UNIVERSITÉ PARIS, SCIENCES & LETTRES

Jean Barré
Diplômé de licence de Lettres
Diplômé de licence d'Informatique

ENTRE CANON ET ARCHIVE, UNE ÉTUDE DES DYNAMIQUES TEXTUELLES À GRANDE ÉCHELLE

La valeur littéraire au révélateur des méthodes quantitatives

Mémoire de première année du master
« Humanités Numériques »

2021

Résumé

Le canon littéraire est une notion très artificielle, construite avec les biais de la société et modelée par les réceptions successives. L'objectif de ce rapport est de mettre en lumière l'existence de dynamiques textuelles qui assurent à certaines oeuvres une longévité exceptionnelle et menacent au contraire la transmission d'une majorité d'autres. Les méthodes quantitatives nous permettent réaliser une lecture distante sur un corpus de 3000 romans. Nous utilisons des mesures sur la variété lexicale - entropie de Shannon, ratio type-token - pour faire l'exégèse des tendances historiques entre les romans appartenant ou non au canon littéraire.

Mots-clés : canon littéraire ; classique ; littérature ; textométrie ; humanités numériques ; lecture distante ; traitement automatique de la langue

Informations bibliographiques : Jean Barré, *Entre Canon et Archive, étude des dynamiques textuelles : La valeur littéraire au révélateur des méthodes quantitatives*, mémoire de master 1 « Humanités Numériques », dir. [Thierry Poibeau, Jean-Baptiste Camps], Université Paris, Sciences & Lettres, 2021.

Abstract

The literary canon is a very artificial notion, constructed with the bias of society and shaped by successive receptions. The objective of this report is to highlight the existence of textual dynamics that ensure an exceptional longevity to some novels and threaten the transmission of a majority of others. Quantitative methods allow us to perform a distant reading on a corpus of 3000 novels. We use measures of lexical variety - Shannon entropy, type-token ratio - to exegete historical trends between canons and archives.

Keywords : canons ; archives ; literature ; textometry ; digital humanities ; distant reading ; natural langage processing

Bibliographic Information : Jean Barré, *Between Canon and Archive, large scale dynamics in the literary field : Literary value assessed by quantitative methods*, M.A. thesis « Digital Humanities », dir. [Thierry Poibeau, Jean-Baptiste Camps], Université Paris, Sciences & Lettres, 2021.

Table des matières

Introduction

Le mot canon vient du grec ancien Κανων qui signifie « tige de roseau ». En littérature, cette étymologie prend tout son sens et donne au canon sa dimension de robustesse, de stabilité. Le canon est une règle, un modèle, une norme à imiter. Il est le fruit d'une hiérarchisation et désigne un ensemble de textes qui font autorité et auquel le monde littéraire fait référence. Les études littéraires s'intéressent à cet ensemble qui ne représente qu'une partie infime de la production littéraire des siècles passés.

Sélectionner des textes est une vieille tradition dans le monde occidental. Dans le canon biblique, l'Église instaure une liste de textes formant les Saintes Écritures. Le canon avait alors une valeur d'autorité qui permettait à l'Église de contrôler un récit commun pour la postérité.

Ce modèle de hiérarchisation des oeuvres a été importé en littérature au XIX^e^ siècle, à l'époque de la montée des nationalismes, « quand les grands écrivains sont devenus les héros de l'esprit des nations » [1]. Les critiques littéraires et les universitaires ont hérité de cette tradition en construisant des listes d'auteurs et d'oeuvres qui valaient la peine d'être lues ou introduites dans un processus de recherche. Un exemple symptomatique de la naissance du canon est la parution en 1860 des « Grands Écrivains de France » aux éditions Hachette. Dès lors, les éditeurs multiplient les publications des auteurs et des ouvrages susceptibles de leur assurer de larges ventes.

Le canon est ainsi le résultat d'une longue construction, entre des politiques d'éditions obéissant à une logique de rentabilité économique, une transmission assurée par l'enseignement et une monumentalisation par la critique littéraire. Par ailleurs, si un des rôles de la critique est de juger et de comparer, elle ne peut éviter le processus de canonisation - et de décanonisation - des oeuvres littéraires qu'engendre son activité. Derrière ces critères se trouvent des enjeux esthétiques, idéologiques et institutionnels [2]. En effet, ces listes de classiques sont souvent le reflet d'une représentation idéologisée de l'histoire littéraire qui exclut sauf exception les genres et les ethnies. Le canon littéraire permet malgré tout de simplifier une première approche à la littérature. John Guillory soutient que la formation de listes d'auteurs et d'œuvres permet

1. Antoine Compagnon, *Le démon de la théorie : littérature et sens commun*, OCLC : 803876805, Paris, 2007.

2. Christopher Lucken, *Le canon littéraire*, OCLC : 1136466474, 2019.

une distribution du "capital culturel" [3] dans les écoles, et ainsi réguler l'accès à la littérature, aux pratiques de lecture et d'écriture. Ce qui pose problème dans cette hiérarchisation des textes littéraires est le manque de diversité et d'inclusion dans les listes, comme l'ont montré les recherches récentes sur le sujet [4].

Le canon littéraire est donc une notion très artificielle, construite avec les biais de la société et modelée par les réceptions successives. L'objectif de ce rapport est de s'affranchir d'une analyse politique, historique ou économique de la constitution du canon, et de regarder si les oeuvres canoniques comportent des indices textuels qui leurs sont propres. On peut s'inscrire dans plusieurs démarches pour étudier le canon littéraire. Une approche diachronique consisterait à étudier l'évolution du canon au cours des siècles, en prenant en compte le long processus de mise à jour du canon. Nous allons ici étudier le canon en synchronie, autrement dit en prenant notre temporalité comme point d'ancrage. Nous nous focaliserons sur le canon littéraire qui est parvenu jusqu'à nous et prendrons comme paramètre d'analyse la réception et la constitution contemporaine d'un canon littéraire.

Il s'agit de mettre en lumière l'existence des dynamiques textuelles qui assurent à certaines oeuvres une longévité exceptionnelle et menacent au contraire la transmission d'une majorité d'autres. Une première hypothèse naïve pour expliquer ce phénomène serait de dire que les lecteurs privilégient les textes informatifs aux textes redondants. Cela permet aux premiers de rester imprimés et d'être republiés au fil des années et condamne les seconds à l'extinction. Cette simplification des processus à l'oeuvre dans la formation d'un canon nous permet de rentrer dans les textes et nous donne des éléments quantifiables pour essayer de confirmer ou d'infirmer cette idée reçue.

L'objectif est d'observer par le biais d'une lecture distante - chère à Franco Moretti [5] - des tendances historiques qui mettraient au jour des dissimilarités morphosyntaxiques entre deux sous-corpus composés d'une part de romans « canoniques » et d'autre part de romans « non-canoniques ». Les méthodes quantitatives nous permettent de traiter des corpus les plus larges

3. John Guillory, *Cultural capital : the problem of literary canon formation*, Chicago, 1993.

4. José Eduardo González, Elliott Jacobson, Laura García García et Leonardo Brandolini Kujman, « Measuring Canonicity : Graduate Reading Lists in Departments of Hispanic Studies », *Journal of Cultural Analytics* (, mars 2021), DOI : 10.22148/001c.21599.

5. Franco Moretti, *Graphs, maps, trees : abstract models for a literary history*, London ; New York, 2005.

possibles, ce qui ouvre la possibilité d'une histoire empirique de la littérature. Dans ce rapport, nous essaierons de reproduire les résultats de l'équipe de recherche du Stanford Literary Lab. Dans leur pamphlet 11 [6] les chercheurs se sont intéressés à la variété lexicale. La première étape de ce rapport consiste à mesurer la quantité de redondances et d'informations présentes dans les deux sous-ensembles de notre corpus. Après une présentation du corpus d'étude et de la démarche quantitative du rapport, nous analyserons nos résultats, puis nous discuterons des dynamiques historiques mises en lumière par nos résultats.

6. Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti et Hannah Walser, « Canon/Archive. Large-scale Dynamics in the Literary Field »–11 (janv. 2016), p. 14, URL : `https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf` (visité le 01/06/2021).

Chapitre 1

Présentation de la démarche quantitative

1.1 Le corpus de travail

Le corpus du projet *ANR Chapitres*[1] sera notre base pour cette recherche. Ce corpus est composé de 2968 romans assemblés à partir de ressources trouvées sur internet (wikisource, ebook gratuit, corpus pré-existant sous licence-libre). Il est encodé en XML-TEI, car cela répondait au besoin de structuration des textes pour le projet *ANR chapitres* qui analysait l'évolution de la longueur des chapitres au cours du temps. Comme on peut le voir avec la figure 1.1, les textes sont bien répartis dans le temps, avec une dominance de la deuxième partie du XIX^{e} siècle.

Le corpus est particulièrement pertinent pour notre étude puisqu'il a été assemblé avec la contrainte de former un équilibre entre les œuvres appartenant ou non au canon. La répartition est de l'ordre de 39% de romans canoniques et 61% de non-canoniques comme on peut le constater sur la figure 1.2.

1. `https://chapitres.hypotheses.org/`

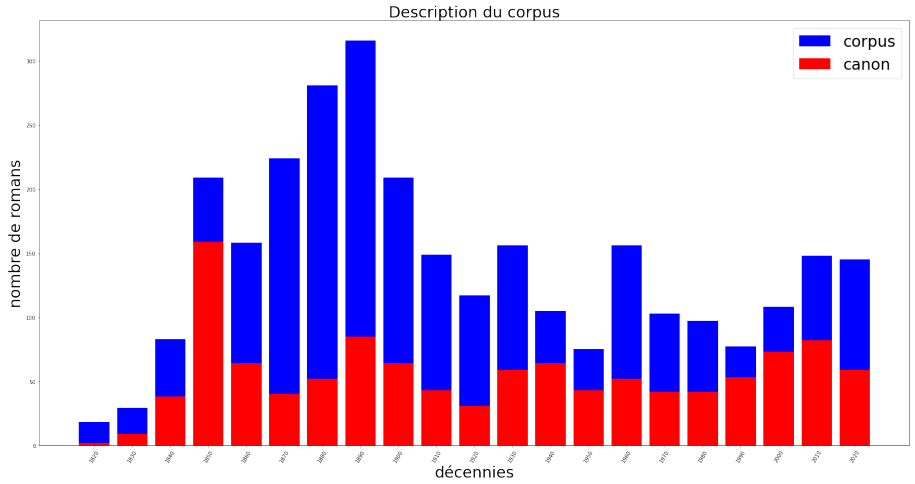


FIGURE 1.1 – Répartition du corpus dans le temps

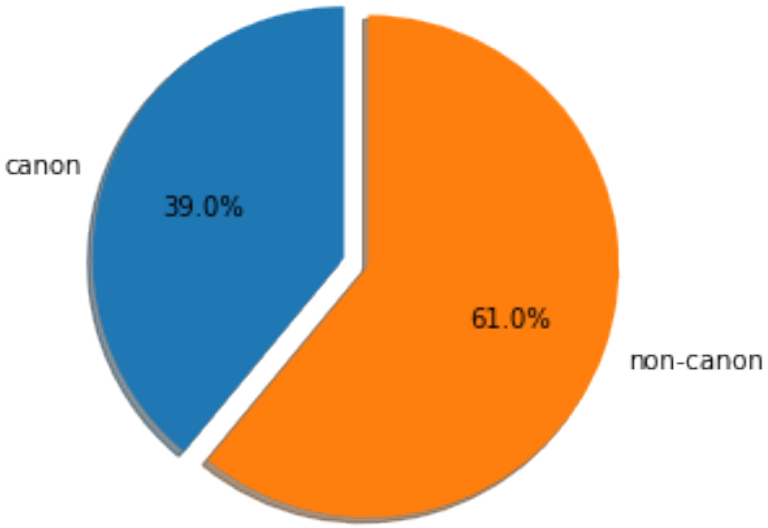


FIGURE 1.2 – Répartition du corpus entre canon et non-canon

Le critère utilisé par l'équipe du projet *ANR Chapitres* pour définir l'appartenance ou non d'un texte donné au canon littéraire est le nombre de résultats d'une requête par auteur dans la revue littéraire en ligne Acta Fabula [2]. Cela s'inscrit dans la perspective du rapport, qui est d'étudier le canon en synchronie, et plus précisément celui qui est parvenu jusqu'à nous.

2. `https://www.fabula.org/index.php`

1.2 Les mesures utilisées

Les chercheurs du Stanford Literary Lab se sont intéressés à une mesure linguistique de la variété lexicale, le ratio type-token. La Longman Grammar of Written and Spoken English définit le ratio type-token ainsi : "La relation entre le nombre de formes de mots différentes, ou types, et le nombre de mots courants, ou tokens, est appelée ratio type-token (ou TTR). En pourcentage, le ratio type-token est égal à (types/tokens) * 100". [3] Plus il y a de types de lemmes différents, plus la redondance des lemmes et le TTR d'un texte sont faibles, et plus sa variété doit être élevée.

La deuxième mesure utilisée par les chercheurs de Stanford est l'indice de Shannon, qui est tiré de la théorie de l'information [4]. Cette dernière est une théorie mathématique de la communication. Cet indice évalue la quantité d'information transmise par les structures linguistiques et l'efficacité de la communication linguistique. De cette manière, Claude Shannon et Warren Weaver cherchaient à mesurer le nombre de bit minimum pour encoder une certaine communication en langage naturel.

Si l'indice de Shannon a été conçu à l'origine pour mesurer l'entropie, il est devenu un indice très populaire dans divers domaines d'études tels que l'écologie, la géographie et l'anthropologie. Son but est de prédire le type ou l'espèce d'un individu choisi au hasard dans un ensemble de données. Mark Algee-Hewitt propose d'utiliser l'indice de Shannon sur les bigrammes de lemmes [5]. L'indice détermine le contenu informationnel de nos textes en évaluant le degré de prévisibilité de chaque transition de mot à mot, compte tenu de l'éventail des transitions possibles. Ainsi, l'indice de Shannon va nous permettre d'évaluer le niveau de diversité des bigrammes dans chaque roman.

$$H\prime = -\sum_{i=1}^{R} pilnpi \tag{1.1}$$

La formule est assez simple : R est le nombre de types (ou d'espèces, en écologie), pi est la proportion d'individus pour chaque type dans un échantillon.

3. Douglas Biber, Susan Conrad et Geoffrey N. Leech, *Longman student grammar of spoken and written English. Hauptbd. ...* 9. impression, OCLC : 838972202, Harlow, 2011.

4. C. E. Shannon, « A Mathematical Theory of Communication », *Bell System Technical Journal*, 27–3 (juil. 1948), p. 379-423, DOI : 10.1002/j.1538-7305.1948.tb01338.x.

5. Mark Algee-Hewitt, *Discourse, Design, Disorder*, URL : http://markalgeehewitt.org/index.php/main-page/projects/discourse-design-disorder/ (visité le 13/06/2021).

L'indice de Shannon est connu pour l'importance qu'il accorde à la richesse des espèces (ici des lemmes). On attend dans nos textes que cette mesure soit plus grande pour les canons que pour les archives.

1.3 Les outils du TAL

Pour quantifier les fréquences d'apparition de mots, on réduit les lexèmes sujets à flexion (les verbes, les substantifs, les adjectifs) à leur unité lexicale commune. On appelle ce processus la lemmatisation. Par exemple, cela permet de compter les différentes formes du verbe être comme autant d'occurrences d'un même lemme, le verbe être. Pour ce traitement, nous utilisons la librairie python Spacy. Cette dernière est très performante pour une analyse sur de grandes quantités de données et couvre tous les traitements d'une chaîne de TAL classique. Cette librairie a aussi l'avantage d'être très bien documentée et comporte plusieurs modèles pour le français. Au vu des performances des différents modèles, nous prenons la décision d'utiliser le modèle fr_core_news_lg qui a un très bon rapport temps d'exécution / performance. Spacy nous permet de tokeniser, lemmatiser et de nettoyer les romans en contrôlant l'étiquetage morphosyntaxique des tokens. Voici la fonction qui va permettre de lemmatiser nos textes avec l'aide de Spacy :

```
def lemmatize(path:str)->list:
    list_lemma = []
    with open(path, encoding="utf8") as file:
        #on récupère l'arbre du document xml
        tree = etree.parse(file)
        #on récupère l'indice de canonicité du roman
        tag = est_canon(tree)
        if tag == True:
            print("canon")
        else:
            print("non_canon")
        #on récupère les paragraphes avec un xpath
        if tree.findall(".//p"):
            for paragraphe in tree.findall(".//p"):
                if paragraphe.text:
```

```
                    #on nettoie le texte et on le met dans la pipeline de spacy
                    clean_text = normalize("NFKD", paragraphe.text)
                    docs = nlp(clean_text)
                    for token in docs:
                        #si le token est bien un mot on récupère son lemme
                        if token.pos_ != "PUNCT" and "SPACE" and "X" and "SYM":
                            list_lemma.append(token.lemma_)
    return list_lemma, tag
```

On récupère d'abord l'indice de canonicité avec un xpath, puis on récupère le texte qui est encodé dans des balises <p> de paragraphes. On le met dans la pipeline de Spacy qui va tokeniser notre texte, c'est à dire le découper en mots. Ensuite nous contrôlons l'étiquetage morphosyntaxique des tokens avec le Part Of Speech tagging (POS) de Spacy. Cela nous permet de nous assurer que le texte est bien propre : les lemmes récupérés ne sont ni des espaces, ni des chiffres, ni des signes de ponctuation. De cette façon, les calculs réalisés sur ces lemmes seront pertinents. On peut ainsi former des listes de lemmes, ou de n-grams qui seront la base de notre analyse.

1.4 Quelques considérations avant l'implémentation

Avant de lancer les calculs sur notre corpus, il est important de définir les paramètres de nos algorithmes pour que les mesures soient les plus pertinentes possibles. La taille des romans du corpus du projet *ANR Chapitres* varie énormément, entre des nouvelles d'une cinquantaine de pages et des romans fleuves de plus de 3000 pages. Nous avons donc découpé nos textes en morceaux de 1000 mots pour normaliser les mesures. Cela nous a permis de ne pas écraser les résultats des fréquences d'apparition de lemmes ou de bigrammes et de pouvoir comparer les textes entre eux. Pour un texte de 30 000 mots, 30 mesures seront réalisées. Ce n'est qu'après que l'on pourra simplifier nos résultats par une simple moyenne entre les 30 mesures prises.

En plus de la fenêtre de 1000 mots, nous avons décidé de mesurer l'indice de Shannon sur les 100 bigrammes les plus fréquents de chaque fenêtre. C'est une valeur assez arbitraire néanmoins elle est aussi utilisée par l'équipe de Stanford dans leur publication.

Après les premières implémentations de l'indice de Shannon sur nos textes, nous avons fait face à un comportement que nous n'avions pas anticipé : La proportion pi d'individus pour les bigrammes de lemmes rares est trop petite (de l'ordre de 10^{-3}). Le logarithme de cette proportion donne un résultat cohérent mais celui-ci multiplié par cette même proportion donne un résultat du même ordre, insignifiant dans le calcul. Pour simplifier, l'indice de Shannon donne dans notre cas plus de poids à un bigramme redondant qu'à un bigramme rare. Or un bigramme rare améliore le niveau de diversité du passage. Cela devrait faire augmenter par la même occasion l'indice de Shannon.

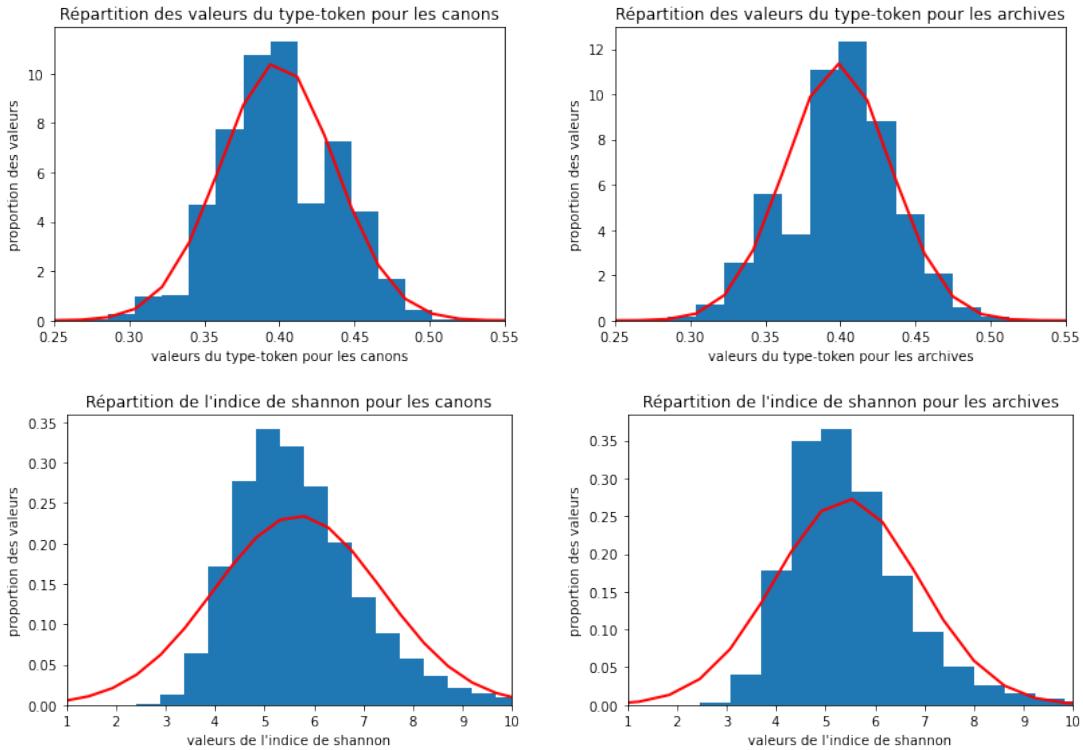
On peut expliquer cela par le fait que nos textes possèdent une population de bigrammes très grande (1001 bigrammes par morceaux de 1000 mots) et des fréquences d'apparition très petite. Néanmoins, il nous semblait pertinent de conserver l'indice de Shannon en prenant en compte qu'il ne mesure pas dans notre corpus le contenu informationnel et la diversité mais bien la redondance de second ordre. En d'autres termes, plus l'indice de Shannon est grand, plus les bigrammes de nos textes sont redondants.

Chapitre 2

Analyse des résultats

2.1 Résultats

Après avoir lancé les deux algorithmes sur les 2968 romans, des premiers résultats sont générés sous la forme de tableurs de 2968 colonnes.



Même si elles ne correspondent pas à une loi normale, nous obtenons une bonne répartition des données pour nos quatre mesures. Ce constat est important car il montre que les valeurs limites portent une signification particulière que nous analyserons dans la suite du rapport.

2.1.1 Type-token ratio

Pour le ratio type-token, un nombre de type de lemme élevé va nécessairement augmenter le résultat, puisque le nombre de token reste le même avec nos fenêtres de 1000 mots. Un type-token élevé implique donc un texte peu redondant et inversement. Nous avons projeté dans la figure 2.1 nos résultats de cette première mesure sur un graphique avec en ordonnées les moyennes des type-token ratio de chaque roman et en abscisses les années de parutions de nos romans.

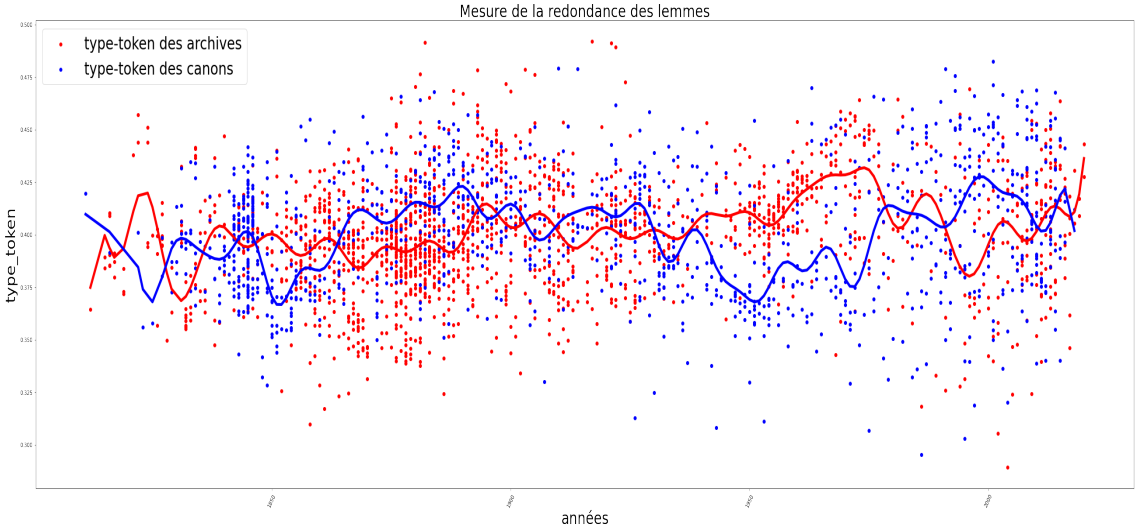


FIGURE 2.1 – Fréquence d'apparition des lemmes

Pour rendre plus lisibles les résultats, nous avons implémenté une régression non-linéaire qui trace la moyenne ajustée aux valeurs de proche en proche. La figure 2.1 représente donc la redondance des lemmes au cours du XIX^{e} et du XX^{e} siècle dans les romans du corpus *ANR Chapitres*. S'il n'y a pas de tendance qui se dégage entre nos deux sous-corpus au XIX^{e} et dans la première moitié du XX^{e} siècle, on remarque que nos données se séparent en deux groupes distincts aux alentours des années 1940 puis se rejoignent dans les années 1980. Ces quarantes années correspondent aussi à un minimum local très marqué pour nos romans canoniques. On peut en effet constater une baisse du type-token ratio et donc une augmentation de la redondance sur cette période. Pour vérifier notre première impression, nous avons effectué avec la figure 2.2 un diagramme en boîte en sélectionnant la période 1940-1980.

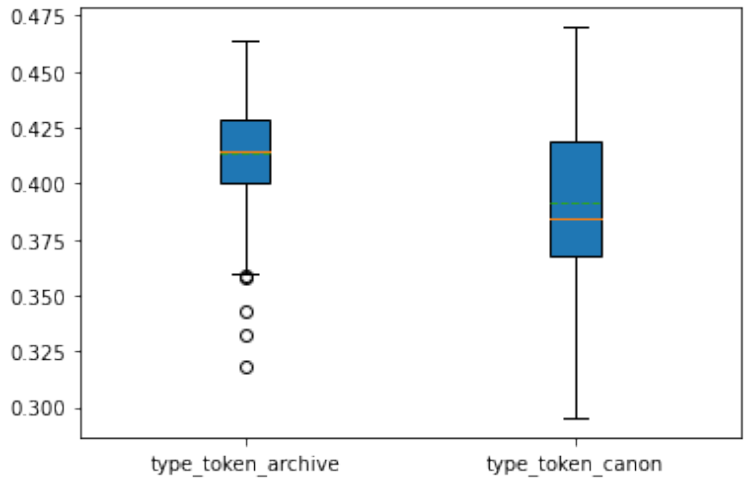


FIGURE 2.2 – diagramme en boîte sur la période discriminante pour le ratio type-token

Sur cette période donnée, nos archives sont d'une part beaucoup plus concentrées que nos canons et d'autre part les résultats sont plus élevés pour nos archives que pour nos canons. Le rectangle bleu des archives, qui correspond à 50% des valeurs (du premier au troisième quartile) se place au dessus de la médiane de celui des canons. Il y a donc une différence notable sur cette période.

2.1.2 Indice de Shannon

Grâce à l'indice de Shannon, l'équipe de Stanford a trouvé des différences significatives entre les deux sous-corpus. Dans notre cas, la différence entre canon et non-canon n'est pas aussi marquée. Comme on le voit sur la figure 2.3, les deux régressions qui tracent la moyenne de proche en proche, pour les canons en bleu et pour les archives en rouge, sont très homogènes. Cependant, nous retrouvons la même période discriminante que la figure 2.1 du type-token ratio. Sur cette période, si les mesures pour les archives restent cohérentes avec le siècle passé, celles pour les canons augmentent rapidement et marquent un plateau d'une cinquantaine d'années.

Pour confirmer cette analyse, nous avons réalisé un diagramme en boîte en figure 2.4 pour les mesures de Shannon sur la période de 1940 à 1990. Il se trouve que 50% des valeurs des romans canoniques dominent le troisième quartile des archives. Nous retrouvons ainsi le même phénomène que pour la

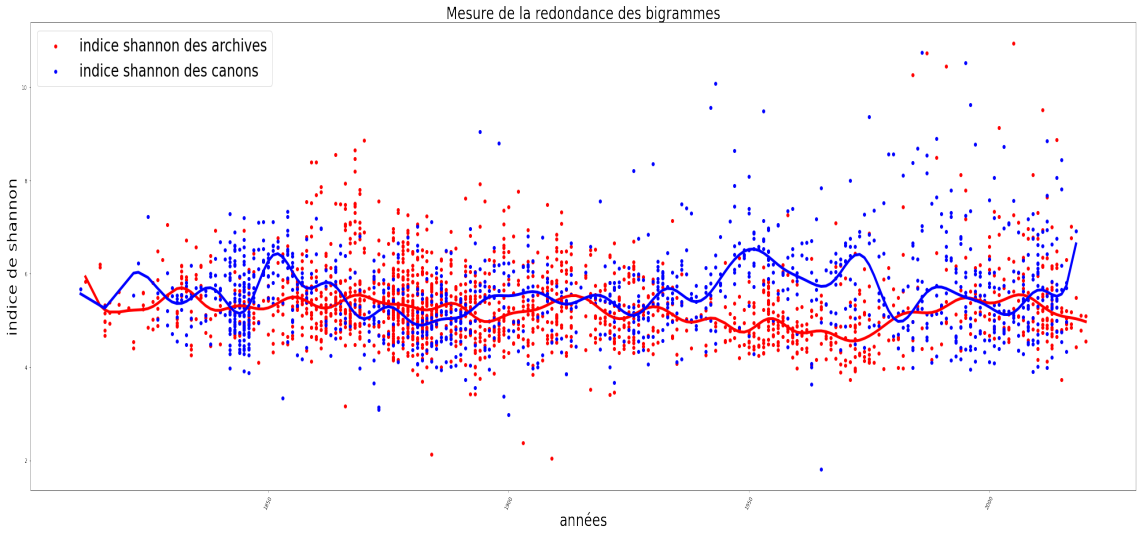


FIGURE 2.3 – Mesure de la redondance des bigrammes

mesure du type-token ratio.

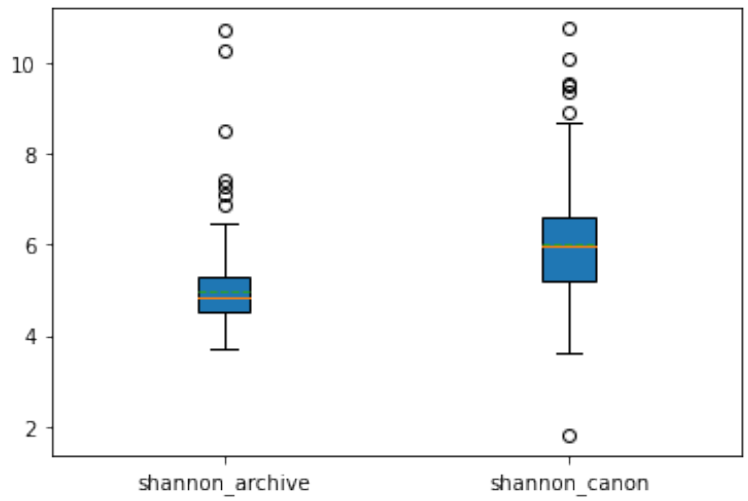


FIGURE 2.4 – diagramme en boîte sur la période discriminante pour l'indice de Shannon

C'est un constat intéressant qui conforte les premiers résultats. Il y a dans la deuxième partie du XX^{e} siècle un niveau de redondance des lemmes et des bigrammes qui sort de l'ordinaire pour nos romans canoniques.

2.2 Analyse des résultats

Les romans classifiés comme canoniques dans le corpus *ANR Chapitres* sont donc plus redondants que les archives sur une période d'environ 50 ans. C'est un résultat auquel on ne s'attendait pas, puisque notre hypothèse de départ penchait vers le contraire. Maintenant que nos résultats montrent ce phénomène, il est légitime de se demander pourquoi le type-token ratio et l'indice de Shannon séparent si bien nos romans ? Et pour quelles raisons sur cette période donnée en particulier ? Il est assez compliqué de répondre à ces questions sans revenir vers une analyse qualitative, mais on peut trouver des éléments de réponses dans la littérature scientifique sur les fréquences de mots dans les romans.

En effet, les travaux de Susan Conrad et de Douglas Biber [1] montrent que les résultats du type-token ratio diffèrent selon les registres utilisés dans le texte. L'opposition fondamentale se situe entre l'oral et l'écrit, avec un rapport type-token plus élevé pour l'écrit que pour l'oral. Cela s'explique en partie par le niveau de langue, qui est plutôt soutenu dans les textes littéraires et familiers ou communs dans les discours rapportés. Plus précisément, Susan Conrad et de Douglas Biber ont travaillé sur de vastes corpus d'anglais parlé et écrit dans quatre registres ; conversation, fiction, nouvelles et prose académique. Ils caractérisent la conversation comme étant co-construite par deux ou plusieurs locuteurs, qui adaptent de manière dynamique leurs expressions à l'échange en cours. Les locuteurs répètent souvent ce qui a été dit, que ce soit pour relier des concepts, focaliser l'attention ou simplement gagner du temps de réflexion. Ainsi, selon cette étude, la conversation serait nettement plus répétitive que les trois autres registres écrits qu'ils examinent.

Mais si l'archive présente donc une plus grande variété lexicale que le canon, ce n'est pas forcément qu'elle tend vers un registre plus "écrit" que le canon. En effet les travaux de Marissa Gemma, Frédéric Glorieux et Jean-Gabriel Ganascia sur les discours familiers [2] montrent que la répétition est aussi une caractéristique du discours familier dans les romans :

Au fur et à mesure que nous avançons dans le 20e siècle, et que la ré-

1. D. Biber et S. Conrad, *Register, Genre, and Style*, 2e éd., 2019, DOI : 10.1017/9781108686136.

2. Marissa Gemma, Frédéric Glorieux et Jean-Gabriel Ganascia, « Operationalizing the Colloquial Style : Repetition in 19th-Century American Fiction », *Digital Scholarship in the Humanities* (, déc. 2015), fqv066, DOI : 10.1093/llc/fqv066.

> pétition de mots et de phrases devient une caractéristique plus importante du discours romanesque, les répétitions deviennent aussi nettement plus familières ; cette tendance est parfaitement illustrée par la façon dont les personnages romanesques en viennent à s'adresser et à se référer les uns aux autres en se tutoyant.

Dans notre graphique du type-token ratio, nous retrouvons cette tendance observée par Marissa Gemma, Frédéric Glorieux et Jean-Gabriel Ganascia. Les romans entre 1940 et 1980 sont très redondants par rapport au siècle précédent. Cela pourrait donc s'expliquer par une utilisation plus prononcée dans nos canons de l'oralité et de la familiarité.

Chapitre 3

Recherches de conjectures

3.1 Les valeurs aberrantes

Il faudrait maintenant repérer dans le corpus les romans les plus caractéristiques de nos résultats, en repérant par exemple les valeurs "aberrantes" du type-token ratio ou de l'indice de Shannon. Une lecture qualitative sur les passages et les romans les plus redondants pourraient infirmer ou confirmer nos hypothèses sur la présence du registre familier et des paroles rapportées pour expliquer ce niveau de redondance de la variété lexicale.

3.1.1 A l'échelle des morceaux de texte

En filtrant nos résultats, nous obtenons les passages limites du corpus. Le passage le moins redondant (celui avec la mesure du type-token ratio la plus grande) se trouve chez Georges Perec, dans un recueil posthume de textes intitulé « *Cantatrix sopranica L* » publié en 1991. Ce dernier compile des pseudo-études scientifiques ou littéraires. Le passage qui nous intéresse obtient un ratio type-token de 61%, ce qui est remarquable puisque la moyenne du corpus est d'environ 40%.

Comme on peut le voir dans la figure 3.1, le passage en question est la bibliographie de l'article scientifique parodique. Le nombre de type très élevé s'explique par la présence de nombreux noms propres différents dans les références. Un autre élément important est qu'il n'y a presque pas de connecteurs logiques ni de mots outils.

Alka-Seltzer, L. Untersuchungen über die tomatostaltische Reflexe beim Walküre. Bayreuth Monatschr. f. exp. Biol. 184 , 34-43, 1815. Attou, J. & Ratathou, F. Laminar configuration of the thalamo-tomatic relay nuclei. Experimental study with Fink-Heimer-Gygax methods. In : The Hyperthalamus , ed. by V. Cointreau and M. Brizard, Cambridge, Oxford U.P., pp. 32-88, 1974. Balalaïka, P. Deafness caused by tomato injury. Observations on half a case. Acta. pathol. marignan. 1, 1-7 , 1515. Beulott, A., Rebeloth, B. & Dizdeudayre, C.D. Brain designing . Chateauneuf-en-Thymerais, Institute of advanced studies (vol. 17), 1974. Bortsch, B. Saccular disturbances produced by whistling (in russian). Fortschr. Hals-Nasen-Ohrenheilk . 3, 412-417, 1955. Carpentier, H. & Fialip, L. Tomato calibres & swallowing. Bull. diet. gastrom. Physiol. 3 , 141-167, 1964. Chachlik, 1. Vocal performance and binoculars. Covent Gard. J. 307 , 1975-1080, 1959-1960. Chou, O. & Lai, A. Tomatic inhibition in the decerebrate baritone. Proc. koning. Akad. Wiss., Amst. 279, 33, 1927a . Chou, O. & Lai, A. Note on the tomatic inhibition in the singing gorilla. Acta laryngol. 8 , 41-42, 1927b . Chou, O. & Lai, A. Further comments on inhibitory responses to tomato splitting in Soloists. Z. f. Haendel Wiss. 17 , 75-80, 1927c. Chou, O. & Lai, A. Faradic responses to tomatic stimulation in the buzzling ouistiti. J. amer. metempsych. Soc. 19 , 100-120, 1928 a . Chou, O. & Lai, A. Charlotte's syndrome is not a withdrawal reflex. A reply to Roux & Combaluzier. Folia pathol. musicol . 7, 13-17 1928 b . Chou, O. & Lai, A. Tomatic excitation and inhibition in awake Counteralts with discrete or massive brain lesions. Acta chirurg. concertgebouw., Amst. 17 , 23-30, 1929 a . Chou, O. & Lai, A. Musicali effetti del tomatino jettatura durante il reprezentazione dell' opere di Verdi. In : Festschrift am Arturo Toscanini , herausgegeb. vom A. Pick, I. Pick, E. Kohl & E. Gramm., München, Thieme & Becker, pp. 145-172, 1929b. Chou, O & Lai, A. Suprasegmental contribution to the yelling reaction. Experiments with stimulation and destruction. Ztschr. f. d. ges. Neur. u. Psychiat. 130 , 631-677, 1930. Colle, E., Etahl, E & Others, S. Leguminase pathways in the brain. A new theory. J. Neurochem. Neurocytol. Enzymol. 1 , 8-345, 1973. Dendritt, A. & Haxon, B. Synaptic contacts in the Lily Pons. Brain Res ., 1975 (in the press) . Donen, S. & Kelly, G.

FIGURE 3.1 – Passage avec le TTR le plus grand du corpus

Le passage le plus redondant (celui avec la mesure du type-token ratio la plus petite) vient du roman *L'Innommable* de Samuel Beckett publié en 1953. C'est un roman très particulier qui s'inscrit dans la lignée des oeuvres de Beckett et de son travail sur l'absurdité de la condition humaine. Il n'y a pas d'intrigue, seulement une suite de digressions autour du néant. Le type-token ratio de ce passage est de 16%, la figure 3.2 en montre un extrait.

pour soi, chacun devant soi, et nous écoutons, tout un peuple, parlant et écoutant, en même temps, ça ex, non, je suis seul, peut-être le premier, ou peut-être le dernier, seul à parler, seul à écouter, seul à être seul, les autres sont partis, ils sont comme partis, ils se sont tus, tus de parler, tus d'écouter, l'un après l'autre, au fur et à mesure des arrivées, un autre viendra, je ne serai plus le dernier, je serai avec les autres, je serai comme parti, dans le silence, ce ne sera pas moi, ce n'est pas moi, je n'y suis pas encore, je vais y aller, je vais essayer d'y aller, pas la peine d'essayer, j'attends mon tour, mon tour d'y aller, mon tour d'y parler, mon tour d'y écouter, mon tour d'y attendre mon tour de partir, d'être comme parti, c'est long, ce sera long, parti où, où va-t-on de là, on doit aller ailleurs, attendre ailleurs, attendre son tour de partir encore, et ainsi de suite, l'un après l'autre, tout un peuple, ou moi tout seul, pas besoin d'autre peuple, ainsi de suite, moi tout seul, et revenir ici, et recommencer, non, continuer, c'est un circuit, un long circuit, je le connais bien, je dois le connaître, ce n'est pas vrai, je ne peux pas bouger, je n'ai pas bougé, je lance la voix, j'entends une voix, il n'y a qu'ici, il n'y a pas deux endroits, il n'y a pas deux prisons, c'est mon parloir, c'est un parloir, je n'y attends rien, je ne sais pas où c'est, je ne sais pas comment c'est, je n'ai pas à m'en occuper, je ne sais pas s'il est grand, ou s'il est petit, ou s'il est fermé, ou s'il est ouvert, c'est ça, réitère, ça fait continuer, ouvert à quoi, il n'y a que lui, ouvert au vide, ouvert au rien, je veux bien, ce sont des mots, ouvert au silence, donnant sur le silence, de plain-pied, pourquoi pas, tout ce temps, au bord du silence, je le savais, sur un rocher, ficelé sur un rocher, au milieu du silence, sa grande houle s'élève vers moi, j'en ruisselle, c'est une image, ce sont des mots, c'est un corps, ce n'est pas moi, je savais que ce ne serait pas moi, je ne suis pas dehors, je suis dedans, dans quelque chose, je suis enfermé, le silence est dehors, dehors, dedans, il n'y a qu'ici, et le silence dehors, que cette voix, et le silence tout autour, pas besoin de murs, si, il faut des murs, il m'en faut, bien épais, il me faut une prison, j'avais raison, pour moi tout seul, je vais y aller, je vais m'y mettre, j'y suis déjà, je vais m'y chercher, j'y suis quelque part, ce ne sera pas moi, ça ne fait rien, je dirai que c'est

FIGURE 3.2 – Passage avec le TTR le plus petit du corpus

Le roman est un flux de pensées ininterrompues. En effet, Samuel Beckett va très loin dans la représentation de la parole intérieure. L'état psychologique

du narrateur est montré sans filtres et sans vouloir correspondre aux conventions littéraires. On assiste au délitement de l'intériorité du narrateur, qui perd contrôle sur lui-même et ses pensées. Le narrateur répète ses mots mais aussi des structures de phrases entières, ce qui renforce la redondance. Le registre familier est ici utilisé, mais il n'explique pas à lui seul la redondance élevée du roman. En effet, le type-token ratio est sensible au traitement de la parole intérieure, d'autant plus quand elle est hors de contrôle et débridée. Le passage est uniquement au discours direct libre. Par ailleurs, on a montré que l'oralité était un facteur fondamental dans la baisse de la variété lexicale. Cela se confirme dans ce passage précis, construit entièrement sur des paroles rapportées.

L'indice de Shannon met en avant tous les passages d'un ouvrage en particulier. Il s'agit de *Je me souviens* de Georges Perec. C'est un recueil de bribes de souvenirs de la vie de l'auteur. L'ouvrage est construit sur une anaphore de « Je me souviens »au début des 480 fragments. On peut en voir un extrait avec la figure 3.3.

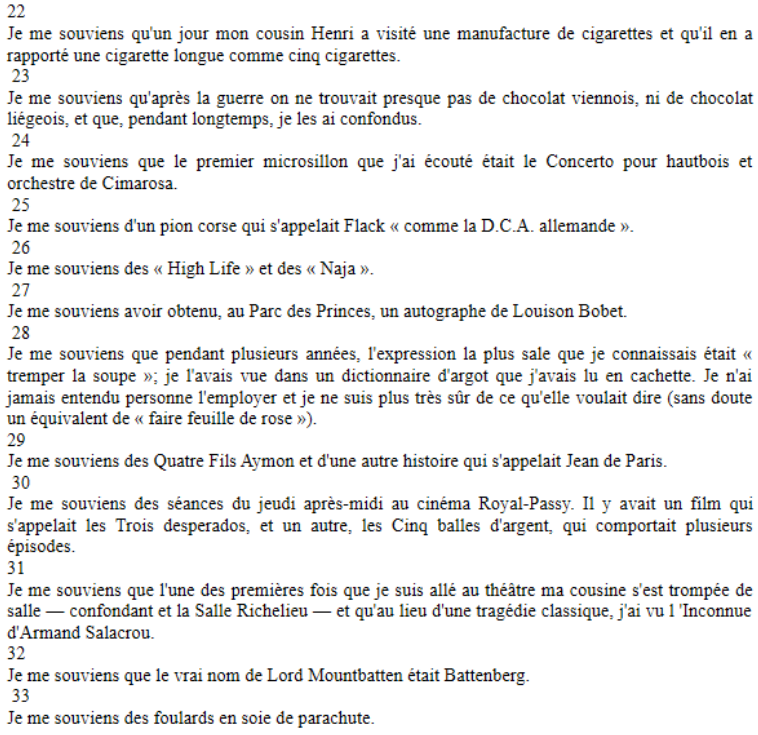

22
Je me souviens qu'un jour mon cousin Henri a visité une manufacture de cigarettes et qu'il en a rapporté une cigarette longue comme cinq cigarettes.
23
Je me souviens qu'après la guerre on ne trouvait presque pas de chocolat viennois, ni de chocolat liégeois, et que, pendant longtemps, je les ai confondus.
24
Je me souviens que le premier microsillon que j'ai écouté était le Concerto pour hautbois et orchestre de Cimarosa.
25
Je me souviens d'un pion corse qui s'appelait Flack « comme la D.C.A. allemande ».
26
Je me souviens des « High Life » et des « Naja ».
27
Je me souviens avoir obtenu, au Parc des Princes, un autographe de Louison Bobet.
28
Je me souviens que pendant plusieurs années, l'expression la plus sale que je connaissais était « tremper la soupe »; je l'avais vue dans un dictionnaire d'argot que j'avais lu en cachette. Je n'ai jamais entendu personne l'employer et je ne suis plus très sûr de ce qu'elle voulait dire (sans doute un équivalent de « faire feuille de rose »).
29
Je me souviens des Quatre Fils Aymon et d'une autre histoire qui s'appelait Jean de Paris.
30
Je me souviens des séances du jeudi après-midi au cinéma Royal-Passy. Il y avait un film qui s'appelait les Trois desperados, et un autre, les Cinq balles d'argent, qui comportait plusieurs épisodes.
31
Je me souviens que l'une des premières fois que je suis allé au théâtre ma cousine s'est trompée de salle — confondant et la Salle Richelieu — et qu'au lieu d'une tragédie classique, j'ai vu l 'Inconnue d'Armand Salacrou.
32
Je me souviens que le vrai nom de Lord Mountbatten était Battenberg.
33
Je me souviens des foulards en soie de parachute.

FIGURE 3.3 – Passage le plus redondant du corpus

Notre calcul de l'entropie souligne à quel point la liaison d'un mot à un autre est prévisible et les résultats pour les passages de « Je me souviens »sont

6 à 7 fois plus importants que pour le reste de notre corpus. Les bigrammes de l'anaphore « Je me »« me souviens »et « souviens que »réalisent à eux seuls des scores importants. Les résultats du type-token ratio sont eux dans la moyenne, ce qui montre bien que c'est la construction des phrases qui est remarquée ici.

3.1.2 A l'échelle des ouvrages

Pour tenter de comprendre à plus grande échelle de quelle nature sont faits les romans redondants, on peut essayer de compiler nos deux mesures pour faire l'exégèse des cas limites Pour cela, on projette sur un graphique l'ensemble des romans en fonction de leurs moyennes de l'indice de Shannon ainsi que du type-token ratio.

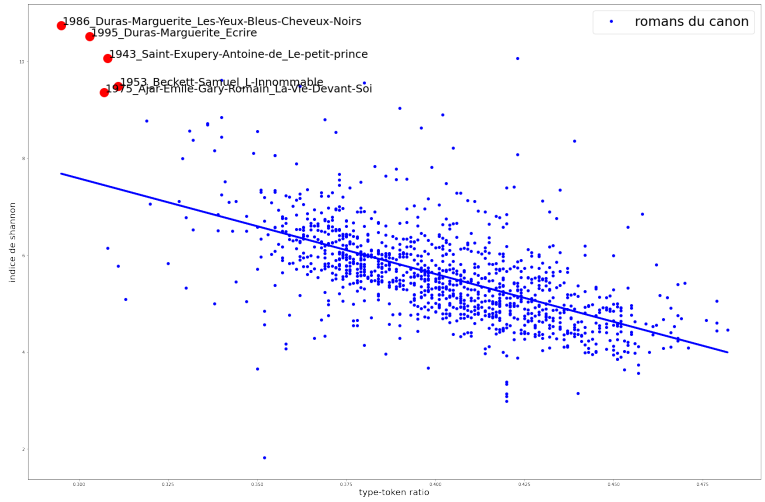


FIGURE 3.4 – Redondance des romans canoniques

La figure 3.4 montre une corrélation entre nos deux mesures, ce qui est logique puisqu'elles calculent chacune à leur manière la redondance dans les textes. Les textes possédant un résultat pour l'indice Shannon très grand ainsi qu'un type-token ratio très faible sont les plus redondants du corpus. La figure 3.4 désigne cinq romans très redondants : il s'agit de « *Le Petit Prince* » d'Antoine de Saint-Exupéry, « *L'Innommable* » de Samuel Beckett, « *Écrire* » et « *Les yeux bleus cheveux noirs* » de Marguerite Duras et enfin « *La Vie Devant Soi* » d'Émile Ajar.

On remarque que les cinq ouvrages appartiennent à la période discriminante (1940-1990) vue en figure 2.1 et 2.3. Ces cinq ouvrages sont donc les plus

symptomatiques de cette période. Pour autant, il n'est pas facile de mettre au jour une unité littéraire tangible entre ces cinq romans. Le premier est une œuvre poétique et philosophique sous l'apparence d'un conte pour enfants. Le deuxième pose des questions existentielles et traite de la contingence de la condition humaine. Le troisième et le quatrième sont deux ouvrages très différents de Marguerite Duras. « *Écrire* » est une compilation des tenants et aboutissants du métier d'écrivain, tandis que « *Les yeux bleus cheveux noirs* » est un roman sur la passion et la fascination : toute l'intrigue se fonde sur un traumatisme intérieur que le récit ne peut que suggérer par ses répétitions et ses silences, ses images obsédantes, ses ruptures énonciatives. Enfin, « *La Vie Devant Soi* » est un portrait sombre d'un enfant immigré orphelin. Nos cinq romans sont donc trop différents pour conclure sur des changements potentiels dans la nature des textes et dans la manière de faire littérature à cette période.

3.2 Conjectures sur cette hausse de redondance

Si l'on met en perspective nos résultats avec le contexte historique et les remous de la théorie littéraire on peut commencer à comprendre les enjeux de cette époque donnée. Tout d'abord, il faut signaler que cette hausse de la redondance dans nos textes s'effectue à la fin des années 1940, aux lendemains de la seconde guerre mondiale. Les remous de cette période sont illustrés par l'affirmation en 1949 de Theodor Adorno « Écrire un poème après Auschwitz est barbare » [1]. Par cette formule cinglante, ce dernier refuse l'amnésie collective d'une « culture ressuscitée » et appelle à la prise de conscience de l'enjeu métaphysique que représente la rupture radicale de la Shoah, qui constitue une négation de l'humanité par elle-même. La période est ainsi marquée par une remise en question de tous les champs culturels.

En littérature, les critiques et les écrivains cherchent à rénover le roman en s'affranchissant des règles de composition dramatique et des objectifs didactiques du roman balzacien [2]. Le renouveau de la critique littéraire française, emmené par les structuralistes Roland Barthes et Gérard Genette, bouleverse

1. Theodor Wiesengrund Adorno, Geneviève Rochlitz et Rainer Rochlitz, *Prismes*, OCLC : 1119500311, Paris, 2003.

2. Denis Labouret, *Littérature française du XXe siècle (1900-2010)*, OCLC : 1191043062, 2013, URL : `http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9782200290184` (visité le 14/06/2021).

la théorie littéraire. Des thèses comme la mort de l'auteur [3] ou le degré zéro de l'écriture [4] remettent en cause les conceptions ancestrales de la littérature. Barthes rend au texte toute sa vivacité et le soustrait de l'emprise de la figure paternelle de l'auteur pour faire place aux interprétations successives des lecteurs. Ce renouveau critique crée un climat propice à la création de nouvelles formes littéraires. C'est dans ce contexte qu'est publié « Pour un nouveau roman » [5], où Alain Robbe-Grillet résume les motivations du monde littéraire de l'époque. Robbe-Grillet décline les « notions périmées » dont il faut faire selon lui table rase : le personnage, l'histoire, la forme et le contenu. Les auteurs de ce courant refusent en effet la caractérisation psychologique des personnages et l'histoire perd toute organisation logique et chronologique. Les auteurs s'en remettent à un lecteur actif, qui comble les blancs du récit et participe à la construction du sens. Se constitue alors un courant littéraire, Le Nouveau Roman, qui inscrit le roman dans une dynamique créatrice de possibilités, repoussant les conventions du roman traditionnel tel qu'il s'était imposé depuis le XIX^{e}siècle. Les auteurs de ce courant très particulier ont tous publié chez le même éditeur, les Éditions de Minuit. Notre corpus dispose d'une partie de ces auteurs : Alain Robbe-Grillet, Claude Simon, Samuel Beckett et Nathalie Sarraute. Nous pouvons comparer leurs oeuvres avec le reste du corpus, pour regarder si ce mouvement littéraire a un impact sur les mesures de variétés lexicales que l'on a mené dans ce rapport. La figure 3.5 met en lumière la place qu'occupe les ouvrages du Nouveau Roman dans la redondance de notre corpus canonique.

Il est intéressant de constater que les résultats de notre analyse sur la redondance lexicale parviennent à capter les remous de cette période. Les ouvrages du Nouveau Roman obtiennent pour la plupart des résultats de redondance assez importants. Ils se placent dans la partie redondante du sous-corpus canonique, même s'ils ne se démarquent pas totalement. On peut conjecturer que l'émancipation des formes plus classiques du roman entraîne une augmentation de l'emploi de niveaux de langue différents, et une nouvelle liberté sur la retranscription des paroles intérieures. L'augmentation de la redondance repérée en figure 2.1 et 2.3 dans nos textes canoniques sur la période 1940-1990

3. Roland Barthes, *Le bruissement de la langue*, OCLC : 919622763, Paris, 2015.

4. Id., *Le Degré zéro de l'écriture*, Paris, 1972 (Points. Littérature, 35).

5. Alain Robbe-Grillet, *Pour un nouveau roman*, OCLC : 840608421, Paris, 2013 (Collection "Double", 88).

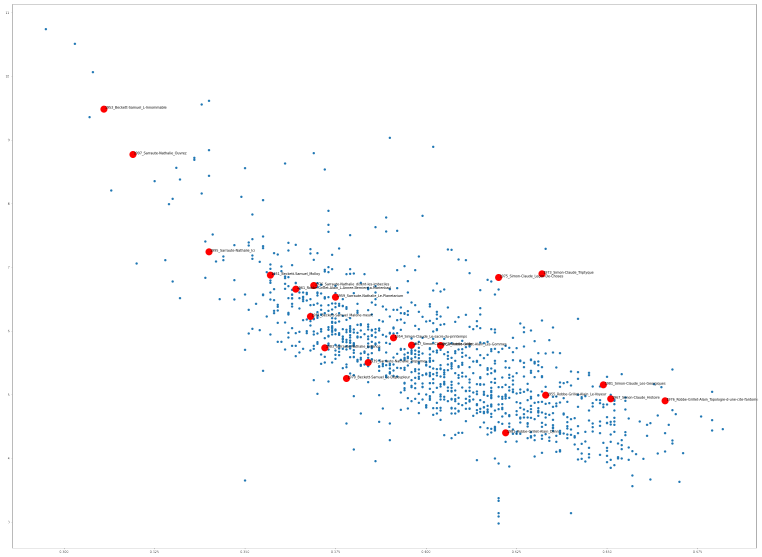


FIGURE 3.5 – Redondance dans le Nouveau Roman

trouverait son explication avec le renouveau littéraire qu'apporte le Nouveau Roman.

Il faudrait étoffer ces premiers résultats avec d'autres analyses. La mesure de la variété lexicale introduit une problématique connue de la critique mais jamais vérifiée empiriquement. Quantifier les différences qu'apporte le Nouveau Roman dans la sphère littéraire n'est pas une chose aisée, et ce travail ouvre simplement une possibilité de recherche plus avancée sur ce sujet.

Conclusion

Pour conclure, l'implémentation de méthodes quantitatives nous a permis d'augmenter la focale pour percevoir des tendances historiques. Notre hypothèse de base était fondée sur un lieu commun de la littérature qui est de dire que les lecteurs privilégient les textes informatifs aux textes redondants. Cela s'est révélé ne pas être pertinent sur notre corpus, et l'hypothèse est probablement fausse vu notre approche empirique. Nos résultats étaient donc homogènes, tant la mesure du ratio type-token que celle de l'indice de Shannon. Néanmoins, une période d'une cinquantaine d'année a retenu toute notre attention. Les ouvrages canoniques se trouvaient être plus redondants que les non canoniques au lendemain de la seconde guerre mondiale. Nous avons donc cherché ce que montraient véritablement nos mesures sur nos morceaux de textes avant d'essayer de trouver des réponses à l'échelle des oeuvres. Si on peut expliquer l'augmentation de la redondance lexicale dans nos romans par une utilisation plus importante du registre familier ou des discours rapportés, ces éléments ne suffisent pas à conclure sur la nature de nos ouvrages. Il faut chercher une ébauche d'explication dans l'histoire littéraire du début de la deuxième partie du XX^{e}siècle. Cette époque est marquée par un renouveau critique qui témoigne de l'éclosion d'un mouvement littéraire à part entière, le Nouveau Roman. Les auteurs de ce courant participent à la remise en cause des formes classiques du roman. Nous avons pu mettre au jour des liens entre nos mesures et ce courant littéraire. Le traitement quantitatif d'un vaste corpus nous a permis de montrer que la variété lexicale était un premier témoin de cette tendance. Des implémentations plus avancées pourrait confirmer ces premiers résultats, comme par exemple l'extraction de motifs et de constructions morphosyntaxiques avec des ngrams de parties du discours serait une approche possible pour la suite des recherches.

Table des figures

Bibliographie

Adorno (Theodor Wiesengrund), Rochlitz (Geneviève) et Rochlitz (Rainer), *Prismes*, OCLC : 1119500311, Paris, 2003.

Algee-Hewitt (Mark), *Discourse, Design, Disorder*, URL : `http://markalgeehewitt.org/index.php/main-page/projects/discourse-design-disorder/` (visité le 13/06/2021).

Algee-Hewitt (Mark), Fredner (Erik) et Walser (Hannah), « The Novel as Data », dans *The Cambridge Companion to the Novel*, 1re éd., 2018, p. 189-216, DOI : `10.1017/9781316659694.013`.

Barthes (Roland), « Introduction à l'analyse structurale des récits », *Communications*, 8–1 (1966), p. 1-27, DOI : `10.3406/comm.1966.1113`.

— *Le Degré zéro de l'écriture*, Paris, 1972 (Points. Littérature, 35).

— *Le bruissement de la langue*, OCLC : 919622763, Paris, 2015.

Anne-Catherine Baudoin et Marion Lata (éd.), *Sacré canon : autorité et marginalité en littérature*, OCLC : on1031847721, Paris, 2017 (Actes de la recherche à l'ENS, 22).

Biber (Douglas) et Conrad (Susan), *Register, Genre, and Style*, 2e éd., 2019, DOI : `10.1017/9781108686136`.

Biber (Douglas), Conrad (Susan) et Leech (Geoffrey N.), *Longman student grammar of spoken and written English. Hauptbd. ...* 9. impression, OCLC : 838972202, Harlow, 2011.

Boukhaled (Mohamed Amine), « On Computational Stylistics : mining Literary Texts for the Extraction of Characterizing Stylistic Patterns » (), p. 163.

Claire Colin, *et al.* (éd.), *Pratiques et poétiques du chapitre du XIXe au XXIe siècle*, OCLC : 985886156, Rennes, 2017 (Interférences).

Compagnon (Antoine), *Le démon de la théorie : littérature et sens commun*, OCLC : 803876805, Paris, 2007.

Erlin (Matt), Piper (Andrew), Knox (Douglas), Pentecost (Stephen), Drouillard (Michaela), Powell (Brian) et Townson (Cienna), « Cultural Capitals : Modeling 'Minor' European Literature », *Journal of Cultural Analytics* (, févr. 2021), doi : 10.22148/001c.21182.

Franco Moretti, *Conjectures on World Literature, NLR 1, January–February 2000*, en, url : https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature (visité le 17/02/2021).

Gemma (Marissa), Glorieux (Frédéric) et Ganascia (Jean-Gabriel), « Operationalizing the Colloquial Style : Repetition in 19th-Century American Fiction », *Digital Scholarship in the Humanities* (, déc. 2015), fqv066, doi : 10.1093/llc/fqv066.

Glorieux (Frédéric), *Auteurs canoniques, graphes et conjonctions (and, et, und, y, e)*, fr-FR, Billet, url : https://resultats.hypotheses.org/388 (visité le 01/02/2021).

González (José Eduardo), Jacobson (Elliott), García (Laura García) et Kujman (Leonardo Brandolini), « Measuring Canonicity : Graduate Reading Lists in Departments of Hispanic Studies », *Journal of Cultural Analytics* (, mars 2021), doi : 10.22148/001c.21599.

Guillory (John), *Cultural capital : the problem of literary canon formation*, Chicago, 1993.

Labouret (Denis), *Littérature française du XXe siècle (1900-2010)*, OCLC : 1191043062, 2013, url : http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9782200290184 (visité le 14/06/2021).

Legallois (Dominique), Charnois (Thierry) et Poibeau (Thierry), « Repérer les clichés dans les romans sentimentaux grâce à la méthode des « motifs » », *Lidil. Revue de linguistique et de didactique des langues*–53 (mai 2016), ISBN : 9782843103261 Number : 53 Publisher : ELLUG, p. 95-117, doi : 10.4000/lidil.3950.

Lucken (Christopher), *Le canon littéraire*, OCLC : 1136466474, 2019.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti et Hannah Walser, « Canon/Archive. Large-scale Dynamics in the Literary Field »–11 (janv. 2016), p. 14, url : https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (visité le 01/06/2021).

Molinié (Georges), « Style et littérarité », *Littératures classiques*, 28–1 (1996), Publisher : Persée - Portail des revues scientifiques en SHS, p. 69-74, doi : 10.3406/licla.1996.2519.

MORETTI (Franco), *Atlas of the European novel : 1800-1900*, Country : GB ill., jaquette ill. en coul. 23 cm. Notes bibliogr. Index., London New York, 1998.

— *Graphs, maps, trees : abstract models for a literary history*, London ; New York, 2005.

— *Distant reading*, Country : GB Contient des textes déjà publiés. Includes bibliographical references and index. Vendor-supplied metadata., London, 2013.

— « « L'opérationnalisation » ou, du rôle de la mesure dans la théorie littéraire moderne », *Critique*, n° 819-820–8 (sept. 2015), Publisher : Éditions de Minuit, p. 712-734, URL : https://www.cairn.info/revue-critique-2015-8-page-712.htm (visité le 20/01/2021).

RIGUET (Marine) et BOUKHALED (Mohamed Amine), « La correspondance de motifs, un outil pour l'analyse du discours ? », *Humanités numériques*–1 (janv. 2020), DOI : 10.4000/revuehn.312.

ROBBE-GRILLET (Alain), *Pour un nouveau roman*, OCLC : 840608421, Paris, 2013 (Collection "Double", 88).

SHANNON (C. E.), « A Mathematical Theory of Communication », *Bell System Technical Journal*, 27–3 (juil. 1948), p. 379-423, DOI : 10.1002/j.1538-7305.1948.tb01338.x.

TED UNDERWOOD et JORDAN SELLERS, » *The Emergence of Literary Diction Journal of Digital Humanities*, en-US, URL : http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/ (visité le 02/06/2021).