# Bouvard : 618, Pécuchet : 598. Etude de statistique textuelle.

*Yvan Leclerc*
*Université de Rouen*
*Centre Flaubert*
*Yvan.Leclerc@univ-rouen.fr*

*Pierre Cubaud*
*Conservatoire national des arts & métiers, Paris*
*Centre d'études et de recherche en informatique (CEDRIC)*
*cubaud@cnam.fr*

Raymond Queneau était bien placé pour observer que dans l'encyclopédie de Bouvard et Pécuchet, manquaient les mathématiques. Est-ce parce que Flaubert ne s'y connaissait pas, et avait renoncé à l'étudier pour ses «□onshommes » ? Ou bien parce que cette science pure, décidément trop abstraite, ne se prêtait pas à une représentation plastique, en échappant au «□omique des idées » ? Ou encore que les erreurs et les bêtises, dans cette science exacte, n'auraient été imputables qu'aux seuls personnages, sans mettre en cause la validité du savoir lui-même, selon la double cible toujours visée dans le roman ?

L'un des deux pourtant possède l'instrument de la science, et paraît savoir s'en servir, au moment où il s'agit d'arranger le jardin : «□écuchet fit plusieurs épures, en se servant de sa boîte de mathématiques□ (II §307[1]). La boîte de mathématiques va nous servir à compter, non les objets du monde, mais les maîtres-mots du roman, ceux qui figurent dans le titre.

Nous allons utiliser pour cela la transcription numérique du texte effectuée par Eric Dubreucq et diffusée sur le site de l'ABU (http://abu.cnam.fr). Cette transcription semble différer sur certains points (ponctuation, inversions, points d'exclamations supprimés ou ajoutés) de l'édition Charpentier de 1891 numérisée en mode image sur Gallica (http://gallica.bnf.fr), mais nous l'avons jugée de qualité suffisante pour l'extraction et le comptage des occurrences des deux mots **Bouvard** et **Pécuchet**.

La transcription se termine à la fin du chapitre X. Nous la rediffusons ici intégralement, découpée par chapitres et en signalant au long du texte :
- un numéro d'ordre des paragraphes, utilisé dans notre étude pour les références des citations
- les occurrences de **Bouvard** (en bleu) et de **Pécuchet** (en rouge)
- leur numéro d'ordre
- la différence courante entre le nombre cumulé d'occurrences **Bouvard** et celui de **Pécuchet** (notée nB-nP)
- la somme de ces deux nombres (nB+nP)

[1] Le numéro du paragraphe renvoie au texte colorisé de *Bouvard et Pécuchet*, accessible à partir de la version HTML du présent travail.

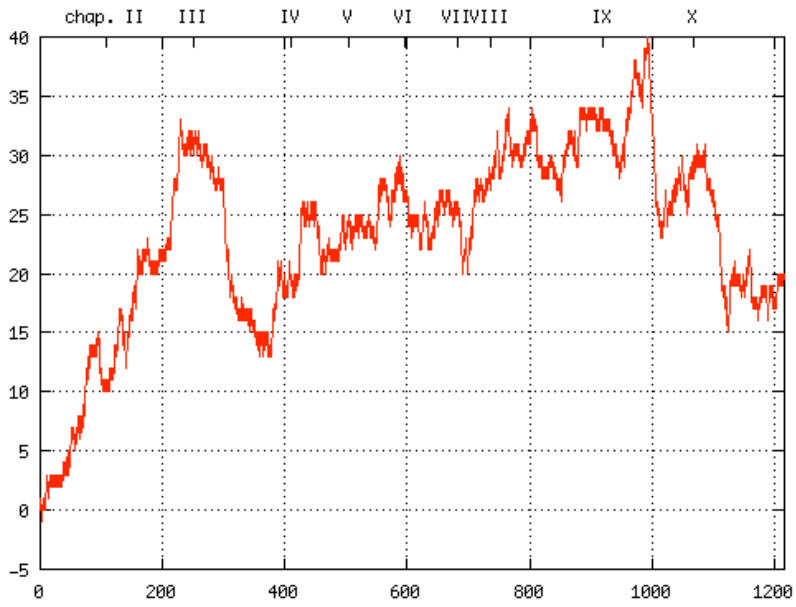
Le tableau ci-dessous résume les comptages obtenus chapitre par chapitre :

| Chapitre | nB | nP | paragraphes | mots |
|---|---|---|---|---|
| I | 60 | 50 | 145 | 6000 |
| II | 81 | 60 | 292 | 11503 |
| III | 74 | 86 | 431 | 12277 |
| IV | 50 | 45 | 352 | 9786 |
| V | 46 | 44 | 266 | 6227 |
| VI | 43 | 44 | 360 | 8309 |
| VII | 28 | 25 | 135 | 3251 |
| VIII | 94 | 89 | 506 | 13928 |
| IX | 73 | 77 | 410 | 11461 |
| X | 69 | 78 | 387 | 10531 |
| total | 618 | 598 | | |

On peut extraire du texte du roman les occurrences de **Bouvard** et **Pécuchet** pour en produire la séquence, notée dans ce qui suivra {BP}. De ce point de vue, le texte de Flaubert se résume à :

```
BPBPPPBBPBPBBBPPBBPBPBPBPBPBPBPBPBPBPBPBBPBPBPBBPPBBBBPBPBPPBBPBBPBPPBBPBBPBBBBPBBPBBPBPB
PBPBPBBPBPPPPBPPBPBPBPBPBPBPBBPBPBPBBBPBPBBBBPBPBPPPBPPPBBBPBBBPBPBBPBBBPBBBBPPBPBPBBPBPB
BPPBPPPBPBPBPBPBPBBPBPBPBPBPBPBPBPBPBPBBBBBBPBPBPBBBBBBPPBPPBPBPBPBBPPBBPBPPBBPBPPBBPBPB
PBPBBPBPBPPBPBPPBBPPBPPBPBPBBPPBPBPBPPPPPPPBPPPPBBPPBPPBPBPPBPBPBPBBPPBPBPBPBPPBBPBPPBPB
PPBPBPPBBPBPBPPBBPBPBPPBPBPBBBBPBBPBBBBPPBBPPPBBPPBPBPBBBPPBPPBBPPBBPBPBBBBBBPBPPBPBBPP
BPBBPBPPBBPBPPPBPPPPBPBBPBPPBBPBBPPBPBPBPBPBPBPBPBPBBPBBPBPPPBBPBBPBPPPBBPBPBBPBPPBBPBPB
PBPBPBPBBPBPPBPBPBPPBPBBBBBBPPBBPBPBPBPPBPPPBPBBBPBPBBBPPBBBPPBBPPBPBPPPBPBPPBBPBPPBPBPB
BPBPPPBPBBBPBBPPBPPPBPBPBBPBPBBBPPBBPBBPBPBPPBPBBPBPBPPBPPBBPBPBPPBPPBBPBPBPBPPBPPPPBBPBBPPPBBBP
BBPBBBBPBBPBPPBBPBPPBPBPBBPBPBBPPBBPBPBBPBBBPPPPBPBBPBBPBBBPBBPPPPBPPBBPBPBPBPBPPBPBPBPBBP
BPBBPBPBPBBPBBPPBPBPPPPBPBPBPPBPBPBPBPBPBPBBPBPBPPBPBPPBPBPBPBPPBBBBPBPBBPBPBPPBPPBBPPBPB
BBBBPBPBPBPBPPBBPBPBPBPBPBPBPPBPBPBBPBPBPPBPBPBPBPPBPBPPBPBPPBPBPPBPBPPBPBPPBBPBPPBBBBPBBPBPBBB
PBBBPBPPBPBPPBPPBBBBBPBPBBPPPPPPPPPPPPPPPBPPBPPBPBBPBBPPPBBPBPBPBBPBPPBBPBPPBBBPPBPBPPBBPPP
BPPBBBBPBPBBPPBBPBBPPBPBPBPBPBBPPPPBPBPBPPBPBPPBPBPPBPPPPPPPBPBPPBPPPBBBBBPBPBBPPBPBPBPB
PPBPBBPPBBPBBPBBPPPPPBPBPBPBPPBBPBBPBPBPBPBPPPBBBPBPBPPBPBPBBBPBPBPBPBPB
```

Enfin, la courbe ci-dessous représente l'évolution chronologique de la séquence {BP}. Nous indiquons en abscisse le numéro d'ordre de l'occurrence B ou P (en bas) et les divisions de chapitres (en haut) . En ordonnée, on trouve la différence nB-nP – très brièvement négative au début.



**Tentative d'interprétation globale**

En regardant la courbe et les statistiques globales, chapitre par chapitre, on est amené à faire plusieurs constations simples :

— Aucun chapitre ne présente une égalité absolue. Comme si quelque chose contrariait la parité, l'équilibre. Dans les chapitres du milieu toutefois, le chapitre V et le chapitre VI, on frôle l'identité numérique, à une ou deux unités près. Il y est question de Littérature et de Politique. Les goûts artistiques et les positions idéologiques entre les deux personnages sont bien tranchés, mais aucun ne l'emporte sur l'autre. On pourrait interpréter ce point d'équilibre comme l'expression positive d'une exigence d'impartialité, ou d'égalisation des contraires, dans des domaines particulièrement «sensibles» (ceux du jugement, de l'opinion).

— La domination générale de **Bouvard** dans l'ensemble du roman se retrouve au niveau des chapitres, pris séparément : **Bouvard** est majoritaire dans six chapitres sur dix. Quatre font donc exception : les chapitres III, VI, IX et X. Dans le chapitre VI (la Politique), le plus proche de l'égalité à un point près, l'écart n'est pas vraiment significatif. Il faudra en revanche l'interpréter dans les trois autres cas, représentés sur la courbe par des minima vers nB + nP = 300, 1000 et 1100. Est-ce lié à la matière étudiée ? A la part respective que prennent les deux protagonistes dans la conduite de l'action, dans l'exploration du savoir ?

Le chapitre III est celui des sciences. Mais il est constitué de multiples sous-séquences : chimie, anatomie, physiologie, médecine, hygiène, astronomie, histoire naturelle, zoologie, géologie. Il convient donc d'interroger plus précisément la distribution B / P séquence par séquence pour établir si la domination de P (86 à 74) est constante ou localisée. En fait, elle se concentre principalement sur la séquence de la médecine pratique, lorsqu'il s'agit de soigner concrètement les Chavignollais : «Une fièvre typhoïde se répandit aux environs : Bouvard déclara qu'il ne s'en mêlerait pas. [...] / Pécuchet se dévoua.» (III § 574 et 575). En conséquence, celui-ci se trouve seul au chevet de leur fermier Gouy : on compte 7 **Pécuchet** à la suite (III § 575 à 604) pendant la consultation, et 11 **Pécuchet** pour 2 **Bouvard** si l'on poursuit jusqu'à l'abandon de la médecine (III § 630). Le premier creux de la courbe, vers l'occurrence 300 (III §635), marque donc, dans l'ensemble du chapitre sur les sciences, la pratique de la médecine. Ce n'est pas que Pécuchet ait l'esprit plus scientifique que Bouvard, malgré «son crâne élevé» (I § 16), signe d'intellectualité et de goût de l'abstraction d'après la phrénologie.

Le chapitre IX traite de la religion. Pécuchet serait-il plus religieux que Bouvard (77 contre 73) ? On pourrait le penser, puisqu'il se montre «désireux de la perfection» (IX § 2497) au point de recourir à la flagellation, qu'il fait preuve de plus de zèle : «La dévotion de Pécuchet s'était développée» (IX § 2605). Malgré ce déséquilibre d'investissement religieux, pourrait-on dire, l'écart quantitatif en faveur de Bouvard croît au début du chapitre (passant de 33 à 39) jusqu'à la dispute avec le curé Jeufroy. Comme auprès du fermier Gouy au chapitre III, Pécuchet est alors seul sous le parapluie en face de l'ecclésiastique : son nom est inscrit 13 fois à la suite (IX § 2651 à 2713), accumulation traduite dans la courbe par le creux vers nB + nP = 1000 (en fait, 1015).

De ce qui précède, on peut tirer deux conclusions : 1) dans les deux chapitres III et IX, **Pécuchet** ne devance au total **Bouvard** qu'en tirant bénéfice d'une séquence où il se trouve seul ; 2) la médecine et la religion se trouvent ainsi liées par le personnage de Pécuchet.

Le chapitre X et dernier ferme la boucle romanesque du savoir sur la pédagogie. Son statut est un peu particulier, puisqu'il est inachevé. On ne peut donc pas tirer de conclusions définitives.

Remarquons que le texte numérisé s'arrête à la fin du manuscrit rédigé de Flaubert, sans inclure le plan des dernières pages prévues, comme le font généralement les éditions sur papier. Pour la dernière fois, Pécuchet l'emporte aux points sur Bouvard : 78 à 69. Contrairement aux deux chapitres précédents (III et IX), il ne prend pas un avantage décisif en une séquence unique, et à la faveur d'une situation où il trouve seul, sans Bouvard. L'écart se creuse dans trois « disciplines » qu'il prend particulièrement en charge : la phrénologie (3 **Pécuchet** de suite aux § 2917-2926), l'astronomie (7 **Pécuchet** aux § 2967-2986, correspondant au troisième et dernier creux de la courbe vers nB + nP = 1100) et le chant (3 **Pécuchet** aux § 3102-3106).

Dans tous les autres chapitres, six sur dix, Bouvard domine. Il gagne des points progressivement, un à un, sauf en quelques séquences d'avancées plus massives. Si l'on regarde ces passages de forte progression (trois **Bouvard** à la suite et plus), on s'aperçoit qu'ils coïncident avec des scènes où il se trouve avec Mme Bordin, la veuve avenante et cupide :
— premier rapprochement opéré à l'occasion du repas que les deux nouveaux habitants offrent aux Chavignollais, au chapitre II : 6 **Bouvard** à la suite (§ 344-359) ;
— visite du Muséum par Mme Bordin et le notaire, au chapitre IV : 7 **Bouvard** (§ 902-933) ;
— Bouvard s'adresse à Mme Bordin spectatrice quand il joue Hernani, au chapitre V : 6 **Bouvard** (§ 1290-1323) ;
— quasi accouplement entre les draps séchant en plein air, au chapitre X : 5 **Bouvard** (§ 3016-3037).

Bouvard prend l'avantage parce qu'il est un « homme à femmes ». La Phrénologie l'établit bien : « Bouvard présentait la bosse de la bienveillance, de l'imagination, de la vénération et celle de l'énergie amoureuse ; vulgo : érotisme » (X § 2899). Dans ce roman encyclopédique, ce n'est pas le savoir qui fait la différence entre les deux personnages, mais l'amour. Flaubert pouvait confier à une correspondante : « les femmes y tiennent peu de place et l'amour aucune. [...] Ceux qui lisent un livre pour savoir si la baronne épousera le vicomte seront dupés » (à Gertrude Tennant, [16 décembre 1879]). Sous l'angle très particulier de cette analyse quantitative, l'amour pourtant prend sa revanche, comme une irruption du romanesque dans l'encyclopédie. Les femmes ont failli être à l'origine du différend entre les deux amis : « C'était le désir d'en avoir qui avait suspendu leur amitié » (§ 1964), concluent-ils à la fin du chapitre VII. Et c'est bien l'amour qui creuse l'écart.

Ce chapitre VII, celui de l'expérimentation amoureuse, se caractérise par une supériorité modérée de **Bouvard** (28) sur **Pécuchet** (25). Pendant que Bouvard, le roquentin, poursuit un projet matrimonial avec la veuve Bordin, Pécuchet, encore puceau, s'initie : la longue scène entre Gorgu et Mme Castillon, à laquelle il assiste caché dans un fossé, entame le chapitre en citant 5 fois son nom (§ 1842-1871), alors que la scène de déclaration à la veuve aligne 4 **Bouvard** (§ 1897-1904). Ce n'est donc pas dans le chapitre exclusivement consacré à l'amour, comme domaine de savoir et d'expérience spécifique, que la différence majeure s'établit, mais tout au long du roman, en additionnant les séquences de l'intrigue Bouvard-Bordin.

**Perturbations de l'alternance dans le chapitre I**

Grâce au texte colorisé, on peut re-parcourir le texte en sautant d'un personnage à l'autre. Le premier chapitre est un bon observatoire : c'est là que le couple se constitue. La rencontre est un modèle de non hiérarchie. Bouvard et Pécuchet, ce n'est pas Don Quichotte et Sancho, ou Don Juan et Sganarelle, le couple éternel du maître et du valet. Quand il n'y a pas reprise de la séquence du titre, un chiasme rééquilibre les comptes : « l'aspect aimable de Bouvard charma de suite Pécuchet » (§12) / « l'air sérieux de Pécuchet frappa Bouvard » (§15). Ou encore, en plus ramassé :

« Pécuchet contracta la brusquerie de Bouvard, Bouvard prit quelque chose de la morosité de Pécuchet » (§ 79). Match nul : 1 partout.

Et pourtant, il y a 10 **Bouvard** de plus que de **Pécuchet** à la fin de ce premier chapitre. Le plus fort déséquilibre se trouve entre les paragraphes 91 et 111 : 10 **Bouvard** contre 4 **Pécuchet**. Ce qui correspond à la séquence motrice du récit : l'héritage qui va rendre possible la retraite à la campagne et la quête encyclopédique. Or, c'est Bouvard qui hérite. Et il hérite d'un père naturel qui porte évidemment son nom : sur les 10 **Bouvard** de cette séquence, 2 occurrences désignent le père Bouvard (§ 95 et 106), creusant un peu plus l'écart, en face d'un Pécuchet « sans parents » (§ 64).

Anticipant cet héritage qui surdétermine le nom du bénéficiaire, la présentation du portrait de son ancêtre par Bouvard donne lieu à l'une des premières ruptures de l'alternance : « La chambre de Bouvard » / « Mon oncle ! dit Bouvard » (§ 51 et 52). Marque de possession : on répète le nom de celui qui a de quoi, du propriétaire, de l'héritier. Le tout premier dérapage est aussi placé sous le signe du « plus » :
« Bouvard l'engagea à mettre bas sa redingote. Lui, il se moquait du qu'en dira-t-on ! Tout à coup un ivrogne traversa en zigzag le trottoir ; — et à propos des ouvriers, ils entamèrent une conversation politique. Leurs opinions étaient les mêmes, bien que Bouvard fût peut-être plus libéral. » (§ 21 et 22).
Il eût suffi d'écrire : « Bien que Pécuchet fût peut-être moins libéral », et le tour était joué. Pour le sens, cela revenait au même. Si Bouvard prend l'avantage dans le premier chapitre, à partir de ce point, c'est par ce « plus », cet excès, déterminé par la pression sémantique du personnage tout entier, placé dans sa description physique et morale sous le signe du trop plein, alors que Pécuchet se caractérise par la petitesse, la maigreur, le creux, le repli, etc.

Dans la séquence des micro-biographies des deux personnages, aux paragraphes 64 et 65, on constate que le déséquilibre entre les noms provient de l'usage différent des pronoms personnels : la biographie de Pécuchet ne comporte pas une seule fois son nom (une série de pronoms reprend le nom employé deux paragraphes plus haut), alors que celle de Bouvard inscrit deux fois le sien. Si l'on compare ces deux paragraphes sur le plan de la « masse » textuelle, c'est-à-dire du nombre de mots ou de signes, on s'aperçoit que Bouvard emporte encore la mise avec une biographie un peu plus longue (152 mots et 740 caractères contre 120 mots et 637 caractères à Pécuchet). Mais rien de comparable au score Bouvard 2 - Pécuchet 0 que donne le décompte des occurrences des noms. Autrement dit, la supériorité numérique de Bouvard sur Pécuchet ne vaut que pour le calcul des occurrences des noms. Une quantification, autrement plus délicate à conduire, qui prendrait en compte la totalité des signes rapportées différentiellement à chaque protagoniste (en faisant intervenir les pronoms personnels et autres marques de la personne), ne donnerait sans doute pas exactement le même résultat. Opération délicate d'extraction : il faudrait enlever du texte tous les passages au pluriel (**ils**, **nous**) et tous les lieux d'indécision dans lesquels on ne sait pas exactement qui des deux parle ou agit (**l'un / l'autre** ; paroles rapportées sans énonciateur) pour ne retenir que les énoncés narratifs, descriptifs, dialogués, clairement « signés » par chacun des protagonistes. On pourrait alors déterminer le « poids » textuel des deux héros, au-delà des seuls comptages des mots **Bouvard** et **Pécuchet** .

Quel que soit le résultat d'une répartition du texte « entre » Bouvard et Pécuchet faisant intervenir les pronoms, il n'est pas sûr qu'elle rende plus fidèlement compte d'un effet de lecture. C'est très certainement à partir des désignateurs forts représentés par les noms propres que le lecteur perçoit la place et le « poids » respectifs de deux entités qui reprennent et modulent les éléments du titre.

L'analyse du premier chapitre de *Bouvard et Pécuchet* conduit à deux autres remarques. La première concerne la répétition rapprochée des noms. Le phénomène se produit une fois : «Bouvard ! Monsieur Bouvard !», crie Pécuchet (§ 58). Est-ce parce que **Bouvard** comporte deux syllabes qu'il a tendance à se dupliquer ? C'est un hapax dans le roman.
On pourra remarquer d'autre part que ces deux **Bouvard** répétés par Pécuchet précèdent immédiatement un passage (le petit dialogue dans lequel Pécuchet apprend à Bouvard qu'il a suivi son conseil en ôtant sa flanelle) où se rencontrent deux **Pécuchet** (§ 59 et 62). De la même manière, on repère deux **Pécuchet** voisins immédiats, en fin et en début de phrase :
«Il fit demander Pécuchet. // Pécuchet parut. » (§ 87-88)
Et cette répétition très appuyée (l'anadiplose en rhétorique) ouvre la séquence de l'héritage dont nous avons parlé, saturée de **Bouvard**. Dans ces deux cas, il semble que le doublement rapproché agisse comme phénomène de compensation vis-à-vis de la séquence à suivre. Localement, le texte paraît procéder ainsi à quelques rééquilibrages.

**Bouvard / Pécuchet dans la Correspondance**

Dans ses lettres, Flaubert parle presque toujours de Bouvard et Pécuchet associés, personnages ou œuvre, le plus souvent réduits aux initiales liées par le signe alpha : «B  P». Il est très rare que les personnages soient mentionnés indépendamment l'un de l'autre. Quelques passages font cependant exception.

— La lettre à Guy de Maupassant du 5 novembre 1877 résume à grands traits la séquence de l'excursion géologique. Flaubert interroge son jeune ami sur les falaises susceptibles de servir de cadre. Les noms de **Bouvard** et de **Pécuchet** sont naturellement dissociés, selon les paroles et les actions de l'un et de l'autre. Dans la lettre, on compte 5 **Bouvard** pour 5 **Pécuchet**, et dans le roman, pour la séquence développée, 10 **Bouvard** et 10 **Pécuchet** (III § 734-761). Egalité absolue, dans le texte de programmation comme dans l'état final.

— A deux reprises, Pécuchet est cité seul : «Pécuchet vient de perdre son pucelage, dans sa cave ! (avant huit jours mon chapitre "de Amore" sera fini)» (à Maupassant, [15-16 décembre 1878]) ; «Maintenant UN SERVICE LITTERAIRE relatif à Bouvard et Pécuchet. / Pécuchet apprend à un enfant le dessin et commence (suivant la recommandation de J.-J. Rousseau dans Émile) par le dessin d'après nature. Il sait fort peu dessiner lui-même, et il doit barboter d'une manière grotesque, la perspective surtout le démonte. Donc voici ma question : "Quelles sont les bêtises qu'il peut faire, et pourquoi ?" Il y a chez lui défaut de vision et d'aptitude, et chez l'enfant encore plus, bien entendu. Rêve un peu à cela, et réponds-moi d'une manière catégorique » (à sa nièce Caroline, [28 janvier 1880]). Le passage consacré au dessin, dans le dernier chapitre, est en effet réservé à Pécuchet (2 occurrences aux § 2985-2986). On pourrait donc en déduire que le nom de Pécuchet devance celui de Bouvard dans les lettres, mais d'une très courte tête, si un autre phénomène ne venait contrebalancer et contredire cette première observation.

— On sait que Flaubert recourt à de nombreux surnoms pour signer ses lettres. Il lui arrive d'emprunter le nom de l'un de ses personnages, par exemple Aulus Vitellius, le jeune adolescent goinfre d'*Hérodias*, dans plusieurs lettres à Edmond Laporte. C'est à ce même correspondant, très actif pour la documentation de *Bouvard et Pécuchet*, qu'il envoie au moins deux lettres, toutes deux relatives à son voyage au Havre pour localiser l'excursion géologique dont il était question plus haut, et signées... Bouvard. Pourquoi l'un et non l'autre ? Parce qu'il est le premier, et qu'il vient plus naturellement sous la plume que le second ? Parce qu'en signant «Bouvard», il entend que Laporte occupe la place de Pécuchet, dans une recomposition imaginaire du couple (comme Flaubert signant Aulus appelle son destinataire «mon Asiatique») ? Parce que Flaubert s'identifie

plus spontanément à Bouvard, pour des raisons de ressemblances physiques, psychologiques, etc ? Bref, Flaubert préfère-t-il Bouvard ?

Numériquement, dans l'ensemble des lettres, on rétablit la parité absolue entre **Bouvard** et **Pécuchet**, mais avec un avantage qualitatif accordée au premier, car la signature occupe une position stratégique dans la lettre : Pécuchet est deux fois nommé seul, en tant que personnage ; Bouvard par deux fois se substitue à son auteur.

**Analyse statistique de la séquence {BP}**

Considérons maintenant la séquence des occurrences {BP} comme une série chronologique, réalisation d'un processus aléatoire binaire (stationnaire !). Nous allons essayer de déterminer les caractéristiques statistiques de ce processus en vue d'en modéliser le comportement. Rappelons que la séquence contient N = 1216 termes, répartis en 618 B (soit 50,82%) et 598 P (soit 49,18%).

Le premier modèle venant à l'esprit pour une telle séquence est le tir à pile ou face : chaque symbole B (resp. P) est tiré au sort avec une probabilité $p$ (resp. $q = 1 - p$). Si $p = 1/2$, le jeu est dit équilibré. On peut estimer $p$, probabilité d'obtenir le symbole B, à partir de M(p), la proportion observée de symboles B, avec :

$$p = M(p) \pm 1{,}96\sqrt{\frac{M(p)(1-M(p))}{N}} \quad \text{pour un risque d'erreur de 5\%}$$

Comme M(p) = 0,5082, on obtient p = 0,5082 ± 0,0281 soit encore p = 51 ± 3 %. L'écart constaté entre le nombre de B et celui de P n'aurait donc rien de significatif : un résultat digne des sondages de campagnes présidentielles !

Peut-on alors considérer la séquence {BP} comme le résultat d'une répétition de tirs à pile-ou-face équilibrés ? Sans doute non, du fait de l'auto-corrélation apparente de la séquence, qui interdit d'avoir recours à l'estimateur précédent de $p$. Pour évaluer l'auto-corrélation, nous allons utiliser une séquence binaire équivalente, où chaque symbole B reçoit la valeur 1 et chaque P, la valeur nulle. En utilisant comme estimateur (frustre) du coefficient d'auto-corrélation :

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad \text{avec } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \text{M(p) et xi} \in \{0,1\}$$

on obtient, pour les coefficients des premiers rangs :

| r(1) | r(2) | r(3) | r(4) | r(5) |
|---|---|---|---|---|
| - 0,3564 | 0,1363 | - 0,0225 | 0,0343 | 0,0187 |

Avec une incertitude de $\pm 2/\sqrt{N}$, soit ± 0,06, les deux premiers rangs sont significatifs. Il existe de nombreux autres tests pour l'indépendance des termes d'une série chronologique (voir par ex. Michel Vaté, *Statistique chronologique et prévision*, Economica, 1993 – chapitres 1 et 3). Nous pouvons par exemple utiliser le rapport $\delta$ de von Neuman :

$$\delta = \frac{d^2}{s^2} = \frac{1}{N-1}\sum_{i=1}^{N-1}(x_{i+1} - x_i)^2 \Big/ \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \text{ avec } E(\delta) = \frac{2N}{N-1} \text{ et } V(\delta) = \frac{4(N-2)}{(N-1)^2} \to \frac{4}{N}$$

On constate ici que $\delta$ = 0,6782 / 0,2499 et $|\delta - E(\delta)|/\sqrt{V(\delta)} = 12{,}41 > 1{,}96$. On peut donc rejeter l'hypothèse d'indépendance des termes de la séquence (avec un risque d'erreur de 5%).

– Les chaînes de Markov paraissent un modèle plus adapté à notre problème. On se rappelle en effet l'étude fondatrice des successions voyelles / consonnes dans le roman *Eugène Onéguine* (1913). Nous allons suivre pour cela les travaux de Micheline Petruszewycz (*Les chaînes de Markov dans le domaine linguistique*, Slatkine, 1981) en conservant autant que possible ses notations, elles-mêmes d'ailleurs héritées des textes de Markov. Dans le chapitre 1 de ce texte, M. Petruszewycz reprend en détail la démarche suivie par Markov. Nous allons procéder de même pour la séquence {BP}, avec une réserve du fait de la petitesse de notre séquence (1216 termes). Markov utilisait les 20000 premières lettres du roman de Pouchkine.

Nous groupons dans un premier temps les 1216 symboles en 38 groupes de 32 symboles successifs (Markov : 200 groupes de 100 lettres) et effectuons le comptage du nombre de B dans ces groupes. Le tableau ci-dessous donne la répartition obtenue :

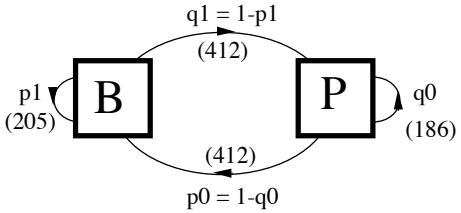
| Valeur | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Effectif | 2 | 1 | 0 | 0 | 3 | 5 | 9 | 5 | 7 | 5 | 1 |

Cet histogramme va nous permettre d'obtenir M(p), valeur moyenne de *p* et sa variance V(p). Le mode de l'histogramme est la valeur 16. La moyenne arithmétique pondérée par la distance au mode est 16,2632 donc M(p) = 16,2632 / 32 = 0,5082. La somme des carrés des écarts à 16,2631 est 205,3684 donc V(p) = 205,3684 / 38 / 32 = 0,1689. Or si la séquence pouvait être considérée comme une suite de réalisations indépendantes, la variance serait M(p)(1-M(p)) = 0,2499 et on a :

$$\eta = \frac{\text{var obs.}}{\text{var théo.}} = \frac{V(p)}{M(p)\left(1 - \mathrm{M}(p)\right)} = \frac{0{,}1689}{0{,}2499} = 0{,}676$$

Pour une répartition en 64 groupes de 19 symboles, on obtient des résultats très proches : V(p) = 0,1698 et $\eta$ = 0,679. Il est intéressant de noter que Markov et Petruszewycz obtiennent pour leur séquences voyelles/consonnes des valeurs plutôt autour de $\eta$ = 0,2 ou 0,3, donc des séquences où la dépendance entre les termes est plus forte.

Pour la chaîne d'ordre 1, le schéma ci-dessous identifie les deux états, les probabilités de transitions et, entre parenthèses, les digrammes décomptés dans la séquence. L'égalité des transitions BP et PB peut surprendre : elle est cependant inévitable (au premier symbole de la séquence près).




Connaissant les taux de transitions, on déduit immédiatement les probabilités stationnaires des deux états. En notant toujours *p* (resp. *q)*, la probabibilité stationnaire du symbole B (resp. P), on a :

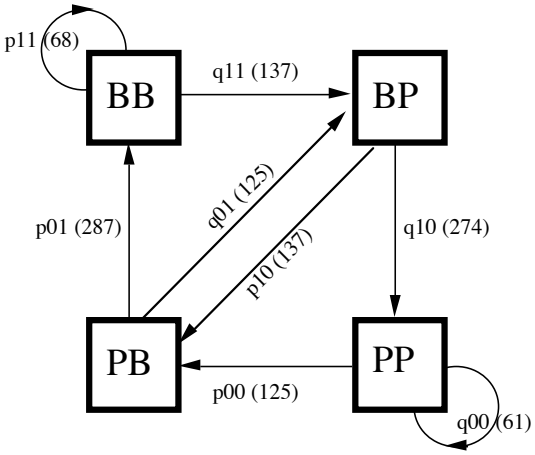
$$p = pp_1 + qp_0 \text{ et } p + q = 1 \Rightarrow p = \frac{p_0}{1 - p_1 + p_0}$$

Comme $p_1$ = 205 / 616 et $p_0$ = 412 / 598, on vérifie que p ≈ 0,508. En introduisant $\delta = p_1 - p_0$, le coefficient dispersif de Markov est pour notre exemple :

$$C_1 = \frac{1+\delta}{1-\delta} = 0{,}474$$

valeur assez sensiblement différente (30%) du rapport $\eta$ = 0,676 attendu. On peut donc supposer que la chaîne d'ordre 1 traduit mal les dépendances observées dans la séquence. Passons maintenant

à la chaîne d'ordre 2 en décomptant cette fois les trigrammes de la séquence. Le schéma ci-dessous en donne le résultat obtenu :




Avec $p_{11} = 68/205$ et $q_{00} = 61/186$ et en introduisant deux variables auxilaires :

$$\alpha = \frac{p_{11} - p_1}{1 - p_1} \text{ et } \beta = \frac{q_{00} - q_0}{1 - q_0}$$

on peut alors calculer le coefficient dispersif de Markov pour l'ordre 2 :

$$C_2 = C_1 \left[ \frac{1+\alpha}{2(1-\alpha)} + \frac{1+\beta}{2(1-\beta)} \right] + \frac{(q-p)(\alpha-\beta)}{(1-\alpha)(1-\beta)} = 0{,}486$$

Là encore, l'accord avec $\chi^2$ est mauvais. Donc, soit la séquence {BP} ne se réduit pas à ce type de chaîne de Markov (voire n'a pas de caractère markovien marqué), soit nous subissons ici la petitesse de l'échantillon.

Pour évaluer ce dernier point, on peut comparer les résultats obtenus en faisant une étude des successions voyelles/consonnes avec les 1216 premières lettres du texte à notre disposition. La nouvelle séquence commence par « Comme il faisait une chaleur de trente-trois degrés » et se termine par le dialogue inachevé « – Mon Dieu, oui, on pourrait prendre le mien à mon bureau – C'est comme moi ».

```
CVCCVVCCVVCVVCVCVCCVCVVCCVCCVCCVCCVVCCVCCVCCVCVVCVCVCCCVVCCVCCVCCVVCVVCVCCVCVCVCCCCVCVCCC
CVCCVCCVCVCVCCVVCCCVCCVCCVCCVCCVCVCCVCCVVCVCCVCVCVCVCVVCVCCVCCCVCCVVCVCVCVVVCVVCVVCCVCCCCVVCV
VCVVCVVCVCVVVVCCVCVVVCCVVCCVCVVCVCCVCCVCVCCVCVVCCVCCCCVCVCCVCVVCVCVCVCVCVCVCVCVCCCCVCVCCVV
CVCCCVVCVCVCVCCCVCCCVCCVVCCCVCCVCCCVVCCVCCVCVCVCVVCVVCVCCCVCVVCCVVCCVCVCVCCCVVCCVCVCVCCVCVC
VVCCVCVCVVCCVCCVCVCVCCCVCCCVCCVCCVVCCCVCCVVCVCCVCCCVVVCCVCCVCVCVCCVVVCCVVVCCVCVCVCVVCCVCC
VCVCVCCVVCCVCVVCCVCCCVCCVCCCVCVCVVCVVCCVVCCVCCCCVVCVCCVVCCVCVCCVCVCVVVCCVCVCCCVCVCVCCCVVC
CVCCVCCVCCVCVCCVVCCCVCVCVVCCVCCVCCVCVCVCCCVCCVCVVCCVCVCVCCVCCVCVVCCVCVCVCCVCCCVCCCVCCVCCV
CCVCCCVCCCVCVCVCVVCVCVCCCVVCCVCCVCVVVVCVCCVVCVCVCVCVCVCCVCVVCVCCVVCCVCCCVCVCVVCVCVVCCVCCVCC
VCVCCVCCCVCVCCCCVCCVCVVCCVVCCVCCVCVCVCVCCVCVCVCCVCCVVCCVVCCVCVCVCVVCVCVCVCCCVVCCVVCVCVVCV
CVVCCVVCVVCCVCCCVCVCCVCCVCVCVVCVCVVVCVCVVCVCVCCCVCCCVCCVCVCCVCVCVCVCVCVCVCVCCVCVCVCVCCCCVV
CCVCCVVVCCVCCVCCVCCCVCVCVCVCCCVVCCCVVCCVCVCCVVCCVCVCCVCVCCVCCVCVVVCCVCVCVCCCVCCVVCVCCVCVC
CVCCVCCCVCCVCVVVCVCVCCVVCVCCVVCVCCCVCCVCCCVVCVCVVCVCVCCVCCVVVCVVCVCVCVCCCVCCCVCVCCVVCCVCVC
VCCVCVCVVCVCCVCVCCVCVCVCVCCVCVCCVCCVVCCCVCVCCVVCVCVCCVVCVCVCVVCVVCVCCCVCVCVCVCCVCVCCVCVCC
VCCCVCCVVCCVCCCVCCCVCCVVVVVVVCCVVCCVVCCCCVCCCVCVCVVCVCVCCCVCVVVCVCCCCVCCVCVV
```

On compte 674 consonnes. L'histogramme obtenu en construisant 38 groupes de 32 symboles est plus resserré que pour la séquence {BP} :

| Valeur | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|
| Effectif | 1 | 1 | 7 | 7 | 8 | 10 | 4 |

Le mode est 19. Ici, M(p) = 0,5543 et V(p) = 0,0686. La variance théorique M(p)(1-M(p)) = 0,2471 et on a□ = 0,28. Après comptage des digrammes et des trigrammes, on obtient les résultats suivants□

| p1 | p0 | p11 | q00 | C1 | C2 |
|---|---|---|---|---|---|
| 0,3858 | 0,7634 | 0,1962 | 0,1641 | 0,45 | 0,29 |

Ces résultats sont bien cohérents avec ceux obtenus par Markov et Petruszewycz sur d'autres corpus. L'accord de la chaîne d'ordre 2 est correct pour cette séquence.

– A ce stade, nous aurions dû continuer l'expérience avec des chaînes d'ordre supérieur. Nous aurions pu également transformer la séquence {BP} en une séquence trinaire, où l'on agrège les digrammes BP et PB en une forme unique. Nous n'avons malheureusement pas su généraliser le calcul des coefficients dispersifs effectué par Markov à ces types de chaînes.

Pour rester dans le contexte des travaux cités, nous allons étudier la séquence binaire {ES} obtenue en groupant dans un état E (ensemble) les digrammes BP et PB, puis dans un état S (seuls) les symboles B et P restants. La séquence devient□

```
EEEEESSEEEEEEEEEEEEESEEESEESSEEEEESEEEESESSSESESEEEEESEESSSESEEEEESEEESSEESSEEEESSESESES
SEESSEESSSESEESEEESESESEEEEESEEEEESEEEESSSSSEEESSSSSESESEEESEEEEEEEESESESEEESEEEESEESEE
SESEEESSSSSSESSEESESEESEEESESEEEEEEESEESEEEEEEEEESEESSSESESSSEEESEESEESSESEEEEESSSSSSE
ESESESESEEEEESSESSSESEEEESESEEEEEESESSEESEESEESEEESEEEEESEEESEESEEESESSSSSEEEEEESESSES
SEESSEESESEESSEEESSEEEEEEESSESSESESESSEESEESSEEESEEESESEEESEEEEEESESSSEESESESESESSSSESE
EEEESEESEESEEEESESSESSSESESESSESESSSEEEEEEESEESEESEEESESESEESSSEEESEEEEESEEESEESEEEESSEE
SESEEEEESSESSSSEEEEEEEEEEEEESEESEEESEEEEESEESEESEEEESEESSESEESSESEESEESEESSSEESESSSSSS
SSSSSSSSESESESESSESEEEESESEEESEESESSESEESSEESEEESESEEEESESSSEEESEESESEESSSSSSEESESESSSEE
SESEEEESESEEESESESSSSEEEEEESEEEEESESEEESEESSEEEEES
```

On compte 754 symboles, dont 462 E (soit 61,3%) et 292 S (soit 38,7%). L'auto-corrélation de la séquence est ici négligeable et le test de von Neuman nous conduit à accepter l'hypothèse d'indépendance□ = 0,4927 / 0,2373 et $|\square - E(\square)| / \sqrt{V(\square)} = 1{,}01 < 1{,}96$.
On peut donc modéliser les alternance Ensemble/Seuls par des tirs à pile ou face déséquilibrés. Nous pouvons alors prédire la quantité de séquences répétées EE…E et SS…S. On sait que dans ce cas, les séquences de $k$ = 1,2,… symboles identiques E (resp. S) apparaisent avec une probabilité $P[X = k] = p^k(1 - p)$ avec p = 0,613 (resp. = $(1 - p)^k p$). Nous indiquons ci-dessous les effectifs observés, avec en regard l'effectif théorique en prenant $p$ = 0,613. Un test de chi-2 montre que les répartitions coïncident bien pour les séquences répétées de E (avec 5 cellules, D = 2,15 < $\square^2_{[95\%,3]}$ = 7,8). En revanche, l'accord est moins bon pour les séquences répétées de S (D = 8,31).

| Séquence | obs. | théo. | (obs-théo)^2/théo |
|---|---|---|---|
| S | 106 | 113,1 | 0,446 |
| ES | 76 | 69,3 | 0,097 |
| EES | 48 | 42,4 | 0,739 |
| EEES | 27 | 26,0 | 0,038 |
| EEEES | 9 | 15,9 | 0,838 |
| EEEEES | 10 | 9,8 | |
| +E…ES | 16 | 15,15 | |

| Séquence | obs. | théo. | (obs-théo)^2/théo |
|---|---|---|---|
| E | 277 | 283,1 | 0,131 |
| SE | 131 | 109,6 | 4,178 |
| SSE | 30 | 42,5 | 3,676 |
| SSSE | 15 | 16,4 | 0,120 |
| SSSSE | 2 | 6,4 | 0,214 |
| SSSSSE | 3 | 2.5 | |
| +S…SE | 4 | 1.6 | |

Par ailleurs, en *supprimant* les digrammes BP et PB, la séquence {BP} devient :

```
BBBBBBBBBBBBBBPPPBBBBBPPPBBBBBBBBBPBBBBBBBBBBPPPPPPPPPPPPPPPPPBBBBBBPBPBBBBBBPPPPBBBPB
PBPBBBBPPBBBBPPPPPPBBPPBBBPPPPPBPBBBBBBBBBBPPBBBBPPPBBBPPPPPBBBBPBBPPPPBBBBBBBBPPBBBBP
PPPPPPPPPPPPPPBBPBBBBPPPBBBBPPPPPPPPPPPPPPBBBBPBBPPPPBPBBB
```

Cette nouvelle séquence, notée {BsPs}, compte 234 termes, dont 190 B. On peut encore (car N>>30) utiliser le test de von Neuman. On obtient ici $\delta$ = 0,2403 / 0,2482 et $|\delta - E(\delta)| / \sqrt{V(\delta)} = 7{,}96 > 1{,}96$. On ne peut donc pas cette fois accepter l'hypothèse d'indépendance. Etudions maintenant si une chaîne de Markov modélise la séquence. L'histogramme obtenu pour 26 groupes successifs de 9 symboles est :

| Valeur | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Effectif | 3 | 1 | 0 | 1 | 6 | 2 | 7 | 3 | 1 | 2 |

Le mode est 6. Ici, la proportion de B est M(p) = 0,5427 et V(p) = 0,6866. La variance théorique M(p)(1-M(p)) = 0,2482 et on a $\delta$ = 2,78. Après comptage des digrammes et des trigrammes, on obtient les résultats suivants :

| p1 | p0 | p11 | q00 | C1 | C2 |
|---|---|---|---|---|---|
| 0,777 | 0,262 | 0,755 | 0,785 | 3,13 | 3,49 |

On pourrait donc modéliser la séquence {BsPs} par une chaîne d'ordre 1, avec toutes les réserves qu'impose la petite taille de l'échantillon.

Ensuite, en renommant X les sous-séquences BP et Y les sous-séquences PB, puis en supprimant les symboles B et P restants, on obtient :

```
XXYXXXYXXXXXXXXXXXXXXXXYXXXYXXXYXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYXXXXYXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXYXXYXXYXXXXXXXXYXXXXXXXXXXYXXXXXXXXXXXXYXXXXXXYXXXYXXXXXXXXYXYXXXXXXYXYXXXXXXXXXYXX
XXXXYXXXXXXXXXXXXYXXXYXXXXYXXXXXXXXXXXXXXYXXXXXXXXXYXYXXXXXXYXXXXXXXXXXXXXXYXXXXXXXXXYXX
XXXYXXYXXXXXYXXXXXXXYXXXXXXXXXXXYXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYXXXXXXYXXXXXXXYXXXXXXX
XXXXXXXXXXXXXXXXYXXYXXXXXXXXXXYXXXXXXXXXXYXXXXXYXXYXXXYXXXYXXXXXXXXXXXXXXXXXXYXXXXXXXXXYXXX
XXXXYXXXXXXXYXXXXXXXXX
```

Cette séquence, notée {XY}, compte 462 termes, dont 412 X. On notera l'absence de répétition de Y. Le test de von Neuman échoue à prouver l'hypothèse d'indépendance. En construisant un histogramme à partir de 42 groupes de 11 termes, on obtient :

M(p) = 0,8918 V(p) = 0,0660 $\delta$ = 0,68

Les caractéristiques de chaîne sont :

| p1 | p0 | p11 | q00 | C1 | C2 |
|---|---|---|---|---|---|
| 0,878 | 1 | 0,867 | 0 | 0,78 | 0,65 |

L'accord entre C2 et $\delta$ est très correct (mais, là encore, l'échantillon est petit).

En conclusion, on pourrait modéliser la séquence {BP} par un modèle markovien résultat de l'agrégation des 3 processus simples {ES}, {BsPs} et {XY}. Pour reconstruire une trajectoire de {BP}, on effectuera un choix au hasard entre une situation « ensemble » (E) avec p = 0,613 ou « seul » (S) avec q = 1-p. Puis, si E est obtenu, choisir entre les digrammes BP et PB selon une chaîne de Markov d'ordre 2. Sinon, choisir entre les symboles B et P selon une chaîne d'ordre 1.

– L'alternance entre les symboles dans la séquence {BP} obéit donc à des règles complexes, que l'on ne peut visiblement pas réduire à une chaîne de Markov d'ordre 2. Les symboles B et P sont liés de manière plus subtile. Pour qui connaît le texte d'origine de la séquence, cela n'est bien sûr pas une découverte. Il est cependant intéressant de voir émerger une structure des informations brutes d'occurrence. Par ailleurs, on peut modéliser les alternances «□B et P ensemble□ / «□B ou P seul□ par un tirage au sort – et cela permet d'en déduire correctement le nombre de répétitions. En fait, ce résultat se retrouve dans tout phénomène «□sans mémoire□□c'est le cas par exemple de la durée des conversations téléphoniques. Ici, l'auteur déterminerait son choix de poursuivre (par ex. à traiter «□B et P ensemble□) à chaque étape, indépendamment des choix pris dans le passé. Il n'est bien sûr pas dans notre intention de conclure formellement que Flaubert procède de cette manière, puisque des processus plus complexes pourraient conduire aux mêmes résultats. Le principe d'économie – fondement de la modélisation de systèmes physiques – ne s'utilise pas sans danger en matière de stylométrie. Il faudrait également poursuivre ce type de mesures sur d'autres textes où apparaissent des noms de personnages en couple.

**Pour conclure**

Le titre du roman, un double nom propre, place les deux personnages éponymes côte à côte ou face à face, dans une position de stricte égalité. Le compteur n'affiche pourtant pas le même nombre d'occurrences pour l'un et pour l'autre : le premier domine le second sur la totalité du roman, et dans la plupart des chapitres : **Bouvard** 618 / **Pécuchet** 598.

Dans ce décompte global, il ne faut cependant pas négliger le nombre de fois où «□Bouvard et Pécuchet□ apparaissent liés dans le texte comme dans le titre : 69 occurrences au total. 69 fois le titre réinscrit la coordination. Une fois sur dix, **Bouvard** et **Pécuchet** forment donc couple syntaxique, toujours dans cet ordre, celui du titre, jamais dans l'ordre inverse. C'est probablement cette permanence du binôme que le lecteur retient.

La supériorité numérique de **Bouvard** sur **Pécuchet**, que le lecteur ne perçoit probablement pas sur l'ensemble du roman, va en tout cas dans le sens d'une intuition ou d'une présomption fondée sur l'ordre alphabétique (**Bouvard** vient avant **Pécuchet** comme B avant P), sur la progression métrique (dite «□adence majeure□ : deux syllabes, puis trois) et sur la représentation que le texte donne des deux hommes, Bouvard étant le plus gros, le plus grand, le plus expansif, le plus généreux, etc. Bref, pour toutes ces raisons, Bouvard est le premier et Pécuchet le second.

Dans un essai déjà ancien, *La Spirale et le monument* (SEDES, 1988), l'un d'entre nous écrivait un peu vite : «□Flaubert dépense des trésors de savoir-faire pour rendre, dans la linéarité du texte, la simultanéité des actions et pour ménager malgré les subordinations syntaxiques et rhétoriques une complète égalité des personnages » (p. 55). Cette étude permet aujourd'hui de nuancer ce propos trop affirmatif.