

Les phrases de Marcel Proust

Cyril Labbé, Dominique Labbé

▶ **To cite this version:**

Cyril Labbé, Dominique Labbé. Les phrases de Marcel Proust. 14th International Conference on Textual Data Statistical Analysis, Department of Mathematics and Statistics - University of Roma "la Sapeienza", Jun 2018, Roma, Italie. pp.400 - 410. halshs-01818296

HAL Id: halshs-01818296
https://shs.hal.science/halshs-01818296

Submitted on 18 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cyril Labbé[1], Dominique Labbé[2]

Les phrases de Marcel Proust

Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 2018, p. 400-410.

1. Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, F-38000 Grenoble France (cyril.labbe@imag.fr)

2. Univ. Grenoble Alpes, PACTE (dominique.labbe@umrpacte.fr)

Abstract

Analysis of sentence lengths in Marcel Proust's *A la recherche du temps perdu*. Counting standards and the various available measures are presented. For most of his reading time, the reader of this novel is confronted with very long and syntactically-complex sentences. A comparison with other writers shows that these sentences are atypical but not unique and that some of their characteristics can be observed in a number of other works, some of which are cited in the *Recherche du temps perdu.*

Résumé

Analyse des longueurs de phrases dans *A la recherche du temps perdu* de Marcel Proust. Présentation des normes de dépouillement et des différentes mesures possibles. Durant la majorité de sa lecture, le lecteur se trouve confronté à des phrases très longues et syntaxiquement complexes. Une comparaison avec un large panel d'écrivains montre qu'il s'agit d'un phénomène exceptionnel mais pas unique et que certaines caractéristiques se retrouvent dans quelques œuvres dont certaines sont citées dans la *Recherche du temps perdu*.

Keywords: lexicometry - stylometry - sentence length – French literature – Proust

Manuscrit des auteurs. Toute citation doit être faite à partir de l'ouvrage cité ci-dessus.

Les corpus utilisés pour cette communication sont disponibles auprès des auteurs.

1. Introduction

Les phrases de Marcel Proust (1871-1922) sont-elles exceptionnelles ? La question a été surtout traitée sous l'angle qualitatif (notamment Curtius 1970). Il existe quelques estimations quantitatives (Bureau 1976, Brunet 1981, Milly 1986), avec des résultats divergents pour des raisons qui seront explicitées au début de cette communication. Mais surtout, nous présentons une comparaison statistique avec d'autres écrivains qui permettra de juger de l'exceptionnalité de la phrase proustienne.

L'analyse des phrases soulève plusieurs des problèmes auxquels est confrontée la lexicométrie (statistique appliquée au langage). En premier lieu, ici, il y a le choix de l'édition de référence. En effet, pour la *Recherche du temps perdu*, ce choix existe et introduit une légère incertitude concernant la ponctuation de l'oeuvre (discussion dans Ferré 1957 et Serça 2010), spécialement pour les trois derniers volumes. Nous nous sommes tenus au principe général selon lequel fait foi l'ultime version révisée par l'auteur ou, à défaut, la plus proche de sa mort. Il s'agit ici de l'édition originale chez Gallimard (annexe 1). De plus, cette édition originale s'impose puisqu'elle est dans le domaine public et peut être communiquée librement aux chercheurs soucieux de reproduire nos résultats et d'aller plus loin dans cette analyse.

2. Le mot et la phrase

Le mot est défini comme l'occurrence d'un vocable, c'est-à-dire une entrée dans le lexique de la langue française selon la norme présentée par Muller 1963. Cette norme est fondée notamment sur la nomenclature de Hatzfeld et al. 1898. Son implémentation est décrite dans Labbé 1990. Par exemple, "aujourd'hui", "parce que" ou "Saint-Loup" sont des mots uniques et non deux "formes graphiques". Il y a 1 449 "parce que" dans la *Recherche*, soit plus d'un mot pour mille ; et 787 fois "Saint-Loup" (l'un des principaux personnages du roman). A l'inverse, les formes graphiques "le", "la", "les" ont deux entrées (pronom ou article) ; "du" ou "des" sont la contraction de deux entrées du lexique - préposition "de" et article "le". En fonction de la norme retenue (vocable ou formes graphiques), le nombre de mots dans un texte peut varier de près 10%. Selon cette "norme Muller", la *Recherche* compte 1 327 859 mots (N dans la suite) et 21 836 vocables différents.

Quant à la phrase, il y a un accord général pour la définir comme l'empan de texte dont le premier mot comporte une majuscule initiale et qui se trouve compris entre deux ponctuations majeures. Les ponctuations majeures sont le point, les points d'interrogation et d'exclamation, les points de suspension. Cependant, aucun de ces 4 signes typographiques ne marque automatiquement une fin de phrase :
- le point dans « M. Verdurin » ne termine pas une phrase même s'il est suivi d'un mot à majuscule initiale. Il y a dans la *Recherche* 3 152 « monsieur » écrits "M.". C'est le deuxième substantif le plus fréquent dans la *Recherche* (juste derrière "Mme"), soit 2,4 pour mille mots.

Ce point "non-terminal" se retrouve dans les initiales que Proust utilise pour "anonymiser" certains noms (Mme X.) ou derrière des abréviations (etc.).
- dans la *Recherche,* plus de trois points d'interrogation sur 10 sont internes à la phrase (721).
- il y a 1 201 points d'exclamation internes à la phrase et 190 points de suspension également dans cette situation. Proust a plusieurs fois déclaré son hostilité envers ces derniers mais il les utilise parfois. Par exemple : « La duchesse émit très fort, mais sans articuler : « C'est l'... i Eon l... b... frère à Robert. » (*la Prisonnière*).

Cette rapide discussion permet de comprendre qu'on ne peut rapporter le nombre de mots d'un ouvrage à l'effectif de ses ponctuations – comme le fait Brunet (1981) - pour obtenir une longueur moyenne. Il y a lieu de localiser précisément chacune de ces fins de phrases. Pour ce faire, un automate place, dans le texte, des balises localisant les fins de phrase et, en cas de doute, l'opérateur choisit : fin de phrase ou ponctuation interne ? A condition que l'opérateur suive toujours la même norme, le dépouillement est fait sans erreur et, surtout, les résultats obtenus sur un auteur sont comparables à ceux de tous les autres.

Ce recensement établit le nombre de phrases de la *Recherche* (voir tableau en annexe). Au total, P = 37 336 phrases.

Comment caractériser ces phrases en fonction de leurs longueurs ?

3. Les indices statistiques usuels.

Les P phrases sont rangées par longueur croissante, dans des classes d'intervalles égaux (ici 1 mots). Par exemple, la première classe (1 mot, généralement une exclamation) contient 124 phrases, soit 0,37% du total. L'effectif de chaque classe est ainsi recensé et son poids relatif est calculé. Ce recensement fournit les informations suivantes :

- **Etendue** de la distribution : 1 à 931 mots. La plus longue phrase est celle sur les homosexuels au début de *Sodome et Gomorrhe*. Les phrases de la *Recherche* ne sont pas réparties uniformément sur cet intervalle. La seconde plus longue – celle sur les chambres au début de *Combray* – compte 542 mots ; la troisième (le salon des Verdurins dans la *Prisonnière*) : 430 ; la quatrième (l'église de *Combray*) : 399. Ensuite, il n'y a plus de "trou" important dans l'étalement des longueurs.

- Le **mode** est la classe la plus peuplée, ou longueur de phrase que le lecteur a le plus de chance de rencontrer : 11 mots. Il y a donc, dans la *Recherche,* une prédominance des phrases courtes et syntaxiquement simples. Il en est ainsi dans la plupart des textes en français.

- La **médiane** est la valeur de la variable pour l'individu du milieu ou individu "médian". Dans les P phrases rangées par longueurs, l'individu médian est celui qui occupe la place (P+1)/2. Lorsque l'effectif total de la population (P) est pair, la médiane est la moyenne des valeurs de la variable pour les 2 individus situés de part et d'autre. Dans un texte étendu comme la *Recherche,* la médiane se trouve dans une classe dont l'effectif est assez élevé. Dans ce cas, la valeur est interpolée en divisant l'intervalle de la classe où se situe l'individu médian par l'effectif de cette classe. Dans la *Recherche,* ce calcul aboutit à une médiane de 26,28 mots. Etant donné que la variable "longueur de phrase" ne prend que des valeurs entières, les décimales indiquent le sens de l'arrondi et la position de la borne. La longueur

médiane des phrases de la *Recherche* est donc de 26 mots. Ou encore la moitié des phrases ont une longueur inférieure ou égale à 26 mots et l'autre moitié une longueur supérieure à 26.

- La **moyenne** (N/P) : 35,57 mots. A cet indice est associée une déviation "standard" des valeurs de la variable autour de la moyenne (écart-type) : racine carrée de la variance (moyenne des carrés des écarts de chaque valeur de la variable à la moyenne arithmétique). L'écart type de la longueur des phrases de la *Recherche* est de 31,42 mots.

La **dispersion** des valeurs autour de la moyenne mesurée par le coefficient de variation relative : rapport de l'écart-type à la moyenne arithmétique (ici 89%). Etant donné l'effectif considéré (37 336 phrases), si les valeurs de la variable "longueur de phrase" étaient distribuées normalement autour de la moyenne (cas d'une population homogène), ce coefficient serait d'environ 4%. Autrement dit, les observations sont extrêmement dispersées. Dans ce cas, la moyenne n'est pas représentative de la série et, en particulier, il n'est pas possible de considérer que cette moyenne se situe à peu près "au milieu" de la population. Dès que la dispersion relative approche les 50% de la moyenne, celle-ci est située dans la partie basse de l'étendue de la distribution qui est fortement asymétrique. Le profil de la distribution des longueurs de phrases dans la *Recherche* est donné par la figure 1 dans laquelle l'effectif relatif de chaque classe est représenté par la hauteur du bâton correspondant (histogramme).

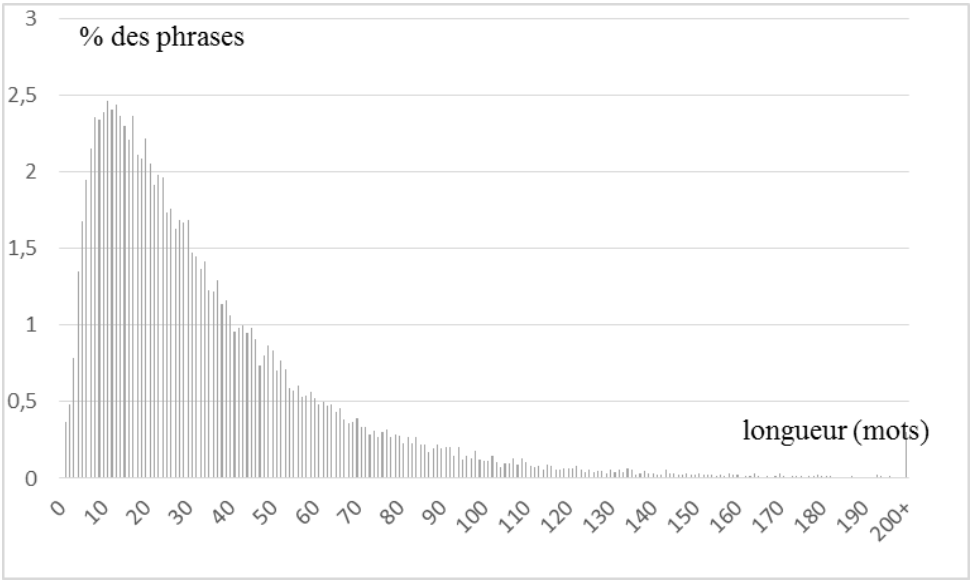


Figure 1. Histogramme de la distribution des longueurs phrases

D'une part, le graphique s'interrompt à la classe 200+ mots et le bâton pour cette classe – à l'extrême-droite du graphique - correspond aux 96 phrases longues de 200 mots et plus (0,3% du total des phrases mais 2,1% de la surface du texte). Le graphique complet est encore plus étalé sur la droite, la grande masse des phrases apparaissant serrées sur la gauche… D'autre part, le bâton le plus haut correspond au mode principal (11 mots) mais l'on observe de nombreux modes secondaires (17, 20, 24, etc.) : plusieurs populations sont donc mélangées.

La plupart des phénomènes sociaux présentent des caractéristiques semblables et, en premier lieu, la distribution des revenus ou des patrimoines. Dans de pareils cas, l'analyse ne se

contente pas des valeurs centrales. Elle se centre sur la distribution du caractère étudié (ici la surface du texte) au sein de la population (ici les phrases).

4. L'inégal partage de la surface du texte entre les phrases

Ce renversement de perspective présente un avantage : la surface de texte correspond grosso-modo à la durée de la lecture. Deux méthodes sont possibles pour l'évaluer.

4.1 Quantile et médiale

Les phrases étant classées par longueurs croissantes, la surface du texte qu'elles couvrent est découpée en masses égales (tableau 1).

Surface divisée en quantiles	Longueur (mots)	% des phrases (cumulé)
Premier décile	18.58	33,8
Deuxième décile	26.70	49,6
Premier quartile	29.53	54,5
Troisième décile	33.30	60,6
Quatrième décile	41.35	70,1
Deuxième quartile (médiale)	**49.93**	**77,5**
Sixième décile	60.20	84,6
Septième décile	72.93	89,7
Dernier quartile	81.13	92,3
Huitième décile	90.57	94,2
Neuvième décile	121.00	97,8

Tableau 1. Partage de la surface du texte en fonction de la longueur des phrases

Dans ce tableau, le premier décile est la borne supérieure de l'intervalle comprenant les phrases les plus courtes couvrant en tout 10% de la surface du texte et la borne inférieure du 2e décile. Il indique que les phrases de longueurs inférieures ou égales à 18 mots couvrent 10% du texte et représentent plus du tiers du total des phrases (33,8%). Le lecteur n'y passe au mieux qu'un dixième du temps de la lecture. Or c'est au-dessus de cette longueur que l'on commence à rencontrer des phrases syntaxiquement complexes. Autrement dit, au mieux, le lecteur de la *Recherche* se trouve face à des phrases simples pendant un dixième de sa lecture (ou il est face à des phrases plus ou moins complexes pendant les neuf dixièmes !)

A l'opposé, 2,2% des phrases (700) comptent plus de 121 mots (9e décile). Elles couvrent également 10% du texte, c'est-à-dire la même surface que le tiers évoqué ci-dessus. Cela signifie que le lecteur de la *Recherche* passe (au moins) autant de temps à lire des phrases très longues – dont la construction est nécessairement complexe -, qu'il n'en consacre à la masse des phrases les plus brèves et structurellement simples.

Dans cette perspective, la valeur centrale la plus caractéristique est la longueur de la phrase qu'il faut atteindre pour avoir lu la moitié du texte. Pour éviter les confusions, cette seconde médiane est appelée **médiale** (Ml). Elle correspond à la borne haute du cinquième décile (ou du deuxième quartile). Dans la *Recherche*, elle est égale à 49,93 mots, soit 50 mots. Le tableau indique que 77,5% des phrases (près de 8 sur 10) sont inférieures à cette médiale. Autrement dit, les lecteurs de la *Recherche* passent au moins la moitié de leurs temps confrontés à des phrases de 50 mots et plus, ce dont la plupart d'entre eux n'ont guère l'habitude. Malgré le talent de l'écrivain, c'est évidemment cela que les lecteurs retiennent.

4.2 Mesure de l'inégalité

Deuxième méthode, un indice unique mesure l'inégale répartition de la surface du texte entre les phrases (en fonction de leurs longueurs). Deux calculs sont proposés :
- le rapport entre la médiane (26,28) et la médiale (49,93) soit 0,90. Autrement dit la médiale est de 90% supérieure à la médiane (pour des comparaisons avec d'autres écrivains, voir l'annexe 2). Cet écart considérable suffit à attester la prédominance des phrases longues dans la *Recherche*.
- le second calcul est utilisé en science économique pour étudier la distribution des revenus ou des patrimoines. Il s'agit de l'indice de Gini qui mesure l'écart entre la situation réelle et celle qui serait observée en cas d'égale répartition du caractère (ici la surface du texte) entre les individus (les phrases) composant le livre. En cas d'équidistribution, toutes les phrases de la *Recherche* auraient la longueur moyenne (≈ 36 mots). Pour chaque centile, on calcule la proportion de la surface de texte couverte et l'écart par rapport à ce que serait cette surface dans l'hypothèse d'équidistribution. L'indice de Gini est la somme de ces écarts. Ici, il est égal à 55,4%. Autrement dit, dans la *Recherche*, les longueurs de phrases s'écartent de plus de 55% de ce qui serait constaté dans une population homogène.

Le "diagramme de Gini" permet de visualiser cette situation. Les phrases étant rangées par longueurs croissantes, on compte le nombre qu'il faut lire pour atteindre 1% de la surface (premier centile), puis 2%, etc. jusqu'à 100%. Les valeurs observées pour chaque centile sont reportées sur la figure 2 où la diagonale représente l'hypothèse d'équidistribution. L'indice de Gini est la surface comprise entre la diagonale et la courbe. Deux auteurs contemporains, et importants pour M. Proust, sont ajoutés sur le diagramme afin d'en illustrer les propriétés.

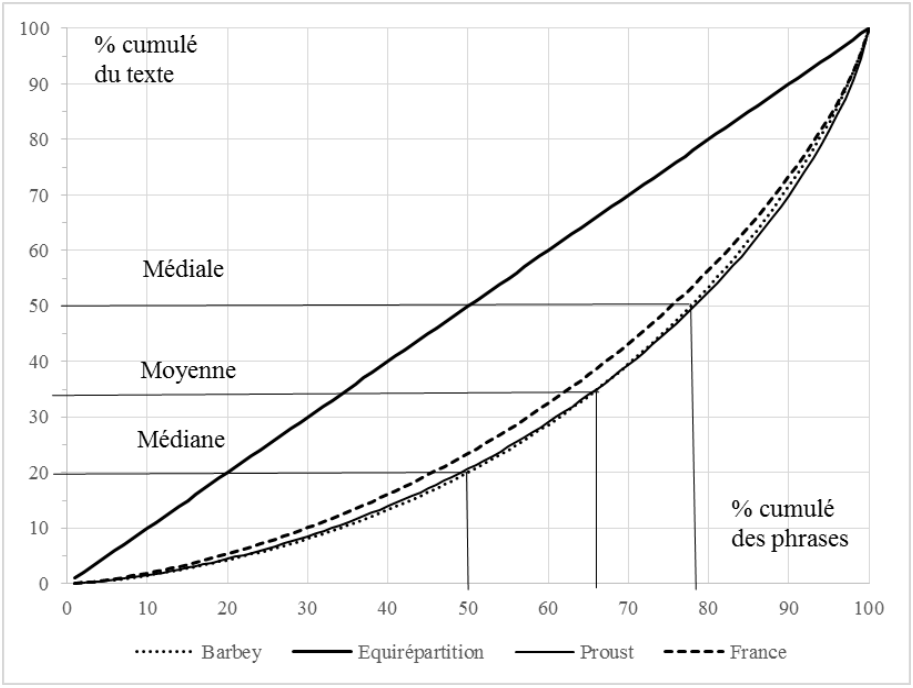


Figure 2 Diagramme de concentration (Gini) de la surface de la *Recherche* sur les phrases longues, comparée à celle de J. Barbey d'Aurevilly et de A. France.

Ce diagramme permet de comprendre pourquoi la médiane ou la moyenne rendent mal compte des distributions fortement asymétriques comme les longueurs de phrase. Par exemple, les deux tiers des phrases ont des longueurs inférieures à la moyenne et pourtant ces phrases ne couvrent qu'à peine plus d'un tiers du texte (34,5%).

La figure 2 montre également que, si les phrases de la *Recherche* sont singulières par rapport à certains écrivains du XIXe - à commencer par A. France qui aurait fourni le modèle de Bergotte (Levaillant 1952) –, elles semblent très proches de quelques livres comme *Une vieille maîtresse* (1851) de Barbey d'Aurévilly, écrivain que Proust cite à plusieurs reprises (Rogers 2000). C'est la dernière question abordée dans cette communication.

5. Singularité de Proust ?

Pour juger de cette singularité : à qui le comparer ? Et comment décider si les écarts constatés sont statistiquement significatifs ?

Premièrement, il faut comparer Proust à lui-même. Un de ses ouvrages se trouve dans le domaine public : *Les Plaisirs et les jours* (1896) dont les valeurs centrales sont indiquées en première ligne dans le tableau 2.

	Etendue	Mode	Médiane	Moyenne	Médiale	Me/Ml	Gini
Plaisirs et jours	1-250	7	21,30	27,87	37,16	0,754	0,542
Recherche	1-931	11	26,28	35,57	49,93	0,900	0,554

Tableau 2. Caractéristiques des phrases des *Plaisirs et les jours* comparés à la *Recherche*

Toutes ces valeurs sont significativement inférieures à celles observées dans la *Recherche*. Cependant, l'indice de Gini indique que le jeune Proust avait déjà tendance à concentrer une proportion importante du texte dans les phrases longues.

Deuxièmement, il faut comparer Proust aux auteurs qu'il cite explicitement ou par allusion, non seulement dans la *Recherche* (Nathan 1968) mais aussi dans ses autres œuvres et dans sa correspondance (Chantal 1967). Dans la *Recherche*, Racine et Mme de Sévigné sont les plus cités, puis en seconde position : Balzac et Saint-Simon ; en troisième : Chateaubriand, Hugo, Molière, Musset, Sand et Vigny. La singularité des phrases théâtrales (Labbé & Labbé 2010) ne permet pas de comparer la *Recherche* (qui est un roman) avec les pièces produites par Molière, Hugo, Musset, Racine ou Vigny.

Enfin, il faut le comparer aux autres romanciers contemporains : ont été ajoutés les principaux écrivains du XIXe et du début du XXe - comme Bourget, Giraudoux, Flaubert, Maupassant, Zola – et quelques auteurs moins connus mais singulièrement proches de Proust.

L'annexe 2 présente un échantillon des résultats. Chaque écrivain est singulier et parfois les indices peuvent varier selon ses oeuvres. La *Recherche* se situe dans la partie haute pour tous les indices et notamment pour la propension à concentrer une proportion importante du texte dans les phrases les plus longues (Gini). Cependant, on observe des caractéristiques supérieures à celle de Proust dans quelques œuvres - Huysmans (*A rebours*), les frères Goncourt (*Mme Gervaisais*) - ou proches dans Barbey d'Aurevilly, mais aussi dans les *Lettres* de Mme de Sévigné ou les *Mémoires* de Saint-Simon.

6. Conclusions

Lorsque, dans une population – ici les phrases d'un texte -, un caractère (la surface de ce texte) est très inégalement réparti, la moyenne et l'écart-type sont de peu d'utilité. Cette remarque s'applique à la plupart des "données textuelles", spécialement la fréquence des mots dans un texte dont la distribution présente un profil assez semblable à celui de la longueur des phrases que nous venons d'étudier. Naturellement, on ne remédie pas à cela en rebaptisant "barycentre", la moyenne et "distance", la variance.

Pour ce type de distributions, l'indice statistique le plus éclairant est la seconde médiane ou médiale. Pour mesurer le degré de dispersion de la série autour de cette valeur centrale, de nombreux indices sont concevables, notamment les rapports entre quantiles extrêmes. Cependant, le rapport entre médiane et médiale, ou l'indice de Gini paraissent les plus aptes à donner une indication de la concentration du caractère sur une proportion plus ou moins restreinte de la population totale.

Ces indices montrent que, durant la majorité du temps, le lecteur de la *Recherche* se trouve

confronté à des phrases très longues (50 mots et plus) et syntaxiquement complexes. Ils confirment que M. Proust a une propension à concentrer une proportion importante du récit dans les phrases les plus longues. Enfin, dans la littérature française moderne, les phrases de la *Recherche* sont exceptionnelles. A part trois autres romans du XIXe siècle – de Barbey d'Aurevilly, Goncourt et Huysmans – nous n'avons, pour l'instant, trouvé aucun ouvrage de fiction présentant des caractéristiques statistiques comparables à celles observées dans la *Recherche*.

Ces conclusions ont été acquises grâce à un dépouillement rigoureux, à des indices statistiques adaptés et à une vaste base de textes traités selon des procédures exactement semblables. A ce prix, la statistique lexicale peut être une auxiliaire utile de l'analyse littéraire.

Enfin, dans une œuvre littéraire, il n'existe pas un type de phrase unique mais plusieurs qui ont chacun leurs particularités lexicales et stylistiques (Monière et al. 2008 ; Labbé & Labbé 2010). Une prochaine publication présentera ces types de phrases avec leurs singularités lexicales, stylistiques et thématiques. Elle répondra aussi à une question pendante : comment déterminer que les écarts entre œuvres et auteurs sont ou non significatifs ?

Remerciements

Cette communication a bénéficié d'une relecture attentive de : Edouard Arnold (Trinity College, Dublin), Denis Monière (Université de Montréal), Jacques Savoy (Université de Neuchâtel) ainsi que des deux relecteurs anonymes du comité de programme des JADT que les auteurs remercient pour leurs remarques très utiles.

References

Brunet E. (1981). La phrase de Proust. Longueur et rythme. *Travaux du cercle linguistique de Nice*, p. 97-117.

Bureau C. (1976). Marcel Proust ou le temps retrouvé par la phrase. *Linguistique fonctionnelle et stylistique objective*. Paris : PUF, p. 178-231.

Curtius E.-R. (1971). Etude de lilas. Le rythme des phrases. In Tadié J.-Y. (dir.). *Lectures de Proust.* Paris : A. Colin.

Milly J. (1975). *La phrase de Proust. Des phrases de Bergotte aux phrases de Vinteuil*. Paris : Larousse.

Ferré A. (1957). La ponctuation de M. Proust. *Bulletin de la Société des Amis de Marcel Proust*, 7, p 171-192.

Hatzfeld A., Darmeisteter A., Thomas A. (1898). *Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours*. Paris : Delagrave.

Labbé C., Labbé D. (2010). Ce que disent leurs phrases. In Bolasco S., Chiari I., Giuliano L. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto. Vol 1, p. 297-307.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Levaillant J. (1952). Note sur le personnage de Bergotte. *Revue des sciences humaines*. Janvier-Mars 1952, p 33-48.

Milly J. (1986*). La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine.

Monière D., Labbé C. & Labbé D. (2008). Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, p. 43-69.

Muller C. (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Nathan J. (1969). *Citations, références et allusions de Marcel Proust dans* A la recherche du temps perdu. Paris : Nizet (Première édition : 1953).

Rogers B. (2000). *Proust et Barbey d'Aurevilly. Le dessous des cartes*. Paris : Champion.

Serça I. (2010). *Les coutures apparentes de la Recherche. Proust et la ponctuation*. Paris : Champion.

Annexe 1 Corpus *A la Recherche du temps perdu* (Marcel Proust. Paris Gallimard 1919-1927)

Livre	Longueur	Vocabulaire	N phrases
Combray	79 906	6 502	1 727
Un amour de Swann	84 142	5 859	2 226
Noms de pays : le nom	19 434	2 823	374
Du côté de chez Swann (1919)	**183 482**	**9 347**	**4 327**
Autour de Mme Swann	91 451	6 532	2 511
Noms de pays : le pays	134 192	8 283	3 334
A l'ombre des jeunes filles en fleur (1919)	**225 643**	**10 396**	**5 845**
Le côté de Guermantes 1	75 494	6 281	1 903
Le côté de Guermantes 2, chapitre 1	84 354	6 368	2 781
Le côté de Guermantes 2, chapitre 2	89 727	6 707	2 700
Le côté de Guermantes (1920-21)	**249 575**	**6 707**	**7 384**
Sodome et Gomorrhe	13 512	2 476	271
Sodome et Gomorrhe 2, chapitre 1	30 699	3 779	2 082
Sodome et Gomorrhe 2, chapitre 2	117 774	7 822	3 056
Sodome et Gomorrhe 2, chapitre 3	57 603	5 311	1 811
Sodome et Gomorrhe 2, chapitre 4	8 137	1 373	250
Sodome et Gomorrhe (1921-22)	**227 725**	**10 972**	**7 470**
La prisonnière (1923)	173 409	9 062	5 124
La fugitive (1925)	115 866	6 456	3 255
Le temps retrouvé (1927)	152 159	8 708	3 931
Dernier volume (posthume)	**441 434**	**13 518**	**12 310**
Total général (*A la recherche du temps perdu)*	**1 327 859**	**21 837**	**37 336**

Annexe 2 Longueur des phrases chez quelques écrivains antérieurs ou contemporains de Proust

	Etendue	Mode	Médiane	Moyenne	Médiale	Me/Ml	Gini
Recherche	**931**	**11**	**26,28**	**35,57**	**49,93**	**0,900**	**0,554**
Balzac*	391	10	17,27	21,88	29,00	0,680	0,511
Barbey d'A. (*Chevalier*)	192	7	21,92	29,4	43,00	**0,964**	**0,557**
Barrès*	195	8	17,86	21,94	28,59	0,601	0,497
Bourget*	201	7	16,62	21,34	29,58	0,780	0,539
Chateaubriand (*Mémoires*)	195	22	24,46	28,5	34,28	0,401	0,437
Daudet*	203	5	13,14	17,84	25,26	0,923	0,549
Dumas*	243	7	14,90	20,28	29,00	0,947	0,567
Flaubert*	231	7	13,75	18,37	25,24	0,837	0,528
France	394	8	15,79	19,98	26,06	0,651	0,504
Gautier*	282	18	27,11	33,07	41,90	0,546	0,493
Giraudoux*	466	4	18,60	25,77	37,76	1,031	0,580
Goncourt (*Gervaisais***)**	**670**	**8**	**24,17**	**34,05**	**51,47**	**1,130**	**0,597**
Goncourt (*Journal*)	373	3	19,80	25,37	37,62	**0,900**	**0,580**
Hugo*	828	6	11,39	16,89	23,68	1,079	0,561
Huysmans (*A rebours*)	254	**28**	**44,24**	**51,49**	**65,82**	0,488	**0,557**
Maupassant*	168	6	14,44	18,98	26,39	0,828	0,542
Musset*	197	16	19,56	23,82	29,57	0,512	0,485
Nerval*	136	12	19,93	24,21	31,27	0,569	0,499
Saint-Simon	361	18	27,89	34,15	44,14	0,523	0,506
Sand (*Champi*)	117	21	22,11	26,19	32,56	0,473	0,477
Sévigné (*Lettres*)	307	11	25,72	31,99	40,96	0,593	0,490
Stendhal*	235	18	20,18	23,92	29,79	0,477	0,463
Vigny*	315	17	20,82	27,47	37,41	0,797	0,538
Zola*	153	8	15,80	19,91	25,66	0,624	0,491

* Uniquement les romans