Une prédominance des paragraphes très courts et une corrélation entre la longueur des paragraphes avec le genre littéraire ou la période d'écriture : les résultats de "Un Atlas des spectres de textes littéraires"

Claude Patricia Tardif

Narra

Une prédominance des paragraphes très courts et une corrélation entre la longueur des paragraphes avec le genre littéraire ou la période d'écriture : les résultats de *Un Atlas des spectres de textes littéraires*

Lien vers l'*Atlas* :
https://hal.archives-ouvertes.fr/hal-02986259/

Claude Patricia Tardif

Résumé

La longueur des paragraphes dans un texte littéraire ne dépend-elle que de l'auteur·e ? Afin d'y répondre, *Un Atlas des spectres de textes littéraires* comprenant plus de 1 255 textes de 162 auteur·e·s a été réalisé avec des algorithmes en fonction de la fréquence de leurs paragraphes alinéaires selon leur longueur. Il ressort de cette classification visuelle une prédominance des paragraphes très courts d'une ligne et une corrélation entre la longueur des paragraphes et le genre littéraire ou la période d'écriture. En revanche, le lien entre la longueur des paragraphes et la popularité n'est pas attesté. Plus avant, l'*Atlas* permet de connaître en un instant le degré de développement des unités thématiques et dialogales des textes ainsi étudiés.

Mots-clés : longueur des paragraphes, fréquence, corrélation, genre littéraire, période d'écriture, popularité, développement des unités textuelles, séquences.

Abstract

Does the length of paragraphs in a literary text only depend on the author ? In order to answer this, *An Atlas of literary texts spectra* including more than 1,255 texts by 162 authors was carried out with algorithms based on the frequency of their paragraphs according to their length. This visual classification shows a predominance of very short one-line paragraphs and a correlation between the length of the paragraphs and the literary genre or period of writing with disparities. However, the link between paragraph length and popularity is not proven. Further on, the *Atlas* allows us to know instantly the degree of development of the thematic and dialogal units of studied texts.

Keywords : paragraph length, frequency, correlation, literary genre, period of writing, popularity, development of textual units, sequences.

1. Introduction

La longueur des paragraphes dans un texte littéraire ne dépend-elle que de l'écrivain·e ? Afin d'y répondre, *Un Atlas des spectres de textes littéraires* a été réalisé en fonction de la longueur des paragraphes. Il comprend 1 255 textes de périodes et de genres littéraires différents de 162 auteur·e·s, pour l'essentiel occidentaux/ales. L'objectif de la présente étude consiste à analyser cette classification et à déterminer s'il existe une corrélation entre la longueur des paragraphes alinéaires et le genre littéraire, la période d'écriture ou la popularité.

Si l'on trouve réunis dans l'*Atlas* les textes d'un même genre littéraire ou d'une même période d'écriture, l'on pourra conclure que la longueur des paragraphes est liée au genre littéraire ou à la période d'écriture, quel/quelle que soit l'auteur·e. De la même manière, si les œuvres les plus populaires sont regroupées dans l'*Atlas*, l'on pourra conclure que la popularité des œuvres est liée à la longueur de leurs paragraphes.

Plus avant, vu que le paragraphe narratif est une unité thématique[1] (Algee-Hewitt et al., 2015) (Piérard & Bestgen, 2006), dans le cas d'une corrélation confirmée entre la longueur des paragraphes et le genre littéraire ou la période, l'on pourra conclure que le développement des unités thématiques et dialogales est lié au genre littéraire et à la période d'écriture, quel/quelle que soit l'auteur·e.

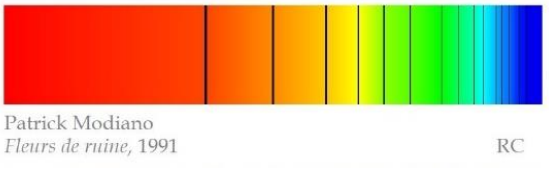
Dans le cas contraire d'une répartition des textes relevant d'un même genre littéraire ou d'une même période dans tout l'*Atlas*, l'on pourra en déduire qu'il n'existe pas de lien entre la longueur des paragraphes et le genre littéraire ou la période d'écriture.

2. Le corpus des textes littéraires étudiés

Le corpus de l'*Atlas* se compose surtout de romans dits de littérature générale : 21 % de romans classiques et 15 % de romans contemporains, puis 13 % de romans policiers (voir *Tableau 2*). L'*Atlas* se compose largement (72 %) d'œuvres publiées entre 1850 et 2000. À l'inverse, 4 % seulement sont publiées avant 1800 et 14 % après 2000.

Chaque texte dans l'*Atlas* se présente sous la forme d'un spectre de couleurs (*Figure 1*) de 7,5 cm de long, quels que soient le nombre de paragraphes et la taille du texte. En-dessous du spectre sont indiqués le titre de l'œuvre, la date de première parution et le genre littéraire (ici RC). Les œuvres les plus connues d'un·e auteur·e sont signalées par un titre en caractères gras et un point placé à gauche du spectre.

Figure 1 : Spectre *Fleurs de ruine*



Les légendes des genres littéraires sont les suivantes.

C	pour un conte,
N	pour une nouvelle
R	pour un roman

Puis,

RA / NA	R et N d'amour
RAv / NAv	R et N d'aventure
RB / NB	R et N auto/biographiques
RC / NC	R et N contemporains
RClass / NClass	R et N classiques
REr / NEr	R et N érotiques
RF / NF	R et N fantastique
Rfa / NFa	R et N de fantasy
RH / NH	R et N historiques, dont cape et d'épée
Rho / Nho	R et N d'horreur et de terreur
RJ / NJ	R et N jeunesse
RP / NP	R et N policiers et thrillers
RSF / NSF	R et N de science-fiction
R / N	R et N « autre »

Les textes dont le ou les genres sont difficiles à déterminer sont classés dans la catégorie « Autre ».

3. Les outils, la méthode

Les atlas sont devenus au cours du XIXe des outils scientifiques. Ils sont une source féconde de connaissances des « objets » étudiés, tels que la composition des étoiles ou la matière. Celui-ci, composé des spectres de textes littéraires, est inspiré des atlas de spectres stellaires[2] (Daston & Galison, 2012), majeurs dans l'histoire de l'astronomie. Ici, les spectres permettent d'étudier les textes sous le prisme de la longueur de leurs paragraphes alinéaires, de les

comparer, de visualiser la fréquence de leurs paragraphes selon leur longueur et de voir ainsi la proportion des paragraphes de 1 ligne, 2 lignes… dans chacun d'entre eux en un instant. Nous allons présenter la façon dont ils ont été réalisés.

3.1. Définition des paragraphes alinéaires et mesure de leur longueur

Les paragraphes alinéaires, ci-après désignés « paragraphes », s'entendent des paragraphes *stricto sensu* ainsi que du texte situé après un tiret de dialogue avec alinéa servant à différencier les interlocuteurs. Ils incluent donc aussi bien les paragraphes narratifs que dialogaux.

Les paragraphes sont mesurés en nombre de caractères avec le logiciel *Narra 2.0*. Pour une saisie immédiate, leur longueur est exprimée en nombre de lignes sur la base selon laquelle 1 ligne équivaut à 60 caractères[3]. Puis, à l'aide d'un second programme, l'effectif des paragraphes en fonction de leur longueur est rapporté à l'effectif total des paragraphes dans le texte. Enfin, à partir de ces fréquences relatives, le spectre est généré en utilisant le code couleur ci-après présenté.

3.2. Le code couleur utilisé

À une longueur de paragraphe est attribuée une couleur, du rouge pour un paragraphe court de 1 ligne, en passant progressivement par la couleur centrale jaune-vert correspondant à la longueur moyenne des paragraphes, soit 340 caractères ou 6 lignes (arrondies à l'unité près) sur un échantillon de 250 textes, jusqu'à un bleu de plus en plus foncé pour les paragraphes longs de plus de 10 lignes. Le code couleur est ainsi le suivant :

¶ de 1 ligne –> rouge…
¶ de 3 lignes –> orange …
¶ de 5 lignes –> jaune ...
¶ de 7 lignes –> vert ...
¶ de 9 lignes –> turquoise ...
¶ de 12 lignes –> bleu.

Prenons l'exemple de la *Figure 1. Fleurs de ruine* comprend plus d'un tiers de paragraphes de 1 ligne, plus d'1/10e de paragraphes de 2 lignes, 1/10e de paragraphes de 3 lignes, etc. Les seuls paragraphes de 1 et 2 lignes composent près de la moitié de l'œuvre de Patrick Modiano.

Notons que les raies noires diversement réparties dans le spectre servent de repère visuel ; elles délimitent la part des paragraphes de 1 ligne, de ceux de 2 lignes, 3 lignes, etc.

*3.3. Construction de l'*Atlas

En raison de la prédominance des paragraphes de 1 ligne, le critère retenu pour réaliser l'*Atlas* a été leur proportion dans les textes. Sept classes ont ainsi été définies.

Classe I : textes en comprenant plus de 55 %
Classe II : textes en comprenant entre 55 et 45 %
  Deux sous-classes : + ou – 5 % de ¶ de + 10 lignes
Classe III : textes en comprenant entre 45 et 33 %
  Deux sous-classes : + ou – 5 % de ¶ de + 10 lignes
Classe IV : textes en comprenant entre 33 et 21 %
  Deux sous-classes : + ou – 5 % de ¶ de + 10 lignes
Classe V : textes en comprenant moins de 21 %
Classe VI : répartition équilibrée
Classe VII : quasi-monopole des paragraphes de 1 ligne et de ceux de plus de 10 lignes, donc très peu de paragraphes de taille intermédiaire.

Deux sous-classes subdivisent les classes II, III et IV, lesquelles comptent de nombreux textes : une notée "B" pour bleu, comprenant les textes avec une part non négligeable (plus de 5 %) de paragraphes longs de plus de 10 lignes, et une sans le "B" du bleu, comprenant les textes avec peu (moins de 5 %) de paragraphes longs de plus de 10 lignes.

Une fois les spectres des textes placés dans une classe ou une sous-classe en fonction de la proportion de leurs paragraphes de 1 ligne, il a fallu les ordonner. À cette fin, un second critère distinct a été utilisé : pour les classes I, II et III, par ordre décroissant de la part des paragraphes de 1 ligne ; pour classes IV et V, par ordre croissant des paragraphes de plus de 10 lignes.

Quelques explications méritent d'être apportées s'agissant du premier critère de subdivision des classes en sous-classes (i) et du second critère d'ordonnancement des spectres (ii).

(i) Le critère qui subdivise les classes II, III et IV – la part de 5 % de paragraphes de plus de 10 lignes – peut sembler faible. Or, compte tenu de l'importance des paragraphes de 1 ligne dans le corpus, cette proportion de 5 % est significative.

(ii) Le critère qui permet d'ordonner les textes est différent selon que l'on se trouve dans les classes de plus de 33 % de paragraphes de 1 ligne ou dans celles de moins de 33 % de paragraphes de 1 ligne. Le fait que la proportion de 33 % de paragraphes de 1 ligne soit le critère qui inverse la façon de disposer les spectres, et non la proportion de 50 % comme l'on pourrait s'y attendre, a été dicté par une cohérence visuelle. En effet, pour les classes comprenant moins de 33 % de paragraphes de 1 ligne, continuer à ordonner les spectres avec la part décroissante des paragraphes de 1 ligne aboutissait à une cacophonie visuelle, alors qu'en retenant la part croissante des paragraphes de plus de 10 lignes, on obtient une progression visuelle cohérente. Il était plus pertinent et aisé de choisir comme critère d'ordre la part des paragraphes dont l'amplitude de variation est la plus large. Pour les classes comprenant moins de 33 % de paragraphes de 1 ligne, la plus grande amplitude est celle des paragraphes de plus de 10 lignes : 25 points d'écart (5 % à 30 % des paragraphes) contre 12 points d'écart pour les paragraphes de 1 ligne (33 % à 21 % des paragraphes). Un autre facteur a pu également intervenir. Hormis quelques exceptions, la part maximale des paragraphes de 1 ligne dans un texte est de 65 % et la part minimale de 10 %. Cet atlas a été ainsi construit, non sur une échelle de 1 à 100 % de paragraphes de 1 ligne, mais sur une échelle de 10 % à 65 %.

4. Les résultats

Le corpus de l'*Atlas* se caractérise d'abord par la prépondérance des paragraphes de 1 ligne. Quelques éléments (voir *Tableau 2*) donnent la mesure de cette primauté : les textes comprenant plus d'un tiers de paragraphes de 1 ligne représentent 63 % du corpus, tandis que ceux comprenant plus d'un tiers de paragraphes de 10 lignes en représentent 5 %.

Ajoutons à cela que les textes comprenant plus de la moitié de paragraphes de 1 ligne représentent 22 % du corpus, tandis que ceux comprenant plus de la moitié de paragraphes longs de plus de 10 lignes en représentent 2 %, dix fois moins.

Par ailleurs, nous remarquons que plus il y a de paragraphes très courts dans un texte littéraire, moins il y a de paragraphes très longs. 83 % des textes comprenant plus de 45 % de paragraphes courts de 1 ligne comprennent moins de 5 % de paragraphes de plus de 10 lignes. Ce sont les autres paragraphes qui compensent.

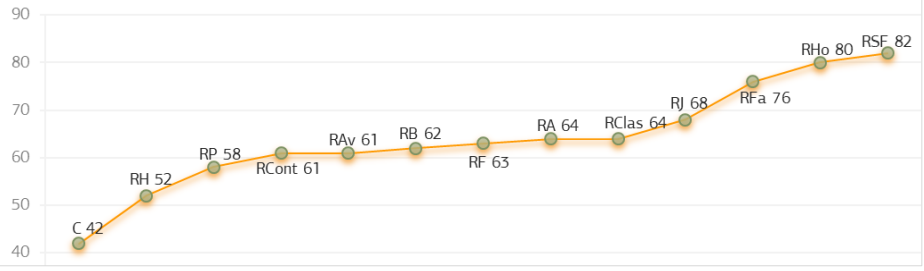
Nous constatons également que les deux tiers des œuvres se concentrent dans les classes II, III et IV (entre 55 % et 21 % de paragraphes de 1 ligne) et un tiers dans la classe III (entre 45 % à 33 % de paragraphes de 1 ligne), ce qui en fait la classe la plus importante.

4.1. Une corrélation entre la longueur des paragraphes et le genre littéraire, mais aussi d'importantes disparités

Notons au préalable que si la longueur des paragraphes n'avait aucun lien avec le genre littéraire ou la période d'écriture, les œuvres auraient été équiréparties dans tout l'*Atlas*. Or, le taux de concentration des textes d'un genre littéraire dans deux classes (*Figure 2*) est au minimum de 42 % et aller jusqu'à 82 %, un taux qui en toute logique baisse, à 67 %, si l'on prend une classe et une sous-classe, à savoir si l'on retient aussi le critère de la part des paragraphes de plus de 10 lignes. Les romans de science-fiction, d'horreur et de fantasy sont les plus concentrés, à environ 80 % dans deux classes. À l'inverse, les romans historiques se concentrent à 52 % et les contes à 42 %.

La *Figure 3* ci-après confirme ces écarts en se divisant en deux : d'un côté, les œuvres comprenant une large

Figure 2 : Taux de regroupement dans deux classes par genre littéraire en %



part de paragraphes courts, parmi lesquelles sont les romans policiers, d'amour, de fantasy, d'horreur et de science-fiction ; de l'autre côté, les œuvres comprenant une part significative de paragraphes longs, parmi lesquelles sont les romans classiques, biographiques et fantastiques.

Les histogrammes montrent aussi que les romans historiques, les contes et, dans une moindre mesure, les romans contemporains et d'aventure sont répartis dans tout le « spectre » de l'*Atlas* (sans pics distinctifs) tant dans les classes avec paragraphes longs que « sans[4] ».

L'analyse des diagrammes des *Figures 3 et 4* permet d'affiner les résultats par genre littéraire.

Figure 3 : Histogrammes de répartition des genres littéraires par classe

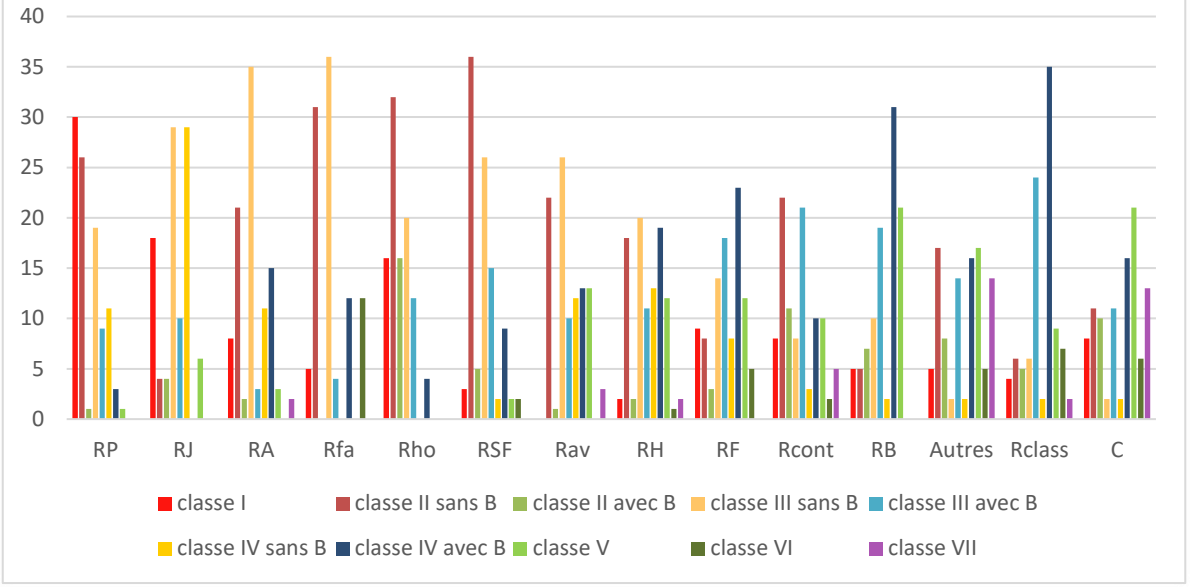
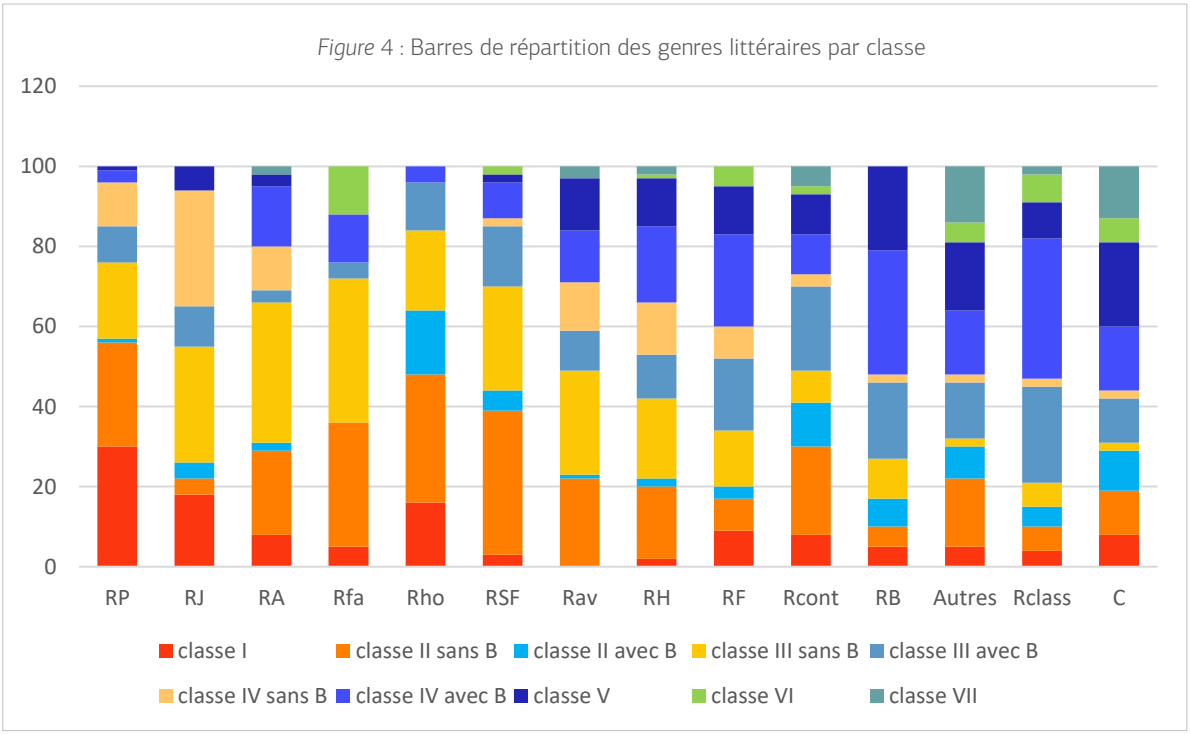


Figure 4 : Barres de répartition des genres littéraires par classe



Les romans de science-fiction, d'horreur, de fantasy, jeunesse et d'amour

Il existe une corrélation étroite entre la longueur des paragraphes et les romans de fantasy, de science-fiction, d'amour et jeunesse. Ils se concentrent dans les classes comprenant une large part de paragraphes très courts de 1 ligne et dans les sous-classes « sans » paragraphes longs : 85 % des romans de science-fiction se situent dans les classes de plus d'un tiers de paragraphes de 1 ligne « sans » paragraphes longs, 76 % des romans de fantasy, 68 % des romans jeunesse et 64 % des romans d'amour.
Les romans jeunesse se distinguent de tous les autres genres littéraires, non par l'absence de paragraphes longs, mais par une grande proportion de paragraphes de taille intermédiaire.

Les romans classiques, biographiques et fantastiques

Trois quarts des romans classiques, biographiques et fantastiques se concentrent dans les sous-classes III, IV et V avec paragraphes longs.
La corrélation entre la longueur des paragraphes et les romans classiques est étroite. Ils sont en effet quasi-absents des classes comptant plus de 45 % de paragraphes courts ainsi que des sous-classes « sans » paragraphes longs et 60 % d'entre eux se situent dans les sous-classes III et IV avec paragraphes longs, où sont réunis Zola, Dickens, Dostoïevski, Balzac, Melville, Sand, Stendhal… S'agissant des romans biographiques et fantastiques, la corrélation est moyenne. Certes, ils se concentrent dans les deux mêmes sous-classes III et IV avec paragraphes longs que les romans classiques, mais respectivement à 52 % et à 41 %.

Les romans contemporains, historiques et les contes

La relation entre la longueur des paragraphes et ces œuvres n'est pas établie ; ils occupent toute l'étendue de la classification.
Les romans contemporains notamment se situent dans les sous-classes à la fois avec paragraphes longs et « sans » paragraphes longs, laissant présumer une diversité d'écriture. Ils sont néanmoins plus présents dans la sous-classe II « sans » paragraphes longs et dans la sous-classe III avec paragraphes longs.

Les romans policiers

Bien que les romans policiers soient répartis dans différentes classes, ils se distinguent des autres par une surreprésentation dans la classe I qui comprend plus de 55 % de paragraphes de 1 ligne dont ils forment à eux seuls 43 %, alors qu'ils représentent 13 % du corpus. Par ailleurs, ils se caractérisent, comme les romans de science-fiction, d'horreur et de fantasy, par une quasi-absence de paragraphes longs. Il en ressort une corrélation étroite entre la longueur des paragraphes et les romans policiers.

4.2. Une corrélation entre la longueur des paragraphes et la période d'écriture

Une première analyse de l'*Atlas* avait été réalisée en distinguant les textes publiés avant 1900 de ceux publiés après 1900 sans qu'une corrélation entre la période d'écriture et la longueur des paragraphes n'ait été constatée. Une seconde analyse, plus fine, s'est avérée nécessaire ; elle se fonde sur une échelle plus resserrée de six périodes : < 1800 ; 1800-1850 ; 1850-1900 ; 1900-1950 ; 1950-2000 ; > 2000 et atteste du lien entre la longueur des paragraphes et la période d'écriture.
Sur l'ensemble de la période étudiée, on observe (voir la « droite de régression » du *Tableau 1*) une réduction progressive de la taille moyenne des paragraphes en raison d'une hausse de la part des paragraphes courts de 1 ligne et d'une réduction concomitante de la part de paragraphes longs de plus de 10 lignes.
Les romans publiés entre 1800 et 1850 se concentrent dans la classe comprenant entre 21 et 33 % de paragraphes de 1 ligne, ceux publiés entre 1850 et 1900 dans les classes entre 21 et 45 % de paragraphes de 1 ligne, ceux publiés entre 1900 et 1950, période la moins concentrée, dans les classes entre 33 et 55 % de paragraphes de 1 ligne, enfin ceux publiés entre 1950 et 2000 dans les mêmes classes que les précédents mais en étant plus concentrés sur celle comprenant entre 45 et 55 % de paragraphes de 1 ligne.
Les œuvres publiées avant 1800 et celles publiées après 2000 se situent aux extrêmes du *Tableau 1 :* 58 % des premières se situent dans la classe V avec

paragraphes longs de plus de 10 lignes et VII avec très peu de paragraphes de taille intermédiaire et 70 % des secondes – après 2000 – se situent dans les classes II et III « sans » paragraphes de plus de 10 lignes.

4.3. Une corrélation faible entre la longueur des paragraphes et la popularité

La moitié des textes les plus connus se situe dans les classes II et III (entre 33 et 55 % de paragraphes de 1 ligne), aussi bien avec paragraphes longs que « sans ». La présence marquée des paragraphes longs n'est donc pas un frein à la popularité. On constate néanmoins un décalage : la proportion des paragraphes de 1 ligne est en moyenne de 44 % pour l'ensemble des œuvres populaires contre 33 % pour l'ensemble du corpus. De plus, les classes dont la proportion des paragraphes de 1 ligne est supérieure à 45 % regroupent un grand nombre de textes populaires.

5. Conclusion et perspectives

Cette étude permet de conclure que la longueur des paragraphes dans un texte littéraire ne dépend pas que de l'auteur·e. Outre la prépondérance des paragraphes très courts de 1 ligne dans l'ensemble du corpus de l'*Atlas*, nous avons observé que la longueur des paragraphes dépend aussi, à des degrés divers, des genres littéraires et, dans une moindre mesure, de la période d'écriture, indépendamment de l'auteur·e, lequel interviendrait alors dans la combinaison de ces longueurs. En revanche, aucun lien entre la popularité des œuvres et la longueur des paragraphes n'a été constaté.

5.1. Une prédominance des paragraphes courts croissante : une potentielle loi de puissance

La fréquence d'usage des paragraphes très courts de 1 ligne est élevée : nous avons vu que 63 % des textes contiennent plus d'un tiers de paragraphes de 1 ligne et 22 % en contiennent plus de la moitié. Il serait intéressant de connaître la part des paragraphes de 1 et 2 lignes.

La prédominance des paragraphes très courts nous amène à penser à l'application d'une loi de puissance, celle de Benford[5] et/ou celle de Zipf[6] (Zipf, 1935/1974, p. 9-50). Contrairement à une loi uniforme selon laquelle la probabilité d'apparition d'un élément est égale, elles énoncent une probabilité d'apparition différenciée, selon une proportionnalité dégressive particulière. Toutefois, la croissance progressive de la part des paragraphes de 1 ligne dans les textes jette un doute quant à l'application de ces lois de puissance, à moins que la part des autres paragraphes changent dans les proportions.

5.2. Une corrélation confirmée non généralisée

Il ressort des premiers résultats une grande variabilité de corrélation entre la longueur des paragraphes et les genres littéraires. D'un côté, une corrélation étroite de deux ordres : les paragraphes sont courts dans les romans policiers, de science-fiction, d'horreur et de fantasy ; longs dans les romans classiques et biographiques. Les romans jeunesse se distinguent par une forte présence des paragraphes intermédiaires, avec très peu de paragraphes longs, et les romans policiers par une large part des paragraphes de 1 ligne. De l'autre, une corrélation faible, voire nulle, entre la longueur des paragraphes et les romans contemporains, historiques et les contes.

Relevons que la corrélation est étroite pour les genres littéraires homogènes et faible pour les genres littéraires diversifiés. La question se pose de savoir si une absence de corrélation ne résulte tout simplement pas du caractère hétérogène du genre littéraire tel que les romans contemporains, historiques et les contes, lesquels regroupent des textes très divers, à l'exemple de ceux de cape et d'épée dans le genre historique.

Une corrélation étroite entre la longueur des paragraphes et le genre littéraire indique non seulement un degré de dépendance fort entre ces deux variables, mais elle permet aussi de présumer de la façon dont les auteur·e·s développent leurs unités thématiques et dialogales.

Quant à la corrélation confirmée entre la longueur des paragraphes et la période d'écriture, il est utile de mentionner une étude de Matthew L. Jockers effectuée sur un large corpus de littérature anglaise de 1780 à 1870 (Jockers, 2013) qui démontre qu'un genre littéraire vit en moyenne le temps d'une génération, donc, qu'il est lié à une période. Cette

analyse ne trouverait cependant pas à s'appliquer à la littérature française (Glorieux, 2015). Toutefois, elle souligne le lien entre le genre littéraire et la période que nous retrouvons, ne fût-ce qu'en raison de la taille réduite des paragraphes des romans contemporains par rapport aux romans classiques.
Par ailleurs, le différentiel de concentration – 58 % dans deux classes pour les œuvres publiées avant 1800, contre 70 % dans deux autres classes pour les œuvres publiées après 2000 – pose l'hypothèse d'une uniformisation de l'écriture, et la dispersion des œuvres publiées entre 1900 et 1950 celle d'une diversité d'écriture inégalée.
S'agissant de la popularité, à ce stade de la recherche, aucune corrélation avec la longueur des paragraphes n'a été constatée. Il ne s'agit cependant que d'un premier résultat.

5.3. Les limites : deux biais

Le corpus de l'*Atlas* présente deux biais. Le premier concerne le corpus d'œuvres par genre littéraire. Les romans d'horreur, de fantasy et biographiques sont insuffisants. Chacun représente au plus 3 % du corpus, avec un maximum de 42 ouvrages. Il conviendra d'en ajouter de sorte à obtenir un nombre équivalent d'œuvres par genre littéraire. En outre, la catégorie « Autres » dont les textes sont répartis dans tout l'*Atlas* ne présente pas d'intérêt pour l'étude et pourra être supprimée. Sa présence a néanmoins eu le mérite de confirmer le lien entre la longueur des paragraphes et le genre littéraire par contraposition.
Le second biais concerne le corpus d'œuvres par auteur·e qui varie de 1 à 24. Par exemple, dans le cas où il y a 20 textes dans l'*Atlas* pour un·e auteur·e spécialisé·e dans un genre littéraire dont le corpus est peu fourni, la part de cet/cette auteur·e est trop élevée et peut fausser les résultats.
Ces biais ne sont pas de nature à compromettre la présente étude car ils ne concernent que les romans biographiques, d'horreur et de fantasy.

5.4. Les perspectives

Cette première analyse nous incite à approfondir l'étude réalisée. Il serait intéressant de savoir si la corrélation observée entre la longueur des paragraphes alinéaires (narratifs et dialogaux) et certains genres littéraires existe aussi pour les seuls paragraphes narratifs, donc sans tenir compte des dialogues. Afin d'y répondre, deux atlas complémentaires devront être réalisés, l'un à partir des paragraphes narratifs, l'autre à partir des paragraphes dialogaux. Dans l'hypothèse d'une corrélation entre la longueur des paragraphes narratifs et le genre littéraire identique à celle des paragraphes alinéaires, les spectres exposeraient comment les auteur·e·s développent leurs unités thématiques et, ainsi, la nature de leur écrit. Toujours dans l'hypothèse d'une corrélation confirmée, mais différente de celle entre la longueur des paragraphes alinéaires et le genre littéraire, une analyse complémentaire devra être effectuée. En l'absence de corrélation entre la longueur des paragraphes narratifs et le genre littéraire, les spectres exposeraient la proportion des dialogues dans les textes, à supposer que les paragraphes de dialogues soient plus courts que les autres, ce qui au demeurant reste à vérifier car on peut envisager la possibilité d'une proportion constante des paragraphes de dialogues (à ne pas confondre avec les paragraphes de 1 ligne) dans les textes et d'un changement général de leur longueur.
Enfin, les spectres réalisés pourront être assemblés par ordre chronologique dans un corpus visuel pour quelques auteur.e.s afin d'étudier leur évolution d'écriture en termes de longueurs de paragraphes.

Pour clore cette étude, précisons que, par son caractère exploratoire, elle constitue une ébauche prometteuse d'une recherche qui, outre la constitution et l'analyse de corpus visuels, nécessite une étude mathématique de long terme, statistique plus précisément, laquelle sera menée dans les prochaines années sur un large corpus de textes, peut-être d'ailleurs pas uniquement littéraires.

Références bibliographiques

Algee-Hewitt, M., Heuser, R., & Moretti, F. (2015). On Paragraphs. Scale, Themes, and Narrative Form – Stanford Literary Lab. Stanford Literary Laboratory. https://litlab.stanford.edu/LiteraryLabPamphlet10.pdf

Daston, L., & Galison, P. L. (2012). *Objectivité* (S. Renaut & H. Quiniou, Trad.; Vol. 1-1). les Presses du réel.

Glorieux, F. (2015). Bibliothèques informatisées et lectures distantes. In A. Gefen (dir.), *Des chiffres et des lettres : les humanités numériques* (Vol1. &-1), *Critique*, *819-820*(8-9), 653-666.

Jockers, M. L. (2013). *Macroanalysis : Digital methods and literary history*. University of Illinois Press.

Piérard, S., & Bestgen, Y. (2006). A la pêche aux marqueurs linguistiques de la structure du discours. *Actes des huitièmes Journées internationales d'Analyse statistique des Données Textuelles - JADT*, *2*, 749-757. http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-067.pdf

Zipf, G. K. (1935/1974). *La psychobiologie du langage : Une introduction à la philologie dynamique* (S. Deflandre & M.-R. Delorme, Trad.). Retz-C.E.P.L.

Notes

1. Dans le pamphlet n°10 du 10 d'octobre 2015 « On Paragraphs. Scale, Themes, and Narrative Form », le laboratoire Litlab de l'université de Stanford précise que leur étude démontre que non seulement le paragraphe est une unité thématique mais qu'il est aussi la meilleure unité textuelle pour appréhender les thèmes du texte.
« […] les changements de paragraphe (ou alinéas). L'auteur d'un texte est en effet censé les introduire pour signaler une discontinuité thématique. […] [Le paragraphe] est un indice pour identifier dans un corpus de textes des marqueurs de structure textuelle » Piérard, S., & Bestgen Y.

2. Je renouvelle mes remerciements à la bibliothèque de l'Observatoire de Paris, site de Paris, pour son accueil et la mise à disposition d'anciens catalogues de spectres stellaires qui m'ont inspiré pour la réalisation de cet atlas.

3. Que l'on prenne 60, 40 ou 70 caractères ne change pas les résultats (les rapports sont identiques) dès lors que ce critère s'applique à tous les textes et que l'unité de mesure des paragraphes est le nombre de caractères ou, plus précisément, le nombre de signes, espaces compris.

4. « Sans » n'est pas à prendre au sens littéral de « sans paragraphes de plus de 10 lignes », mais au sens de l'*Atlas* lorsque le texte comprend très peu, c'est-à-dire moins de 5 % de paragraphes de plus de 10 lignes.

5. La loi de Benford énonce que le premier chiffre non nul le plus fréquent débutant les nombres d'une série statistique quelconque ne suit pas la loi uniforme où chaque chiffre (de 1 à 9) a une égale probabilité d'apparition de 1/9e, soit 11,1 %. En effet, la probabilité d'apparition du chiffre 1, la plus élevée est de 30 %, puis celle du chiffre 2 est de 17 %, etc., enfin, celle du chiffre 9, le moins fréquent, est de 4,6 %.

6. La loi de Zipf énonce que la fréquence d'emploi (f) du n-ième mot le plus employé est inversement proportionnelle à son rang (r) : f varie en fonction de 1/r. Les mots les plus courts sont les plus utilisés, inversement, les plus longs sont les moins utilisés. Dans le roman *Ulysses* de James Joyce, il observa que le mot le plus courant (*the*) apparaît 8 000 fois, le deuxième (*of*) deux fois moins que le 1er, le troisième trois fois moins, ..., le centième cent fois moins que le 1er, etc. Par exemple, le douzième mot apparaît 1/12e fois le premier. Cette loi empirique et contestée pendant des décennies a ensuite été théorisée par Benoît Mandelbrot qui démontra qu'elle était un cas particulier d'une loi de puissance plus générale appelée loi de Mandelbrot.

Tableau 1 : Répartition des œuvres par période en nombre et en pourcentage

	Classe I > 55 %	Classe II : 55-45 %			Classe III : 45-33 %			Classe IV : 33-21 %			Classe V < 21 %	Classe VI Équil.	Classe VII Ø méd	TOTAL
		Sans B	Avec B	Total	Sans B	Avec B	Total	Sans B	Avec B	Total				
TOTAL	**120**	205	60	**265**	193	187	**380**	85	209	**294**	**111**	**40**	**45**	1 255
en %	**9 %**	16 %	5 %	**21 %**	16 %	15 %	**30 %**	7 %	17 %	**24 %**	**9 %**	**3 %**	**4 %**	100 %
Connus	**25**	42	13	**55**	22	38	**60**	11	37	**48**	**24**	**8**	**9**	229
en %	**22 %**	20 %	22 %	**21 %**	11 %	20 %	**16 %**	13 %	18 %	**16 %**	**22 %**	**22 %**	**23 %**	18 %
Périodes plus fines que < 1900 ou > 1900 en nombre														
< 1 800	**4**	2	3	**5**	:	7	**7**	:	6	**6**	**18**	**1**	**14**	55
1 800- 1 850	**5**	15	3	**18**	12	28	**40**	1	64	**65**	**29**	**17**	**5**	179
1 850- 1900	**20**	14	18	**32**	26	66	**92**	19	71	**90**	**18**	**9**	**3**	264
1 900- 1 950	**35**	50	22	**72**	34	41	**75**	11	36	**47**	**34**	**6**	**18**	287
1 950- 2 000	**30**	79	8	**87**	71	26	**97**	39	25	**64**	**9**	**5**	**3**	295
> 2 000	**26**	4	6	**51**	50	19	**69**	14	8	**22**	**3**	**2**	**2**	175
TOTAL	**120**	205	60	**265**	193	187	**380**	84	210	**294**	**111**	**40**	**45**	1255
en % du total de chaque période														
< 1 800	**7 %**	4 %	5 %	**9 %**	:	13 %	**13 %**	:	11 %	**11 %**	**33 %**	**2 %**	**25 %**	100 %
1 800- 1 850	**3 %**	8 %	2 %	**10 %**	7 %	16 %	**22 %**	1 %	36 %	**36 %**	**16 %**	**9 %**	**3 %**	100 %
1 850- 1900	**7 %**	5 %	7 %	**12 %**	10 %	25 %	**35 %**	7 %	27 %	**34 %**	**7 %**	**3 %**	**1 %**	100 %
1 900- 1 950	**12 %**	17 %	8 %	**25 %**	12 %	14 %	**26 %**	4 %	13 %	**16 %**	**12 %**	**2 %**	**6 %**	100 %
1 950- 2 000	**10 %**	27 %	3 %	**29 %**	24 %	9 %	**33 %**	13 %	8 %	**22 %**	**3 %**	**2 %**	**1 %**	100 %
> 2 000	**15 %**	26 %	3 %	**29 %**	28 %	11 %	**39 %**	8 %	5 %	**13 %**	**2 %**	**1 %**	**1 %**	100 %
en % du total de chaque classe														
< 1 800	**3 %**	1 %	5 %	**2 %**	:	4 %	**2 %**	:	3 %	**2 %**	**16 %**	**3 %**	**31 %**	4 %
1 800- 1 850	**4 %**	7 %	5 %	**7 %**	6 %	15 %	**11 %**	1 %	30 %	**22 %**	**26 %**	**43 %**	**11 %**	14 %
1 850- 1900	**16 %**	7 %	30 %	**12 %**	13 %	35 %	**24 %**	23 %	34 %	**31 %**	**15 %**	**23 %**	**7 %**	21 %
1 900- 1 950	**30 %**	24 %	37 %	**27 %**	18 %	22 %	**20 %**	13 %	17 %	**16 %**	**31 %**	**15 %**	**40 %**	23 %
1 950- 2 000	**24 %**	39 %	13 %	**33 %**	37 %	14 %	**26 %**	46 %	12 %	**22 %**	**8 %**	**31 %**	**7 %**	24 %
> 2 000	**22 %**	22 %	10 %	**19 %**	26 %	10 %	**18 %**	17 %	4 %	**7 %**	**3 %**	**8 %**	**4 %**	14 %
TOTAL	**100 %**	100 %	100 %	**100 %**	100 %	100 %	**100 %**	100 %	100 %	**100 %**	**100 %**	**100 %**	**100 %**	100 %

Lecture :
« classe I > 55 % » signifie que cette classe regroupe les œuvres contenant plus de 55 % de paragraphes d'une ligne.

Note :
Les classes II, III et IV sont scindées en deux sous-classes : avec "B"leu (avec paragraphes longs de plus de 10 lignes) et sans "B"leu (sans paragraphes longs de plus de 10 lignes).
Classe I : plus de 55 % de paragraphes de 1 ligne
Classe II : entre 55 et 45 % de paragraphes de 1 ligne
Classe III : entre 45 et 33 % de paragraphes de 1 ligne
Classe IV : entre 33 et 21 % de paragraphes de 1 ligne
Classe V : moins de 21 % de paragraphes de 1 ligne
Classe VI : répartition de la longueur des paragraphes équilibrée
Classe VII : peu de paragraphes de taille intermédiaire

Tableau 2 : Répartition des œuvres par genre littéraire en nombre et en pourcentage

	Classe I > 55 %	Classe II : 55-45 %			Classe III : 45-33 %			Classe IV : 33-21 %			Classe V < 21 %	Classe VI Équil.	Classe VII Ø méd	TOTAL
		Sans	Avec	Total	Sans	Avec	Total	Sans	Avec	Total				
RA/REr	**5**	14	1	**15**	24	2	**26**	7	10	**17**	**2**	**:**	**1**	66
RAv	**:**	14	1	**15**	17	7	**24**	8	9	**17**	**9**	**:**	**2**	**67**
RB	**2**	2	3	**5**	4	8	**12**	1	13	**14**	**9**	**:**	**:**	**42**
RCont.	**17**	42	22	**64**	16	40	**56**	7	21	**28**	**21**	**4**	**11**	**201**
RClass.	**13**	16	15	**31**	17	68	**85**	7	96	**103**	**24**	**21**	**8**	**285**
RF	**6**	5	2	**7**	9	12	**21**	5	15	**20**	**8**	**3**	**:**	**65**
RFa	**2**	13	:	**13**	15	2	**17**	:	5	**5**	**:**	**5**	**:**	**42**
RH	**2**	19	2	**21**	21	12	**33**	14	19	**33**	**13**	**1**	**2**	**105**
RHo	**4**	8	4	**12**	5	3	**8**	:	1	**1**	**:**	**:**	**:**	**25**
RJ	**9**	2	2	**4**	15	5	**20**	15	:	**15**	**3**	**:**	**:**	**51**
RP	**53**	46	2	**48**	31	17	**48**	19	5	**24**	**2**	**:**	**:**	**175**
RSF	**2**	23	3	**26**	17	10	**27**	1	6	**7**	**1**	**1**	**:**	**64**
C	**5**	7	6	**13**	1	7	**8**	1	10	**11**	**14**	**1**	**8**	**60**
Autres	**5**	15	7	**22**	2	12	**14**	2	15	**17**	**16**	**5**	**13**	**92**
TOTAL	**125**	226	70	**296**	194	205	**399**	87	225	**312**	**122**	**41**	**45**	**1 340**

RA/REr	**8 %**	21 %	2 %	**23 %**	35 %	3 %	**38 %**	11 %	15 %	**26 %**	**3 %**	**:**	**2 %**	**100 %**
RAv	**:**	22 %	1 %	**23 %**	26 %	10 %	**36 %**	12 %	13 %	**25 %**	**13 %**	**:**	**3 %**	**100 %**
RB	**5 %**	5 %	7 %	**12 %**	10 %	19 %	**29 %**	2 %	31 %	**33 %**	**21 %**	**:**	**:**	**100 %**
RCont.	**8 %**	22 %	11 %	**33 %**	8 %	20 %	**28 %**	3 %	10 %	**13 %**	**10 %**	**2 %**	**5 %**	**100 %**
RClass.	**4 %**	6 %	5 %	**11 %**	6 %	24 %	**30 %**	3 %	35 %	**38 %**	**9 %**	**7 %**	**2 %**	**100 %**
RF	**9 %**	8 %	3 %	**11 %**	14 %	18 %	**32 %**	8 %	23 %	**31 %**	**12 %**	**5 %**	**:**	**100 %**
RFa	**5 %**	31 %	:	**31 %**	36 %	4 %	**40 %**	:	12 %	**12 %**	**:**	**12 %**	**:**	**100 %**
RH	**2 %**	18 %	2 %	**20 %**	20 %	11 %	**31 %**	13 %	19 %	**32 %**	**12 %**	**1 %**	**2 %**	**100 %**
RHo	**16 %**	32 %	16 %	**48 %**	20 %	12 %	**32 %**	:	4 %	**4 %**	**:**	**:**	**:**	**100 %**
RJ	**18 %**	4 %	4 %	**8 %**	29 %	10 %	**39 %**	29 %	:	**29 %**	**6 %**	**:**	**:**	**100 %**
RP	**30 %**	26 %	1 %	**27 %**	19 %	9 %	**28 %**	11 %	3 %	**14 %**	**1 %**	**:**	**:**	**100 %**
RSF	**3 %**	36 %	5 %	**41 %**	26 %	15 %	**41 %**	2 %	9 %	**11 %**	**2 %**	**2 %**	**:**	**100 %**
C	**8 %**	11 %	10 %	**21 %**	2 %	11 %	**13 %**	2 %	16 %	**18 %**	**21 %**	**6 %**	**13 %**	**100 %**
Autres	**5 %**	17 %	8 %	**25 %**	2 %	14 %	**16 %**	2 %	16 %	**18 %**	**17 %**	**5 %**	**14 %**	**100 %**
TOTAL	**9 %**	17 %	5 %	**22 %**	15 %	15 %	**30 %**	7 %	17 %	**23 %**	**9 %**	**3 %**	**3 %**	**100 %**

Lecture :
« classe I > 55 % » signifie que la classe I regroupe les œuvres contenant plus de 55 % de paragraphes d'une ligne.

Note :
Certains textes relèvent de deux genres littéraires, par exemple RC/RH. Aussi, le total des textes compris dans les genres littéraires (1 340) est supérieur au total des œuvres présentes dans le corpus (1 255). Les romans contemporains (RCont.) et les romans classiques (RClass.) se différencient par le fait que les romans contemporains ont été publiés après 1914 et les romans classiques avant 1914.