine : la ponctuation rentre ainsi dans le rang commun des unités.

Dans le même souci topologique, toute unité rencontrée, analysée ou non, est comptabilisée dans la mesure des temps de retour : ainsi, les chiffres parmi les graphèmes, la catégorie éliminatoire parmi les parties du discours, et les mots vides parmi les concepts.

6.1 Graphémologie

Comme dans les scopies précédentes, chaque œuvre du corpus est préalablement ramenée à un seul paragraphe par l'élimination des retours à la ligne.

6.1.1 Moments et spectres

L'organigramme de la figure 3.5 schématise le cheminement du programme écrit en *C*.

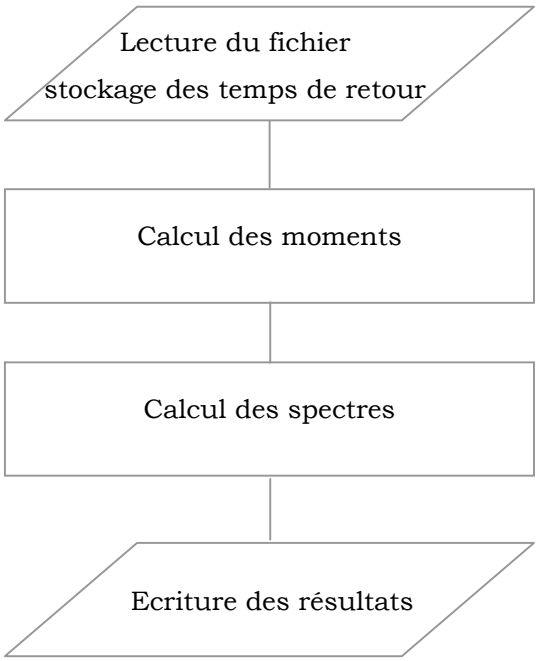


Figure 3.5 : calcul des moments et des spectres

- la lecture du fichier est analogue à celle employée pour le comptage des graphèmes (fig. 3.2) : les temps de retour d'une unité sont stockés dans une matrice, dont la première dimension est égale au nombre d'unités et la seconde au nombre d'intervalles créés ;
- les moments sont calculés par une méthode classique, en premier lieu la moyenne qui sert de référence, puis les moments centrés d'ordre 2 et 3 sous leurs formes réduites ;
- les temps de retour d'une unité sont stockés puis classés par ordre croissant ; pour chaque valeur, les occurrences sont dénombrées, d'où la distribution des fréquences ; les spectres unitaires forment des raies discrètes, alors que le spectre de synthèse se fonde sur des

intervalles continus de longueur h paramétrable ;

– enfin les spectres de chaque unité se répartissent dans des fichiers spécifiques, tandis que les moments et le spectre de synthèse sont écrits dans une sortie générale.

6.1.2 Séquences

Les spectres tracés, des points remarquables ressortent, d'où l'intérêt d'analyser les séquences associées à un temps de retour.

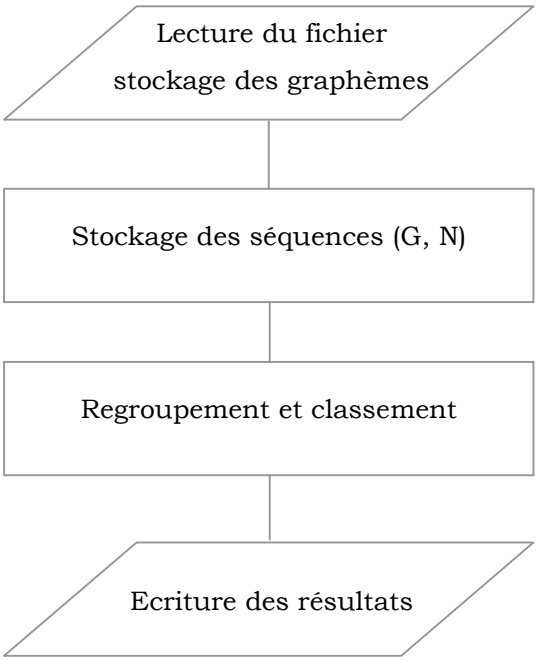


Figure 3.6 : calcul des séquences

Pour le graphème G et la période N, un utilitaire écrit en *C* permet de connaître les n-grammes les plus courants (fig. 3.6) :

– la lecture du fichier et le stockage des graphèmes sont analogues à ce qui précède ;

– les séquences (G, N) sont mémorisées par ordre d'arrivée ;
– enfin, les séquences sont regroupées et classées par fréquences décroissantes.

6.2 Syntaxe

Le traitement de ce niveau est semblable au précédent : il sufit de coder les parties du discours par des entiers. Le fichier littéral des étiquettes se ramène donc à une suite de chiffres assimilables par le programme de la graphémologie.

Pratiquement, un script en *Perl* permet d'associer la catégorie éliminatoire à 0, les adjectifs à 1, les adverbes à 2, les conjonctions à 3, les déterminants à 4, les noms à 5, les prépositions à 6, les pronoms à 7, les verbes à 8, et la typographie à 9.

6.3 Sémantique

De même qu'au niveau syntaxique, on se ramène au processus de la graphémologie : les concepts ainsi codés de 1 à 28[341].

Cependant, des adaptations sont nécessaires pour tenir compte des spécificités de l'étiquetage sémantique : certaines unités sont associées à plusieurs concepts.

Une solution consiste à éliminer ces affectations multiples, réduisant sensiblement la base des statistiques. Finalement, le parti est pris de noter chaque apparition d'un concept, même si la place est partagée en certains lieux.

---

[341] Le format de sortie de *Cordial* en annexe 3 donne la table d'équivalence.

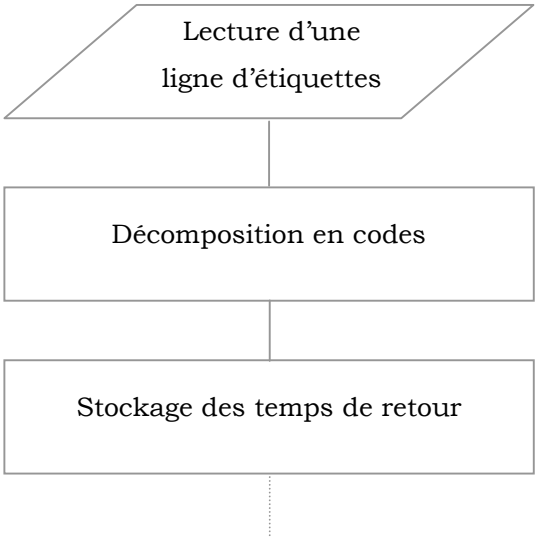


Figure 3.7 : entrées sémantiques

La figure 3.7 précise le processus écrit en *C* :

- la lecture d'une ligne se fait avec la fonction *fgets* ;
- *strtok* permet de récupérer les codes individuels d'une ligne, séparés par des espaces ;
- à chaque occurrence d'un code, un nouvel intervalle est créé pour l'unité correspondante et un temps de retour supplémentaire est engrangé ;

Les traitements ultérieurs sont analogues à ceux des autres niveaux linguistiques.

7 Nanoscopie

Les temps de retour d'une unité sont lus comme dans la microscopie par un programme en *C*, puis stokés dans un fichier spécifique.

En aval, le logiciel *S-Plus* lit ces donnés à l'aide de la fonction *scanf*,

puis calcule les corrélations avec *acf* selon les options « *correlation* » et « *partial* ».

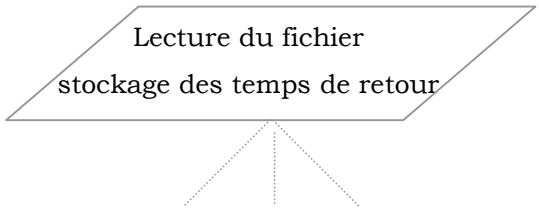


Figure 3.8 : nanoscopie

8 Télescopie

Le calcul de la distance entre deux textes est réalisé par un programme écrit en *C*.

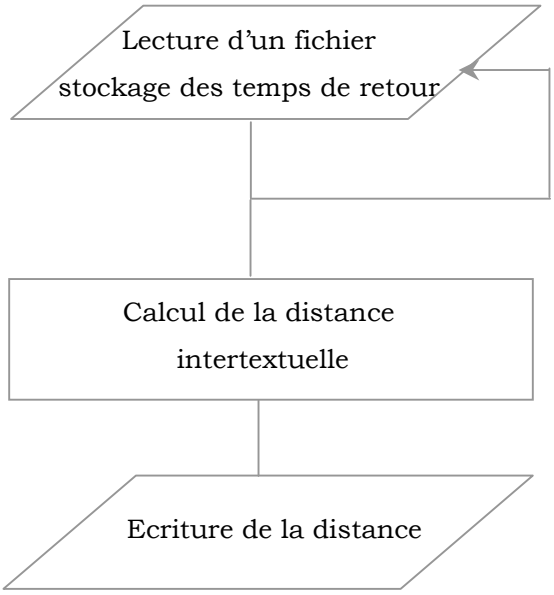


Figure 3.9 : télescopie

Sur la figure 3.9, la première étape de lecture et de stockage est semblable aux précédentes.

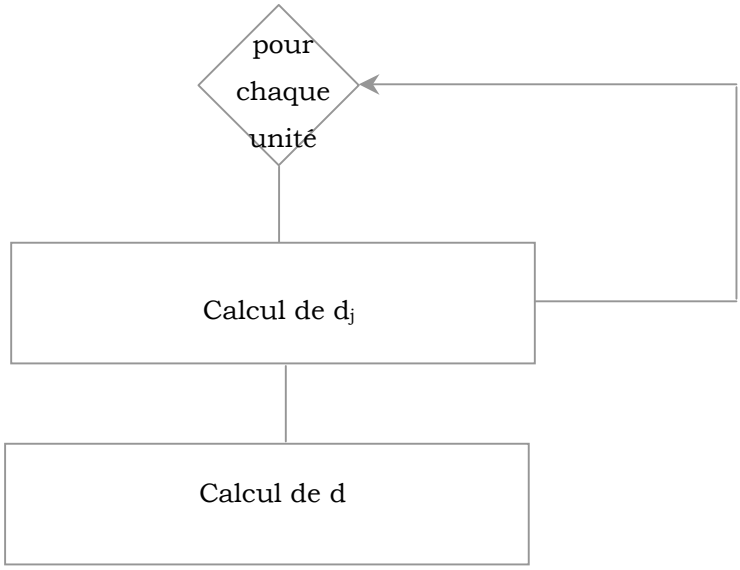


Figure 3.10 : Calcul des distances

La distance globale intègre les distances sur une unité j (fig. 3.10) :

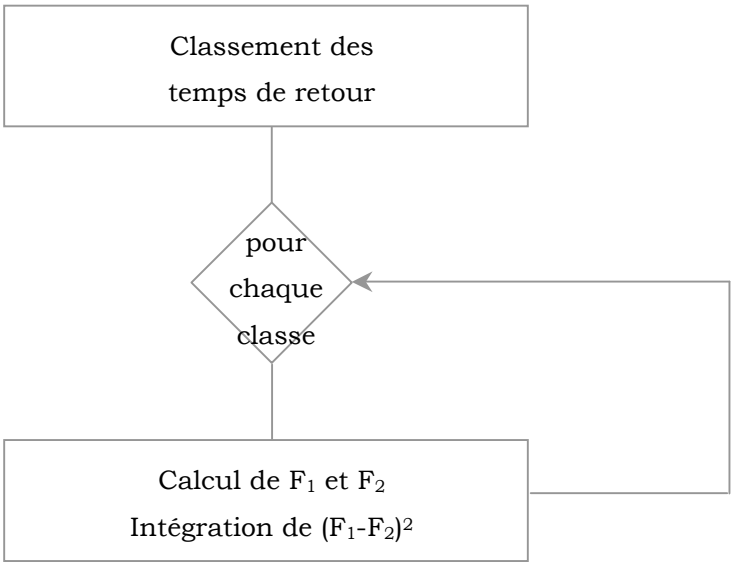


Figure 3.11 : calcul de d_j

Le calcul de dj est spécifié sur la figure 3.11.

Afin de rendre le programme performant, le classement des temps de retour par grandeurs croissantes se fait à l'aide d'une fonction de tri

rapide. La méthode développée par Hoare[342] revient à diviser pour régner : les termes à classer sont réorganisés de part et d'autre d'un pivot, puis le procédé est appliqué récursivement à ces deux sous-ensembles. Si n désigne le nombre d'éléments à trier, le nombre d'opérations se réduit en moyenne à n log n, au lieu de n^2 pour un algorithme classique.

Le calcul des fonctions de répartition F_1 et F_2 associées aux deux textes consiste à faire défiler les temps de retour ainsi triés. Des fronts montants apparaissent à chaque classe non vide de la population considérée. Les écarts sont mesurés par la différence des carrés entre F_1 et F_2, pondérée par le total des élements de la classe.

[342] Hoare, « Quicksort ».

Seconde partie : observations

✯

Ici débute la phase pratique de cette étude. Toute expérimentation se fonde sur la variété des observations et sur la reproductibilité des phénomènes.

Cette partie n'échappe donc pas à un certain systématisme et à des répétitions, prix d'une démarche scientifique. Le lecteur à l'œil rapide pourra cependant se focaliser sur les synthèses qui ponctuent chaque scopie et forment la régularité de la structure qui vient.

La géométrie parle souvent mieux que les mots : les figures sont délibérement et largement mises à contribution pour présenter les chiffres placés en annexe. Afin d'alléger le texte, *Mémoires d'Hadrien* est simplement désigné par *Hadrien ; Vendredi ou les limbes du Pacifique* est abrégé par *Vendredi* ; *Désert* reste inchangé.

Chapitre 4 : macroscopie

1 Introduction

Dans ce chapitre, chaque œuvre est vue comme un bloc opaque : seule sa composition est connue à travers les fréquences des unités.

Le chapitre est découpé selon les plans graphémologiques, syntaxiques et sémantiques. Les comptages qui nourrissent les stylogrammes sont joints dans l'annexe 5.

2 Graphémologie

2.1 Tailles

C'est la mesure la plus élémentaire. Evaluée à partir du nombre de graphèmes, espaces compris, elle sert de référence pour le calcul des fréquences relatives.

Figure 4.1 : tailles des œuvres

Sur la figure 4.1, les écarts autour de la moyenne restent limités à 25 % environ.

2.2 Espaces

Parfois négligé, ce graphème est souvent le plus dense dans un texte[343], et c'est le cas dans notre corpus. Rapport du vide au plein, la proportion des espaces varie peu, mais elle est plus forte dans *Désert* (fig. 4.2).

Coïncidence ou intention plus ou moins consciente de l'auteur ? La question reste de savoir si une différence de l'ordre de 10 %, perçue par l'œil de l'analyste, l'est aussi par le lecteur.

Figure 4.2 : espaces

2.3 Ponctuation

Si l'espace est du vide, la ponctuation évoque le silence. Sa proportion[344] est sensiblement plus forte dans *Désert* que dans les autres œuvres (fig. 4.3) : ce lieu muet paraît s'exprimer par la bouche du style. Avec l'alternance des silences et des sons, Le Clézio rejoint le rythme de la musique et de la danse.

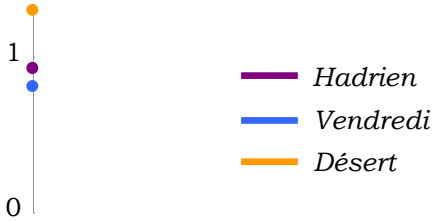


Figure 4.3 : ponctuation

[343] Le texte semble fait à l'image de l'univers physique, essentiellement fait de vide.
[344] Le nombre de signes détaillés dans la figure suivante, divisé par le total des graphèmes.

La figure 4.4 montre la distribution de la ponctuation selon ses symboles.

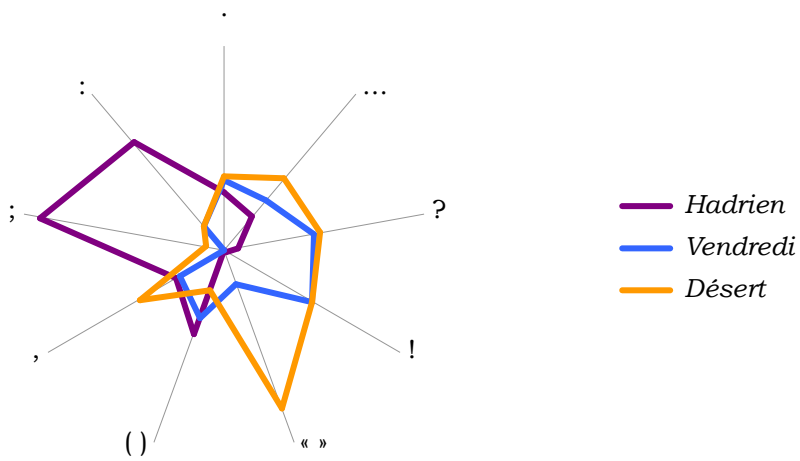


Figure 4.4 : ponctuation

– *Hadrien* privilégie les deux-points et le point-virgule, et dans une moindre mesure les parenthèses, avec un propos résolument didactique. Maîtrisé, son discours laisse peu de place à l'émotion, à la suspension, l'interrogation et l'exclamation. Lettre fictive, le récit est sans dialogue ni guillemets.
– *Vendredi* paraît trop jeune pour apprécier des points-virgules menacés de désuétude.
– *Désert* est marqué par la présence des dialogues et des guillemets. Spontanée, l'écriture passe outre l'usage des parenthèses. Des phrases rythmées par des virgules fréquentes s'achèvent dans le flou de la suspension.

2.4 Lettres

Globalement, le poids des lettres dans l'ensemble des graphèmes est identique d'une œuvre à l'autre (fig. 4.5).

Figure 4.5 : lettres

2.4.1 Alphabet

Le E prédomine dans le corpus comme dans la langue française en général[345]. Les spécificités de chaque œuvre données par la figure 4.6.

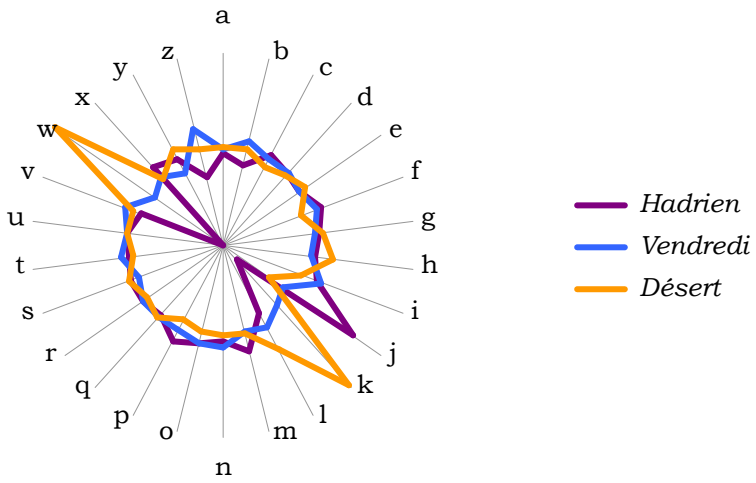


Figure 4.6 : lettres

- *Hadrien* privilégie le J dans un récit écrit à la première personne, mais ignore le K et le W.
- *Désert* se caractérise au contraire par un suremploi de K et de W. La première lettre est fréquente dans une langue arabe souvent gutturale, et Le Clézio emploie volontiers certains de ses mots pour plonger le lecteur dans l'univers de l'Homme Bleu. Spécifiquement, le W est lié au nom de l'héroïne, Lalla Hawa.
- *Vendredi* se tient au milieu de ces deux éléments, avec un pic sur la

[345] Voir Brunet, *Le vocabulaire français de 1789 à nos jours*, p. 182. Cette articulation centrale désigne aussi l'élément neutre d'un groupe mathématique.

lettre Z. L'île Speranza et l'archipel Fernandez n'y sont sans doute pas étrangers.

En l'absence de véritable comptage phonologique, il est hasardeux d'évaluer la répartition des consonnes (antérieures ou postérieures, constrictives ou occlusives, sonores ou sourdes) et des voyelles (antérieures ou postérieures, orales ou nasales, ouvertes ou fermées).

2.4.2 Voyelles et consonnes

Plus synthétiquement, le ratio voyelle/consonne est trop grossier et ne permet pas de distinguer les trois formes (fig. 4.7) :

Figure 4.7 : voyelle/consonne

2.5 Typographie

Un autre axe d'analyse est la forme des caractères. Si l'éditeur fixe la police et la taille, le choix entre les lettres majuscules, minuscules, italiques ou droites est du ressort de l'écrivain.

2.5.1 Majuscule et minuscule

Le ratio majuscule/minuscule, lié aux phrases courtes et à l'emploi de noms propres est assez stable, mais *Désert* se distingue une nouvelle fois d'*Hadrien*, et dans une moindre mesure de *Vendredi* (fig. 4.8).

Figure 4.8 : majuscule/minuscule

2.5.2 Italique

La figure 4.9 illustre la diversité de son emploi. *Vendredi* sort cette fois du lot : les lettres italiques apparaissent d'emblée dans le prologue, mais elles sont disséminées dans le récit pour mettre l'accent sur une portion de la phrase.

Figure 4.9 : italiques

2.6 Synthèse graphémologique

Après cette avalanche de chiffres et de figures, il est temps de faire une première halte pour goûter du panorama. La méthode adoptée jusqu'ici analyse les mesures axe par axe, et se place successivement dans des espaces de dimension un. Le passage à des dimensions supérieures permet d'intégrer les résultats.

L'ensemble des unités est formé des espaces, de la ponctuation et

des lettres. Suivant une approche vectorielle, chaque œuvre est représentée par un point dans un espace multidimensionnel, d'où le calcul de distances deux à deux. Comme ces trois points forment un plan, il est possible de les représenter sur une carte.

Les distances de la figure 4.10 font voir :

– un maximum entre *Hadrien* et *Désert*, qui définissent l'axe du corpus ;
– un minimum entre *Hadrien* et *Vendredi*, qui constituent un sous-groupe.

Figure 4.10 : synthèse graphémologique

3 Syntaxe

Les comptages de ce nivau portent principalement sur les parties du discours et les sorties de *Syntex*.

Cependant, certaines mesures plus fines sont le fruit de *Cordial* : ainsi, les variétés d'adjectifs, d'articles, de noms, de pronoms, les temps et les modes verbaux.

3.1 Parties du discours

3.1.1 Parties inconnues

La catégorie éliminatoire groupe les éléments inconnus de *Syntex*. Paradoxalement riche d'enseignements, elle signe l'écriture libre *Désert*[346] et le classicisme d'*Hadrien* (fig. 4.11) :

Figure 4.11 : parties inconnues

3.1.2 Parties classiques

Sans surprise, les parties les plus fréquentes sont les noms et les verbes. Les particularités stylistiques apparaissent sur la figure 4.12 :

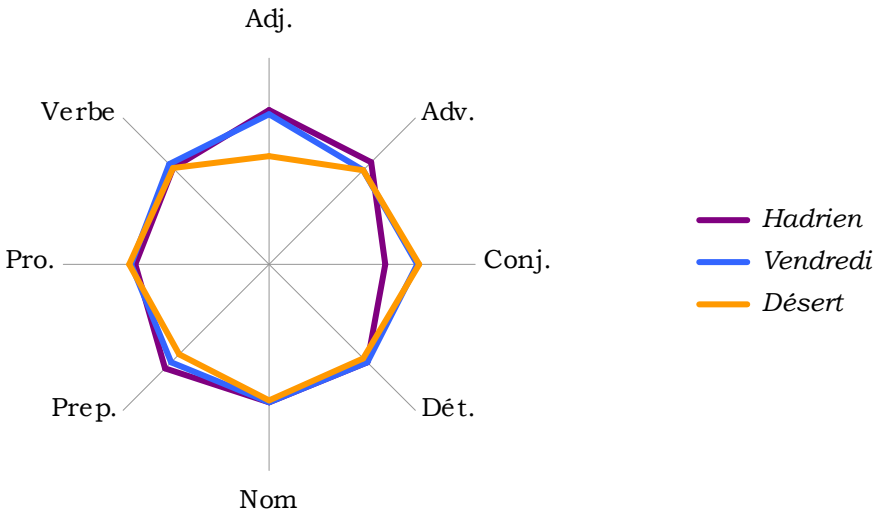


Figure 4.12 : parties du discours

[346] En l'occurrence, ces éléments sont souvent des interjections redoublées.

– *Hadrien* dédaigne les conjonctions : manifestation d'une pensée autocrate peu encline à la relation ?
– *Désert* consomme avec parcimonie les adjectifs : porteurs de jugements, ils sont mal aimés de Le Clézio[347] ;
– *Vendredi* reste dans la moyenne du corpus.

3.1.3 Etre et faire

Plus synthétiquement, quels sont les poids respectifs des groupes nominaux et verbaux ? Additionnons d'une part adjectifs, déterminants, noms et pronoms, d'autre part adverbes et verbes : les écarts sont imperceptibles (fig. 4.13).

Figure 4.13 : groupe nominal/verbal

3.1.4 Genre et nombre

Un autre axe d'analyse porte sur le genre et le nombre. *Syntex* distingue ces aspects pour les adjectifs, les déterminants, les noms et les participes passés.

Paradoxalement, *Vendredi* comporte le plus de formes féminines : le fait étonne avec des acteurs exclusivement masculins. A contrario, *Désert* montre la syntaxe la plus virile malgré une héroïne féminine (fig. 4.14). Les noms communs semblent prédominer sur les noms propres dans les statistiques.

[347] Cf. chapitre 1, section 4.1.2.

Figure 4.14 : féminin/masculin

Vendredi se voit le plus souvent écrit au singulier (fig. 4.15) : effet de la solitude ? Au contraire, *Désert* est ouvert au nombre : la communion avec les éléments naturels, la solidarité de la vie nomade y contribuent sans doute.

Figure 4.15 : pluriel/singulier

3.2 Adjectifs

Le ratio entre les adjectifs démonstratifs et possessifs est donné par la figure 4.16[348].

- *Hadrien* se situe clairement dans l'explication : l'empereur à la vue large sait dépasser l'intérêt personnel et l'avidité propriétaire. En arrière-plan se trouve l'aspiration de Yourcenar pour l'universel.
- *Désert* est ancré dans le réel, d'où la résurgence de la possession, paradoxale dans un espace sans limite et chez un écrivain qui rejette cette entrave : ceci trahit peut-être un attachement inconscient.
- *Vendredi* occupe une position intermédiaire.

[348] Les faibles pourcentages doivent cependant inciter à la prudence.

Figure 4.16 : adjectif démonstratif/possessif

3.3 Articles

Soit à présent le rapport entre les articles définis et indéfinis (fig. 4.17). *Désert* cultive la précision, et c'est un nouveau paradoxe connaissant l'aversion de Le Clézio pour le calcul et la logique : il faut éventuellement y voir la trace d'une écriture concrète. Par contre, *Hadrien* et *Vendredi* diffèrent relativement peu.

Figure 4.17 : article défini/indéfini

3.4 Noms

3.4.1 Noms communs et propres

La figure 4.18 illustre le ratio entre les noms communs et propres. *Vendredi* se révèle le plus abstrait, à l'image d'un Tournier philosophe. En revanche, Le Clézio en quête d' « extase matérielle » recourt aux noms propres pour donner un effet de réel.

Figure 4.18 : nom commun/propre

3.4.2 Noms composés

Vendredi use largement de noms composés : serait-ce la manifestation du binarisme de Tournier ? Au contraire, un Le Clézio à la recherche de l'unité en emploie peu dans *Désert* (fig. 4.19).

Figure 4.19 : noms composés

3.5 Pronoms

La proportion de déictiques est donnée par le ratio 1-2e personne/3e personne (fig. 4.20). La lettre d'*Hadrien* est portée par un « je » omniprésent, tandis que les personnages de *Désert* s'effacent derrière la trame de l'histoire. Une nouvelle fois, *Vendredi* est dans une situation intermédiaire, dans l'alternance du « log-book » et du récit mythique.

Figure 4.20 : pronoms personnels : 1-2/3

3.6 Verbes

3.6.1 Temps

Le futur est absent d'un corpus résolument orienté vers le passé ou le présent, mais des différences apparaissent (fig. 4.21).

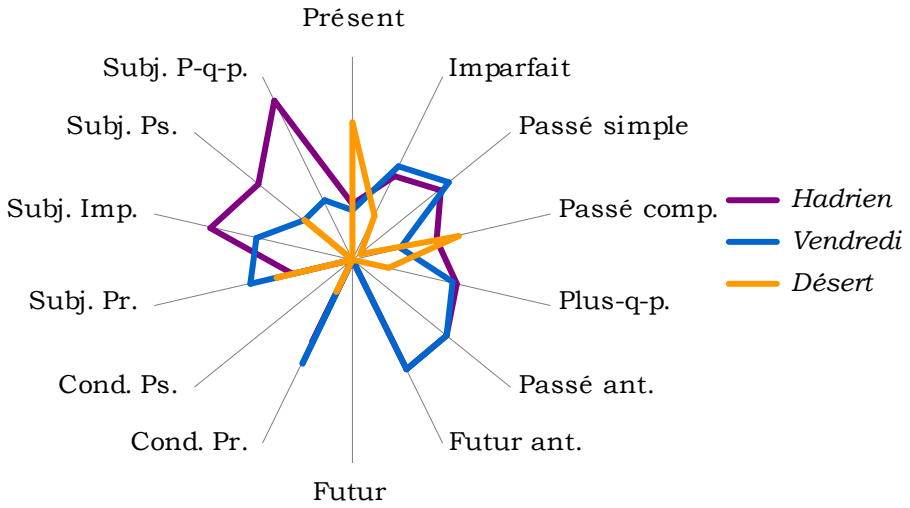


Figure 4.21 : temps verbaux

- Avec une langue soutenue, *Hadrien* se distingue par les temps passés du subjonctif.
- *Vendredi* emploie généreusement le présent du subjonctif et du conditionnel. Le récit de facture classique fait appel au couple du passé simple et de l'imparfait.
- *Désert* s'inscrit dans une tendance contemporaine. Il recourt largement au présent, et dans une moindre mesure au passé composé. Corrélativement, les imparfaits, passés simples, plus-que-parfaits sont rares, ainsi que le présent du conditionnel. Des temps désuets sont même absents, comme les formes antérieures du passé et du futur, ou les imparfaits et plus-que-parfaits du subjonctif.

3.6.2 Modes

Plus synthétiquement, comment se distribuent les modes de ces temps ? Sur la figure 4.22 :

- *Hadrien* emploie volontiers le subjonctif, reflet d'une langue relevée ;
- *Vendredi*, riche en conditionnels, exprime une pensée logique ;

– *Désert* ressort nettement du corpus et privilégie un indicatif ancré dans le réel.

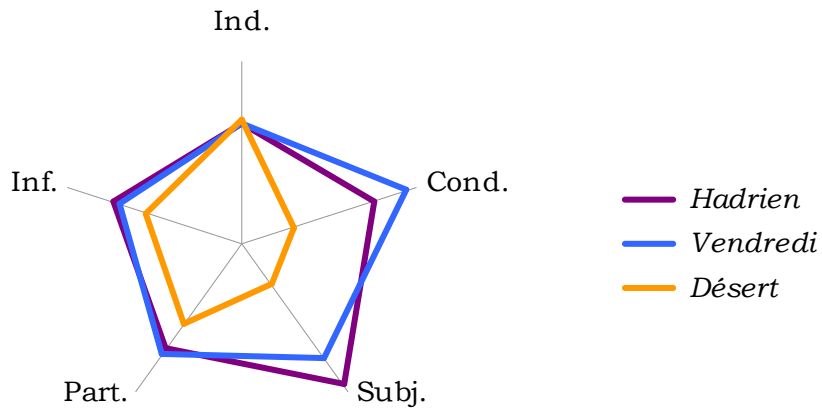


Figure 4.22 : modes verbaux

3.6.3 Orientation temporelle

La répartition entre le passé, le présent et le futur[349] est illustrée par la figure 4.23 : *Hadrien* et *Vendredi* se confondent, tandis que *Désert* se caractérise par la prédominance du présent, la rareté du passé et l'absence du futur.

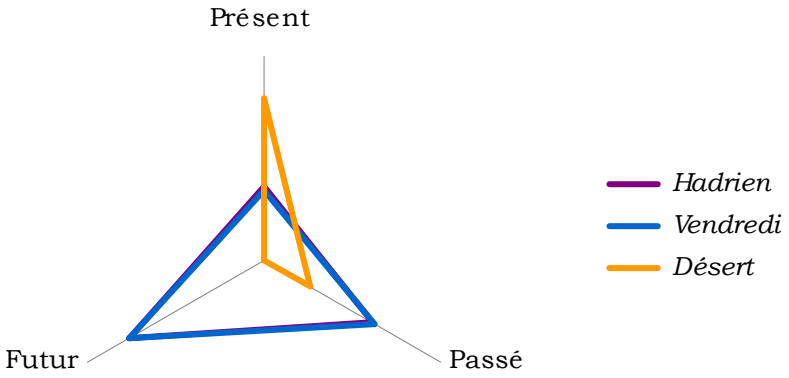


Figure 4.23 : orientation temporelle

3.7 Synthèse syntaxique

La carte de la figure 4.24 est obtenue de façon analogue à la

[349] Le futur antérieur est inclus dans cette dernière catégorie.

graphémologie, par la considération des parties du discours.

Figure 4.24 : synthèse syntaxique

La configuration du niveau graphémologique réapparaît dans ses grandes lignes. En regardant de plus près, *Vendredi* s'est éloigné d'*Hadrien*, rapproché de *Désert*, et semble hésiter plus encore entre ces deux sphères d'influence.

4 Sémantique

Du général au particulier, ce plan est abordé par sa structure — la richesse du vocabulaire et son niveau —, pour analyser ensuite la répartition des concepts. Certains thèmes plus spécifiques sont creusés à la fin.

4.1 Richesse de vocabulaire

Le traitement informatique ne comptabilise que les mots connus du dictionnaire : il exclut de fait la plupart des archaïsmes et des néologismes.

La première idée consiste à calculer le rapport V/N, où V est le nombre de lemmes différents et N le nombre total d'occurrences. Le chapitre suivant montre que ce rapport n'évolue pas linéairement au fil d'une oeuvre : une comparaison rigoureuse prendrait en compte la

longueur du texte. En première approche, l'effet est négligé pour évaluer simplement la richesse.

Nomino extrait les noms, les adjectifs, les verbes et les adverbes sous forme de lemmes, puis estime leurs fréquences. Le rapport recherché en résulte.

Figure 4.25 : richesse du vocabulaire

Sur la figure 4.25, le vocabulaire de *Désert* apparaît singulièrement plus pauvre qu'ailleurs dans le corpus.

Si cette mesure parle, elle ne renseigne que partiellement sur la distribution des fréquences. La richesse du vocabulaire fait retrouver l'inverse du taux de répétition moyen. Une information supplémentaire est donnée par la variabilité : dans le cas idéal où cette grandeur est nulle, un auteur choisit avec équanimité un terme de son registre. A contrario, la sélection du vocabulaire est moins équilibrée dans *Désert* qu'ailleurs dans le corpus : Le Clézio favorise ou délaisse certains mots (fig. 4.26).

Figure 4.26 : variabilité du choix de vocabulaire

La distribution mène au-delà des statistiques d'ordres 1 et 2 : en abscisse se trouve la fréquence d'occurrence d'un terme, et en ordonnée la proportion de ce dernier dans l'ensemble du registre[350]. De gauche à droite sur les courbes, *Désert* compte moins de mots rares et plus de répétitions (fig. 4.27) :

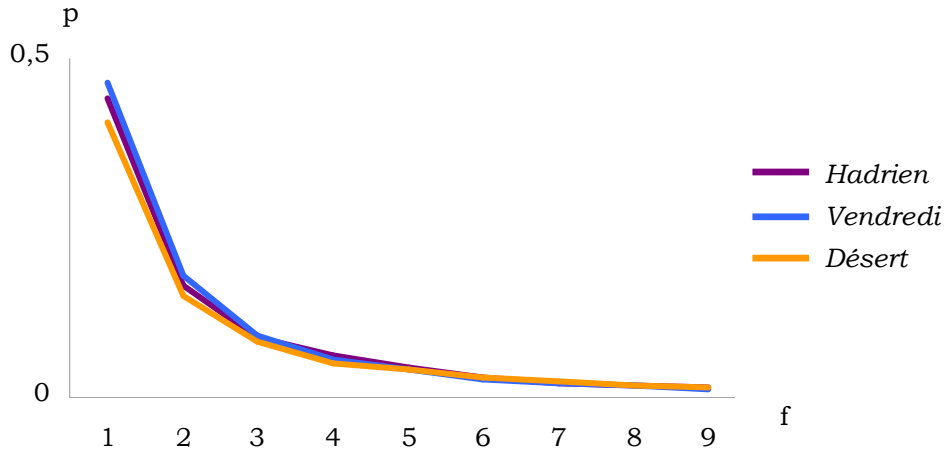


Figure 4.27 : distributions de fréquences

4.2 Niveau de vocabulaire

Cordial reprend la tripartition classique entre les niveaux basiques, usuels et rares. *Hadrien* et *Vendredi* sont sur la même ligne, là où *Désert* tranche par un vocabulaire dépouillé, à l'image du Sahara ou de la vie nomade (fig. 4.28)

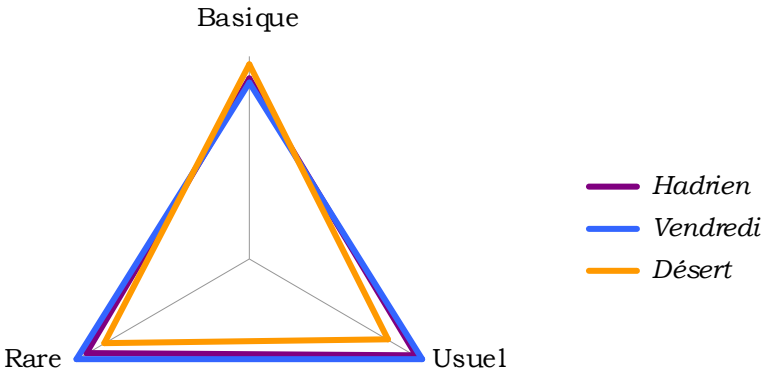


Figure 4.28 : niveau de vocabulaire.

[350] Pour une meilleure lisibilité, seules les basses fréquences sont représentées et la loi discrète est extrapolée entre les abscisses.

4.3 Concepts

De façon générale, le concept « ordre et mesure » domine dans le corpus. Cependant, des différences significatives se dessinent entre les œuvres (fig. 4.29).

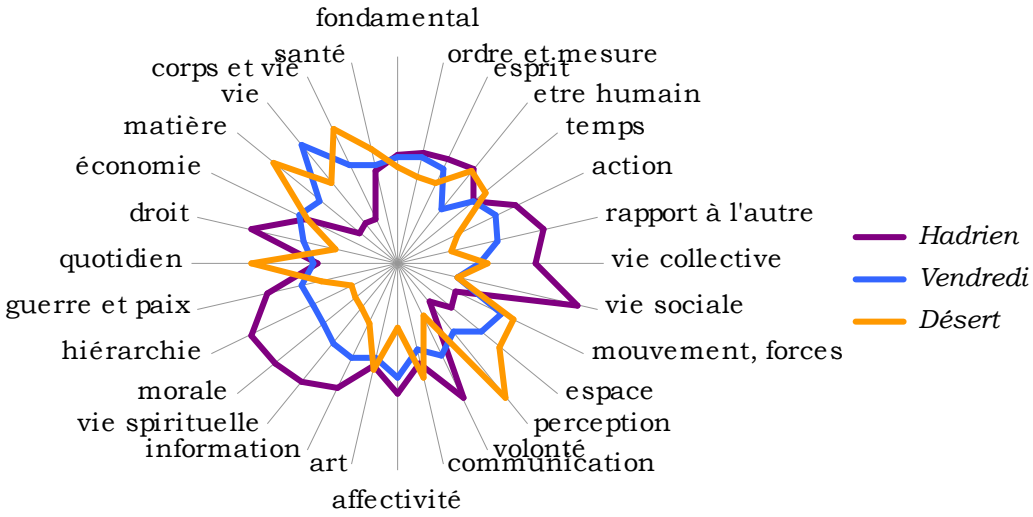


Figure 4.29 : concepts

– *Hadrien* campe dans l'action, le rapport à l'autre, la vie collective et sociale, la volonté, l'information, la vie spirituelle et morale, la hiérarchie, la guerre et la paix, ainsi que le droit. En revanche, le mouvement et les forces, l'espace, la perception, la matière, la vie, le corps jouent un rôle mineur.
– *Vendredi* vit au contraire au contact de son corps, et reste le plus souvent dans une position médiane, sauf en ce qui concerne l'être humain : sa vie est solitaire.
– *Désert* se place dans le mouvement et les forces, l'espace, la perception, le quotidien, la matière, le corps, la santé. Il est en revanche pauvre dans les registres de l'ordre et de la mesure, de l'action, du rapport à l'autre, de la volonté, de l'affectivité, de l'information, de la vie spirituelle, de la morale, de la hiérarchie, du

droit.

4.4 Thèmes

Sur une note ludique, voici les thèmes alchimiques et chromatiques des trois œuvres. Les éléments et les couleurs sont recensés à l'aide de *Cordial*.

4.4.1 Alchimie[351]

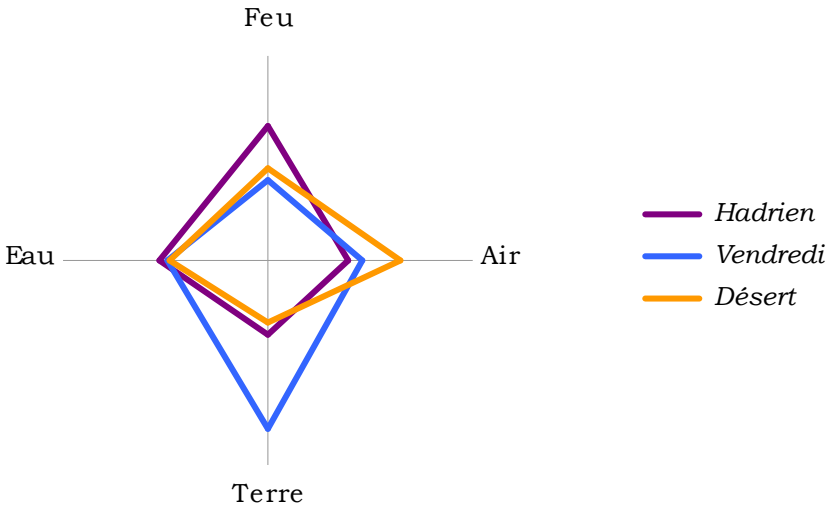


Figure 4.30 : thème alchimique

La figure 4.30 révèle les éléments suivants :

- *Hadrien* est porté vers le feu : la lumière, la lucidité sont des thèmes récurrents chez Yourcenar, et le roman s'achève par ces mots : « les yeux ouverts » ;
- *Vendredi* n'est pas tourné vers l'océan, mais vers son île ; la souille, mélange de terre et d'eau, attire Robinson.
- *Désert* fixe le ciel : sur la terre nue, le regard oublie le divertissement

[351] Bachelard a creusé ce thème dans plusieurs ouvrages *: La Terre et les rêveries du repos, La Terre ou les rêveries de la volonté, L'eau et les rêves, L'air et les songes, La psychanalyse du Feu.*

et revient à l'essentiel ; paradoxalement, l'eau y est autant présente que dans l'île de *Vendredi*.

4.4.2 Couleurs

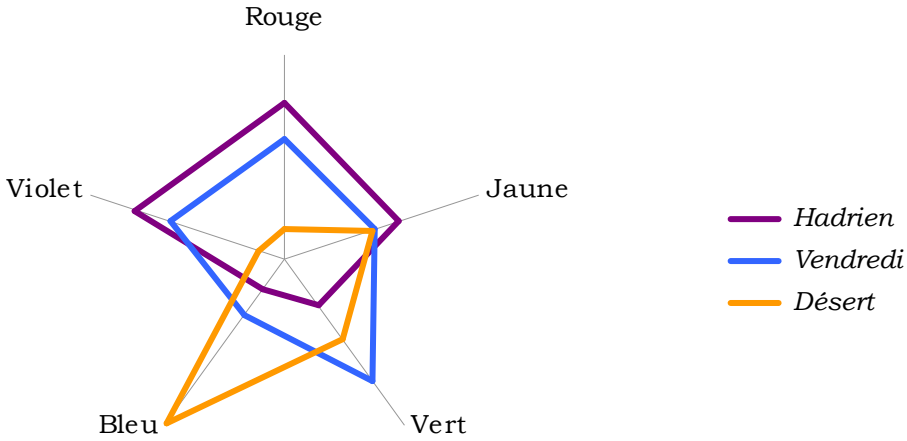


Figure 4.31 : thème chromatique

Les couleurs[352] résonnent avec les éléments (fig. 4.31) :

- *Hadrien* privilégie le violet, le rouge et le jaune : si la première couleur est le privilège de l'empereur, les trois sont celles du feu ; le bleu et le vert de la nature sont peu présents dans un monde centré sur l'humain ;
- *Vendredi* aime le vert de la végétation qui couvre son île ; entre les pôles colorés, le centre du spectre séduit un Tournier en quête de symétrie et d'équilibre ;
- *Désert* reflète l'azur du ciel, les Hommes Bleus ou la mer de sable : Le Clézio exprime d'ailleurs cette préférence lors d'un entretien[353] ; par contre, le rouge et le violet sont peu visibles dans ce lieu..

---

[352] L'orange et l'indigo ne sont malheureusement pas recensés par *Cordial*. Cependant, le comptage brut des mots de couleur ne fournit pas d'échantillon significatif. Un spectre lacunaire est donc retenu pour l'analyse.
[353] Ezine, *Ailleurs*, p. 82.

4.5 Synthèse sémantique

La carte qui intègre les concepts ressemble aux précédentes (fig. 4.32). Cependant, *Vendredi* se rapproche simultanément d'*Hadrien* et de *Désert*.

Figure 4.32 : synthèse sémantique

5 Synthèse macroscopique

Comment dessiner une carte qui rassemble les niveaux graphémologiques, syntaxiques et sémantiques ?

Ce mélange cru de racines a de quoi inquiéter. On postule ici que les différentes strates obéissent fondamentalement à des lois analogues — ce point de vue sera renforcé par les scopies suivantes. La difficulté est pragmatiquement levée en moyennant les distances évaluées sur les trois plans.

Sur la figure 4.33, *Hadrien* et *Désert* sont les romans les plus éloignés du corpus. *Vendredi* est dans une position intermédiaire, plus proche du premier que du dernier.

Figure 4.33 : synthèse macroscopique

Chapitre 5 : mésoscopie

1 Introduction

Le principe général de cette phase consiste à sectionner chaque œuvre et à appliquer successivement la méthode de la macroscopie.

Le découpage d'un texte n'est pas une opération anodine, et la procédure mérite d'être interrogée : c'est l'objet d'une section préliminaire.

1.1 Divisions

Une première méthode consiste à segmenter chaque livre en parties de longueurs égales, par exemple 1000 caractères[354]. Mais ce démembrement numérique hache cruellement les organes de nos créatures littéraires.

Une autre voie épouse la partition de l'auteur.

- *Hadrien* comprend six chapitres titrés.
- *Vendredi* se compose d'un prologue et de douze chapitres titrés et numérotés.
- Les choses se compliquent avec *Désert* : les deux en-têtes, *Le Bonheur* et *La vie chez les esclaves*, ne définissent pas à eux seuls une partition convaincante du roman. Il semble plus pertinent de

[354] Cette problématique est approfondie dans les travaux de Longrée, Luong et Mellet : *Temps verbaux, axe syntagmatique, toopologie textuelle : analyse d'un corpus lemmatisé (2004)* ; *Distance intertextuelle et classement des textes d'apères leut structure : méthodes de découpage et analyses arborées (2006).*

suivre l'entrelacement des récits historiques et fictifs. Dans ce duo vocal, l'alternance est clairement perçue par l'œil grâce à la mise en page et la marge. Neuf séquences sont ainsi identifiées.

1.2 Organisation

Concrètement, le chapitre suit comme précédemment les trois plans de la linguistique. Cependant, on se limite pour simplifier :

- aux espaces, signes de ponctuation et lettres pour la graphémologie ;
- aux parties du discours pour la syntaxe ;
- à la richesse du vocabulaire et aux concepts pour la sémantique.

Pour chacun de ces plans et de ces unités, l'étude comporte deux phases.

- La dynamique ou le suivi des mesures en fonction du temps, vise à situer une division par rapport à l'autre. Pour faciliter la lecture des graphes, les variables sont ramenées à des références communes : l'axe temporel, indexé par le nombre de graphèmes, atteint ainsi toujours finalement la valeur 1 ; de même, les fréquences locales sont divisées par leurs valeurs moyennes sur l'ensemble de l'œuvre[355].
- Plus condensée, la variabilité est le fondement de stylogrammes analogues à ceux de la macroscopie, qui représentent les excentricités par rapport à la moyenne du corpus.

[355] Ces opérations sont sans incidence sur le calcul de la variabilité.

2 Graphémologie

2.1 Tailles

2.1.1 Dynamique

Pour évaluer la taille d'une division, plusieurs voies sont envisageables. Le choix s'est porté sur le nombre de graphèmes, en raison de la finesse de ces grains.

La figure 5.1 représente les évolution au fil du temps[356].

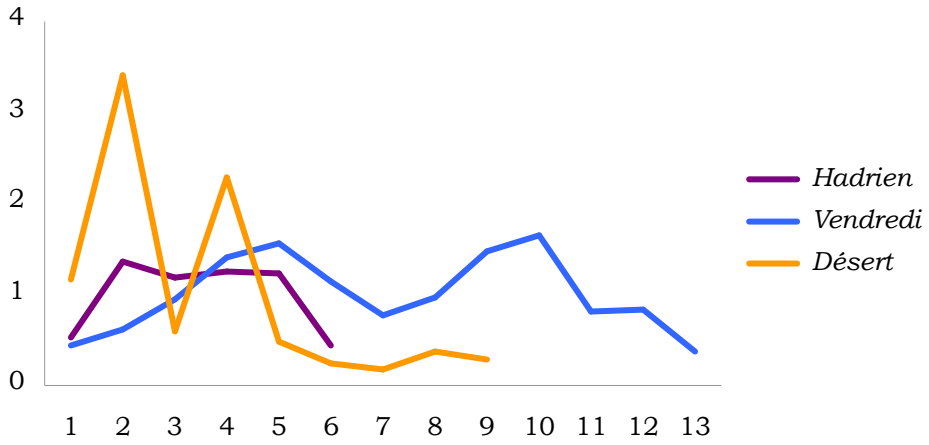


Figure 5.1 : taille des divisions

– Le paysage de *Désert* est accidenté. De fortes oscillations apparaissent initialement et s'amortissent avec le temps, si bien que le livre semble organisé en deux zones : la montagne des quatre premières séquences qui s'achève par l'évanouissement de Lalla, paroxysme de son éloignement ; puis la plaine des cinq suivantes, le retour vers le désert initié par : « la mort est venue ».
– Si le relief de *Désert* est alpestre, celui de *Vendredi* évoque des monts

[356] Chacune a été ramenée à sa valeur moyenne dans l'œuvre correspondante pour faciliter la comparaison.

anciens. Une remarquable symétrie s'opère autour du chapitre 6, avant l'arrivée du compagnon de Robinson.

– La forme d'*Hadrien* est monolithique et classique : un prologue et un épilogue brefs encadrent un massif homogène.

2.1.2 Variabilité

Sur la figure 5.2, *Désert* manifeste une nouvelle fois son atypie.

Evidemment, la dispersion des tailles induit mécaniquement celle des fréquences, qui fluctuent davantage sur des petits échantillons que sur des grands. Cet effet est compensé dans la variabilité puisque les écarts sont pondérés par la taille des divisions[357].

Hormis ces considérations numériques, des facteurs littéraires jouent aussi. Les développements qui suivent précisent ces aspects selon les unités.

Figure 5.2 : variabilité des tailles

2.2 Espaces

2.2.1 Dynamique

Les espaces sont les graphèmes les plus fréquents, d'où la régularité des tracés. *Désert* présente une intéressante alternance, sans doute liée

[357] La variance d'une moyenne emprique décroit linérairement avec la taille de l'échantillon, en supposant que les observations soient indépendantes et suivent la même loi.

à la structure d'un livre partagé entre récit historique et roman : les mots sont plus longs dans le premier registre que dans le second (fig. 5.3) :

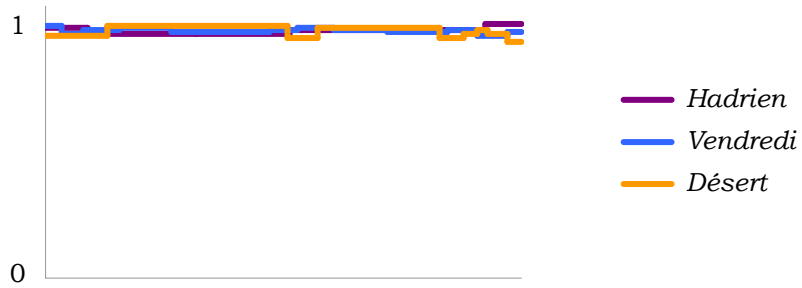


Figure 5.3 : espaces

2.2.2 Variabilité

La gradation entre *Hadrien* et *Désert* apparaît une nouvelle fois sur la figure 5.4 :

Figure 5.4 : espaces

2.3 Ponctuation

2.3.1 Dynamique

De fortes fluctuations apparaissent a contrario sur les unités rares et les parties courtes.

Les points se font plus nombreux au début et à la fin de *Vendredi*, les phrases y sont courtes ; *Désert* fait retrouver l'alternance déjà évoquée, par des phrases plus longues dans le récit historique que dans le roman (fig. 5.5) :

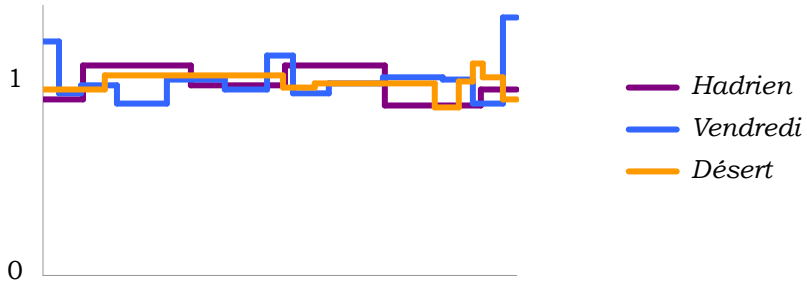


Figure 5.5 : points

Les points de suspension sont particulièrement denses dans le dernier chapitre d'*Hadrien*, symboles d'une vie qui hésite à franchir un cap. Ils sont par ailleurs courants dans le prologue et le chapitre central de *Vendredi* ; leur fréquence semble globalement diminuer depuis l'ouverture de *Désert*, avec un pic furtif dans la cinquième division, déjà pressentie comme l'amorce d'une seconde phase dans le livre (fig. 5.6).

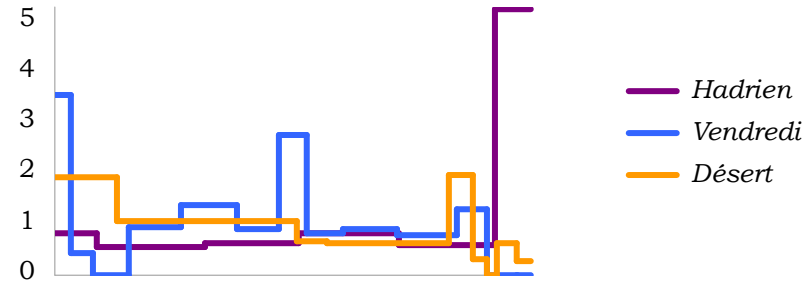


Figure 5.6 : points de suspension

Les interrogations sont pressantes au début d'*Hadrien*, traduisant les doutes de l'Empereur à l'approche de la mort, et paradoxalement dans « Sæculum aureum », apogée de sa vie. Dans *Désert*, cette ponctuation est surtout présente au cours de « La vie chez les esclaves », exprimant le désarroi, la perte de repères d'une Lalla égarée dans le monde moderne. Enfin dans *Vendredi*, son emploi fluctue et culmine lors du prologue, de la sixième division et du chapitre

final[358] (fig. 5.7) :

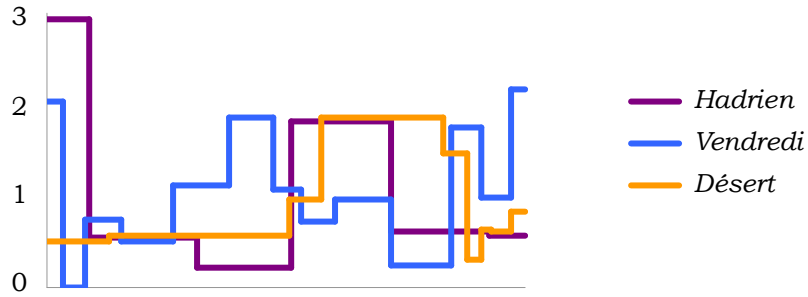


Figure 5.7 : points d'interrogation

Les exclamations, rares dans le monde mesuré d'*Hadrien*, sont cantonnées au début du livre. Dans *Vendredi*, elles se font surtout entendre dans le prologue et après l'irruption du jeune homme dans la vie de Robinson. Elles s'espacent progressivement au fil de *Désert* (fig. 5.8) :

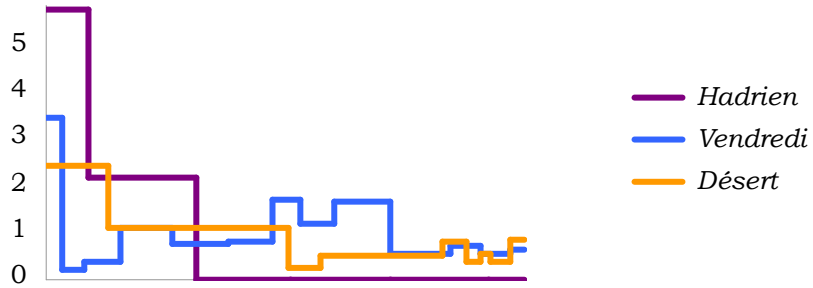


Figure 5.8 : points d'exclamation

Les guillemets sont absents du monologue d'*Hadrien*. Dans la solitude de *Vendredi,* il exprime souvent une distanciation, notamment au cours du « log-book » du chapitre 10. Dans *Désert*, ce signe ponctue la prise de parole : fréquente dans le chapitre d'ouverture, elle se raréfie dans un second temps pour s'effondrer finalement (fig. 5.9) :

[358] Nous retrouvons, sous une forme inversée, le relief observé sur la taille des divisions.

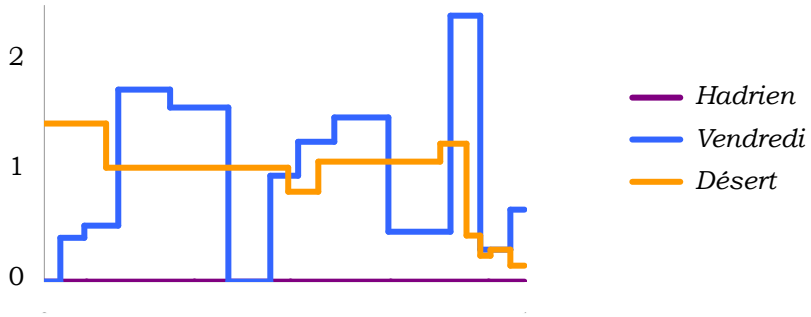


Figure 5.9 : guillemets

Les parenthèses peu nombreuses déclinent au fil d'*Hadrien*, et se manifestent dans le chapitre 8 de *Vendredi*. Elles ne s'ouvrent que dans les parties les plus longues de *Désert*, « Le Bonheur » et « La vie chez les esclaves » (fig. 5.10) :

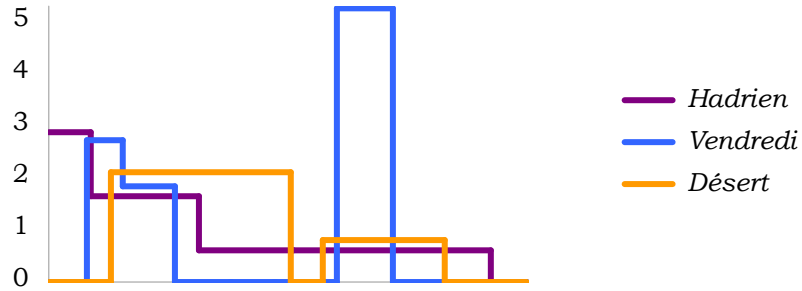


Figure 5.10 : parenthèses

Les virgules se raréfient dans *Hadrien*, et se font au contraire plus fréquentes du début à la fin de *Désert* (fig. 5.11) :

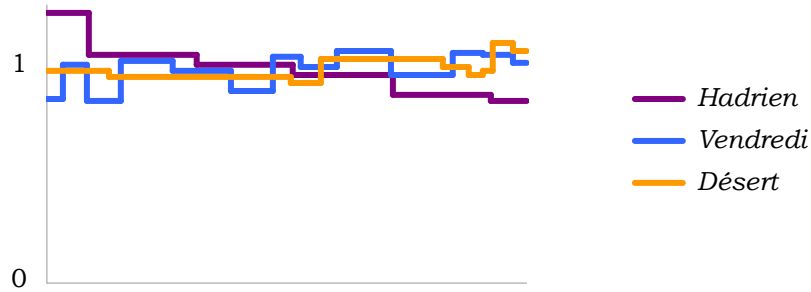


Figure 5.11 : virgules

Le recours au point-virgule s'accroît sûrement dans *Hadrien*, et plus

irrégulièrement dans *Désert*. Il n'apparaît dans *Vendredi* que lors du prologue et du chapitre 8 (fig. 5.12) :

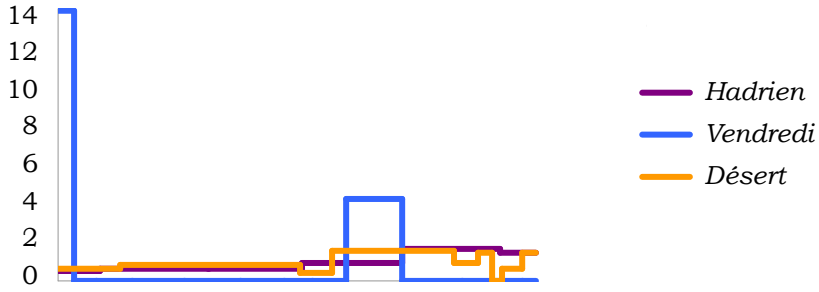


Figure 5.12 : point-virgule

Les deux-points semblent à l'image d'*Hadrien* et suivent curieusement la trajectoire de sa vie : expansion, apogée, chute, lente remontée (fig. 5.13) :

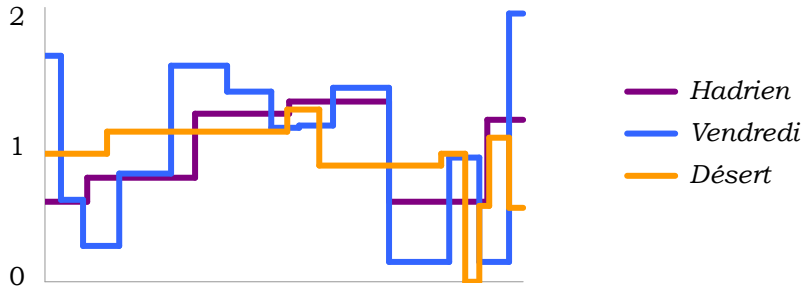


Figure 5.13 : deux-points

2.3.2 Variabilité

La figure 5.14 montre que :

– *Hadrien* fait varier les points de suspension, d'interrogation et d'exclamation, ainsi que les virgules ; en ce qui concerne les guillemets absents de cette œuvre, la variabilité est indéfinie et fixée à 0 pour la commodité de la représentation[359] ;

[359] Le quotient de deux zéros plonge vers l'incertitude. Avec cette convention, la contribution de cette unité est nulle dans la synthèse de la variabilité, si bien que cette unité absente reste sans effet.

– *Vendredi* fait varier les guillemets, les parenthèses, les points-virgules et les deux-points ;
– *Désert* est plus stable que les autres éléments du corpus.

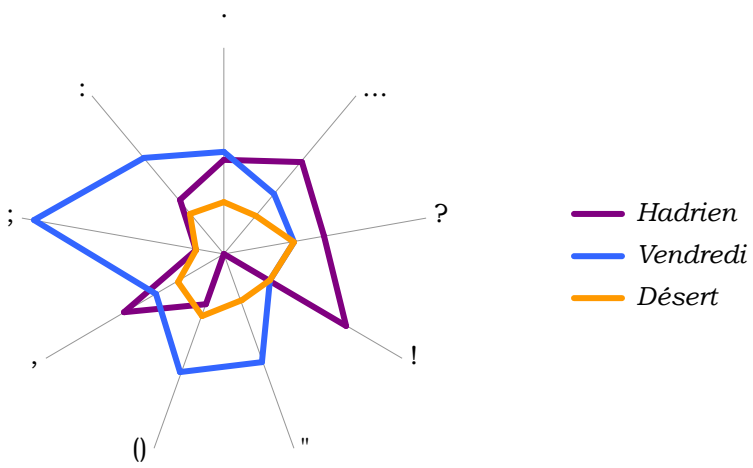


Figure 5.14 : ponctuation

2.4 Lettres

Dans ce qui suit, les minuscules et les majuscules sont confondues tandis que les caractères accentués sont ignorés.

2.4.1 Dynamique

Le passage en revue de l'ensemble des lettres pourrait rendre cet exposé fastidieux, et l'objectif de ce chapitre n'est pas de refaire la macroscopie à une échelle plus petite. Seuls les éléments les plus variables du corpus sont donc analysés. Les autres graphes, néanmoins pris en compte dans la synthèse de ce plan, sont tracés en annexe 5.

Le J est sensible dans *Vendredi* au cours du chapitre 10 écrit à la première personne (fig. 5.15) :

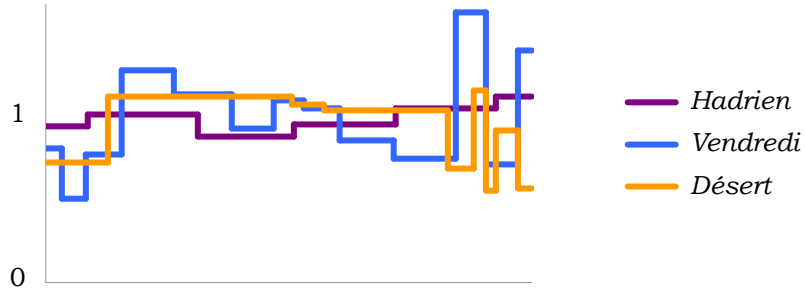


Figure 5.15 : J

Le K se fait entendre par la figure du grand cheikh dans le récit historique de *Désert*,; au cours du chapitre 10 de *Vendredi*, il prend les traits d'un « log-book » onze fois remis en page. Dans *Hadrien* enfin, il se fait plus présent avec le juif Akiba, pour disparaître totalement lors de l'épilogue (fig. 5.16) :

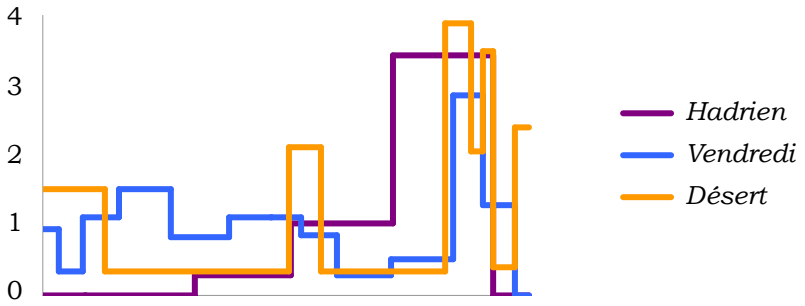


Figure 5.16 : K

Le W est absent d'*Hadrien* ; il surgit vigoureusement dans *Vendredi* avec l'arrivée du Whitebird anglais au chapitre 11. Dans *Désert*, il renvoie à l'héroïne, dénommée par Hawa durant son séjour en France, et par Lalla dans son pays d'origine (fig. 5.17) :

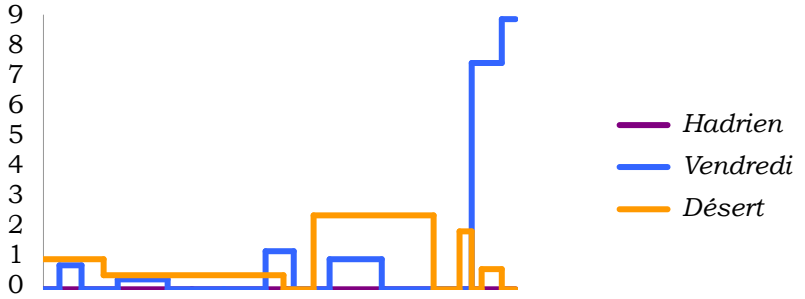


Figure 5.17 : W

Le Z, très présent dans le prologue de *Vendredi*, se conjugue avec le capitaine Van Deysel, vouvoyant Robinson. Dans *Désert*, son emploi est lié à la mort de Radicz au cours de la sixième division (fig. 5.18) :

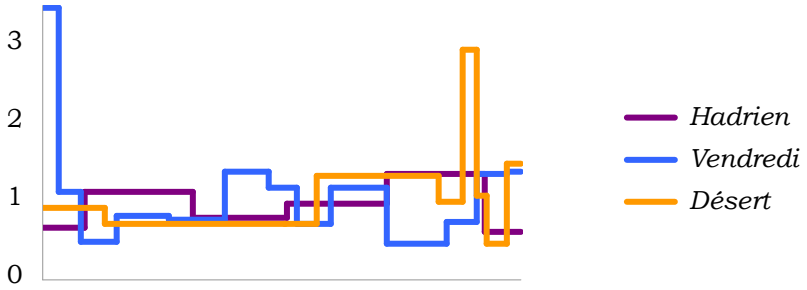


Figure 5.18 : Z

2.4.2 Variabilité

Sur la figure 5.19 :

– *Hadrien* est généralement plus stable que les autres œuvres ; la variabilité du W absent est conventionnellement représentée par 0, de façon analogue aux guillemets ;
– *Vendredi* fait varier les B, F, J, M, U, V, X, Y, Z ;
– *Désert* fait varier les E, H, I, L, P, Q, R, T.

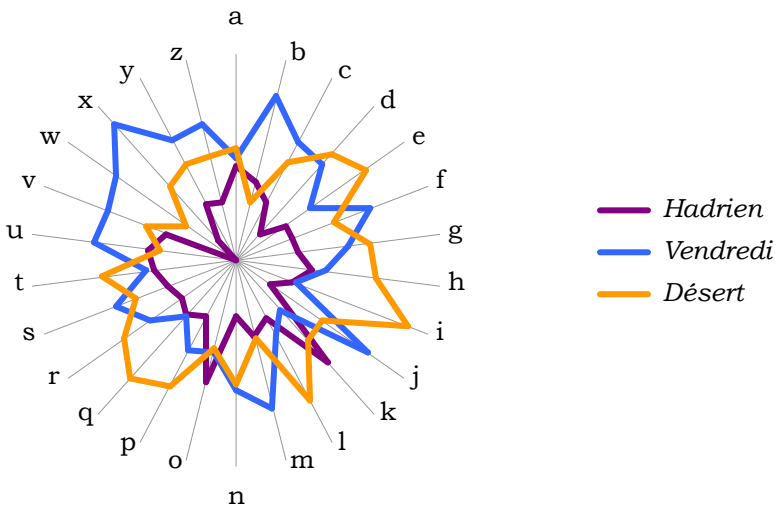


Figure 5.19 : lettres

2.5 Synthèse graphémologique

2.5.1 Dynamique

Elle représente la distance de chaque division par rapport à l'œuvre complète.

Pour faciliter les comparaisons, les valeurs sont normalisées. La distance de référence est liée à la variabilité du paragraphe suivant (cf. chapitre 2, section 4.2.1).

La figure 5.20 trace ces distances en fonction du temps :

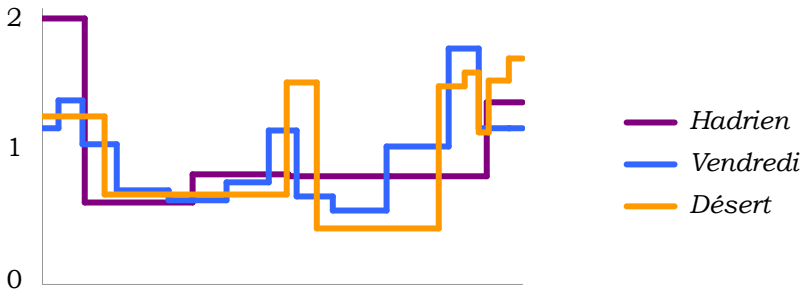


Figure 5.20 : dynamique

– les débuts et les fins des œuvres sont excentrés ;
– un pivot central apparaît dans *Vendredi* au chapitre 6, et dans *Désert* lors du premier exode vers le nord.

2.5.2 Variabilité

La mesure globale intègre les variabilités de ce plan, et fait retrouver la gradation déjà observée (fig. 5.21) :

Figure 5.21 : variabilité

3 Syntaxe

Les regroupements des étiquettes de *Syntex* sont analogues à ceux de la macroscopie, et ramènent aux parties du discours.

3.1 Parties du discours

3.1.1 Dynamique

Les adjectifs mal aimés de Le Clézio pointent à la surface des dernières divisions de *Désert*. Dans *Hadrien*, ils sont en excès dans « Tellus Stabilita » et en déficit dans « Patientia » (fig. 5.22) :

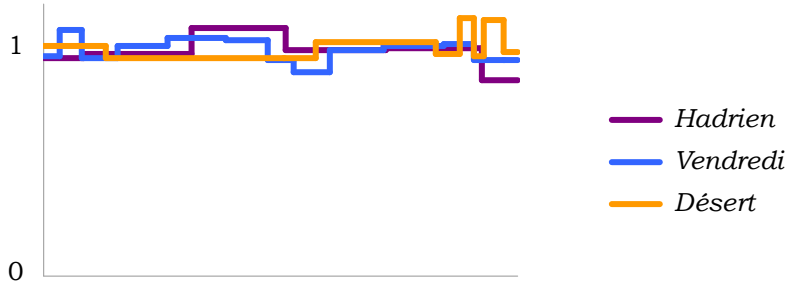


Figure 5.22 : adjectifs

Les adverbes et les conjonctions sont plus présents dans la partie romanesque de *Désert* que dans le récit historique. Dans *Hadrien*, ces parties se font de plus en plus rares pour atteindre un minimum dans « Sæculum aureum » et remonter à la fin : la configuration est symétrique par rapport à celle des deux-points (fig. 5.23 et 5.24) :

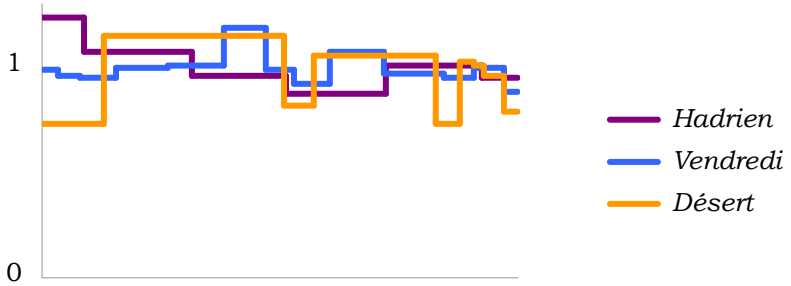


Figure 5.23 : adverbes

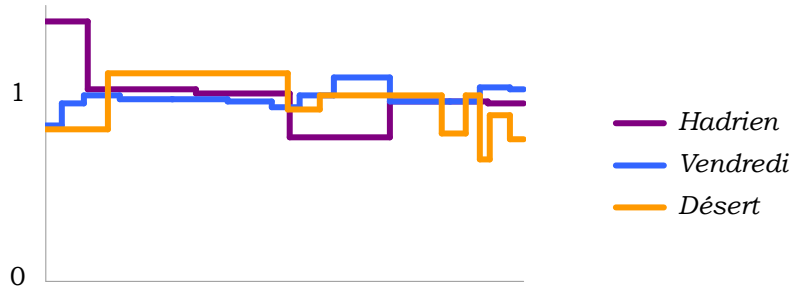


Figure 5.24 : conjonctions

L'alternance déjà évoquée à propos de *Désert* se retrouve sur le tracé des conjonctions ; paradoxalement, le récit historique censé véhiculer l'action est plus nominal que la partie romanesque (fig. 5.25 et 5.26) :

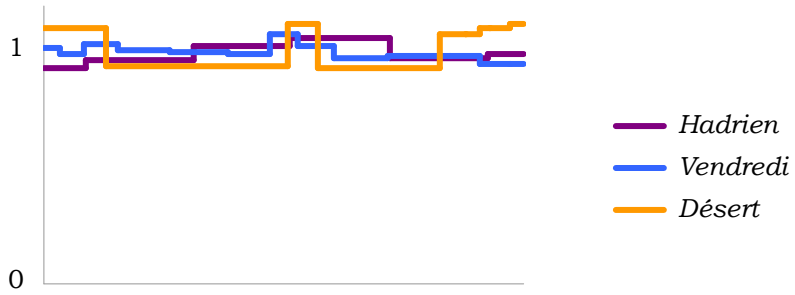


Figure 5.25 : déterminants

Figure 5.26 : noms

Les prépositions retracent le balancement prédédemment observé dans *Désert* (fig. 5.27) :

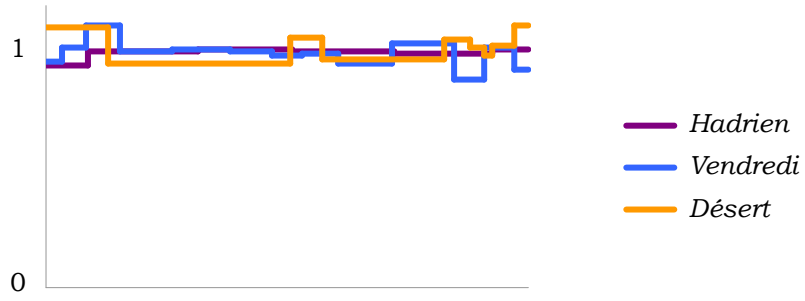


Figure 5.27 : prépositions

Les pronoms suivent une loi analogue aux adverbes et aux conjonctions ; dans *Hadrien*, le minimum est cette fois atteint au cours de « Tellus stabilita » (fig. 5.28) ;

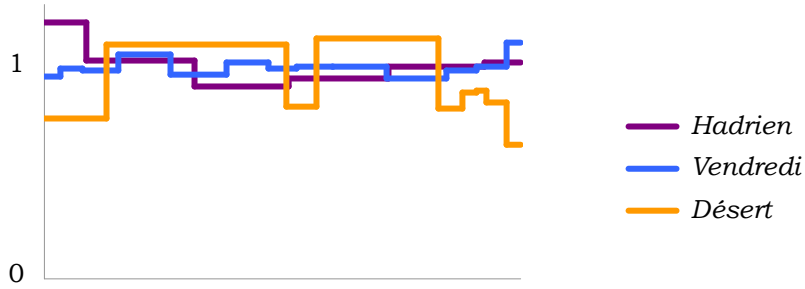


Figure 5.28 : pronoms

Les verbes sont logiquement en opposition de phase par rapport au noms dans *Désert* (fig. 5.29) :

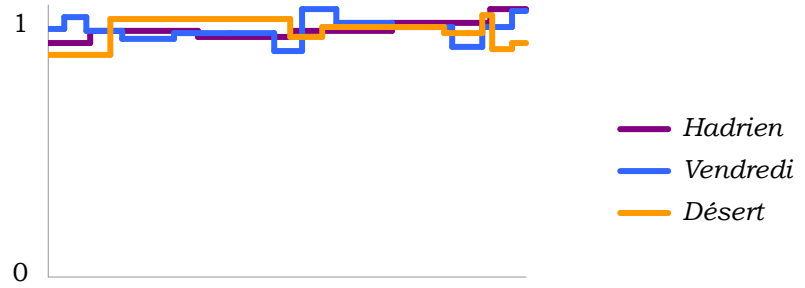


Figure 5.29 : verbes

3.1.2 Variabilité

La figure 5.30 donne une vue générale de ces parties :

– *Hadrien* fait varier les adverbes et les conjonctions ;

– ailleurs, *Désert* se montre le plus variable.

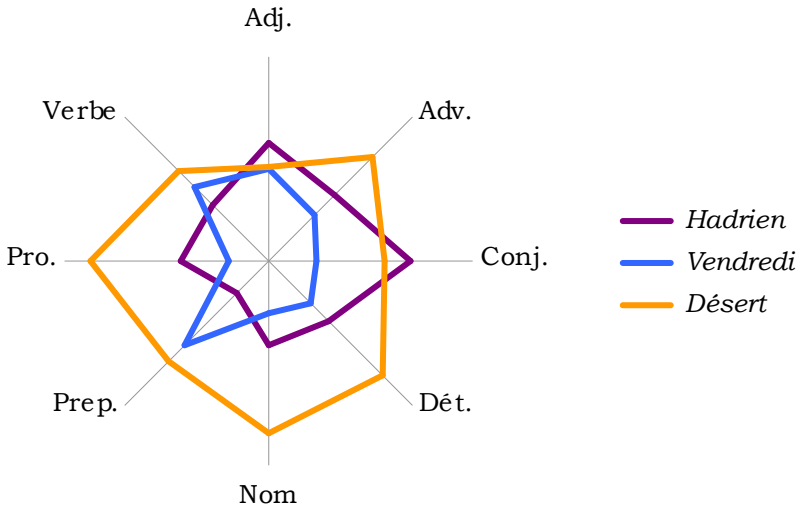


Figure 5.30 : parties du discours

3.2 Synthèse syntaxique

3.2.1 Dynamique

La dynamique des divisions est indiquée par la figure 5.31 :

– à nouveau, les débuts et les fins des œuvres sont excentrés ;

– un pivot central apparaît dans chaque cas : « Sæculum aureum » pour *Hadrien*, le chapitre 6 pour *Vendredi*, et le premier exode vers le nord pour *Désert*.

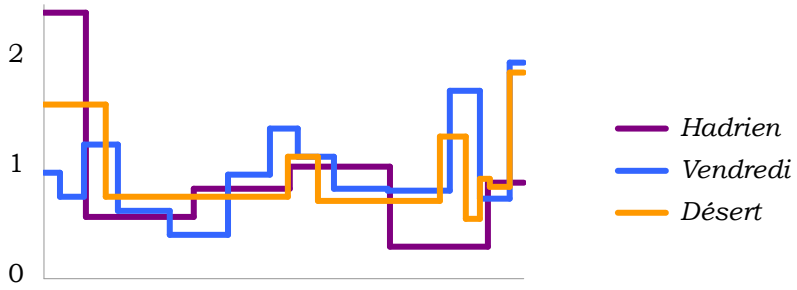


Figure 5.31 : dynamique

3.2.2 Variabilité

Les contributions partielles sont intégrées afin d'évaluer cette grandeur à l'échelle de l'œuvre : *Désert* montre clairement sa plus grande dispersion (fig. 5.32).

Figure 5.32 : variabilité

4 Sémantique

Comme au chapitre précédent, la structure de ce plan précède l'analyse de ses éléments.

4.1 Richesse du vocabulaire

4.1.1 Dynamique

Plus que la richesse globale du vocabulaire, son évolution au fil du livre est porteuse de sens. Le processus est généralement non linéaire : si la première forme est nécessairement nouvelle, la progression se fait plus difficile au cours de l'œuvre, en fonction de l'étendue du vocabulaire de l'auteur.

Le nombre de formes (V) est mis en regard du nombre d'occurrences (N) pour l'ensemble des noms, adjectifs, verbes et adverbes. Comme l'enrichissement réel se fait régulièrement à chaque mot, les courbes de la figure 5.33 lissent les paliers formés par la partition du texte.

Hadrien et *Vendredi* sont proches, et l'enrichissement de leur vocabulaire est rapide. Dans *Désert*, une cassure s'opère à la fin du premier récit mythique, puis le roman se poursuit sur un mode plus prosaïque.

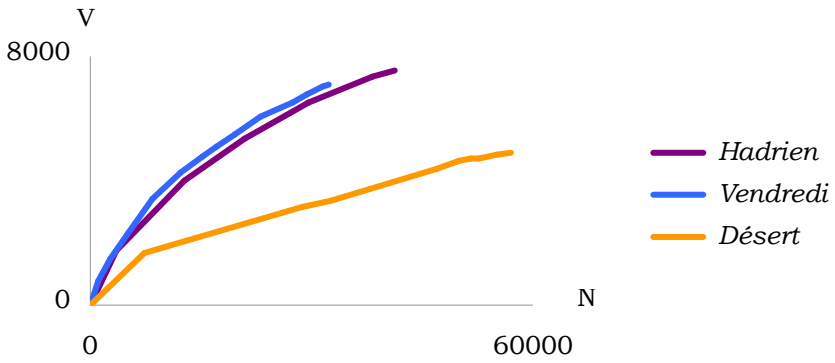


Figure 5.33 : enrichissement du vocabulaire

4.1.2 Variabilité

Son estimation suppose des mesures homogènes et un processus stationnaire. Pour un enrichissement non linéaire, ces conditions ne sont pas vérifiées et l'analyse n'est pas réalisée.

4.2 Concepts

4.2.1 Dynamique

Seuls les éléments les plus variables sont présentés pour alléger la section. Les autres graphes sont tracés en annexe 5.

La vie sociale prend une forte ampleur à la fin de *Désert*, dans les trois dernières divisions du récit historique. Elle procède par flux et reflux dans *Vendredi*, les chapitres 4[360], 7 et 11 formant les crêtes de

[360] Paradoxalement, avant l'arrivée de Vendredi.

trois vagues. Dans *Hadrien*, elle culmine discrètement au cours de « Varius multiplex multiformis » (fig. 5.34) :

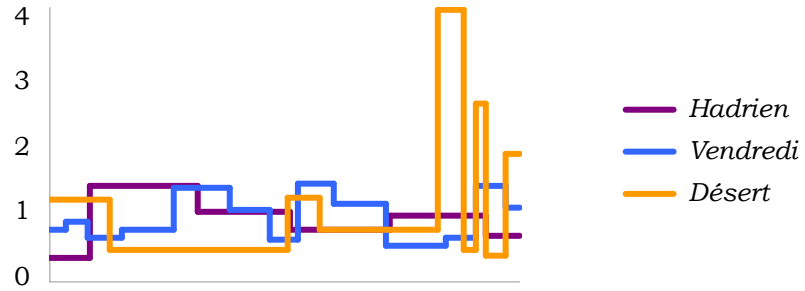


Figure 5.34 : vie sociale

L'information se dissémine généralement dans la partie romanesque de *Désert* pour se dévoiler massivement lors de « La vie chez les esclaves ». Dans *Vendredi*, elle frappe trois grands coups au moment des chapitres 3, 8 et 12. Enfin dans *Hadrien*, elle prend de l'ampleur, atteint son âge d'or au cours de « Disciplina augusta » avant de retomber dans l'oubli (fig. 5.35) :

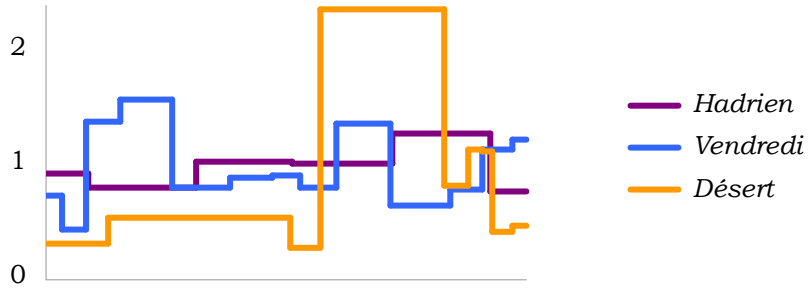


Figure 5.35 : information

La vie spirituelle prend son envol au fil d'*Hadrien*, atteint un premier sommet avec « Sæculum aureum », tombe et reprend son souffle avant de gagner d'autres cieux. Dans *Vendredi*, elle privilégie le prologue ainsi que les chapitres 6 et 10. Enfin, son influx est sensible au début et à la fin de *Désert* (fig. 5.36) :

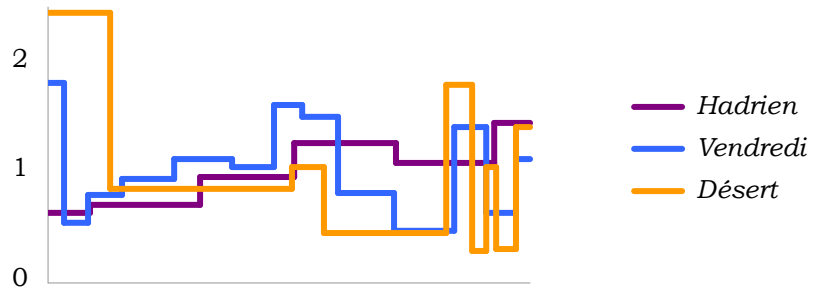


Figure 5.36 : vie spirituelle

La guerre et la paix, larvées durant le récit historique de *Désert*, éclatent au cours des dernières divisions. Si elles épargnent le prologue, l'épilogue et l'âge d'or d'*Hadrien*, elles se font entendre dans « Varius multiplex multiformis » (fig. 5.37) :

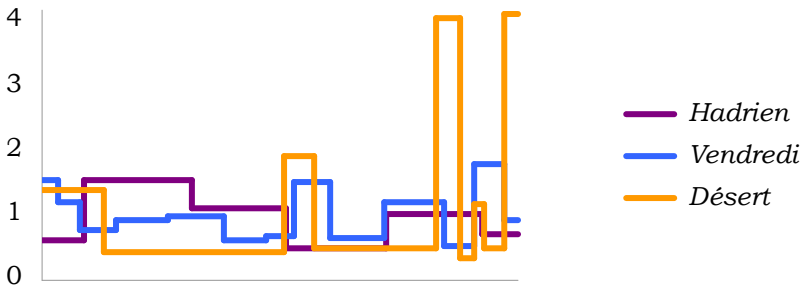


Figure 5.37 : guerre et paix

Dans *Hadrien*, le droit suit deux phases ascendantes qui scellent « Tellus stabilita » et « Patientia ». Dans *Vendredi*, il fonde la charte de l'île de Speranza. Il marque enfin le second exode vers le nord et les opérations militaires dans *Désert* (fig. 5.38) :

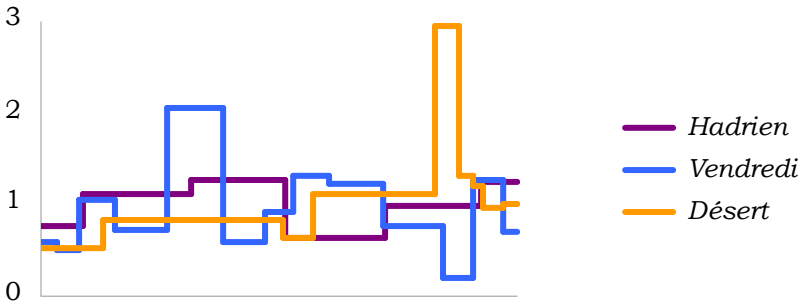


Figure 5.38 : droit

4.2.2 Variabilité

Sur la figure 5.39 :

- *Hadrien* est stable et ne fait surface que dans les domaines de la perception et de l'affectivité ;
- *Vendredi* fait varier le fondamental, l'être humain, le temps, le mouvement et les forces, l'espace, la morale, l'économie, le corps et la vie, ainsi que la santé ;
- *Désert* fait varier l'ordre et la mesure, le rapport à l'autre, la vie sociale, l'information, la vie spirituelle, la guerre et la paix, le quotidien et la matière.

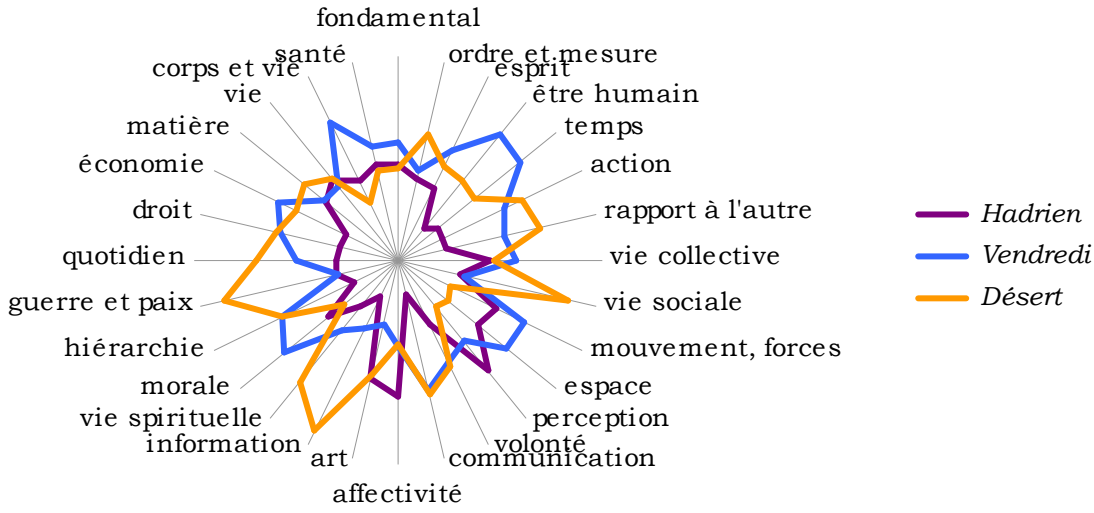


Figure 5.39 : concepts

4.3 Synthèse sémantique

4.3.1 Dynamique

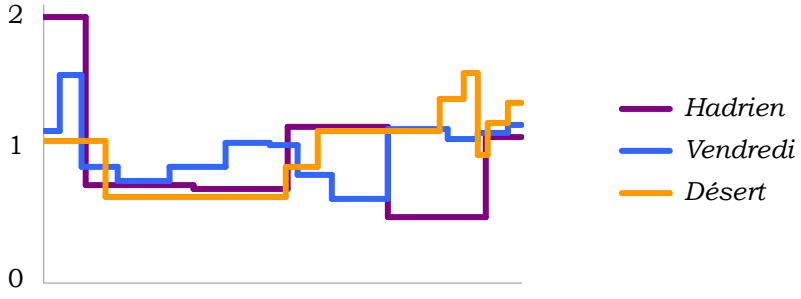


Figure 5.40 : dynamique

La figure 5.40 représente la position de chaque division par rapport à sa référence :

– le début, la fin et un pivot central se détachent dans *Hadrien* et *Vendredi* ;
– *Désert* met principalement en relief ses dernières divisions.

4.3.2 Variabilité

Sur l'ensemble des concepts, *Hadrien* apparaît comme l'œuvre la plus stable du corpus (fig. 5.41) :

Figure 5.41 : concepts

5 Synthèse mésoscopique

Les résultats de ce chapitre sont rassemblés comme dans la macroscopie par la moyenne des chiffres recueillis sur les trois plans linguistiques.

5.1 Dynamique

La figure 5.42 fait apparaître une structure en W étrangement familière entre les trois œuvres, composée de leur début, de leur fin et d'un pivot central[361] : « Sæculum aureum » dans *Hadrien*, le chapitre 6 qui précède l'arrivée de Robinson dans *Vendredi*, et le premier exode vers le nord dans *Désert*.

[361] Paradoxalement excentrique, ce pivot peut être assimilé à une phase de transition.

Cette structure rappelle, mais avec plus de finesse et en l'inversant, la forme en M dessinée par la taille des divisions (fig. 5.1). Cette explication quelque peu désenchantée n'est sans doute que partielle, car la distribution numérique pourrait refléter une organisation profonde et un mode de pensée.

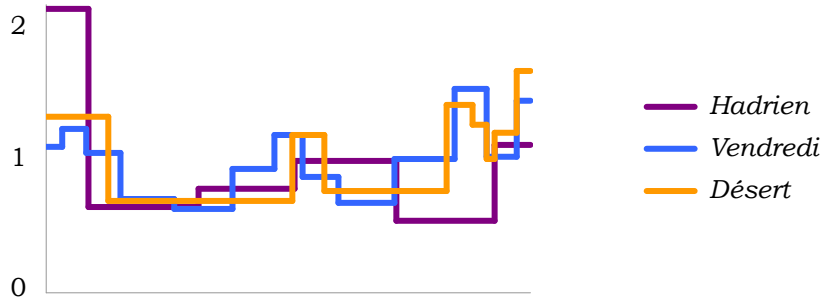


Figure 5.42 : dynamique

5.2 Variabilité

La figure 5.43 compare la variabilité des différentes œuvres : la gradation déjà observée dans la macroscopie réapparaît.

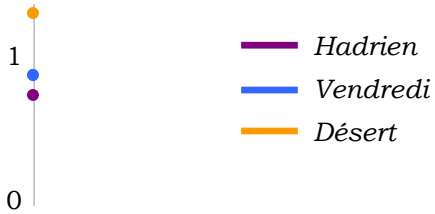


Figure 5.43 : variabilité

Chapitre 6 : microscopie

1 Introduction

Cette phase consiste essentiellement à étudier les temps de retour d'une unité.

Le chapitre est organisé comme les précédents, selon les trois plans linguistiques. Les moments et les spectres sont présentés unité par unité, avant d'être synthétisés.

1.1 Moments

Les statistiques principales pour chaque œuvre et chaque unité sont chiffrées en annexe 6 : le nombre d'intervalles, la moyenne qui sert de référence, ainsi que la variabilité et l'asymétrie sous forme réduite.

1.2 Spectres

Les valeurs extrêmes et bruitées sont filtrées pour se limiter à 80 % des individus, soit un échantillon qui reste représentatif de la population.

Sur les spectres absolus, les abscisses marquent un temps de retour brut, pour fixer les ordres de grandeurs physiques ; si les temps expérimentaux sont discrets, des valeurs intermédiaires sont interpolées pour améliorer la lisibilité et se ramener aux lois continues de la fiabilité.

Sur les spectres relatifs, les abscisses représentent un temps normé par sa valeur moyenne, dans le but d'homogénéiser les résultats et de faciliter les comparaisons entre les unités.

L'annexe 6 fournit en outre :

– les modes ou périodes les plus fréquentes, parfois difficilement lisibles sur les spectres qui donnent une vision globale : ces valeurs se distinguent des moyennes en fonction de l'asymétrie des distributions ;
– les séquences mises en évidence sur les courbes : un temps de retour extrême N couvre un ensemble de n-grammes dont le plus fréquent est retenu ; ce dernier élu se manifeste généralement sous différentes formes, d'autant plus cernées que N est grand.

2 Graphémologie

2.1 Espaces

2.1.1 Moments

Les statistiques sur ce graphème donnent une information sur la longueur des mots. Mais il s'agit ici des écarts par rapport à la moyenne, et non de grandeurs brutes comme dans la macroscopie.

Les espaces apparaissent régulièrement dans *Désert,* quand les paysages de *Vendredi* ou *Hadrien* sont plus heurtés (fig. 6.1). Les asymétries répondent à la même logique (fig. 6.2).

Figure 6.1 : variabilité

Figure 6.2 : asymétrie

2.1.2 Spectres

Pour toucher les textes du doigt, examinons les spectres absolus à gauche de la figure 6.3. Après une première apparition, l'espace ressurgit couramment trois temps plus tard : le mode traduit notamment la séquence « de » . Puis les courbes divergent et *Désert* se démarque du corpus.

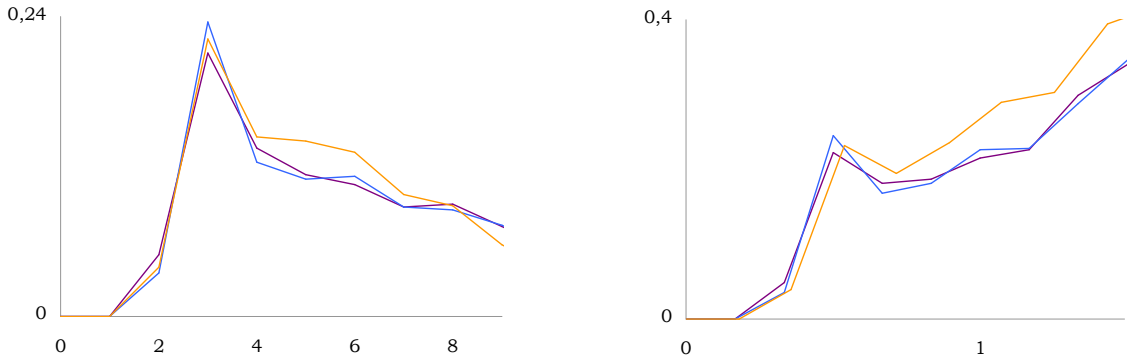


Figure 6.3 : espaces

Les spectres relatifs de droite sont plus abstraits mais aussi plus éclairants : ceux-ci croissent régulièrement depuis l'origine, selon une distribution de Rayleigh. Une surtension pointe au voisinage du mode, qui peut être prise en compte par une loi linéaire.

2.2 Ponctuation

La ponctuation forte occupe une place privilégiée et donne une indication sur la longueur de la phrase.

Seules les parenthèses ouvrantes sont comptabilisées dans le calcul des temps de retour.

Les guillemets ouvrants et fermants sont indifférenciés par la langue anglaise, inspiratrice des codes *ASCII*. Il faut donc se résoudre à des mesures quelque peu discutables, mais cohérentes au sein du corpus.

2.2.1 Moments

Les guillemets sont absents d'*Hadrien* : les moments relatifs sont donc indéfinis et fixés à zéro par convention[362]. Par ailleurs, le point d'exclamation apparaît deux fois dans *Hadrien*, de même que le point-virgule dans *Vendredi* : un seul temps de retour est donc mesuré, d'où la nullité des moments centrés.

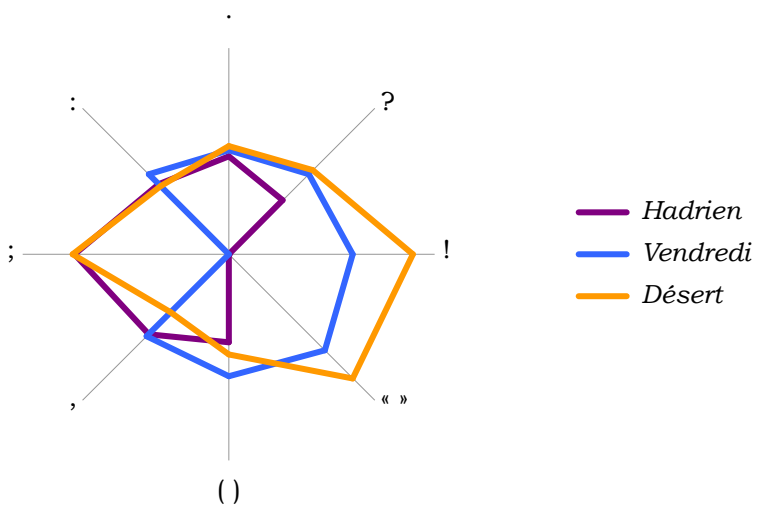


Figure 6.4 : variabilité

[362] Sa contribution dans les moments de synthèse est ainsi nulle.

Dans *Désert*, les points d'exclamation et les guillemets varient ; dans *Vendredi*, les éléments remarquables sont les parenthèses et les deux-points ; *Hadrien* est stable pour les parenthèses (fig. 6.4).

Quant à l'asymétrie (fig. 6.5), *Désert* domine pour les points, l'exclamation et les guillemets ; *Vendredi* pointe son nez sur les parenthèses ; enfin *Hadrien* se tient en retrait dans les interrogations et les parenthèses avant de se rebeller vers les points-virgules. L'asymétrie suit donc les tendances de la variabilité en les amplifiant.

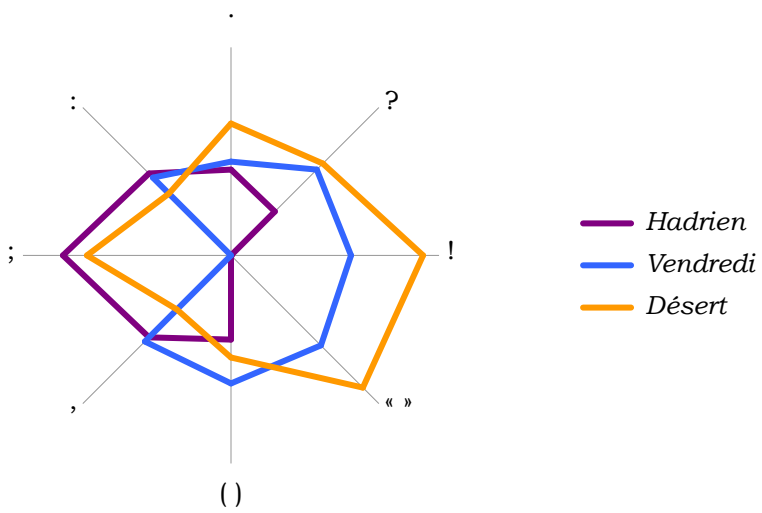


Figure 6.5 : asymétrie

2.2.2 Spectres

Le calcul de ces entités est plus pointu que celui des moments : les signes rares sont exclus pour garder les protagonistes, le point et la virgule.

Sur le spectre des points, la raie placée au premier temps est liée aux suspensions. Plus spécifique est le pic de *Vendredi*, associé au « . Log-book. ». Un mode commun se dessine vers le centième temps (fig. 6.6).

A droite, les courbes relatives comportent deux phases : la croissance linéaire jusqu'à une abscisse voisine de 1, caractéristique d'une distribution de Rayleigh ; puis la stabilisation selon une distribution exponentielle.

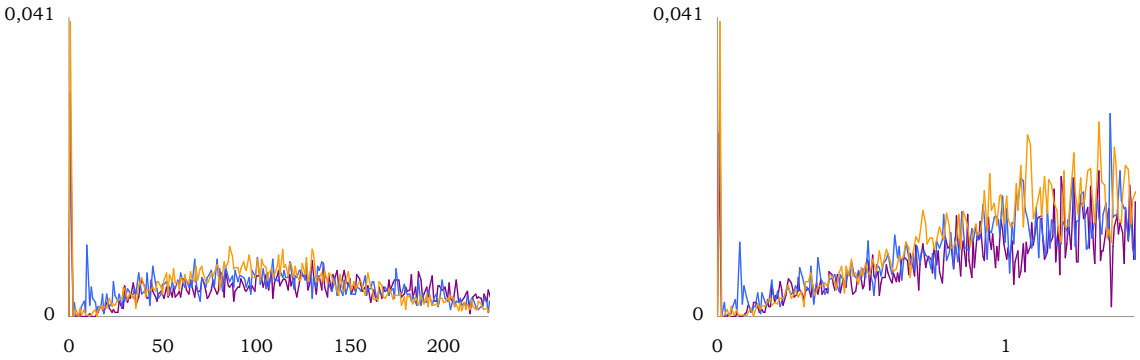


Figure 6.6 : points

Les virgules atteignent leur maximum au douzième temps avec la séquence «, comme cela, » dans *Désert* (fig. 6.7) :

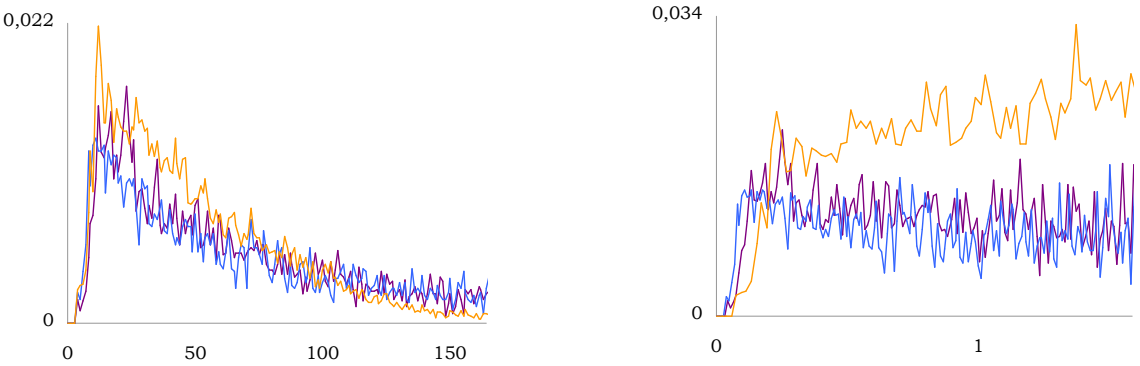


Figure 6.7 : virgules

Sur les courbes relatives, *Désert* manifeste une nouvelle atypie : si tous les spectres s'initient par une distribution de Rayleigh, ceux d'*Hadrien* et de *Vendredi* se prolongent par une phase exponentielle, tandis que celui de *Désert* poursuit une course solitaire vers la croissance.

2.3 Lettres

De même que dans la mésoscopie, les lettres minuscules et majuscules sont confondues, tandis que les caractères accentués sont ignorés.

2.3.1 Moments

Le W absent d'*Hadrien*, les moments relatifs sont fixés à zéro par convention.

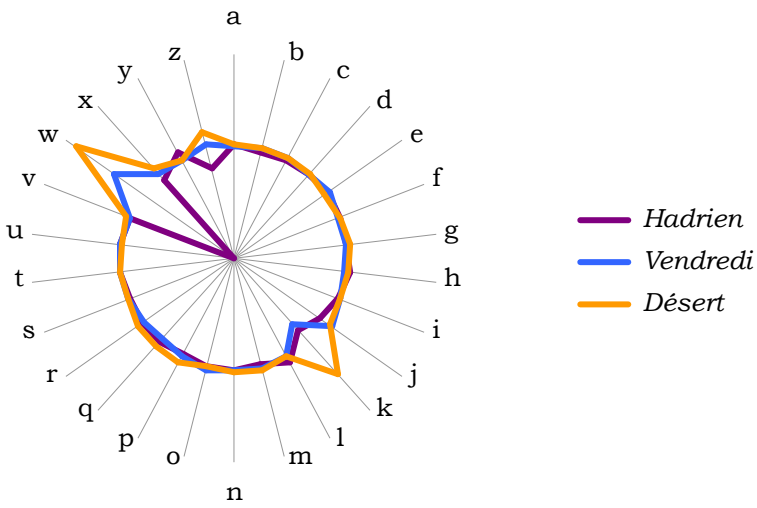


Figure 6.8 : variabilité

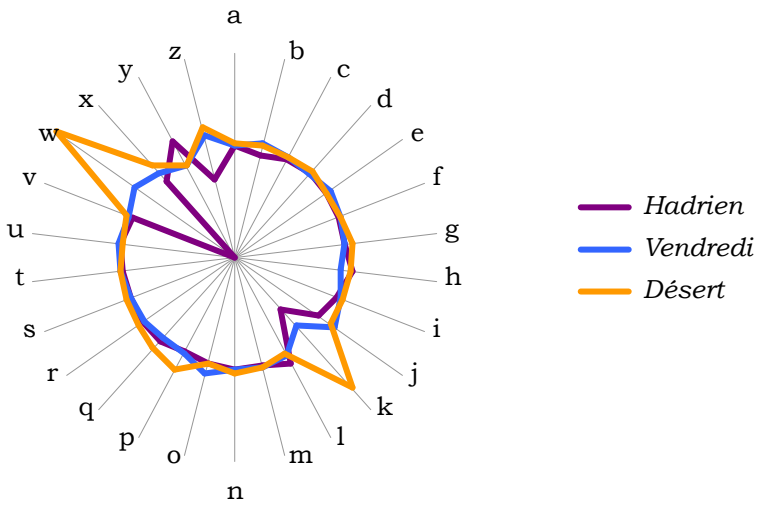


Figure 6.9 : asymétrie

Désert est plus variable pour W et K, lettres rares et hors de la norme ; *Hadrien* est au contraire plus stable pour Z (fig. 6.8).

Les résultats sont semblables sur l'asymétrie (fig. 6.9).

2.3.2 Spectres

De façon générale, les lettres courantes forment des courbes régulières, alors que d'autres plus rares engendrent des tracés erratiques. Certaines lettres comme K, W, Y ou Z ne sont pas représentées : leurs temps de retour extrêmes produisent des fichiers volumineux et peu pertinents.

Sans entrer dans l'exhaustivité, voici quelques singularités cueillies sur le défilé des courbes. Les spectres relatifs placés en regard sont modélisés après ce parcours.

Le spectre de A culmine avec le troisième temps. Le mode est particulièrement aigu pour *Désert* : la séquence « alla » est la trace de l'héroïne Lalla et du verbe « aller » (fig. 6.10) :

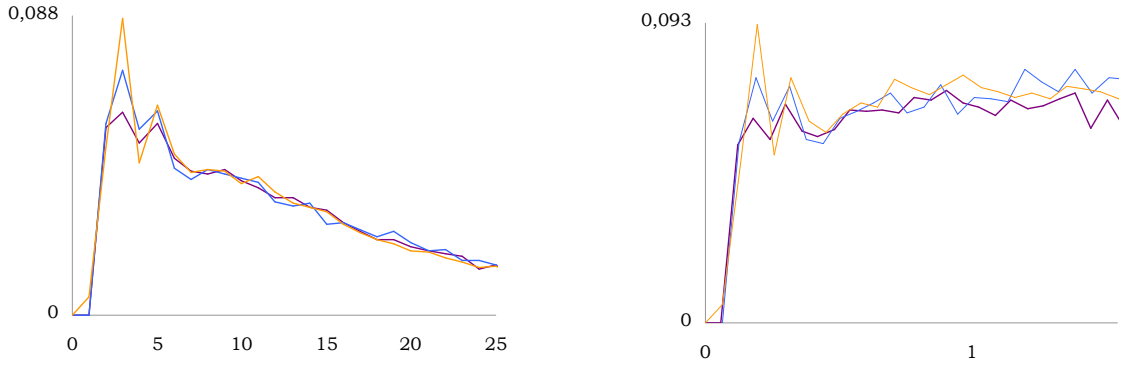


Figure 6.10 : A

Le B ressurgit en trois temps dans les « barbares » d'*Hadrien* et la « barbe » de *Vendredi* (fig. 6.11) :

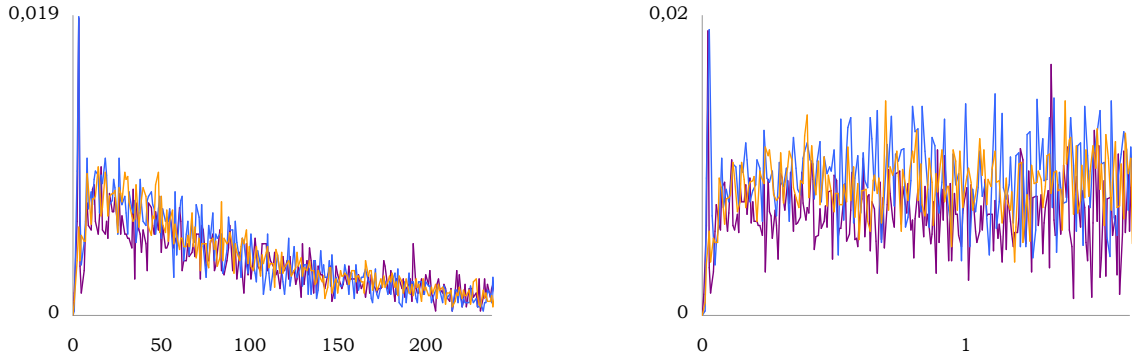


Figure 6.11 : B

Le C revient en six temps dans *Désert* par l'emploi de « comme » suivi d'un démonstratif, et celui de « commencer » (fig. 6.12) :

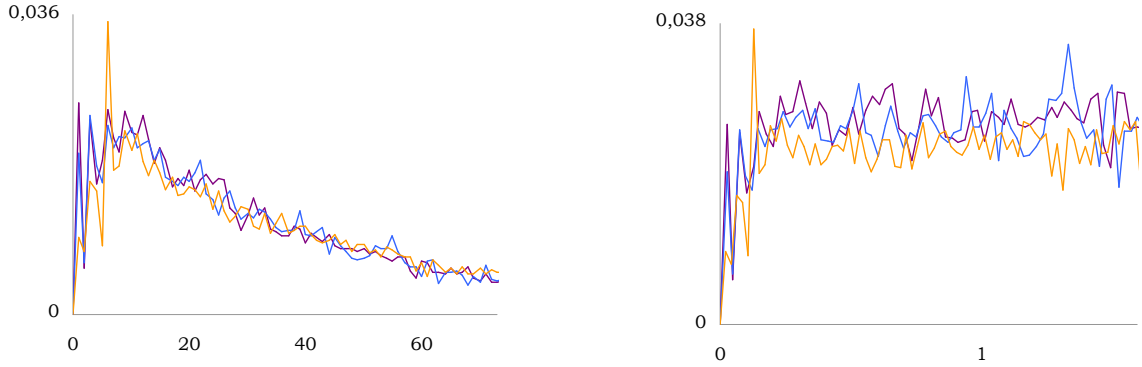


Figure 6.12 : C

Le D est relancé au bout de trois temps dans *Vendredi* : le nom de l'araucan est en cause (fig. 6.13) :

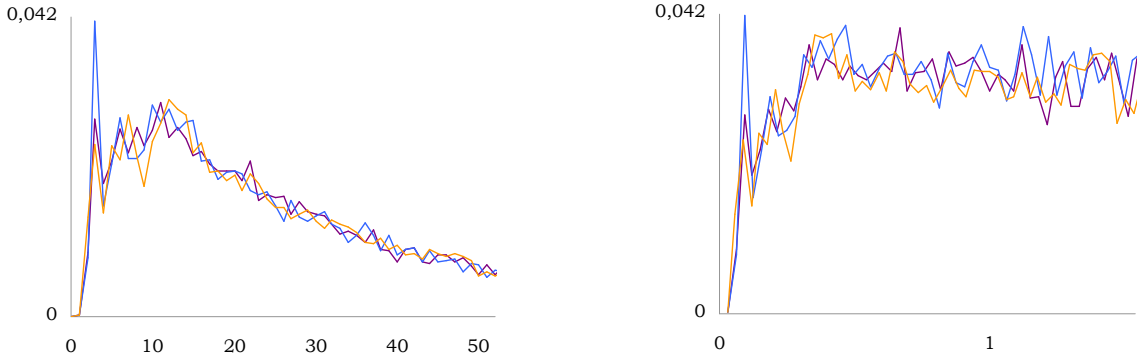


Figure 6.13 : D

Peu de différences apparaissent sur les spectres de E, graphème le

plus courant. Le mode du troisième temps s'exprime dans *Hadrien* et *Vendredi* par un « e » final enchaîné à « de » ; dans *Désert*, il correspond à « elle » (fig. 6.14) :

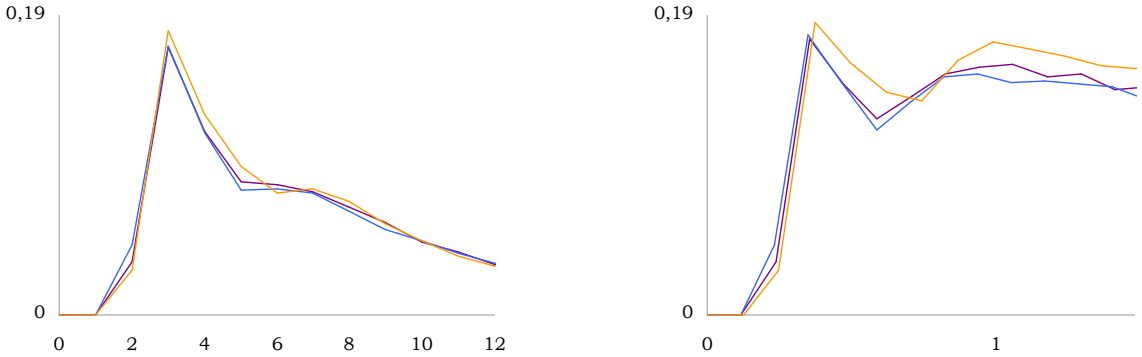


Figure 6.14 : E

Le doublement de F soulève un pic vertigineux (fig. 6.15) :

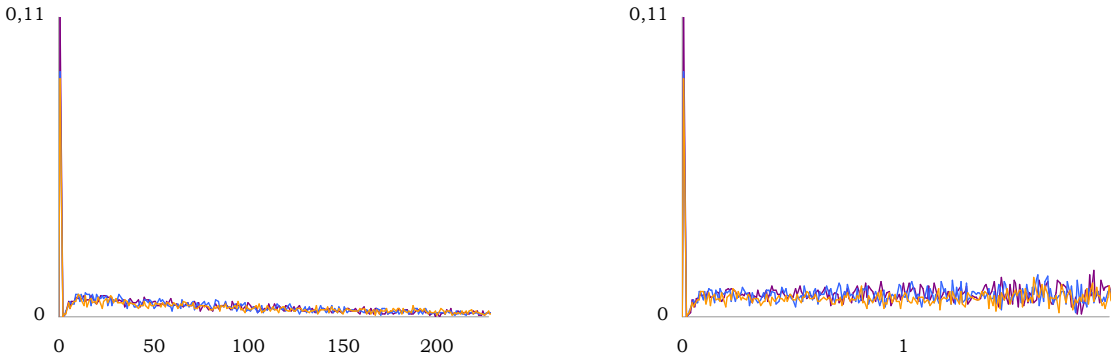


Figure 6.15 : F

Sur le spectre de G, la raie du deuxième temps résulte des emplois de « dégager », « engager », « gagner » dans *Vendredi*. Dix rangs plus loin, une autre résonance se fait entendre dans *Désert* sous les traits du « guerrier aveugle » (fig. 6.16) :

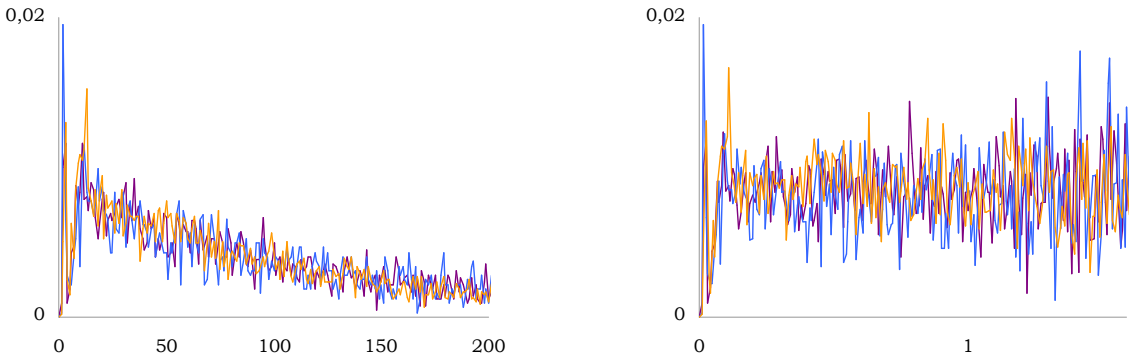


Figure 6.16 : G

Au quatrième temps, le mode de H est piquant et se manifeste dans *Désert* par « cheikh » et « chercher » (fig. 6.17) :

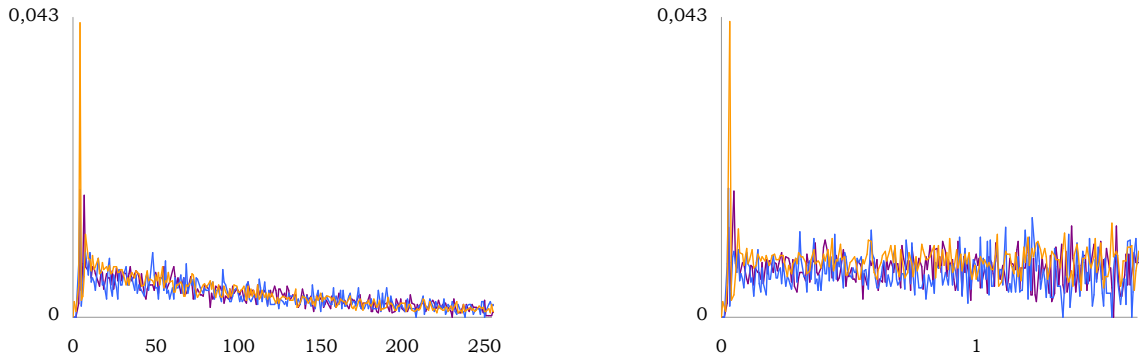


Figure 6.17 : H

Le sommet principal de I se place au septième temps : dans *Hadrien* et *Vendredi*, avec « première fois » et « dernière fois », dans *Désert* avec « ils étaient ». *Hadrien* cultive l'« ici » et soulève un pic primitif au deuxième rang (fig. 6.18) :

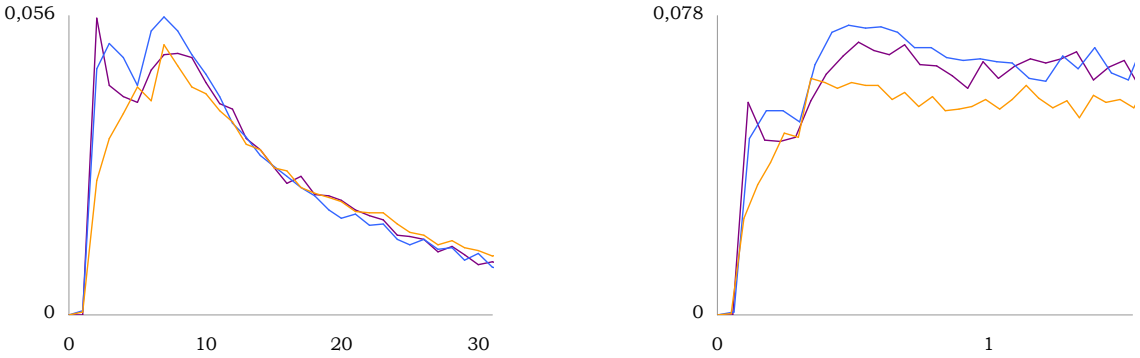


Figure 6.18 : I

Perceptible au milieu du bruit, le cri de J au treizième temps signe le refrain hynotique de *Désert* : « un jour, oh, un jour, … » (fig. 6.19) :

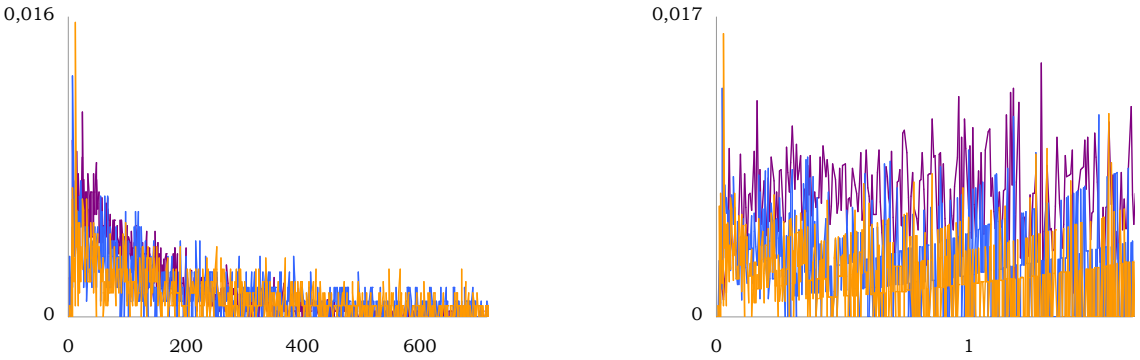


Figure 6.19 : J

Le doublement de L est marqué dans *Désert* par les références à l'héroïne, « Lalla » et « elle » (fig. 6.20) :

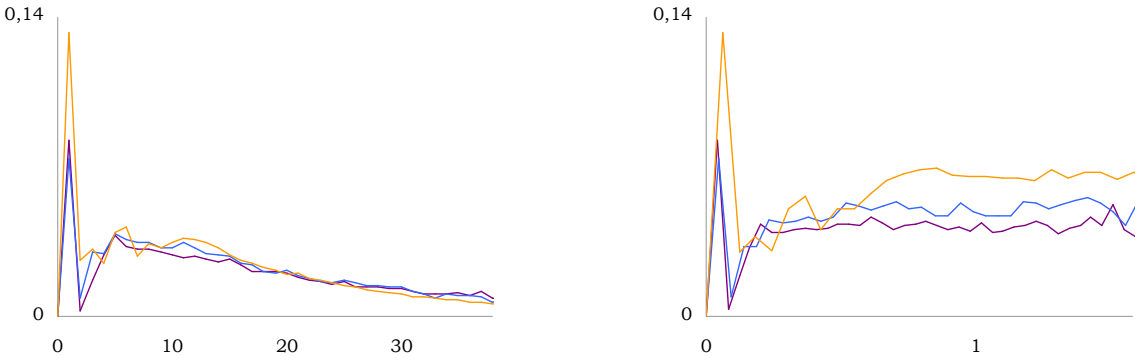


Figure 6.20 : L

Le doublement de M est lui aussi accru dans *Désert* et se manifeste par l'emploi de « comme » (fig. 6.21) :

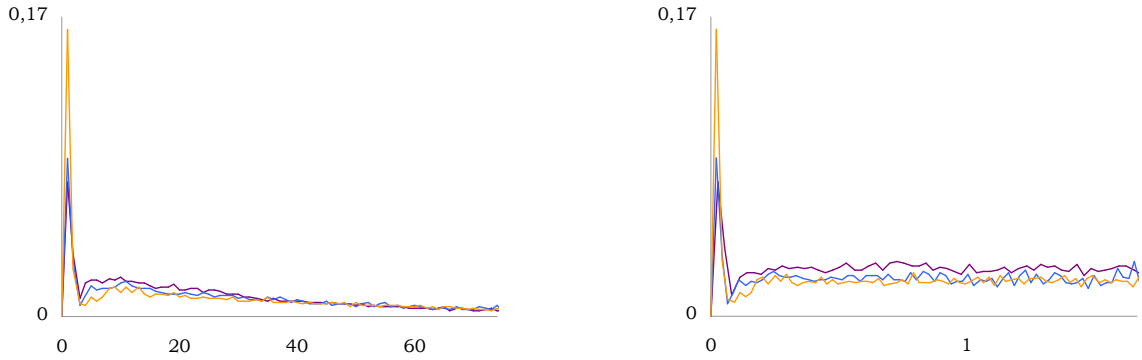


Figure 6.21 : M

Les spectres de N sont homogènes : le premier pic correspond au doublement de la lettre. Au troisième temps, le mode de *Vendredi* est lié à « Robinson » (fig. 6.22) :

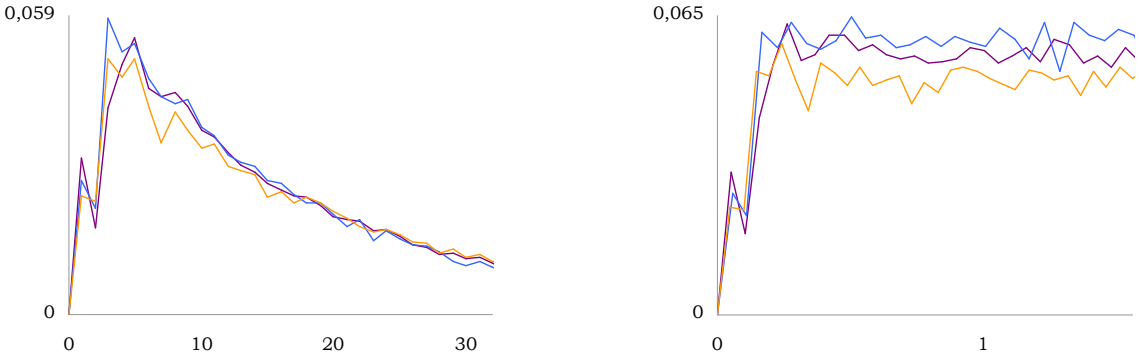


Figure 6.22 : N

L'éminence de O sur le cinquième temps est encore le fait de « Robinson » (fig. 6.23) :

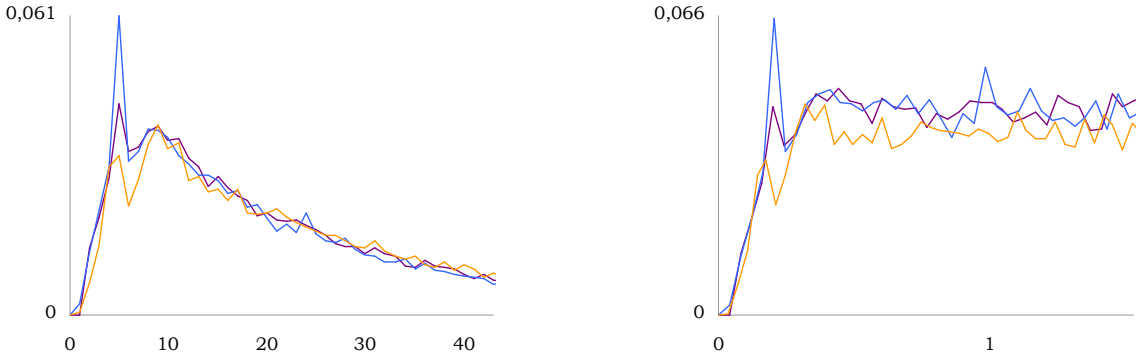


Figure 6.23 : O

La répétition de P est banale (fig. 6.24) :

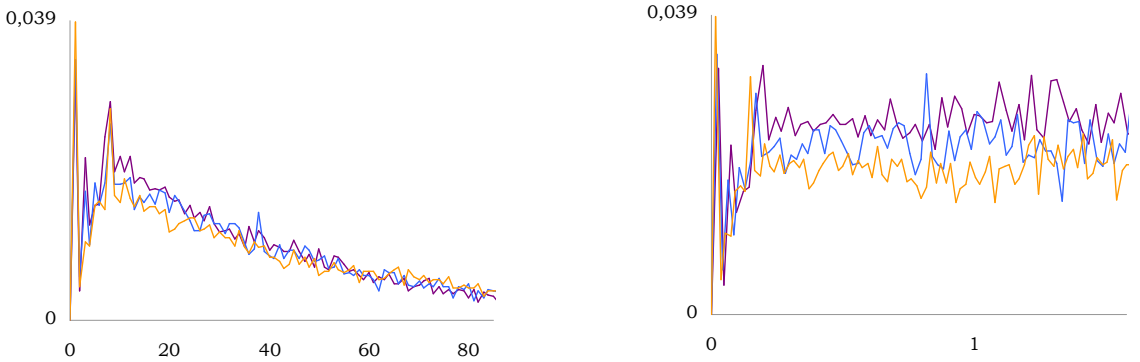


Figure 6.24 : P

Q surgit au quatrième temps avec « quelque » (fig. 6.25) :

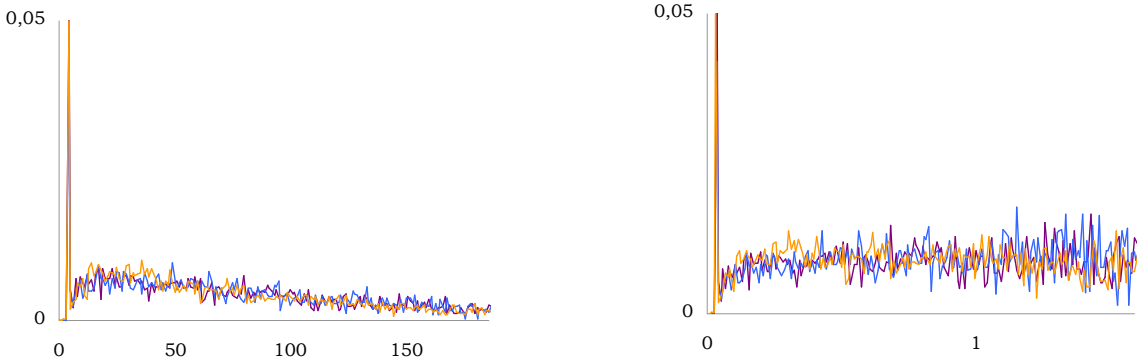


Figure 6.25 : Q

Le doublement de R est sensible dans *Désert*, en relation avec la

« terre » et la « pierre ». Au quatrième temps, le mode est l'effet du « regard » (fig. 6.26) :

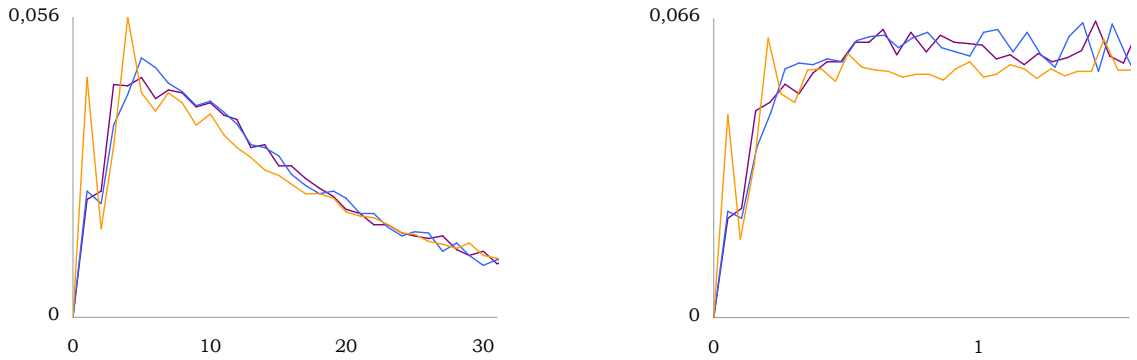


Figure 6.26 : R

Les premiers sommets de S sont soulevés par le doublement de cette lettre. Le mode principal se présente au septième temps : à noter dans *Désert* les références aux « hommes » précédées d'un déterminant pluriel (fig. 6.27) :

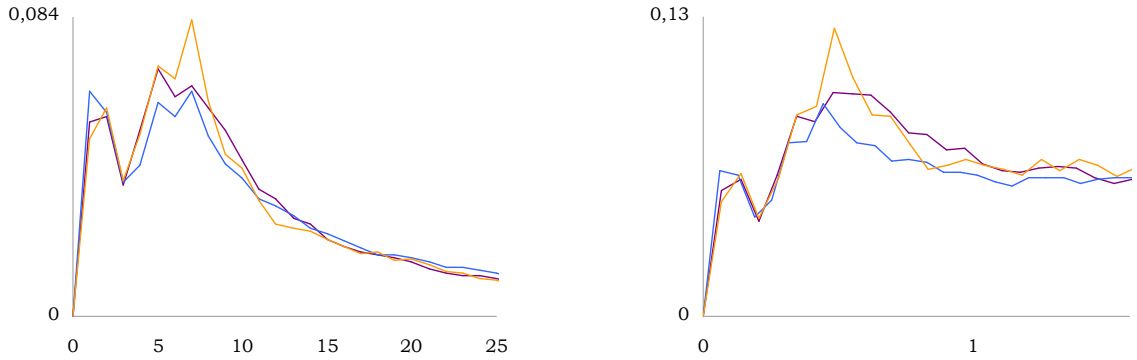


Figure 6.27 : S

Le mode de T est atteint au troisième temps : dans *Hadrien* et *Vendredi* par la marque de l'imparfait « tait », dans *Désert* par l'emploi de « tout » (fig. 6.28) :

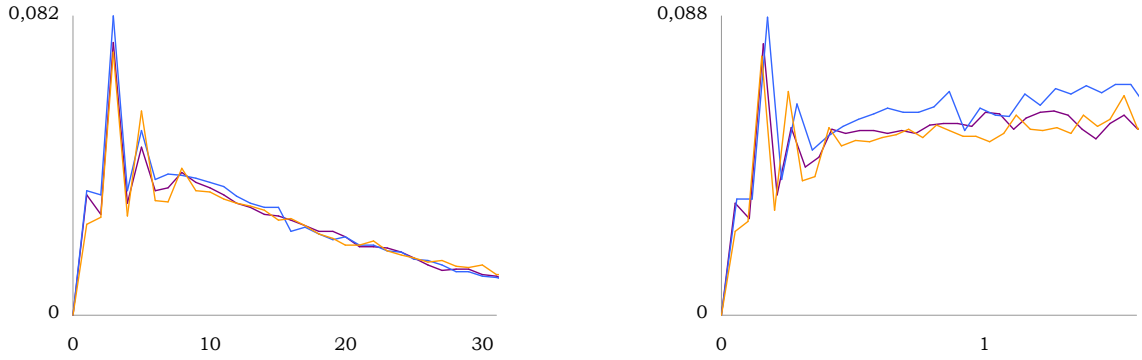


Figure 6.28 : T

Le U se fait entendre au quatrième temps avec « quelque » (fig. 6.29) :

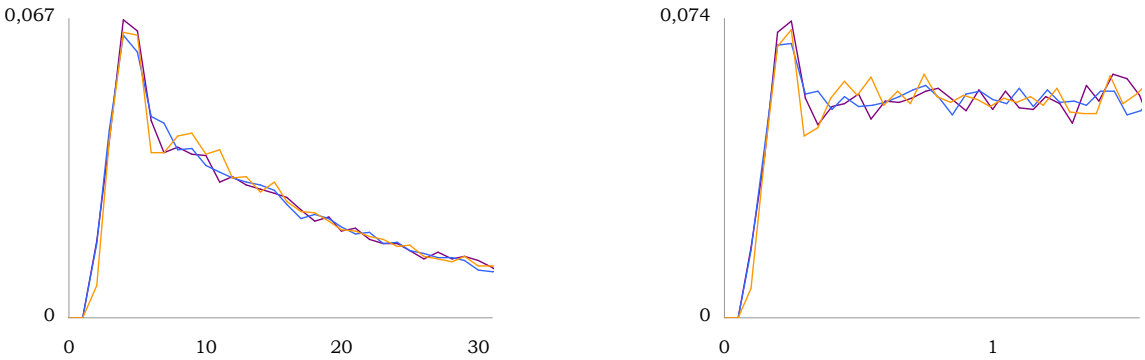


Figure 6.29 : U

Sur le spectre de V, la raie du second temps est le fruit de « vivre ». Au huitième rang, le sommet de *Désert* est partagé entre « vers la ville » et « vers la vallée » (fig. 6.30) :

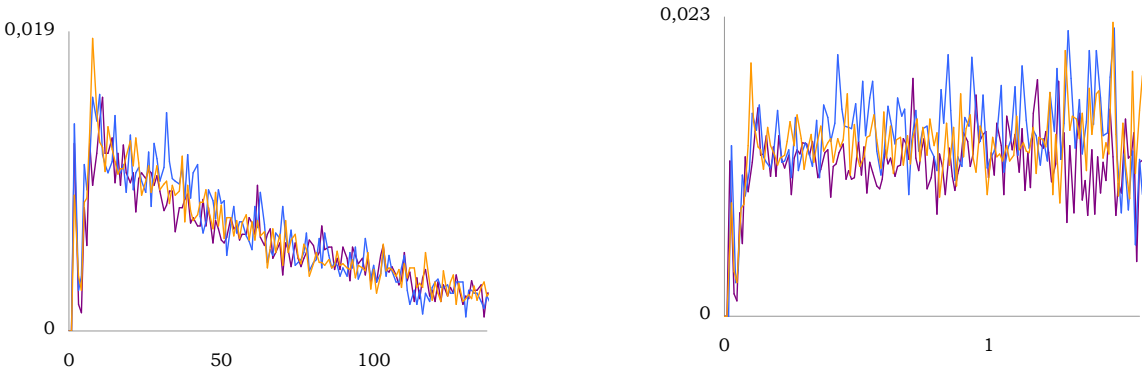


Figure 6.30 : V

Sur X, une raie traverse le cinquième temps de *Désert* par l'expression « aux yeux » (fig. 6.31) :

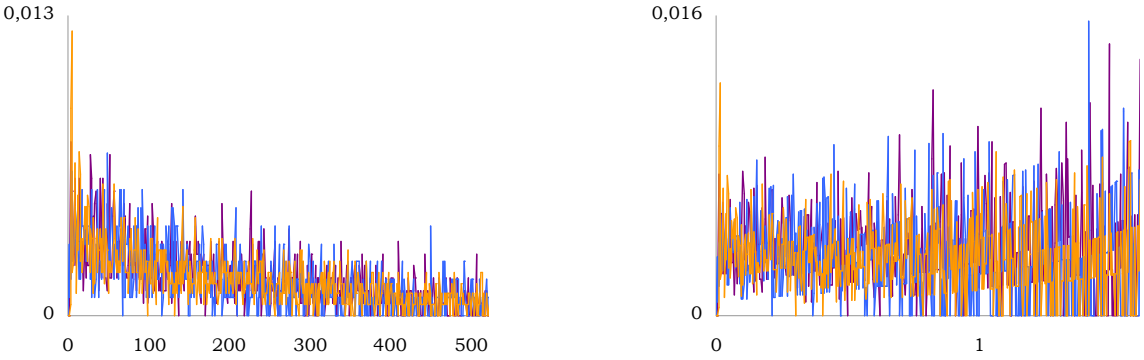


Figure 6.31 : X

Plus linéaires, les spectres relatifs suivent néanmoins les anomalies de leurs frères absolus. D'un point de vue plus linguistique, ébauchons un modèle :

– les courbes commencent par une courte phase de Rayleigh et s'achèvent par une longue période exponentielle ;
– la première phase est parfois perturbée par le doublement de la lettre comme pour F, L, M, P, R et S, ou plus tardivement pour G, H, Q, T et V ;
– une surtension s'observe aux modes des lettres E, S et U.

Ces deux derniers phénomènes peuvent être pris en compte par l'insertion de distributions linéaires entre les lois initiales et finales.

2.4 Synthèse graphémologique

2.4.1 Moments

Les différentes unités sont intégrées à l'aide des algorithmes du chapitre 2.

Figure 6.33 : variabilité

Figure 6.34 : asymétrie

La variabilité ne permet pas de discerner les œuvres (fig. 6.33). En revanche, l'asymétrie détache *Désert* du corpus (fig. 6.34).

2.4.2 Spectres

Les espaces se comportent différemment des lettres : dans le premier cas, le modèle se réduit à une loi de Rayleigh. En revanche, les lettres combinent une brève distribution de Rayleigh et une longue période exponentielle. La ponctuation se situe à mi-chemin : elle commence par une phase de Rayleigh et se termine par une loi linéaire.

Ces observations se rapprochent des conclusions de Mandelbrot :

> In other terms, the properties of the recurrence of « space » seem more intrinsic in some ways than those of the recurrence of any proper letter[363].

Dans son article, il modélise un discours comme une séquence de lettres aléatoires, coupées en mots par une lettre impropre, l'espace.

Nos mesures révèlent un aléa plus marqué pour les espaces : avec une probabilité relative constante, la distribution exponentielle oublie le temps qui passe. A l'inverse, la distribution Rayleigh garde en mémoire la date de la dernière occurrence pour évaluer la probabilité de la suivante. Sans âge, la première reste libre tandis que la seconde vieillit inexorablement.

Quantitativement, un spectre sur la population entière des unités est bâti à partir des temps de retour relatifs x/m (fig. 6.35). La largeur

[363] « En d'autres termes, les propriétés de la récurrence d'un espace semblent plus intrinsèques par certains côtés que celles de la récurrence d'une lettre proprement dite » (Jakobson & alii, *Structure of Langage and Its Mathematical Aspects*, « Word frequencies and Markovian Models of discourse », p. 205).

des classes — le paramètre h évoqué dans le chapitre 4 — est fixé à 0.2, afin de concilier précision et régularité. Avec des effectifs plus nombreux que dans le cas des unités isolées, les statistiques restent significatives sur 99 % de la population.

Désert se singularise visiblement sur les courbes de gauche. A droite apparaît une phase de Rayleigh, puis une loi linéaire décroissante qui se stabilise en exponentielle. Fait remarquable, la distribution synthétique suit la même logique que celle des unités : le mélange des populations se justifie donc a posteriori.

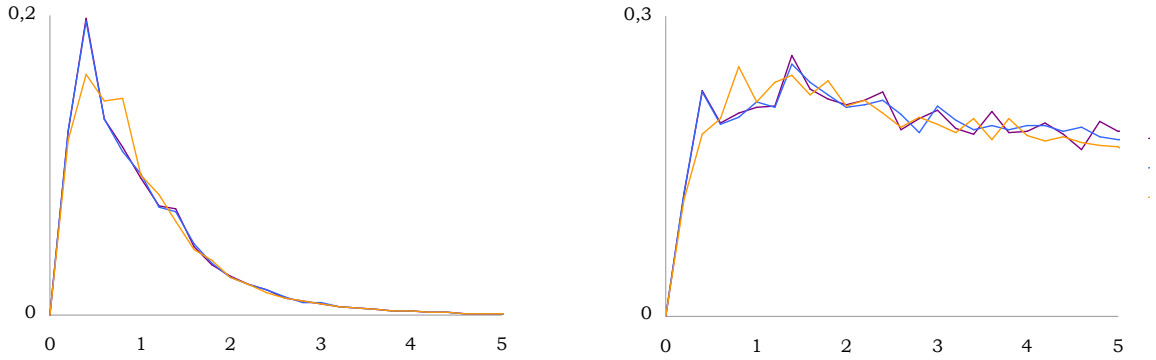


Figure 6.35 : spectres

3 Syntaxe

Comme dans la mésoscopie, l'étude porte uniquement sur les parties du discours. Les catégories typographiques (Typo) et éliminatoires (Elim) de *Syntex* ne sont pas analysées, mais elles sont comptabilisées dans les temps de retour afin de respecter la topologie du texte.

3.1 Parties du discours

3.1.1 Moments

Désert est généralement le plus instable, mais *Hadrien* prend le

dessus pour les conjonctions (figure 6.36) :

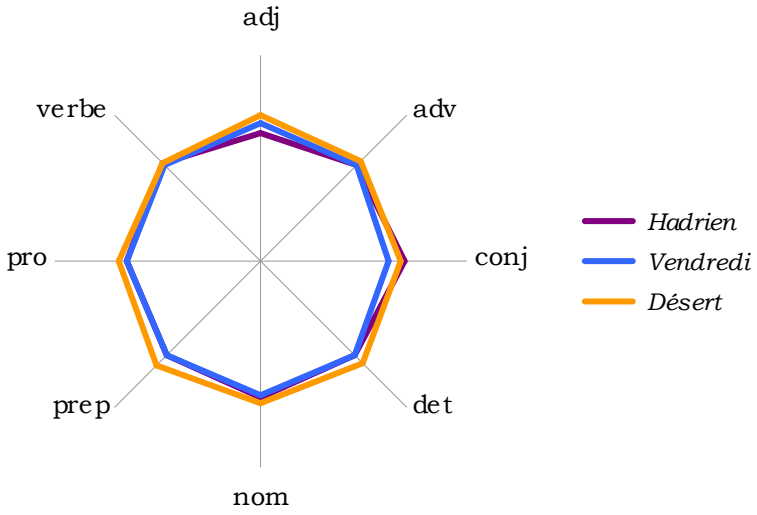


Figure 6.36 : variabilité

En ce qui concerne l'asymétrie, *Désert* prend un net ascendant mais *Hadrien* le menace ponctuellement sur les conjonctions (fig. 6.37) :

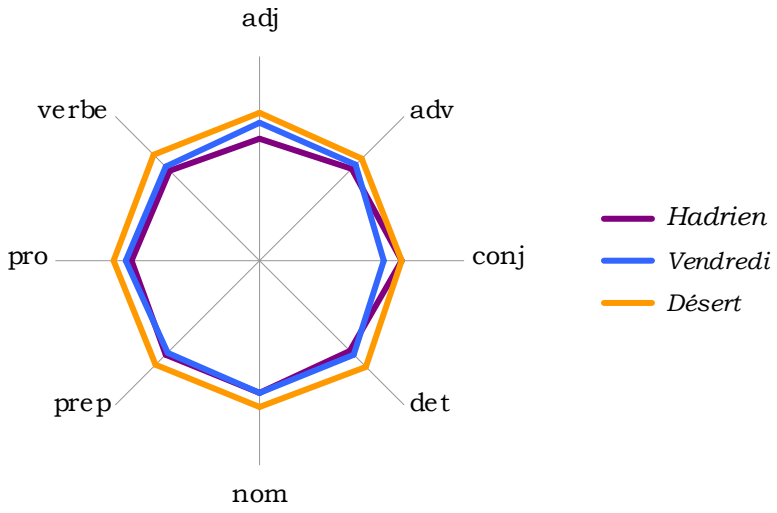


Figure 6.37 : asymétrie

3.1.2 Spectres

Parcourons les courbes pour cerner les points particuliers.

Les adjectifs culminent au deuxième temps dans *Hadrien*, portés par la séquence « adjectif nom adjectif ». Un rang plus loin, les modes de

Vendredi et *Désert* traduisent la séquence « adjectif typographie adverbe adjectif » (fig. 6.38) :

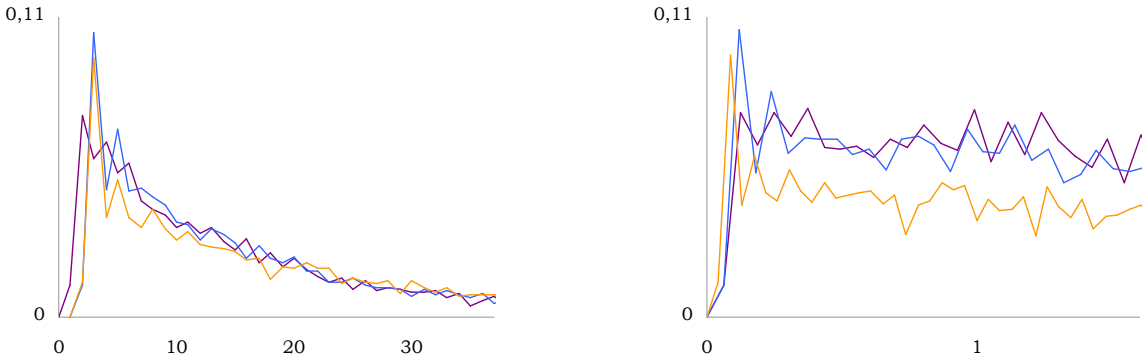


Figure 6.38 : adjectifs

Les adverbes résonnent en deux temps avec la séquence « adverbe verbe adverbe », très courante dans *Désert* (fig. 6.39) :

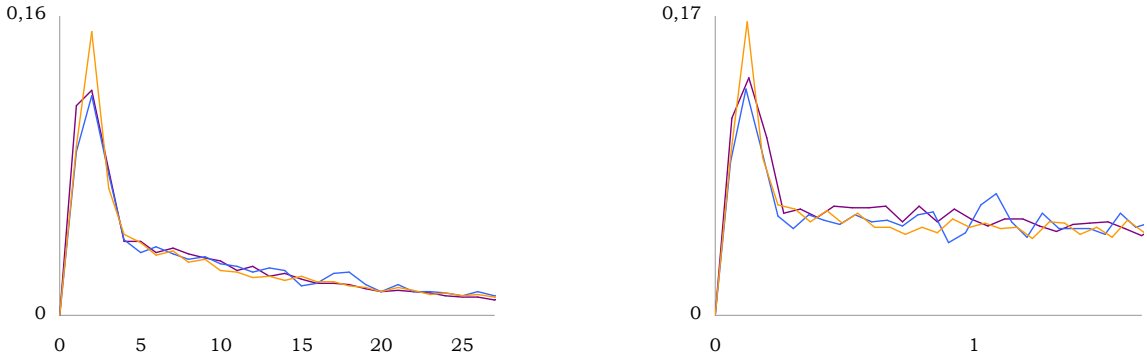


Figure 6.39 : adverbes

Le doublement des conjonctions engendre une raie primitive et commune. *Hadrien* connaît son apogée au sixième temps, puis *Vendredi* et *Désert* au rang suivant (fig. 6.40) :

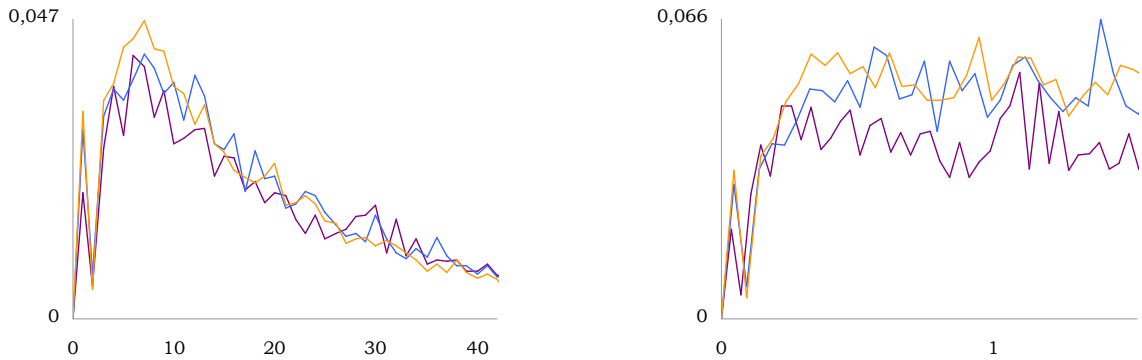


Figure 6.40 : conjonctions

Les courbes des déterminants sont semblables entre les œuvres. Le mode situé au troisième temps est lié à la séquence « déterminant nom préposition déterminant » (fig. 6.41) :

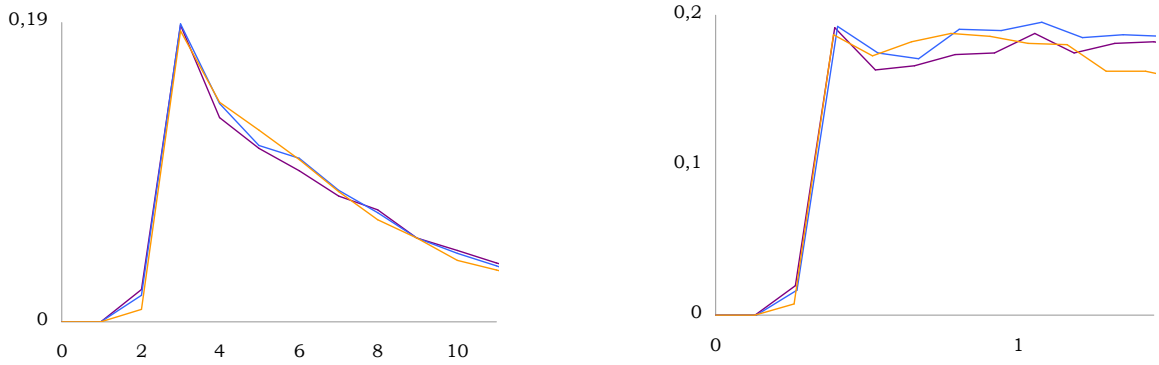


Figure 6.41 : déterminants

De même, les spectres des noms sont homogènes. Placé au troisième temps, le sommet s'incarne dans la séquence « nom préposition déterminant nom » (fig. 6.42) :

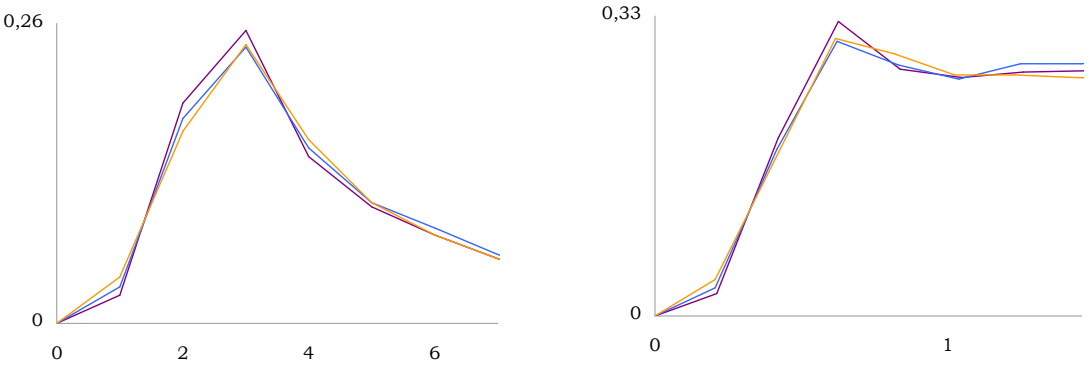


Figure 6.42 : noms

Les spectres des prépositions sont encore proches. Le mode du troisième temps est le fait de la séquence « préposition déterminant nom préposition » (fig. 6.43) :

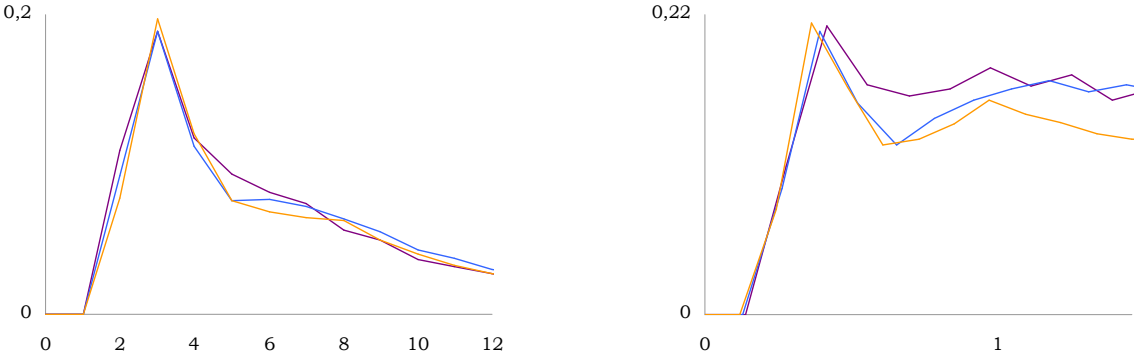


Figure 6.43 : prépositions

Le pic primaire des pronoms signe leur doublement, avant un rebond peu sensible au troisième temps (fig. 6.44) :

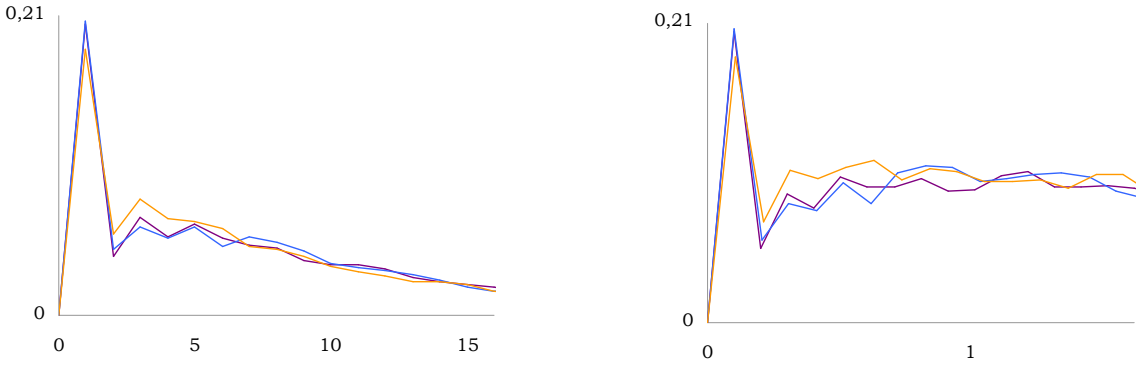


Figure 6.44 : pronoms

Les verbes voient un pic analogue, suivi d'un mamelon secondaire : au cinquième temps pour *Hadrien* et *Désert* avec la séquence « verbe adverbe nom typographie pronom verbe » ; un rang plus loin pour *Vendredi* avec la séquence « verbe préposition déterminant nom typographie pronom verbe » (fig. 6.45) :

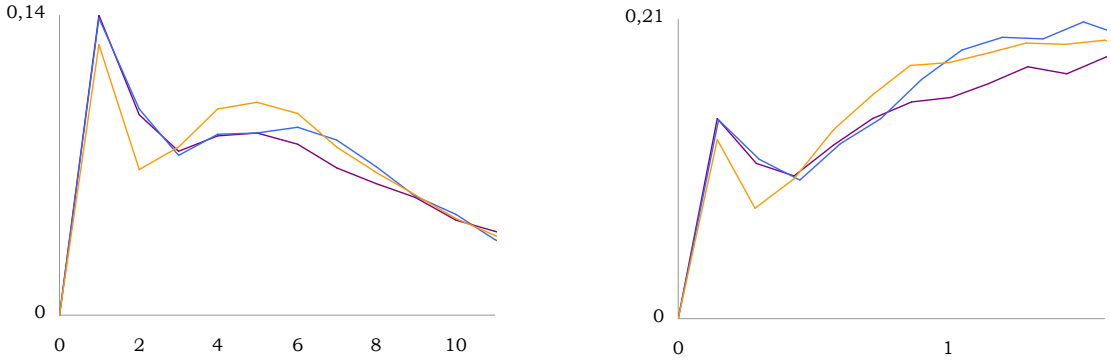


Figure 6.45 : verbes

Le spectres relatifs présentent le gabarit suivant :

– ils commencent par une phase de Rayleigh pour s'achever par une période exponentielle ;
– la phase initiale est perturbée par un pic dans le cas des conjonctions ou des verbes ;
– une surtension s'observe sur les modes des adjectifs, adverbes, prépositions et pronoms.

3.2 Synthèse syntaxique

3.2.1 Moments

Globalement, *Désert* se montre le plus variable, suivi d'*Hadrien* et de *Vendredi* (fig. 6.46) :

Figure 6.46 : variabilité

Les résultats sont analogues pour l'aymétrie avec des écarts plus marqués (fig. 6.47) :

Figure 6.47 : asymétrie

3.2.2 Spectres

La distribution générale rassemble les unités (fig. 6.48). L'intervalle entre les classes est fixé à 0.2 comme pour la graphémologie.

L'asymétrie positive se manifeste sur les courbes absolues, orientées vers la gauche. Les courbes relatives montrent une loi de Rayleigh suivie d'une loi linéaire décroissante.

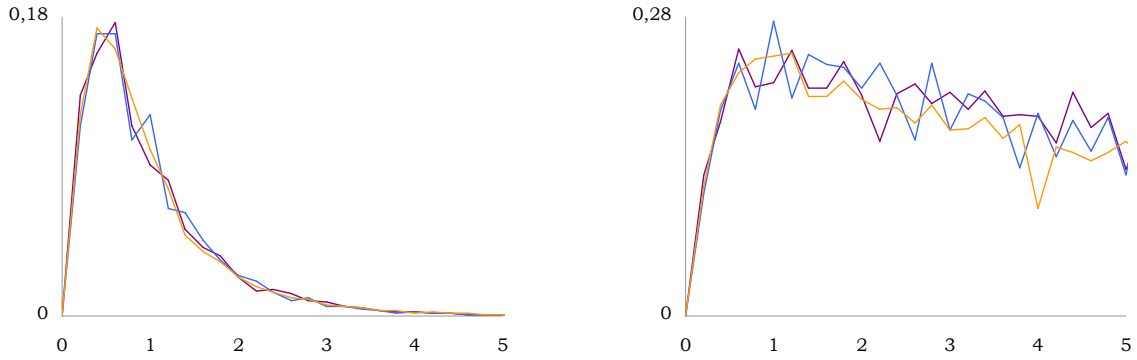


Figure 6.48 : spectres

4 Sémantique

La grille de lecture est la même que celle de la macroscopie et de la mésoscopie : l'univers sémantique est découpé en vingt-huit concepts.

Cordial n'attribue pas certains termes du texte, tandis qu'il donne plusieurs étiquettes à d'autres éléments : la notion de séquence perd donc son sens et n'est pas étudiée à ce niveau.

L'annexe 6 fournit pour chaque œuvre les statistiques de base et les modes des spectres.

4.1 Concepts

4.1.1 Moments

La figure 6.49 analyse la variabilité :

- *Désert* se montre le moins stable, notamment pour la vie sociale, l'information, la vie spirituelle, la hiérarchie, la guerre et la paix, le quotidien, le droit, la matière et la vie ;
- *Vendredi* pointe légèrement avec le corps ;
- *Hadrien* reste le plus stable dans le cas de l'être humain, la

communication, l'économie, la vie et le corps.

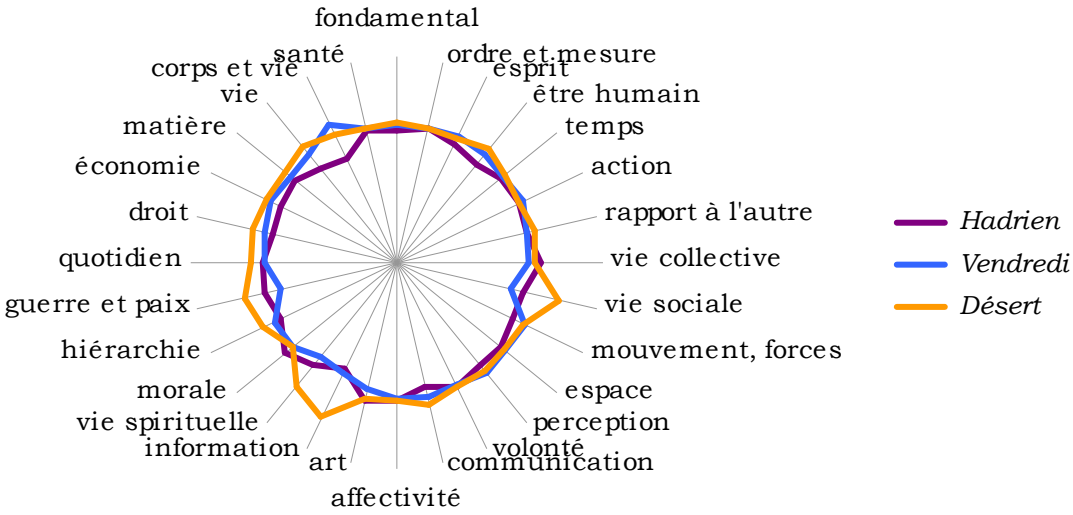


Figure 6.49 : variabilité

L'asymétrie suit le même schéma en l'amplifiant (fig. 6.50) :

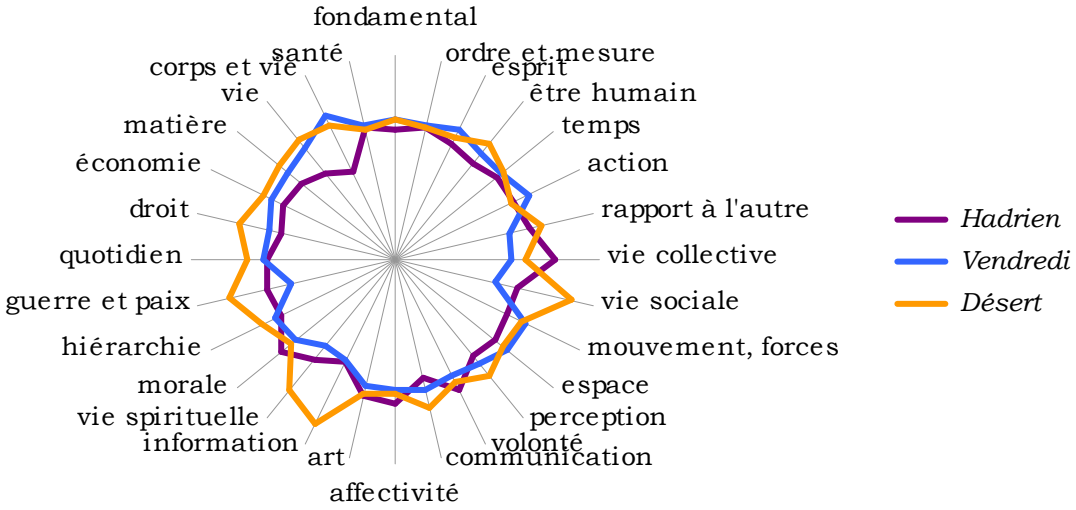


Figure 6.50 : asymétrie

4.1.2 Spectres

L'analyse limitée par l'absence de séquences, seuls les concepts les plus fréquents sont représentés. Les autres spectres sont placés en

annexe 6.

Après le dénuement des origines, le fondamental trouve son apogée en un temps ; *Désert* se détache des autres romans par un sommet moins marqué (fig. 6.51) :

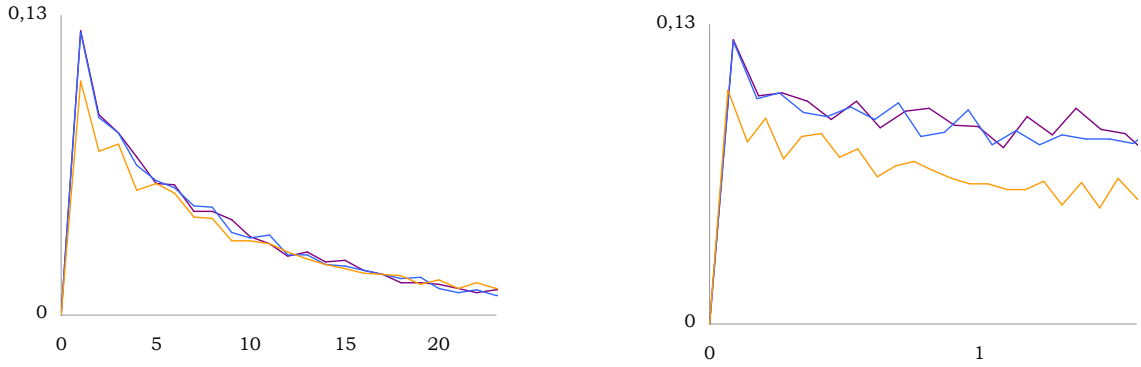


Figure 6.51 : fondamental

L'ordre et la mesure suivent la même logique (fig. 5.52) :

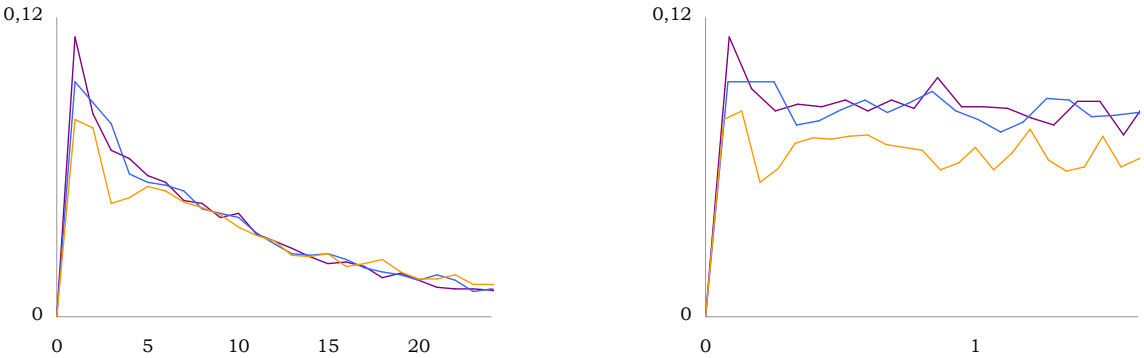


Figure 6.52 : ordre et mesure

Les modes de l'esprit naissent au premier temps, tandis qu'un second élan affleure au troisième rang ; les courbes s'étagent selon la gradation habituelle (fig. 6.53) :

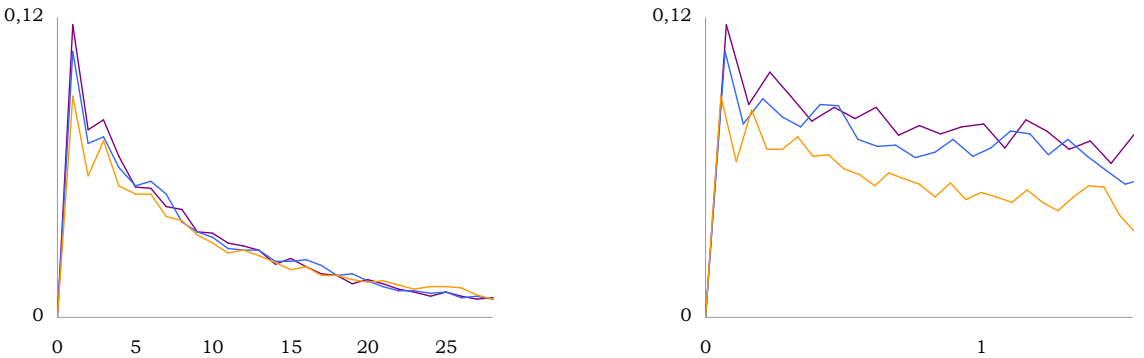


Figure 6.53 : esprit

Les modes du temps s'échelonnent : 1 dans *Vendredi* , 2 dans *Hadrien* et 3 dans *Désert* (fig. 6.54) :

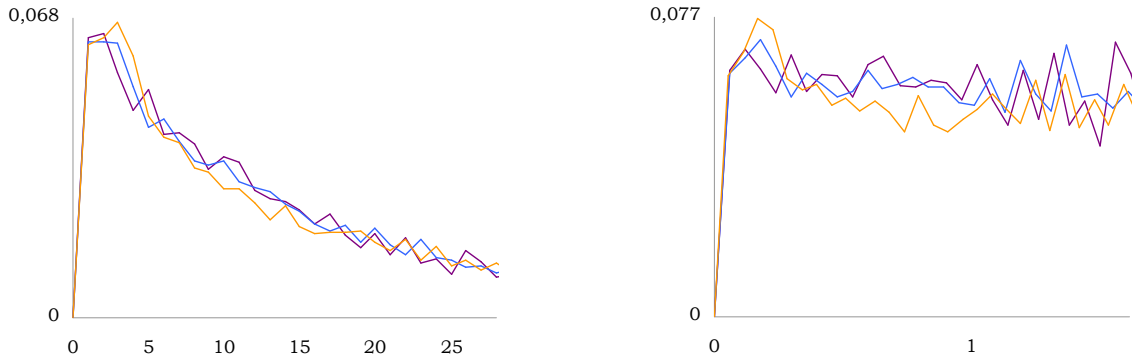


Figure 6.54 : temps

Désert culmine au deuxième temps et *Vendredi* au suivant ; au second rang, l'apogée d'*Hadrien* reste plus modeste (fig. 6.55) :

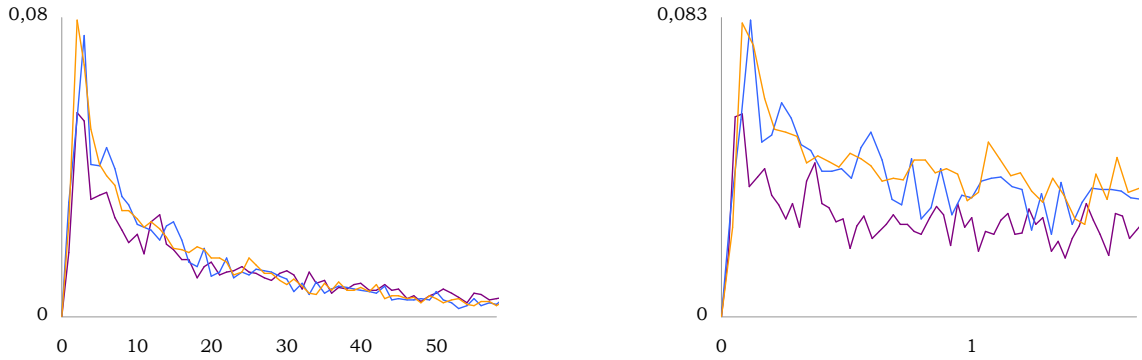


Figure 6.55 : espace

La matière structure ses modes entre les deuxièmes et troisièmes temps, faisant succéder *Hadrien*, *Vendredi* et *Désert* dans un bel ordonnancement (fig. 6.56) :

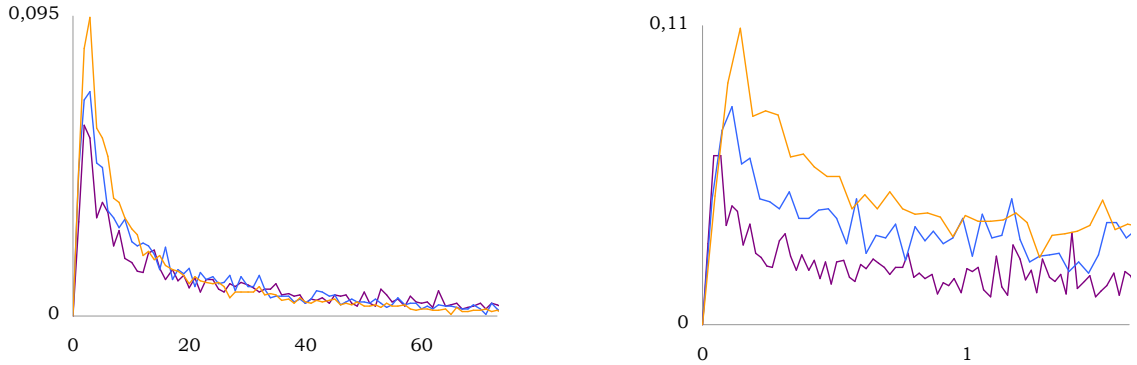


Figure 6.56 : matière

Les spectres relatifs reproduisent le schéma des niveaux précédents. Des surtensions accompagnent certains modes, et des phases linéaires décroissantes précèdent une stabilisation exponentielle.

4.2 Synthèse sémantique

4.2.1 Moments

Dans l'ensemble, la gradation entre les trois romans réapparaît, plus visible dans le cas de l'asymétrie que de la variabilité (fig. 6.57 et 6.58).

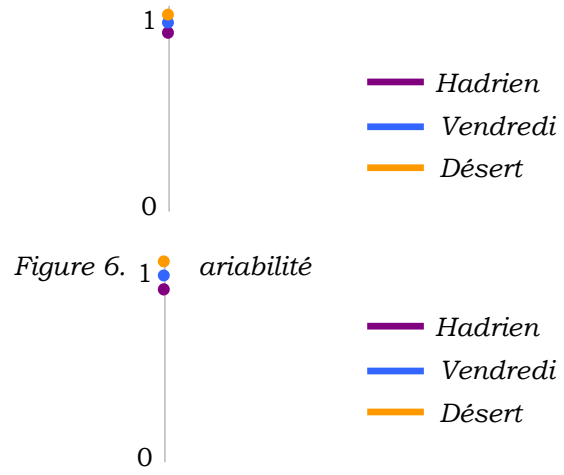


Figure 6. *ariabilité*

Figure 6.58 : asymétrie

4.2.2 Spectres

Les différentes distributions sont intégrées à l'aide des temps de retour relatifs (fig. 6.59) :

La gradation entres les trois œuvres est visible, notamment sur les modes. Les spectres relatifs commencent avec une phase de Rayleigh, continuent par une loi linéaire décroissante, pour finir sur une exponentielle.

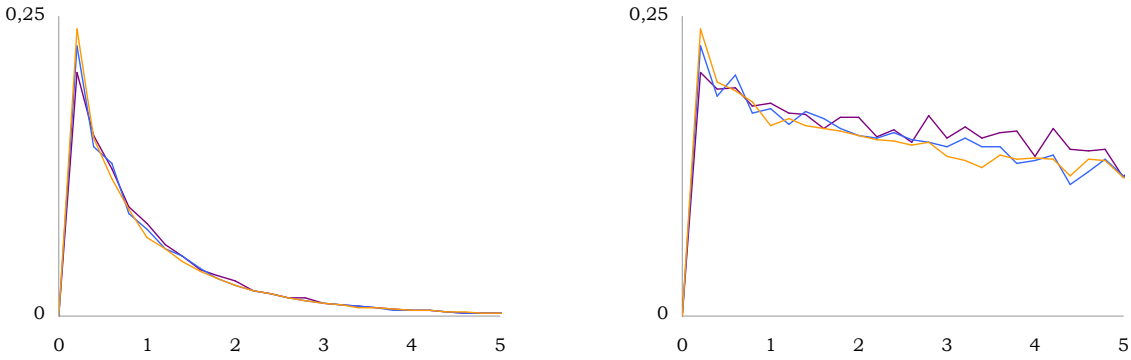


Figure 6.59 : spectres

5 Synthèse microscopique

5.1 Moments

Les mesures de ce chapitre sont rassemblées par la moyenne des chiffres obtenus sur les trois plans linguistiques.

La variabilité confond presque les trois œuvres, toutefois *Désert* semble légèrement plus irrégulier (fig. 6.60) :

Figure 6.60 : variabilité

L'asymétrie montre la même tendance en l'amplifiant (fig. 6.61) :

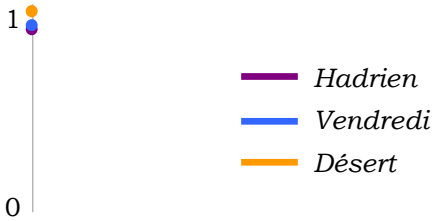


Figure 6.61 : asymétrie

D'un point de vue plus général, il est intéressant de comparer la variabilité et l'asymétrie des trois plans linguistiques. Si l'évaluation se fonde sur un corpus limité, elle n'en donne pas moins quelques indications.

Le niveau sémantique varie le plus, devant le syntaxique et le graphémologique (fig. 6.62) :

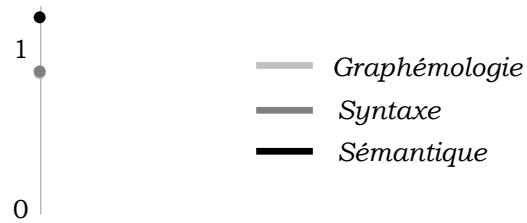


Figure 6.62 : variabilité

La figure 6.63 présente des résultats analogues pour l'asymétrie :

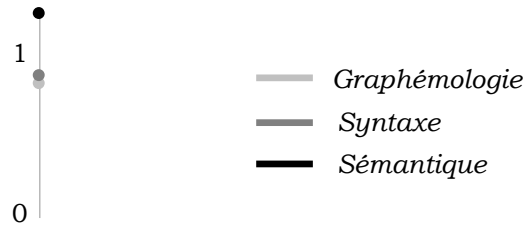


Figure 6.63 : asymétrie

Ces résultats doivent être pris avec précaution, en raison des particularités du niveau sémantique : le nombre d'intervalles analysés est plus faible, et le repérage des concepts est plus complexe.

De façon générale, l'asymétrie est plus discriminante que la variabilité : elle distingue les valeurs positives des négatives, et amplifie les écarts importants.

5.2 Spectres

D'un niveau linguistique à l'autre, les courbes sont plus ou moins régulières, en fonction de la richesse des effectifs de chaque classe. Deux paramètres interviennent principalement :

– le nombre total d'intervalles pour une unité ;
– la variété des temps de retour, qui détermine le nombre de classes d'accueil.

Cette dernière grandeur se mesure par les écarts-types, qui diffèrent de la variabilité par le jeu des moyennes. Le tableau qui suit résume les caractéristiques d'ensemble de chaque plan :

	Intervalles par unité	Ecarts-types
Graphémologie	14726	386
Syntaxe	11645	10
Sémantique	2672	88

Figure 6.64 : caractéristiques des spectres

Les courbes de la syntaxe sont en général plus sereines que celles de la graphémologie. D'une part, le nombre d'intervalles reste élevé : il y a moins de mots que de lettres, mais la distribution en huit parties est plus généreuse que celle en trente cinq graphèmes : les drames locaux sont ainsi prévenus. D'autre part, les temps de retour sont peu variés : la population des intervalles se concentre sur une aire limitée, d'où des classes fournies et des histogrammes lisses.

Relativement au niveau syntaxique — a fortiori graphémologique — les populations sémantiques sont faibles : certains mots ne sont pas étiquetés, tandis qu'un découpage en vingt-huit cicrconscriptions affaiblit le recensement. De plus, les intervalles entres les unités sont

éventés vers de larges plages, même si la graphémologie se disperse encore plus. Finalement, les courbes ont une qualité comparable à celles de la graphémologie.

Quantitativement, le spectre suivant rassemble les trois niveaux linguistiques : il suffit de moyenner les distributions absolues[364] et d'en déduire les courbes relatives. Son sens physique est limité, mais il donne une image générale des courbes rencontrées (fig. 6.65) :

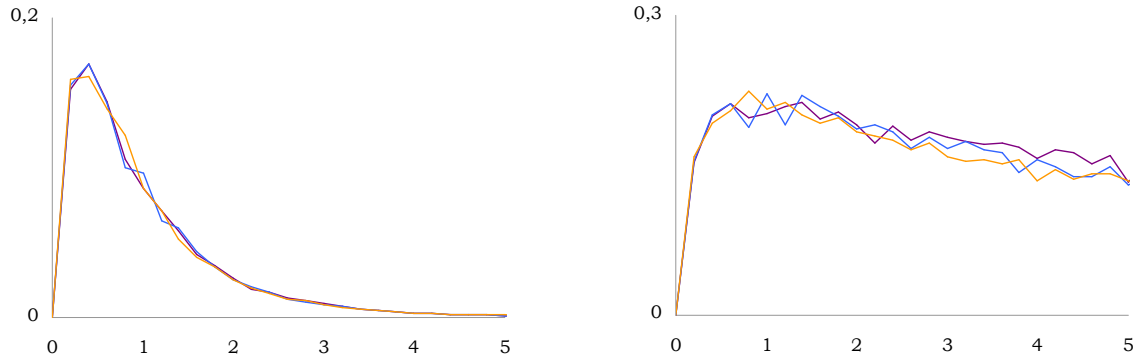


Figure 6.65 : spectres

Ces spectres quelque peu artificiels sont de la même famille que les précédents et traduisent l'unité profonde des phénomènes en jeu.

Sur les courbes absolues, *Désert* se distingue des autres oeuvres, notamment par une plus grande asymétrie. Les courbes relatives commencent par une loi de Rayleigh, suivie par une loi linéraire décroissante.

[364] La moyenne de plusieurs distributions reste une distribution : la somme de ses composantes est mécaniquement égale à 1.

Chapitre 7 : nanoscopie

1 Introduction

L'objectif de ce chapitre est de mesurer les corrélations entre les temps de retour successifs d'une unité, selon les trois plans linguistiques considérés.

Seules les unités les plus marquantes sont analysées ici, les autres figurent en annexe 7.

Les valeurs expérimentales des corrélations sont à comparer avec leurs intervalles de confiance, compte tenu de la taille des échantillons utilisés : elles sont significatives au-delà d'un seuil, mais explicables par le hasard en-deçà. De nombreux travaux ont été consacrés à la détermination de ces intervalles, notamment ceux de Bartlett[365] pour les corrélations totales et de Quenouille[366] pour les partielles. La formule précise dépend de l'ordre de la corrélation, mais une valeur approchée de $2/n^{1/2}$ est retenue pour une confiance de 95 %, où n désigne la taille de l'échantillon. Les valeurs pour chaque unité et chaque œuvre sont données en annexe.

[365] « On the Theoretical Specifications of Sampling Properties of Autocorrelated Time Series ».
[366] « The joint distribution of serial correlation coefficients »

2 Graphémologie

2.1 Espaces

La figure 7.1 montre un comportement similaire pour l'ensemble des œuvres :

– la corrélation d'ordre 1 est négative : les mots courts alternent avec les longs[367], le trait est commun à de nombreuses langues ; par définition, les corrélations totales (à gauche) et partielles (à droite) sont égales à ce stade ;
– pour les ordres supérieurs, les deux versions des corrélations divergent théoriquement, mais elles se rejoignent pratiquement par leur nullité.

Le phénomène est donc indifféremment représenté par un modèle auto-régressif ou à moyenne mobile d'ordres 1 :

– AR(1) : $x(t) = \phi_1 x(t-1) + w(t)$
– MA(1) : $x(t) = \theta_1 w(t-1) + w(t)$.

où w(t) désigne un bruit, i.e. un signal aléatoire de moyenne nulle.

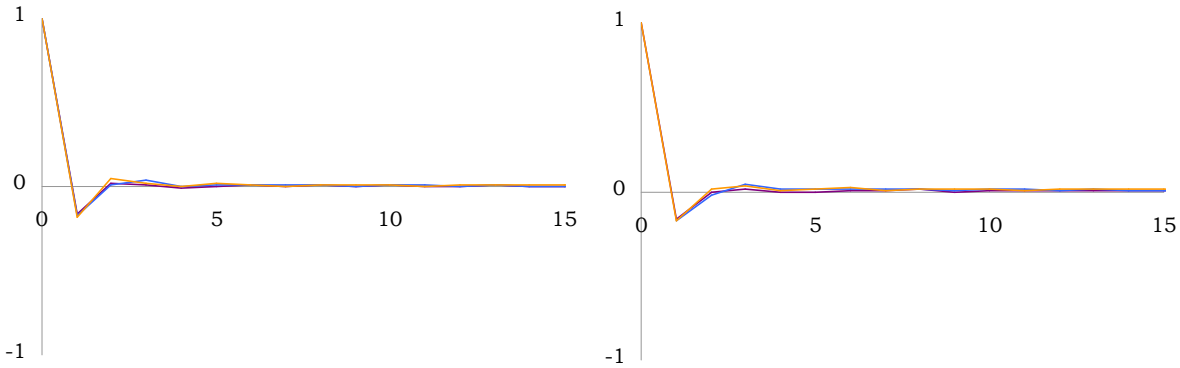


Figure 7.1 : espaces

[367] Par court et long, il faut entendre respectivement inférieur et supérieur à la valeur moyenne, selon la formule de la corrélation.

2.2 Ponctuation

Les points présentent le phénomène inverse (fig. 7.2) : la corrélation d'ordre 1 est positive, en d'autres termes une phrase longue appelle une longue, et inversement pour une courte.

Pour les ordres supérieurs, les corrélations décroissent progressivement et tendent vers 0 en restant généralement positives.

Les corrélations partielles semblent décroître plus vite que les totales : le modèle AR est donc privilégié par rapport au MA. En l'absence de chute franche vers 0, l'ordre est peu déterminé. Avec un intervalle de confiance compris entre 0.03 et 0.04 selon les œuvres, ce paramètre est voisin de 6.

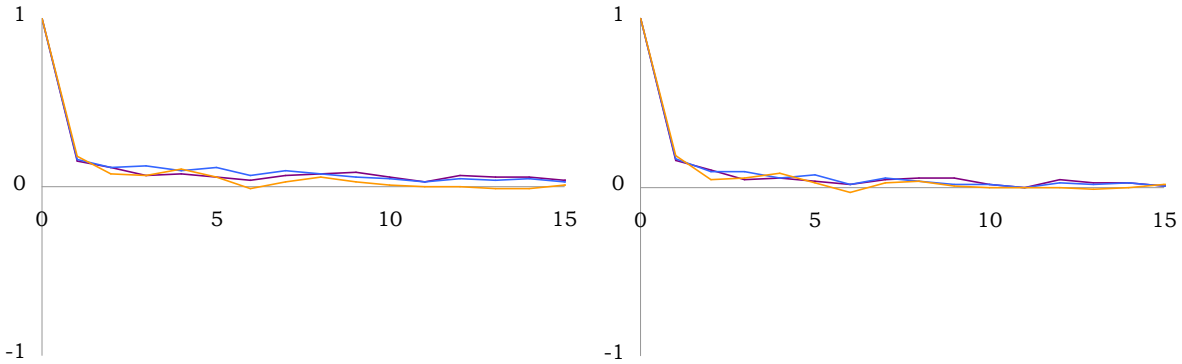


Figure 7.2 : points

Quant aux virgules (fig. 7.3), leurs corrélations sont quasi nulles, si bien que le processus est purement aléatoire. Formulé par x(t) = w(t), il est vu indifféremment comme un AR(0) ou un MA(0).

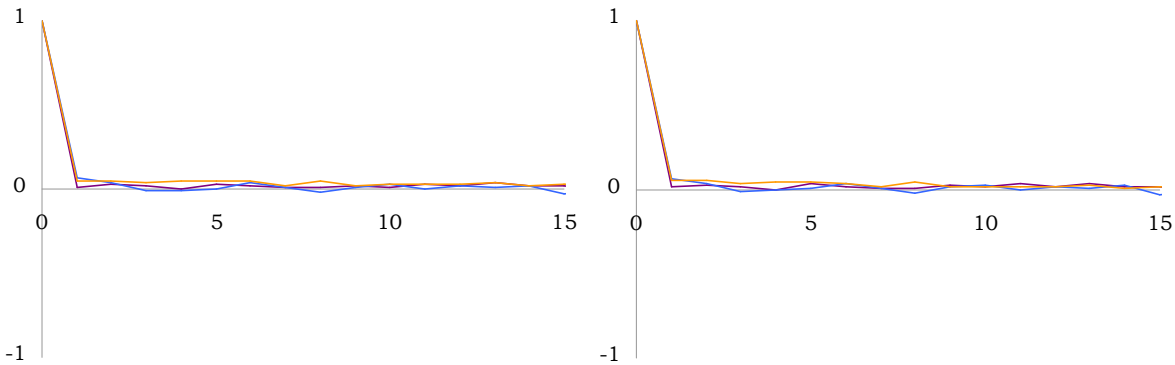


Figure 7.3 : virgules

2.3 Lettres

La lettre la plus fréquente — E — est clairement sans corrélation (fig. 7.4), et répond au même modèle purement aléatoire que les virgules.

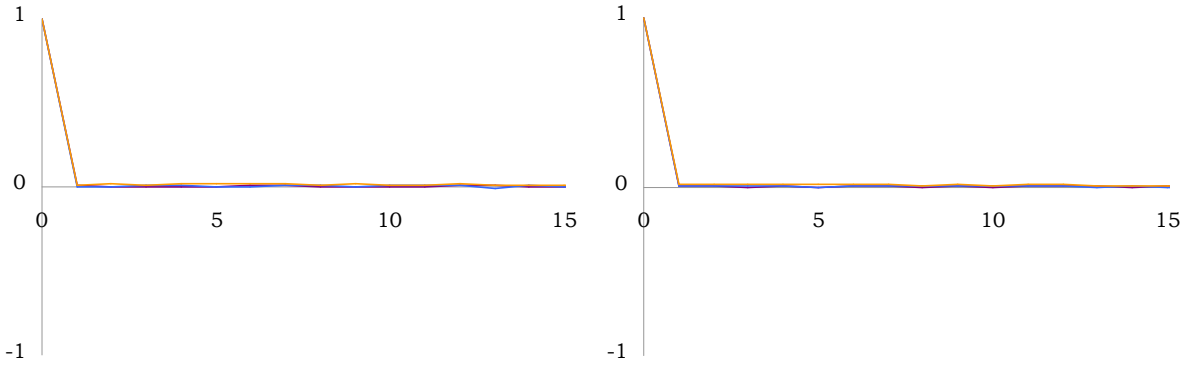


Figure 7.4 : E

Le spectre de la lettre la moins fréquente — K — est plus chaotique. Les valeurs restent en-deçà des seuils de confiance (respectivement 0.48, 0.28 et 0.13 pour *Hadrien*, *Vendredi* et *Désert)*.

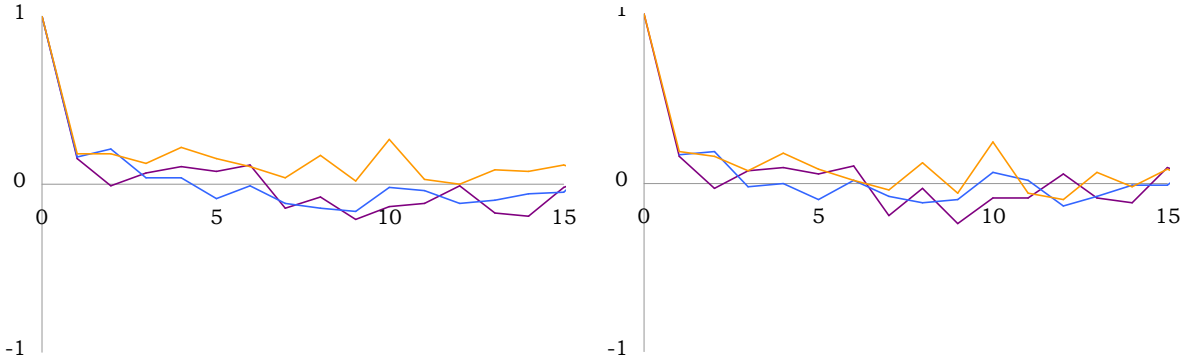


Figure 7.5 : K

De même, les corrélations des autres lettres sont négligeables (cf. annexe 7).

2.4 Synthèse graphémologique

Considérons les temps de retour relatifs x/m, soit à une constante

près l'arythmie[368]. D'après la figure 7.6, ces grandeurs sont totalement décorrélées, et le phénomène est modélisé par un AR(0) ou un MA(0) :

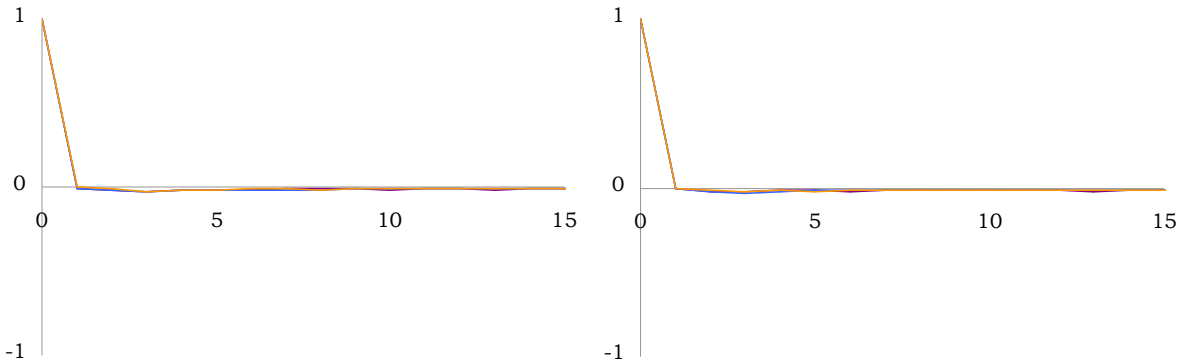


Figure 7.6 : arythmie

3 Syntaxe

3.1 Parties du discours

La catégorie grammaticale la plus courante — le nom — présente des corrélations faibles (fig. 7.7). Guère au-delà du seuil de confiance (entre 0.01 et 0.02 selon les œuvres), elles restent légèrement positives. Le choix d'un modèle AR ou MA ainsi que de son ordre est incertain.

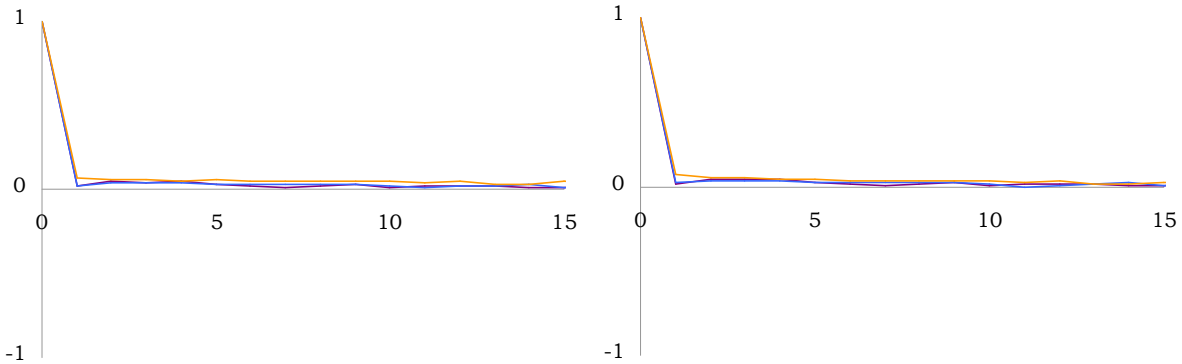


Figure 7.7 : noms

[368] Le coefficient de corrélation est insensible à ces transformations linéaires.

En ce qui concerne la partie la plus rare — la conjonction — les corrélations sont également faibles (fig. 7.8). Le plus souvent positives, elle ne se démarquent guère des seuils de confiance (entre 0.03 et 0.02). Un modèle purement aléatoire AR(0) ou MA(0) semble le plus adapté.

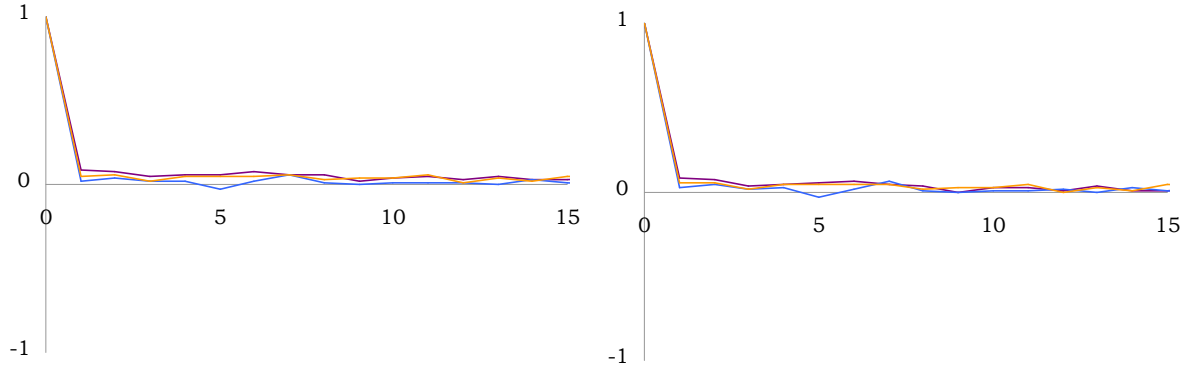


Figure 7.8 : conjonctions

Données en annexe, les corrélations des autres parties du discours sont elles aussi faibles.

3.2 Synthèse syntaxique

De même qu'au niveau graphémologique, la figure 7.9 analyse les corrélations des arythmies (fig. 7.9).

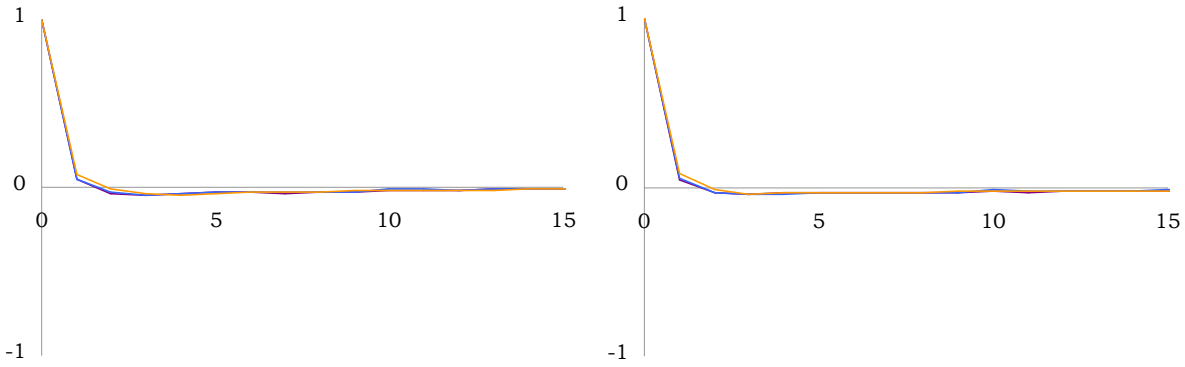


Figure 7.9 : arythmie

Les populations importantes induisent un intervalle de confiance réduit à 0.01 : la corrélation, légèrement positive à l'ordre 1, devient

négative à partir de l'ordre 3. La décroissance vers 0 semble plus franche sur la courbe de gauche, si bien qu'un processus MA limité à l'ordre 10 modélise finement le phénomène. Mais dans une première approche, un processus purement aléatoire reste valide.

4 Sémantique

4.1 Concepts

Le concept le plus fréquent — le fondamental — présente une faible corrélation positive, significative pour les ordres 1 et 2, avec un seuil de confiance de 0.02 (fig. 7.10) :

Linguistiquement, le phénomène se concoit aisément : le texte est organisée en sphères d'influences qui privilégient un thème donné.

Mathématiquement, un modèle MA(2) ou AR(2) explique ces observations, sans qu'il soit possible de trancher entre ces deux familles.

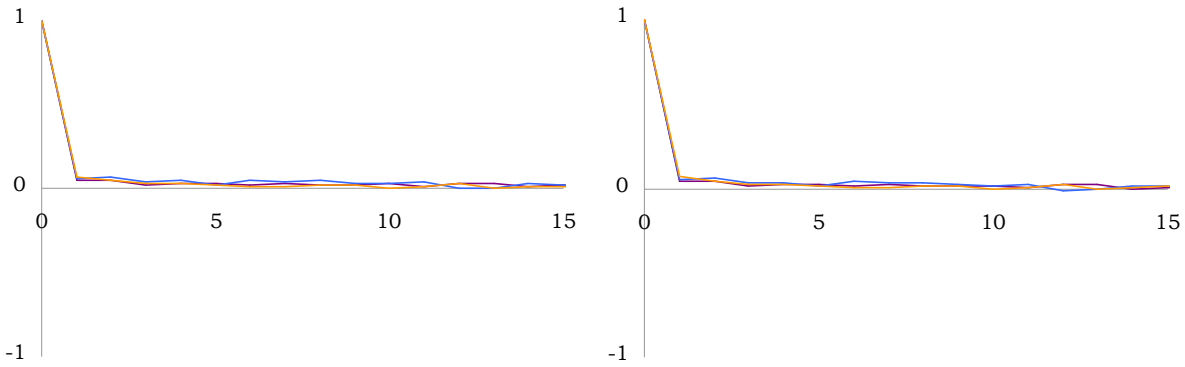


Figure 7.10 : fondamental

Le concept le plus rare — l'information — expose des corrélations plus perturbées (fig. 7.11), en général peu significatives et en deçà du

seuil de confiance (entre 0.12 et 0.15 selon les cas). Négligeables dans *Hadrien*, elles se font légèrement sentir dans *Vendredi* et dans *Désert* jusqu'à l'ordre 3. Les courbes de droite décroissent généralement plus vite que celles de gauche, si bien que les processus MA paraissent mieux adaptés.

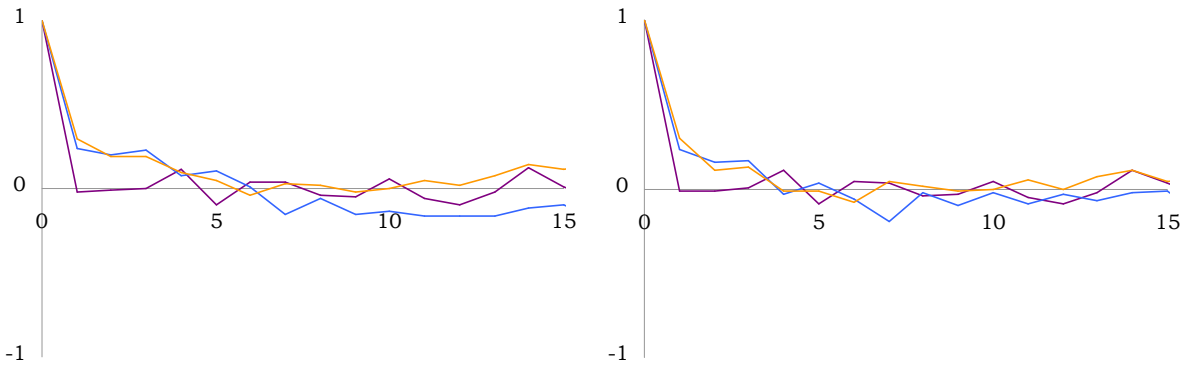


Figure 7.11 : information

Les autres tracés donnés en annexe montrent eux aussi des corrélations faiblement positives, notamment sur les premiers ordres.

4.2 Synthèse sémantique

La figure 7.12 montre sans ambiguité la décorrélation des temps de retour relatifs et de l'arythmie. Ces variables se modélisent donc par un processus purement aléatoire.

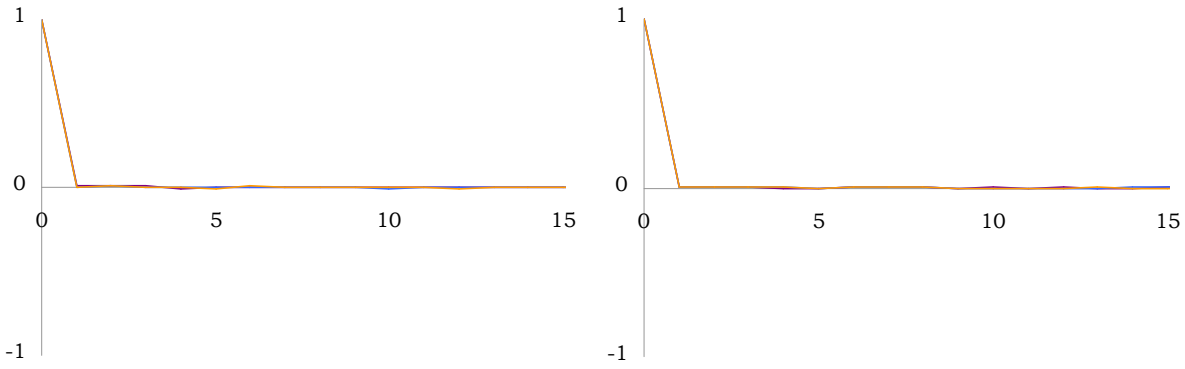


Figure 7.12 : arythmie

5 Synthèse nanoscopique

Moyenne des résultats obtenus sur les trois plans linguistiques, la figure 7.13 montre l'absence de corrélation sur l'arythmie. A nouveau, le modèle aléatoire s'impose.

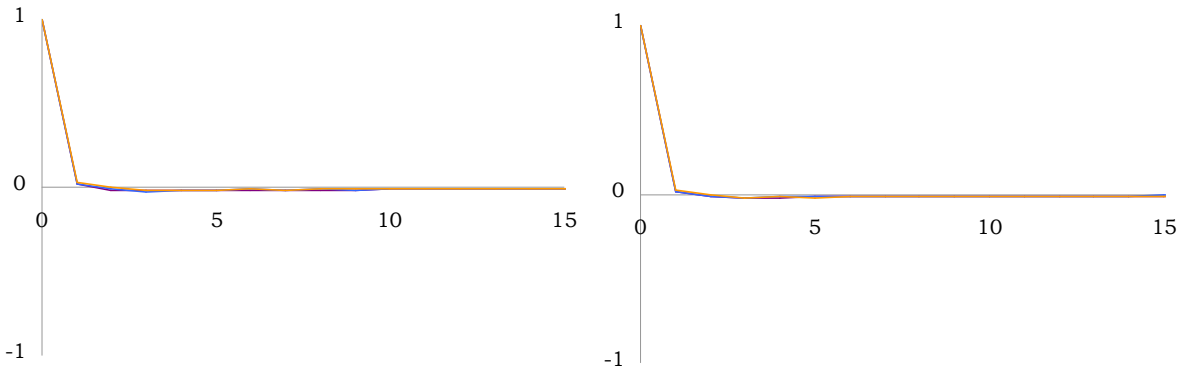


Figure 7.13 : arythmie

Chapitre 8 : télescopie

1 Introduction

La logique de ce chapitre est l'inverse des précédents : il ne s'agit plus d'analyser un corpus connu, mais de classer et d'étiqueter des textes à partir de leurs caractéristiques statistiques, dans une perspective d'attribution d'auteur.

Ce chapitre est d'abord l'occasion de toucher du doigt la mesure qui sous-tend le classement. La distance généralisée intégrant le rythme est comparée à sa version classique, fondée sur les fréquences. Les écarts obtenus sont mis en relation avec les asymétries des distributions[369]. Puis, cette mesure est exploitée pour situer globalement les œuvres du corpus.

Une véritable démarche d'attribution conduirait à considérer le corpus complet de Yourcenar, Tournier et Le Clézio. Or tous leurs livres ne sont pas numérisés. Il paraît donc plus simple de tester la méthode à une échelle plus petite et de simuler le processus sur les divisions de la mésoscopie.

Le test consiste à affecter une division à l'œuvre dont elle est la plus proche. Les distances internes et externes sont comparées, soit les écarts entre une division et son livre d'origine d'une part, les livres collatéraux d'autre part.

---

[369] Les asymétries sont préférées aux variabilités, les moments d'ordre 3 étant plus sensibles à l'arythmie que ceux d'ordre 2.

Afin de se placer dans des conditions réelles d'attribution et d'éliminer une consanguinité confondante, chaque division est rendue orpheline, chassée de son œuvre mère qui se retrouve ainsi quelque peu perturbée. Cette séparation troublante est néanmoins le prix d'un jugement impartial.

L'annexe 8 rassemble les valeurs numériques des distances, tandis que les mesures fréquentielles sont tirées de l'annexe 5. Il comprend également les asymétries de chaque division.

Comme dans les autres chapitres, l'étude suit les trois plans linguistiques considérés.

2 Graphémologie

2.1 Divisions

2.1.1 Distances internes

La distance généralisée qui prend en compte le rythme est tracée en traits gras, tandis que la version classique fondée sur les fréquences est en traits fins.

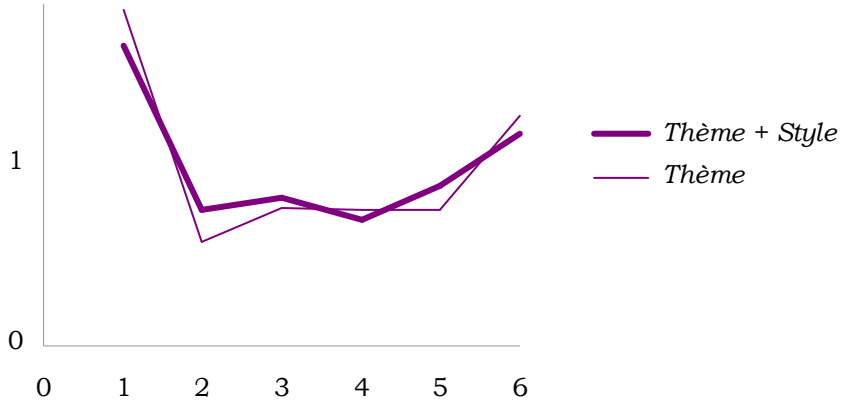


Figure 8.1 : Hadrien

Dans *Hadrien*, les deux mesures sont très corrélées. L'écart le plus important est obtenu par la division 1, dont l'arythmie est maximale : la distance totale est alors réduite par rapport à sa version classique (fig. 8.1).

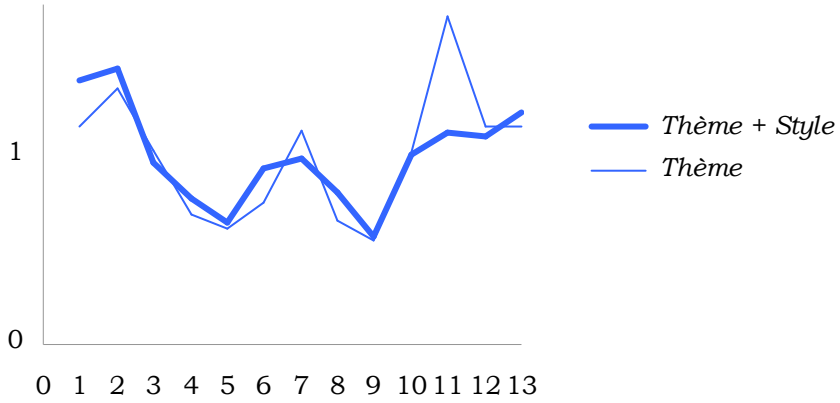


Figure 8.2 : Vendredi

En ce qui concerne *Vendredi*, la corrélation entre les deux mesures reste importante, mais des écarts considérables apparaissent localement, en particulier sur la division 11. L'asymétrie est alors quasiment maximale (1.13 pour un extrême de 1.14). Ici également, la distance totale est plus faible que son homologue classique (fig. 8.2).

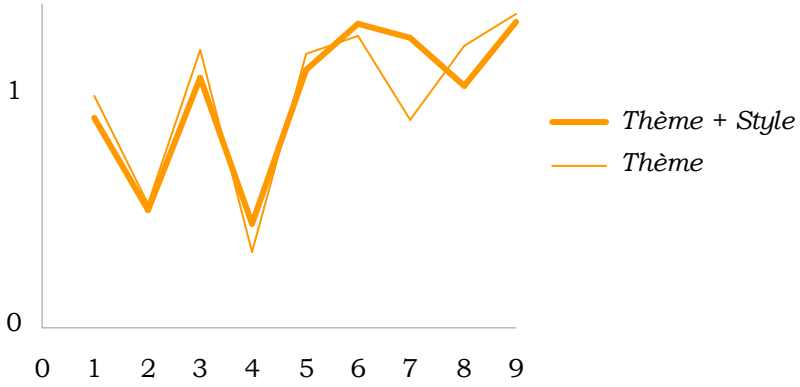


Figure 8.3 : Désert

Dans *Désert*, le phénomène s'inverse : sur la division 7, la distance

totale est significativement accrue par rapport à la distance classique : l'asymétrie est alors minimale (fig. 8.3).

2.1.2 Attribution

La figure 8.4 représente les écarts entre une division d'*Hadrien* (axes 1 à 6) et une œuvre du corpus (polygones colorés), avec les précautions évoquées dans l'introduction : pour calculer les distances internes, chaque partie d'*Hadrien* est successivement retirée de l'œuvre mère. Le problème ne se pose pas pour les distances externes par rapport à *Vendredi* et *Désert*.

Sans ambiguïté, chaque division d'*Hadrien* est plus proche d'*Hadrien* que de *Vendredi* ou de *Désert*. L'attribution se fait donc aisément.

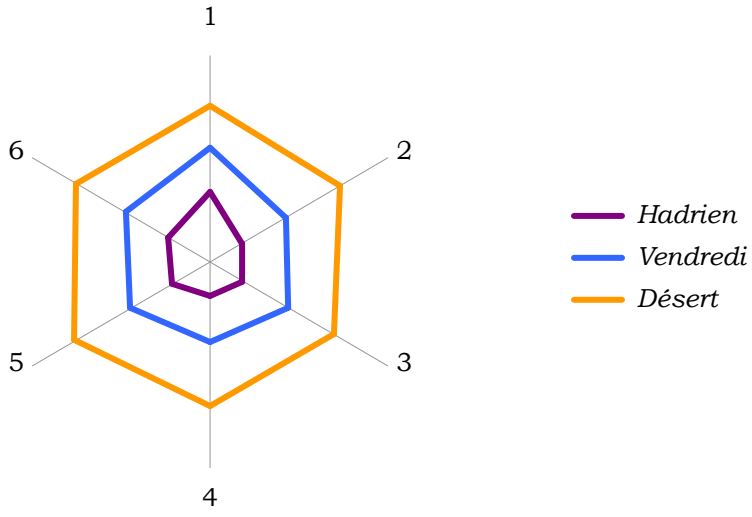


Figure 8.4 : attribution d'Hadrien

Les résultats sont semblables pour *Vendredi* (fig. 8.5) et *Désert* (fig. 8.6).

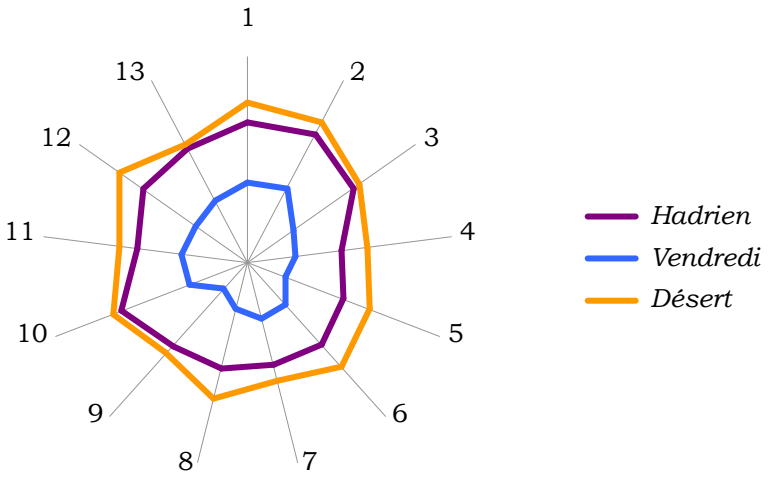


Figure 8.5 : attribution de Vendredi

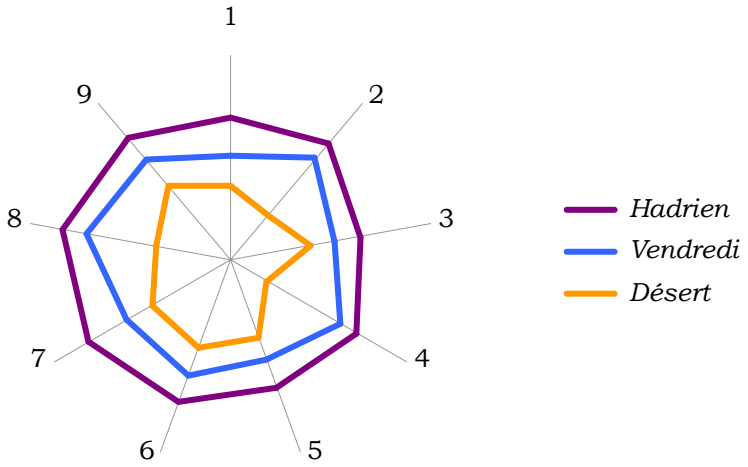


Figure 8.6 : attribution de Désert

Certaines différences apparaissent cependant : les écarts entres les distances internes et externes sont plus marqués dans *Hadrien* que dans *Vendredi,* et a fortiori que dans *Désert* (fig. 8.7). La première œuvre forme un ensemble homogène qui ressort clairement du corpus, tandis que la dernière se répand et se fond volontiers dans son environnement. Bien entendu, l'attribution est d'autant plus facile que la marge est forte, et devient a contrario délicate quand les écarts s'amenuisent.

Figure 8.7 : distances internes/externes

2.2 Corpus

L'application de la distance généralisée aux éléments du corpus permet d'établir une carte pour le niveau graphémologique. Affinée par rapport à celles de la macroscopie, elle trace un contour néanmoins connu : *Hadrien* et *Désert* sont les œuvres les plus éloignées, tandis qu'*Hadrien* et *Vendredi* sont les plus proches (fig. 8.8) :

Figure 8.8 : carte du corpus

3 Syntaxe

3.1 Divisions

3.1.1 Distances internes

Comme au niveau des graphèmes, les distances totales et classiques sont généralement corrélées.

Dans *Hadrien*, les écarts apparaissent notamment sur la division 1, dont l'asymétrie est maximale. La distance totale est alors en-deçà de

son homologue (fig. 8.9) :

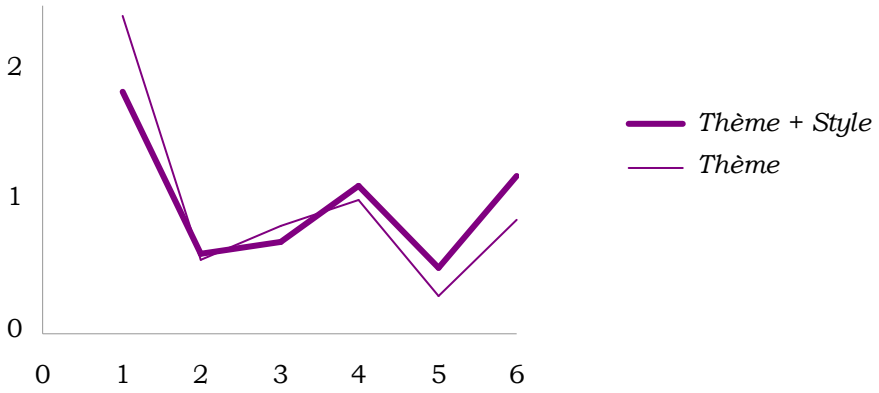


Figure 8.9 : Hadrien

Dans *Vendredi*, le phénomène est analogue, mais avec des écarts plus importants, en particulier pour la division 11 dont l'asymétrie est maximale ; le phénomène s'inverse avec l'asymétrie minimale de la division 2, et la courbe en gras passe au-dessus de son alter ego (fig. 8.10) :

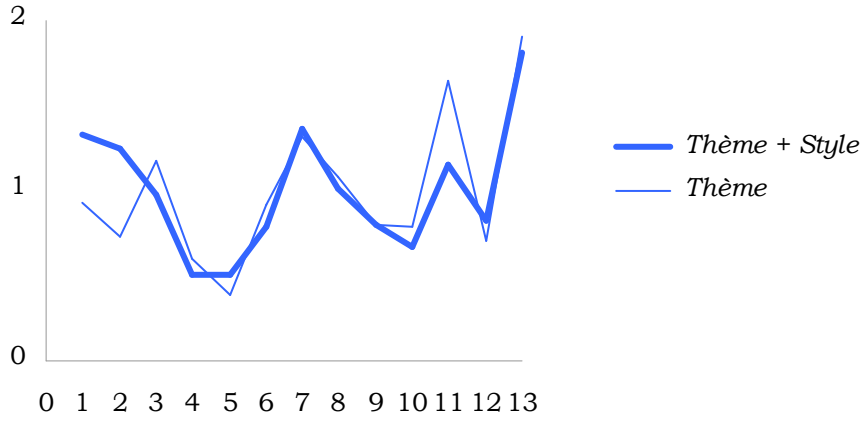


Figure 8.10 : Vendredi

Enfin dans *Désert*, l'écart est sensible sur la division 7, qui présente une asymétrie proche des minima (1.05 pour un extrême de 1.04). La distance totale est alors plus forte que la distance classique (fig. 8.11) :

Figure 8.11 : Désert

3.1.2 Attribution

Dans *Hadrien*, le test échoue pour la division 1 : partie la plus excentrée, elle est associée à *Désert*[370]. Le test réussit toutefois dans les autres cas (fig. 8.12) :

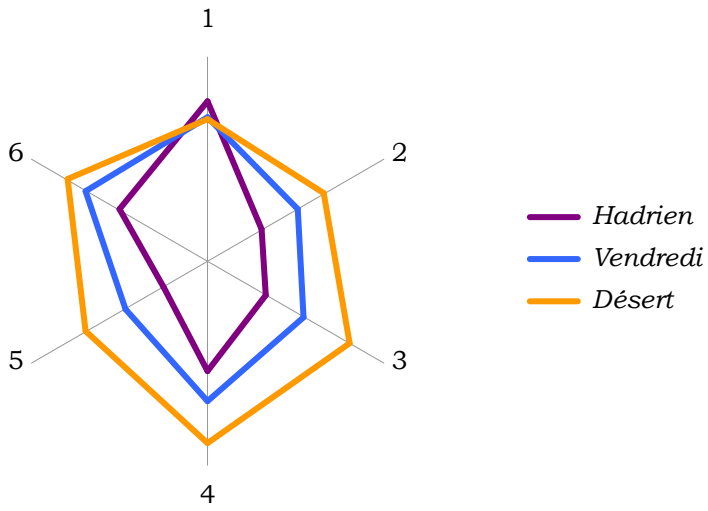


Figure 8.12 : attribution d'Hadrien

Dans *Vendredi*, l'attribution est défectueuse pour les divisions 1 et 7,

[370] Si l'éloignement rend parfois l'attribution incertaine, il peut dans d'autres cas la confirmer : imaginons deux points fixes d'une droite et un élément que l'on associe à la plus proche de ces références. Il suffit de déplacer l'élément d'une extrémité à l'autre de la droite pour voir les interactions entre ces deux influences.

elles aussi atypiques. L'erreur menace en outre les divisions 3 et 13. Ailleurs, la discrimination se fait dans de bonnes conditions (fig. 8.13) :

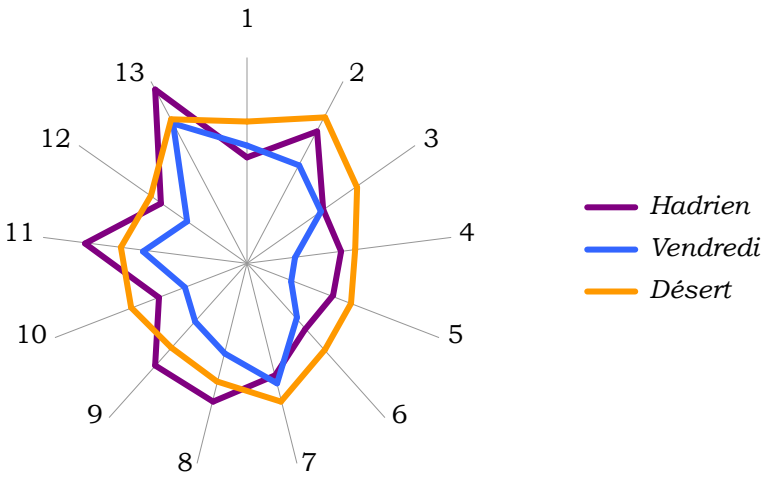


Figure 8.13 : attribution de Vendredi

Dans *Désert*, l'attribution devient problématique : elle avorte pour les divisions 1, 5, 7 et 9 qui comptent parmi les plus excentrées, et réussit difficilement ailleurs (fig. 8.14) :

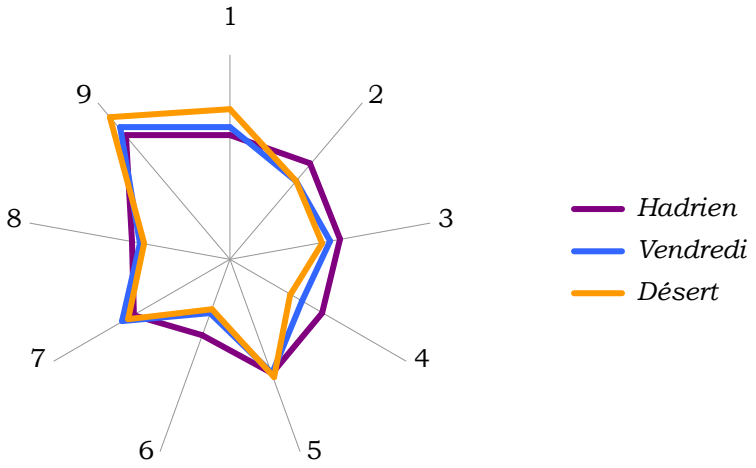


Figure 8.14 : attribution de Désert

La figure 8.15 illustre ce phénomène : l'écart entre les distances internes et externes est quasiment nul dans *Désert*, qui peine à se forger une identité dans le corpus. En revanche, cette marge reste

confortable pour les autres œuvres, dont les contours sont plus marqués.

Figure 8.15 : distances internes/externes

3.2 Corpus

La carte établie à l'aide des données syntaxique est proche de celle du niveau précédent. *Vendredi* s'éloigne d'*Hadrien*, mais plus proche de ce dernier que de *Désert*.

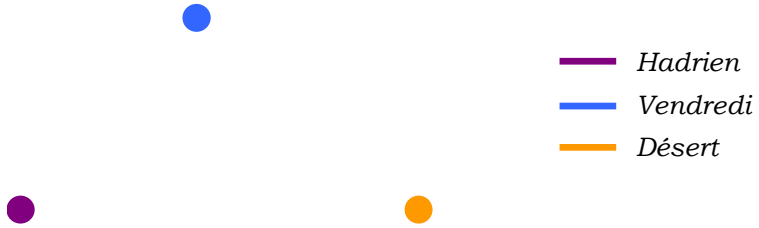


Figure 8.16 : carte du corpus

4 Sémantique

4.1 Divisions

4.1.1 Distances internes

Les distances totales et classiques restent étroitement corrélées dans *Hadrien.* Les écarts les plus importants surviennent dans la division 1 dont l'asymétrie est quasiment maximale (1.55 pour un extrême de 1.56). La courbe épaisse est alors en dessous de la courbe fine (fig.

8.17) :

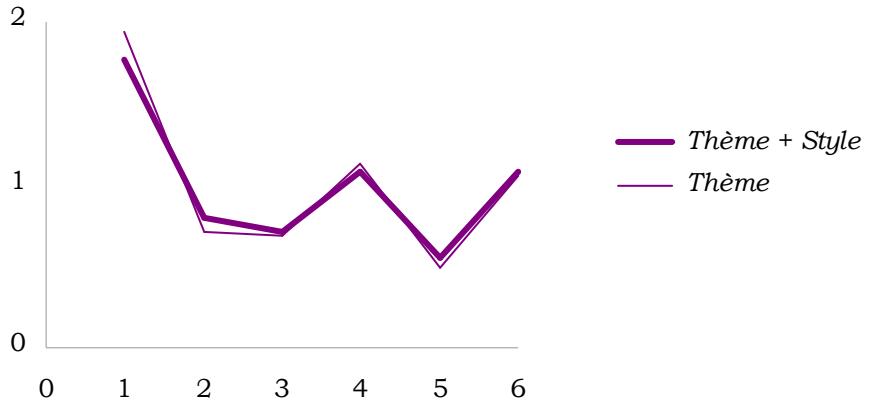


Figure 8.17 : Hadrien

Les écarts sont plus sensibles dans *Vendredi*. Leur maximum est atteint par la division 1 tandis que l'asymétrie est minimale. La distance totale dépasse alors sa version classique (fig. 8.18) :

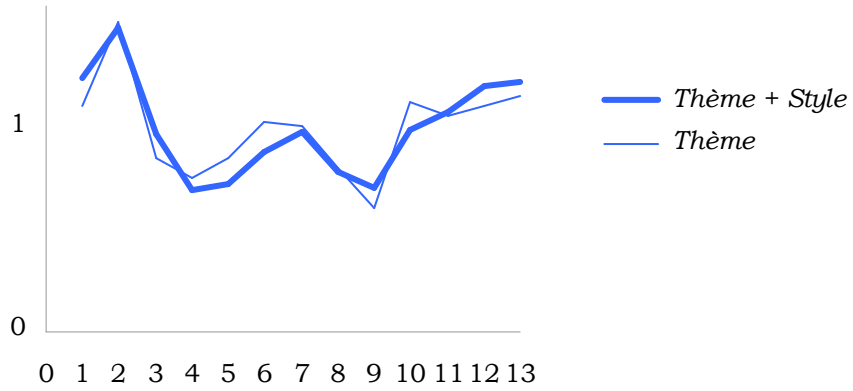


Figure 8.18 : Vendredi

Dans *Désert*, les différences les plus fortes se font sentir sur la division 6, en relation avec une asymétrie minimale qui fait passer la courbe épaisse au-dessus de la courbe fine (fig 8.19) :

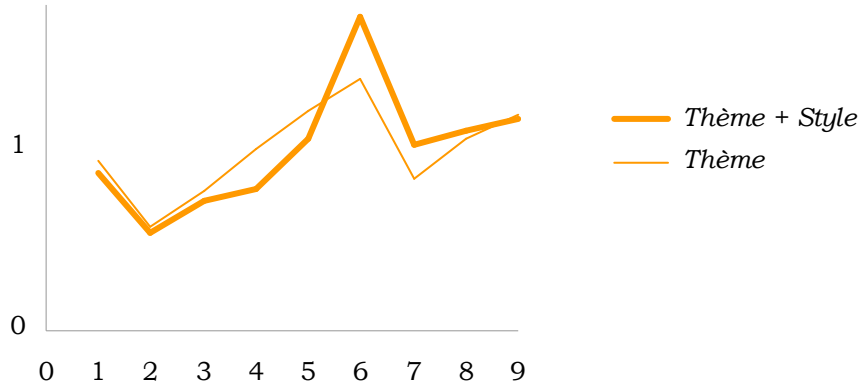


Figure 8.19 : Désert

4.1.2 Attribution

L'attribution des partie d'*Hadrien* se fait sans encombre (fig. 8.20) :

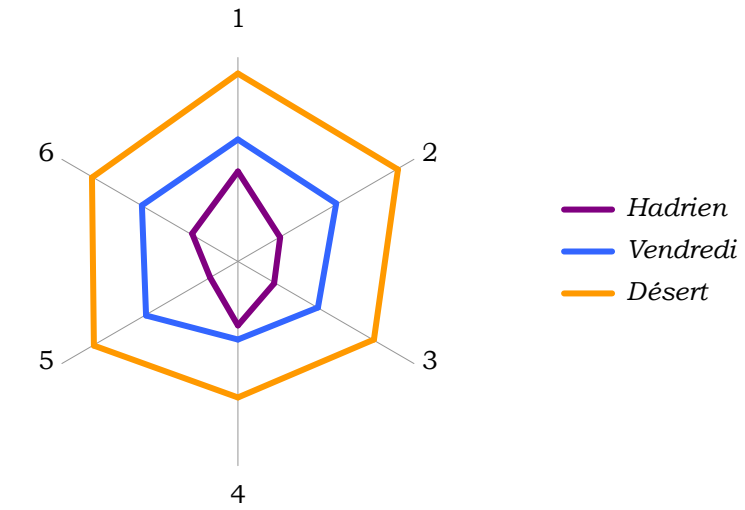


Figure 8.20 : attribution d'Hadrien

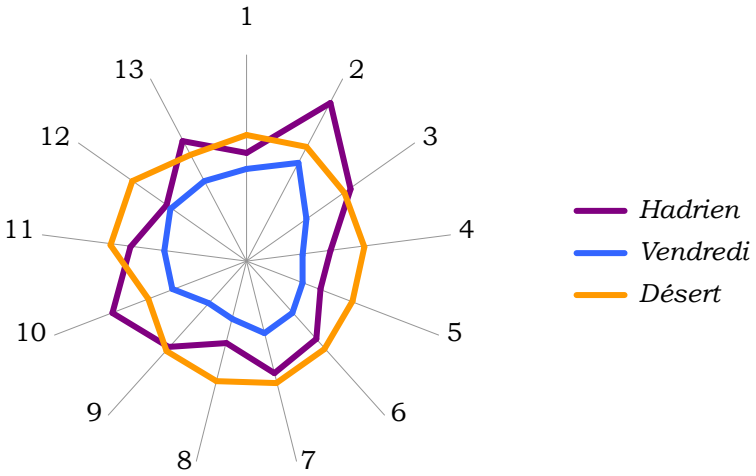


Figure 8.21 : attribution de Vendredi

Celle de *Vendredi* se montre un peu plus délicate, en particulier pour la division 12 assez excentrée (fig. 8.21).

Elle reste sans ambiguïté pour *Désert*, quelle que soit la division considérée (fig. 8.22) :

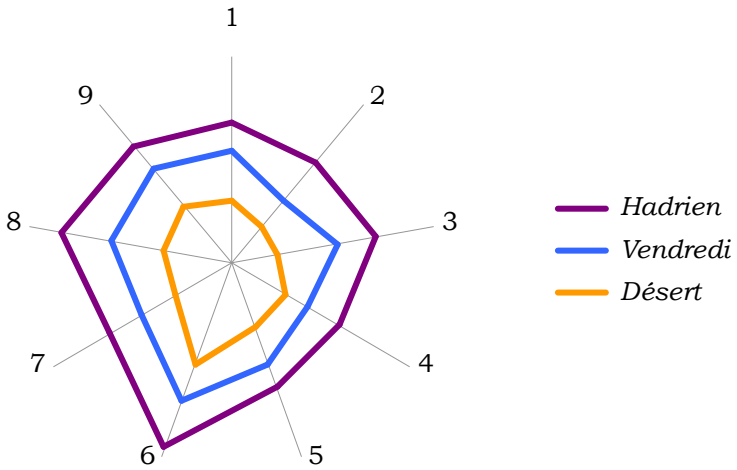


Figure 8.22 : attribution de Désert

La figure 8.23 illustre les remarques qui précèdent : la marge entre les distances internes et externes est maximale dans *Hadrien*, et mimimale dans *Vendredi*. Les contours du premier sont donc plus nets pour le premier que pour le second.

Figure 8.23 : distances internes/externes

4.2 Corpus

La carte du niveau sémantique confirme les précédentes, allongeant la distance entre les pôles du corpus, *Hadrien* et *Désert* (fig. 8.24) :

Figure 8.24 : carte du corpus

5 Synthèse télescopique

5.1 Divisions

Les distances généralisées et classiques évoluent globalement dans le même sens. Fait remarquable, les écarts les plus forts apparaissent sur les mêmes divisions, indépendamment du plan linguistique : ainsi la première dans *Hadrien*, la onzième dans *Vendredi* et la septième dans *Désert*. Une organisation profonde semble donc sous-tendre ces phénomènes de surface.

Les écarts reflètent pour une grande part l'arythmie des divisions. Selon une métaphore physique, tout se passe comme si un mouvement secondaire de rotation venait s'ajouter à la translation d'ensemble de la structure en désordonnant ses unités. Cette « rotation » se rapproche d'ailleurs curieusement de l'étymologie de l'entropie, concept étroitement lié à l'arythmie[371]. Il s'agit évidemment d'une explication simplifiée destinée à faire parler les différentes formulations. Par ailleurs, la taille des divisions ne semble pas jouer de rôle dans ces écarts.

Pour compléter l'analyse, essayons d'interpréter ces distances en

[371] Cf. chapitre 2, section 5.1.3.1.

termes littéraires et linguistiques.

5.1.1 Distances internes

La figure 8.25 résume les résultats obtenus sur *Hadrien*. Pour les deux mesures, « Disciplina augusta » est la partie la plus caractéristique, ce qui se conçoit à la lumière des personnalités d'Hadrien et de Yourcenar. A l'opposé, « Animula vagula blandula » et « Patientia », au début et à la fin du livre, sont vues communément comme les parties les plus atypiques.

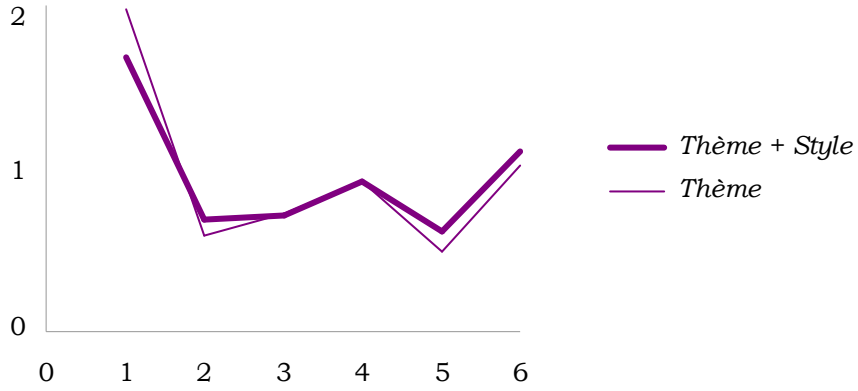


Figure 8.25 : Hadrien

Les mesures relatives à *Vendredi* sont illustrées par la figure 8.26. Le quatrième chapitre, et pour une moindre part le troisième et le huitième, sont les plus typiques. En revanche, les éléments les plus excentriques sont le prologue, le premier et le dernier chapitre, alors que la considération des fréquences désigne un chapitre interne, le dixième. Or les parties introductives et conclusives d'un livre sont a priori plus distantes que celles du récit principal.

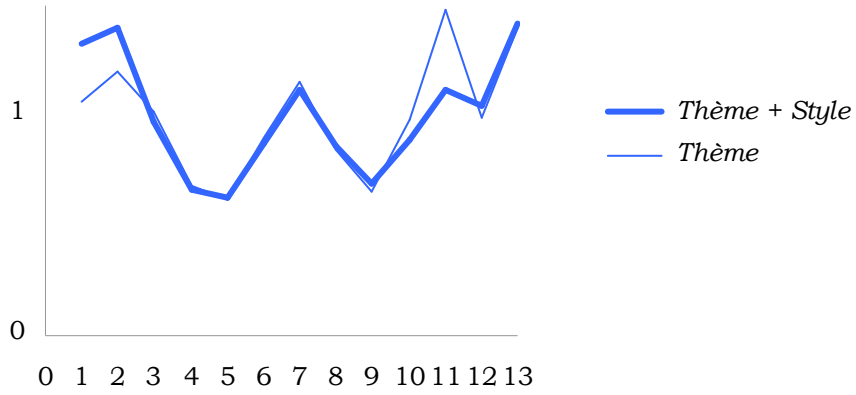


Figure 8.26 : Vendredi

Enfin, les écarts de *Désert* sont indiqués sur la figure 8.27. Sans surprise et pour les deux mesures, « Le Bonheur » et « La vie chez les esclaves » sont les parties les plus caractéristiques du livre. Par contre, la mesure généralisée fait ressortir la dernière division, tandis que le distance fréquentielle pointe vers une division interne, la cinquième.

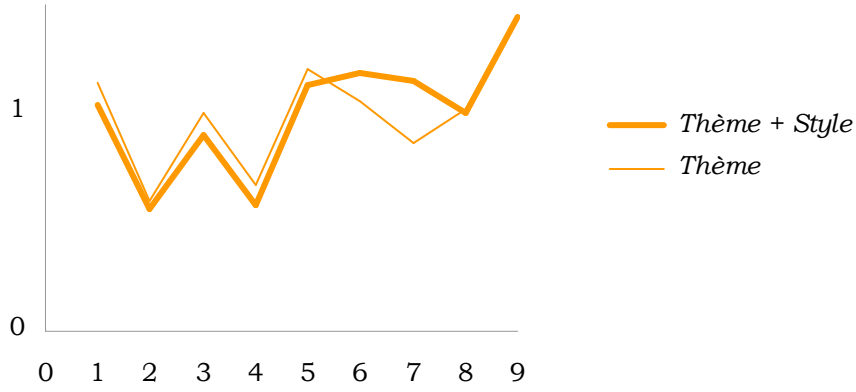


Figure 8.27 : Désert

5.1.2 Attribution

Les meilleures performances sont obenues pour les niveaux graphémologiques et sémantiques. Il n'est donc pas pertinent d'intégrer les mesures syntaxiques pour procéder à la synthèse. La figure 8.29 traduit plus précisément ce phénomène à l'aide du ratio des distances internes et externes : les graphèmes semblent les plus adaptés pour

caractériser un auteur, suivis par les concepts. En revanche, des parties du discours dont l'usage est normé discriminent plus difficilement les auteurs.

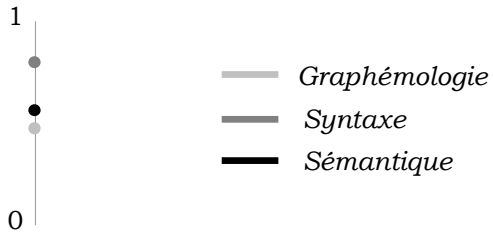


Figure 8.29 : distances internes/externes

Il est d'autre part intéressant de constater qu'*Hadrien* possède la plus forte individualité, tandis que *Vendredi* et surtout *Désert* tendent à se fondre dans le corpus (fig. 8.28) :

Figure 8.28 : distances internes/externes

5.2 Corpus

La carte globale et finale du corpus, moyenne des distances sur les trois plans linguistiques, confirme les conclusions déjà tirées par de multiples approches : une gradation d'*Hadrien* à *Désert, Vendredi* se tenant légèrement plus proche du premier que du dernier (fig. 8.30) :

Figure 8.30 : carte du corpus

Conclusion

Au fil des pages, cette thèse a suivi un double objectif : d'une part, la mise au point de mesures destinées à comparer les textes ; de l'autre, l'étude du corpus dans ses aspects linguistiques et stylistiques.

En suivant ce découpage, examinons les principaux enseignements de cette recherche et les ouvertures qui s'en dégagent.

1 Méthode et mesure

1.1 Bilan

De la première scopie à la dernière, les mesures s'affinent et tentent de refléter au mieux la réalité littéraire. Leur ontogénèse suit en quelque sorte la phylogénèse de la stylométrie.

La macroscopie et l'analyse des fréquences restent les fondements de ces mesures, et donnent une image claire de la composition d'un texte, ou d'un « thème » généralisé.

Selon les divisions d'un livre, ces fréquences fluctuent. A partir de ce mouvement, la mésoscopie trace un portrait plus fidèle du texte, qui prend en compte l'organisation des unités linguistiques et ébauche donc le style.

La microscopie passe de l'autre côté du miroir : il ne s'agit plus de mesurer une abondance héritée d'une conception linguistique, mais une rareté au parfum plus stylistique. Les temps de retour d'une unité et

leur distribution font toucher du doigt un rythme longtemps espéré. Autour des valeurs moyennes qui centrent les spectres ainsi obtenus, les arythmies définissent leur forme. Si les premières traduisent le thème, les secondes reflètent le style. Selon une technique éprouvée du domaine de la fiabilité, ces distributions complexes se résolvent à l'aide de probabilités conditionnées par un passé connu. Les spectres relatifs qui en résultent se laissent linéariser par intervalles, d'où des modèles et des simulations simplifiés.

Au sens propre, le rythme dépend en outre de la succession des intervalles, d'où une nanoscopie qui analyse les corrélations temporelles et tente de les expliquer à l'aide de modèles aléatoires issus du domaine économique, appelés « Auto-Regressive » et « Moving Average ».

Dans une perspective d'attribution d'auteur, la téléscopie fait le tri entre ces mesures pour retenir l'information utile, et adopte une approche descriptive sans se référer à un modèle probabiliste. Elle se fonde sur les répartitions des unités, qui intègrent les facteurs thématiques et stylistiques tout en lissant les irrégularités des distributions. La distance généralisée qui en résulte suit globalement les évolutions de son homologue classique basée sur les fréquences, mais des différences significatives apparaissent localement, selon l'intensité de l'arythmie.

1.2 Ouvertures

D'inspiration musicale, l'étude privilégie le rythme et le temps pour mettre au second plan les interactions spatiales entre les unités. D'un point de vue technique, la généralisation de la méthode précédente permet de les prendre en compte : il suffit de substituer aux temps de retour d'une unité les temps de transition entre deux unités. Le coût

d'une telle opération est loin d'être négligeable[372], et les excursions souvent infructeuses de la nanoscopie suggèrent qu'une analyse trop fine ne se justifie pas toujours. La question reste néanmoins ouverte.

Au-delà du champ littéraire et linguistique, l'approche suivie est applicable à d'autres domaines, comme la musique ou les arts plastiques : formellement et mathématiquement, seule la répartition des unités intervient, et non les unités elles-mêmes, d'où la simplicité de la transposition. De l'esthétique au poétique, le point de vue est d'autre part susceptible de s'inverser : les processus aléatoires se détournent alors de l'analyse pour synthétiser des formes artistiques.

2 Etude du corpus

Les principaux enseignements de l'étude sont classés en deux catégories : les convergences linguistiques entre les œuvres, et leurs particularités stylistiques.

2.1 Bilan linguistique

Les comptages de la macroscopie affichent des résultats connus : les graphèmes les plus courants sont les espaces et la lettre E, tandis que les noms et les verbes prévalent parmi les parties du discours. Selon le découpage du monde adopté, le concept majeur est « ordre et mesure ».

La mésoscopie fait apparaître une structure commune, indépendante du plan linguistique considéré — graphémologique, syntaxique ou

[372] Si J désigne le nombre d'unités, la charge augmente proportionnellement à J^2, contre J dans le cas précédent.

sémantique : dans chaque œuvre, le début, la fin et un pivot central saillissent, en relation avec la taille réduite de ces divisions. Coïncidence ou organisation générique ? Ce point mériterait d'être creusé.

Dans la microscopie, les distributions des temps de retour forment des courbes en cloche asymétriques. Les spectres relatifs dégagent deux familles principales qui traduisent le phénomène suivant : au voisinage d'une première occurrence, la venue d'une seconde est peu probable, mais cette aversion diminue au fur et à mesure, si bien qu'en dehors de cette sphère d'influence, l'apparition d'une unité à un instant donné devient indépendante du temps passé. Au prix de quelques adaptations, ce modèle s'applique à l'ensemble des plans linguistiques considérés. Dans le détail, le niveau sémantique présente une arythmie plus forte : aériens et nonchalants, les concepts flottent au-dessus d'éléments plus terriens. Ce résultat doit cependant être tempéré au regard de la faiblesse relative des populations envisagées et de la complexité de l'étiquetage des concepts.

Des corrélations significatives entre les temps de retour n'apparaissent qu'au niveau graphémologique pour les espaces et les points, et dans une moindre mesure au niveau sémantique pour les concepts. Ailleurs, la nanoscopie fait le constat désabusé de la décorrélation de ces intervalles, et semble faire un pas de trop vers des modèles stochastiques sophistiqués. Tout se passe comme si chacun de ces intervalles était le fruit d'un tirage purement aléatoire, entièrement défini par sa distribution. Cette scopie qui paraît d'emblée un peu vide est finalement riche d'enseignements : elle révèle la nature profondément aléatoire d'un « rythme » aux accents étranges, et justifie a posteriori les statistiques de la microscopie, ainsi fondées sur des observations pratiquement indépendantes.

Dans la télescopie, l'affectation des divisions à leur œuvre-mère se révèle fiable pour la plupart des cas. Elle résulte d'une unité organique de chaque oeuvre au sein du corpus, traduite par des écarts intra-classes sensiblement inférieurs aux écarts inter-classes. En d'autres termes, la notion de style d'auteur se vérifie ici, et justifie a posteriori la démarche d'attribution. Celle-ci se révèle la plus sûre à partir des graphèmes, et dans une moindre mesure des concepts[373]. On s'interroge alors : le sens est-il véritablement premier, ou la petite musique des lettres mène-t-elle secrètement la danse ? Les deux niveaux semblent en tous cas pouvoir être mis sur un même plan. En revanche, une syntaxe cristallisée par les règles de la grammaire laisse peu de place à la différenciation et rend l'attribution incertaine. Dans cette optique, le niveau graphémologique offre des avantages certains : quantitativement, il fournit les populations les plus nombreuses, donc les statistiques les plus sûres ; qualitativement, le processus devient objectif et simple en l'absence de prétraitement linguistique. Ce plan ouvre d'intéressantes perpectives, comme l'analyse des langues oubliées ou inconnues.

Globalement et au-delà des différences qui peuvent apparaître, les différents niveaux linguistiques semblent se répondre mutuellement et obéir profondément aux mêmes lois.

2.2 Bilan stylistique

Les mesures effectuées dans la macroscopie confirment l'intuition littéraire : le style résolument classique de *Mémoires d'Hadrien* tranche avec les traces assagies du Nouveau Roman dans *Désert. Vendredi ou*

[373] De même, les systèmes de reconnaissance de la parole utilisent les sons sans connaître leur signification.

les limbes du Pacifique paraît hésiter entre ces deux pôles, mais reste plus proche de la tradition.

Les fluctuations entre les divisions prises en compte par la mésoscopie placent le livre de Tournier à la charnière entre un âge classique attaché à l'ordre et une ère moderne éprise de liberté voire de chaos.

Dans la microscopie, les niveaux de l'arythmie sont proches pour les trois oeuvres et ne sont perceptibles qu'avec l'asymétrie, plus sensible que la variabilité. L'analyse précise pose un regard familier sur notre corpus : l'axe qui oppose *Hadrien* à *Désert* est tracé une nouvelle fois.

La nanoscopie ne permet pas de différencier efficacement les œuvres : dans les rares cas où les corrélations sont significatives, les niveaux obtenus sont communs, d'où une approche impropre à la caractérisation.

Les attributions d'auteur testées dans la télescopie montrent la forte homogénéïté des divisions de *Mémoires d'Hadrien*. A l'opposé, celles de *Désert* se dispersent et tendent à se fondre dans le corpus.

Les chiffres de l'ensemble des scopies concordent, et font apparaître la même gradation entre les œuvres du corpus, en suivant leur chronologie.

2.3 Ouvertures

Le cadre de cette étude est évidemment limité, et les résultats qui précèdent doivent être éprouvés sur des corpus plus vastes, afin de faire jouer pleinement les paramètres de l'auteur, du genre, de la chronologie, de la forme poétique ou prosaïque.

Testée sur le théâtre classique de Corneille, Molière et Racine[374], la distance généralisée conclut à de bonnes attributions d'auteur sauf dans certains cas remarquables : lorsque les tragédiens écrivent des comédies et inversement ; lorsque Racine à ses débuts s'inspire de Corneille. En revanche, la forme poétique ou prosaïque ne semble pas jouer un rôle significatif.

[374] Ce corpus a fait l'objet de nombreuses publications, aux conclusions souvent discordantes. Citons ici : Beaudouin & Yvon, « Contribution de la métrique à la stylométrie » ; Labbé & Labbé, « Intertextual Distance and Authorship Attribution : Corneille and Molière » ; Viprey, « Analyse séquencée de la micro-distribution lexicale ».

Bibliographie[375]

Corpus

LE CLEZIO J.M.G., *Désert*, Paris, Gallimard, 1980 (Frantext, Gallimard 1995).

TOURNIER M., *Vendredi ou les limbes du Pacifique,* Paris, Gallimard, 1967 [Frantext, Gallimard 1995].

YOURCENAR M., *Mémoires d'Hadrien*, Paris, Plon, 1951 [Frantext, Gallimard 1991].

Œuvres associées

LE CLEZIO J.M.G., *Le Procès-Verbal*, Paris, Gallimard, 1963.

LE CLEZIO J.M.G., *Le Déluge,* Paris, Gallimard, 1966.

LE CLEZIO J.M.G., *L'extase matérielle*, Paris, Gallimard, 1967.

LE CLEZIO J.M.G., *La guerre*, Paris, Gallimard, 1970.

LE CLEZIO J.M.G., *Les géants*, Paris, Gallimard, 1973.

LE CLEZIO J.M.G., *Vers les Icebergs*, Saint-Clément, Fata Morgana, 1978.

LE CLEZIO J.M.G., *L'inconnu sur la terre,* Paris, Gallimard, 1978.

LE CLEZIO J.M.G., *Le chercheur d'or*, Paris, Gallimard, 1985.

LE CLEZIO J.M.G., *Voyage à Rodrigues*, Paris, Gallimard, 1986.

LE CLEZIO J.M.G., *Le Rêve mexicain ou la pensée interrompue* (Paris,

[375] Quand l'édition utilisée n'est pas l'originale, elle est spécifiée entre les symboles [ ].

Gallimard, 1988.

LE CLEZIO J.M.G, *La fête chantée*, Paris, Le Promeneur, 1997.

LE CLEZIO J. et J.M.G., *Gens des Nuages*, Paris, Stock, 1997.

TOURNIER M., *Le Roi des Aulnes*, Paris, Gallimard, 1970.

TOURNIER M., *Vendredi ou la vie sauvage*, Paris, Flammarion, 1971.

TOURNIER M., *Les Météores*, Paris, Gallimard, 1975.

TOURNIER M., *Le Vent Paraclet*, Paris, Gallimard, 1975 [Folio, 1977].

TOURNIER M., *Le Coq de Bruyère*, Paris, Gallimard, 1978.

TOURNIER M., *Le Vol du vampire*, Paris, Gallimard, 1981.

TOURNIER M., *Le Vagabond immobile*, Paris, Gallimard, 1984.

TOURNIER M., *Journal extime*, Paris, La Musardine, 2002.

YOURCENAR M., *Feux*, Paris, Grasset, 1936.

YOURCENAR M, *Nouvelles orientales*, Paris, Gallimard, 1938.

YOURCENAR M., *Le coup de grâce*, Paris, Gallimard, 1939.

YOURCENAR M., *L'Œuvre au Noir*, Paris, Gallimard, 1968.

YOURCENAR M., *Souvenirs pieux*, Paris, Gallimard, 1974.

YOURCENAR M., *Archives du Nord*, Paris, Gallimard, 1977.

YOURCENAR M, *Le temps, ce grand sculpteur*, Paris, Gallimard, 1983.

Lettres, philosophie et musique

ALTHEN G., *Jean-Marie Le Clézio*, Marseille, Sud, 1990.

ANTOINE G., *Revue d'Enseignement supérieur*, I-1959, p. 49-60.

APOLLINAIRE, *Calligrammes*, Paris, Gallimard, 1925.

ARISTOTE, *La Poétique* [traduction Magnien M., Librairie Générale de France, 1990].

ARISTOTE, *La Rhétorique* [traduction Ruelle C.E., Librairie Générale de France, 1991].

BACHELARD G., *La psychanalyse du Feu*, Paris, Gallimard, 1938.

BACHELARD G., *L'Air et les songes*, Paris, Corti, 1943.

BACHELARD G., *L'Eau et les rêves*, Paris, Corti, 1942.

BACHELARD G., *La Terre et les rêveries de la volonté*, Paris, Corti, 1947.

BACHELARD G., *La Terre et les rêveries du repos*, Paris, Corti, 1948.

BALLY C., *Traité de stylistique française*, Paris, Klincksieck, 1909.

BERGSON H., *Durée et simultanéité. A propos de la théorie d'Einstein*, Paris, Alcan, 1922 [http://classiques.uqac.ca/classiques].

BONHOMME B.& SYGMINTON M, *Le rythme dans la poésie et les arts*, Paris, Champion, 2005.

BOULOUMIE A., *Michel Tournier : le roman mythologique*, Paris, Corti, 1988.

BOULOUMIE A., *Vendredi ou les limbes du Pacifique de Michel Tournier*, Paris, Gallimard, 1991.

CAPPELLO S., *Le réseau phonique et le sens : l'interaction phono-sémantique en poésie*, Bologna, Clueb, 1990.

CASTAREDE M.F, KONOPCZYNSKI G. & alii, *Au commencement était la voix*, Ramonville-Sainte-Agne, Erès, 2005.

CHANDA T., *Entretien avec Jean-Marie Le Clézio*, www.diplomatie.gouv.fr, 2001.

CHOUVEL J.M.& LEVY F., *Observation, analyse, modèle : peut-on parler d'art avec les outils de la science ?*, Paris, IRCAM, 2002.

CICERON, *Œuvres complètes* [traduction sous la direction de Nisard C., Paris, Didot, 1858].

CONDILLAC E., *Œuvres complètes*, Paris, Houel, 1798.

CROCE B., *Estetica come scienza dell'espressione e linguistica generale: teoria e storia,* Milano, Sandron, 1902.

DEFOE D, *Robinson Crusoé*, Paris, Gallimard, 1950.

DELAS D. & alii, *Rythme et écriture*, vol. I-IV, Université de Paris X, 1988-1994.

DELAS D., *Les enjeux de la stylistique*, PARIS, LAROUSSE, 1995.

DESSONS G.& MESCHONNIC H., *Traité du rythme : des vers et des proses*, Paris, Dunod, 1998.

DOMANGE S., *Le Clézio ou la quête du désert*, Paris, Imago, 1993.

DUCROT O. & SCHAEFFER J.M., *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris, Seuil, 1995.

EPINETTE-BRENGUES F., *Vendredi ou les limbes du Pacifique*, Paris, Ellipses, 1998.

EZINE J.L., *Ailleurs*, Paris, Arléa, 1995.

FONAGY I, *La Vive Voix*, *Essai de psychophonétique*, Paris, Payot, 1983.

FONTANIER P. *Les figures du discours*, 1827 [Paris, Flammarion, 1977].

FRAISSE P, *Psychologie du rythme*, Paris, Presses Universitaires de France, 1974.

GALEY M., *Marguerite Yourcenar, Les yeux ouverts*, Paris, Le Centurion, 1980.

GUILLAUME G., "*Le problème de l'article et sa solution dans la langue française*, Paris, Hachette, 1919.

GUIRAUD P. & KUENTZ P., *La Stylistique*, Paris, Klincksieck, 1970.

GUIRAUD P., *La Stylistique*, Paris, Presses Universitaires de France, 1955.

GUSLEVIC C., *Mémoires d'Hadrien*, Paris, Ellipses, 1999.

HERSCHBERG PIERROT A., *Stylistique de la prose*, Paris, Belin, 2003.

HUMBOLDT W., *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts,* Berlin, Königliche Akademie der Wissenschaften, 1836.

INSTITUT JACQUES DALCROZE, *3e Congrès International du Rythme* , Genève, Papillon, 1999.

JACOB M., *Le cornet à dés*, Paris, Gallimard, 1945.

JAKOBSON R., *Essais de Linguistique générale* [traduction Ruwet N., Paris, Editions de Minuit, 1963].

JAUSS H.R., *Pour une esthétique de la réception* [traduction Mailard C., Paris, Gallimard, 1978].

JENNY L., *La parole singulière*, Paris, Belin, 1990.

JOLLIN-BERTOCHI S. & THIBAULT B., *J.M.G. Le Clézio*, Nantes, Editions du temps, 2004.

JOUSSET P., *Anthropologie du style*, Presses Universitaires de Bordeaux, 2008.

JULIEN A.Y., *Marguerite Yourcenar*, Paris, Presses Universitaires de France, 2002.

KARABETIAN E., *Histoire des stylistiques*, Paris, Armand Colin, 2000.

LABOV W., *Principles of linguistic change*, Oxford, Blackwell, 1927.

LARTHOMAS P., *Le Français moderne*, Paris, Artrey, 1964.

LEIBNIZ G.W., *Opera omnia*, Genève, Dutens, 1768.

LEVILLAIN H., *Mémoires d'Hadrien de Marguerite Yourcenar*, Paris, Gallimard, 1992.

LEVI-STRAUSS, C., *L'homme nu*, Paris, Plon, 1971.

LHOSTE P., *Conversations avec J.M.G. Le Clézio*, Paris, Mercure de France, 1971.

LUK F.L., *Michel Tournier et le détournement de l'autobiographie,* Editions Universitaires de Dijon, 2003.

MERLEAU-PONTY M., *Phénoménologie de la perception*, Paris, Gallimard 1945 [Gallimard, 1999].

MERLEAU-PONTY M., *Signes*, Paris, Gallimard, 1960.

MERLEAU-PONTY M., *La prose du monde*, Paris, Gallimard 1969.

MERLLIE F., *Michel Tournier*, Paris, Belfond, 1988.

MESCHONNIC H. & alii, *Le rythme*, Colloque d'Albi, Ecole normale, 1983.

MESCHONNIC H., *Critique du rythme*, Lagrasse, Verdier, 1982.

MICHELS U., *Guide illustré de la musique,* 1977 [traduction Gribenski J., Leothaud G., Paris, Fayard, 1988].

MOLINIE, G., *La stylistique*, Paris, Presses Universitaires de France 1989.

MOLINIE G. & CAHNE P., *Qu'est-ce que le style ?*, Paris, Presses Universitaires de France, 1994.

NIETZSCHE F., *Die Geburt der Tragödie*, Leipzig, Fritzsch, 1872.

NOVALIS, *Œuvres complètes* [traduction Guerne A., Paris, Gallimard, 1975].

ONIMUS J, *Pour lire Le Clézio*, Paris, Presses Universitaires de France, 1994.

OUAKNIN M.A., *Les mystères de l'alphabet*, Paris, Assouline, 1997.

PAPUS, *Le Sepher Jesirah*, Paris, Cariscript, 1987.

PEZECHKIAN-WEINBERG P., *Michel Tournier, Marginalité et création*, Bern, Peter Lang, 1988.

PLATON, *Lois* [traduction Robin L., Paris, Gallimard, 1950].

RAMBURES J.L., *Comment travaillent les écrivains*, Paris, Flammarion, 1978.

RASTIER F., *Arts et sciences du texte*, Paris, Presses Universitaires de France, 2001.

RIFFATERRE M., *Essais de stylistique structurale* [traduction Delas D., Paris, Flammarion, 1970].

SALLES M., *Désert*, Paris, Ellipses, 1999.

SARRAUTE N., *L'ère du soupçon*, Paris, Gallimard, 1956.

SATPREM, *Le mental des cellules*, Paris, Laffont, 1981.

SAUSSURE F., *Cours de linguistique générale*, Paris, Payot, 1916.

SAUVANET P. *Le rythme grec*, Paris, Presses Universitaires de France, 1999.

SCHLEIERMACHER F., *Sämmtliche Werke*, Berlin, Reimer, 1834-1864.

SCHLOCKER G., *Equilibre & Symétrie dans la Phrase française moderne*, Paris, Klincksieck, 1957.

SOULAGE M., *Le solfège*, Paris, Presses Universitaires de France, 1969.

SPINOZA B., *L'Ethique*, 1677 [traduction Misrahi R., Paris, Presses Universitaires de France, 1990].

SPITZER L., *Etudes de style*, Paris, Gallimard, 1970.

SZAMBIEN W. *Symétrie, Goût, Caractère*, Paris, Picard, 1986.

TADIE J.Y., *Le récit poétique*, Paris, Presses Universitaires de France, 1978 [Paris, Gallimard, 1994].

TODOROV T., *Théories du symbole*, Paris, Seuil, 1977.

TROUVE A., *Leçon littéraire sur Mémoires d'Hadrien de Marguerite Yourcenar*, Paris, Presses Universitaires de France, 1996.

VIERNE S., *Rite, roman, initiation*, Presses Universitaires de Grenoble, 1973.

WUNDT W., *Essays*, Leipzig, Engelman, 1885.

Linguistique statistique

AZAR M. & KEDEM B., « Some Time Series in the Phonetics of Biblical Hebrew », *Association of the Literary and Linguistic Computing*, 1979, vol. 7, n° 2, p. 111-129.

BENZECRI J.P., *Pratique de l'analyse des données, Linguistique et lexicologie*, Paris, Dunod, 1981.

BENZECRI J.P., *Histoire et préhistoire de l'analyse des données*, Paris, Dunod, 1982.

BEAUDOUIN V. & VVON F., « *Contribution de la métrique à la stylométrie* » , *Journées internationales d'Analyse statistique des Données Textuelles*, 2004.

BORODA M.G., « Complexity oscillations in a coherent text : towards the rhythmic foundations of text organization », *Journal of quantitative linguistics*, 1994, vol. 1, p. 87-97.

BRATLEY P. & ROSS D., « Syllabic Spectra », *Association of the Literary and Linguistic Computing*, 1981, vol. 2, n° 2, p. 41-50.

BRUNET E, *Le vocabulaire français de 1789 à nos jours,* Paris, Champion, 1981.

BRUNET E., *Méthodes quantitatives et informatiques dans l'étude des textes*, Paris, Champion, 1986.

BRUNET E., « Qui lemmatise dilemme attise », *Lexicométrica*, 2000.

CLEMENT R. & SHARP D., « Ngram and Bayesian Classification of Documents for Topic and Authorship », *Literary and Linguistic Computing*, 2003, vol. 18, p. 423-447.

CORDUAS M., « La struttura dinamica dei dati testuali », *Journées internationales d'Analyse Statistique des Données Textuelles*, 1995, p. 345-352.

CZELLAR J., « Attribution d'auteur : application des méthodes de qsums au français », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2006.

DOLEZEL L., *Statistics and Style*, New-York, Elsevier, 1969.

DREHER J., YOUNG E., NORTON R. & MA J. « Power spectral densities of literary speech rhythms », *Computer Studies in the humanities and verbal behavior*, 1969, vol. 2, p. 170-191.

GARRETTE R., *La phrase dans l'œuvre dramatique de Racine : étude stylistique et stylométrique*, Université de Toulouse, Thèse, 1988.

GAUDARD F.C., *Contribution à l'analyse des discours littéraires : exploration stylistique de l'espace poétique Baudelairien,* Université de Toulouse, Thèse, 1989.

HARRIS Z., *Mathematical Structures of Language*, New York, Interscience, 1968.

HARRIS Z., *A theory of Language and Information : a Mathematical approach*, Oxford, Clarendon Press, 1991.

HERDAN G., *Language as Choice and Chance*, Den Haag, Mouton, 1956.

HJORT N.L., « And Quiet Does Not Flow the Don : Statistical Analysis of a Quarrel between Nobel Laureates », Centre for Advanced Study at the Norwegian Academy of Sciences and Letters, *Consilience*, Interdisciplinary Communications 2005/2006, p. 134-140.

HOLMES D.I., « The Federalist revisited : New Directions in Authorship Attribution », *Literary and Linguistic Computing*, 1995, vol. 10, n° 2, p. 111-127.

HOLMES D.I., « The Evolution of Stylometry in Humanities Scholarship», *Literary and Linguistic Computing*, 1998, vol. 13, n° 3, p. 117-117.

JAKOBSON R. & alii, *Structure of Language and Its Mathematical Aspects*, New-York, American Mathematical Society, 1961.

JARDINO M., « Identification des auteurs de textes courts avec des n-grammes de caractères », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2006.

JULESZ B., « Visual pattern discrimination », *IRE Transactions on Information Theory*, 1962, vol.8, n° 2, p. 84-92.

KASTBERG M., *L'écriture de J.M.G. Le Clézio, une approche*

lexicométrique, Université de Nice, Thèse, 2002.

KHMELEV D. & TWEEDIE F.J., « Using Markov Chains for Identification of Writers », *Literary and Linguistics Computing*, 2001, vol. 16, n° 4, p. 299-307.

LABBE C & LABBE D., « Intertextual Distance and Authorship Attribution : Corneille and Molière », *Journal of Quantitative Linguistics*, 2001, 8-3 : 213-231.

LAMALLE C. & SALEM A., « Types généralisés et topographie textuelle dans l'analyse d'un corpus lemmatisé », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2002.

LEBART L. & SALEM A., *Statistique textuelle*, Paris, Dunod, 1994.

LEBART L., PIRON M. & STEINER J.F., *La sémiométrie*, Paris, Dunod, 2003.

LELU, A., « Clusters and factors : neural algorithms for a novel representation of huge and highly multidimensional data sets » (in Diday E., Lechevallier Y. & alii (eds.), *New Approaches in Classification and Data Analysis*, Berlin, Springer-Verlag, 1994, p. 241-248).

LONGREE D., LUONG X. & MELLET S., « Temps Verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2004.

LONGREE D., LUONG X. & MELLET S., « Distance intertextuelle et classement des textes d'après leur structure : méthode de découpage et analyses arborées », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2006.

LUONG X & alii, « La distance intertextuelle », *Corpus*, n° 2, 2003.

MANDELBROT B., « On the Theory of Word frequencies and on Related Markovian Models of Discourse », *Proceedings of Symposia in Applied Mathematics*, 1961, vol. 12, p. 190-219.

MANDELBROT B., *Langage, logique et théorie de l'information*, Paris, Presses Universitaires de France, 1957.

MARKOV A., « Primer statisticeskogo issledovanija nad tekstom "Evgenija Onegina", illjustrirujuscii svaz ispytanii v cep », *Bulletin de*

l'Académie Impériale des Sciences, 1913, p 153-162.

MAYAFFRE D., *Paroles de président*, Paris, Champion, 2004.

MEALAND D.L., « Measuring Genre Differences in Mark with Correspondence Analysis », *Literary and Linguistic Computing*, 1995, vol. 12, p. 227-245.

MENDENHALL T.C., « The characteristic curves of composition », *Science*, 1887, vol. 11, p. 237-249.

MOLES A., *Théorie de l'information et perception esthétique*, Paris, Flammarion, 1958.

MORTON A.Q., *Literary Detection : How to Prove Authorship and Fraud in Literature and Documents*, Epping, Bowker, 1978.

MÜLLER C., *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, 1967.

MÜLLER C., *Principes et méthodes de statistique lexicale,* Paris, Champion, 1977.

PAGET R. & LONGSTAFF I.D., « Texture synthesis via a Non-parametric Markov Random Field » , ETH Zürich, *Dicta'95*.

PAWLOWSKI A., *Séries temporelles en linguistique*, Paris, Champion, 1998.

PETRUSZEWYCZ M., *Les chaînes de Markov dans le domaine linguistique*, Genève, Slatkine, 1981.

REINERT M., « Un logiciel d'analyse lexicale [ALCESTE] », *Cahiers de l'Analyse des Données*, 1985-4, p. 471-484.

SCHMID H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, 1994, p. 44-49.

SHANNON C.E., « Prediction and Entropy of Printed English », *Bell Systems Technical Journal*, 1951, vol. 30, p. 50-64.

SICHEL H.S., « On a Distribution Law for Word Frequencies », *Journal of the American Statistical Association*, 1975, vol. 70, p. 542-547.

SMITH J.B. & ROSENBERG B.A., « Rhythms in Speech : the Formulaic Structure of Four Fundamentalist Sermons », *Computer Studies in the humanities and verbal behavior*, 1973, vol. 4, p. 166-173.

VIPREY J.M., « Analyse séquencée de la micro-distribution lexicale », *Journées internationales d'Analyse Statistique des Données Textuelles*, 2004.

WORONCZAC J., « Statistische Methoden in der Verslehre », *Poetics*, 1961, vol. I (A16), p. 607-624.

YULE G.U., « On sentence-length as a statistical analysis of style in prose, with application to two cases of disputed authorship », *Biometrika*, 1938, vol. 30, p. 363-390.

YULE G.U., *The statistical study of literary vocabulary*, Cambridge University Press, 1944.

ZIPF G.K., « Selected studies of the Principle of Relative Frequency in Language », Harvard University Press, *Language*, 1932, vol. 9, n° 1, p. 89 -92.

Sciences naturelles et formelles

ADANSON M., *Histoire naturelle du Sénégal*, Paris, Bauche, 1757.

ANDERSON T.W., « On the distribution of the two-sample Cramer-Von Mises criterion », *The Annals of Mathematical Statistics*, 1962, vol. 33, n° 3, p. 1148-1159.

ANDERSON T.W., *The Statistical Analysis of Time Series*, New York, Wiley, 1970.

BACRY H., *La symétrie dans tous ses états,* Paris, Vuibert, 2000.

BARTLETT M.S, « On the Theoretical Specification of Sampling Properties of Autocorrelated Time-Series », *Journal of the Royal Statistical Society*, 1946, B8, p. 27-41.

BAYES T., « An essay towards solving a Problem in the Doctrine of Chances », *Philosophical Transactions of the Royal Society of London*, vol. 53, p. 370-418, 1763 [www.stat.ucla.edu/history/essay.pdf]

BENZECRI J.P., *L'analyse des données*, Paris, Dunod, 1973.

BENZECRI J.P., « Histoire et préhistoire de l'analyse des données », *Les cahiers de l'analyse des données*, 1976, n° 1-2.

BERNOULLI J., *Ars conjectandi,* Basileae, 1713.

BIRRIEN J.Y, *Histoire de l'informatique*, Paris, Presses Universitaires de France, 1990.

BOURBONNAIS R. & TERRAZA M., *Analyses des séries temporelles en économie*, Paris, Presses Universitaires de France, 1998.

BOX G., JENKINS G. & REINSEL G., *Time series analysis : forecasting and control*, San Francisco, Holden-Day, 1970.

BRETON P., *Une histoire de l'informatique*, Paris, La Découverte, 1987.

CLAUSIUS R., *Abhandlungen über die mechanische Wärmetheorie*, Braunschweig, Vieweg, 1864.

CLOSE F., *Asymétrie, la beauté du diable*, Les Ulys, EDP Sciences, 2001.

CORAZZA M., *Techniques mathématiques de la fiabilité prévisionnelle*, Toulouse, Cépaduès, 1975.

DEWEY, M., *Classification and subject index for a library,* Amherst, Mass, 1876.

DRAPER H., *The Draper Catalogue*, Harvard Observatory, 1890.

DROESBEKE J.J & TASSI P., *Histoire de la statistique*, Paris, Presses Universitaires de France, 1997.

DURBIN J., « Efficient Estimation of Parameters of Moving Average Models », *Biometrika*, 1959, vol. 49, p. 306-316.

EINSTEIN A., *La théorie de la relativité restreinte et générale,* 1916 [traduction Solovine M., Paris, Dunod, 1999].

FISHER R.A., « Theory of Statistical Estimation », *Proceedings of the Cambridge Philosophical Society*, 1925, vol. 22, p. 700-725.

GALLOIS E., *Oeuvres mathématiques*, *Journal de mathématiques pures*

et appliquées, 1846, tome XI, p. 381-444.

GOURIEROUX C. & MONFORT A., *Séries temporelles et modèles dynamiques*, Paris, Economica, 1995.

GRAUNT J., *Nature and Political Observations upon the Bills of Mortality*, London, Roycroft, 1662.

HOARE C.A.R., « Quicksort », *Computer Journal*, 1962, vol. 5, p. 10-15.

HOLLERITH H., *In connection with the electric tabulation system which has been adopted by U.S. government for the work of the census bureau,* Columbia University School of Mines, Thèse, 1890.

KOLMOGOROV A., *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, Springer, 1933.

KLEIN F., *Vergleichende Betrachtungen über neuere geometrische Forschungen,* Erlangen, Deichert, 1872).

LAPLACE P.S, *Théorie analytique des probabilités*, Paris, Courcier, 1812 [Sceaux, Gabay, 1995].

LINNE C., *Systema Naturae*, Leyden, Haak, 1735.

MAC QUEEN J.B., « Some Methods for classification and Analysis of Multivariate Observations », *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. , p. 281-297.

MAIRESSSE J. & alii, *Pour une histoire de la statistique*, Paris, INSEE, 1987.

MANDELBROT B., *Les objets fractals : forme, hasard et dimension*, Paris, Flammarion, 1975.

NAPIER J., *Mirifici logarithmorum canonis descriptio*, Lyon, Vincentium, 1620.

NEUMANN J., « First Draft of a Report on the EDVAC» , University of Pennsylvania, 1945 [*IEEE Annals of the History of Computing*, 1993, vol. 15, n° 4, p. 27-75].

NEWTON I., *Philosophiae naturalis principia mathematica*, London, Pepys, 1687.

NEYMAN J., « Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability », *Philosophical Transactions of the Royal Society of London, 1937,* Series A, vol. 236, n° 767, p. 333–380.

PAC J.L., *Processus aléatoires*, Toulouse, Sup'Aéro, 1985.

PEARSON K, « Contributions to the Mathematical Theory of Evolution », *Proceedings of the Royal Society of London*, 1893, vol. 54, p. 329-333.

PEARSON K., « On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling », *Philosophical magazine*, 1900, vol. 50, n° 302, p. 157-176.

PEARSON K., « On Lines and Planes of Closest Fit to Systems of Points in Space », *Philosophical Magazine*, 1901, vol. 2, n° 6, p. 559–572.

PEIRCE C.S., *Writings of Charles S. Peirce,* Indiana University Press, 1993.

PETTY W., *Political arithmetic*, London, Clavel & Mortlock, 1690.

PLAYFAIR W., *The commercial and political atlas and Statistical breviary,* Cambridge University Press, 2005.

RENYI A., *Calcul des probabilités*, 1962 [traduction Bloch C., Paris, Dunod, 1966].

QUENOUILLE M.H. , « The joint distribution of serial correlation coefficients », *Annals of Mathematical Statistics*, 1949, 10.

SAMUELIDES M. & TOUZILLIER L., *Analyse harmonique*, Toulouse, Sup'Aéro, 1984.

SAPORTA G, *Probabilités, Analyse de données et Statistiques*, Paris, Technip, 1990.

SHANNON C. *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, 1949.

SLUTZKY E., « The Summation of Random Causes as the Source of Cyclic Process », *The Conjuncture Institute*, 1927, vol. 3, n° 1.

TURING A., « On Computable Numbers, With an Application to the Entscheidungsproblem », *Proceedings of the London Mathematical*

Society, 1937, series 2, vol. 42, p. 230-265.

VAN FRAASEN B.C., *Lois et symétrie*, Paris, Vrin, 1994.

WEYL H., *Symétrie et mathématique moderne*, Paris, Flammarion, 1964.

WIENER N., *Cybernetics, or control and communication in the animal and the machine*, The MIT Press, 1948.

WILLIAMS R., *The Mac Is Not a Typewriter*, Berkeley, Peachpit Press, 2003.

WOLD H., *A study in the analysis of stationary time series*, University of Stockholm, Thèse, 1938.

XIAO Y., GORDON A. & YAKOVLEV A., « The L1-Version of the Cramér-von Mises Test for Two-Sample Comparisons in Microarray Data Analysis », *EURASIP Journal on Bioinformatics and Systems Biology,* 2006, Article ID 85769.

YULE G.U., « On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers » *Philosophical Transactions*, 1927, vol. 226, p. 267-298.

Revues et colloques en France

CHAMPS DU SIGNE : w3.lla.univ-tlse2.fr/champsdusigne

CORPUS : revel.unice.fr/corpus

JOURNEES D'ANALYSE STATISTIQUE DES DONNEES TEXTUELLES : www.jadt.org

LEXICOMETRICA : www.cavi.univ-paris3.fr/lexicometrica

TEXTO : www.revue-texto.net

Logiciels et bases de données

ALCESTE : Société Image, www.image-zafar.com

BDLEX : Université de Toulouse, www.irit.fr

CORDIAL-ANALYSEUR : Société Synapse, www.synapse-fr.com

HYPERBASE : Université de Nice, www.unice.fr/bcl

NEURONA*V* : Société Diatopie, www.diatopie.com

NOMINO : Université de Québec, www.ling.uqam.ca/ato

S-PLUS : Société Insightful, www.insightful.com

SYNTEX : Université de Toulouse Le Mirail, www.univ-tlse2.fr/erss

TREETAGGER : Université de Stuttgart, www.ims.uni-stuttgart.de

TROPES-ZOOM : Société Acétic, www.acetic.fr

Lexique

La liste rencense principalement des notions scientifiques, ainsi que des termes linguistiques employés dans un sens atypique.

ARYTHMIE
[Statistique] : traduit la dispersion d'une variable autour de sa valeur moyenne temporelle. Afin de s'affranchir des effets d'échelle, elle se mesure par un écart relatif.

ASYMETRIE
[Statistique] : moment réduit d'ordre 3 de l'arythmie. Il diffère quelque peu du coefficient d'asymétrie usuel afin de rendre sa définition homogène à celle de la variabilité, mais traduit le même gauchissement de la distribution sous-jacente.

CONDITIONNEMENT
[Statistique] : probabilité qu'une variable prenne une valeur en tenant compte d'un évènement connu.

CONTINU
[Topologie] : par exemple, l'ensemble « plein « des nombres réels : la distance entre deux éléments voisins peut être rendue aussi petite que voulue.

CORRELATION
[Statistique] : estime le lien entre les observations de deux variables, sans préjuger de leur causalité. Précisément, la corrélation mesure la dépendance linéaire de ces observations.

DIMENSION
[Géométrie] : degré de liberté d'un espace : un pour une droite, deux pour un plan, trois pour notre monde physique. Les mathématiques ont imaginé des lieux exotiques aux dimensions élevées, infinies, voire non entières avec les fractales.

DISCRET
[Topologie] : par exemple, l'ensemble « troué » des nombres entiers : la distance entre deux voisins est toujours égale à 1.

DISTANCE
[Géométrie] : dissimilarité symétrique vérifiant l'inégalité triangulaire. La plus connue est euclidienne.

ECART RELATIF
[Statistique] : si une variable X a pour moyenne m, l'écart relatif est donné par X/m -1.

ENTROPIE
[Etymologie] : du grec εν–τροπειν, « tourner dedans ».
[Thermodynamique] : mesure le désordre dans la matière.
[Théorie de l'information] : quantité d'information véhiculée par un signal aléatoire.

DISTRIBUTION
[Statistique] : sa courbe représente la probabilité qu'une variable prenne une valeur, en fonction de cette valeur.

ERGODICITE
[Statistique] : décroissance rapide des corrélations temporelles.

ESPACE
[Géométrie] : ensemble formel d'objets mathématiques. Dans cette acception, le temps est un espace.

INDEPENDANCE
[Statistique] : absence de lien entre les observations de deux variables. Cette notion est plus forte que la décorrélation, qui exclut uniquement les liens linéaires.

INFERENCE
[Statistique] : partant d'un échantillon, l'inférence fait l'hypothèse d'une population qui l'engendre. Elle dépasse donc la simple description des données.

LOI DES GRANDS NOMBRES
[Statistique] : fondement de la pratique des sondages, la loi exprime que les statistiques d'un échantillon convergent vers celles de la population totale.

LOI NORMALE
[Statistique] : mise en évidence par Gauss, cette loi permet de modéliser un grand nombre de phénomènes. Sa distribution a la forme d'une courbe en cloche symétrique.

MESURE
[Musique] : division du temps en sections d'égales durée.

MODE
[Statistique] : classe dominante d'une distribution.

MOMENT
[Statistique] : le moment d'ordre 1 d'une variable X est sa moyenne m ; sa version centrée est celui de la variable X-m.
Le moment d'ordre k est celui de la variable X^k ; sa version réduite le ramène à la puissance 1/k.

PROCESSUS STOCHASTIQUE
[Etymologie] : du grec στοχος, « but » ou « conjecture ».
[Statistique] : modélise des observations successives, en mêlant les principes de causalité et d'incertitude : le futur est partiellement déterminé par le passé.

REPARTITION
[Statistique] : la répartition intègre la distribution. Sa courbe représente la probabilité qu'une variable reste inférieure à une valeur, en fonction de cette valeur.

RYTHME
[Etymologie] : du grec ρυθμος, « mouvement réglé et mesuré ».
[Musique] : son périodique.
[Analyse harmonique] : tout signal, périodique ou non, est représentable dans l'espace des fréquences à l'aide d'une transformation de Fourier, d'où un rythme généralisé.
[Statistique] : un rythme dégradé peut se définir à partir de la distribution des temps de retour, en faisant abstraction de leur succession.

SIMILITUDE
[Géométrie] : transformation spatiale conservant la forme d'un objet. Par exemple, la translation, la rotation, l'homothétie ou la symétrie.

SPECTRE
[Physique] : composition de la lumière blanche en fonction des couleurs, et par extension distribution spatiale d'une variable temporelle.

STATIONNARITE
[Statistique] : un processus stochastique est stationnaire si sa loi ne change pas avec le temps. Au sens faible, il suffit que les moments

statistiques vérifient cette propriété.

STRUCTURE DE GROUPE
[Algèbre] : ensemble stable d'éléments entretenant des relations d'opposition symétriques.

STYLE
[Etymologie] : du grec στυλος, « colonne », puis « le burin pour écrire ».
[Linguistique] : organisation des unités dans un texte. Cette description complète le thème, qui fixe la composition.

STYLOMETRIE
[Linguistique] : mesure du style à l'aide de méthodes statistiques. Les observations peuvent s'interpréter en termes littéraires ou probabilistes.

SYMETRIE
[Etymologie] : du grec συμ–μετρος, « de même mesure ».
[Géométrie] : cette notion s'est infiltrée dans de nombreux domaines de la science à l'art, et pourrait même fonder notre pensée. Par rapport à son opposé, elle possède un avantage décisif, la simplicité.

TEMPO
[Musique] : durée d'une mesure.

THEME
[Etymologie] : du grec θεμα, « ce qui est posé ».
[Linguistique] : composition d'un texte selon ses unités. Cette définition étend à l'ensemble des plans linguistiques l'acception commune, plutôt restreinte au niveau sémantique.

VARIABILITE
[Statistique] : moment réduit d'ordre 2 de l'arythmie.

Index nominum

Index rerum

M

N

O

P

R

S

T

V

Table des matières

Seconde partie : observations

Erratum

— J'aurais bien fait un errata pour les fautes qu'une impression achevée en hâte a laissées dans mon livre ; mais — qui est-ce qui lit un errata ? — personne.

Balzac, *Les Chouans*, introduction de la première édition, 1829.

Je ne suis ni de l'est ni de l'ouest**,**
ni de la mer ni de la terre,
je ne suis ni matériel ni éthéré,
ni composé d'éléments.
Je n'existe pas
je ne suis une part de ce monde ni d'un autre
je ne descends ni d'Adam ni d'Eve
ni d'aucune origine.
Ma place n'a pas de place,
une trace de ce qui n'a pas de trace
ni corps, ni âme.
J'appartiens au Bien-Aimé
j'ai vu les deux mondes réunis en un seul
le premier ; le dernier, celui du dehors, celui du dedans,
simples comme le souffle d'un homme qui respire.

Rumi, *Mathnawi*, Livre premier