***Chapitre 8 (numérique) de : Étienne BRUNET,* Tous comptes faits.**
Écrits choisis, tome III. Questions linguistique*s, Bénédicte PINCEMIN (éd.),*
*Paris : Éditions Champion, sous presse (publication prévue en 2016).*
*Publié en ligne par la revue Texto ! Textes & Cultures,* http://www.revue-texto.net
*Volume XXI – n°1 (2016). Coordonné par Christophe GÉRARD.*


# La ponctuation et le rythme du discours (d'après les données du *Trésor de la langue française*)[1]

## – I –

**1 –** La ponctuation pourrait constituer un domaine privilégié de la linguistique quantitative. Elle permet en effet d'échapper au cercle étroit du mot et d'ouvrir une perspective sur la phrase et le rythme du discours. Tant que la reconnaissance du sens – qui conditionne dans une large part celle de la syntaxe – n'aura pas trouvé une solution acceptable dans le traitement automatique des textes, la ponctuation restera l'un des seuls accès au supralexical dont on puisse tirer parti. L'ordinateur en effet est encore une machine bien lourde pour l'analyse du contenu. La machine se déplace vite, mais il lui faut une route. Et dans le maquis du sens, c'est l'homme qui doit frayer le passage, préparer le texte, réduire les ambiguïtés, simplifier les constructions, établir les équivalences et réaliser les substitutions. Si on prétend lancer la machine dans le texte brut, on risque l'accident au moindre homographe et pour contourner le moindre caillou d'infinies manœuvres seront nécessaires. Or à cette machine dont la vue est myope, la ponctuation propose des jalons précieux et une signalisation simple et sûre qui permet un survol rapide. Sans doute en ne prenant appui que sur ces repères, la saisie reste incomplète. On ne voit de la route que les bornes et l'on ignore le paysage. Mais si élémentaire et si rapide qu'elle soit, cette vue dynamique du discours en mouvement apporte un utile complément aux méthodes strictement lexicales et un peu de liant à la poussière des mots. D'un fractionnement atomique et lexical du langage, nous passons à une segmentation naturelle et fonctionnelle du discours.

---

1. NDÉ : Article publié dans *CUMFID* n°13, Université de Nice, 1981, p. 1-28 (1981d), et correspondant également au chapitre VI du *Vocabulaire français* (1981a).

**2 –** La ponctuation a d'autre part l'avantage d'un système simple et peu équivoque. Certes quelques signes peuvent remplir plusieurs fonctions et l'environnement immédiat doit être examiné avant de décider si le point est celui de la siglaison, de la suspension, de l'abréviation ou de la fin de phrase. Mille nuances du sentiment peuvent s'exprimer par le point d'exclamation, de l'admiration à l'indignation, de l'émotion à l'indifférence. Même le signe le plus léger, la virgule, peut jouer des rôles bien différents, tantôt en liant, tantôt en isolant les termes. Et que dire des deux points qui peuvent introduire une explication ou une énumération ou une citation ? Mais chaque signe de ponctuation est autonome (mis à part ceux qui fonctionnent par paires comme les guillemets, les parenthèses ou certains tirets). Le sens d'un signe de ponctuation ne dépend pas de celui qui le précède, non plus que de celui qui le suit. Il y a peu de combinatoire, peu de syntaxe de ces éléments du langage, dont l'agencement est quasi libre. Au surplus ces éléments sont très peu nombreux, une dizaine au maximum, dont le système apparaît bien pauvre si on le compare aux mille ressources dont dispose le langage parlé pour marquer les intonations, les pauses, les accents, la mélodie, la mimique. Et cet appauvrissement représente un avantage méthodologique. Comment pourrait-on éviter l'arbitraire s'il s'agissait d'étalonner l'intonation, de mesurer les pauses, de définir les accents et de cataloguer les mille nuances de la voix et du geste ? L'abstraction et la simplification du codage écrit, par quoi l'on transpose l'accompagnement métalinguistique du discours parlé, permettent en outre, en réduisant les variétés étudiées, d'augmenter les effectifs. Et la statistique se complaît dans le désert des grands nombres.

**3 –** Au surplus la ponctuation dans un corpus littéraire contemporain n'est plus chose méprisable. On sait que le système ne s'est constitué qu'à une date relativement récente et qu'à part le point aucun signe ne remonte au delà du XVI$^{e}$ siècle. La gamme comportait alors quatre signes (. , : ?), auxquels le XVII$^{e}$ ajoute le point et virgule et le point d'exclamation et le XVIII$^{e}$ les points de suspension. Si nous écartons quelques signes secondaires de création plus récente encore (les tirets), on peut dire que le système est complet au moment où commence la première tranche de notre corpus. Il serait probablement stupide d'étudier la ponctuation d'un texte plus ancien, le plus souvent restituée et normalisée après coup. Ce ne l'est pas, à partir du XIX$^{e}$ siècle, d'autant que les éditions choisies ont en général respecté l'orthographe et la ponctuation originales. En particulier la ponctuation respiratoire et non logique qui règne encore à la fin du XVIII$^{e}$ a laissé des traces dans la

prose de Chateaubriand[2], et l'édition du centenaire préparée par Maurice Levaillant a maintenu ces particularités dans les *Mémoires d'outre-tombe*. En même temps l'écrivain – et Chateaubriand le premier qui s'est corrigé sur ce point en s'accommodant au goût du jour – a prêté plus d'attention aux marques extérieures du rythme et ne s'en est plus remis là-dessus aux soins de l'imprimeur. Certaines écoles littéraires ont eu recours à la ponctuation pour affirmer l'originalité du style. En multipliant les virgules, ou les points, on a pu obtenir un effet de nivellement, de distanciation, d'émiettement ou de précipitation selon les cas. À la limite même l'absence de ponctuation – et l'on songe à Apollinaire – est encore une manière d'utiliser le système en s'en écartant pour produire un effet. Il en est ainsi de l'artifice typographique qui proscrit les majuscules et qui les impose d'autant plus fortement à l'attente déçue du lecteur. En s'affinant, la ponctuation libère la phrase, elle assouplit, module le rythme et y introduit ce qui dans la musique est aussi essentiel que la musique même : le silence. Il n'est que de comparer un texte de rhétorique classique où le discours a horreur du vide, et où les mots et les phrases s'enchaînent sans discontinuité, et telle page de Céline qui joue en virtuose du cri et du silence et utilise à chaque ligne les points d'exclamation et de suspension. Le cri a en effet plus d'écho et plus de force quand la phrase a des trous, des ruptures, des espaces quasi vides où la ponctuation se substitue aux mots.

**4 –** En outre la ponctuation permet des études dynamiques où le mouvement même du rythme est étudié dans son déroulement. Certaines méthodes statistiques permettent en effet de traiter les données sérielles où le temps est essentiel. Or le plus souvent la statistique lexicale s'occupe de fréquences où la localisation des mots importe peu puisque seul compte le nombre d'occurrences. Le temps y est donc aboli au moins dans le texte traité, même s'il peut réapparaître quand on compare entre eux plusieurs textes d'époque différente. Sans doute les méthodes sérielles pourraient elles être appliquées aux unités lexicales. Encore faut-il que les éléments dont on note l'apparition régulière ou irrégulière aient une fréquence telle que la régularité ait lieu de s'exercer – ce qui réserve le traitement aux mots grammaticaux ou à certains mots sémantiques particulièrement répandus dans le discours. La très grande fréquence des signes de ponctuation facilite grandement ces calculs de périodicité – que nous avons mis en œuvre dans l'*Émile*[3].

2. Par exemple la virgule peut séparer le verbe et son sujet, ou le verbe et l'objet.

3. Voir « Considérations sur la ponctuation et le rythme de l'*Émile* », dans notre *Index*

**– II –**

**1 –** Malheureusement nos données initiales ne nous ont pas permis cette dernière étude, puisque nous manquaient les références précises des mots et des signes. Nous pouvions calculer un espace moyen entre les unités mais non leur dispersion. On peut seulement supposer que les phénomènes de périodicité observés dans l'*Émile* ne sont pas propres à ce texte et qu'on les retrouve dans l'exercice naturel du langage. Le discours n'est pas un milieu homogène : les variations thématiques, rythmiques, stylistiques rendent irrégulière l'apparition des mots et des signes. Les occurrences de la même unité s'accumulent en certains endroits, et se raréfient ailleurs. Appliqué aux signes de ponctuation, le phénomène de la spécialisation linguistique aboutit à des mouvements rythmiques, tantôt précipités, tantôt ralentis, qui excluent pareillement le hasard – du moins dans les cas soumis à l'étude.[4]

**2 –** D'autre part nos données n'étaient pas aussi pures qu'il eût été souhaitable. Ainsi le point n'a pas été désambiguïsé. Sur les 4 millions d'occurrences qui ont été relevées, certaines appartiennent à la suspension, d'autres à la siglaison, d'autres à l'abréviation. Seules ces deux dernières espèces peuvent être repérées indirectement.

Comme le point a été conçu uniformément comme séparateur, des lettres se sont trouvées isolées qui désignent soit l'initiale d'un prénom soit les composants d'un sigle. Le détail pour chacune des initiales figure dans le tableau 1. Dans plusieurs cas l'abréviation se mêle à d'autres phénomènes beaucoup plus fréquents : l'existence d'une forme non abrégée (A du verbe AVOIR, adverbe Y), ou d'une lettre de liaison (le T euphonique). En d'autres occasions certaines lettres servent à la transcription des chiffres romains, ce qui est particulièrement gênant pour I, V et X. Enfin il y a des perturbations imprévues lorsqu'un prénom ou une initiale pour des raisons thématiques reviennent avec insistance dans tel ou tel texte. C'est le cas de H dans la 6ème tranche (1068 occurrences sur 2668) et de P dans la 11ème tranche (2925 occurrences sur 3525). On trouvera dans le tableau 2 ces distributions particulières. Il reste à peu près 40 000 points qui servent à désigner sigles et abréviations (en tenant compte aussi de quelques abréviations qui recourent à plusieurs lettres

---

*de l'Émile*, Slatkine, p. 569-583 (1980a).

4. NDÉ : Une telle étude de la répartition intratextuelle des ponctuations est présentée au chapitre suivant (« La phrase de Zola ») ; et dans les *Écrits choisis*, un autre exemple est donné dans « Hugocentric Tendancies or Can One Approach Hugo Counting Words » (1988b), tome I, chapitre 5, p. 121-122.

comme ETC, CF, VS). Notons que le coefficient chronologique est presque toujours positif. Les 3 signes négatifs du tableau 1 s'expliquent aisément, I et V par l'interférence des chiffres romains, M par la fréquence de M. pour MONSIEUR (la distribution de H est d'ailleurs parallèle à celle de MONSIEUR). Sur les 19 lettres peu ambiguës, 11 montrent un coefficient significatif, ce qui prouve le progrès d'ensemble de la siglaison et de l'abréviation.

**Tableau 1. Abréviation et siglaison. Variation selon le genre**

| | tendance | prose | Prose poétique | Vers | Techn. | Solil. | Dial. | Reste | fréquence |
|---|---|---|---|---|---|---|---|---|---|
| b | 0,40 | -36,9 | -3,8 | -6,1 | 46,8 | -2,6 | -12,3 | 12,0 | 1618 |
| c | 0,52 | -20,4 | -3,5 | -5,1 | 27,2 | 0,4 | - 5,2 | 3,8 | 1559 |
| d | 0,04 | -9,4 | -2,9 | -3,4 | 13,6 | 7,3 | -5,7 | -1,3 | 833 |
| e | 0,35 | -15,9 | -1,8 | -4,1 | 21,0 | 0,1 | -4,2 | 3,3 | 950 |
| f | 0,60 | 7,2 | -3,2 | -4,8 | -4,6 | 24,9 | -2,8 | -17,9 | 1003 |
| g | 0,65 | -1,0 | 0,7 | -2,9 | 2,6 | 5,1 | -5,9 | 0,6 | 371 |
| h | 0,02 | 18,8 | -4,6 | -8,1 | -15,5 | 55,6 | -9,1 | -37,7 | 2688 |
| i | -0,10 | -31,1 | -6,8 | -1,2 | 38,3 | 12,7 | -8,2 | -3,7 | 7666 |
| j | 0,63 | 3,3 | -1,1 | -1,1 | 2,9 | 2,8 | -0,2 | -2,1 | 114 |
| k | 0,65 | -2,2 | -1,1 | -2,2 | 4,1 | -3,2 | -2,2 | 4,4 | 123 |
| l | 0,16 | 3,1 | -1,1 | -2,4 | -1,9 | 1,0 | 6,0 | -6,1 | 566 |
| m | -0,57 | 24,7 | -6,4 | -12,9 | -18,9 | 9,6 | 11,9 | -17,3 | 5548 |
| h | 0,01 | 2,3 | -3,3 | -5,6 | 1,6 | 23,8 | 1,2 | -20,2 | 1467 |
| o | 0,59 | -8,8 | 5,7 | 10,0 | 2,6 | -2,6 | 26,1 | -18,7 | 1561 |
| p | 0,50 | -26,0 | -5,6 | -10,7 | 37,5 | 4,9 | -8,9 | 3,2 | 3525 |
| q | 0,32 | 4,2 | -1,2 | -2,3 | -3,1 | 10,1 | -3,2 | -5,6 | 140 |
| r | 0,65 | 6,7 | -2,7 | -5,5 | -3,7 | 21,4 | 1,8 | -18,8 | 1101 |
| s | 0,15 | -14,3 | -4,4 | -4,5 | 20,2 | 10,9 | -5,1 | -4,7 | 3509 |
| t | 0,16 | 21,1 | -8,3 | -12,6 | -14,4 | -15,0 | 29,9 | -11,8 | 39082 |
| u | 0,57 | -3,9 | -0,5 | -1,6 | 5,5 | -2,7 | 0,3 | 2,0 | 157 |
| v | -0,48 | -4,0 | -2,8 | -1,5 | 6,3 | 15,6 | -8,2 | -6,1 | 1468 |
| w | 0,62 | 2,1 | -1,0 | -1,3 | -1,4 | -0,1 | -1,6 | 1,3 | 92 |
| x | 0,26 | -7,2 | -1,1 | -0,6 | 8,8 | -1,1 | -3,8 | 3,9 | 683 |
| z | 0,50 | -3,2 | -1,2 | 5,9 | 0,7 | 2,1 | -0,2 | -1,5 | 149 |

De plus ces phénomènes sont liés au genre littéraire. On les rencontre rarement en poésie (tous les écarts sauf un sont négatifs dans la prose poétique et dans les vers), souvent dans le corpus technique et plus encore dans le soliloque (4 écarts négatifs seulement).[5]

5. NDÉ : Ces genres correspondent aux sous-ensembles définis à l'époque dans le

**Tableau 2. Abréviation et siglaison. Variation selon la tranche chronologique. Cas particuliers des lettres H, I, M, P, T, V. (*f* = fréquence, *z* = écart réduit)**

| | h | | i | | m | | p | | t | | v | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | f | z | f | z | f | z | f | z | f | z | f | z |
| 1800 | 37 | -13,0 | 1008 | 15,2 | 330 | -6,4 | 99 | -11,9 | 2830 | -7,8 | 209 | 8,2 |
| 1825 | 295 | 7,6 | 515 | -1,8 | 719 | 16,4 | 116 | -9,0 | 2746 | -1,6 | 207 | 10,1 |
| 1835 | 72 | -9,0 | 455 | -4,2 | 491 | 4,8 | 94 | -10,4 | 2934 | 2,5 | 232 | 12,8 |
| 1845 | 65 | -5,7 | 302 | -7,0 | 792 | 26,9 | 83 | -8,8 | 2476 | 4,4 | 37 | -5,4 |
| 1855 | 59 | -8,2 | 317 | -6,9 | 400 | 3,8 | 81 | -9,3 | 2611 | 5,7 | 57 | -3,4 |
| 1865 | 1068 | 72,5 | 594 | 5,6 | 464 | 6,9 | 92 | -8,8 | 2150 | -5,7 | 114 | 2,5 |
| 1875 | 99 | -4,5 | 714 | 13,4 | 256 | -3,6 | 96 | -7,7 | 2458 | 4,6 | 72 | -1,4 |
| 1885 | 34 | -11,5 | 359 | -7,8 | 156 | -12,1 | 181 | -4,2 | 2510 | -4,0 | 77 | -2,6 |
| 1900 | 23 | -12,6 | 423 | -5,6 | 571 | 9,0 | 406 | 10,0 | 2882 | 1,5 | 103 | -0,2 |
| 1913 | 43 | -9,6 | 486 | 1,2 | 258 | -4,3 | 298 | 6,1 | 1939 | -8,8 | 42 | -5,1 |
| 1922 | 169 | -1,1 | 630 | 4,7 | 288 | -4,9 | 1025 | 52,2 | 2925 | 4,9 | 61 | -4,1 |
| 1930 | 496 | 28,1 | 445 | -0,1 | 120 | -11,7 | 68 | -9,9 | 2016 | -5,7 | 47 | -4,3 |
| 1935 | 69 | -7,6 | 194 | 13,1 | 303 | -2,1 | 63 | -10,7 | 2468 | 1,5 | 41 | -5,3 |
| 1942 | 76 | -8,0 | 387 | -5,9 | 143 | -12,4 | 514 | 18,5 | 2675 | 0,7 | 77 | -2,3 |
| 1955 | 63 | -10,4 | 837 | 10,4 | 257 | -8,7 | 309 | 2,2 | 3462 | 8,2 | 92 | -2,1 |

corpus du *Trésor de la langue française*. Deux partitions sont disponibles : une première distingue « la prose, les vers, la prose poétique et la prose technique, soit respectivement 83 %, 3,6 %, 1 % et 12,3 % du corpus. L'absence de toute classification à l'intérieur de la prose littéraire est évidemment regrettable mais cette lacune est compensée en partie par une autre partition du corpus entier en trois sous-ensembles, fondés sur la personne dominante, et mieux équilibrés, puisque la première personne (Soliloque) regroupe 18,8 % des textes, la seconde (Dialogue) 18,2 %, et la troisième (Reste) 63 %. » (1981a), p. 775-777, en corrigeant l'interversion des pourcentages de vers et de prose poétique par recoupement avec les tableaux et d'autres passages, par exemple p. 773.

« La définition de ces trois sous-ensembles a été confiée à la machine, qui a classé dans la 1re catégorie (Soliloque) les textes ou parties de texte où les pronoms et les adjectifs de la 1re personne représentaient 40 % au moins de l'ensemble des possessifs et personnels, dans la seconde (Dialogue) ceux des textes qui faisaient intervenir les pronoms-adjectifs de la seconde personne, dont la proportion devait atteindre ou dépasser 16 %, dans la dernière enfin (Reste) les textes qui n'avaient pas trouvé place dans les deux premières, et qui avaient recours plus volontiers à la troisième personne, sans démêler si la tournure est personnelle ou impersonnelle et s'il s'agit d'un roman ou d'un traité. » (1981a), p. 2, note 2. « Le calcul a été fait au niveau des sous-textes (le plus souvent des chapitres), pour affiner la pureté des classes et empêcher la neutralisation mutuelle des parties d'un texte », à la différence de la première partition qui opère « sur des textes entiers » (1981a), p. 777.

Concernant la prose technique, « ne nous méprenons pas toutefois sur la « technicité » des œuvres ainsi cataloguées : il s'agit de textes dont la valeur littéraire le dispute à l'intérêt scientifique et qui exploitent le vocabulaire plus général de la philosophie, de l'économie politique, de la morale et de certaines grandes synthèses scientifiques. Ce sont le plus souvent des essais ou des traités. » (1981a), p. 7-9.

**3 –** Les points de suspension échappent totalement à notre contrôle. On ne peut en effet les identifier qu'en recourant aux textes ou aux index et nous ne disposions ni des uns ni des autres. Nous pouvons cependant faire une estimation de leur importance, à partir d'un texte, l'*Émile*, dont nous avons désambiguïsé la ponctuation. Dans cet ouvrage sur 1836 occurrences du point 209 étaient à porter au compte de la suspension, soit moins de 3 %. En tenant compte d'une faveur plus grande des points de suspension au XIX$^{e}$ et au XX$^{e}$ siècles, on peut estimer leur pourcentage à 4 %, à quoi il faut ajouter les 40 000 occurrences du point d'abréviation et de siglaison. Les données du point doivent donc être diminuées de 5 %, si l'on veut isoler la ponctuation forte qui finit la phrase.

## – III –

**1 –** Le calcul de la longueur moyenne de la phrase devrait tenir compte aussi des noms propres et des mots étrangers qui ont été écartés lors du processus de lemmatisation. Les premiers comptent 2 234 448 occurrences, les seconds 434 379. Il faut enfin admettre – de façon quelque peu abusive – que les signes de ponctuation forte (point, interrogation et exclamation) finissent toujours la phrase et que la phrase a toujours besoin de l'un d'eux pour prendre fin. Le résultat auquel on aboutit est de 15,82 mots en moyenne par phrase pour la totalité du corpus (tableau 3) :

**Tableau 3. Longueur moyenne de la phrase**

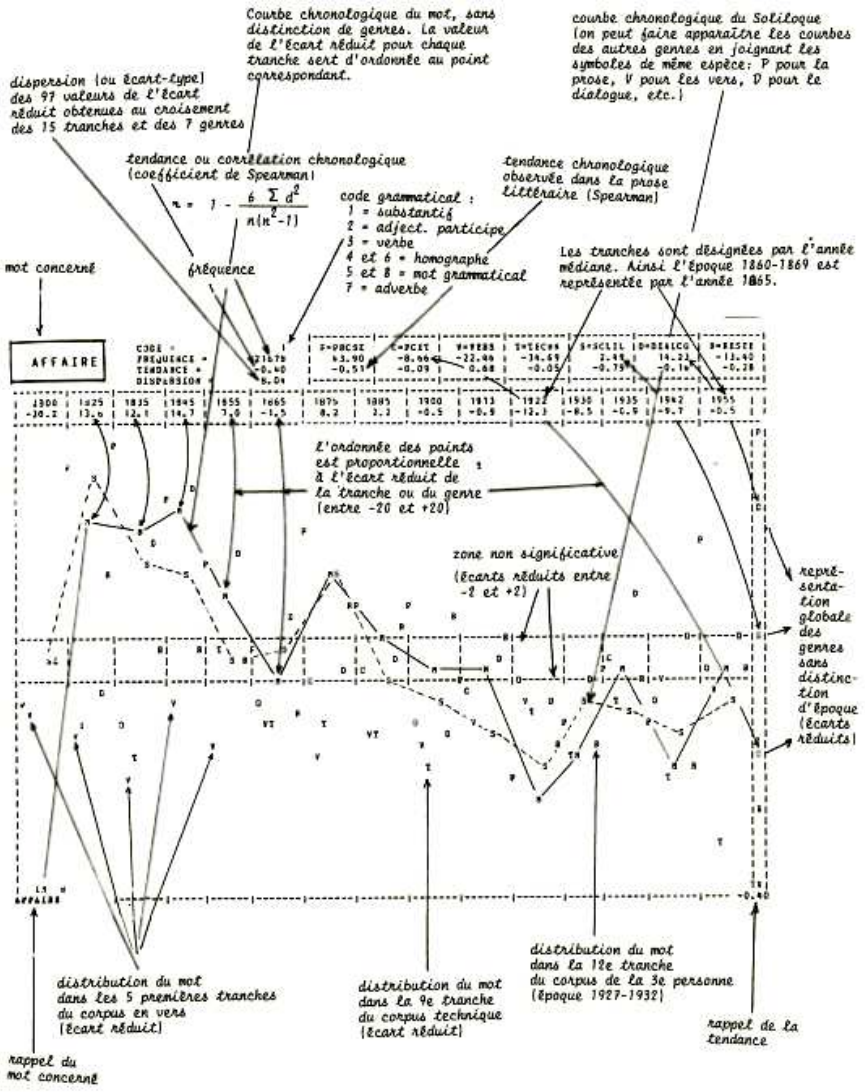
| | | | | |
|---|---|---|---|---|
| vocabulaire commun | 70 273 552 | point désambiguïsé | 3 835 858 | |
| noms propres | 2 234 448 | point d'interrogation | 308 920 | longueur moyenne |
| mots étrangers | 434 379 | point d'exclamation | 466 654 | de la phrase |
| Total mots | 72 942 379 | ponctuations fortes | 4 611 432 | 15,82 mots |

Or cette longueur varie grandement selon le genre et l'époque, ce qu'on peut montrer en faisant la somme des 3 signes de ponctuation forte dans chacun des sous-ensembles. Il revient au même en effet de calculer la densité des ponctuations fortes ou la longueur moyenne des phrases. Le tableau 4a reproduit la répartition observée.

**Tableau 4a. La densité des ponctuations fortes selon l'époque et le genre**

| | f | z | | f | z | | f | z |
|---|---|---|---|---|---|---|---|---|
| 1800 | 210487 | -314 | 1865 | 345995 | 91 | 1922 | 310035 | -36 |
| 1825 | 261354 | -153 | 1875 | 323475 | 92 | 1930 | 327653 | 91 |
| 1835 | 313710 | -56 | 1885 | 334185 | 0,4 | 1935 | 400864 | 202 |
| 1845 | 247931 | -62 | 1900 | 383175 | 66 | 1942 | 350200 | 40 |
| 1855 | 265794 | -44 | 1913 | 364997 | 145 | 1955 | 342199 | -53 |
| Prose | 4265137 | 325 | Solil. | 932680 | 35 | | | |
| P.Poet. | 45404 | -13 | Dial. | 1292140 | 491 | | | |
| Vers | 183624 | 20 | Reste | 2588499 | -421 | | | |
| Techn. | 319154 | -379 | | | | | | |

La densité des ponctuations fortes est plus faible au début du XIX^e^ siècle et plus forte à partir de 1865 (sauf dans deux tranches : 1922 et 1955). La phrase se raccourcit donc depuis 1789 : le coefficient de corrélation chronologique (0,60) est largement significatif. La phrase est beaucoup plus longue dans les textes techniques ($z$ = -379) que dans la prose littéraire ($z$ = +325), plus longue dans la prose poétique (-13) que dans les vers (+20), plus longue enfin quand domine la troisième personne (-421) que lorsqu'interviennent la 1^re^ personne (+35) et la 2^e^ (+491). Ces résultats n'ont rien de surprenant, sinon l'ampleur des écarts. Mais l'étude individuelle de chaque signe de ponctuation apporte des précisions intéressantes.




**Encart 4b. Comment lire les graphiques (ici pour le mot AFFAIRE, mais dans les graphiques suivants pour les différentes ponctuations considérées tour à tour)** [6]

6. NDÉ : Cet encart 4b ne fait pas partie de l'article original. C'est une figure annotée tirée des premières pages du *Vocabulaire français* (1981a), tome 3, donnant les clés de lecture des figures suivantes.

**2 –** Commençons par le point (figure 5) toutes fonctions confondues. Avec ses 4 millions d'occurrences, il ne représente que les deux tiers de la virgule. Mais il progresse, à telle enseigne qu'il rejoint presque la virgule dans la tranche 13 (345 725 contre 355 843). Ce progrès d'ensemble (corrélation chronologique : 0,68) est plus sensible dans la prose (0,72) que dans les vers (0,16) ou dans le corpus technique (0,27). La dispersion est l'une des plus élevées qu'on ait rencontrée dans le corpus (82,66). Même en réduisant d'un facteur 4 l'échelle des courbes, les écarts réduits sont si considérables qu'ils sortent des limites du graphique. La même amplitude des variations est observée dans le temps et dans le genre. Dans le temps l'écart maximum sépare l'époque 1800 ($z$ = -227) de l'époque 1935 ($z$ = +204). Dans le genre l'opposition la plus forte éloigne la seconde personne (+347) de la troisième (-315). Comme le point domine de loin l'ensemble des ponctuations fortes, il détermine la longueur de la phrase et les commentaires que nous avons faits là-dessus sont ceux qu'appelle la distribution du point.

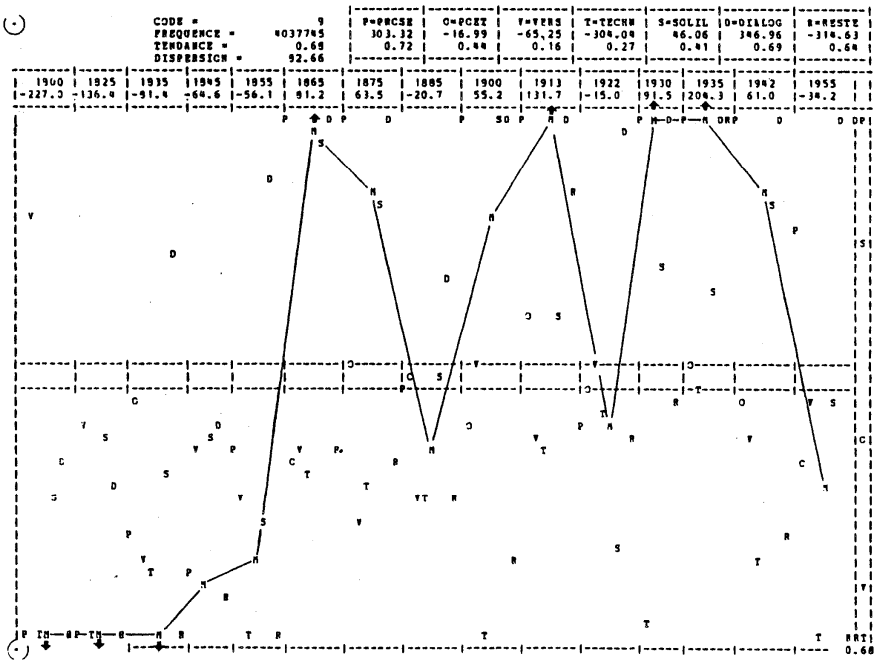



**Figure 5. Distribution du point**

**3 –** La virgule (figure 6) manifeste une courbe en cloche tout-à-fait singulière. Une montée régulière au XIX$^{e}$ siècle conduit à une apogée en 1885 (+166), suivie d'une régression au XX$^{e}$, la dernière tranche étant la

plus déficitaire (-84). Bien entendu le coefficient de Spearman (-0,15) est impuissant à rendre compte de ce double mouvement. L'examen des genres explique en partie cette évolution chronologique. Le discours technique répugne en effet à l'emploi de la virgule (-85) et comme les textes techniques sont particulièrement nombreux dans les deux tranches extrêmes, 1800 et 1955, il en résulte un affaissement de la courbe à ces deux endroits. La virgule accompagne volontiers la troisième personne (+75) et fuit la première (-89). Elle foisonne dans le vers (+88) mais aussi dans la prose littéraire (+40). En fait elle a partie liée avec le style descriptif ou narratif qui multiplie les circonstances et les adjectifs. La courbe de l'adjectif a le même profil que celui de la virgule et la même progression s'y remarque quand le goût croissant du pittoresque est transmis par le romantisme au réalisme puis au naturalisme.

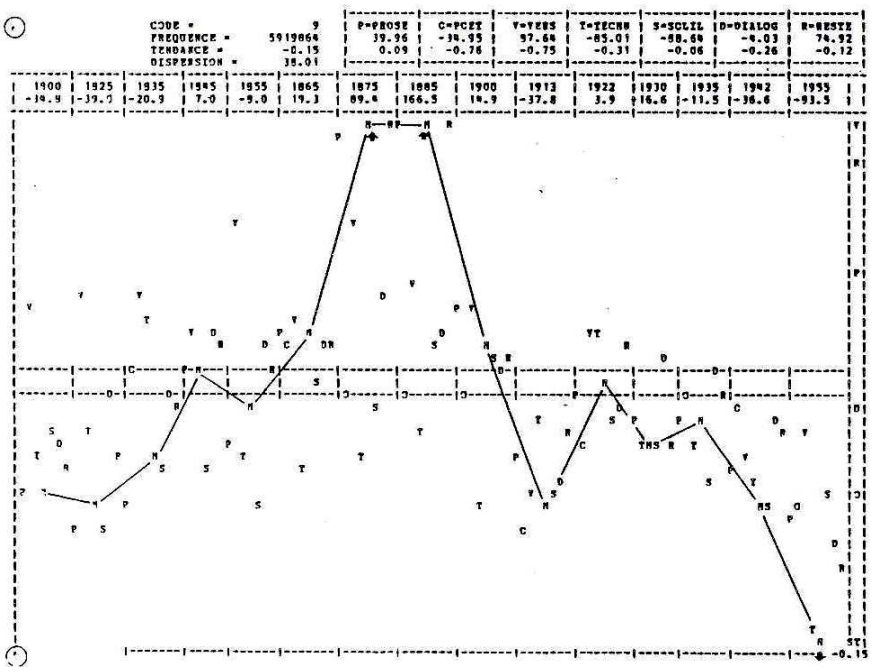



**Figure 6. Distribution de la virgule**

**4 –** Signe intermédiaire entre la faible pause marquée par la virgule et l'arrêt ferme indiqué par le point, le point et virgule (figure 7) a une position ambigüe et vulnérable qui explique son déclin constant depuis 1789. Le déclin est en effet général et s'observe dans la prose (-0,95), dans les vers (-0,77), et dans le technique (-0,89). Une simple

comparaison entre les deux tranches extrêmes résume cette régression : alors que ces deux tranches ont sensiblement la même étendue, la première contient presque trois fois plus de points et virgules (74 867 contre 27 056). On peut s'inquiéter sur l'avenir réservé à ce signe, pourtant l'un des plus anciens du système. Si l'on calcule l'espace moyen entre deux points et virgules dans chaque tranche, on voit qu'en 175 ans cet espace a triplé (tableau 8). L'extrapolation à partir de la droite de régression conduit à la disparition de ce signe dans quelques siècles.

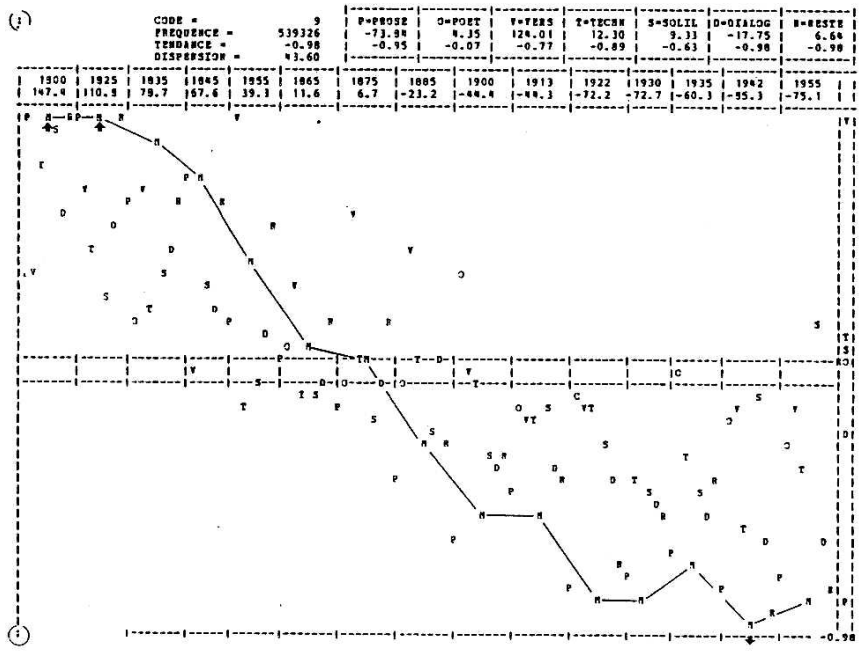



**Figure 7. Distribution du point-virgule**

**Tableau 8. Intervalle moyen entre deux points et virgules.**

| | | | | | |
|---|---|---|---|---|---|
| 1800 | 78,24 | 1865 | 122,74 | 1922 | 204,30 |
| 1825 | 84,58 | 1875 | 125,66 | 1930 | 216,29 |
| 1835 | 94,05 | 1885 | 147,31 | 1935 | 192,10 |
| 1845 | 95,08 | 1900 | 166,54 | 1942 | 228,40 |
| 1855 | 107,53 | 1913 | 171,11 | 1955 | 201,35 |
| 2022 | 283 | 2139 | 394 | 2955 | 1176 |

Le point et virgule réagit moins à l'endroit des genres qu'à l'égard du temps. Il faut noter toutefois une aptitude particulière à ponctuer le rythme poétique ($z$ = 124 dans le corpus en vers). Il est vrai que les vers étant trois fois plus nombreux au XIX$^{e}$ qu'au XX$^{e}$ siècle, la chronologie interfère avec le genre.

**5 –** Les deux points (figure 9) occupent dans le système une place particulière qui n'est nullement menacée. La courbe de ce signe est régulièrement montante et le coefficient de corrélation chronologique est positif et significatif ($r$ = 0,79), même si la tendance paraît à la baisse dans le corpus en vers (-0,45). La dispersion est faible (16,51) et les points de la courbe ne s'écartent guère de la ligne générale. Cette solidité repose en fait sur une diversité d'emplois spécifiques où ce signe ne rencontre pas de concurrents. Qu'il s'agisse d'introduire des propos rapportés, de présenter l'explication d'un fait ou d'annoncer une énumération, le recours aux deux points s'impose. S'y ajoute peut-être une nuance stylistique de simplicité rapide et nette qui voue ce signe plutôt à la prose littéraire (+21) qu'à la prose technique (-30).

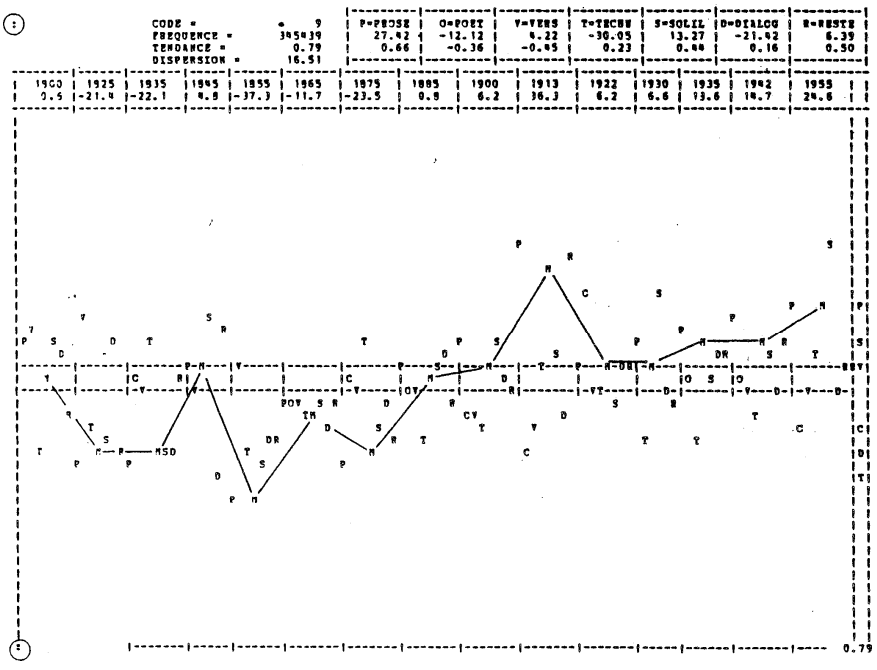



**Figure 9. Distribution des deux-points**

**6 –** Les caractéristiques du genre sont plus nettes dans le cas des signes affectifs, points d'interrogation et d'exclamation (figures 10 et 11). L'un et l'autre sont multipliés dans le dialogue (respectivement +261 et +344), Tous les deux préfèrent la prose littéraire à la prose technique (92 et 79 contre -119 et -227). Mais le point d'exclamation est plus fréquent en poésie, comme si le lyrisme comportait moins de demandes que d'émotions. Et surtout le mouvement de la chronologie s'y exerce différemment : le point d'interrogation est en progression ($r$ = 0,49) surtout en prose (0,65), alors que la courbe du point d'exclamation est parabolique comme celle de l'interjection à laquelle elle est étroitement corrélée ($r$ = 0,91). De même le point d'interrogation est associé au progrès des adverbes, adjectifs ou pronoms interrogatifs ($r$ = 0,60)[7].

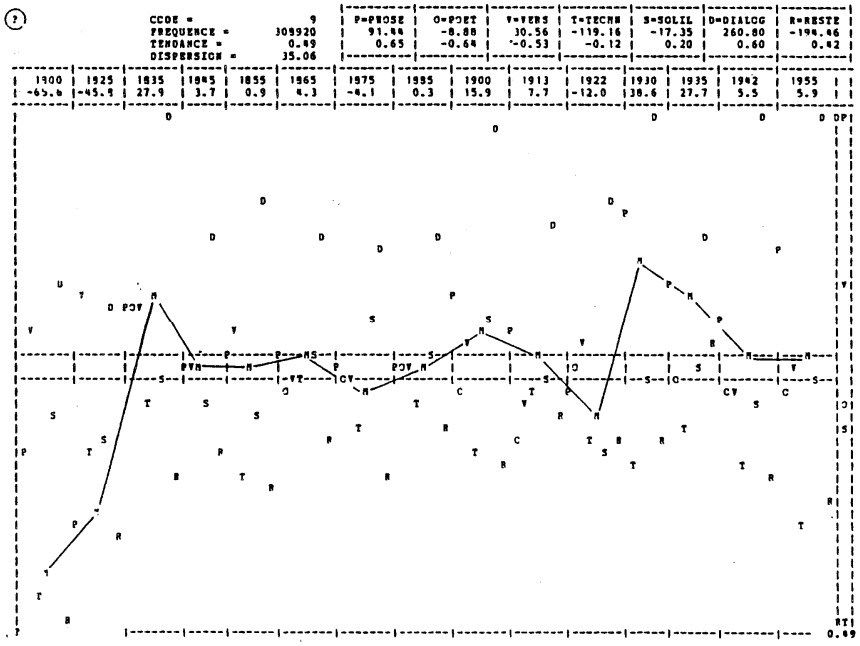



**Figure 10. Distribution du point d'interrogation**

La dispersion enfin du point d'exclamation est beaucoup plus forte (60,40 contre 35,06) et le nuage des points y remplit tout l'espace du graphique. C'est que le point d'exclamation est le signe d'une situation

7. La corrélation n'est pas parfaite parce que beaucoup des mots interrogatifs (QUEL, COMBIEN, COMMENT, etc.) peuvent introduire un mouvement exclamatif aussi bien qu'interrogatif.

extrême où le sentiment se risque à visage découvert – ce qui se produit dans la poésie et le dialogue. Mais là où domine la raison, ce signe trop violent et presque impudique est systématiquement rejeté. De là vient le fossé béant qui sépare les vers (+231) du technique (-227) et la deuxième personne (+344) de la troisième (-267).

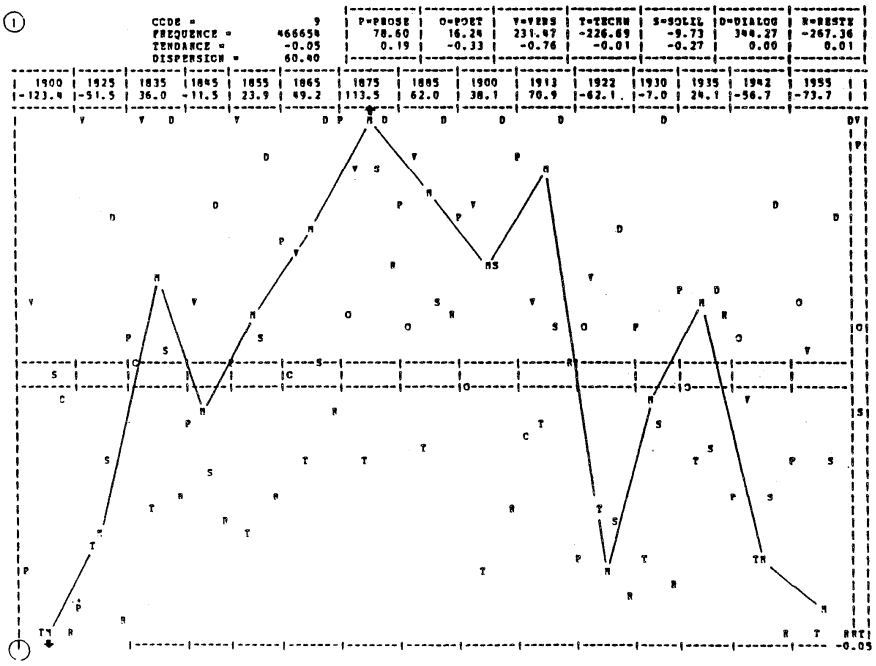



**Figure 11. Distribution du point d'exclamation**

## – IV –

Les six signes dont l'étude précède constituent le système de la ponctuation. Il aurait fallu ajouter les points de suspension – qui échappaient à notre pouvoir. Mais d'autres signes affectent le discours, dont nous avons pu faire le relevé. Les uns complètent la segmentation, notamment les parenthèses, d'autres sont des marques typographiques et métalinguistiques qui désignent un changement de locuteur (guillemets, italiques) ou tout au moins, pour certaines valeurs de l'italique, une distance plus proche – c'est l'italique de mise en relief – ou plus lointaine – c'est l'italique de l'allusion – que l'auteur établit avec son texte.

**1 –** Les parenthèses (figure 12) sont de plus en plus nombreuses dans le corpus. Le coefficient chronologique est significatif ($r$ = 0,61) et l'évidence visuelle reconnait une diagonale dans le graphique de ce signe,

qui progresse partout : en prose (+0,52), en vers (0,64) et dans le technique (0,57). Mais dans ce corpus littéraire la densité des parenthèses est de 408 mots, ou plus exactement de 816 si l'on tient compte du fait que les parenthèses vont par paires. Nul doute qu'elles seraient plus répandues dans une langue moins relevée, plus technique ou plus proche du langage parlé. Dans le discours spontané les parenthèses servent d'exutoire aux réserves, aux repentirs, aux précisions multiples qui accompagnent la pensée pensante. Mais dans le discours mûri à vocation littéraire, le souci du style a le plus souvent retouché ces esquisses trop rapides ou trop peu liées et notre corpus contient plus de tableaux que d'ébauches. Il est vrai que certains auteurs, bien loin de fondre les parenthèses dans le texte, en ajoutent de nouvelles au moment de la correction, jugeant le premier jet insuffisamment précis ou insuffisamment exact. Il en est ainsi souvent du texte de Proust.

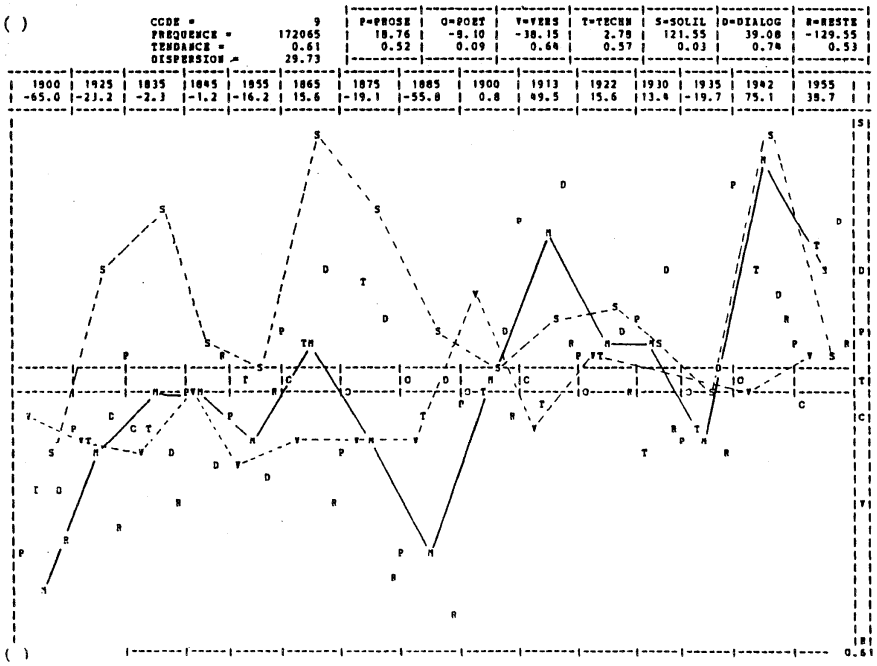



**Figure 12. Distribution des parenthèses**

La valeur esthétique de la parenthèse est peu appréciée des poètes (prose poétique -8 ; vers -38) et peu prisée par la troisième personne (-10). Dans le dialogue (+38) les parenthèses trouvent à s'employer, pour un aparté ou une indication scénique. Mais c'est la première personne qui recourt le plus volontiers à cette facilité d'écriture (+122). Le ton familier

de la correspondance, la volonté d'éviter le rythme oratoire et le désir de glisser une précision, un rapprochement, une restriction, un trait d'esprit, multiplient les parenthèses dans le soliloque, où seule devant son miroir la coquetterie de la première personne se permet le négligé.

On peut supposer que l'emploi du tiret est assez voisin de celui de la parenthèse. Mais ce symbole est ambigu : simple liaison graphique quand il précède le pronom personnel, ou quand il marque le début d'une réplique au théâtre, il acquiert en d'autres occasions le statut logique de la parenthèse. Nos données ne nous ont pas permis d'isoler les diverses valeurs de ce symbole, dont nous ne pouvons rien dire.

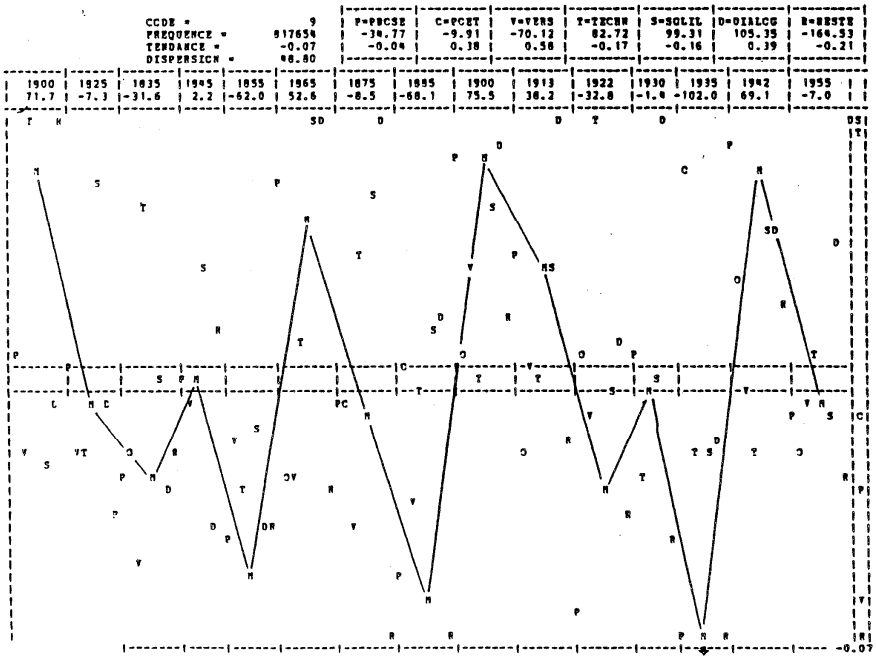



**Figure 13. Distribution de l'italique**

**2 –** Restent l'italique et les guillemets. L'italique (figure 13) se marque par un changement de police typographique qui ne peut avoir d'équivalent sur un support informatique. On utilise un symbole au début et à la fin de la zone en italique laquelle peut être un mot, un groupe de mots, une phrase ou un paragraphe. Quand cette zone recouvre plusieurs lignes, le signal est répété à l'entrée de chaque ligne. Le nombre d'occurrences de l'italique n'est donc pas celui des mots mis en italique, ni celui des passages, mais un nombre intermédiaire dont la précision ne peut être poussée plus avant. Pour un mot isolé il faut compter deux

symboles mais pour une ligne deux suffiront aussi et pour un passage plus long il y aura autant de symboles que de lignes. Mais il n'est pas nécessaire de raisonner sur le nombre des mots en italique. L'étude quantitative du symbole lui-même suffit. On remarquera dans la courbe qui lui est propre l'extrême dispersion de sa distribution (sur 97 sous-ensembles 1 sur 3 atteint un écart réduit de 50 en valeur absolue). On ne constate pas de tendance chronologique affirmée mais une série de variations fortes, en dents de scie. La réaction au genre est vive : positive à l'égard de la première et de la seconde personne (soliloque $z = 99$, dialogue $z = 105$), négative à l'égard de la troisième ($z = -165$). On n'attendait pas l'italique dans le dialogue mais il faut se souvenir que les indications scéniques s'écrivent en italique dans une pièce de théâtre. L'italique est rare en poésie, comme prévu, mais pourquoi constate-t-on la même rareté dans la prose littéraire et un excédent dans le corpus technique – ce qui s'accorde mal avec le choix des trois personnes. Ces faits manquent de clarté, ils correspondent sans doute aux emplois fort divers de l'italique qui peut marquer une insistance, une ironie, une citation, un mouvement scénique, un emprunt avoué, un néologisme déclaré.

**3 –** Les guillemets (figure 14) peuvent partager beaucoup de ces fonctions, surtout sur les machines à écrire usuelles qui disposent d'une seule police de caractères. Mais dans la tradition typographique ils servent surtout à encadrer un mot prêté à quelqu'un d'autre ou des propos rapportés expressément. Comme les parenthèses ils vont par paires et en principe on ne les répète pas à chaque ligne. Leur répartition dans les genres est assez claire : les guillemets interviennent dans la prose littéraire et principalement dans le roman ($z = 70$ en prose et $z = 14$ pour la troisième personne). Qu'on ne s'étonne pas de leur absence dans le dialogue ($z = -74$) où les propos sont reproduits sans être introduits : les guillemets marquent la séparation du discours narratif et des paroles rapportées et ils perdent leur nécessité quand le récit disparaissant, seuls subsistent les échanges. Le fait saillant du graphique est d'ordre historique : c'est la progression très nette et très régulière de ce symbole depuis 1789. Le coefficient de corrélation ($r = 0{,}93$) est hautement significatif. Alors que les guillemets n'apparaissaient qu'au bout de 400 mots en moyenne dans la première tranche, l'intervalle diminue constamment et se situe aux alentours de 100 mots dans la dernière tranche. On en rencontre un par page en 1800 et quatre à l'époque actuelle. Quelle conclusion tirer de là ? L'importance croissante du discours direct est sans doute l'indice d'une nouvelle relation de

l'écrivain au monde qu'il crée ou qu'il décrit. Dans le discours contemporain les gens parlent plus volontiers. Il est vrai que les héros de Homère ou de Chrétien de Troyes prennent aussi souvent la parole. Mais dans la prose de notre temps les guillemets ont un rôle plus complexe : non seulement ils jouent le rôle du micro dans une prise directe, mais ils servent aussi à l'enregistrement et à la restitution différée. Il s'agit alors de la citation. De ces deux fonctions l'une est plus fréquente dans le récit et l'autre dans les essais. Nous n'avons pas hélas le moyen de les distinguer. Et nous ne pouvons que constater un effet global dont le progrès chronologique est néanmoins fort net.

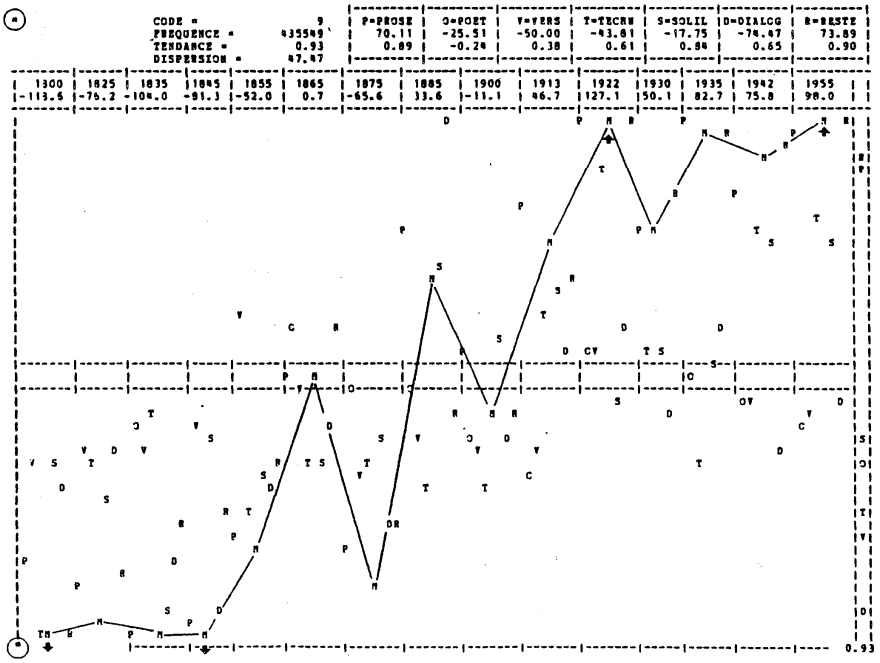



**Figure 14. Distribution des guillemets**

**4 –** L'étude de la majuscule (figure 15) est à la fois très proche et très éloignée de la ponctuation. S'il s'agit de l'initiale de la phrase, l'effet est mécanique et son étude serait redondante, venant après celle des signes de ponctuation forte. Cette majuscule a été écartée, n'ayant d'autre valeur que typographique.

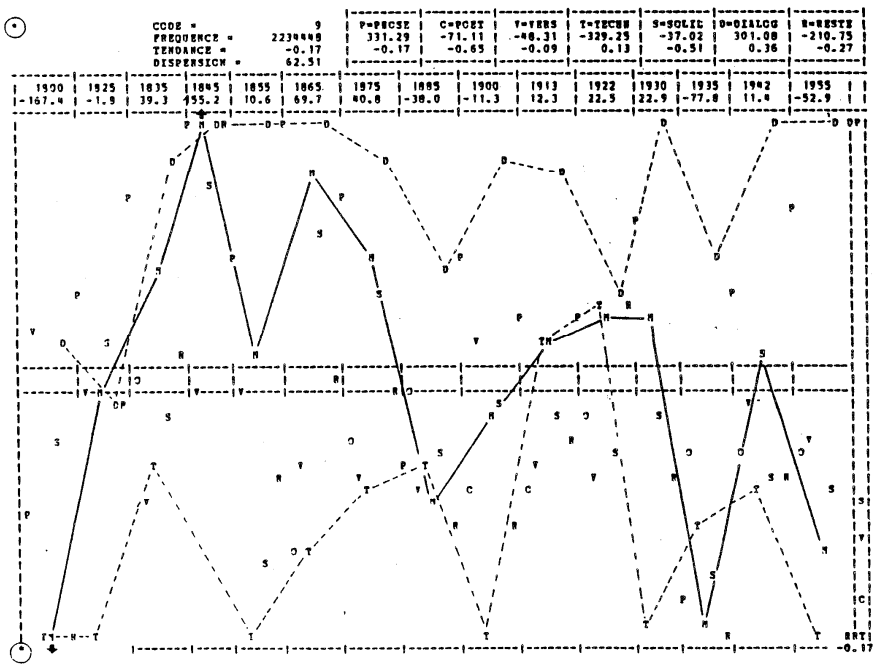


**Figure 15. Distribution des noms propres**

Reste la majuscule des noms propres, que nous considérerons ici. Ces derniers auraient mérité une étude indépendante[8] : mais nous n'avons dans ce corpus qu'une trace de leur présence, un code (l'astérisque) dont on les avait pourvus à l'enregistrement. La statistique s'applique rarement à cette catégorie lexicale et ils ont été écartés du *Trésor de la langue française*. Si nous abordons leur étude rapide ici, après l'étude du discours cité, ce n'est pas seulement par un rapprochement superficiel suggéré par l'artifice des codes typographiques. L'italique (ou les guillemets) et la majuscule que signalent dans l'écrit des symboles surajoutés, ont en commun quelque chose de plus profond que Jakobson analyse dans un article consacré aux embrayeurs[9]. Il s'agit d'une circularité dans les relations qu'entretiennent le code et le message. Mais tandis que le discours cité est un message qui renvoie au message

8. NDÉ : voir par exemple dans ce volume le chapitre 13, « Les noms propres chez Zola » (1987a).

9. R. Jakobson, *Essais de linguistique générale*, p. 176-196, Paris, 1968. La première partie de l'article « Embrayeurs et autres structures doubles » (p. 176-181) reprend en l'abrégeant le texte d'une communication de 1950, malheureusement peu accessible : « Overlapping of code and message in language », University of Michigan.

(circularité M/M), le nom propre met en jeu un code qui renvoie au code (circularité C/C). Sauf le cas des noms propres devenus des types (un Néron), la signification des noms propres « inclut une référence patente au code » : « Si beaucoup de chiens s'appellent Fido, ils n'ont en commun aucune propriété spéciale de Fidoïté ».

Nous avons compté plus de deux millions de noms propres dans notre corpus, soit un mot sur 32 en moyenne. La dispersion est considérable : certains genres les attirent, d'autres les repoussent. Parmi les premiers la prose littéraire ($z$ = +331) et le dialogue ($z$ = +301) ; parmi les seconds la poésie (prose poétique -71, vers -48), le technique (-329), et la troisième personne (-211). Comment expliquer cette répartition ? Il faut prendre garde d'abord que la majuscule, peut désigner un nom de personne (ou d'être individualisé), ou un nom de lieu, que dans les deux cas les noms fictifs cohabitent avec les noms réels, que les uns et les autres ont des substituts, les embrayeurs, (IL peut désigner Dupont et ICI peut renvoyer à Paris). La confusion dans le même ensemble de fonctions et de classes si diverses rend l'interprétation difficile. On croit pouvoir avancer cependant que le genre technique, qui évolue le plus souvent dans l'abstrait, s'accorde mal avec la valeur concrète attachée au nom propre. Mais cela tient en partie à la nature des textes rassemblés sous la rubrique technique. Si les ouvrages historiques y étaient plus nombreux, la proportion des noms propres y serait fort différente, puisque les hommes et les lieux – et les noms propres qui les désignent – deviendraient la matière même du discours. Quant à la corrélation entre le dialogue et les noms propres (surtout les noms de personne), on peut supposer qu'elle tient à une situation de relation interpersonnelle où interviennent non seulement le TU et le MOI mais aussi les tierces personnes nommément citées. La situation de dialogue engendre aussi, parmi les interjections et les exclamations, des interpellations ou vocatifs qui multiplient les noms personnels et plus encore les prénoms. Si l'on rapproche la distribution de ceux-ci de celle des pronoms TU, TE, TOI et VOUS, on obtient un coefficient de corrélation de 0,69 largement significatif, tandis que le classement selon la proportion des textes techniques dans les tranches s'écarte significativement de celui des noms propres ($r$ = -0,55) (tableau 16).

**Tableau 16. Évolution comparée des noms propres, de la proportion de textes techniques et des pronoms de 2e personne**

| | a | b | c | | a | b | c | | a | b | c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1800 | 15 | 15 | 1 | 1865 | 2 | 5 | 12 | 1922 | 5 | 14 | 8 |
| 1825 | 10 | 13 | 6 | 1875 | 3 | 4 | 11 | 1930 | 6 | 7 | 3 |
| 1835 | 4 | 1 | 10 | 1885 | 12 | 2 | 9 | 1935 | 14 | 11 | 13 |
| 1845 | 1 | 2 | 15 | 1900 | 11 | 10 | 4 | 1942 | 8 | 8 | 5 |
| 1855 | 9 | 3 | 7 | 1913 | 7 | 6 | 14 | 1955 | 13 | 9 | 2 |

en a classement selon les noms propres
en b classement selon les pronoms de la deuxième personne
en c classement selon l'importance des textes techniques

La chronologie montre d'amples variations et surtout une rapide montée dans la première moitié du XIXe qui correspond à un goût de plus en plus affirmé pour l'individualité des personnes et des lieux. Le mythe du héros, la légende napoléonienne, la libération de l'individu, mais aussi l'exploration dans le temps, et la faveur retrouvée des grandes figures du temps jadis, des lieux oubliés et des coutumes locales et tout un mouvement de découverte de l'histoire et de l'espace géographique provoquent en peu d'années la multiplication des noms propres, à laquelle le mouvement romantique contribue autant que le réalisme balzacien. L'exemple le plus frappant de cette tendance est constitué par les *Mémoires d'outre-tombe*, où l'espace moyen entre deux noms propres est de 22,76 mots, inférieur de 40 % à celui de l'ensemble du corpus[10], et presque égal à celui de l'apogée (1845 : 22,18). À l'opposé cet intervalle tombe à 50 dans la période 1800 où le mouvement des Idéologues fréquente plus souvent les idées que les êtres et les lieux réels.

10. Cf. notre article « Les noms propres dans l'œuvre de Chateaubriand », *Annales de la Faculté des Lettres de Nice*, n° 38, 1979, p. 83-94 (1979b).

**– V –**

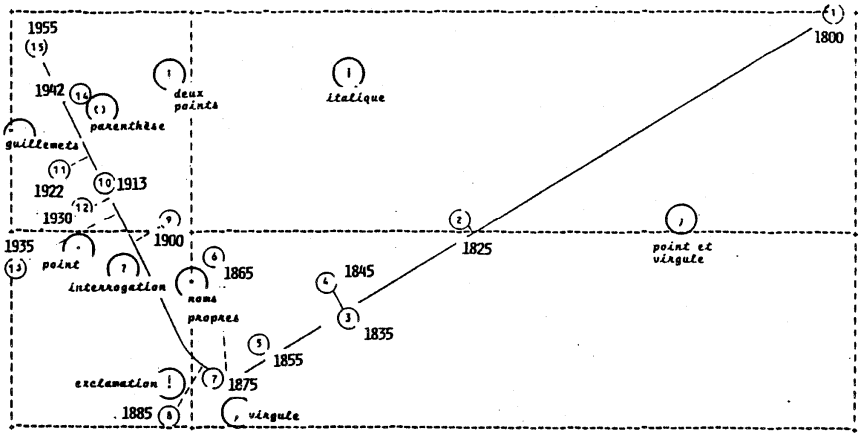



**Figure 17 Analyse factorielle selon la chronologie**

Pour résumer toutes les observations qui précèdent concernant la ponctuation et certaines caractéristiques typographiques du corpus, nous avons fait appel une nouvelle fois à l'analyse factorielle. Celle de la figure 17, qui porte sur la chronologie, fait apparaître un cortège orienté de droite à gauche où les tranches se suivent dans un ordre régulièrement chronologique. Sur le chemin les signes de ponctuation se dispersent selon leur préférence, le point et les deux points rejoignent la tête du cortège (époques récentes), et le point et virgule la queue de la colonne (entre 1800 et 1825) tandis que la virgule choisit une position intermédiaire en compagnie du point d'exclamation. C'est au milieu de la colonne que se place aussi la majuscule, tandis que guillemets et parenthèses se portent à l'avant-garde. Nous retrouvons bien là les enseignements des courbes individuelles.

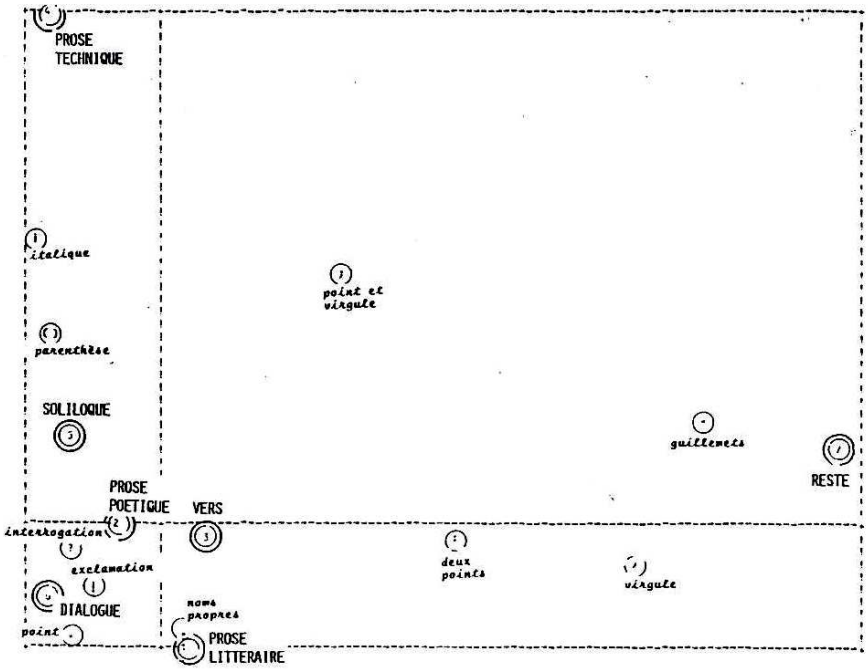



**Figure 18. Analyse factorielle selon le genre**

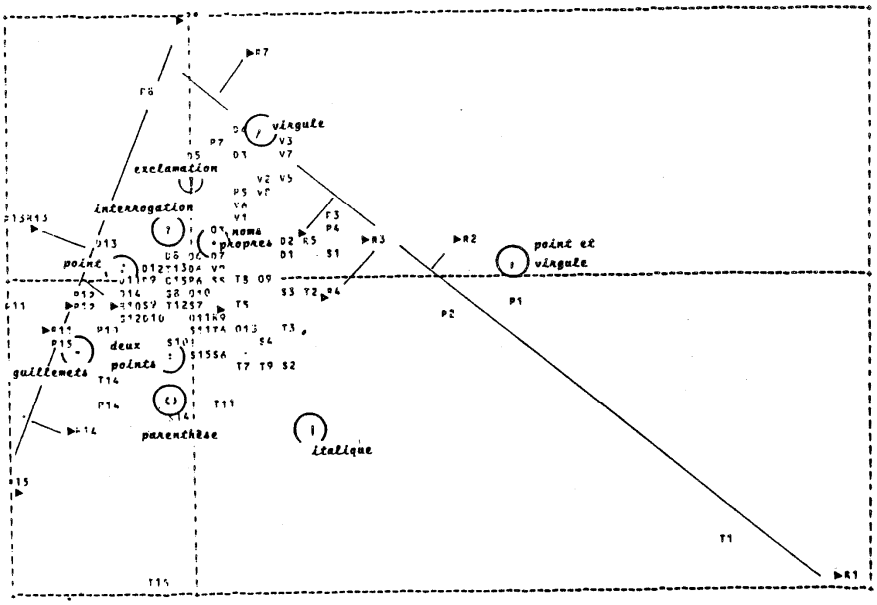



**Figure 19. Analyse factorielle selon l'époque et le genre**

La figure 18 rend compte des forces stylistiques qui structurent les genres et qui opposent violemment la troisième personne (à droite) aux deux autres (à gauche) et la prose technique (en haut du graphique) à la prose littéraire (en bas). Les signes faibles (virgule, point et virgule, deux points) se portent à droite, avec les guillemets, dans la zone du récit (3e personne), les signes forts (point, interrogation, exclamation) entourent le dialogue et la prose littéraire. Aucun signe ne rejoint la position excentrique du technique sinon, à quelque distance, l'italique et la parenthèse.

Reste à démêler les deux influences, ce qu'on peut faire en présentant à la fois l'époque et le genre à l'analyse factorielle (figure 19). Contrairement à ce que nous avions coutume d'observer, c'est ici la chronologie qui intervient d'abord et s'installe dans le premier facteur, là où nous trouvions généralement le genre. À droite se portent les sous-ensembles du XIX$^{e}$ siècle, à gauche ceux du XX$^{e}$. Et en se dirigeant de la gauche à la droite on rencontre les signes de ponctuation dans l'ordre même où ils s'étaient présentés quand la chronologie était seule en cause (graphique 17). L'intervention du genre a donc moins d'influence dans ce domaine que la chronologie. On peut s'expliquer cet asservissement au temps, si l'on se souvient que le système de la ponctuation est encore flottant à l'orée du XIX$^{e}$ siècle. Si le système est alors complet, il n'est pas encore stabilisé. Deux siècles après, il ne semble pas encore définitivement fixé : et un de ses éléments (le point et virgule) semble en voie de disparition. Quand on songe à la fossilisation de l'orthographe, on mesure le jeu et la souplesse du système des ponctuations. Cela est si vrai que la notion de faute de ponctuation existe à peine. Dans le plus répandu des exercices littéraires, la dictée, seules les fautes d'orthographe sont pourchassées et le maitre d'école précise aimablement et à haute voix la ponctuation : « Les poules étaient sorties du poulailler *virgule* en rangs serrés *virgule* dès qu'on[11] avait ouvert la porte *point* ». Dans la dictée musicale au contraire, parce que la ponctuation rythmique y a des règles sûres, le maitre de musique tait les silences et les soupirs.

11. La légende veut qu'un gendarme ait fait une faute à cet endroit.