Goncourt 2020 : mais qu'a-t-il de plus que les autres ?

Goncourt Prize Winner 2020: Why Is It the Best?

Michel Bernard

Édition électronique

URL : <https://journals.openedition.org/revuehn/2297>

DOI : 10.4000/revuehn.2297

ISSN : 2736-2337

Éditeur

Humanistica

Référence électronique

Michel Bernard, « Goncourt 2020 : mais qu'a-t-il de plus que les autres ? », *Humanités numériques* [En ligne], 4 | 2021, mis en ligne le 01 décembre 2021, consulté le 29 novembre 2022. URL : <http://journals.openedition.org/revuehn/2297> ; DOI : <https://doi.org/10.4000/revuehn.2297>

Creative Commons - Attribution 4.0 International - CC BY 4.0
<https://creativecommons.org/licenses/by/4.0/>

Goncourt 2020 : mais qu'a-t-il de plus que les autres ?

Goncourt Prize Winner 2020: Why Is It the Best?

Michel Bernard

Résumés

Peut-on rendre compte des choix successifs de l'académie Goncourt – qui ont abouti en 2020 à distinguer *L'Anomalie* d'Hervé Le Tellier – sur la base de critères purement textuels ? En utilisant les ressources de la textométrie sur un corpus constitué des 15 romans de la première sélection Goncourt et de 32 autres romans de la rentrée littéraire 2020, on essaiera de déterminer quels marqueurs lexicaux ou stylistiques peuvent expliquer les quatre phases de sélection. Le constat final est négatif, en ce sens qu'aucun des critères identifiés ne présente de continuité concluante sur les quatre étapes. Cela renvoie dans un premier temps à la prise en compte de caractéristiques extratextuelles pour l'attribution du prix mais aussi au caractère indécidable des choix successifs opérés par des lecteurs professionnels dans le cadre d'une négociation.

Can we explain the successive choices of the académie Goncourt – which in 2020 led *L'Anomalie* by Hervé Le Tellier to win the Goncourt prize – on the basis of purely textual criteria? Using textometric tools on a corpus made up of the 15 novels of the first Goncourt selection and 32 other novels from the 2020 literary season, we will try to determine which lexical or stylistic markers can explain the four phases of selection. The final finding is negative, in the sense that none of the criteria identified has conclusive continuity over the four stages. This refers to the consideration of extra-textual characteristics for awarding the prize but also to the undecidability of the successive choices made by professional readers in the context of a negotiation.

Entrées d'index

MOTS-CLÉS : littérature, analyse textuelle, roman, stylistique

KEYWORDS: literature, textual analysis, novel, stylistics

« Nul n'est à l'abri du succès. De toute façon, Miesel s'en moque. *Des échecs qui ont raté*, son dernier roman, s'était retrouvé dans les premières listes du Médicis, du Goncourt et du Renaudot, pour disparaître quinze jours plus tard des deuxième sélections [...] »

Hervé Le Tellier, *L'Anomalie*, Gallimard, 2020

¹ Comme chaque année, moyennant quelques ajustements dus à la situation sanitaire, le mois de novembre a vu la proclamation rituelle du lauréat annuel du prix Goncourt, en l'espèce, Hervé Le Tellier pour son roman *L'Anomalie*.

² Malgré ce titre inquiétant, rien cependant que de très normal dans le processus qui a abouti à cette distinction : les jurés du prix ont d'abord sélectionné, le 15 septembre, parmi les très nombreux ¹ romans de la rentrée littéraire, une première liste de quinze titres :

1. *L'Anomalie* d'Hervé Le Tellier (Gallimard)
2. *L'Historiographe du royaume* de Maël Renouard (Grasset)
3. *Les Impatientes* de Djaili Amadou Amal (Emmanuelle Collas)
4. *Thésée, sa vie nouvelle* de Camille de Toledo (Verdier)
5. *Héritage* de Miguel Bonnefoy (Rivages)
6. *Les Funambules* de Mohammed Aïssaoui (Gallimard)
7. *Mes fous* de Jean-Pierre Martin (L'Olivier)
8. *Un crime sans importance* d'Irène Frain (Seuil)
9. *Chavirer* de Lola Lafon (Actes Sud)
10. *L'Enfant céleste* de Maud Simonnot (L'Observatoire)
11. *La Chambre des dupes* de Camille Pascal (Plon)
12. *La Société des belles personnes* de Tobie Nathan (Stock)
13. *Les Roses fauves* de Carole Martinez (Gallimard)
14. *Saturne* de Sarah Chiche (Seuil)
15. *Yoga* d'Emmanuel Carrère (P.O.L)

³ Après quelques semaines de lecture, le 7 octobre, les jurés ont publié une seconde liste, ne contenant que les huit premiers titres de la liste ci-dessus, puis, le 27 octobre, une liste comprenant les quatre premiers titres. Enfin, le 30 novembre, le lauréat était élu, par huit voix contre deux, au terme de cette sélection progressive.

⁴ Quels sont les critères qui président à ce processus ? Sur quelles règles se base-t-on pour passer de quinze à huit, de huit à quatre puis à un ? Qu'est-ce que ce livre a de plus que tous les autres, qui pourrait expliquer le choix final ? Comme pour toute opération relevant du goût, et spécialement du goût littéraire, la subjectivité règne ici en maître, et il serait

vain de demander des comptes au jury, même si chaque année son choix donne lieu à de nombreuses contestations et supputations. Cette opération est, à l'évidence, influencée par de nombreux facteurs extérieurs au texte : personnalité et carrière des auteurs, thèmes d'actualité, puissance des éditeurs, etc. Ils se réfèrent en outre à un « talent littéraire » auquel il serait impossible d'associer des caractéristiques objectivables.

5 Cependant, on partira dans la présente étude d'une hypothèse naïve : est-ce que les différences textuelles entre ces romans permettent de comprendre les choix du jury Goncourt ? Il ne s'agira pas ici de rechercher quelque formule miraculeuse qui permettrait de fabriquer des Goncourt à coup sûr, mais de se demander comment, en 2020, les jurés ont établi leur choix, au moins partiellement, sur des critères purement textuels. Je ne pousserai toutefois pas la naïveté jusqu'à me demander si ces critères textuels sont les seuls à entrer en ligne de compte, ni si le talent d'un romancier ou d'une romancière peut se subsumer à l'addition des mots qu'il ou elle a utilisés, ni encore si la résolution des divergences d'avis entre les jurés a pu influencer sur le choix final, ni même si chaque membre du jury a vraiment lu l'intégralité des quinze romans... L'important ici est d'appliquer une méthode, celle de la statistique textuelle, à un corpus fermé au sein duquel on peut opérer une partition nette : première sélection, deuxième sélection, troisième sélection, choix final. Il s'agit ensuite de se demander si les différences textuelles entre ces parties du corpus peuvent expliquer le processus de sélection. Pour l'exprimer autrement : est-ce que les corpus correspondant aux quatre étapes de choix du jury pourraient être différenciés de manière automatique en trouvant les critères pertinents ? Peut-on constater un affinement progressif de la liste en fonction de critères textuels mesurables et cohérents de la première à la dernière sélection, le roman primé étant ainsi celui qui présenterait au plus haut point les caractéristiques retenues ? J'ai donc réuni, à cet effet, un ensemble constitué des textes intégraux des quinze romans élus par le jury du Goncourt² et je lui ai appliqué les traitements habituels de la textométrie³.

6 Cependant, pour une première approche, j'ai réuni un corpus plus large, contenant 32 romans de la rentrée littéraire 2020 qui n'ont pas été retenus par le jury. Il s'agit ainsi de vérifier si les 15 romans sélectionnés dans un premier temps montrent déjà des caractéristiques repérables, indicatrices des goûts du jury. Même si ce corpus est assez étendu (2 732 402 mots pour les 32 romans non sélectionnés, 1 132 115 pour les 15 sélectionnés), il ne s'agit encore que d'un échantillon assez faible des 511 romans identifiés par *Livres Hebdo*. J'ai simplement choisi des romans signalés par la critique et éligibles au prix Goncourt, au sens où ils remplissent les critères de ce prix : roman écrit en français, auteur non encore primé (*Le Pays des autres* de Leïla Slimani a, par exemple, été écarté) et non-membre du jury (*Fille* de Paule Constant, par exemple). Je supposerai donc que ce corpus est représentatif des romans écartés dès la première sélection mais qui auraient pu y figurer⁴. En voici la liste :

- *Art nouveau* de Paul Greveillac (Gallimard)
- *Broadway* de Fabrice Caro (Gallimard)
- *Buveurs de vent* de Franck Bouysse (Albin Michel)
- *Ce qu'il faut de nuit* de Laurent Petitmangin (La Manufacture des livres)

- *Comédies françaises* d'Éric Reinhardt (Gallimard)
- *Comme un empire dans un empire* d'Alice Zeniter (Flammarion)
- *Crénom, Baudelaire !* de Jean Teulé (Mialet-Barrault)
- *Du côté des Indiens* d'Isabelle Carré (Grasset)
- *Histoires de la nuit* de Laurent Mauvignier (Éditions de Minuit)
- *La Famille Martin* de David Foenkinos (Gallimard)
- *La Fièvre* de Sébastien Spitzer (Albin Michel)
- *La Grande Épreuve* d'Étienne de Montety (Stock)
- *Le Bon, la brute et le renard* de Christian Garcin (Actes Sud)
- *Le Cœur synthétique* de Chloé Delaume (Seuil)
- *Le Grand Vertige* de Pierre Ducrozet (Actes Sud)
- *Le Palais des orties* de Marie Nimier (Gallimard)
- *Le Petit Polémiste* d'Ilan Duran Cohen (Actes Sud)
- *Le Tailleur de Relizane* d'Olivia Elkaim (Stock)
- *Le Temps gagné* de Raphaël Enthoven (L'Observatoire)
- *Les Aérostats* d'Amélie Nothomb (Albin Michel)
- *Les Évasions particulières* de Véronique Olmi (Albin Michel)
- *Les Nuits d'été* de Thomas Flahaut (Éditions de l'Olivier)
- *Liv Maria* de Julia Kerninon (L'Iconoclaste)
- *Nature humaine* de Serge Joncour (Flammarion)
- *Rachel et les siens* de Metin Arditi (Grasset)
- *Sabre* d'Emmanuel Ruben (Stock)
- *Térébenthine* de Carole Fives (Gallimard)
- *Un jour viendra couleur d'orange* de Grégoire Delacourt (Grasset)
- *Une chance sur un milliard* de Gilles Legardinier (Flammarion)
- *Une fille de rêve* d'Éric Laurrent (Flammarion)
- *Une piscine dans le désert* de Diane Mazloum (JC Lattès)
- *Une rose seule* de Muriel Barberly (Actes Sud)

La première sélection

7

Si l'on compare les deux corpus (15 romans sélectionnés contre 32 romans non sélectionnés), on voit apparaître quelques caractéristiques textuelles intéressantes :

- Les romans sélectionnés sont en moyenne plus courts que ceux qui n'ont pas été retenus : 75 474 mots en moyenne contre 85 386. Mais les variations sont considérables. Les romans non sélectionnés vont de 26 262 mots (Nothomb) à 190 182 (Mauvignier) et les romans sélectionnés de 28 316 (Simonnot) à 150 377 (Pascal).
- Les romans sélectionnés contiennent davantage⁵ de pronoms et d'adjectifs de la première personne (*je, me, mon, nous*) mais beaucoup moins de la deuxième personne (*tu, te*) et de la troisième (*il, elle, se*). En particulier, le pronom *elle* est particulièrement rare dans les romans sélectionnés, suggérant une prédominance des personnages masculins⁶.

- On observe qu'ils contiennent davantage de guillemets mais beaucoup moins de tirets de dialogue.
- Le point-virgule et les deux-points sont également caractéristiques de la sélection Goncourt. En revanche, ils contiennent moins de points, de points d'interrogation, de points de suspension, ce qui indique des phrases plus longues.
- Ils contiennent moins de tournures négatives (*pas, non*), moins de comparaisons (*comme*).
- Quelques indices laissent entendre que les romans sélectionnés usent d'un français plus châtié : *pas* est par exemple plus fréquent dans les romans non sélectionnés mais pas l'adverbe *ne*, ce qui indique des négations simples, la forme *ça* est plus courante dans les non sélectionnés, ainsi que des mots comme *gueule, foutre, mec, chose* (mais ce mot est quand même très fréquent chez Emmanuel Carrère), *con, gars*. Ils utilisent aussi plus d'interjections. À l'inverse, les romans de la sélection utilisent davantage des mots comme *car, aussitôt, cependant, ô, désormais, dont, quelquefois*, qui relèvent d'une langue plus soutenue.
- Les romans sélectionnés ont des groupes nominaux plus développés : plus de noms communs, de prépositions, de déterminants (possessifs, articles), d'adjectifs numéraux. Symétriquement, ils contiennent moins d'adverbes, moins de pronoms personnels.
- Aucun des deux groupes n'utilise plus de verbes que l'autre mais les temps et les modes sont diversifiés : plus de participes passés (et donc de temps composés) et de présents pour les romans sélectionnés, plus d'imparfaits, de participes présents, de conditionnels dans les romans non sélectionnés.

8

Sur le plan thématique, il est plus difficile de distinguer les deux groupes car certaines spécificités ne tiennent en réalité qu'à un seul roman. Ainsi, les mots *roi, duc, duchesse, Majesté, Cour, reine* sont plus fréquents dans le groupe des romans sélectionnés mais cela tient presque exclusivement au roman *La Chambre des dupes*, qui contient par exemple à lui seul 1 232 des 2 020 occurrences du mot *roi*. En ne retenant que les mots qui sont répartis sur plusieurs romans, on peut seulement remarquer que ceux qui sont plus employés par les romans sélectionnés sont *mort* (il est particulièrement fréquent chez Carole Martinez, Camille de Toledo et Irène Frain) et *mourir* (chez les mêmes auteurs et Sarah Chiche), *soldat* (particulièrement chez Maël Renouard, Miguel Bonnefoy et Tobie Nathan), *médecin* (chez Camille Pascal, Sarah Chiche et Irène Frain, mais aussi chez Sébastien Spitzer, qui n'est pas dans les auteurs sélectionnés), *avion* (essentiellement dans les romans d'Hervé Le Tellier et de Miguel Bonnefoy), *famille* (principalement chez Djaili Amadou Amal mais aussi chez Mohammed Aïssaoui et Irène Frain), *récit, secret*. À l'inverse, et en prenant les mêmes précautions, on trouve moins, dans les romans sélectionnés, les mots *envie, bière, bras et main, voir et regarder, air*. Mais ces ensembles sont trop disparates pour tracer une limite nette entre les deux corpus, pour trouver ce qui serait systématiquement présent dans les romans sélectionnés et absent dans les romans non sélectionnés (ou l'inverse).

Ce que nous venons de repérer ne trace qu'un profil moyen des romans des deux groupes, et ne rend pas bien compte des variations individuelles. Par exemple, la prédominance des noms communs, que j'ai signalée plus haut, correspond en réalité à la distribution suivante : certains des 15 romans sélectionnés (figure 1, à droite, encadré) contiennent une proportion très importante de noms (*La Chambre des dupes*, *Héritage*, *Chavirer*). Mais les autres n'en contiennent pas des quantités remarquables, et la balance ne penche de ce côté qu'en raison des nombreux romans non sélectionnés qui, eux, présentent des sous-effectifs importants de noms communs, alors qu'aucun des romans sélectionnés ne présente de tels déficits. Mais cela n'empêche pas par exemple *Sabre*, qui n'a pas été sélectionné pour le Goncourt, de présenter une forte proportion de noms communs.

FIGURE 1. SPÉCIFICITÉS DES NOMS COMMUNS

Un autre procédé statistique consiste à représenter les romans dans un graphique d'analyse factorielle des correspondances (AFC). Pour le dire vite, il s'agit de cartographier les ressemblances de vocabulaire des 47 romans du corpus dans un graphique où la proximité des points indique une proximité lexicale⁷. Mais le résultat n'est guère lisible (figure 2) : pour résumer, la quasi-totalité des 47 romans, sélectionnés ou non, se présente comme une masse homogène sans véritables spécificités. Les seuls romans qui sortent du lot, et dont on peut dire qu'ils présentent, au moins sur le plan du lexique, une vraie originalité sont deux romans qui figurent d'ailleurs dans la troisième sélection du Goncourt (le chiffre qui précède le titre indique dans quelle liste il se trouve, de 0 pour les romans non sélectionnés à 4 pour le lauréat) : *Thésée, sa vie nouvelle* et *L'Historiographe du royaume*⁸. On distingue aussi *La Chambre des dupes* (dont nous avons déjà vu qu'il utilisait un vocabulaire très spécifique) mais également deux romans qui n'ont pas été retenus par le jury : *Sabre* et *La Fièvre*. Les romans d'Amélie Nothomb et de Laurent Mauvignier sortent aussi – très légèrement – de la masse centrale. En revanche, on notera dès à présent que le Goncourt 2020, le roman d'Hervé Le Tellier, se situe au milieu de ce conglomerat.

[illegible]

Une manière plus lisible de présenter le même traitement consiste à classer les romans par ressemblance, sous la forme d'un arbre (figure 3⁹) ; l'algorithme procède par bipartition, sur la base des ressemblances lexicales¹⁰ : il isole d'abord *Thésée, sa vie nouvelle* par rapport à tous les autres romans, puis il sépare un groupe constitué de quatre romans (Nothomb, Mauvignier, Olmi, Kerninon) de tous les autres, et ainsi de suite. L'affichage sous forme d'arbre permet plus aisément de repérer des ensembles de romans proches par le vocabulaire. Hélas, on ne trouve pas de *cluster* qui regrouperait miraculeusement les trois sélections du Goncourt. Tout au plus peut-on noter que le roman de Le Tellier se trouve dans un ensemble de sept romans (encadré dans la figure 3) qui contient trois romans de la deuxième sélection et un roman de la première sélection, et que deux autres romans de la deuxième sélection, *Un crime sans importance* et *Les Funambules*, sont appairés par l'algorithme.

La deuxième sélection

12

Mais ces résultats peu concluants pourraient tenir à ce corpus assez hétérogène, où l'on a empilé des romans certes publiés à la rentrée 2020 mais dont la liste n'a été établie qu'un peu au hasard des avis de divers critiques et ne peut même pas se réclamer d'une quelconque exhaustivité. Voyons à présent si l'étude des seuls quinze romans sélectionnés permet de trouver des spécificités plus régulières, puisqu'il s'agit là d'un corpus plus légitime, au sens où il a été délimité par le jury en tant que tel. Dans un premier temps, nous comparerons les huit romans de la deuxième sélection avec les sept de la première sélection qui n'ont pas été finalement retenus.

13

Dans une première approche, il peut être intéressant de reprendre les spécificités que nous avons mises en évidence pour distinguer les quinze romans de la sélection des autres romans de la rentrée littéraire, dans le but de vérifier si ces critères ont pu être prolongés dans la deuxième sélection :

- Les romans de la deuxième sélection sont ainsi plus courts que ceux de la première sélection (64 862 mots en moyenne contre 87 603). De ce point de vue, les romans abandonnés après la première sélection sont très proches de ceux qui n'avaient pas été sélectionnés du tout (85 386 mots en moyenne). Il y a donc là une continuité.

- De même, ces huit romans sont caractérisés par un usage plus important de la première personne (*je, me, mon, moi*) et le pronom *elle* y est plus rare. En revanche, ils contiennent plus de mots de la deuxième personne (*tu, te, toi, ton*) et de *il*. Les personnages y sont donc davantage masculins et il y a davantage de dialogues. Au demeurant, plusieurs de ces romans sont très largement des narrations à la première personne, ce qui explique en partie les fréquences constatées : *Les Funambules*, *Les Impatientes*, *L'Historiographe du royaume*, *Mes fous*, *Un crime sans importance* et, partiellement, *Thésée, sa vie nouvelle*.
- Les guillemets y sont plus fréquents, mais aussi les tirets de dialogue, ce qui confirme l'hypothèse d'une plus grande quantité de dialogues et de discours rapportés à la première personne.
- Les points-virgules et les deux-points sont plus fréquents mais aussi, à la différence de la première comparaison, les points d'interrogation (sans doute, une fois de plus, en raison des dialogues). Le point y est un peu plus fréquent que dans la première sélection mais dans des proportions assez modestes, ce qui veut dire que les phrases y sont légèrement moins longues, mais sans que cela soit très net. En réalité, c'est *Thésée, sa vie nouvelle* qui brouille ici la statistique car sa typographie très particulière exclut presque systématiquement le point.
- L'usage de la négation et de la comparaison (*comme*) ne présente pas de différence dans les deux ensembles.
- Il n'y a plus de différenciation quant au niveau de langue. Les négations sont presque toujours doubles et il n'y a plus de mots familiers ou soutenus en spécificité. C'est sans doute le résultat de la première sélection.
- Pour ce qui est des groupes nominaux, la spécificité est ici inversée. C'est-à-dire que les huit romans de la deuxième sélection ont beaucoup moins de noms, d'articles et de prépositions que les sept autres. La raison tient ici encore principalement à un seul roman, *La Chambre des dupes*, un roman de la première sélection, qui contient une quantité extrêmement élevée de groupes nominaux, ce qui a suffi à conférer à l'ensemble des romans sélectionnés cette caractéristique. En revanche, les huit romans de la deuxième sélection ont beaucoup plus de pronoms personnels que les sept autres.
- Comme pour la première comparaison, aucun des deux groupes n'a plus de verbes que l'autre mais les temps verbaux sont à nouveau très contrastés : présent, imparfait du subjonctif, futur et passé simple pour la deuxième sélection, imparfait et participe présent pour la première. Cela prolonge donc, en gros, la toute première sélection, puisque nous avons vu que les quinze romans sélectionnés avaient plus de verbes au présent et moins d'imparfaits. En revanche, les participes passés et le conditionnel ne sont plus ici déterminants. Il est d'ailleurs possible que la prédominance des présents soit en lien avec celle des dialogues et du discours direct. Au demeurant, il faut là encore tenir compte des variations individuelles. Ainsi, dans la deuxième sélection, ce sont *L'Anomalie* et *Les Funambules* qui contiennent le plus de verbes au présent, et à l'in-

verse dans *Héritage* et *L'Historiographie du royaume*, deux romans historiques, ils sont particulièrement rares, au profit des temps du passé.

14 Nous nous heurtons donc ici à la même difficulté que dans notre première comparaison : certaines des caractéristiques mises en évidence par le traitement statistique sont en réalité attribuables à un seul ou éventuellement à deux ou trois romans, et l'on ne peut donc pas y voir un critère explicatif du choix puisque les autres romans de la sélection ne présentent pas la même particularité. Ce phénomène est même plus important encore dès lors que le corpus ne contient plus que quinze romans au lieu de quarante-sept. Nous devons donc chercher des spécificités qui soient communes à tous les romans d'une des sélections.

15 C'est par exemple le cas du mot *devoir* (verbe ou nom), caractéristique de la deuxième sélection¹¹. Voici comment se présente le graphique des spécificités (figure 4) : comme on le voit, tous les romans de la deuxième sélection (encadrés) utilisent ce mot en excès ou avec une spécificité négative négligeable¹². De l'autre côté, tous les romans de la première sélection l'emploient peu ou très peu, à l'exception d'un seul roman, où la spécificité est là encore négligeable. Un mot comme celui-ci correspond donc bien à ce que nous recherchons : un critère qui permettrait d'expliquer l'ensemble de la sélection.

FIGURE 4. OCCURRENCES DU MOT *DEVOIR*

16 Avant de tenter une analyse, cherchons de manière systématique tous les mots qui correspondent à ce schéma. Il y en a assez peu, à vrai dire, et encore faut-il parfois admettre des exceptions : *cependant* est plus fréquent dans la deuxième sélection (avec toutefois une exception notable, celle de *L'Anomalie*, le Goncourt 2020...), *chambre* dans la première (mais il reste rare dans *La Société des belles personnes*), ainsi que *pied* (mais là encore *L'Anomalie* fait exception).

17 On avouera que la récolte est maigre, et la conclusion dérisoire... Le jury aurait-il choisi des romans qui parlent de *devoir*, qui utilisent l'adverbe *cependant* et qui évitent de parler de *chambre* et de *pied* ? À l'inverse, il est plus sûr d'avancer qu'il n'y a pas de caractéristiques textuelles qui permettent d'expliquer la deuxième sélection du Goncourt. L'AFC nous confirme cette impossibilité de discrimination textuelle, tout au moins sur la base des 200 mots les plus fréquents (figure 5) : on retrouve les mêmes caractéristiques que pour l'AFC portant sur les 47 ro-

mans. Les deux mêmes romans (de la deuxième sélection) sont décalés par rapport à une masse compacte de romans assez semblables du point de vue lexical.

FIGURE 5. ANALYSE FACTORIELLE DES CORRESPONDANCES DES 15 ROMANS DE LA PREMIÈRE SÉLECTION

18

Même constats avec la classification (figure 6) : il n'y a évidemment toujours pas deux classes bien définies avec les deux sélections regroupées dans une même branche... Les romans des deux sélections sont mêlés et appairés de manière indistincte. Notons dès à présent que le Goncourt est rapproché d'un roman de la première sélection, celui de Tobie Nathan.

La troisième sélection

19

Nous en arrivons à la dernière liste établie par le jury, celle des quatre romans qui restaient en lice au 27 octobre. Nous avons déjà pu constater que deux de ces romans, *Thésée, sa vie nouvelle* et *L'Historiographe du royaume*, présentaient des caractéristiques lexicales qui les distinguaient des 45 autres œuvres. On observera également dans la classification ci-dessus (figure 6) que trois des romans de la troisième sélection (à l'exception notable du lauréat de l'année, comme nous venons de le noter) sont classés à part, dans une branche individuelle de l'arbre. Cela laisse attendre des textes atypiques, avec de fortes identités lexicales. Voici quelques constats établis en comparant les quatre romans qui n'ont pas dépassé le stade de la deuxième sélection aux quatre qui ont figuré dans cette dernière *short list*. On se concentrera surtout, une fois de plus, sur les différences entre les critères possibles de chaque niveau de sélection.

- Les romans de la troisième sélection sont plus volumineux (76 956 mots en moyenne contre 52 768), ce qui inverse le constat précédent. Mais ces moyennes recouvrent des écarts très importants puisque *Les Impatientes*, avec ses 52 210 mots – soit quasiment la taille moyenne des romans de la deuxième sélection – est deux fois plus mince que *L'Anomalie* ou *L'Historiographe du royaume*.

- On retrouve la forte spécificité des pronoms et adjectifs de la deuxième personne (*tu, te, vous, ton, toi, votre*) mais *je* est passé en spécificité négative (ce qui est d'autant plus remarquable que trois des romans sont, au moins partiellement, des récits à la première personne). Le pronom *elle*, en revanche, est toujours particulièrement rare, sans d'ailleurs que *il* soit plus fréquent. On notera aussi la grande fréquence de *nous, notre* et, à l'inverse, la faible fréquence de *on*.
- Les tirets sont encore plus fréquents mais ce n'est plus le cas des guillemets.
- Sur le plan de la ponctuation, nous retrouvons la fréquence du point-virgule mais les deux-points sont déficitaires, alors qu'on trouve davantage de points d'exclamation et d'interrogation, plus de virgules. Le point est plus rare mais c'est à nouveau *Thésée, sa vie nouvelle* qui fausse la statistique, et de manière encore plus nette puisqu'il représente ici un quart du corpus.
- *Comme* est plus rare dans cette dernière sélection, ce qui semble renouer avec un des critères de la première sélection.
- Le déficit en groupes nominaux est encore très net dans cette troisième sélection : moins de noms, d'articles indéfinis. Mais, contrairement à ce qui avait été noté pour la seconde sélection, les pronoms personnels y sont plus rares.
- Pour ce qui est des temps verbaux, la troisième sélection se caractérise par la surfréquence du futur, de l'imparfait du subjonctif, du conditionnel et de l'infinitif, et par la sous-fréquence du passé simple et du participe passé.
- On constate également des phénomènes qui n'étaient pas apparus jusqu'alors : une plus grande utilisation des conjonctions *et, car*, des relatifs *qui* et *que* ; mais aussi la rareté de *avec* (c'est *Héritage* qui en fait un usage très abondant).

Je ne reviendrai pas ici sur le rôle des particularités individuelles dans ces comparaisons par blocs. Il va de soi que plus le corpus se réduit et plus il est sensible à ces variations. En revanche, et pour la même raison, ce corpus réduit commence à faire apparaître des positionnements significatifs sur l'AFC (figure 7) : trois des quatre romans restés en deuxième sélection sont groupés et un seul (*Héritage*) s'en distingue nettement. Les quatre romans de la troisième sélection, à l'opposé, sont tous bien espacés et présentent donc de fortes particularités lexicales. Il est intéressant de constater que les deux qui sont les plus proches (*L'Historiographe du royaume* et *L'Anomalie*) sont précisément les deux finalistes, au sens où ils sont les deux seuls à avoir reçu des suffrages lors du dernier vote.

FIGURE 7. ANALYSE FACTORIELLE DES CORRESPONDANCES DES 8 ROMANS DE LA DEUXIÈME SÉLECTION

21

Ce positionnement est confirmé par la classification (figure 8) : cette fois, chacun des romans de la troisième sélection est seul sur sa branche, à l'exception de *L'Historiographe du royaume*, rapproché d'*Héritage*.

FIGURE 8. ANALYSE ARBORÉE DES 8 ROMANS DE LA DEUXIÈME SÉLECTION

Le choix final

22

Notre dernière étape consistera à comparer le roman de Le Tellier avec les trois autres romans de la troisième sélection pour essayer une dernière fois de repérer des spécificités qui pourraient expliquer cet ultime choix du jury. Ce faisant, je récapitulerai chaque fois les caractéristiques déjà notées lors de trois premières sélections pour vérifier si ce dernier choix confirme l'hypothèse d'une continuité des critères textuels.

- Le critère de la taille du roman va dans le même sens que celui que nous avons constaté précédemment : le Goncourt est le roman le plus gros de la troisième sélection, et même de la deuxième¹³.
- *L'Anomalie* se caractérise par la grande fréquence du pronom *vous*, ce qui apparaissait dans le comparatif précédent. De même, les pronoms et adjectifs de la première personne du singulier (*me, mon, je, moi*) ainsi que *on* sont particulièrement rares dans ce roman. En revanche, *elle* est cette fois beaucoup plus fréquent et *il* beaucoup moins¹⁴. Les mots de la deuxième personne du singulier (*tu, ton, te*) sont très peu présents et *nous* n'a pas de spécificité notable. Il n'y a donc pas, sur ce point, de continuité avec ce que nous avons constaté jusque-là.
- En revanche, il y a bien une grande quantité de tirets de dialogue dans le roman de Le Tellier, et cela le différencie nettement des autres œuvres.
- Cette fois, ce sont les deux-points et le point d'interrogation qui sont en surnombre mais le point-virgule est ici particulièrement rare (seulement 36 dans tout le roman), ainsi que le point d'exclamation. Par contre, Le Tellier utilise plus que les autres la virgule. Pour ce qui est de la longueur des phrases, les statistiques sont encore perturbées par *Thésée, sa vie nouvelle*. Voici dans le tableau 1 la longueur moyenne des phrases dans la deuxième sélection¹⁵.

TABLEAU 1. LONGUEUR MOYENNE DES PHRASES DANS LA DEUXIÈME SÉLECTION

<i>Thésée, sa vie nouvelle</i> de Camille de Toledo	100,1
<i>L'Historiographe du royaume</i> de Maël Renouard	37,5
<i>Héritage</i> de Miguel Bonnefoy	27,7
<i>L'Anomalie</i> d'Hervé Le Tellier	18,9
<i>Les Funambules</i> de Mohammed Aïssaoui	17,9
<i>Un crime sans importance</i> d'Irène Frain	17,6
<i>Mes fous</i> de Jean-Pierre Martin	15,2
<i>Les Impatientes</i> de Djaïli Amadou Amal	13,7

23

L'Anomalie a donc des phrases d'une longueur comparable à celle des autres romans.

- Comme dans la sélection précédente, le roman utilise moins que les autres toute une gamme de conjonctions, de prépositions, de conjonctions et de pronoms relatifs (*de, que, qui, car, dont, quand*). Cela suggère des phrases plus simples, sans compléments.

- L'examen des catégories grammaticales confirme qu'il y a dans le roman de Le Tellier, plus globalement, moins de prépositions, moins de pronoms relatifs, moins de conjonctions mais aussi moins d'adjectifs possessifs et de déterminants que dans les autres. En revanche, il se caractérise par de hautes fréquences de noms propres, d'interjections et de numéraux. Ainsi, rien de ce que nous avons observé précédemment sur les groupes nominaux ne se trouve confirmé par cet ultime choix. De plus, on constate que les pronoms personnels sont très rares dans le roman. Sans doute faut-il mettre cela en perspective avec la fréquence des noms propres : les personnages sont désignés par leur nom plutôt que par un pronom, ce qui est compréhensible dans un roman qui compte un grand nombre de personnages (parfois doubles !) que le lecteur pourrait confondre.
- *L'Anomalie* comporte peu de verbes et ces verbes sont préférentiellement au présent. L'auteur semble à l'inverse éviter le passé simple et l'imparfait.
- On notera aussi que le roman est beaucoup plus riche en adjectifs qualificatifs.
- Il est inutile, dans ce cas, d'étudier les spécificités thématiques. Elles feraient simplement apparaître les thèmes propres au roman primé : *passager, président, protocole, simulation, agent* et, évidemment, *avion*.

24

Pour résumer, il y a donc peu de spécificités qui plaideraient pour une continuité des quatre choix du jury. Il ne reste guère que les tirets de dialogue, les deux-points et les verbes au présent qui pourraient suggérer une permanence des critères de choix. Ce ne sont donc pas les phénomènes lexicaux pris en compte jusqu'à maintenant qui peuvent donner des indications stables sur les critères du jury.

Facilité de lecture

25

Essayons un autre angle : la facilité de lecture. Il s'agit évidemment d'une notion très controversée¹⁶ mais il peut être intéressant d'en tester certains principes pour vérifier sa pertinence pour notre corpus. Après tout, un prix littéraire ne vise-t-il pas à recommander au public une œuvre littéraire qui doit satisfaire le plus grand nombre ? L'existence du Goncourt des lycéens, basé sur la même première sélection effectuée par le jury Goncourt, peut aussi inciter les jurés à sélectionner, au moins dans un premier temps, des œuvres accessibles à un public plus jeune.

26

Un indicateur courant de lisibilité est l'indice de Flesch. Il s'agit d'un calcul basé sur la longueur moyenne des mots (en nombre de syllabes) et la longueur des phrases (en nombre de mots). Cet indice peut évidemment donner lieu à de nombreuses contestations mais il est souvent utilisé dans les logiciels (Microsoft Word, par exemple) en raison de sa commodité. J'ai donc calculé pour les 48 romans du premier corpus le *Flesch Reading Ease Level*. C'est un indice qui va de 0 à 100, du plus difficile au plus facile¹⁷. En l'espèce, les indices s'échelonnent de 5,48 (*Histoires de la nuit* de Laurent Mauvignier¹⁸) à 83,05 (*La Fièvre* de Sébastien Spitzer). Les comparaisons de moyennes ne montrent pas des diffé-

rences spectaculaires, mais une réelle progression : les 32 romans non sélectionnés ont un indice moyen de 63,6 et les 15 romans sélectionnés de 67,4¹⁹. Ces derniers sont donc effectivement un peu plus « faciles » à lire que les autres. Même évolution entre les deux sélections : les 7 romans restés en première sélection ont un indice moyen de 65,6 et les 8 de la deuxième sélection de 69. Ils sont donc encore plus « faciles » à lire, le roman finalement primé ayant pour sa part un indice de 70,11, ce qui le place encore légèrement au-dessus en matière de lisibilité²⁰.

27 Voilà qui nous donne une indication mais l'indice de Flesch ne prend pas en compte la difficulté du vocabulaire. Un mot peut en effet être court, figurer dans une phrase courte mais présenter quand même des difficultés de lecture en raison de sa rareté. Nous avons déjà vu que les œuvres sélectionnées usaient d'un français plus soutenu et l'on peut se demander si cela ne va pas de pair avec un vocabulaire plus recherché, moins courant.

28 La première piste explorée sera d'étudier la proportion de vocabulaire courant dans chaque roman. J'ai utilisé pour cela une liste de 1 500 mots établie par Étienne Brunet et proposée aux enseignants sur le site du ministère de l'Éducation nationale²¹. Il s'agit ensuite de calculer la proportion de ces mots très courants dans le vocabulaire des romans²². Le premier constat est très surprenant : le roman qui contient le plus de mots courants (75,56 %) est *Histoires de la nuit* de Laurent Mauvignier, que l'indice de Flesch donnait comme le roman le plus « difficile » des 47 œuvres étudiées, et celui qui en contient le moins (62 %) est *L'Anomalie*, c'est-à-dire le Goncourt 2020, c'est-à-dire celui dont l'indice de Flesch suggérerait qu'il avait été choisi au terme d'une élimination progressive des romans les moins lisibles... L'examen des moyennes par groupe de sélection confirme cette difficulté : les 32 romans non sélectionnés ont un ratio de 67,99 % et les 15 sélectionnés de 68,96 %. La différence, qui indique que les romans de la première sélection sont un peu plus « faciles » à lire, est donc tout à fait minime. Et la comparaison des deux premières sélections va dans le sens inverse : les 7 romans qui n'ont pas dépassé le stade de la première sélection ont un ratio de 69,25 % et les 8 de la deuxième sélection sont à 68,70 %, ce qui serait confirmé par les 62 % du choix final. Mais les 4 romans de la troisième sélection ont une moyenne de 69,16 %. Cet indicateur est donc peu fiable et ses conséquences peu claires.

29 Une dernière étude²³ se basera sur ce que l'on appelle des hapax, c'est-à-dire des mots qui n'ont qu'une seule occurrence dans un corpus. Il s'agit donc, par définition, de mots rares, qui seraient en quelque sorte l'inverse de ceux que nous venons de prendre en compte. Mais il ne s'agit pas de mots rares dans l'absolu, ou disons dans la langue. Voici, par exemple, quelques hapax du roman de Le Tellier : *démouler*, *discrétionnaire*, *dépressurisation*, *doctoresse*, *duplication*, *éclectisme*, *envahissement*, *eucaryote*, *formalisation*. Il est le seul à les utiliser dans tout le corpus et il ne les utilise qu'une seule fois. Il ne s'agit pas de termes toujours incompréhensibles (dans le cadre d'un texte scientifique ou technique, par exemple, ils ne seraient sans doute pas des hapax) mais ils peuvent être inhabituels pour de nombreux lecteurs et, dans tous les cas, ne font pas partie du langage romanesque contemporain courant puisqu'un seul auteur sur 47 les emploie. Certains s'expliquent par la thématique du roman, d'autres par les choix esthétiques de l'auteur.

J'ai tout d'abord dénombré le nombre d'hapax absolus dans chaque roman. J'entends par là, comme dans l'exemple précédent, des mots qui, dans tout notre corpus, n'ont qu'une seule occurrence dans un seul roman. Ce nombre a ensuite été rapporté à la taille du roman, sous la forme d'un simple ratio²⁴. Il va de 0,11 % (*Histoires de la nuit* de Laurent Mauvignier, à nouveau) jusqu'à 0,87 % (*Une fille de rêve* d'Éric Laurrent). En examinant les moyennes par groupes de sélection, nous observons exactement les mêmes phénomènes qu'avec les mots les plus courants (ce qui est assez intuitif puisque les deux mesures devraient être à peu près complémentaires : plus un roman utilise de mots courants, moins il devrait comporter de mots rares). Ainsi, les romans sélectionnés contiennent un peu plus d'hapax que ceux qui n'ont pas été sélectionnés (0,47 % contre 0,46 %), les romans des deuxième et troisième sélections nettement plus que ceux de la première sélection (0,49 % contre 0,44 %) et le roman primé est très au-dessus, avec 0,84 % d'hapax.

Ainsi, les trois indices de lisibilité retenus ici ne donnent pas tous les mêmes résultats, le cas le plus extrême étant sans doute celui du roman de Laurent Mauvignier, qui peut être désigné comme le texte le plus difficile à lire par l'indice de Flesch et comme le plus simple par les deux autres indicateurs.

Quelles conclusions ?

Il est temps de conclure ce parcours textométrique, en commençant par reconnaître son échec à déterminer des critères textuels stables expliquant les choix du jury Goncourt. Tous les indicateurs examinés se révèlent inopérants ou contradictoires. On peut en résumer quelques-uns dans le tableau 2 (les + et les – indiquent le sens de la comparaison avec la sélection précédente ; une case vide signale qu'aucune spécificité n'est relevée).

TABLEAU 2. SYNTHÈSE DES CRITÈRES ÉVALUÉS

Sélection	Première	Deuxième	Troisième	Lauréat
Nombre de romans comparés	15 vs 32	8 vs 7	4 vs 4	1 vs 3
Longueur moyenne	–	–	+	+
<i>je</i>	+	+	–	–
<i>tu</i>	–	+	+	–
<i>il</i>	–	+		–
<i>elle</i>	–	–	–	+
<i>nous</i>	+		+	
<i>vous</i>			+	+
<i>on</i>			–	–
Guillemets		+		
Tirets de dialogue	–	+	+	+
Points-virgules	+	+	+	–
Deux-points	+	+	–	+
Points	–	+	–	
Points d'interrogation	–	+	+	+

Points d'exclamation			+	–
Points de suspension	–			
Virgules			+	+
Négations	–			
Comme	–		–	
Langage familier	–			
Noms communs	+	–	–	
Noms propres				+
Prépositions	+	–		–
Déterminants	+	–		–
Pronoms relatifs				–
Conjonctions				–
Adjectifs numéraux	+			+
Adjectifs possessifs				–
Adjectifs qualificatifs				+
Adverbes	–			
Pronoms personnels	–	+	–	–
Interjections				+
Verbes				–
Participes passés	+		–	
Présents	+	+		+
Imparfais	–	–		–
Imparfais du subjonctif		+	+	
Participes présents	–	–		
Conditionnels	–		+	
Futurs		+	+	
Passés simples		+	–	–
Infinitifs			+	
Lisibilité (Flesch)	+	+	–	+
Mots courants	+	–	+	–
Hapax	+	+		+

33 Comme on le voit, il n'y a aucun critère continu, de la première à la dernière sélection. C'est pourtant ce qui était attendu pour expliquer un choix qui se serait basé sur un critère stable, qui aurait mené progressivement à un roman primé qui aurait ainsi présenté au plus haut point les caractéristiques retenues.

34 Mais il serait trop aisé d'attribuer cette impasse à la prise en compte par les jurés de critères extratextuels. Nous avons dit en ouverture que ceux-ci existent, bien entendu, et qu'ils jouent même un rôle important dans la décision, mais ils ne servent au final qu'à différencier des œuvres qui ont été lues, et lues par des professionnels du roman. C'est donc sur cette lecture, fût-elle rapide, que se basent les choix successifs et les débats au sein du jury²⁵. Il y a donc bien des critères littéraires, et donc textuels, qui entrent en jeu.

Examinons par exemple la déclaration du président de l'académie Goncourt, Didier Decoin, qui justifie le prix 2020 en saluant un roman « à l'écriture recherchée, transparente et structurée [...] un livre qui rend heureux, un véritable plaisir de lecture²⁶. » Les quatre critères évoqués ici sont très difficiles à traduire de manière textuelle, et donc statistiquement repérable :

- « écriture recherchée » : cela se traduit-il par un vocabulaire rare ? Nous avons vu que *L'Anomalie* utilisait en effet un grand nombre d'hapax. Mais le terme « écriture » peut aussi renvoyer à d'autres phénomènes : syntaxe complexe, narration élaborée, élégance du style, procédés rhétoriques, etc.
- « écriture [...] transparente » : s'agit-il de la facilité de lecture ? Outre le fait que cette « transparence » s'accorde mal avec une « écriture recherchée » (à moins que la recherche ne portât justement sur l'accessibilité du texte), nous avons vu que les indicateurs de lisibilité sont difficiles à interpréter, et que le lauréat du Goncourt, en outre, était celui qui utilisait le moins de mots faciles et le plus d'hapax. Mais là encore, l'écriture peut désigner le schéma narratif, la psychologie des personnages, etc.
- « écriture [...] structurée » : la structure ici évoquée peut être narrative (enchaînement des séquences, chronologie complexe de la reduplication des biographies, intrigues parallèles) mais pourrait être également typographique (organisation des paragraphes, des chapitres), référentielle (vraisemblance des situations, des psychologies, de la topographie, etc.) Dans chacun de ces cas, quels indices textuels permettraient-ils d'en rendre compte de manière objective ?
- « livre qui rend heureux [...] véritable plaisir de lecture » : voici par excellence le type de réaction de lecteur qui ne peut être référé à aucun phénomène textuel quantifiable. C'est tout d'abord que ce sentiment est subjectif, et par conséquent renvoie au lecteur autant qu'au texte lui-même. On peut aisément imaginer que cette histoire fantastique mettant en cause la réalité de notre monde pourrait tout aussi bien susciter l'effroi ou une profonde détresse chez d'autres lecteurs. La recherche de motifs thématiques (*topic modeling*, *sentiment analysis*) ne serait ici d'aucun secours car les sentiments éprouvés par le lecteur ne sont pas forcément formulés par le texte (il n'y a que 16 occurrences de *heureux*, 11 de *bonheur* et 8 de *plaisir* dans *L'Anomalie*).

On pourrait certes multiplier les études : longueur des chapitres, système des personnages, toponymie, chronologie, organisation narrative, prise en compte de tout l'historique du choix des dernières années, etc. ; multiplier les traitements statistiques : cooccurrences, segments répétés, *deep learning*, etc. Il y a fort à parier que les résultats en seraient tout aussi décevants. Des travaux récents sur les *best-sellers* ont certes cherché à montrer que l'on pouvait détecter sûrement, voire automatiquement, les livres à succès (en particulier, Archer et Jockers 2016 ; Hall 2012). Ce qui est proposé par ces auteurs n'est toutefois pas applicable dans le cas qui nous occupe ici. Tout d'abord, un roman primé par le Goncourt n'est souvent un *best-seller* que parce qu'il a été primé, et n'en avait pas forcément *a priori* les caractéristiques. Le Goncourt est un prix

exigeant du point de vue littéraire, et le jury ne choisit pas, ou pas toujours, un roman qui serait par lui-même devenu un grand succès de librairie. Il a souvent été signalé que *Boussole* de Mathias Énard (Goncourt 2015) s'était mal vendu puisqu'il n'avait pas dépassé les 200 000 exemplaires mais on peut avancer, à l'inverse, que ce livre difficile n'aurait même pas approché un tel chiffre s'il n'avait pas été primé.

37

Mais c'est aussi une question probabiliste : il est plus difficile de prédire ce que sera le choix de dix lecteurs expérimentés et théoriquement autonomes que celui de centaines de milliers de lecteurs amateurs influencés par des mécanismes promotionnels. Il est, symétriquement, plus facile de fabriquer un *best-seller* en appliquant quelques recettes que de fabriquer un Goncourt. Les lauréats et leurs éditeurs sont souvent suspectés d'avoir formaté leurs romans dans le but d'obtenir le prix. Ils s'en défendent toujours et je crois qu'il faut prendre au sérieux un auteur comme Hervé Le Tellier quand il déclare « On ne s'attend jamais à un prix comme le Goncourt. On n'écrit pas pour avoir un prix et on ne peut pas s'imaginer l'avoir » (Botticelli 2020). Notre étude a montré combien il était difficile de mettre en évidence des critères textuels explicatifs des choix du jury. On imagine qu'il est à l'inverse tout aussi difficile de les anticiper pour écrire un roman qui correspondrait parfaitement aux attentes des jurés.

38

La complexité des choix provient aussi de deux facteurs bien connus de la théorie des jeux :

- Les délibérations et les votes étant secrets, le mérite du choix rejaillit sur l'ensemble de l'académie Goncourt, sans qu'aucun de ses membres puisse s'en prévaloir. C'est ce que l'on appelle un « jeu collaboratif », où tous les joueurs retirent le même bénéfice du résultat, ou encourent les mêmes reproches. Ainsi, chaque juré doit tenir compte des autres avis pour que le choix final soit conforme à ce qu'il ou elle estime le meilleur choix, voire le moins mauvais. Les négociations qui en résultent peuvent donc expliquer que certains partis pris soient abandonnés au cours du processus, quand certains jurés acceptent d'abdiquer une exigence au profit d'une autre, qu'ils jugeront plus importante.
- Les délibérations se font en plusieurs étapes. C'est ce que l'on appelle un « jeu répété ». À chaque palier de sélection, chaque juré prend connaissance des choix des autres jurés et reconsidère ses priorités en fonction de ce choix collectif. Cela peut également expliquer des modifications dans les critères retenus.

39

Ainsi, le jury ne poursuit pas, de la première à la dernière sélection, la recherche d'un certain nombre de critères fixés d'avance. Tel juré qui préfère, par exemple, les romans courts, ou les dialogues, devra éventuellement choisir entre ces deux critères ou même les abandonner tous les deux au profit d'un autre critère qui lui sera apparu plus important. Ces mécanismes sont du reste valables autant pour les critères textuels que pour les autres considérations prises en compte au moment du choix.

40

Impalpables et indécis, finalement indécidables, les choix du jury sont donc impossibles à cerner par des moyens statistiques.

Bibliographie

- Archer, Jodie et Matthew L. Jockers. 2016. *The Bestseller Code. Anatomy of the Blockbuster Novel*. New York : Saint Martin's Publishing Group.
- Bernard, Michel et Baptiste Bohet. 2017. *Littérométrie. Outils numériques pour l'analyse des textes littéraires*. Paris : Presses Sorbonne Nouvelle.
- Botticelli, Manon. 2020. « Hervé Le Tellier lauréat du Goncourt 2020 avec "L'anomalie" : le jury récompense un roman exigeant et divertissant ». *franceinfo:culture*. 30 novembre. https://www.francetvinfo.fr/culture/livres/prix-litteraires/le-prix-goncourt-2020-est-attribue-a-herve-le-tellier-pour-lanomalie-gallimard_4201337.html.
- Boura, Olivier. 2003. *Un siècle de Goncourt*. Paris : Arléa.
- Contreras, Isabel. 2020. « Le Goncourt 2020 pour Hervé Le Tellier ». *Livres Hebdo*. 30 novembre. <https://www.livreshebdo.fr/article/le-goncourt-2020-pour-herve-le-tellier>.
- Cossette, André. 1994. *La Richesse lexicale et sa mesure*. Paris : Honoré Champion.
- Gary, Nicolas. 2020. « La vérité sur les fameux 511 romans de la rentrée littéraire 2020 ». *Actualité*, 21 août. <https://actualitte.com/article/6073/reportages/la-verite-sur-les-fameux-511-romans-de-la-rentree-litteraire-2020>.
- Hall, James W. 2012. *Hit Lit. Cracking the Code of the Twentieth Century's Biggest Bestsellers*. New York : Random House Trade Paperbacks.
- Richaudeau, François. 1976. « Faut-il brûler les formules de lisibilité ? » *Communication et langages* 30 (1) : 6-19. <https://doi.org/10.3406/colan.1976.4293>.
- Thomas, Vincy. 2020. « 511 romans pour la rentrée littéraire 2020 ». *Livres Hebdo*, 2 juillet. <https://www.livreshebdo.fr/article/511-romans-pour-la-rentree-litteraire-2020>.
- Zamanian, Mostafa et Pooneh Heydari. 2012. « Readability of Texts : State of the Art ». *Theory and Practice in Language Studies* 2 (1) : 43-53. <https://doi.org/10.4304/tpls.2.1.43-53>.

Notes

- 1 511 d'après *Livres Hebdo* (Thomas 2020) mais au moins le double selon le site *Actualité* (Gary 2020).
- 2 Ont été ôtés des textes : indications éditoriales, listes des œuvres du même auteur, tables des matières, numéros de pages, quatrièmes de couverture. J'ai en revanche conservé les remerciements, épigraphes et notes des auteurs. Je ne peux évidemment rendre public ce corpus, pour des raisons obviées de propriété intellectuelle, mais je puis le communiquer aux chercheurs qui m'en feraient la demande, dans le respect des articles 3 et 4 de la directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019.
- 3 C'est le logiciel TXM (version 0.7.9) qui a été utilisé ici. Je renvoie, pour en savoir plus sur les divers traitements textométriques du texte littéraire, au livre que j'ai cosigné avec Baptiste Bohet (Bernard et Bohet 2017).
- 4 Signalons que les caractéristiques biographiques des auteurs sont assez comparables dans les deux listes : 40 % de femmes sélectionnées contre 34 % non sélectionnées, âge moyen de 53 ans dans la première liste et de 50 dans la seconde.
- 5 On utilise ici le calcul des spécificités, en ne considérant que les exposants supérieurs à 10, c'est-à-dire très significatifs.
- 6 Il faudrait évidemment considérer les cas où une narratrice s'exprime à la première personne du singulier...
- 7 L'AFC porte ici sur les 200 lemmes les plus fréquents.
- 8 À l'inverse, ces deux romans au vocabulaire particulier contribuent à écraser les données de tous les autres et à donner ainsi l'impression qu'ils se ressemblent tous. Mais si l'on enlève ces deux romans pour l'AFC, le graphique obtenu n'est qu'un *zoom* sur le nuage compact des autres romans. On verra, dans les AFC qui suivent, ce phénomène de différenciation augmenter au fur et à mesure que l'on traite moins de romans.
- 9 Outil « Classification » de TXM, avec 2 *clusters*. Le graphique a été pivoté pour faciliter la lecture.

10 Dans les deux cas, on a basé les calculs sur les 200 lemmes les plus fréquents : la comparaison porte donc sur l'utilisation de mots très courants.

11 Spécificité de 17,8.

12 On entend par là une spécificité en dessous du seuil de significativité (fixé à 2 et matérialisé sur le graphique par des lignes de couleur).

13 Saluons au passage l'exploit oulipien d'y avoir utilisé exactement 100 300 mots !

14 Au demeurant, les personnages féminins ne semblent pas être plus nombreux. Sur les dix personnages les plus cités (Victor Miesel, André Vannier, Adrian Miller, David Markle, Lucie Bogaert, Joanna Wasserman, Blake, Adriana Becker, Sophia Kleffman, Patrick Silveria), il n'y a que quatre femmes.

15 Il s'agit en fait du ratio entre les ponctuations de fin de phrase (?!...) et le nombre de mots du roman.

16 Voyez, par exemple, cet article déjà ancien : Richaudeau 1976. On trouvera un état de l'art plus récent dans Zamanian et Heydari 2012.

17 Plus précisément : $206,835 - (1,015 \times \text{longueur moyenne des phrases}) - (84,6 \times \text{nombre moyen de syllabes par mot})$. J'ai utilisé pour le calcul un petit logiciel appelé *Flesh* [sic] et créé par Jack Frink (<http://flesh.sourceforge.net>). Il est à noter que le logiciel est conçu pour l'anglais et que les décomptes des unités du français peuvent être parfois inadéquats. Mais l'important était ici d'appliquer le même algorithme à tous les textes, ce qui permet de comparer les résultats de manière valide. Notons aussi que le logiciel fournit également un autre indice, le *Flesch-Kincaid Grade Level*. Mais le coefficient de corrélation entre les deux indices étant de plus de 0,99, il est ici entièrement redondant, et donc superflu.

18 Ce score extrême s'explique surtout par la longueur des phrases de ce roman : la première phrase, par exemple, ne compte pas moins de 273 mots...

19 Avec des écarts-types comparables : 10,7 et 9,1.

20 Signalons au passage, pour montrer la relativité de ces indicateurs et de notre hypothèse de départ, que le Goncourt des lycéens, *Les Impatientes* de Djaili Amadou Amal, est le roman qui présente le plus faible indice de la sélection (48,2), c'est-à-dire la plus grande difficulté de lecture...

21 <https://eduscol.education.fr/186/liste-de-frequence-lexicale/>. Je n'ai utilisé ici que la liste alphabétique, et non celle qui précise la nature grammaticale des mots. Cela peut générer quelques erreurs (le substantif « pouvoir » comptabilisé comme le verbe « pouvoir ») mais j'ai estimé qu'elles étaient marginales.

22 Il a été choisi ici d'établir ce ratio par rapport au nombre d'occurrences total dans chaque roman. Il aurait aussi été possible de l'établir par rapport au nombre de lemmes différents : la différence entre les deux indicateurs est peu significative puisque le coefficient de corrélation entre les deux séries est de 0,84.

23 Indiquons que l'on n'a pas utilisé ici les techniques qui visent traditionnellement à évaluer la « richesse du vocabulaire ». Ces algorithmes sont très contestables parce qu'ils sont dépendants de la taille de chaque roman. Sur ces questions, on peut encore lire cette référence ancienne mais qui a gardé toute sa pertinence : André Cossette 1994.

24 On pourrait reprocher à ce ratio d'être quand même dépendant de la taille du roman. Un calcul du coefficient de corrélation montre cependant qu'il n'en est rien : il est en effet de - 0,16, ce qui indique une indépendance importante des deux valeurs. À l'opposé, on n'a pas utilisé ici le classique ratio entre le nombre de formes uniques et la taille du roman car celui-ci présente un coefficient de corrélation de 0,84 avec la taille (c'est l'écueil habituel des études de richesse lexicale, signalé plus haut).

25 Ils sont évidemment confidentiels mais quelques indiscretions permettent de s'en faire une idée. Voir, par exemple, Boura 2003.

26 Cité par Contreras (2020).

Auteur

Michel Bernard

UMR 7172 THALIM, université Sorbonne-Nouvelle, Paris, France

Michel Bernard est professeur émérite de littérature française. Spécialiste d'humanités numériques, il a publié plusieurs ouvrages dans le cadre du centre Hubert de Phalèse et, plus récemment, avec Baptiste Bohet, *Littérométrie. Outils numériques pour l'analyse des textes littéraires* (Presses Sorbonne Nouvelle, 2017).

ORCID [0000-0002-9133-2216](https://orcid.org/0000-0002-9133-2216)

michel.bernard@udpn.fr

Droits d'auteur

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la [Licence Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/).