

UNIVERSITÉ PARIS, SCIENCES & LETTRES

Léo Dray

Licencié en Histoire et en Sciences Politiques

**Représentations et projections dans la littérature
d'anticipation :
Une approche textométrique pour explorer l'histoire
des futurs envisagés**

Mémoire de seconde année de master
« Humanités numériques et computationnelles »

2021

Résumé

Dans ce mémoire, nous cherchons à démontrer l'apport herméneutique de la science-fiction en tant qu'objet de recherche. Nous montrons notamment son intérêt pour étudier l'histoire des représentations et projections associées au futur. Nous commençons par définir clairement la littérature de science-fiction, afin de construire un objet de recherche adapté à notre étude. Puis nous présentons la notion de futurs envisagés : nous montrons ici la pertinence d'utiliser la littérature de science-fiction « anticipatrice » comme support d'investigation pour appréhender l'histoire des futurs envisagés. Nous proposons ensuite d'avoir recours à des algorithmes informatiques de topic modeling pour construire une méthode d'analyse. Nous poursuivons en appliquant cette méthode sur un vaste corpus de littérature d'anticipation. Enfin, nous analysons les résultats, et esquissons une réflexion épistémologique sur notre démarche. Cette réflexion doit permettre de parfaire notre méthode d'analyse, ainsi que d'explorer les perspectives de recherches qui s'offrent à nous.

Mots-clés : Humanités numériques ; Histoire ; Littérature ; Textométrie ; Sémanistique

Informations bibliographiques : Léo Dray, *Représentations et projections dans la littérature d'anticipation : Une approche numérique pour explorer l'histoire des futurs envisagés*, mémoire de master 2 « Humanités numériques et computationnelles », Université Paris, Sciences & Lettres, 2021.

Abstract

In this thesis, I am trying to demonstrate the hermeneutic contribution of science fiction as a research topic. Notably, I show its interest to study the history of futuristic representations and projections. Firstly, I define clearly the chosen science fiction literature, in order to build up an appropriate research topic. Then, I present the notion of expected futures as an appropriate one to study representations and projections among SF literature. I also suggest using digital humanities methods, particularly topic modeling, to build up an analysis conduct. Finally, I apply topic modeling to a large corpus of anticipatory literature. I analyse the results thus obtained, and sketch out an epistemological thought on my conduct. This thought offers a solid basis to improve our analysis method, and explore new research perspectives.

Keywords : Digital humanities ; History ; Literacy ; Textometrics ; Semantics

Bibliographic information : Léo Dray, *Representations and projections in anticipatory literature : a digital approach to explore envisioned futures*, M.A.2 thesis « Digital and computational humanities », Université Paris, Sciences & Lettres, 2021.

Introduction

« *L'Histoire est un roman qui a été ; le roman est de l'histoire qui aurait pu être* », écrivaient d'une même plume Jules et Edmond de Goncourt, relevant toute la socialité du texte littéraire et de ses ressorts narratifs. La littérature résulte, en effet, d'une formalisation problématique de l'imaginaire social, et entretient de fait une relation privilégiée autant qu'ambigüe avec les imaginaires collectifs : « *Un roman est en interaction dynamique majeure avec la narrativité ambiante, c'est-à-dire avec les façons de raconter l'histoire de la société et les événements qu'elle vit* »¹. Ainsi, le roman, en ce qu'il est ‘une histoire qui aurait pu être’, est profondément ancré dans la réalité qui constitue son contexte de production, et intéresse particulièrement l’histoire des imaginaires sociaux. Mais qu’en est-il, plus particulièrement, de la littérature qui raconte des histoires qui pourraient advenir ? De manière plus précise, nous avons souhaité nous intéresser à la littérature d’anticipation, qui s’attache à décrire des futurs possibles, des mondes envisageables, en ce qu’elle constitue un matériau intéressant pour accéder aux imaginaires projectionnistes. En décalant ainsi notre perspective, on s’intéresse non plus à l’histoire des imaginaires sociaux, mais à celle des projections futuristes, des différents scénarios envisagés pour l’avenir. Ce sont ces imaginaires projectionnistes que nous proposerons d’appeler ‘futurs envisagés’ tout au long de ce mémoire.

Commençons donc par définir clairement cette notion, puisqu’elle est au cœur de notre démarche. Les futurs envisagés désignent explicitement les scénarios considérés lorsqu’on essaye de se projeter dans l’avenir, et les multiples possibles et imaginaires qu’ils évoquent. Deux notions relativement proches mais bien distinctes de ce que nous entendons par l’expression “futurs envisagés” doivent ici nous permettre de la clarifier.

D’abord, les futurs envisagés se distinguent de la prospective, cette dernière consistant en la recherche du, ou des scénario(s) le(s) plus plausible(s), d’un point de vue pragmatique. À cet égard, de nombreux travaux intéressants ont été publiés sur l’histoire de la prospective². Or, ces travaux s’attachent davantage à décrire la manière dont les hommes ont interrogé ces scénarios pour éclairer la prise de décision, qu’à sonder les ima-

1. POPOVIC, Pierre. « La sociocritique. Définition, histoire, concepts, voies d’avenir », *Pratiques*, 151-152 | 2011, 7-38.

2. GAUDIN, Thierry. *La prospective*. 2005, Presses Universitaires de France, « Que sais-je ? ».

ginaires qui entourent ces scénarios. La prospective, dans une pure logique de l'action, consiste à objectiver ces scénarios, à en dégager des éléments de prédiction. Au contraire, lorsque nous étudions les futurs envisagés, nous souhaitons sonder les imaginaires auxquels ils renvoient. En ce sens, il n'est pas exclu de prendre en compte les scénarios les plus improbables, tant qu'ils restent concevables. C'est cette différence qui est fondamentale, et qui explique l'intérêt de cet objet pour notre étude : la science-fiction est prospective, mais ne cherche pas nécessairement à être prédictive. La question du probable concerne la discussion scientifique, tandis que le possible, nettement plus ouvert, est un espace d'expression privilégié pour l'imagination.

Ensuite, les futurs envisagés se distinguent de ce qu'on peut appeler le champ des possibles. Cette notion, inventée par Arlette Farge, et développée par les tenants d'une histoire contrefactuelle³, vise à résituer les événements historiques dans un champ des possibles plus large, c'est-à-dire à envisager les différentes possibilités qui s'offraient à l'Histoire, à un moment précis. Ainsi, étudier le champ des possibles, c'est avant tout dégager l'événement historique d'une explication linéaire, parfois téléologique ; c'est, en somme, rendre compte de l'existence d'autres chemins qu'aurait pu emprunter l'histoire. S'intéresser au champ des possibles, c'est donc questionner les possibles historiques, les futurs alternatifs plausibles qui auraient pu suivre un événement historique donné. Lorsque nous évoquons les futurs envisagés, la démarche est toute autre : nous cherchons à appréhender les imaginaires auxquels renvoient ces tentatives de projection, peu importe qu'elles soient plausibles ou non. En ce sens, il n'est pas exclu de prendre en compte les scénarios les plus improbables, tant qu'ils restent concevables. La littérature d'anticipation est prospective, mais ne cherche pas nécessairement à être prédictive. La question du probable concerne la discussion scientifique, tandis que le possible, nettement plus ouvert, est un espace d'expression privilégié pour l'imagination.

Mais comment appréhender l'histoire de ces futurs envisagés ? C'est cette question qui marque le point de départ de ces travaux de recherche. D'abord, comment caractériser et comprendre ces futurs envisagés ? Ensuite, comment étudier leur évolution historique ? Quel matériau utiliser, et quels méthodes employer ? En proposant un support d'investigation, la littérature d'anticipation, et en développant une méthode computationnelle pour explorer thématiquement cet objet, ce mémoire participe à la construction d'un procédé d'investigation devant permettre d'étudier les imaginaires projectionnistes, et leur évolution historique.

A quels éléments un auteur se réfère-t-il lorsqu'il imagine un futur possible ? Quelles thématiques récurrentes observe-t-on lorsque les auteurs d'une même époque décrivent le futur ? Que nous disent ces thématiques de cette époque, et en particulier des imaginaires

3. DELUERMOZ, Quentin, SINGARAVELOU, Pierre. *Pour une histoire des possibles. Analyses contrefactuelles et futurs non advenus*, Paris, Éd. Le Seuil, coll. L'Univers historique, 2016.

associés à l'avenir ? Lorsqu'on analyse les situations décrites par le narrateur, et *a fortiori*, les mots choisis pour décrire cette situation, on sonde les références communes, les fantasmes et angoisses qui structurent sa manière de penser l'avenir. La littérature d'anticipation apparaît, dès lors, comme un objet littéraire de choix pour étudier la question des futurs envisagés. Non seulement car les œuvres d'anticipation sont profondément ancrées dans la réalité des auteurs qui les composent, mais aussi et surtout, en ce qu'elles permettent d'explorer les imaginaires collectifs auxquels se réfèrent ces mêmes auteurs. Ainsi, lorsqu'un auteur écrit, il rend compte de craintes collectives, d'espoirs communs, d'interrogations propres à son époque. Réciproquement, la littérature d'anticipation contribue à nourrir ces imaginaires, à alimenter ces craintes, ces fantasmes. Quelles angoisses, quels fantasmes caractérisent cette époque ? Certaines technologies futures inspirent-elles la crainte ? Au contraire, certaines avancées techniques nourrissent-elles de grands espoirs ? En ce qu'elle est à la fois un réservoir d'expériences de pensée et des formes de problématisations, la littérature d'anticipation offre un matériau pertinent pour tenter de répondre à de telles questions, démontrant tout son potentiel heuristique⁴.

Nous allons maintenant réfléchir aux méthodes par lesquelles nous analyserons ces futurs envisagés. Commençons par évoquer le caractère historique de cette étude. Il s'agit bien ici de faire de l'histoire en utilisant la science-fiction comme support d'investigation, et en ayant recours à des méthodes numériques, comme nous l'exposerons par la suite. À travers la littérature de science-fiction que nous avons définie, nous cherchons donc à appréhender l'évolution des futurs envisagés sur un plan historique. On cherchera vraiment à adopter un regard d'historien sur la question, en mettant ces futurs envisagés et les imaginaires auxquels ils renvoient en perspective avec le contexte de leur production. De plus, notre étude pourra interroger, outre la manière dont ces imaginaires évoluent dans le temps, la manière dont ils évoluent dans l'espace, en comparant de la littérature issue de différentes régions du monde, et empreinte de problématiques propres à l'espace culturel au sein duquel elles ont été produites.

Puisque nous étudions uniquement des œuvres de littérature, nous conviendrons que les représentations associées aux futurs envisagés y sont décrites essentiellement par des mots. Le choix des mots est donc d'une importance considérable : c'est par leur intermédiaire que le lecteur accède à des images, qu'il associe à des imaginaires. C'est donc également par leur intermédiaire que nous proposons d'interroger les futurs envisagés. Ainsi, nous aurons recours à des méthodes d'analyse sémantique. Ce que l'on cherche à faire, c'est donc analyser les mots choisis par les auteurs, en fonction des époques, pour décrire ces futurs envisagés, afin de comprendre leur sens et leur évolution. Ici, c'est au moyen de l'analyse thématique que nous souhaitons explorer ces mots. Sur les

4. RUMPALA, Yannick. « Littérature à potentiel heuristique pour temps incertains », in *Methodos*, 15 | 2015 URL : <http://journals.openedition.org/methodos/4178>.

méthodes d'analyse thématique et leur intérêt, on pourra se référer à l'article de Pierre Zweigenbaum et Benoît Habert, *Accès mesurés aux sens*⁵. Analyser les mots au moyen de méthodes thématiques, c'est essayer de déterminer des thèmes récurrents, et des mots qui reviennent fréquemment pour explorer ces thématiques. On s'attend ainsi à observer des différences thématiques en fonction des époques de production des œuvres, et surtout, au sein même de ces thématiques, des ensembles de mots très variés, rendant compte des différents futurs envisagés.

Il nous faut donc trouver un moyen de traiter un volume très important de données textuelles, correspondant aux mots des œuvres étudiées, de manière systématique. On comprend, en effet, que l'étude a beaucoup plus d'intérêt si elle s'appuie sur un grand nombre de textes. Plus la pluralité des textes et des auteurs est conséquente, plus les résultats seront significatifs. Plus on compte de textes différents pour une même époque, plus on sera en mesure d'apprécier finement les futurs envisagés à cette époque, les imaginaires desquels ils procèdent. Justement, l'utilisation d'une méthode computationnelle permet ici d'envisager un traitement massif des données ; si l'on trouve une méthode efficace et probante, on n'aura donc aucun mal à la rendre systématique et à l'appliquer à un large corpus. De surcroît, nous avons accès à des données gratuites de manière quasi illimitée : lorsqu'ils sont assez vieux, ces ouvrages sont tombés dans le domaine public et le texte est facilement accessible en ligne ; lorsque ce n'est pas le cas, il suffit de trouver une version numérisée du texte en question.

Nous proposons donc d'avoir recours à une méthode relevant du champ des humanités numériques, nous permettant de systématiser notre analyse grâce à un procédé computationnel, et à terme, de traiter un très grand nombre de données. On l'a dit, dans le cadre de notre étude sémantique, la capacité à traiter un grand nombre de données fait vraiment tout l'intérêt de la méthode, d'autant plus qu'on dispose d'un grand nombre de données accessibles relativement facilement. Or, c'est précisément l'intérêt des méthodes informatiques que nous proposons. D'abord, son faible coût et sa réemployabilité nous permettent de prétendre à étendre l'analyse à un très grand nombre d'œuvres. Là où l'analyse littéraire demande pour chaque œuvre un travail fastidieux, les méthodes numériques, une fois bien établies et fonctionnelles, sont facilement réutilisables sur d'autres textes. La capacité d'analyse ne dépend alors que de la disponibilité des œuvres et du temps de traitement des données.

Ensuite, les méthodes informatiques permettent une analyse rigoureusement systématique. Là, il faut évoquer la notion de *distant reading*, développée par Franco Moretti en 2000, dans un article qui a fait date⁶. Dans ce dernier, le professeur propose d'em-

5. ZWEIGENBAUM, Pierre, HABERT, Benoît. « Accès mesurés aux sens », *Mots. Les langages du politique*, 74 | 2004.

6. Pour l'article en question, voir MORETTI, Franco. « Conjectures on World Literature », *New*

ployer des méthodes informatiques et statistiques pour analyser la littérature mondiale, et définit l'idée de *distant reading*, par opposition au *close reading* que constitue l'analyse littéraire par le corps du texte. Initialement conçu comme une méthode complémentaire pour traiter un volume important de littérature hors du corpus de l'auteur, le *distant reading* a progressivement développé des outils puissants, permettant de traiter *per se* un corpus d'oeuvres destiné à l'analyse. Le *distant reading* apparaît donc aujourd'hui comme une approche à part entière des études littéraires, se définissant par l'application de méthodes computationnelles à des données littéraires. De manière intéressante, cette lecture à distance permet aussi de s'affranchir de la subjectivité du chercheur et des erreurs d'attention éventuelles auquel s'expose nécessairement le *close reading*. C'est là un apport majeur de ces méthodes informatiques, qui permettent de systématiser l'analyse, c'est-à-dire de traiter toutes les données uniformément et infailliblement. Bien entendu, l'interprétation des résultats obtenus via ces méthodes relève entièrement du chercheur, et ce travail d'intelligibilité doit rester un espace de subjectivité ; mais le recours à une approche informatique permet d'objectiver l'analyse des données textuelles.

L'analyse statistique de données textuelles, qu'on appelle parfois textométrie, regroupe un ensemble de méthodes aux usages très variés. On distingera, pour ce qui est du champ des sciences sociales et des humanités, deux principaux types d'applications. Le premier, tourné vers l'analyse de la forme linguistique et du style, s'illustre lors de la datation ou de l'attribution de textes : à ce propos, l'étude stylométrique menée en 2013 par Mike Kestemont, Sara Moens et Jeroen Deploige sur l'œuvre d'Hildegarde de Bingen démontre tout l'intérêt de ces méthodes⁷. Le deuxième s'intéresse davantage au contenu lexical et thématique des textes, et correspond mieux à l'usage que nous souhaitons faire de ces méthodes. Nous pensons ici plus précisément aux méthodes exposées par les champs de la lexicométrie, qui propose une étude quantitative du lexique, et de la littérométrie, qui fournit des outils numériques plus spécifiquement adaptés à l'analyse des textes littéraires⁸. Parmi les nombreuses approches existantes, nous retiendrons principalement celles du *topic modeling*, un ensemble de méthodes de modélisation statistiques permettant la recherche de thèmes récurrents dans un corpus de documents. Fondant nos analyses sur ces modèles, nous reviendrons plus longuement sur cette notion.

Pour présenter nos travaux, nous commencerons donc par expliquer plus en profondeur le choix de la littérature d'anticipation comme support d'investigation, en partant

Left Review, 2000. On note aussi que cette publication, ainsi que d'autres articles dans lequel l'auteur poursuit sa réflexion, ont été collectés dans un ouvrage dédié à la question du distant reading. Voir MORETTI, Franco. *Distant Reading*, Verso, 2013.

7. KESTEMONT, Mike, MOENS, Sara, DEPLOIGE, Jeroen. « Collaborative authorship in the twelfth century : A stylometric study of Hildegard of Bingen and Guibert of Gembloux », *Literary and Linguistic Computing*, 2013.

8. Pour une synthèse en français des méthodes existantes, voir BERNARD, Michel, BOHET, Baptiste. *Littérométrie : outils numériques pour l'analyse des textes littéraires*, Les fondamentaux de la Sorbonne Nouvelle, 2017.

du questionnement d'historien initial et en expliquant le cheminement de pensée qui a mené au choix de ce matériau, pour présenter finalement le corpus que nous avons constitué pour cette étude. Ensuite, nous reviendrons sur la méthode du *topic modeling*, en expliquant son fonctionnement, en présentant le module que nous avons développé pour faciliter sa mise en oeuvre, et en rendant compte rigoureusement du procédé suivi pour appliquer cette méthode à notre corpus. Enfin, nous présenterons les résultats ainsi obtenus, que nous commenterons en détail, avant d'ouvrir la discussion épistémologique sur notre méthode et les différentes perspectives d'enrichissement qui s'offrent à nous pour poursuivre en ce sens.

Bibliographie

Sur la science-fiction

BAUDOU, Jacques. *La science-fiction*, Que sais-je ?, 2003.

BERTHELOT, Francis. *Genres et sous-genres dans les littératures de l'imaginaire*, Séminaire de Narratologie, EHESS, 2004/2005.

BESSON, Anne. « Aux frontières du réel : les genres de l'imaginaire », in *La Revue des livres pour enfants* (CNLJ-La Joie par les livres, BNF), n° 274, décembre 2013, dossier « Littératures de l'imaginaire », p. 86-93.

BRÉAN, Simon. *La science-fiction en France : Théorie et Histoire d'une littérature*. 2012, Presses Universitaires Paris-Sorbonne.

COLSON, Raphaël, RUAUD, André-François. *La science-fiction, une littérature du réel*, Klincksieck, 2006.

FITTING, Peter. « Utopies/Dystopie/Science-fiction : l'interaction de la fiction et du réel », paru dans *Alliage*, n°60 - Juin 2007.

GATTEGNO, Jean. *La science-fiction*, Que sais-je, 1971.

KYROU, Ariel. « Nos subjectivités baignent dans un imaginaire de science-fiction », in *Multitudes*, 2016/1 (n° 62), pp. 126-132. <https://www.cairn.info/revue-multitudes-2016-1-page.htm>.

LANGLET Irène. « Étudier la science-fiction en France aujourd'hui », in *ReS Futurae* [En ligne], 1 | 2012 <http://journals.openedition.org/resf/181>.

LEBAS, Frédéric, COUSSIEU Wilfried. « Avant-propos. La science-fiction, littérature ou sociologie de l'imaginaire ? », in *Sociétés*, 2011/3 (n°113), p. 5-13. URL : <https://www.cairn.info/revue-societes-2011-3-page-5.htm>.

RIEDER, John. « On defining SF, or Not : Genre Theory, SF and History » in *science fiction Studies*, vol. 37, n°116, novembre 2010, pp. 191-209.

UMPALA, Yannick. « Littérature à potentiel heuristique pour temps incertains », *Methodos*, 15 | 2015 URL : <http://journals.openedition.org/methodos/4178>.

Sur les questions d'histoire et d'humanités

CHARTIER Roger. « Le monde comme représentation », *Annales. Économies, Sociétés, Civilisations* 44ème année, N. 6, 1989. pp. 1505-1520.

CORBIN Alain. « "Le vertige des foisonnements" Esquisse panoramique d'une histoire sans nom », *Revue d'histoire moderne et contemporaine*, vol. 39, n°1, janvier-mars 1992, p. 117.

LAVOCAT, Françoise. *La théorie littéraire des mondes possibles*, Paris, Éditions du CNRS. 2010.

PAYEN, Pascal. « François Hartog, Régimes d'historicité. Présentisme et expériences du temps », *Anabases*, 1 | 2005, 295-298.

POPOVIC, Pierre. « La sociocritique. Définition, histoire, concepts, voies d'avenir », *Pratiques*, 151-152 | 2011, 7-38.

VOVELLE, Michel, BOSSÉNO, Christian-Marc. « Des mentalités aux représentations », *Sociétés & Représentations*, vol. 12, no. 2, 2001, pp. 15-28.

ZWEIGENBAUM, Pierre, HABERT, Benoît. « Accès mesurés aux sens », *Mots. Les langages du politique*, 74 | 2004.

Sur la méthodologie informatique

AGGARWAL C., ZHAI C. « Mining Text Data », *Springer*, 2012.

ALIGON, Julien, GUILLET, Fabrice, BLANCHARD, Julien, PICAROUGNE, Fabien. « Analyse par Motifs Fréquents et Topic Modeling », paru dans le cadre du défi EGC, Laboratoire Informatique de Nantes Atlantique, 2016.

BERNARD, Michel, BOHET, Baptiste. « Littérométrie : outils numériques pour l'analyse des textes littéraires », *Les fondamentaux de la Sorbonne Nouvelle*, 2017.

BLEI, David M., NG, Andrew Y., JORDAN Michael I. « Latent Dirichlet Allocation », *Journal of Machine Learning Research* 3, 2003, 993-1022.

BLEI, David M., LAFFERTY, John D. « A CORRELATED TOPIC MODEL OF SCIENCE », *The Annals of Applied Statistics* 2007, Vol. 1, No. 1, 17–35.

CARBOU Guillaume. « Analyser les textes à l'ère des humanités numériques. Quelques questions pour l'analyse statistique des données textuelles », *Les Cahiers du numérique*,

2017/3 (Vol. 13), p. 91-114. URL : <https://www.cairn.info/revue-les-cahiers-du-numerique-2017-3-page-91.htm>.

DEERWESTER, Scott, DUMAIS, Susan, FURNAS, George, LANDAUER, Thomas, HARSHMAN, Richard. « Indexing by latent semantic analysis », *Journal of the american society for information science*, 1990.

HAMILTON William, LESKOVEC Jure, JURAFSKY Dan. « Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change », *arXiv e-prints*, Computer Science Department, Stanford University, 2016.

HOFFMANN, Thomas. « Probabilistic latent semantic analysis », UAI'99 *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, July 1999 Pages 289–296.

KESTEMONT, Mike, MOENS, Sara, DEPLOIGE, Jeroen. « Collaborative authorship in the twelfth century : A stylometric study of Hildegard of Bingen and Guibert of Gembloux », *Literary and Linguistic Computing*, 2013.

MONTARIOL Syrielle. *Models of diachronic semantic changes using word embeddings. Document and Text Processing*. Université Paris-Saclay, 2021.

MORETTI, Franco. *Distant Reading*, Verso, 2013.

NIGAM, K., MCCALUM, A.K., THRUN, S. et al. « Text Classification from Labeled and Unlabeled Documents using EM », *Machine Learning* 39, 103–134 (2000). <https://doi.org/10.1023/A:1007692713085>.

Chapitre 1

La littérature d'anticipation comme support d'investigation

1.1 Du questionnement d'historien...

Tout d'abord, revenons sur le questionnement d'historien qui constitue le point de départ de ces recherches : comment appréhender l'histoire des futurs envisagés, c'est-à-dire des représentations associées à ces projections ? Si cet objet d'études apparaît comme encore relativement singulier et peu étudié dans la littérature académique, du moins dans le monde francophone, notre questionnement, lui, s'inscrit dans une longue tradition historiographique de mise en perspective des représentations et des imaginaires. Il convient donc de revenir sur les discussions scientifiques qui ont précédé notre réflexion, pour mieux comprendre le cheminement de pensée qui a présidé à notre démarche, ainsi que de dresser un état de l'art des recherches en la matière.

Vers une histoire des représentations

L'idée de s'intéresser à l'histoire des représentations n'est pas née dans les années 1980 avec le courant historiographique du même nom ; déjà, l'histoire des mentalités, apparue dans les années 1960 sous la houlette de Georges Duby et Robert Mandrou, avait porté un regard privilégié sur les imaginaires, en esquissant l'histoire de toutes formes de pensées, sentiments ou croyances. Mais cette dernière subissait de lourdes critiques concernant sa manière d'attribuer des états mentaux à des classes ou des catégories sociales entières, et son approche quantitative et sérielle de ces états mentaux. En outre, cette histoire des mentalités reposait sur des concepts flous, des objets encore mal cernés, comme en témoigne l'expression consacrée de Robert Mandrou qui définissait l'histoire des mentalités comme une « histoire des visions du monde ». Dans un article consacré à

la transition des mentalités aux représentations, Michel Vovelle l'évoque d'ailleurs comme une définition « belle mais vague ». Pour d'autres historiens dont Vovelle, l'histoire des mentalités était en fait l'histoire des représentations collectives. On voit, à travers ces multiples tentatives de définition, qu'il s'agissait encore d'un champs de recherche balbutiant, hésitant sur les concepts même¹.

Lorsque Jacques Le Goff coordonne la publication d'un ouvrage sur la « nouvelle histoire » en 1978, et même si le terme de représentation n'apparaît pas encore, il est bel et bien question de représentations collectives, abordées à travers l'histoire de l'amour, de la sexualité ou de la mort. Il y a une forme de rupture avec l'histoire politique et sociale telle qu'on la pratiquait. On note aussi, quoiqu'avec un certain décalage, l'influence du monde académique anglo-saxon et en particulier des *cultural studies*. On citera notamment les analyses classiques de Richard Hoggart sur la culture populaire, qui posent de solides bases méthodologiques pour appréhender les représentations, et intéressent tout particulièrement nos recherches. Un de ses ouvrages majeurs, *La culture du pauvre – The Uses of Literacy* en langue originale, dont on note qu'il ne sera publié en France qu'en 1970, soit 13 ans après sa publication en Grande-Bretagne, ouvre quelques pistes de réflexion d'intérêt pour l'analyse de la science-fiction, qui est une littérature fondamentalement populaire.

Finalement, on observe dans la deuxième moitié des années 1980 en France une rupture marquant le passage de l'histoire des mentalités à l'histoire des représentations. Une des figures emblématiques de cette rupture est Roger Chartier, qui propose dans son article de référence de 1989 de « faire une histoire avec un retour du social à partir des représentations que l'on s'en fait et qui le conditionnent »². Les représentations deviennent alors un moyen de faire de l'histoire sociale. Elles sont, comme l'exprimaient joliment Chartier, la manière par laquelle on rend présent ce qui est absent.

Définissons ici clairement ce que l'on entend par histoire des représentations, puisque le terme français de *représentation* peut prêter à confusion – on note sa polysémie. Alain Corbin en donne une définition claire et précise, expliquant qu'elle se concentre sur l'histoire « [des] représentations qu'un individu se fait du monde, d'un éventuel au-delà, de lui-même et de l'autre. Ces représentations qui règlent le jeu du désir et de la répulsion ; qui décident des figures de l'angoisse et de l'horreur. Le système de représentations ne fait pas qu'ordonner le système d'appréciation ; il détermine les modalités de l'observation du monde, de la société »³. En d'autres termes, il s'agit d'une histoire des idées, des images,

1. VOVELLE, Michel, BOSSÉNO, Christian-Marc. « Des mentalités aux représentations », *Sociétés & Représentations*, vol. 12, no. 2, 2001, pp. 15-28.

2. CHARTIER Roger. « Le monde comme représentation », *Annales. Économies, Sociétés, Civilisations*. 44ème année, N. 6, 1989. pp. 1505-1520.

3. CORBIN Alain, « "Le vertige des foisonnements" Esquisse panoramique d'une histoire sans nom », dans *Revue d'histoire moderne et contemporaine*, vol. 39, n°1, janvier-mars 1992, p. 117.

des symboles, des manières dont on a pu se représenter le monde et la société. Pour faire le lien avec notre étude, nous nous intéressons à ces idées, ces images, et ces symboles, à ces ‘représentations’ en somme, lorsqu’elles concernent des projections du monde et de la société dans l’avenir.

Par quels méthodes, par quels procédés construit-on cette histoire des représentations ? Dans un premier temps, les historiens des représentations se sont fondés sur une analyse quantitative afin de visualiser l’évolution et les inflexions significatives des représentations dans le temps. Ce qui est primordial lorsque l’on travaille sur l’histoire des représentations, c’est l’étude de ces dernières avec une vision chronologique de celles-ci. Cette méthode vient de l’influence des travaux d’Ernest Labrousse et François Simiand notamment, sur l’histoire sociale. La manière d’étudier les représentations a ensuite évolué vers l’étude de cas concrets, par opposition avec l’approche quantitative. Cette évolution dans la méthodologie de l’histoire des représentations pousse les historiens à étudier les individus représentatifs de l’« exception normale », théorisée par Carlo Ginzburg. Néanmoins, selon Michel Vovelle ces deux approches ne sont pas nécessairement opposées, mais peuvent être complémentaires.

Les sources disponibles pour l’histoire des représentations peuvent être de différents types. Dans un premier temps, dans la perspective d’analyses quantitatives, elles sont principalement composées de testaments, ex-voto, retables, etc. Aujourd’hui, elles se constituent principalement de témoignages littéraires ou de documents iconographiques. Les sources disponibles pour l’histoire des représentations peuvent être de différents types. Dans un premier temps, dans la perspective d’analyses quantitatives, elles sont principalement composées de testaments, ex-voto, retables, etc. Aujourd’hui, elles se constituent principalement de témoignages littéraires ou de documents iconographiques. L’histoire des représentations s’appuie principalement sur les mots, les images et les symboles.. L’usage des mots est donc un moyen particulièrement apprécié lorsqu’on s’intéresse à l’histoire des représentations associées au futur, et il nous faut ici poser la question fastidieuse du lien existant entre littérature et représentations.

Littérature et représentations

Il existe un lien fort, ou plus précisément, une forme de circularité entre littérature et imaginaire social, représentations collectives. La littérature, en ce qu’elle est le résultat d’une formalisation problématique de l’imaginaire social, entretient une relation privilégiée autant qu’ambigüe avec les imaginaires collectifs. Ainsi, l’objet littéraire constitue un matériau intéressant pour étudier l’histoire des imaginaires sociaux, et comprendre leur évolution, Ici, il nous faut mobiliser les travaux de la sociocritique pour comprendre la nature de ces liens réciproques, et la manière dont on peut accéder à des représentations à travers la littérature. Un article synthétique de Pierre Popovic permet de bien comprendre

les enjeux de la sociocritique, en même temps qu'il en donne une définition claire⁴ :

« La sociocritique n'est ni une discipline ni une théorie. Elle n'est pas non plus une sociologie, de quelque sorte qu'elle soit, encore moins une méthode. Elle constitue une perspective. À ce titre, elle pose comme principe fondateur une proposition heuristique générale de laquelle peuvent dériver de nombreuses problématiques individuellement cohérentes et mutuellement compatibles. Cette proposition se présente comme suit : Le but de la sociocritique est de dégager la socialité des textes. Celle-ci est analysable dans les caractéristiques de leurs mises en forme, lesquelles se comprennent rapportées à la semiosis sociale environnante prise en partie ou dans sa totalité. L'étude de ce rapport de commutation sémiotique permet d'expliquer la forme-sens (thématisations, contradictions, apories, dérives sémantiques, polysémie, etc.) des textes, d'évaluer et de mettre en valeur leur historicité, leur portée critique et leur capacité d'invention à l'égard du monde social. Analyser, comprendre, expliquer, évaluer, ce sont là les quatre temps d'une herméneutique. C'est pourquoi la sociocritique – qui s'appellerait tout aussi bien « sociosémiose » -peut se définir de manière concise comme une herméneutique sociale des textes. »

Cette définition posée, on comprend que la sociocritique ait apporté beaucoup à l'étude des liens entre littérature et discours social, à bien des niveaux différents. Ce qui nous intéresse ici, c'est précisément l'étude des liens entre le texte littéraire et le discours social, mais sur le plan des représentations, des topiques, des motifs récurrents.

Mais une telle étude implique un important effort de définition du concept d'imaginaire social, et des modes d'interaction que la littérature entretient avec celui-ci. Là encore, Pierre Popovic en propose un exposé clair et succinct, que nous aimerions ici mobiliser :

« L'imaginaire social est ce rêve éveillé que les membres d'une société font, voient, lisent et entendent et qui leur sert d'horizon de référence pour tenter d'appréhender, d'évaluer et de comprendre la réalité dans laquelle ils vivent. Il peut être pensé comme suit :

L'imaginaire social est composé d'ensembles interactifs de représentations corrélées, organisées en fictions latentesL'imaginaire social est composé d'ensembles interactifs de représentations corrélées, organisées en fictions latentes, sans cesse recomposées par des propos, des textes, des chromos et des images, des discours ou des œuvres d'art. [...] Dans toute société, quatre de ces ensembles de représentations sont essentiels : le premier concerne l'histoire et

4. POPOVIC, Pierre. Op. Cit.

la structure de la société (représentations du passé, du présent et de l'avenir ; représentations des institutions, des hiérarchies, des collectivités) ; le deuxième la relation entre l'individu et le collectif global (représentation de l'individu, de sa vie, des rapports du privé et du public) ; le troisième la vie érotique (représentations des corps, des affects, des sentiments, du sexe) ; le quatrième le rapport avec la nature (représentations métaphysiques, religieuses ou non religieuses, etc.).

[...] *L'imaginaire social peut être vu comme le résultat de l'action de cinq modes majeurs de sémiotisation de la réalité [...] – narrativité, poéticité, cognitivité, iconicité, théâtralité – signifie que l'imaginaire social est empreint de littérarité. De là découle une façon particulière de reposer la question nodale de la sociocritique. Ce qui est appelé « littérature » est ce qui résulte d'une formalisation problématique de l'imaginaire social, aussi bien au niveau du système générique qu'au niveau des textes : un roman est en interaction dynamique majeure avec la narrativité ambiante [...] L'imaginaire social peut être vu comme le résultat de l'action de cinq modes majeurs de sémiotisation de la réalité [...] – narrativité, poéticité, cognitivité, iconicité, théâtralité – signifie que l'imaginaire social est empreint de littérarité. De là découle une façon particulière de reposer la question nodale de la sociocritique. Ce qui est appelé « littérature » est ce qui résulte d'une formalisation problématique de l'imaginaire social, aussi bien au niveau du système générique qu'au niveau des textes : un roman est en interaction dynamique majeure avec la narrativité ambiante, c'est-à-dire avec les façons de raconter l'histoire de la société et les événements qu'elle vit. La mise en texte peut sans doute simplement calquer ces façons de raconter, mais elle vise plus souvent – dans les sociétés qui le permettent – à les détourner, à les critiquer, à les transformer. Toute mise en forme vivace est alors dite problématique parce que les textes littéraires sont susceptibles d'installer une distance sémiotique à l'intérieur et à l'égard de cet imaginaire social justement pour la raison qu'ils activent, individualisent les cinq modes de sémiotisation prédécrit [...] L'imaginaire social peut être vu comme le résultat de l'action de cinq modes majeurs de sémiotisation de la réalité [...] – narrativité, poéticité, cognitivité, iconicité, théâtralité – signifie que l'imaginaire social est empreint de littérarité. De là découle une façon particulière de reposer la question nodale de la sociocritique. Ce qui est appelé « littérature » est ce qui résulte d'une formalisation problématique de l'imaginaire social, aussi bien au niveau du système générique qu'au niveau des textes : un roman est en interaction dynamique majeure avec la narrativité ambiante [...] L'imaginaire social peut être vu comme le résultat de l'action de cinq modes majeurs de sémiotisation de la réalité [...] – narrativité, poéticité,*

cognitivité, iconicité, théâtralité – signifie que l'imaginaire social est empreint de littérarité. De là découle une façon particulière de reposer la question nodale de la sociocritique. Ce qui est appelé « littérature » est ce qui résulte d'une formalisation problématique de l'imaginaire social, aussi bien au niveau du système générique qu'au niveau des textes : un roman est en interaction dynamique majeure avec la narrativité ambiante, c'est-à-dire avec les façons de raconter l'histoire de la société et les événements qu'elle vit. La mise en texte peut sans doute simplement calquer ces façons de raconter, mais elle vise plus souvent – dans les sociétés qui le permettent – à les détourner, à les critiquer, à les transformer. Toute mise en forme vivace est alors dite problématique parce que les textes littéraires sont susceptibles d'installer une distance sémiotique à l'intérieur et à l'égard de cet imaginaire social justement pour la raison qu'ils activent, individualisent les cinq modes de sémiotisation prédictés. »

Armés de ces précieux concepts, on comprend l'intérêt de la littérature comme support pour étudier l'histoire des représentations et l'évolution du discours social. Or, ce qui nous intéresse, c'est d'arriver à explorer les représentations associées aux projections futuristes, la partie du discours social qui porte sur l'avenir. Pourquoi, alors, ne pas fonder notre analyse sur une littérature qui s'emploie à imaginer des fictions dans des mondes futurs ?

Utiliser la SF pour faire de l'histoire des représentations futuristes

L'idée d'utiliser la littérature de science-fiction pour sonder les représentations n'est pas neuve. Déjà en 2011, dans l'avant-propos d'un numéro de la revue *Sociétés* consacré à la science-fiction, les sociologues Wilfried Coussieu et Frédéric Lebas démontraient que la SF pouvait être appréhendée comme « une sociologie de l'imaginaire »⁵.

« La science-fiction exerce actuellement un champ d'influence que l'on pourrait qualifier de systémique. L'éventail des hallucinations et des connaissances qu'elle met en jeu se situe à l'intersection complexe du savoir scientifique et des représentations sociales, nourrissant l'un et l'autre de rapports vertigineux et contrastés. Sa capacité hybride à mêler le vrai et le vraisemblable, sa mise en scène des concepts dialectaux de la science rendus localement au service de la fiction, ainsi que son recours constant à des thèmes angoissés, anticipateurs et assoiffés d'alternatives (utopiques ou dystopiques), semble également affirmer la porosité de sens caractéristique de la société postmoderne. Par ailleurs, l'évolution et l'impact actuel des franchises médiatiques comme accélétratrices de modes

5. LEBAS, Frédéric, COUSSIEU Wilfried, « Avant-propos. La science-fiction, littérature ou sociologie de l'imaginaire ? », in *Sociétés*, 2011/3 (n°113), p. 5-13. URL : <https://www.cairn.info/revue-societes-2011-3-page-5.htm>.

de consommation élaborés sur des univers science-fictionnels, démontrent l'intérêt d'une véritable analyse à inscrire dans le champ des sciences humaines et sociales. »

Ce court extrait synthétise bien la pensée des deux auteurs, et démontre tout l'intérêt de l'objet pour sonder les imaginaires d'une époque. Mais c'est là le propre de toutes les « littératures de l'imaginaire », expression par laquelle on désigne souvent le fantastique, la fantasy, et la science-fiction. Or, ce qui fait la spécificité de la science-fiction, c'est qu'elle « s'attache à ce qui pourra/pourrait se produire, un jour, ailleurs »⁶. Par conséquent, elle permet de cerner plus précisément, au sein des imaginaires qu'elle explore, des projections d'avenir, des « mondes possibles »⁷ : des futurs envisagés.

On voit donc que la science-fiction est déjà exploitée comme un support d'investigation pour la recherche en sciences humaines et sociales. De manière plus récente, l'objet semble intéresser le prisme des humanités numériques. Le 22 novembre 2018, une journée d'étude consacrée à « l'anticipation au prisme des humanités numériques » se tenait à la maison de la recherche de l'université Paris-Sorbonne, montrant l'intérêt des approches computationnelles et de la culture numérique pour appréhender un objet dont l'étude appelle la transversalité. C'est dans cette perspective que nous souhaitons inscrire nos travaux sur les futurs envisagés.

Il faut ici soulever un problème majeur pour notre démarche. Lorsqu'on étudie les futurs envisagés, lorsqu'on sonde les imaginaires associés à une œuvre de science-fiction, sommes-nous réellement en train d'apprécier les imaginaires collectifs d'une époque ? Dans la mesure où l'œuvre renvoie surtout à l'imaginaire de l'auteur qui l'a composée, peut-on monter en généralité, et mettre en lien cet imaginaire avec celui d'une époque entière ? La science-fiction est souvent décrite comme une littérature très spécifique : on pourrait avancer que les futurs envisagés que nous observerons à travers cette littérature ne procéderont que des visions ésotériques de l'auteur, ou au mieux d'un cercle d'initiés, les *Fandoms*⁸, qui envisageraient le futur de manière très spécifique. C'est là un biais éventuel, dont on doit être conscient pour la suite de notre étude. En effet, le raisonnement voulant que des thématiques présentes dans les romans de SF publiés à une époque donneraient accès aux thématiques présentes dans les représentations collectives reste discutable. On pourrait facilement avancer que les thématiques présentes dans une littérature de grande diffusion comme la SF correspondent avant tout à ce que les acteurs de la production littéraire de cette époque – auteurs en premier lieu, mais également éditeurs, critiques, journalistes – estimaient être des thématiques importantes, en résonnance avec leur temps. Les lecteurs

6. BESSON, Anne. *Aux frontières du réel : les genres de l'imaginaire*, in La Revue des livres pour enfants (CNLJ-La Joie par les livres, BNF), n° 274, décembre 2013, dossier « Littératures de l'imaginaire », p. 86-93.

7. LAVOCAT, Françoise, *La théorie littéraire des mondes possibles*, Paris, Éditions du CNRS. 2010

8. Ce terme désigne les lecteurs assidus de SF qui se réunissent autour de magasins, newsgroups, forums, sites Internet spécialisés et conventions. Il serait apparu lorsque les premiers lecteurs d'Amazing Stories se regroupaient pour partager leur passion.

ne sont pas obligés d'y adhérer, quand bien même ils achètent, aiment, ou utilisent ces livres dans leurs pratiques discursives.

Mais l'article de Wilfried Coussieu et Frédéric Lebas montre justement bien que la science-fiction est désormais partie intégrante de la culture contemporaine, affirmant même que « personne n'est désormais épargné par la vague science-fictionnelle ». Surtout, les deux auteurs s'appuyant sur les concepts de H.R. Jauss sur l'esthétique de la réception⁹, mettent en lumière l'importance des mondes possibles que la science-fiction imagine, des représentations qu'elle forge, qui participent de la construction de consciences collectives et influencent même les dispositifs et conduites humaines. D'autant plus lorsque l'on considère uniquement la littérature de science fiction qui procède de l'anticipation réaliste. En outre, il ne faut pas oublier de mentionner le caractère collectif de cette littérature : les inventions ou concepts de certains auteurs sont réemployés par d'autres auteurs, constituant une sorte de répertoire commun. Ainsi, on observe une certaine circularité entre représentations collectives associées au futur et littérature de science-fiction : en même temps qu'elle témoigne de ces imaginaires, elle les nourrit et contribue à les faire évoluer.

Ainsi, la science-fiction apparaît aujourd'hui comme un genre riche et diversifié qui, malgré son caractère fondamentalement populaire, a acquis ses lettres de noblesse sous la plume de grands auteurs, faisant aujourd'hui autorité au-delà des cercles initiés au genre. En outre, le genre s'est exporté à l'international, et on compte de nombreuses scènes de la littérature de science-fiction à travers le monde, et dans diverses langues. On pourrait notamment citer les frères Strougatski en Russie, ou l'Espagnole Rosa Montero, dont le succès s'est étendu à l'ensemble du monde hispanophone et au-delà. En dehors du monde occidental, également, la science-fiction se développe, notamment en Asie, par exemple au Japon et plus récemment, en Chine¹⁰. Cette internationalisation du genre, amenant souvent des spécificités thématiques régionales, contribue à dynamiser et à complexifier la science-fiction. Surtout, la science-fiction et les imaginaires auxquels elle renvoie semble omniprésente dans la culture contemporaine. À ce propos, le journaliste Ariel Kyrou a consacré un article paru en 2016 dans la revue *Multitudes*, dont le titre résumait bien la pensée : « Nos subjectivités baignent dans un imaginaire de science-fiction »¹¹. Néanmoins, l'article expose surtout, à raison, le rôle du cinéma ou de l'industrie des jeux vidéo, voire d'institutions plus formelles comme la Nasa ou l'entreprise Google dans la profusion des imaginaires de science-fiction¹². Il est vrai que le rôle de la littérature a sans doute été moindre dans cette profusion. Cependant, elle a tissé une toile de fond dans lequel le

9. JAUSS, Hans Robert. *Pour une esthétique de la réception*, Paris, Gallimard, coll. Tel, 1978, p. 55.

10. Pour approfondir sur la science-fiction chinoise, voir ALOÏSO, Loïc. « La science-fiction chinoise », *Impressions d'Extrême-Orient* 8 | 2018, <http://journals.openedition.org/ideo/772>.

11. KYROU, Ariel, « Nos subjectivités baignent dans un imaginaire de science-fiction », in *Multitudes*, 2016/1 (n° 62), pp. 126-132. <https://www.cairn.info/revue-multitudes-2016-1-page-126.htm>.

12. KYROU Ariel, op.cit.

cinéma et les jeux-vidéos sont venus puiser beaucoup de concepts, d'inventions, voire des univers tout entier.

La place de la SF dans le monde académique

Il nous reste désormais à traiter la question de la place de la science-fiction dans le monde académique, en dressant un bref aperçu de l'état de la recherche sur cet objet. En France, il faut souligner le peu d'intérêt qu'a suscité le genre jusque récemment. En 2012, dans un article intitulé « Étudier la science-fiction en France aujourd'hui », la professeure de littérature contemporaine Irène Langlet écrivait : « La science-fiction est une pratique intermédiaire large et diversifiée, aussi massivement présente et reconnue dans la culture contemporaine que discrète et méconnue dans le monde universitaire français »¹³. A une période où la science-fiction faisait une entrée discrète dans le champ universitaire français, la professeure consacrait cet article à exposer les lacunes en la matière. Le genre semble néanmoins connaître une vitalité nouvelle depuis une dizaine d'années dans le champ universitaire. En témoigne l'ouvrage majeur publié en 2012 par Simon Bréan¹⁴ sur l'histoire de la littérature de science-fiction francophone.

En outre, s'il est vrai que l'objet reste assez marginal pour ce qui est des études littéraires et stylistiques, une part croissante d'études s'intéressent à la science-fiction, comme comme support d'investigation dans des domaines variés. Étudiée en sciences politiques¹⁵, en sociologie¹⁶, ou en littérature comparée, avec des travaux remarqués sur la théorie des mondes possibles¹⁷, la science-fiction s'est attribuée une place incontestable dans le paysage universitaire. Mais cette présence est disséminée dans les disciplines les plus diverses et on note un vide dans le cœur de discipline. Louée pour sa capacité à saisir les imaginaires, et à les projeter dans l'avenir en inventant des mondes possibles, elle reste cependant un objet difficile à manipuler. D'abord, parce qu'elle s'est développée en France comme une « *subculture* », la science-fiction se nourrit d'une pratique d'érudition populaire qui peine à dialoguer avec le protocole académique. Ensuite, son étude nécessite un travail d'articulation des différents champs de la culture savante, des sciences exactes aux humanités, qui rend peu probante l'approche disciplinaire classique. Enfin, les institutions académiques peinent à concentrer les résultats de travaux divers et variés, procédant de disciplines différentes. Si l'on sort du cadre français pour donner un aperçu de la recherche sur la science-fiction dans le monde, on ne sera pas surpris de trouver une

13. LANGLET Irène, « Étudier la science-fiction en France aujourd'hui », in *ReS Futurae* [En ligne], 1 | 2012 <http://journals.openedition.org/resf/181>.

14. BRÉAN Simon, *La science-fiction en France : Théorie et Histoire d'une littérature*. 2012, Presses Universitaires Paris-Sorbonne.

15. RUMPALA Yannick, « Ce que la science-fiction pourrait apporter à la pensée politique », *Raisons politiques*, 2010/4 (n° 40), p. 97-113. URL : <https://www.cairn.info/revue-raisons-politiques-2010-4-page-97.htm>.

16. LEBAS, Frédéric, COUSSIEU, Wilfried. op. cit.

17. LAVOCAT, Françoise. op. cit.

littérature américaine relativement abondante en la matière. Mais là encore on note une disparité dans les thèmes étudiés : tandis que la littérature très contemporaine, du cyberpunk à nos jours, est très largement sur-représentée, on note une absence de recherches sur la période des *pulps*, pourtant fondatrice¹⁸. D'une manière générale, si la science-fiction trouve désormais une place incontestable dans le monde universitaire, on voit donc que c'est un objet encore mal défini, aux enjeux mal cernés, et qu'on peine encore à exploiter pleinement.

On l'aura compris, la SF est un genre riche et foisonnant, qui recèle d'images et d'idées, et qui constitue un matériau de choix pour quiconque s'intéresse aux représentations. Bien qu'encore relativement peu étudiée, le genre suscite un intérêt croissant dans le monde académique, et appelle de nombreux travaux. C'est aussi l'un des avatars de la production littéraire « industrielle ». En cela, le recours aux Humanités Numériques pour analyser cette littérature prend tout son sens, car elles permettent d'envisager de travailler plus facilement sur ces vastes corpus de données. Par ailleurs, c'est un genre représentatif d'une certaine viralité, notamment dans le cadre des futurs imaginés, puisqu'on y observe une circulation des thèmes et idées entre les œuvres. Il s'agit donc là d'une ressource littéraire idéale pour faire de l'histoire des représentations associées au futur. Il nous faut maintenant présenter plus longuement la littérature de science-fiction afin de bien définir le support d'investigation sur lequel fonder nos travaux.

18. LANGLET Irène, op. cit.

1.2 ...à la définition d'un support d'investigation adapté

Après avoir bien compris les enjeux et l'intérêt d'une étude plus approfondie des futurs envisagés, il nous faut définir un support d'investigation adapté à notre objectif. Au fur et à mesure de premiers travaux de recherche menés dans le cadre du mémoire de M1, le choix de la littérature d'anticipation s'est progressivement précisé, et il convient ici de résumer brièvement ces travaux pour définir plus clairement cette littérature, et expliquer tout son potentiel en tant que matériau de recherche pour appréhender les imaginaires projectionnistes.

Science-fiction et littérature d'anticipation

Revenons d'abord sur la notion de littérature d'anticipation, qui renvoie à une longue réflexion menée dans le cadre de nos travaux de recherche de M1. Nous souhaitions au départ travailler sur la littérature de science-fiction mais, après avoir analysé les sous-genres de la science-fiction, nous sommes parvenus à la conclusion qu'une part substantielle de la littérature de science-fiction ne permettait pas de sonder les ‘futurs envisagés’, en ce qu’elles décrivaient des mondes tout à fait déconnectés de la réalité. À l’inverse, l’anticipation ne constitue pas véritablement un genre de la littérature, mais plutôt un dispositif de narration. Ainsi, ce que nous appellerons ici littérature d’anticipation, et que nous avons choisi d’étudier, renvoie en fait à une partie de la littérature de science-fiction qui procède de l’anticipation : revenons ici sur le cheminement de pensée ayant mené au choix de la littérature d’anticipation.

Initialement, nous avions l’intuition que la littérature de science-fiction pouvait être un support d’investigation intéressant pour sonder les futurs envisagés. La définition du dictionnaire Larousse explicite d’ailleurs très clairement le lien qu’on peut faire entre projections futuristes et littérature SF, définissant cette dernière comme « un genre littéraire qui invente des mondes, des sociétés et des êtres situés dans des espaces-temps fictifs (souvent futurs) impliquant des sciences, des technologies et des situations radicalement différentes ». Ainsi défini, l’objet de notre étude soulève un problème majeur. Peut-on véritablement affirmer qu’un auteur de science-fiction « invente » des mondes « radicalement différents » ? Dans la mesure où celui-ci décrit un état futur du monde en spéculant sur les avancées possibles de la science ou de la technologie, s’agit-il pour lui d’inventer un monde tout à fait nouveau, ou d’imaginer une suite possible à la réalité dans laquelle il évolue ? Même si l’auteur projette son intrigue dans un monde nouveau, celle-ci reste fortement empreinte de considérations propres au réel, au monde dans lequel l’auteur vit : pour le bien de la narration, ce dernier est tenu de pourvoir son récit de repères, d’éléments plus ou moins compréhensibles auxquels le lecteur peut se référer. Quand bien même le monde décrit dans certaines œuvres de science-fiction semble très lointain du

nôtre, la nature même du genre invite toujours le lecteur à imaginer, lorsqu'elle n'est pas suggérée, une trame historique crédible, une suite d'événements plausibles, faisant office de passerelle entre le monde de référence, dans lequel évolue l'auteur, et auquel peut se référer le lecteur, et le monde fictif, dans lequel se déroule l'histoire.

Ici, il faut ré-insister sur le lien qui existe entre la réalité de l'auteur qui compose ces œuvres, et le monde fictif qu'il imagine. C'est précisément le caractère envisageable du récit qui fait l'intérêt de la SF pour notre étude. Écrire un ouvrage de science-fiction, c'est se projeter dans un ailleurs, souvent futur, à partir de son propre monde : c'est donc d'abord porter un regard sur son propre monde. En fin de compte, les éléments auxquels se réfèrent ces récits renvoient à des problématiques propres au contexte dans lequel ces œuvres ont été produites. La science-fiction doit à ce titre être appréhendée comme une « littérature du réel »¹⁹, démontrant l'intérêt de cet objet d'étude pour étudier les imaginaires associés au futur. Or, et c'est là tout le problème, toute science fiction ne se déroule pas nécessairement dans le futur, et même le cas échéant, elle ne procède pas toujours du réalisme, ou du vraisemblable. Si l'on cherche à apprêhender les futurs envisagés à travers cette littérature, il nous faut donc restreindre notre champ d'étude à la seule littérature qui s'attache à décrire un futur envisageable pour l'homme, qu'on pourrait qualifier de littérature d'anticipation. Mais la littérature d'anticipation n'est pas un genre : elle est un dispositif, qui est, il faut le souligner, bien souvent superposé au genre de la science-fiction, d'où une confusion notoire.

C'est bien la littérature de science-fiction qui procède de l'anticipation qui intéresse nos recherches. Or, on l'a dit, toute science fiction ne procède pas de l'anticipation, et inversement, certaines des œuvres d'anticipation majeures n'appartiennent techniquement pas au genre de la SF : c'est par exemple le cas de *1984* (1949) de George Orwell, ou du *Meilleur des Mondes* (1932), d'Aldous Huxley, qui n'ont pas été publiées dans des collections SF, alors même que les historiens du genre s'accordent à les qualifier de littérature de science-fiction, en ce qu'elles projettent toutes deux leur intrigue dans un monde futuriste aux technologies évoluées²⁰. Ce qu'on voit émerger, à travers cette tentative de définition problématique, c'est l'impossibilité à définir clairement un objet complexe, qui s'est construit empiriquement, et qui appelle en fait une définition par l'histoire. Dressons donc un bref aperçu historique et thématique de la science-fiction pour mieux cerner la part de cette littérature qui relève de l'anticipation, et qui intéresse notre étude.

19. COLSON, Raphaël, RUAUD, André-François. *La science-fiction, une littérature du réel*, Klincksieck, 2006.

20. Ces œuvres sont toutes deux mentionnées comme des œuvres de science-fiction dans les ouvrages suivants :

. GOIMARD Jacques, AZIZA Claude. *Encyclopédie de poche de la science fiction. Guide de lecture*, Presses Pocket, coll. "Science-fiction", 1986.

. GUIOT Denis. *La Science-fiction*, Massin, coll. "Le monde de...", 1987.

Un objet historiquement construit

Sans rentrer dans un débat fastidieux sur l'utilisation même de la notion de genre, il faut ici souligner qu'il n'existe pas véritablement de définition du genre de la science-fiction : davantage qu'un ensemble de textes cohérent par leur nature, il se caractérise par un usage fait de ces textes, une manière de les relier entre eux. Comme le montre très bien l'historien de la littérature John Rieder dans un article remarqué paru en novembre 2010 dans la revue *science fiction Studies*²¹, ce sont les acteurs de la production, de la distribution et de la consommation de cette littérature qui construisent le genre « par des actes de définition, catégorisation, inclusion et exclusion [...] mais aussi par leur usage des protocoles et leurs stratégies rhétoriques qui distinguent le genre d'autres formes d'écriture et de lecture ». En d'autres termes, en s'appropriant cette littérature, ses usages et ses codes, les auteurs, éditeurs, lecteurs contribuent collectivement à la définir. De fait, dresser un aperçu historique du genre, discuter de ses « origines », pour autant qu'elles soient clairement identifiables, constitue en soi une manière de définir le genre. Il faut bien comprendre que la science-fiction est un objet historiquement construit plus qu'une catégorie arrêtée²².

Les historiens de la science-fiction ont longtemps cherché des origines lointaines au genre ; certaines œuvres antiques, à l'instar de l'*Histoire véritable* de Lucien de Samosate, composée au IIe siècle, ont beaucoup intéressé les spécialistes du genre²³,²⁴, qui y ont vu une forme de proto-science-fiction. À bien des égards, il est vrai que la modernité du récit, qui relate une guerre spatiale entre des peuples extraterrestres, le rapproche du genre. Plus tardivement, on peut considérer *Utopia* (1516), de Thomas More, comme un récit proche de la SF, en ce qu'elle projette son intrigue dans un ailleurs, un *non lieu*, un monde différent. Cependant, nous prendrons le parti d'affirmer que la modernité scientifique et technique du XIXe siècle, ainsi que le passage à un régime d'historicité dans lequel « le futur prévaut sur le présent »²⁵, induit une véritable rupture, préalable à la constitution d'une littérature propre. Comme le souligne admirablement Jean Gattégno²⁶, « l'erreur de tout historien de la science-fiction est de négliger qu'il ne peut y avoir de science-fiction tant qu'il n'y a pas de sciences, et même de sciences appliquées ». Certes, certaines œuvres antérieures ont pu présenter des caractéristiques très semblables à celles

21. RIEDER, John. « On defining SF, or Not : Genre Theory, SF and History » in *science fiction Studies*, vol. 37, n°116, novembre 2010, pp. 191-209.

22. RIEDER, John. op. cit.

23. FREDERICKS, S.C. « Lucian's True History as SF », *science fiction Studies*, mars 1976, p. 49-60.

24. SWANSON, Roy Arthur Swanson. « The True, the False, and the Truly False : Lucian's Philosophical science fiction », *science fiction Studies*, Vol. 3, No. 3, novembre 1976, p. 227-239.

25. Pour une bonne synthèse du concept de régime d'historicité chez François Hartog, voir PAYEN, Pascal. « François Hartog, Régimes d'historicité. Présentisme et expériences du temps », *Anabases*, 1 | 2005, 295-298.

26. GATTEGNO, Jean. *La science-fiction*, col. « Que sais-je », n°1426, 1971, p. 9.

des ouvrages que nous étudions, mais ils relèvent davantage de la fantaisie, du fantastique, voire du surnaturel. Or, tandis que ces derniers font intervenir des éléments surnaturels, parfois d'origine magique ou divine, qui ne nécessitent pas de justification rationnelle, la science-fiction, elle, fournit une explication logique, scientifique, aux mondes qu'elle imagine²⁷.

Au XIXe siècle, à l'aune d'une modernité nouvelle, la science-fiction se constitue notamment en littérature d'anticipation. C'est précisément cet aspect qui intéresse nos travaux : l'ancre du récit dans une réalité tangible, dans un monde qui apparaît comme un futur possible, ou du moins, envisageable, puisqu'il est parfois à la limite du vraisemblable. Les premiers récits d'anticipation datent du XVIIIe siècle, mais ne sont pas encore empreints du réalisme scientifique qui apparaît au XIXe siècle²⁸. Aussi, beaucoup considèrent le roman *Frankenstein* de Mary Shelley comme le premier véritable ouvrage de science-fiction, au sens où nous l'entendons : c'est en effet, le premier à faire le récit d'une fiction futuriste qui soit relativement plausible, acceptable, du moins envisageable pour les lecteurs contemporains de l'époque. C'est d'ailleurs la position défendue par Brian Aldiss dans son essai *Billion Year Spree*. Et l'auteur de poursuivre, lorsqu'il évoque *Frankenstein* : « *The event on which this fiction is founded has been supposed, by Dr. Darwin, and some of the physiological writers of Germany, as not of impossible occurrence* »²⁹. Remarquons cependant ici qu'il ne s'agit pas encore tout à fait d'un roman d'anticipation : l'auteur n'indique à aucun moment que l'histoire se déroule dans le futur. Mais l'ouvrage marque une véritable rupture en ce qu'il est le premier à apparaître, aux yeux de ses contemporains, comme scientifiquement probable.

On observe à partir de la parution du roman de Mary Shelley et tout au long du XIXe siècle, non seulement une profusion de la littérature de science-fiction, mais aussi une spécification du genre et de ses thématiques. La projection dans un futur proche, aux sciences et technologies avancées, est un de ces motifs récurrents, qui s'illustre dans des œuvres comme *The Mummy! A Tale of the Twenty-Second Century* (1827) de Jane C. Loudon, *Le Roman de l'Avenir* (1834) de Félix Bodin, ou *Le Monde tel qu'il sera* (1846) d'Emile Souvestre, qui s'attachent à imaginer l'avenir de l'humanité dans un horizon relativement proche, du siècle suivant pour les deux premiers, au XXIIe siècle pour le roman de Jane Loudon. L'apparition plus tardive de la figure du robot, dans un roman de Samuel Butler intitulé *De l'autre côté des montagnes* (1872), est d'ailleurs assez révélatrice de la manière dont l'usage d'une science et d'une technologie de pointe structure le genre – on parle bien de science-fiction, après tout. On retrouve aussi, souvent, la question de l'altérité, qui s'illustre par la rencontre avec une civilisation étrangère, dans bien des cas

27. BERTHELOT, Francis. *Genres et sous-genres dans les littératures de l'imaginaire*, Séminaire de Narratologie, EHESS, 2004/2005.

28. ALKON, Paul. *Origins of Futuristic Fiction*, Georgia UP, 1987.

29. ALDISS, Brian W. *Billion Year Spree*, Doubleday, 1973.

de nature extraterrestre. *Star ou Ψ de Cassiopée* (1854) de Charlemagne Ischir Defontenay, ou *La pluralité des mondes habités* (1862) de Camille Flammarion, en sont de bons exemples. Enfin, il faut relever l'apparition d'un sous-genre plus rare, et plus difficile à traiter dans le cadre de notre étude, mais sur lequel nous reviendrons : celui de l'uchronie. L'ouvrage *Napoléon et la conquête du Monde* (1836) dans lequel Louis Geoffroy imagine le monde tel qu'il serait si Napoléon avait conquis Moscou, et établit une monarchie universelle. L'auteur ne se projette pas véritablement dans l'avenir : il imagine un futur non advenu au passé³⁰. Toujours est-il que la tentative d'imaginer un futur possible à partir d'un état donné du monde reste au cœur de l'idée de science-fiction.

Cette littérature trouve, dans la deuxième moitié du XIXe siècle, de puissants échos, sous les plumes de Jules Verne, et Herbert G. Wells, qui restent aujourd'hui considérés comme les « pères fondateurs du genre »³¹. Ces auteurs sont, en effet, les premiers à avoir recours au procédé narratif de l'anticipation, tout en ancrant leur récit dans des considérations scientifiques. Des œuvres comme *De la Terre à la Lune* (1865) et *Vingt mille lieues sous les mers* (1869) pour le premier, ou *La Machine à explorer le temps* (1895) et *La guerre des mondes* (1898) pour le second, illustrent bien l'omniprésence des thématiques évoquées précédemment. C'est notamment grâce à ces auteurs que le genre s'exporte aux États-Unis, où il trouvera un nouvel essor au siècle suivant. Cette époque est aussi marquée par l'alphabetisation croissante des sociétés occidentales, et le développement d'une littérature populaire grandement diffusée par les revues, qui permettent l'émergence d'un lectorat populaire encore relativement limité, mais toujours plus nombreux.

C'est véritablement à partir du XXe siècle que le genre se popularise pour connaître, entre les années 1930 et 1950, son « âge d'or ». On l'a dit, le genre a été très bien reçu aux Etats-Unis, où il inspire de nombreux auteurs. Or, les débuts du XXe siècle voient naître, dans ce même pays, les *pulp magazines*. Héritage direct des *dime novels* (« romans à trois sous »), ces magazines imprimés sur des matériaux de piètre qualité se caractérisent par leur faible coût et leur grande accessibilité, et proposent des fictions de toutes sortes, destinées à divertir un public très large. Ils publient, entre autres, de nombreuses histoires de SF, qui connaissent un véritable succès auprès du grand public. Grâce à ce moyen de diffusion, la science-fiction conquiert un lectorat de plus en plus large et devient, dès lors, une littérature fondamentalement populaire³². En 1926, Hugo Gernsback crée *Amazing Stories*, le premier magazine consacré exclusivement aux histoires de science-fiction. C'est dans les colonnes de ce magazine que le terme de science-fiction entre en usage. Bien qu'on le retrouve de manière isolée, en 1851, sous la plume de l'auteur britannique William

30. Sur l'uchronie, voir BAUDOU, Jacques. *La science-fiction*, Que sais-je ?, 2003, p. 91.

31. L'expression « pères fondateurs » est ici empruntée à Jacques Baudou, dans un chapitre qu'il consacre à Jules Verne et Herbert G. Wells. Voir BAUDOU, Jacques. op. cit.

32. Pour une bonne synthèse en français sur l'histoire des pulp magazines, voir SAINT-MARTIN, Francis. *Les pulps - l'âge d'or de la littérature populaire américaine*, Encrage, 2000.

Wilson, c'est l'éditeur H. Gernsback qui répand son usage aux États-Unis et ailleurs³³.

A partir des années 1930, on assiste à une multiplication des magazines spécialisés dans la science-fiction. Au total, plus d'une trentaine de revues du genre sont créées aux États-Unis pendant l'entre deux-guerres, faisant du pays le véritable épicentre de la production de science-fiction à cette époque. On l'a dit, les *pulp* sont des magazines à la production peu coûteuse, qui permettent un très grand nombre de tirages. Parfois, on compte jusqu'à 500 000 exemplaires pour une même revue³⁴. Parmi les plus emblématiques, on compte notamment *Weird Tales*, *Wonder Stories* et *Astounding Stories*. Bien que les *pulp magazines* peinent à s'exporter hors du territoire américain, la science-fiction se développe également au Royaume-Uni, en Allemagne, et en France, grâce à l'industrialisation de la presse et à la baisse du coût des tirages. Le caractère populaire de la science-fiction se traduit aussi par l'apparition du genre au cinéma, dès le début du XXe siècle avec *Le Voyage dans la lune* (1902) de Georges Méliès. Sans trop s'étendre sur le sujet, on retiendra de ce passage à l'écran qu'il ouvre le genre à un public toujours plus large.

Entre 1930 et 1950, la science-fiction connaît donc un « âge d'or »³⁵ : l'actualité brûlante de la période avec la montée des totalitarismes, la Seconde Guerre mondiale et les débuts de la Guerre Froide, ainsi que l'émergence de nouvelles technologies destructrices, à l'instar de la bombe nucléaire, inspirent une génération d'auteurs. En 1937, John W. Campbell, un écrivain et éditeur de SF, prend la direction de la rédaction d'*Astounding SF*. Il met en place une nouvelle politique éditoriale, qui s'articule autour d'idées et de spéculations sur les sciences exactes. Il veut une SF d'ingénieur : de manière très imagée, on parle souvent de « hard science-fiction » pour qualifier le style des auteurs de cette période, en référence aux sciences dures (« hard sciences »). On compte par exemple parmi eux Isaac Asimov, inventeur du terme de « robotique » et concepteur des fameuses « trois lois de la robotique », ou encore le britannique Arthur C. Clarke. Les bouleversements politiques de la période sont aussi un cadre propice à l'émergence d'une littérature qui embrasse les thèmes politiques, et que certains qualifient de *politic fiction*³⁶. En 1932 paraît *Le Meilleur des mondes* d'Aldous Huxley, titre emblématique qui pose assez tôt la question des risques de la montée des totalitarismes, de l'uniformisation et de l'eugénisme. L'immédiat après-guerre est tout aussi marqué par cette réflexion, avec la parution en 1949 d'un titre emblématique : *1984*, de George Orwell, aura une très grande portée, attestant l'influence croissante de la science-fiction au sein des sociétés occidentales.

Dans les années 1960, en effet, la science-fiction mute encore profondément, no-

33. Informations tirées du centre national de ressources textuelles et lexicales. URL : <https://www.cnrtl.fr/etymologie/science-fiction>.

34. SAINT-MARTIN, Francis. op. cit.

35. BAUDOU, Jacques. op. cit. p. 32.

36. BERTHELOT, Francis. op. cit.

tamment sur le plan éditorial. Depuis la création d'*Amazing Stories* en 1926, les revues spécialisées ont été le support quasi unique de la littérature de SF. Au début des années 60, le marché des revues reste dominé par celle de J.W. Campbell, rebaptisée *Analog*, et d'autres revues comme *Galaxy SF* et *The Magazine of Fantasy and science-fiction* continuent d'être publiées avec un certain succès³⁷. Mais la science-fiction délaisse progressivement les magazines spécialisés pour investir le monde de l'édition en livres. Jusqu'au début des années 1960, les œuvres de SF ne paraissaient que très rarement en librairie, même si quelques passionnés du genre avaient déjà créé des structures éditoriales spécialisées pour conserver les meilleures œuvres de SF, sous la forme d'ouvrages reliés sous jaquette. Dans les années 1960, justement, les éditeurs généralistes se mettent, eux aussi, à publier de la SF en éditions reliées, à un moment où le livre de poche fait une petite révolution dans le monde de l'édition. De nombreuses œuvres existantes sont d'abord rééditées par des collections de livre de poche, puis progressivement, ces collections publient de l'inédit. La SF rejoint alors le circuit de la grande distribution : bénéficiant d'une meilleure exposition, elle gagne un lectorat encore plus large. Cette mutation induit également une mutation de la forme du récit : jusqu'alors, la science-fiction se composait essentiellement de récits courts, les œuvres longues paraissant sous la forme de feuilleton dans les magazines ; à partir des années 1960, cependant, le roman devient le mode d'expression dominant du genre.

La mutation est aussi d'ordre thématique : la science-fiction s'enrichit ainsi de nombreux sous-genres au cours des années 1960. S'inspirant de plus en plus des problèmes contemporains, elle invite le lecteur à s'interroger sur la société dans laquelle il évolue. En projetant les problématiques de leur temps dans un futur proche et plus que jamais envisageable, ses auteurs portent un regard social, parfois militant, sur leur époque. La multiplication des thématiques abordées, des problèmes envisagés, et de manière plus générale, de la production de science-fiction, font de cette littérature un genre dynamique, riche et diversifié, ramifié en de nombreux sous-genres³⁸. Loin de prétendre à en dresser un tableau exhaustif, nous souhaitons ici évoquer les principaux genres existants et leur histoire, afin de donner un aperçu de cette profusion et de mieux définir, par la suite, la littérature que nous ciblerons pour notre étude. Ce qu'on voit se dessiner, à travers cette description thématique des genres, c'est une littérature de science-fiction différenciée en fonction des aspects du futur qu'elle choisit d'interroger, mais aussi en fonction du niveau de vraisemblance, de ces futurs envisagés.

Un aperçu thématique des genres de la SF : entre anticipation et fiction

Après ce détour nécessaire sur l'histoire de la littérature de science-fiction, nous voilà

37. BERTHELOT, Francis. op. cit.

38. BERTHELOT, Francis. op. cit.

armés pour donner un aperçu thématique des sous-genres de la SF, et bien cerner leur niveau de vraisemblance respectif. En déterminant à chaque fois s'il relève davantage de l'anticipation ou de la fiction, nous tâcherons d'en revenir à notre questionnement initial, et de préciser ensemble la part de la littérature SF qui relève de l'anticipation.

On a déjà évoqué brièvement l'apparition de la **hard science fiction** dans les années 1930, sous l'influence de Campbell et d'une vague d'auteurs qu'incarnent Isaac Asimov et Arthur C. Clarke. Ces œuvres se caractérisent par le fait qu'elles ne développent que des théories plausibles s'appuyant sur les découvertes les plus récentes dans les domaines scientifiques de la physique ou de l'astronomie. Ainsi, le genre impose à l'auteur une rigueur méthodique lorsque ce dernier imagine des mondes. L'usage d'un discours scientifique en est d'ailleurs caractéristique. Au début des années 1960, le marché des revues reste dominé par celle de J.W. Campbell, rebaptisée *Analog. Galaxy SF* et *The Magazine of Fantasy and science-fiction*. Il s'agit pour l'auteur d'explorer le domaine des sciences exactes à travers la fiction, et on en retrouve de bons exemples, dans la revue *Amazing Stories*, notamment avec les auteurs susmentionnés, mais aussi plus tardivement, sous forme de livres de poche. On pense notamment au *Nuage Noir* (1957) de Fred Hoyle, ou à *L'Anneau-monde* (1970) de Larry Niven. Ce sous-genre procède de l'anticipation, et s'attache particulièrement au réalisme scientifique des mondes qu'il décrit : il est donc caractérisé par un haut niveau de vraisemblance.

Il existe également une littérature d'anticipation qui s'attache davantage à la dimension politique des mondes qu'elle anticipe, et que certains qualifient de **politic fiction**. Elle s'applique à imaginer, souvent d'un point de vue critique, les conséquences à court ou long terme de choix politiques et sociaux contemporains. On pense notamment au *Meilleur des mondes* (1932) d'Aldous Huxley, ou à *1984* (1949) de George Orwell, qui projettent tous deux les dangers du totalitarisme et de l'uniformisation dans les mondes qu'ils imaginent. Comme on l'a expliqué, ces œuvres appartiennent bien au genre de la science-fiction, même si elles n'ont pas été publiées dans des collections SF. Notons toutefois que nous souhaitons être assez prudents avec le terme de **politic fiction**, qui ne fait pas consensus, et auquel nous privilierons ici le concept de **littérature d'anticipation politique**. On dira d'elle qu'elle connaît une vitalité nouvelle dans les années 1960-70, notamment en Europe, sous les plumes de l'Anglais Anthony Burgess, avec *L'Orange mécanique* (1962), et du Français Philippe Curval dans *Le Cycle de l'Europe après la pluie* (1976-1983). En outre, remarquons que l'**anticipation politique** présente souvent une vision dystopique d'un monde en proie aux conséquences dramatiques des choix politiques et sociaux contemporains, contrairement à la **hard science fiction** qui loue les progrès scientifiques et technologiques des futurs utopiques qu'elle imagine. On note ici l'existence d'une polarité récurrente entre des conceptions opposées du futur, qui intéressera

nos travaux³⁹.

L'introduction des sciences humaines dans la littérature de science-fiction donne naissance à un nouveau genre dans les années 1960 : la *speculative fiction*. Par opposition à la *hard science fiction*, la fiction spéculative met l'accent sur les idées, sur l'évolution psychosociologique des sociétés qu'elle décrit : les questions qu'elle pose appellent en ce sens des réponses moins tranchées, moins scientifiques et rationnelles, d'où son caractère « spéculatif ». Parmi ce genre, on compte notamment la *Série des quatre apocalypses* de J.G. Ballard qui, entre 1962 et 1966, explore la question de la destruction du monde en quatre volets consacrés aux quatre éléments. La question de la réalité truquée s'illustre également à travers ce genre : on pense notamment aux ouvrages de Philip K. Dick, *Le Temps désarticulé* (1959), et *Ubik* (1969). Le genre inspire beaucoup des grandes œuvres de science-fiction dans les années 1960-70, et explore davantage les aspects sociaux des futurs qu'elle envisage. Ainsi, par exemple, des questions du langage et de la communication, avec *L'Enchâssement* (1973) de Ian Watson, ou de l'ambivalence sexuelle, avec *La Main gauche de la nuit* (1969) d'Ursula le Guin.

Le sous-genre du *Space Opera* connaît également un essor important dans les années 1960. Le terme avait été forgé en 1941 par l'écrivain Wilson Tucker qui, ironisant sur certaines œuvres de SF aux intrigues naïves, faisait l'analogie avec les « soap opera »⁴⁰. Ces drames aux intrigues longues et complexes se caractérisent par un cadre géopolitique futuriste, souvent intergalactique. Aussi, ils prennent souvent la forme de cycles littéraires, en plusieurs opus, à l'instar du célèbre *Cycle de Dune* (1965) de Frank Herbert ou des *Cycle de Mars* (1917-1922) et *Cycle de Vénus* (1934-1946) d'Edgar R. Burroughs. À partir des années 1960, le courant devient très populaire, et on voit apparaître de nombreuses œuvres *space opera* : marqué par une grande vitalité, le sous-genre s'enrichit des apports d'autres courants de la SF, et sera à l'origine de beaucoup des grands classiques de la science-fiction. Néanmoins on voit que le genre ne procède pas de l'anticipation, en ce qu'il n'attache que peu d'importance au niveau de vraisemblance du récit : la littérature *space opera* ne se veut pas nécessairement réaliste. En outre, il décrit souvent un ailleurs lointain, quitte à priver parfois le récit d'un lien direct avec l'Histoire humaine. Signalons ici qu'on retrouvera parfois une variante plus récente du *space opera* sous le nom de *planet opera*, dont le récit est centré non sur l'espace tout entier mais sur une planète unique. On pense notamment à *Lum'en* (2015), de Laurent Genefort.

Il faut ici mentionner un autre sous-genre majeur : le *cyberpunk* (en référence au mouvement culturel *punk*), apparu dans les années 1980. Le terme est employé pour la

39. FITTING, Peter. « Utopies/Dystopie/Science-fiction : l'interaction de la fiction et du réel », paru dans *Alliage*, n°60 - Juin 2007, II. Science fiction et politique : le prisme des utopies, Utopies/Dystopie/Science-fiction : l'interaction de la fiction et du réel, mis en ligne le 01 août 2012, URL : <http://revel.unice.fr/alliage/index.html?id=3494>.

40. BERTHELOT, Francis. op. cit.

première fois en décembre 1984 dans un article du *Washington Post*, intitulé « SF in the Eighties », pour qualifier l'œuvre remarquée de William Gibson, *Neuromancien*, parue la même année⁴¹. Ce sous-genre qualifie des œuvres d'anticipation dystopiques, alimentées par les grands progrès scientifiques et techniques des années 1980, notamment concernant l'informatique, Internet, la robotique et l'aéronautique. Le genre développe, entre autres, les thèmes du cyberspace, et de l'homme augmenté. Outre le roman fondateur évoqué, on compte notamment *Les androïdes rêvent-ils de moutons électriques* (1966) de Philippe K. Dick, plus connu sous le nom de son adaptation au cinéma, *Blade Runner*, ou encore, très récemment, *Des larmes sous la pluie* (2011), de Rosa Montero, qui se situe dans un prolongement du monde imaginé dans *Blade Runner*. Le **cyberpunk** inspire également un autre courant plus marginal dans les années 1980 : celui du **steampunk**, en référence à la machine à vapeur, puisqu'il revisite les progrès technologiques du XIXe siècle. Là, on ne peut évidemment pas parler de littérature d'anticipation, puisque l'intrigue se situe dans le passé. Le monde imaginé ne relève donc pas d'un futur envisagé, mais plutôt d'un paradoxe temporel. Plus délicat à manipuler, il n'intéressera pas nos travaux pour le moment.

Enfin, un sous-genre majeur, encore peu évoqué jusqu'ici, doit être mentionné : celui de l'**uchronie**. De manière rétrospective, l'**uchronie** imagine des passés alternatifs, des futurs non advenus, en revisitant notre histoire. On peut citer *Autant en emporte le temps* (1953) de Ward Moore, qui explore les conséquences qu'aurait eu une victoire des confédérés lors de la guerre de Sécession, ou *Le Maître du haut château* (1962), dans lequel Philippe K. Dick imagine une Amérique partagée en zones d'occupations allemande et japonaise après la victoire de l'Axe. Là encore, cette littérature relève du paradoxe temporel, et non de l'anticipation. Elle sera donc écartée à ce stade de nos travaux. On doit également mentionner l'existence de sous-genres caractérisées par la recherche stylistique, davantage que par le souci de réalisme scientifique, comme la **New Wave**. Né dans les années 1960 au Royaume-Uni, le courant est popularisé par l'écrivain Michael Moorcock à travers le magazine *New Worlds*, dont il a pris la direction éditoriale en 1964. Cette littérature ne relevant pas non plus nécessairement de l'anticipation, nous ne nous étendrons pas plus sur le sujet.

Encore une fois, il convient d'insister sur la complexité à déterminer nettement des catégories de sous-genres qui sont en mutation permanente, et s'influencent mutuellement. Parfois, ils se recoupent même, comme c'est le cas avec la **New Wave** et la **Speculative fiction**⁴². A cet égard, les travaux de Francis Berthelot montrent bien la porosité des

41. BERTHELOT, Francis. op. cit.

42. C'est notamment le cas de l'oeuvre d'Ursula K. Le Guin, orientée de considérations sociales propres à la speculative fiction, mais marquée par une recherche stylistique caractéristique de la New Wave. On peut notamment citer *La Main gauche de la nuit* (1969).

genres de la science-fiction, dont les limites sont souvent floues⁴³. Loin d'être arrêtée, cette catégorisation des sous-genres est perpétuellement remise en cause par l'apparition de nouveaux courants, de nouvelles tendances, et par l'internationalisation de la science-fiction. Toujours est-il qu'ils donnent un aperçu de la grande diversité de sous-genres qui se développent dans la seconde moitié du XXe siècle, et permettent d'expliquer la richesse actuelle de la SF. Très dynamique, le genre est marqué par une recherche permanente de renouveau, et puise son inspiration dans les problématiques socio-politiques contemporaines émergentes, ce qui complique encore notre tentative de catégorisation. Néanmoins, ce travail préalable nous donne un aperçu thématique des genres, et permet de bien cerner la part de cette littérature qui procède de l'anticipation dans son volet le plus réaliste.

43. BERTHELOT, Francis. op. cit.

1.3 Constituer un corpus de littérature d'anticipation

Au terme de ces premières recherches, nous avons donc posé un cadre préalable à la définition d'un support d'investigation adapté à l'étude des futurs envisagés, la littérature d'anticipation, en démontrant tout son potentiel heuristique. Il nous reste alors à constituer un véritable corpus d'analyse. Nous reviendrons ici sur les différents problèmes épistémologiques soulevés par la constitution de notre corpus et sur les éléments de réflexion que nous avons apporté à ces questions, avant de présenter plus en détail le corpus ainsi composé.

Définir clairement la littérature d'anticipation

Il doit désormais nous permettre de définir clairement ce que nous entendrons par littérature d'anticipation, qui se restreindra, à une partie seulement de la littérature de science-fiction. D'abord, écartons les sous-genres de l'*uchronie* et du *steampunk*, puisqu'ils présentent des paradoxes temporels. En effet, les mondes décrits dans ces œuvres ne renvoient pas à des futurs envisagés, mais plutôt à des futurs non advenus pour l'un, et à des passés alternatifs, revisités, pour l'autre. On pourrait avancer qu'ils décrivent des futurs envisagés du point de vue d'un moment historique passé. Mais le regard ultérieur porté par l'auteur sur ce moment historique introduit nécessairement un biais, qui rend très délicate la manipulation de ces œuvres. Aussi, nous préférerons les écarter de notre champ d'étude, pour le moment.

De même, la question du *space opera*, qui constitue une part conséquente de la littérature de science-fiction, mérite d'être posée. Il constitue un sous-genre complexe, parfois à la limite des procédés de l'anticipation, mais d'une anticipation qui n'a pas vraiment de visée réaliste. De plus, s'il décrit ce qui pourrait se produire un jour, si les mondes qu'il invente procèdent toujours d'un futur acceptable, il met bien souvent en scène un ailleurs lointain, qui prive parfois le monde fictif d'un lien direct avec le monde du lecteur. Aussi, la question des futurs envisagés invite un traitement plus rigoureux à travers le genre du *space opera*, prenant en compte cette dimension spécifique. On choisira donc ici d'exclure temporairement le genre de notre corpus d'étude. Soulignons néanmoins que nous comptons, à terme, inclure cette littérature abondante à nos travaux.

Reste alors la littérature qui intéresse véritablement notre objet de recherche, à savoir celle qui procède véritablement d'une littérature d'anticipation à visée réaliste. Que cette dernière soit utopique ou dystopique, elle se caractérise par son haut niveau de vraisemblance. On a déjà évoqué les principaux genres auxquels on pourra se référer en la matière : la *hard science-fiction*, l'*anticipation politique*, la *speculative fiction* et le *cyberpunk*. Comme on l'a déjà expliqué, ces genres ont en commun un caractère anticipatoire, à la limite du prospectif. Comme l'exprime on ne peut mieux Ariel Kyrou

« La science-fiction, lorsqu'elle ne se contente pas de divertir, fabrique des souvenirs du futur. Et ces souvenirs, à l'instar de dispositifs d'art contemporain, deviennent autant de tableaux d'hypothèses sur l'à venir »⁴⁴. La part de la littérature de science-fiction que nous retenons ici, celle qui « ne se contente pas de divertir », constitue donc précisément l'objet de nos recherches.

Le choix de la francophonie

Il faut ici évoquer le choix décisif de constituer un corpus composé exclusivement de littérature francophone. De même que toute la discussion historique et thématique que nous avons précédemment menée, ce choix intervient après une première année de travaux de recherche et d'expérimentations. Dans le cadre des travaux de M1, en effet, nous avions constitué un prototype expérimental composé de deux œuvres seulement - *1984*, de G. Orwell, et *Neuromancer*, de W. Gibson – afin de tester nos méthodes. Nous avions alors choisi de travailler sur les traductions françaises de ces deux ouvrages écrits en anglais. Or, les conclusions de ces premières phases expérimentales avaient soulevé un problème épistémologique central autour de la question de la traduction. Dans le cadre d'une analyse sémantique, en effet, il est préférable de travailler sur des textes en version originale, d'autant plus que les traductions sont parfois bien ultérieures à la composition des œuvres, induisant un biais dans l'analyse historique de notre corpus.

Nous avons donc choisi de travailler sur des versions originales. Mais il restait à choisir entre littérature anglophone et francophone. Nous avons pris la décision de travailler exclusivement sur de la littérature d'anticipation francophone, pour deux raisons principales. D'abord parce qu'il nous a semblé que, dans le cadre d'une analyse sémantique, il était important d'avoir une sensibilité particulière aux mots qui constituaient la base de nos résultats. De fait, nous nous sentions plus à l'aise pour apprécier justement leurs différentes connotations, leur texture, leur portée dans notre langue maternelle, le français. Ensuite, car la littérature d'anticipation, et d'une manière plus générale, la science-fiction francophone, est un objet nettement moins étudié dans le monde universitaire, et qu'il nous paraissait plus intéressant de travailler sur des matériaux encore peu abordés.

Un vaste corpus de littérature contemporaine

Les retours constructifs des premiers travaux de recherches effectués dans le cadre du M1 avaient également attiré notre attention sur deux points essentiels pour la constitution de notre corpus. Tout d'abord, les résultats que nous avions obtenu sur le corpus expérimental – composé, rappelons-le, de seulement 2 œuvres – ne permettaient ni de monter en généralité, ni d'apprécier aussi finement les textes qu'avec une méthode d'analyse littéraire plus classique. Aussi, ils indiquaient que la méthode n'aurait de l'intérêt

44. KYROU Ariel. op.cit.

qu'avec la montée en échelle de notre corpus, puisqu'elle permettrait à la fois de monter en généralité dans l'exploration thématique de nos œuvres, et de surpasser la capacité de traitement de méthodes d'analyse traditionnelles. Il nous paraissait donc indispensable, pour poursuivre nos recherches, de constituer un corpus de grande taille, de l'ordre d'une centaine d'œuvres au moins.

Ensuite, ces retours d'expériences suggéraient de traiter, au moins dans un premier temps, des œuvres concentrées sur une courte période historique, plutôt qu'étalées sur plusieurs siècles, comme nous l'avions envisagé initialement. Il nous fallait donc également choisir la période sur laquelle nous souhaitons porter notre étude des futurs envisagés. Dès les prémisses de ces travaux de recherche, nous souhaitions travailler sur de la littérature très contemporaine : le courant **cyberpunk**, qu'on a déjà présenté auparavant, intéressait particulièrement notre étude, en ce qu'il traite de thématiques et de technologies très actuelles – le cyberspace, notamment. Cependant, il ne s'agit pas d'un courant très représenté dans la littérature de science-fiction francophone, et le seul choix de ce sous-genre n'aurait pas suffi à constituer un corpus de taille suffisante. En outre, en France comme dans le monde anglophone, il s'agit encore d'une littérature très spécialisé et assez peu grand public, et la question de l'accessibilité des œuvres aurait sans doute posé quelques problèmes lors de la récupération des textes.

Ainsi, nous avons choisi d'élargir notre sélection en choisissant des œuvres issues d'autres sous-genres, tant que celles-ci procédaient de l'anticipation. Nous avons également allongé la période étudiée : alors que la littérature cyberpunk apparaît à partir du milieu des années 80, nous sommes allés puiser des œuvres du début des années 50 jusqu'à la fin des années 80 pour constituer ce corpus. On notera tout de même que la littérature étudiée reste très contemporaine, et respecte relativement notre volonté de se concentrer sur une période assez limitée. En outre, cette période se caractérise en France par de nombreuses mutations technologiques (électroménager, conquête spatiale, téléphonie mobile, débuts d'internet...) et socio-économiques (périodes des trentes glorieuses puis chocs pétroliers et tournant néo-libéral, émancipation des femmes, questionnements environnementaux), nourrissant les différents imaginaires associés aux projections futuristes.

Constituer un corpus d'analyse

A partir des éléments de réflexions présentés, nous avons finalement choisi deux collections emblématiques de la littérature de science fiction française : 'Anticipations', de la maison d'édition Fleuve noir, créée en 1951, et 'Présence du futur', des éditions Denoël, lancée en 1954. Revenons d'abord sur les deux collections et leur politique éditoriale, afin de comprendre l'intérêt d'y puiser des ouvrages pour constituer notre corpus. Pour ce qui est de la collection Denoël, il faut savoir qu'elle a d'abord été dirigée par Robert Kinters, traduisant et popularisant de nombreux auteurs anglo-saxons de science-fiction,

comme Brian Aldiss ou Philipp K. Dick, donc certaines œuvres majeures, comme le *cycle de Fondation*. Mais la collection publie aussi de jeunes auteurs français comme Jean-Pierre Andrevon, Serge Brussolo, Emmanuel Jouanne, Pierre Pelot ou Antoine Volodine, notamment sous l'impulsion d'Elisabeth Gille qui en prend la direction en 1976. L'accent est porté sur la science-fiction littéraire, mais la collection édite aussi de la **hard science-fiction** et du **cyberpunk**. De manière analogue, la collection anticipations se fait surtout, à ses débuts, la traductrice et la promotrice de la SF anglo-saxonne. Mais elle publie également de jeunes auteurs francophones, comme Jean-Louis et Doris Le May, ou Louis Thirion. Remarquons que la maison d'édition Fleuve Noir adopte à cette époque une politique encourageant la fidélité de ses auteurs – ces derniers n'étaient publiés qu'à partir du deuxième manuscrit sélectionné et devaient s'engager à ne pas publier ailleurs sous le même nom pour 5 ans, d'où les nombreuses œuvres publiées sous des pseudonymes, comme « Gilles d'Argyre » pour Gérard Klein.

Nous avons donc procédé en listant toutes les œuvres qui avaient été éditées dans ces deux collections entre leur création et la fin des années 80, en écartant les œuvres qui n'étaient pas des versions originales, puis en sélectionnant celles qui, après lecture de leur résumé, procédaient de la littérature d'anticipation. Evidemment, cette sélection introduit un biais, puisqu'elle semble quelque peu arbitraire : les œuvres ne sont pas toujours présentées comme des œuvres d'anticipation à proprement parler, et même si on a défini plus clairement les sous-genres qui relevaient de l'anticipation, ces sous-genres ne sont pas explicitement mentionnés. Cependant, nous nous sommes astreints au suivi de quelques critères pour tenter d'objectiver un tant soit peu cette sélection : de fait, nous avons écarté toutes les œuvres qui s'apparentaient au genre du space opera, et qui faisaient le récit d'aventures intergalactiques et de batailles interstellaires : elles n'intéressent pas cette analyse des imaginaires projectionnistes ; nous avons également écarté les œuvres qui relevaient du paradoxe temporel, c'est-à-dire qui racontaient des voyages dans le temps, et qui là encore n'ont pas vraiment à voir avec les futurs envisagés ; enfin, nous avons écarté les œuvres dans lesquels l'intrigue se déroulait dans des mondes extraterrestres très différents de la Terre, et qui n'avaient aucun lien avec l'Homme.

Nous souhaiterions ici faire remarquer que sur une liste d'œuvres de l'ordre de 2000 œuvres – tous les ouvrages publiés par ces collections, seule une petite partie des œuvres – environ 5% – remplissaient les critères mentionnés. Nous avons donc tout bonnement sélectionné toutes ces œuvres pour constituer notre corpus final. Ce corpus comptait également des recueils de nouvelles, et nous avons choisi d'en extraire les nouvelles qui nous paraissaient adéquates. Finalement, on comptait alors 112 œuvres et nouvelles dans notre corpus. Mais certains détails techniques nous ont conduit à affiner encore légèrement notre corpus. En effet, nous avons mentionné notre volonté d'avoir recours à une méthode informatique pour appréhender notre corpus, le *topic modeling*, sur laquelle nous reviendrons.

Or, il faut un minimum de l'ordre de 1000 mots par document pour que cette méthode soit efficace, et nous avons dû retirer une vingtaine de nouvelles qui n'étaient pas assez longues et qui auraient perturbé le fonctionnement de notre modèle.

Présentation du corpus final

Après ce premier traitement, nous disposons finalement d'un corpus constitué de 94 œuvres littéraires en versions originales françaises, toutes issues des deux collections présentées précédemment. Pour la liste exhaustive des différents ouvrages qui composent notre corpus, nous invitons nos lecteurs à consulter rapidement la bibliographie du corpus⁴⁵. Il convient ici de donner un aperçu des différents ouvrages qui composent notre corpus.

Il faut d'abord souligner la grande variété d'auteurs dont est issue ce corpus : on en compte 39, dont on ne fera pas le répertoire exhaustif ici. Néanmoins, il est intéressant de mentionner cette diversité, qui n'était pas forcément un objectif dans la constitution du corpus, mais qui se révèle une force lorsqu'on souhaite explorer les futurs envisagés. De fait, en s'intéressant à une plus grande variété de plumes, on accède à davantage de points de vue différents, qui sont autant d'images et de questionnements d'intérêt pour notre analyse. Remarquons également que pour la collection 'Anticipations', de nombreux ouvrages sont publiés sous des alias pour les raisons qu'on a évoquées plus haut. Ces alias correspondent au pseudonyme d'un auteur ou, parfois, d'un collectif d'auteurs : on citera notamment « Christopher Stork », l'alias choisi par un collectif d'auteurs belges pour publier certaines œuvres, à l'instar de *L'ordre établi* (1979), ou encore François Richard Bessière, l'alias de François Richard et Henri Bessière. Aussi, il faut réinsister sur la diversité autorale de notre corpus, non seulement sur le plan du nombre d'auteurs, mais également sur celui des modalités de composition des romans, parfois écrits collectivement.

On compte aussi une grande diversité dans la forme et la longueur des textes. On dispose ainsi de 38 nouvelles, dont les plus courtes font à peu près un millier de mots, et de 56 romans, dont le plus long fait environ 79 000 mots. Sous cette forme, on a donc un écart 1 à 80 entre le texte le plus court le plus long, ce qui est encore une fois intéressant pour la diversité de notre corpus, mais soulève, il faut le dire, certains problèmes méthodologiques pour l'implémentation du *topic modeling*. Nous reviendrons sur les problèmes d'harmonisation de la longueur des textes au moment d'expliquer plus en détail notre méthode. Notons toutefois qu'une fois le prétraitement de nos données textuelles réalisé, cet écart de longueur est moins significatif, et qu'en ayant recours à des techniques d'échantillonnage, nous pouvons résoudre la question de la longueur des textes de manière relativement efficace.

45. Voir Bibliographie du corpus en Annexe 1.

Donnons enfin un bref aperçu de la composition littéraire et thématique de ce corpus, en revenant plus en détail sur certaines des œuvres majeures. Là encore, on remarque qu'il est composé aussi bien d'œuvres majeures, qui a posteriori, apparaissent comme de véritables monuments de la SF francophone, que d'œuvres nettement moins connues du grand public. Se côtoient ainsi de petites nouvelles assez peu remarquées, enfouies au sein de larges recueil de nouvelles, comme *Univers-Code* de Jean-Claude Dunyach ou *Soleil de soufre* de Serge Brussolo, aussi bien que des œuvres majeures, parmi lesquelles on pourra notamment citer *Niourk*, de Stéphane Wul. Du point de vue thématique, on a aussi une grande profusion de sujets abordés et de questionnements. Une partie importante de notre corpus interroge des phénomènes apocalyptiques, liés à l'effondrement des sociétés humaines. On peut évoquer, de manière non exhaustive, certains thèmes récurrents qui concernent comme les pandémies, avec par exemple *L'ère des biocybes* de Jimmy Guieu (1960) et *Pandémoniopolis* de Gabriel Jan (1975), le climat, avec *La sortie est au fond de l'espace* de Jacques Sternberg (1956) ou *La Mort de la vie* de Jimmy Guieu (1957) ou encore les guerres robotiques, avec *Territoire robot* de Jean-Gaston Vandel (1954). D'autres ouvrages abordent des thématiques relatives à l'avancée scientifique et l'expansion de la civilisation humaine : reviennent notamment les thèmes de la conquête spatiale, de la ville du futur ou des avancées scientifiques, au sujet desquels on trouvera diverses œuvres dans les recueils de nouvelles *Vue en coupe d'une ville malade* de Serge Brussolo (1980) et *Superfuturs*, de Philippe Curval (1986). Si d'autres thèmes sont également abordés, l'ensemble ainsi décrit résume assez bien les motifs récurrents et questionnements abordés au sein des œuvres qui composent notre corpus.

En fin de compte, on dispose d'un corpus qui rend assez bien compte de la diversité de la scène SF française de notre période, avec différents types d'ouvrages, sur le plan de la forme comme sur celui du fond. Le corpus prend en compte les éléments de réflexion qu'on a pu formuler en vue de sa constitution, et il est en ce sens un support d'investigation adapté à l'étude des futurs envisagés. Notons toutefois qu'on aurait aimé pouvoir agrandir encore la taille de notre corpus, mais que seule une part assez restreinte de la littérature d'anticipation francophone, du moins pour les collections qu'on a utilisées, semble correspondre aux critères de sélection établis. Néanmoins, le corpus ainsi composé constitue déjà un matériau solide pour explorer les imaginaires projectionnistes. Il nous faut donc maintenant expliquer la méthode informatique que nous avons choisie pour explorer sémantiquement ce corpus, et sonder les futurs envisagés.

Chapitre 2

Construire une méthode computationnelle pour explorer notre corpus

En 2017, un article de Guillaume Carbou intitulé « Analyser les textes à l’ère des humanités numériques. Quelques questions pour l’analyse statistique des données textuelles » paraissait dans *Les Cahiers du numérique*¹. Faisant état de l’intérêt des outils statistiques et des méthodes d’analyse informatiques, ce dernier prônait « la reconnaissance de la dimension herméneutique des humanités numériques ». Mais l’article de Guillaume Carbou s’attachait surtout à poser les questions méthodologiques que soulèvent une telle démarche, montrant que le traitement automatique de grands volumes de données nécessitait un effort épistémologique de réflexion sur l’objet étudié, et la manière de l’aborder sur le plan de l’analyse. Actant l’apport majeur des humanités numériques, nous souhaitons employer ces méthodes afin d’appréhender les futurs envisagés à travers la littérature d’anticipation. Nous allons donc par aborder la méthode que nous proposons d’utiliser, le *topic modeling*, d’un point de vue scientifique et technique. Puis, nous présenterons le module que nous avons développé, avant de faire le compte rendu de son utilisation sur nos données, en évoquant les différentes considérations épistémologiques auxquelles nos données ont pu nous mener.

1. CARBOU Guillaume. « Analyser les textes à l’ère des humanités numériques. Quelques questions pour l’analyse statistique des données textuelles », *Les Cahiers du numérique*, 2017/3 (Vol. 13), p. 91-114. URL : <https://www.cairn.info/revue-les-cahiers-du-numerique-2017-3-page-91.htm>.

2.1 Le *topic modeling*

Qu'est ce que le *topic modeling* ?

L'appellation *topic modeling* regroupe un ensemble de modèles probabilistes génératifs, qui fournissent des méthodes pour analyser la fréquence d'occurrence de mots et de thèmes dans un corpus de documents. On note ici que ces méthodes se réfèrent uniquement à la fréquence d'occurrence des mots, leur position dans le document étant parfaitement négligée. On parlera de modèles *bag of words* : on considère chaque document comme un « sac de mots », c'est-à-dire un ensemble faisant abstraction de l'ordre d'apparition des mots. Le *topic modeling* se perfectionne et s'affirme depuis une vingtaine d'années comme l'approche la plus pertinente pour le traitement thématique de données textuelles. Pour mieux cerner ce concept complexe, et pour avoir un meilleur aperçu des méthodes qui sont à notre disposition, revenons brièvement sur l'histoire des principaux modèles et de leur apport.

Dans les années 1990, avec l'arrivée d'internet, la question de la classification de grandes masses de données s'impose avec force. De premiers modèles de traitement statistique du lexique voient le jour : on pense notamment à la Latent Semantic Analysis (LSA) en 1990², et aux méthodes qui se fondent sur le modèle unigramme³. Ces derniers modèles sont, en effet, limités en ce qu'ils attribuent à chaque document un thème unique, et se révèlent incapables de traiter le problème de l'appartenance simultanée de certains mots à plusieurs thèmes, et celui des documents comportant plusieurs thèmes. C'est pour tenter d'apporter une réponse à ces problèmes que de véritables modèles de traitement thématiques sont élaborés. Un proto-modèle de *topic modeling* est proposé par Hofman en 1999, lorsqu'il développe la LSA probabiliste (PLSA)⁴. Mais la Latent Dirichlet Allocation (LDA) qui, en 2003⁵, s'affirme véritablement dans le monde universitaire comme un outil de *topic modeling* pertinent et fonctionnel. Cette méthode permet à la fois de révéler des structures thématiques cachés au sein de grands corpus, et d'analyser les fréquences d'occurrence de ces thèmes au sein de chaque document. Différentes extensions du modèle LDA seront ensuite présentées, et on citera notamment le *correlated topics model* (CTM) en 2007⁶ qui autorise les corrélations entre différents thèmes. La LDA, et les différentes

2. DEERWESTER, Scott, DUMAIS, Susan, FURNAS, George, LANDAUER, Thomas, HARSHMAN, Richard. « Indexing by latent semantic analysis », *Journal of the american society for information science*, 1990.

3. NIGAM, K., MCCALUM, A.K., THRUN, S. et al. « Text Classification from Labeled and Unlabeled Documents using EM », *Machine Learning* 39, 103–134 (2000). URL : <https://doi.org/10.1023/A:1007692713085>.

4. HOFFMANN, Thomas. « Probabilistic latent semantic analysis », *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, July 1999 Pages 289–296.

5. BLEI, David M., NG, Andrew Y., JORDAN Michael I. « Latent Dirichlet Allocation », *Journal of Machine Learning Research* 3 (2003) 993-1022.

6. BLEI, David M., LAFFERTY, John D. « A CORRELATED TOPIC MODEL OF SCIENCE »,

méthodes qui se fondent sur ses principes, sont aujourd’hui le meilleur moyen d’analyser thématiquement de vastes ensembles littéraires ; en ce sens, c’est celles que nous retiendrons dans le cadre de nos travaux, et il convient ici de rentrer davantage dans le détail de leur fonctionnement.

Utiliser la LDA pour étudier les futurs envisagés

La LDA, on l’a dit, est une des méthodes les plus utilisées lorsqu’on cherche à faire du *topic modeling*. Elle offre des solutions intéressantes en matière d’organisation de documents par thèmes et d’analyse du texte. Sans rentrer dans les mathématiques complexes qui se cachent derrière ces méthodes, tâchons ici de comprendre son fonctionnement. La LDA est fondée sur deux principes majeurs. D’abord, chaque document est un mélange de thèmes (*topics*), présents en différentes proportions. Ensuite, chaque *topic* est un mélange de mots, eux-aussi, présents en différentes proportions dans chaque document. De manière intéressante, on remarque que les mots peuvent appartenir simultanément à plusieurs thèmes. La LDA est une méthode mathématique qui permet de trouver le mélange de mots qui correspond à chaque thème, et le mélange de thèmes qui décrit chaque document.

Sans rentrer dans des détails trop techniques, approfondissons tout de même l’explication mathématique de ce modèle, avant de donner un exemple plus parlant. Nous disposons d’un ensemble de documents textuels, notre corpus d’oeuvres d’anticipation, chaque document étant défini par l’ensemble de mots qui le composent. La LDA suppose que chaque document correspond au mélange d’un petit nombre de *topics*, et que la génération de chaque occurrence d’un mot est attribuable à l’un des thèmes du document. Lorsqu’on entraîne un modèle de LDA, on commence par une phase d’initialisation, en attribuant un *topic* à chaque mot de chaque document, selon une distribution de Dirichlet sur un ensemble de N *topics*. Notons donc qu’il faut choisir un nombre de thèmes lors de l’implémentation de la LDA. Ce premier *topic model* est assez peu cohérent, puisqu’il est généré aléatoirement. Intervient alors la phase d’apprentissage au cours de laquelle on cherche à affiner notre modèle. Pour cela, dans chaque document, on prend chaque mot et on met à jour le thème auquel il est lié, en lui attribuant le *topic* qui a la plus forte probabilité de générer le mot donné dans ce document. On répète ces étapes un grand nombre de fois : c’est le nombre d’itérations, qu’il faut également définir lors de l’implémentation de la LDA. À l’issue de cette phase d’apprentissage, on obtient un modèle de sujets affiné, qui comporte la probabilité d’occurrence des *topics* présents dans chaque document – on compte le nombre de représentation de chaque thème dans chaque document, et les probabilités d’occurrence des mots associés à chaque thème – on compte les mots associés à chaque thème dans l’ensemble du corpus.

Pour mieux comprendre ce procédé, prenons l'exemple d'un corpus d'articles de journaux, dans lequel on aurait un premier article relatif à la politique et un second article relatif aux sciences dures. Le premier article pourrait ainsi contenir des mots tels que « référendum », « gouvernement », « parlement », « débat ». Le second article, pourrait quant à lui contenir des mots tels que « équations », « physique », et éventuellement « débat ». En appliquant nos méthodes à ces deux documents, on pourrait donc voir émerger deux *topics*, l'un relatif aux sciences, l'autre à la politique et tous deux contenant notamment les mots leur étant respectivement associés. Maintenant, imaginons qu'un troisième article relate un débat parlementaire, au cours duquel auraient été abordées, entre autres, des problématiques scientifiques. Notre méthode permettrait ici d'attribuer une proportion d'utilisation de chaque *topic* à ce troisième article : ainsi, si l'article invoque principalement un lexique politique, et de manière plus brève, un lexique scientifique, on pourrait grossièrement obtenir quelque chose comme « l'article 3 relève à 85% du *topic* 1, et à 15% du *topic* 2 », le *topic* 1 et le *topic* 2 pouvant, par le travail d'interprétation du chercheur, être caractérisés : le *topic* 1 renvoie à la politique, le *topic* 2 renvoie aux sciences.

Au regard de cet exemple, on imagine aisément l'intérêt de la méthode lorsqu'on souhaite étudier les futurs envisagés à travers la littérature SF d'anticipation. La double approche nous permet d'envisager deux niveaux d'analyse complémentaires. D'abord au niveau du corpus, en repérant des *topics* récurrents au sein d'un corpus constitué de science fiction, nous pouvons faire émerger des thèmes récurrents, correspondant à autant de points de crispation, de considérations qui reviennent souvent lorsque les auteurs du corpus se projettent dans le futur. Cette analyse, menée sur différents corpus, permettra aussi d'apprécier la manière dont ces considérations évolue en fonction des époques, des régions, ou encore des sous-genre de la SF. Ensuite au niveau des textes, elle permet de faire ressortir les mots qui sont utilisés, dans chaque texte, pour évoquer chaque thème. Cette analyse nous permettra, ici, d'analyser les mots choisis par les auteurs pour parler de chacune de ces considérations, là encore, en considérant les différents paramètres que sont le temps, l'espace, et le sous-genre.

Le *topic modeling* appliqué à des corpus littéraires

Qu'en est-il de l'utilisation du *topic modeling* sur des corpus littéraires dans le monde académique ? Existe-t-il des cas précédents d'utilisation sur de la littérature de Science-fiction, ou de manière plus générale, sur des corpus d'oeuvres littéraires ? Le cas échéant, ces recherches avaient-elles également pour but de sonder les représentations ?

Après avoir parcouru longuement parcouru les principales revues et actes de congrès en Humanités Numériques (DSH, DHQ, actes du congrès DH), nous devons souligner l'absence d'exemples d'application de *topic modeling*, ou même plus généralement de méthodes

de Distant Reading, sur des corpus de littérature d'anticipation. À défaut de trouver des exemples sur des corpus de Science Fiction, nous pouvons ici évoquer quelques travaux majeurs dont les perspectives se rapprochent des nôtres, sur d'autres genres littéraires.

Pour ce qui est de l'étude des représentations à proprement parler, il n'existe que peu d'exemples d'utilisation des méthodes de Distant Reading à ces fins. On évoquera ici des travaux assez méconnus, quoique particulièrement intéressants, sur la l'imagination littéraire et les représentations liées à l'Ancien Monde dans la fiction du XIXème siècle⁷. Dans ce compte-rendu allégé d'un cours donné à la première réunion de la *Digital Classicists Association*, Matthew L. Jokers fait l'exposé de recherches mêlant exploration thématique, géographique et sentimentale des représentations liées à l'Ancien Monde, usant notamment du *topic modeling* comme d'une méthode d'exploration thématique. On a ici un exemple intéressant, bien qu'esseulé, d'utilisation du *topic modeling* pour sonder les représentations dans des corpus de littérature. En fait, pour trouver des travaux dont les perspectives se rapprochent des nôtres, il faut bien souvent décaler un peu le regard, en s'écartant de l'exploration thématique des représentations. Lorsqu'on s'intéresse aux exemples d'application du *topic modeling* sur des corpus littéraires, en effet, on trouve nettement plus de recherches liées à des problèmes de distinction et de classification thématiques.

En particulier, on pourra citer l'article remarqué de Christop Schöch sur les genres du théâtre classique⁸. Dans cette contribution le *topic modeling* est utilisé pour analyser un corpus de pièces de théâtre françaises de l'âge classique et de la Renaissance. De manière remarquable, Schöch y mène une démonstration de l'intérêt du *topic modeling* comme méthode d'analyse quantitative des genres pour les textes de cette période. L'étude se fonde sur un corpus de 391 pièces publiées entre 1610 et 1810, représentatives des différents genres traditionnellement attribués aux pièces de l'époque : la comédie, la tragédie et la tragi-comédie. Les recherches de l'auteur parviennent justement à mettre en évidence l'utilité du *topic modeling*, non seulement pour explorer sémantiquement les thématiques abordées dans ces œuvres, mais aussi pour distinguer thématiquement les sous-genres de l'époque à l'aide de critères quantitatifs. Fort de résultats probants, Christop Schöch parvient à opérer une classification de ces sous-genres en distinguant différents schémas thématiques dominants pour chacun d'entre-eux. En comparant ses résultats à une classification conventionnelle des sous-genres de la période, l'auteur confirme quantitativement et nuance des résultats qui avaient été établis par des méthodes d'analyse plus classiques.

7. JOCKERS, Matthew L. « The Ancient World in Nineteenth-Century Fiction; or, Correlating Theme, Geography, and Sentiment in the Nineteenth Century Literary Imagination », University of Nebraska, *DHQ*, Vol. 10 | n°2, 2016. URL : <http://digitalhumanities.org:8081/dhq/vol/10/2/000250/000250.html>.

8. SCHÖCH, Christof. « Topic Modeling Genre : An Exploration of French Classical and Enlightenment Drama », University of Würzburg, *DHQ*, Vol. 11 | n°2, 2017.

URL : <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html#politz2015>.

Plus récemment, on pourra citer un article paru en 2019 dans *digital humanities quarterly* sur une étude computationnelle du *Quan Tang Shi*⁹, anthologie monumental de la poésie tang, produite tout au long du règne de la dynastie Qing (1644-1912) sur la Chine. Le corpus analysé est constitué de plus de 50 000 poèmes et fragments de poème. De fait, ce volume rend difficile une analyse des textes pour l'échelle humaine, et l'article contribue à démontrer l'intérêt des méthodes de distant reading, et en particulier du *topic modeling*, pour cadrer et penser les questions traditionnelles posées par les études d'histoire littéraire sur le *Quan Tang Shi*, et apporte une nouvelle perspective de réponse sur ce que signifie véritablement lire la poésie tang.

De cette brève revue de l'état de l'art en matière de *topic modeling* appliquée à des corpus littéraires, on retiendra donc qu'il existe assez peu d'exemples qui se servent de ces méthodes pour sonder les représentations. Cependant, de nombreuses études, dont on aura cité ici qu'une infime part, montrent l'intérêt bien fondé du *topic modeling* pour des problématiques liées à la classification des textes, et pour changer changer d'échelle dans l'analyse de très vastes corpus littéraires.. D'ailleurs, puisqu'on a longuement évoqué la diversité de la SF et la difficulté à la classifier en sous-genres, il faut souligner qu'utiliser le *topic modeling* à des fins de classification thématique du genre serait une piste intéressante à suivre. Mais il faut bien comprendre que nos travaux se distinguent clairement de cette perspective. Nous cherchons ici explorer thématiquement notre corpus à des fins interprétatives, et non typologiques.

9. BROADWELL, Peter, SHEPARD, David, CHEN, Jack W. « Reading the *Quan Tang shi* : Literary History, Topic Modeling, Divergence Measures », DHQ, Vol. 13 | n°4, 2019. URL : <http://digtalhumanities.org:8081/dhq/vol/13/4/000434/000434.html>.

2.2 Construction d'un module

Nous avions déjà pu utiliser le *topic modeling* sur un prototype expérimental dans le cadre des recherches menées en M1. Les retours plutôt positifs suggéraient de poursuivre dans cette voie, en exploitant néanmoins davantage les possibilités de visualisation des résultats. Nous avons donc décidé de mettre au point un module pour faciliter la mise en oeuvre du *topic modeling* et développer des solutions de visualisation plus adaptées à l'interprétation de nos résultats. Aussi, la création de notre module devait obéir à un cahier des charges défini de manière claire et exigeante, que nous allons maintenant présenter.

Preprocessing des données

Avant toute chose, l'implémentation de la LDA nécessite donc un travail de nettoyage de notre corpus et de lemmatisation des documents. On a besoin d'une liste de tous les mots pour chaque texte, sous une forme lemmatisée. Notre module comporte donc une composante dédiée au preprocessing des données, permettant d'abord de *tokeniser* les données, puis de les lemmatiser, et enfin, d'opérer des opérations de filtrage des mots pour retirer les éléments qui n'intéressent pas notre analyse.

Pour traiter nos données textuelles, il faut d'abord tokeniser notre corpus, c'est-à-dire convertir l'ensemble de nos documents textuelles en des listes de mots, en parsant les différents éléments du texte. Cette opération est assez facile, mais certains problèmes apparaissent rapidement lorsqu'on traite de gros volumes de textes édités pour l'impression, notamment au niveau des tirets en fin de ligne et des mots composés. Notre module utilise le modèle de Spacy, auquel nous avons adjoint une fonction utilisant les expressions régulières pour améliorer le parсage des mots composés et traiter les problèmes de tirets en fin de ligne. Après avoir obtenu une liste de *tokens*, il faut lemmatiser notre corpus. La lemmatisation est une opération qui consiste à rapporter chacun des mots du texte à une forme neutre, canonique. On rapporte les formes conjuguées des verbes à leur radicaux où à une forme infinitive, les mots au pluriel sont mis au singulier, etc. Les lettres capitales sont également mises en lettres minuscules. De nombreuses bibliothèques de fonctions proposent des méthodes efficaces pour lemmatiser les textes en langue anglaise ; en revanche, l'opération est plus délicate et moins bien renseignée pour les textes en français. Là encore, nous utilisons les modèles proposés par le module Spacy, et notamment le modèle le plus lourd performant, ‘fr_core_news_lg’. Nous avions déjà utilisé un modèle plus allégé de spacy, ‘fr_core_news_sm’ dans le cadre de notre approche expérimentale en M1. Au cours des recherches de cette année, nous avons pu comparer les performances des différents modèles de spacy, ainsi que d'un modèle concocté par des chercheurs de l'Ecole nationale des Chartes, pie-extended. Ce dernier semblait plus adapté à la langue française mais, entraîné sur un corpus francophone de l'époque classique, il avait beau-

coup de mal à traiter notre corpus de littérature d’anticipation très contemporaine. En outre, les récentes mises à jour du module Spacy ont permis des progrès considérables des performances de lemmatisation de leurs modèles, nous déterminant finalement à utiliser ces derniers.

A l’issue de la phase de lemmatisation, non obtenons donc une liste de lemmes pour chaque texte de notre corpus. Mais il subsiste de nombreux lemmes qui n’intéressent pas notre analyse thématique : aussi, il nous faut intervenir pour écarter de l’analyse ces éléments. Ils se divisent en deux catégories. D’une part, les Stopwords, ou mot-outils, c’est-à-dire les mots dont le rôle syntaxique est important, mais qui n’ont aucune importance sémantique : prépositions, conjonctions, pronoms, déterminants, etc. On dispose d’une liste de *stopwords* pour la langue française, dont on va se servir comme d’un filtre sur notre corpus lemmatisé. De même, on va exploiter les prédictions morphosyntaxiques (POS) du modèle Spacy pour retirer tous les éléments qui ne nous intéressent pas (pronoms, déterminants, adverbes, etc). D’autre part, les mots de narration, qui là encore ne sont pas porteur de beaucoup de sens d’un point de vue thématique. Il s’agit ici de toutes les entités nommées propres à l’intrigue des ouvrages – noms de personnages, de lieux, dates, etc – ainsi que des verbes de narration, qui servent à décrire l’action mais qui disent peu du fond thématique de nos textes. De surcroît, leur grande fréquence d’occurrence peut biaiser notre modèle de *topic modeling*. Ici, on utilise les fonctions de reconnaissance d’entités nommées proposées par le modèle de Spacy pour retirer ces entités.

Grâce à l’utilisation du modèle de Spacy, nous obtenons de bonnes performances de tokenisation et de lemmatisation, qui permettent d’améliorer considérablement les résultats des analyses de *topic modeling*. Empiriquement, après avoir essayé différents lemmatiseurs, celui de Spacy est un de ceux qui fonctionne le mieux pour la langue française, ou du moins pour notre corpus. À cet égard, il est nécessaire de souligner que c’est un corpus de littérature d’anticipation très contemporaine, et que certains auteurs écrivent parfois de sultats des analyses de *topic modeling* sont dans un premier temps assez peu pertinents. Il est donc nécessaire de refiltrer manuellement notre liste de lemmes, en utilisant une liste de mots de narration constituée *ad hoc*. Or, la constitution de cette liste de lemmes constitue, de fait, un problème pour la généralisation de notre module de *topic modeling*.

Enfin, évoquons ici le problème de l’échantillonnage de nos données. Une fois nos données pré-traitées, et prêtes à l’emploi pour l’implémentation de la LDA, il peut tout de même subsister un problème d’inégalité de longueur des textes. Or, cela constitue un problème pour l’analyse de *topic modeling*, très sensible à la longueur des documents. Aussi, nous avons ajouté à notre module des fonctions d’échantillonnage pour traiter ces cas-là. Concrètement, c’est fonctions permettent de tirer au sort des échantillons de tailles égales pour chaque document du corpus. Remarquons ici qu’il est important d’effectuer

plusieurs analyses de validation sur différents tirages au sort pour confirmer une tendance.

Implémenter la LDA à l'aide de différentes méthodes

Après avoir préparé et échantillonné nos données, nous pouvons entrer dans la phase d'implémentation de la LDA. Nous cherchons d'abord à créer un bag of words à partir de l'ensemble des mots de notre corpus. Concrètement, ce bag of words se présente sous la forme d'une liste de tuples, contenant l'ID de chaque mot et son nombre d'occurrences : c'est-à-partir de ce format que nous pouvons implémenter la LDA. Comme on l'a expliqué précédemment, de nombreuses méthodes existent. Là encore, après une phase d'essais, nous avons retenu les deux méthodes dont les résultats semblaient les plus probants.

Elles se fondent toutes deux sur les fonctions proposées par la bibliothèque gensim. Intéressons-nous de plus près aux fonctions utilisées : concrètement, elles entraînent la machine sur notre bag of words en lui apprenant à classer les mots en *topics*. Initialement, nous avions appliqué cette méthode aux données brutes de notre bag of words. Mais cette fonction retournait, lors de nos premiers essais, des résultats assez étonnantes, et peu cohérents. Nous avons tout de même choisi de conserver la possibilité d'utiliser cette fonction dans notre module puisqu'il semble qu'elle puisse mieux fonctionner avec certaines données. Cependant, nous avons également mis au point une deuxième fonction d'implémentation de la LDA, qui se fonde sur un bag of words dans lequel chaque mot est pondéré grâce à la méthode TF-IDF – pour term frequency / inverted document frequency. Cette méthode permet de mettre en relief les mots discriminants d'un texte au sein d'un corpus. Elle s'est révélée nettement plus probante dans le cas de notre prototype d'expérimentation de M1.

On note ici avec intérêt que notre méthode permet de jouer sur plusieurs paramètres pour affiner notre modèle. D'abord, le nombre de *topics* que la machine est censée repérer. Il n'y a pas vraiment de règles en la matière ; aussi, nous avons mis au point une fonction d'assistance au nombre de choix de *topics*, fondée sur la cohérence du modèle pour différents nombres de *topics*. Elle permet d'obtenir un graphique d'aide au choix du nombre de *topics*, bien que cette décision ne puisse se passer du regard éclairé du chercheur. Ensuite, et c'est là un facteur essentiel, le nombre de passes, c'est-à-dire le nombre d'itérations que la machine va effectuer sur notre jeu de données. Plus la machine s'entraîne, plus les *topics* qu'elle détermine sont pertinents, et mieux elle pourra apprécier chacun des textes au regard de ces *topics*. Ici, on n'est limité que par le temps de traitement de l'ordinateur. Il faut ici souligner qu'à partir d'un certain nombre d'itérations, les résultats semblent se stabiliser. Un nombre trop important d'itérations serait trop coûteux en temps, et n'apporterait finalement rien de plus à notre modèle. On peut également jouer sur les données d'entrées, puisque notre module permet de convertir notre liste de lemmes en liste de n-grammes - c'est-à-dire, d'ensemble de lemmes de longueur n. Par

exemple, si nous souhaitions extraire les trigrammes de la phrase « Nous sommes allés à la plage hier soir » qui, lemmatisée donnera « Nous être aller plage hier soir », nous obtiendrions « nous-être-aller, être-aller-plage, aller-plage-hier, plage-hier-soir ». Or, les n-grammes permettent parfois d'aboutir à des résultats particulièrement probants dans le cadre d'analyses de LDA. Enfin, on dispose dans notre module d'une fonction d'évaluation de la cohérence de notre modèle après entraînement de celui-ci. Ce score de cohérence permet de mesurer la performance générale de notre modèle et éventuellement, de le comparer à d'autres modèles, même si la mesure de la cohérence doit être prise avec beaucoup de prudence.

Offrir des solutions de visualisation de données adaptées à nos résultats

Notre module doit finalement offrir de nombreuses solutions de visualisation de données permettant d'interpréter nos résultats, et de les faire parler de manière plus aboutie. Nous souhaitons notamment implémenter différents types de visualisations permettant d'explorer autant les résultats à l'échelle des *topics* eux-même qu'à une échelle inter-*topic*, pour l'ensemble du corpus ou pour un élément spécifique de ce dernier. Pour répondre à cette nécessité, nous avons choisi d'implémenter différentes fonctionnalités que nous allons ici présenter.

On trouve en premier lieu des visualisations assez simples permettant d'afficher nos résultats de manière plus parlante que la sortie console du programme. On mentionnera ici la possibilité de produire des dataframes, c'est-à-dire des tableaux de données, permettant de synthétiser d'une part les informations relatives aux *topics* établis par notre modèle sur l'ensemble du corpus, et d'autre part celles relatives à l'attribution de ces différents *topics* à chacun des éléments du corpus. Nous avons également ajouté une fonction permettant produire des graphiques à partir des résultats de notre modèle, et notamment d'un graphique permettant de visualiser le nombre d'occurrences de chaque mot et leu poids relatif au sein des différents *topics*. Enfin, le module permet de produire des *wordclouds*, c'est-à-dire des nuages de mots de chaque *topic*, dans lesquels la taille des mots est variable et correspond à l'importance relative du mot au sein du *topic*.

Dans un second temps, nous avons travaillé sur des fonctionnalités de visualisation nettement plus avancées. D'abord, avec le T-SNE Clustering, pour *t-distributed stochastic neighbor embedding*. Derrière ce nom assez complexe se cache une méthode de classification assez originale, qui permet, à l'instar d'une Principal Component Analysis (PCA), de réduire la dimensionnalité d'un problème complexe et d'en offrir une visualisation en deux dimensions nettement plus manipulable. Dans notre cas, le T-SNE Clustering permet de visualiser en deux dimensions la distance thématique entre les différents éléments du corpus, à partir des résultats obtenus par notre modèle. Cette solution de visualisation ouvre de nombreuses perspectives pour évaluer la qualité de notre modèle, en ce qu'elle permet

de porter un regard sur l’homogénéité des *topics* obtenus et sur la présence d’éventuels *outliers*, c’est-à-dire d’éléments très éloignés du reste du corpus, qui pourrait attirer notre attention. Ici, il est d’ailleurs intéressant de noter qu’un tel résultat doit permettre un aller retour plus fin entre notre méthode de distant reading et le corps des textes étudiés, justement en focalisant notre regard sur les éléments situés à la marge du corpus.

Enfin, la dernière fonctionnalité qu’il convient de présenter est LDAvis, un outil particulièrement puissant et adapté à la visualisation de résultats obtenus avec une analyse de LDA. LDAvis est un package permettant une visualisation interactive de notre *topic model* en ligne. Elle permet notamment, de manière très intéressante, de modéliser la distance inter-*topic* en deux dimensions ainsi que le poids respectif de chacun des *topics* sur l’ensemble du *topic*. LDAvis permet également de zoomer sur chacun des *topics*, en observant les 30 mots les plus déterminants du *topic* ainsi que leur poids relatif. En outre, elle permet de faire varier certains paramètres en explorant nos *topics*, afin d’aller plus en profondeur sur l’appréciation de la répartition des différents *topics*. C’est sans doute l’outil le plus abouti dont nous disposons pour visualiser et interpréter nos résultats.

Conclusions sur notre module

Pour conclure, nous disposons désormais d’un module pleinement opérationnel et facile d’utilisation, offrant une base solide pour nos recherches dans le cadre de ce mémoire. En permettant de centraliser et de faciliter l’ensemble de la chaîne du traitement des données d’un corpus littéraire, du preprocessing des données jusqu’à la visualisation des résultats des analyses de *topic modeling*, ce module est parfaitement adapté aux analyses que nous souhaitons produire sur notre corpus.

En outre, il doit permettre de faciliter et de populariser l’utilisation du *topic modeling* auprès d’autres chercheurs en Humanités Numériques. C’est notamment le cas au sein du Master de l’Ecole des Chartes, où plusieurs élèves ont déjà réemployé ce module dans le cadre de leurs propres recherches. D’ailleurs, le script du module est disponible en open source en ligne, sur Github¹⁰, et se veut accessible et consultable par tout le monde. Il a d’ailleurs vocation à être amélioré collectivement, et nous invitons tout lecteur intéressé à consulter le lien indiqué et à ne pas hésiter à émettre remarque et propositions de modification.

10. Voir <https://github.com/leo8/ModuleTM>.

2.3 Elaboration d'un modèle entraîné sur les données de notre corpus

Récupération et normalisation des textes

A l'issue de la phase de constitution de notre corpus, nous avons pu commencer à récupérer les textes et à pré-traiter les données textuelles, l'objectif étant de parvenir à des formes normalisées de nos textes, au format .txt, pour toutes les œuvres de notre corpus. Pour les recueils de nouvelles, nous avons choisi d'extraire chaque nouvelle pour en faire un nouveau document. Après ce premier traitement, nous obtenions un corpus constitué de 94 documents, soit 94 œuvres littéraires toutes issues des deux collections évoquées. L'objectif était alors de parvenir à des formes normalisées de nos textes, au format .txt, pour toutes les œuvres de notre corpus. Pour ce qui est des ouvrages complets, nous avons simplement retiré manuellement les préfaces, tables des matières, mention de ‘chapitres’ et de ‘parties’ – en prenant soin de conserver leur titre. Pour les recueils de nouvelles, en revanche, nous avons choisi d'extraire chaque nouvelle pour en faire un nouveau document. Nous avons ensuite appliqué le même traitement manuel que pour les ouvrages. Ici, il faut également mentionner le retrait d'une vingtaine de nouvelles, qui étaient trop courtes pour que le *topic modeling* soit efficace – il faut un minimum de lemmes par document, de l'ordre de 1000, que ces nouvelles ne permettaient pas d'atteindre.

Il faut souligner la grande variété d'auteurs dont est issue ce corpus – on en compte 39 – qui n'était pas forcément un objectif dans la constitution du corpus, mais qui se révèle une force lorsqu'on souhaite explorer les futurs envisagés. Remarquons également que la longueur des différents documents qui composent le corpus est très variable : de l'ordre de 1 à 20, ce qui constitue un problème pour l'analyse de *topic modeling*, très sensible à la longueur des documents. Aussi, nous avons souhaité enrichir la phase de lemmatisation de nos documents d'un processus d'échantillonage. Concrètement, ces fonctions permettent de tirer au sort des échantillons de tailles égales pour chaque document du corpus. Mais sous cette forme, les données textuelles doivent encore être pré-traitées avant de pouvoir utiliser l'algorithme de LDA.

Echantillonage et lemmatisation

Avant d'implémenter la LDA, il faut tokeniser et lemmatiser nos documents, c'est-à-dire transformer nos textes bruts en listes de mots – tokens – et rapporter tous ces mots à une forme racine et invariable – lemmes – afin de compter le nombre total d'occurrences de ces mots sous leurs différentes formes. Pour ce faire, notre module dispose de deux fonctions distinctes, faisant appel à deux modules différents : spacy et pie-extended. Initialement, nous étions utilisés le lemmatiseur de pie-extended, qui obtient des résultats

nettement meilleurs, même s'il est entraîné sur du français classique. Mais l'utilisation de pie-extended reposant largement sur la POS, il convient de tokeniser au niveau des phrases pour l'utiliser correctement. Ainsi, les premières phases d'essais de lemmatisation ont mené à des résultats assez étranges – tendance, entre autres, à rapporter les noms propres et les noms communs finissant par -a à une forme infinitive, si bien qu nous avons décidé de revenir au lemmatiseur de spacy qui obtenait des résultats convenables. Il faudra envisager, dans un second temps, de revoir la phase de tokenisation afin de permettre l'utilisation du lemmatiseur de pie-extended et de comparer les résultats obtenus à ceux de spacy.

Après cette première phase, il subsiste de nombreux lemmes qui n'intéressent pas notre analyse thématique : aussi, il nous faut intervenir pour écarter de l'analyse ces éléments. Ils se divisent en deux catégories. D'une part, les Stopwords, ou mot-outils, c'est-à-dire les mots dont le rôle syntaxique est important, mais qui n'ont aucune importance sémantique : prépositions, conjonctions, pronoms, déterminants, etc. D'autre part, les mots de narration, qui là encore ne sont pas porteur de beaucoup de sens d'un point de vue thématique. Il s'agit ici de toutes les entités nommées propres à l'intrigue des ouvrages – noms de personnages, de lieux, dates, etc – ainsi que des verbes de narration, qui servent à décrire l'action mais qui disent peu du fond thématique de nos textes. Ici, nous avons donc écarté manuellement ces éléments, en utilisant une liste de Stopwords à partir de ce qui existait déjà en ligne et en la complétant au fur et à mesure de l'analyse, et une liste de mots de narration, constituée ad hoc en explorant la liste de lemmes obtenue avec Spacy. À l'issue de cette phase de lemmatisation, nous avons donc récupéré un vaste ensemble de lemmes, qui conservait la structure de nos documents. Concrètement, il s'agissait d'une liste de listes de lemmes, chaque sous liste représentant les lemmes d'un document. Ces données sont stockées dans un fichier texte, et facilement récupérables pour la phase analytique. Cependant, comme on l'a dit plus haut, le *topic modeling* est très sensible à la longueur des documents fournis en entrée. Aussi, nous avons souhaité échantillonner nos documents pour harmoniser leur longueur. Nous avons donc enrichi notre module d'une fonction d'échantillonnage permettant de découper nos différentes listes de lemmes en échantillons de taille égale. Là encore, le *topic modeling* est très sensible à la question de la taille des documents fournis en entrée. Après différents essais, il semble qu'un minimum de 1000 lemmes par document soit essentiel au bon fonctionnement de nos méthodes. Nous avons donc découpé nos documents en échantillons de 1000 lemmes, pour finalement obtenir environ 700 échantillons lemmatisés, prêts à être fournis en entrée pour la phase analytique.

Obtenir un modèle pertinent

Pour obtenir un modèle pertinent, nous avons commencé par utiliser la fonction

d’assistance au choix du bon nombre de *topics*. Voici le graphique que nous avons obtenu à partir de cette fonction pour 300 passes et un nombre maximal de 20 *topics*.

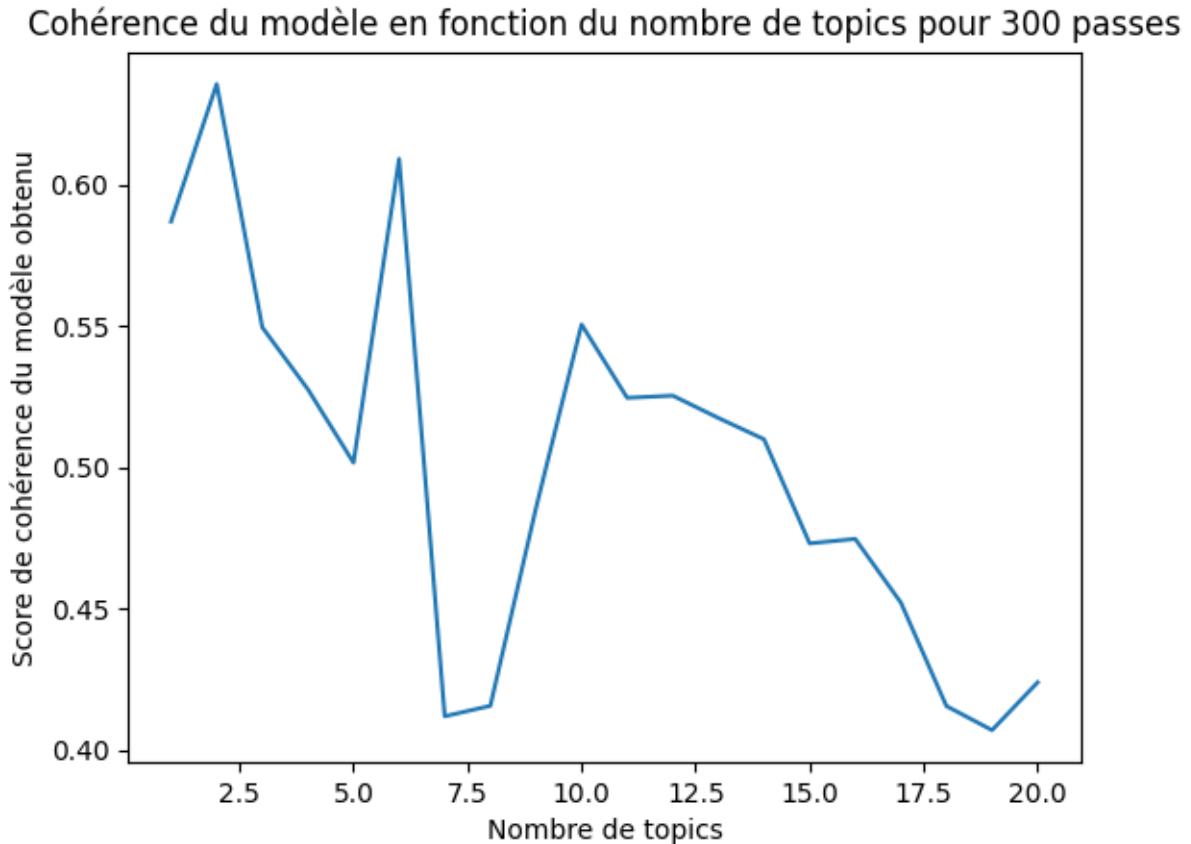


FIGURE 1 – Graphique de cohérence du modèle en fonction du nombre de topics

On observe deux pics de cohérence pour 2 *topics* et 6 *topics*. Le modèle obtenu avec deux *topics* semble d’ailleurs un peu plus cohérent que pour 6 *topics*. Cependant, on a déjà expliqué que cette fonction permettait une assistance au choix du bon nombre de *topics*, mais ne pouvait en aucun cas se substituer au regard du chercheur. Aussi, il est évident que nous préférerons travailler avec 6 *topics* plutôt qu’avec seulement 2, même au prix d’une petite perte de cohérence de notre modèle. C’est ce paramètre que nous retiendrons pour la suite de notre étude

A partir de là, nous avons travaillé à affiner notre modèle, en l’entraînant avec davantage d’itérations sur nos données. Nous avons, à ce stade de nos recherches, rencontré un deuxième questionnement majeur, sur lequel il nous faut ici revenir. En effet, nous avons expliqué plus haut que notre module possédait deux fonctions pour implémenter la LDA, l’une ayant recours aux données brutes de notre bag of words, et l’autre se fondant sur la méthode TF-IDF, que nous tenions pour plus efficace compte tenu d’expérimentations

menées sur un prototype.

Or, en menant de brefs essais à l'aide des deux fonctions, et en confrontant les résultats obtenus avec et sans TF-IDF, il nous est apparu que c'était sans TF-IDF que nous obtenions les meilleurs résultats. Voici deux *wordclouds* obtenus avec les premiers modèles – assez peu entraînés – pour chacune des deux fonctions.



FIGURE 2 – Wordcloud-Model5-6topics-3000passes
(sans TF-IDF)



FIGURE 3 – Wordcloud-Model4-6topics-5000passes
(avec TF-IDF)

En comparant brièvement ces deux nuages de mots, on se rend compte que le modèle qui n'utilise pas la méthode TF-IDF fait beaucoup plus de sens. Le deuxième est très redondant, puisque différents *topics* partagent les mêmes mots (angoisser, chloroforme, désinvolture, insécurité). Si on réfléchit plus longuement au problème qui se pose à nous, on peut entrevoir une piste explicative pour comprendre les moins bonnes performances du modèle ayant recours à TF-IDF. Comme on l'a dit, la méthode TF-IDF s'applique à faire ressortir les mots discriminants au sein de chacun des textes. Or, dans le cadre d'une exploration thématique de la sémantique employée dans notre corpus, nous souhaitons au contraire faire ressortir les thématiques communes qui transcendent notre corpus et se retrouvent de manière abondante dans différents textes. Les mots très spécifiques à chacune des œuvres, au contraire, n'intéressent que dans une moindre mesure notre analyse, puisqu'ils renvoient à des thématiques trop proches des intrigues, et nous disent que très

peu des représentations partagées.

A partir de ces différents éléments, nous avons donc entraîné différents modèles – plus d'une dizaine en tout, que l'on pourra retrouver en annexe - en affinant tout particulièrement un modèle qui n'utilisait pas TF-IDF : le modèle 12, entraîné avec 5000 passes et 5000 itérations. En parallèle, nous avons également entraîné plusieurs modèles sur les différentes listes de n-grams dont nous disposons, quoiqu'avec un entraînement plus allégé - 500 passes et 500 itérations. Pour comprendre comment, à partir des fonctions de notre module, nous avons procédé pour entraîner les modèles et obtenir des résultats, nous invitons le lecteur à consulter le script présent en annexe¹¹. Présentons donc maintenant les résultats que nous obtenons avec ces différents modèles.

11. Cf Script TMAnalysis en annexe 2.6.

Chapitre 3

Résultats et retours méthodologiques

3.1 Présentation des résultats

Les résultats de notre modèle final

Commençons ici par présenter les résultats obtenus sur notre corpus échantillonné avec le modèle 12, c'est-à-dire le modèle le plus affiné, entraîné pour 6 *topics* avec 5000 passes et 5000 itérations. Voici un *wordcloud* et un graphique représentant les mots principaux des différents *topics*, tous deux obtenus à partir de notre modèle.



FIGURE 4 – Wordcloud-Model12-6topics-5000passes
(avec TF-IDF)

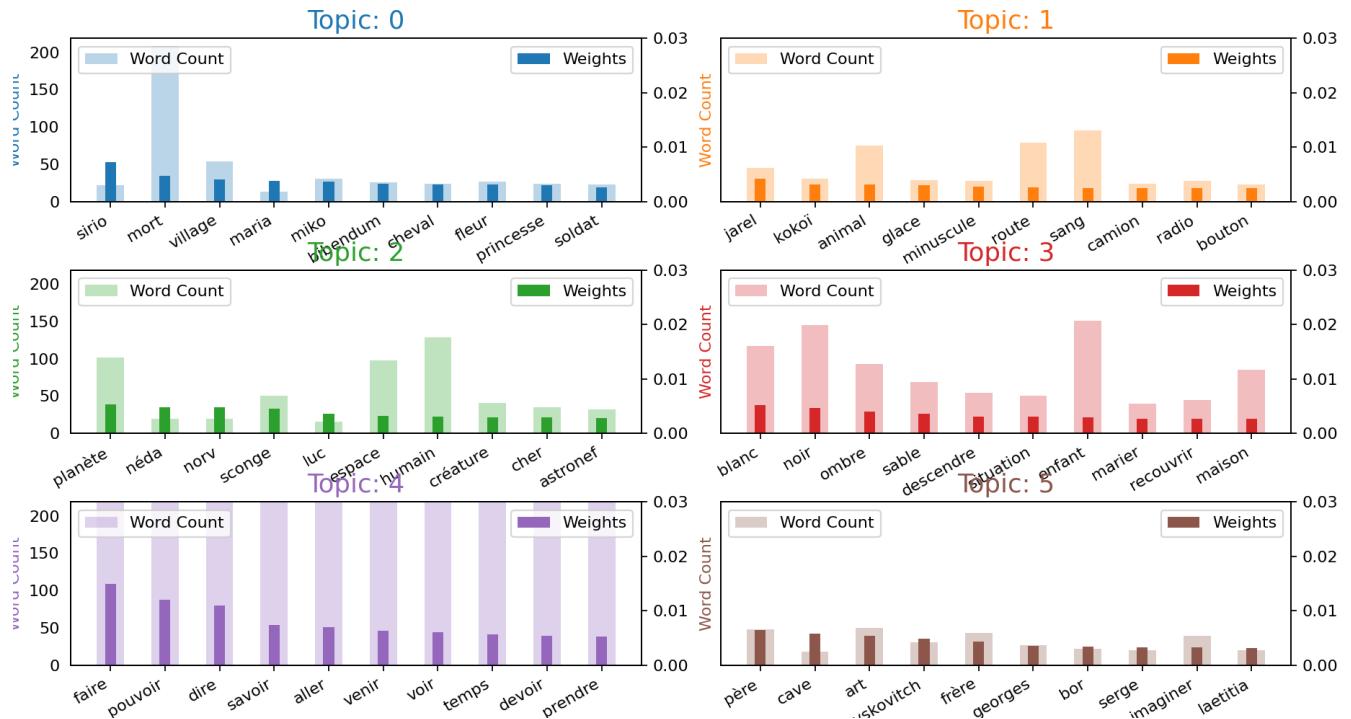


FIGURE 5 – Wordcloud-Model12-6topics-5000passes
(avec TF-IDF)

Ces deux résultats nous permettent ici de faire quelques remarques intéressantes. D'abord, intéressons nous aux 6 *topics* en eux-mêmes.

Le premier est relativement cohérent. Il semble renvoyer à un imaginaire médiéval, chevaleresque, avec des termes comme « princesse », « mort », « village », « soldat », « cheval » et « fleur ». Le mot « bibendum », qui désigne la mascotte de la marque Renault, et dont la figure a été réemployée dans différentes œuvres de science-fiction, n'a pas énormément de cohérence avec le reste du *topic*. En outre, on note la présence de noms propres, qui auraient pourtant dû être écartés lors de la phase de preprocessing : presque toujours présents, quoiqu'en petit nombre, dans les *topics* que nous obtenons en résultats, nous reviendrons plus longuement sur cette question.

Le deuxième *topic*, quant à lui, ne peut que difficilement être associé à une thématique propre. On retrouve quelques mots qui peuvent être, de près ou de loin, associés au thème du voyage, avec par exemple « route », « camion », « radio », « animal », ou « glace ». Mais on retrouve également des termes assez éloignés thématiquement, comme « minuscule », « sang » et « bouton », qu'il est difficile d'associer à de mêmes représentations. Enfin, on trouve les noms propres « jarel » et « kokoï ».

Le troisième *topic*, en revanche, est assez cohérent et renvoie très largement aux thématiques de l'exploration et de la conquête spatiales. On retrouve ainsi les termes « planète », « espace », « humain », « créature » et « astronef ». On retrouve également

le terme « sconge », qui renvoie à une créature extraterrestre. Enfin, on retrouve encore une fois quelques noms propres : « néda », « norv » et « luc ».

Le quatrième *topic* semble renvoyer aux thèmes de la famille, avec des termes comme « situation », « enfant », « marier » et « maison ». On retrouve également les couleurs « blanc » et « noir » qui peuvent être associées sémantiquement au thème évoqué. Enfin, on retrouve les termes « sable », et « ombre », ainsi que les verbes « descendre » et « recouvrir » qui, sans être en pleine adéquation avec le thème de la famille, n'en sont pas franchement éloignés. Aussi, ce quatrième *topic* apparaît également comme relativement cohérent.

Le cinquième *topic* ne rassemble que des verbes de narration, que nous avions pourtant écarté de l'analyse. On trouve ainsi les termes « faire », « dire », « pouvoir », « aller », « savoir », « venir », « voir », « devoir », « prendre », ainsi que le mot assez générique – surtout en français – « temps ». Ici, le *topic* est parfaitement cohérent en ce qu'il rassemble des termes d'une même nature morphosyntaxique ; cependant, ces termes auraient dû être écartés lors de la phase de preprocessing des données. En fait, la présence de ce thème révèle les limites des fonctions de notre module en matière de prétraitement des données : certaines formes des verbes qu'on a mentionnés échappent à notre algorithme, si bien qu'ils se retrouvent en grand nombre dans notre analyse. Nous pouvons donc ignorer ce *topic* dans notre travail d'interprétation.

Enfin, le dernier *topic* n'est pas très cohérent. On retrouve des mots associés au thème de la famille, comme « père » et « frère » – on peut d'ailleurs ici se demander pourquoi ils n'ont pas été rangés dans le quatrième *topic*. On retrouve également, de manière esseulée, les termes « cave », d'une part, et « art » et « imaginer », d'autre part. On retrouve finalement de nombreux noms propres : « georges », « bor », « serge », « laetitia » et « lovskovitch ».

Revenons justement sur la présence de noms propres dans les différents *topics* que nous venons de présenter. De manière analogue aux verbes de narration, ils auraient dû être écartés lors de la phase de preprocessing des données. Cependant, là encore, on assiste aux limites de nos fonctions de prétraitement, qui ne parviennent pas à distinguer tous les noms propres. Nous avons travaillé manuellement à écarter les noms propres en les ajoutant à la liste des mots de narration que nous supprimons lors du preprocessing. Mais leur omniprésence dans l'ensemble du corpus, ainsi que leur grande fréquence d'occurrence – puisqu'il s'agit de romans, les noms des personnages sont cités un grand nombre de fois – rend cette tâche difficile. Nous avons donc dû nous résoudre à ignorer leur présence dans notre interprétation thématique des *topics*.

Afin de se livrer à un examen plus approfondi des résultats du modèle 12 et des *topics* obtenus, nous allons désormais explorer les solutions de visualisation offertes par le

module LDavis. Nous présentons ici les screenshots de ces visualisations, avec les différents *topics* sélectionnés. Pour consulter directement la visualisation en ligne, nous invitons le lecteur à consulter le lien qui mène directement à celle-ci¹.

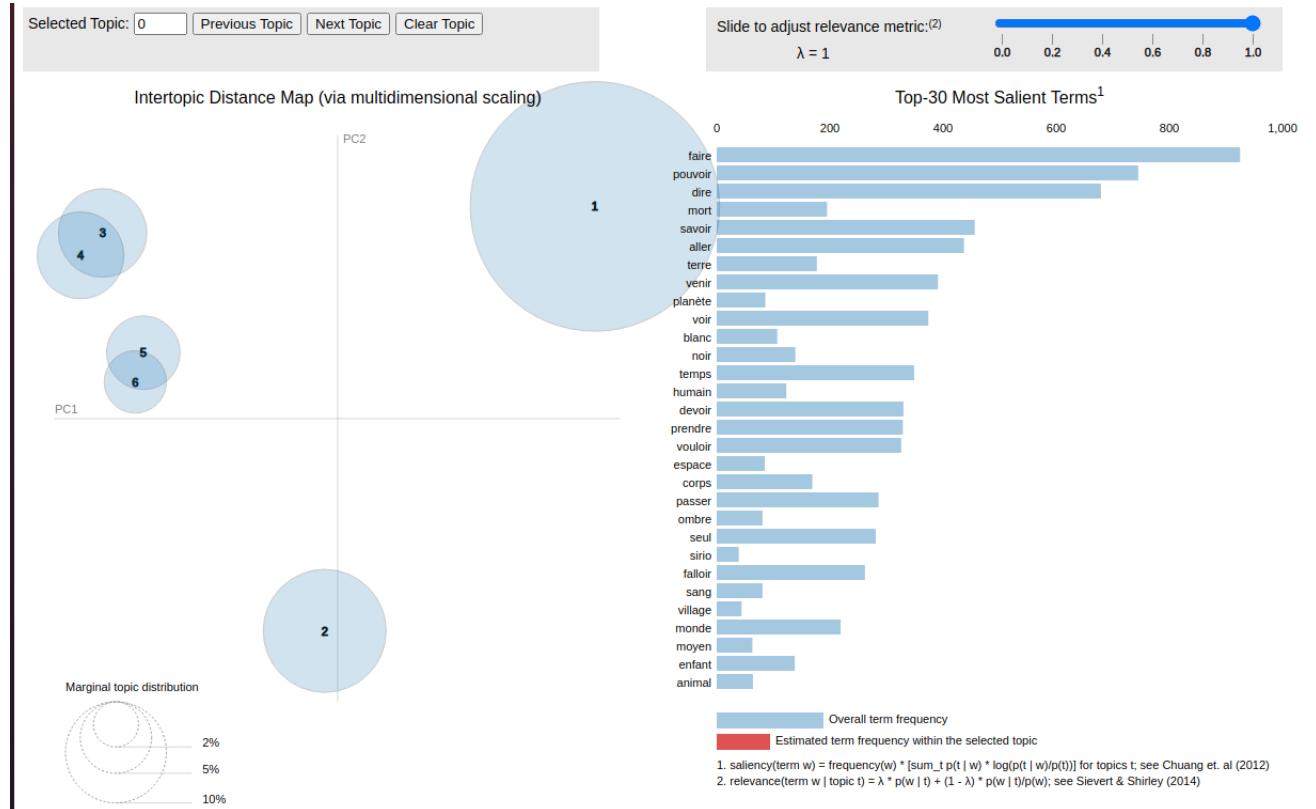


FIGURE 6 – LDavis-model12-notopicsselected

1. file : //home/leo/Desktop/M%C3%A9moire%20M2/Annexes/Results/LDAvis-Model12-6topics-5000passes-5000iter.html.

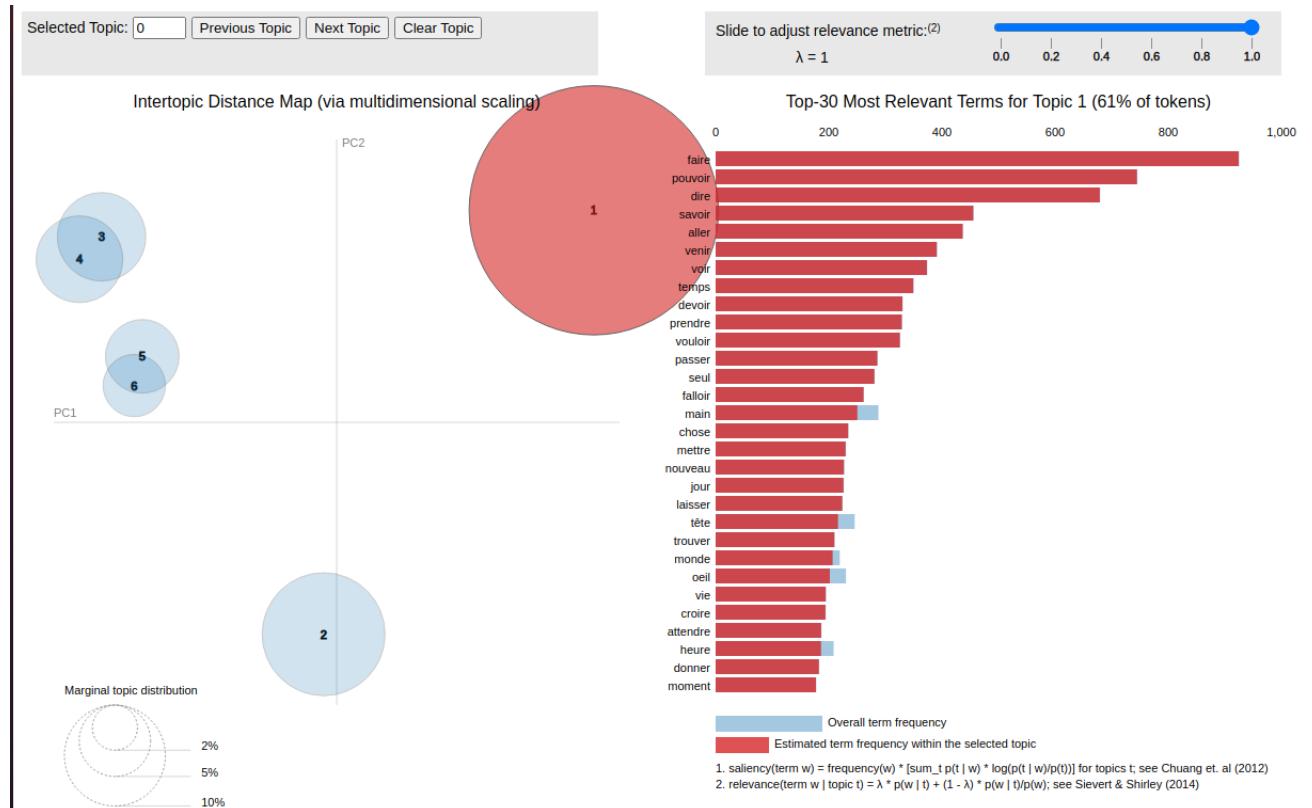


FIGURE 7 – LDAvis-model12-topic1selected

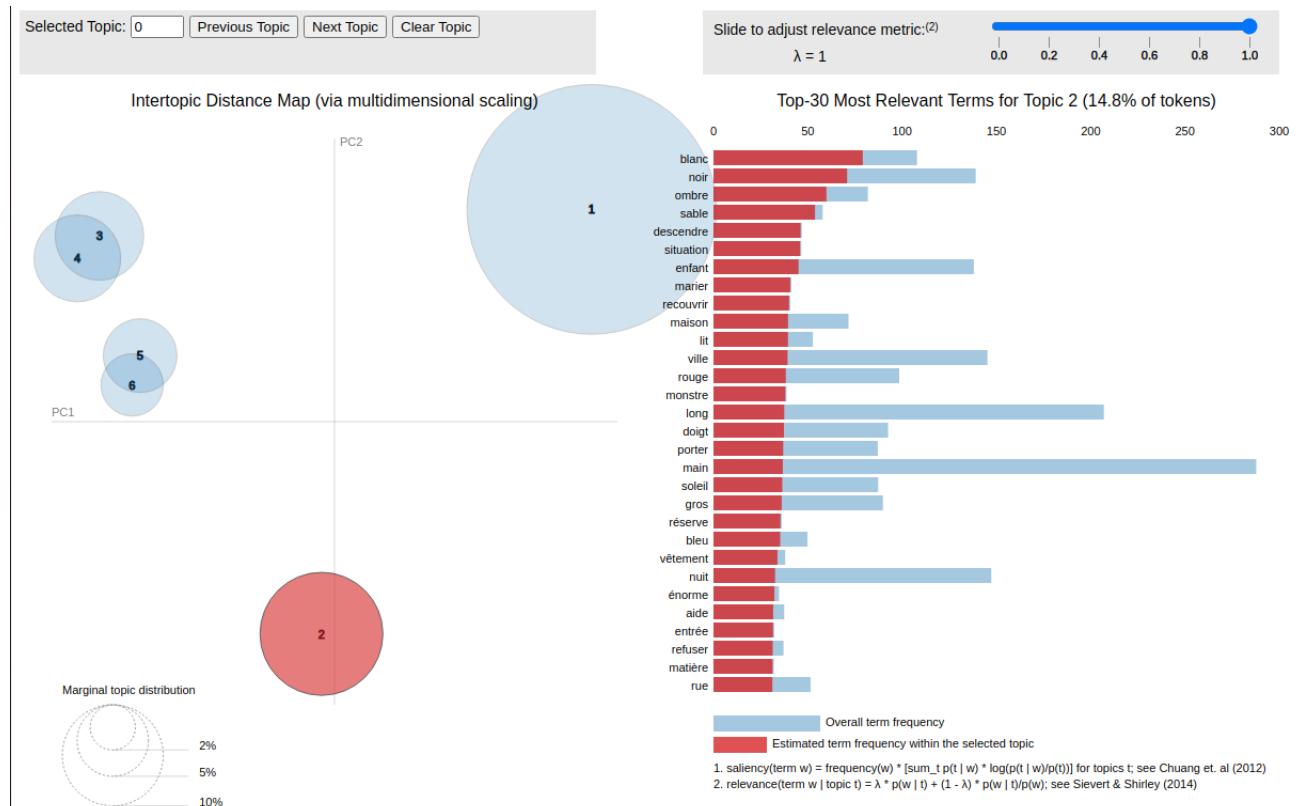


FIGURE 8 – LDAvis-model12-topic2selected

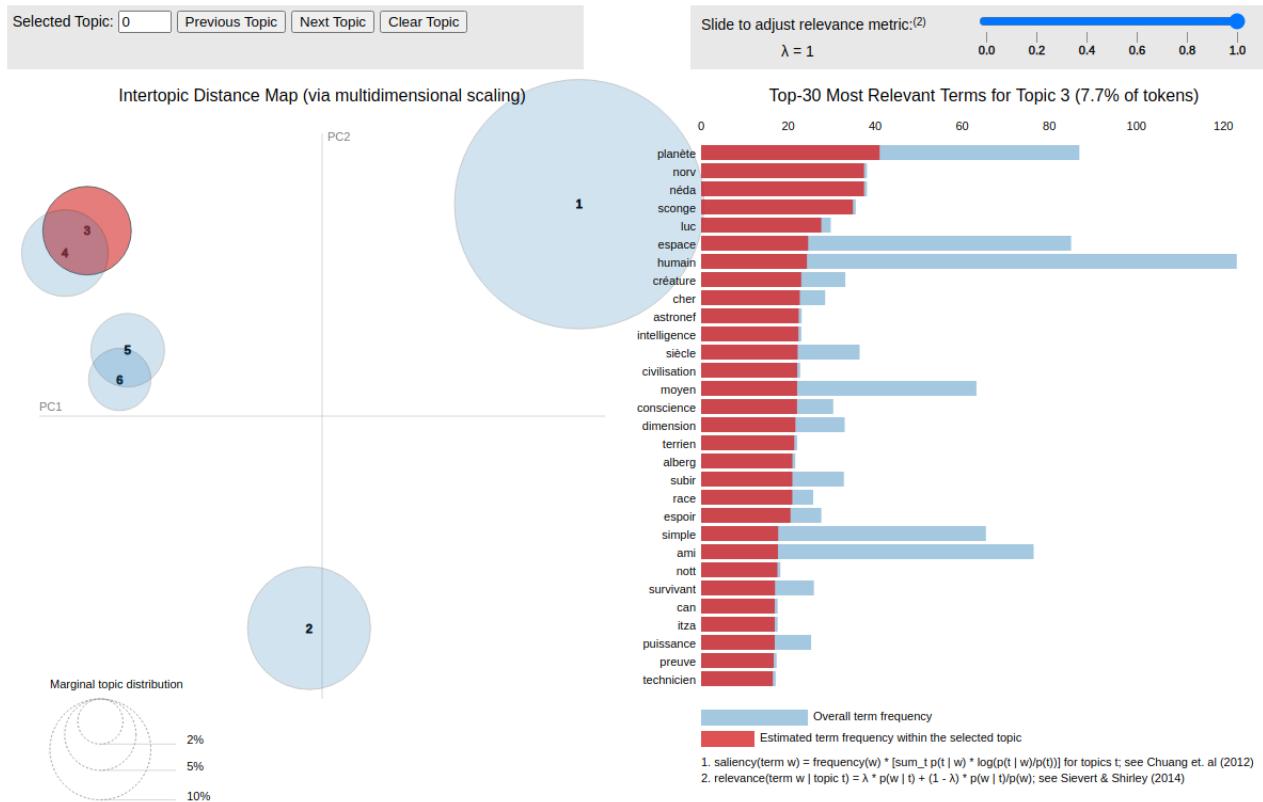


FIGURE 9 – LDAvis-model12-topic3selected

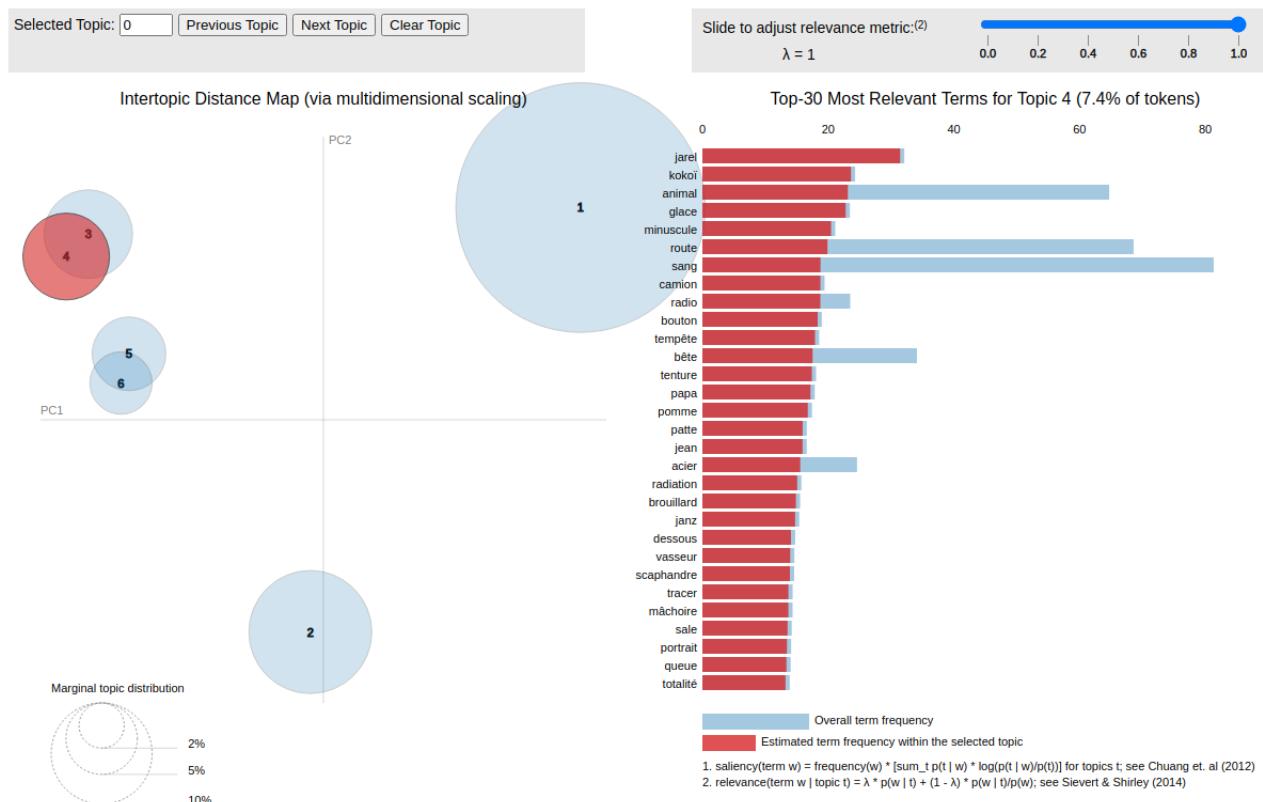


FIGURE 10 – LDAvis-model12-topic4selected

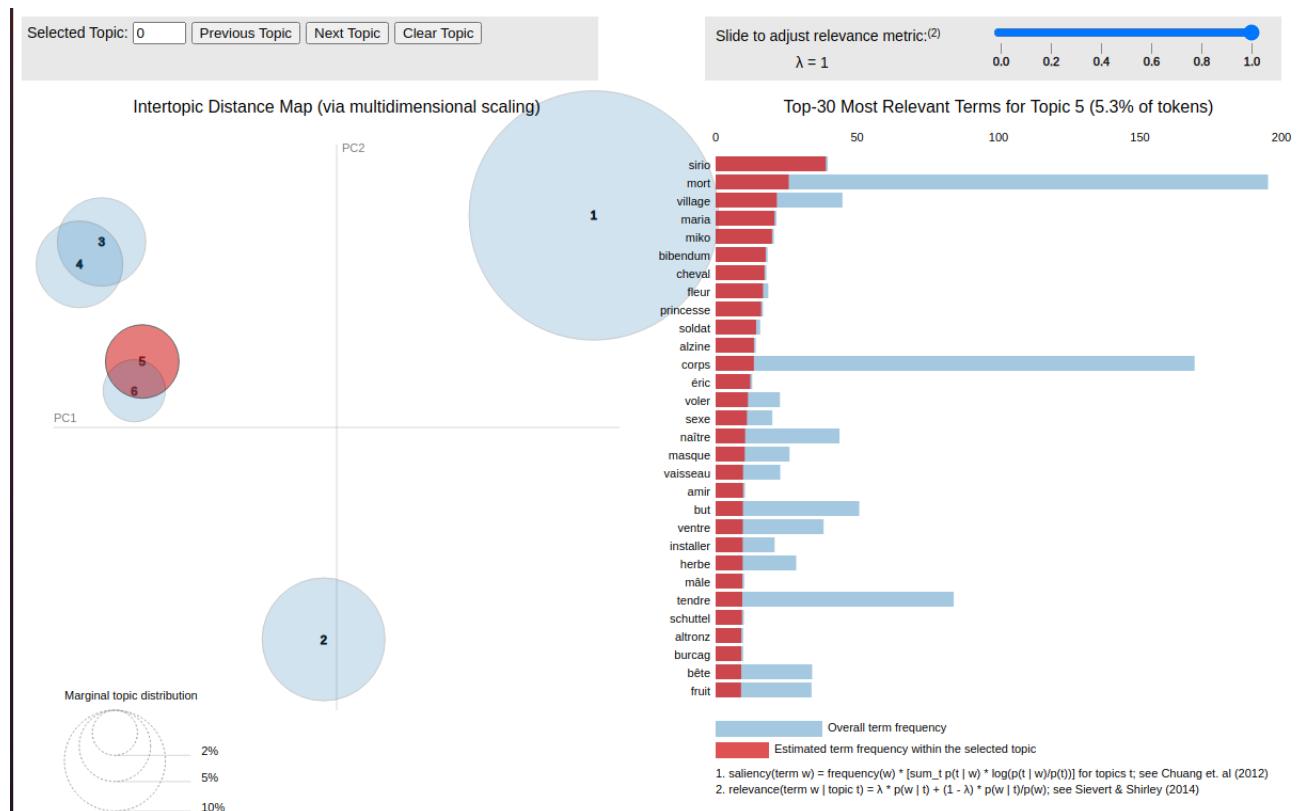


FIGURE 11 – LDavis-model12-topic5selected

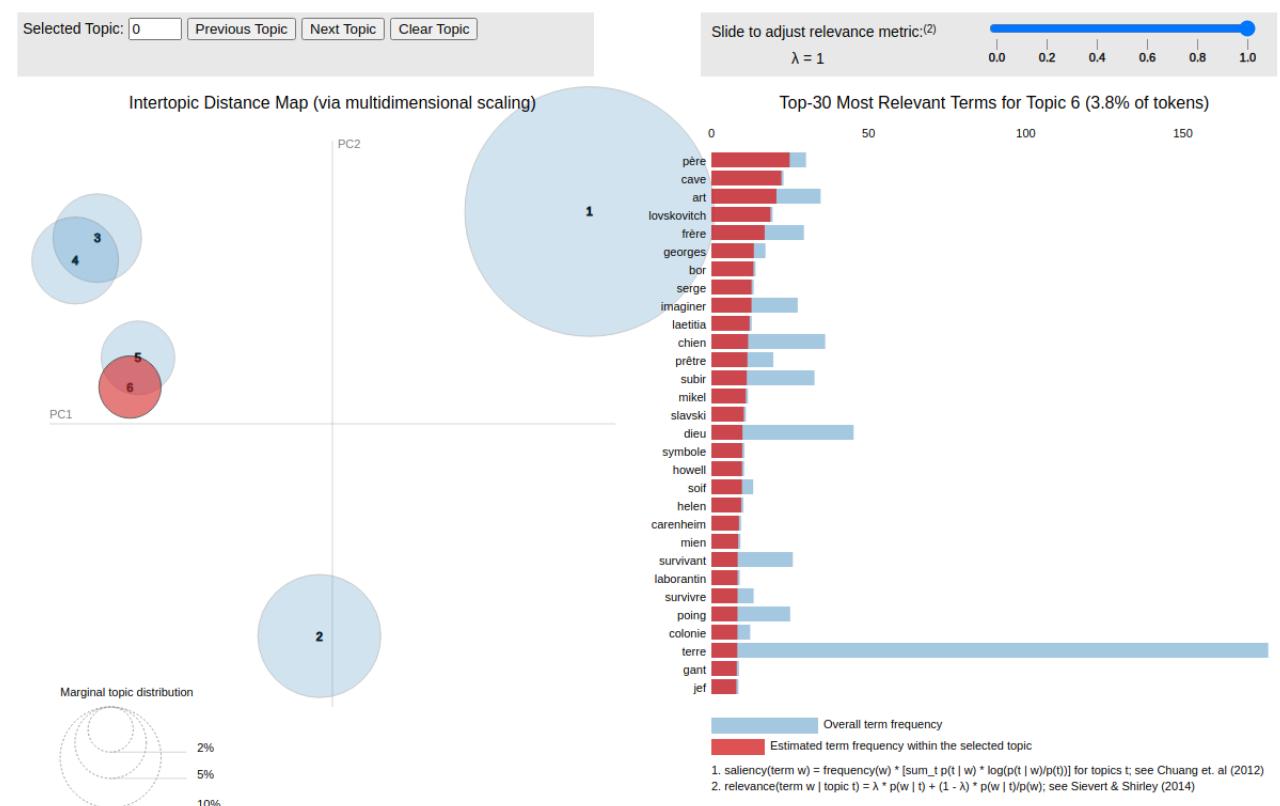


FIGURE 12 – LDavis-model12-topic6selected

Ces visualisations permettent, en premier lieu, d'explorer plus en profondeur chacun des *topics*. En observant davantage de mots pour chaque *topic*, on peut mettre en perspective les interprétations qu'on avait faites, et les affiner. Ainsi du *topic* 3, qu'on avait associé à l'exploration et à la conquête spatiales. À ce dernier s'ajoutent notamment les termes « civilisation », « intelligence », « terrien », « dimension », « race », « espace », « survivant » et « espoir », qui appuient encore notre interprétation.

On mesure également, grâce aux visualisations offertes par LDAvis, l'importance relative des *topics* dans l'ensemble du corpus. Sans surprise, c'est le *topic* associé aux verbes de narration (*topic* 5) qui représente la part la plus conséquente de tokens dans notre corpus. Viennent ensuite, respectivement, celui qu'on avait associé à la famille (*topic* 4), celui de l'exploration et de la conquête spatiales (*topic* 3), celui qui nous semblait vaguement lié au voyage (*topic* 2), puis celui renvoyant à des représentations chevaleresques et médiévales (*topic* 1) et enfin celui dont on n'avait pas pu tirer d'interprétation très claire (*topic* 6).

Enfin, ces visualisations nous permettent de faire quelques commentaires sur la distance inter-*topic* et les axes structurants pour notre modèle. On note que les *topics* 1 et 2 sont tous deux relativement singuliers, et se distinguent particulièrement des deux autres. Les *topics* 3 à 6, en revanche, sont assez proches, et se superposent même par endroits. On peut également noter que le premier axe structurant (PC1) semble opposer thématiques terrestres et spatiales, tandis que le deuxième (PC2) correspond à l'opposition entre termes de narration et termes propres aux intrigues.

Résultats complémentaires obtenus sur nos listes de n-grams

Ici, nous allons présenter plus succinctement les résultats que nous avons obtenus sur les différentes listes de n-grams, et qui révèlent parfois certaines composantes thématiques intéressantes. Sans nous adonner à une analyse aussi aboutie que pour le modèle précédent, nous les commenterons toutefois en relevant les éléments que nous jugerons particulièrement pertinents. Nous procéderons dans l'ordre du nombre de grams, c'est-à-dire en allant des analyses portées sur les bi-grammes à celles qui concernent les heptagrammes. Nous présenterons d'abord l'ensemble des visualisations, avant de faire quelques commentaires.

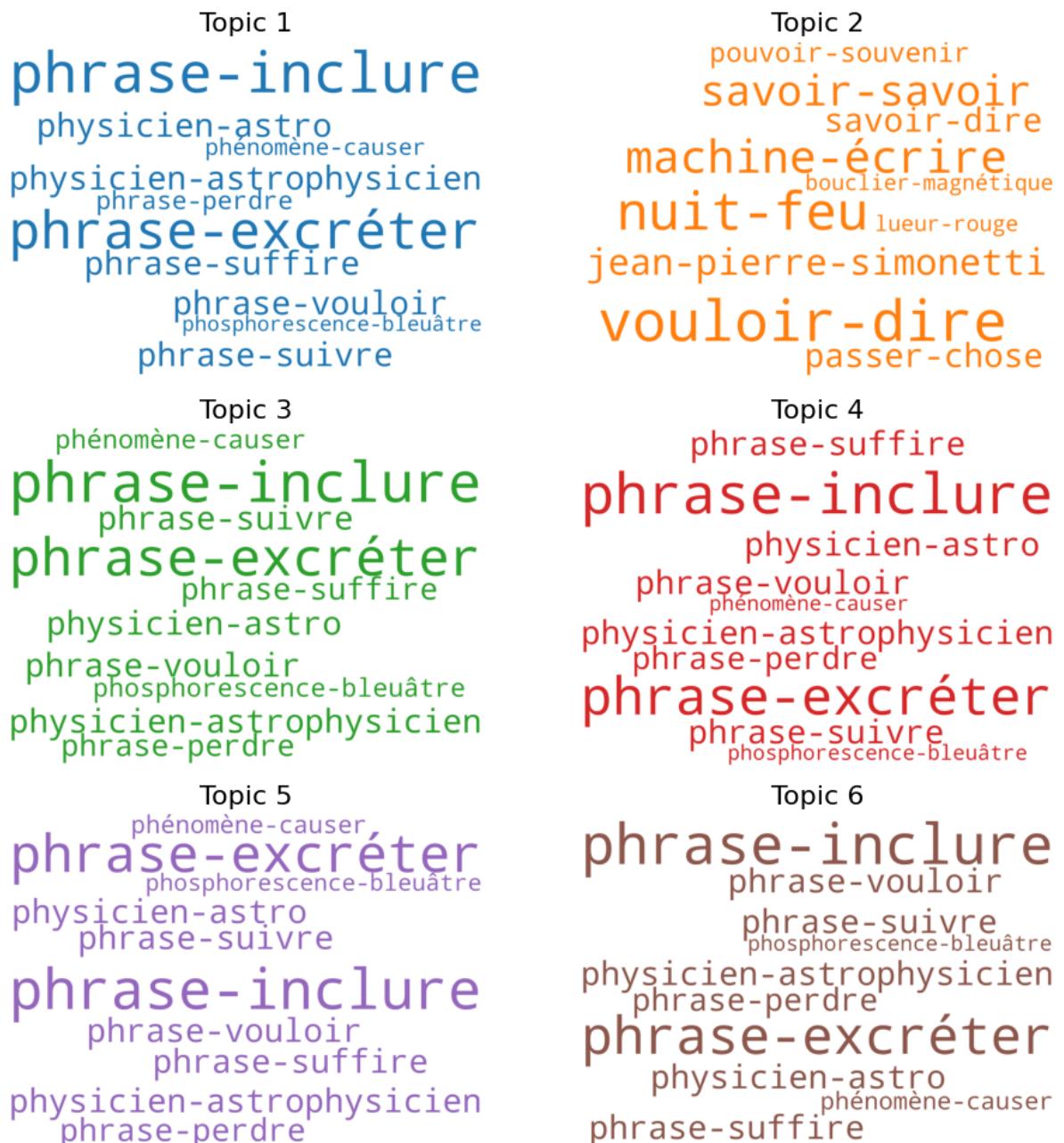


FIGURE 13 – WordCloud-Model6-bigrams-500passes

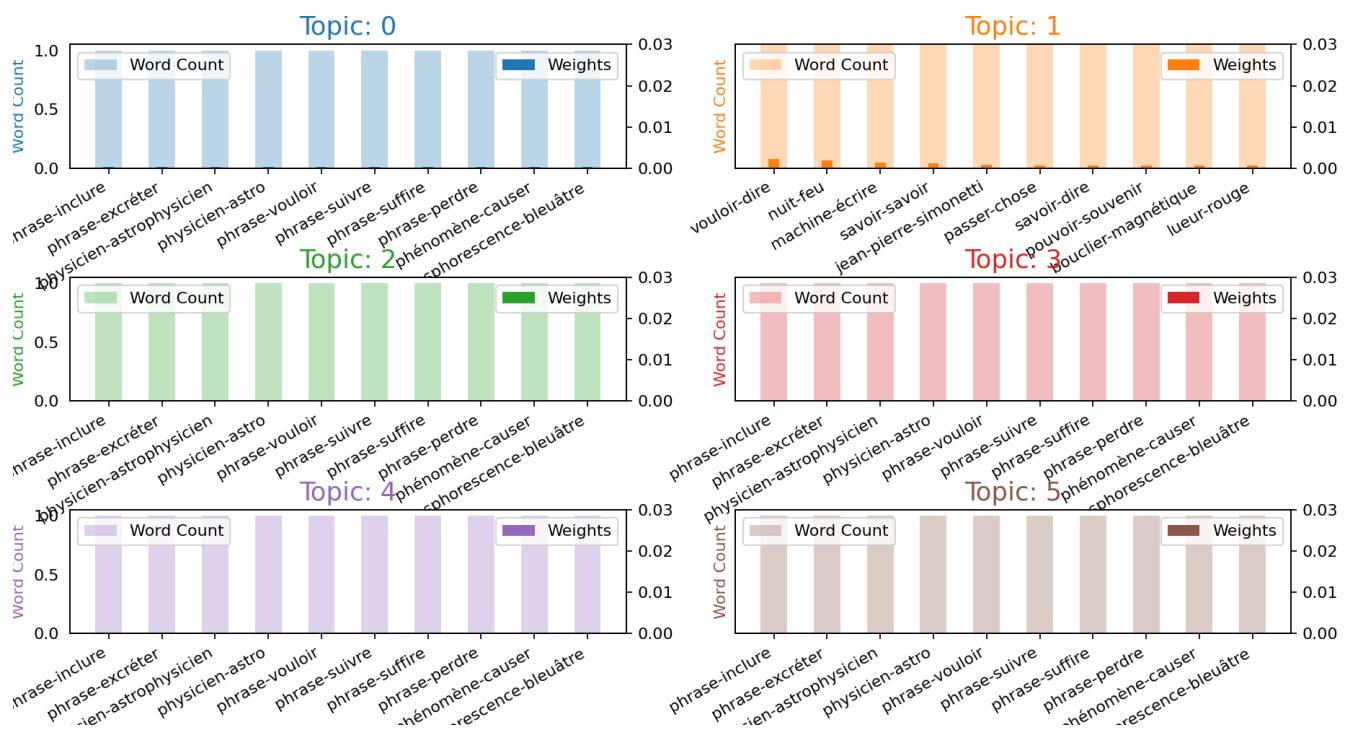


FIGURE 14 – Graphique-Model6-bigrams-500passes



FIGURE 15 – WordCloud-Model7-trigrams-500passes

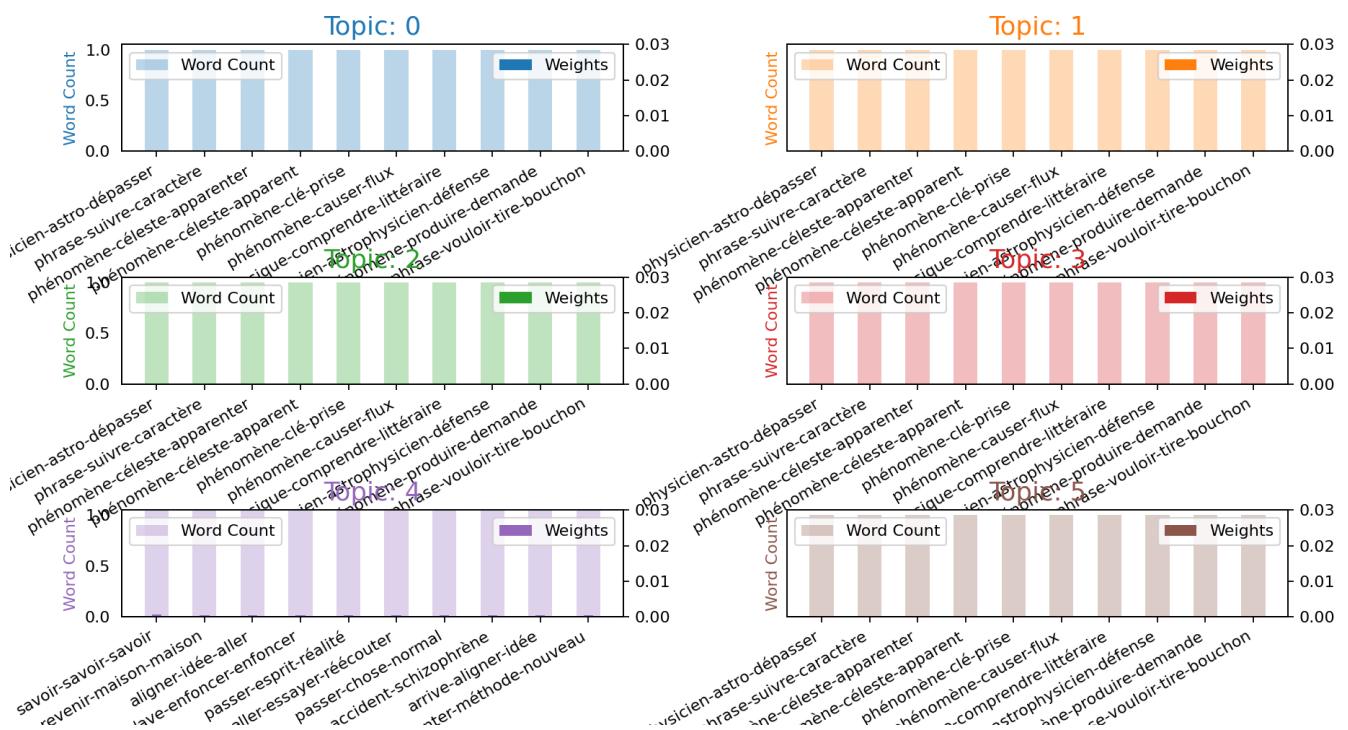


FIGURE 16 – Graphique-Model7-trigrams-500passes



FIGURE 17 – WordCloud-Model8-tetragrams-500passes

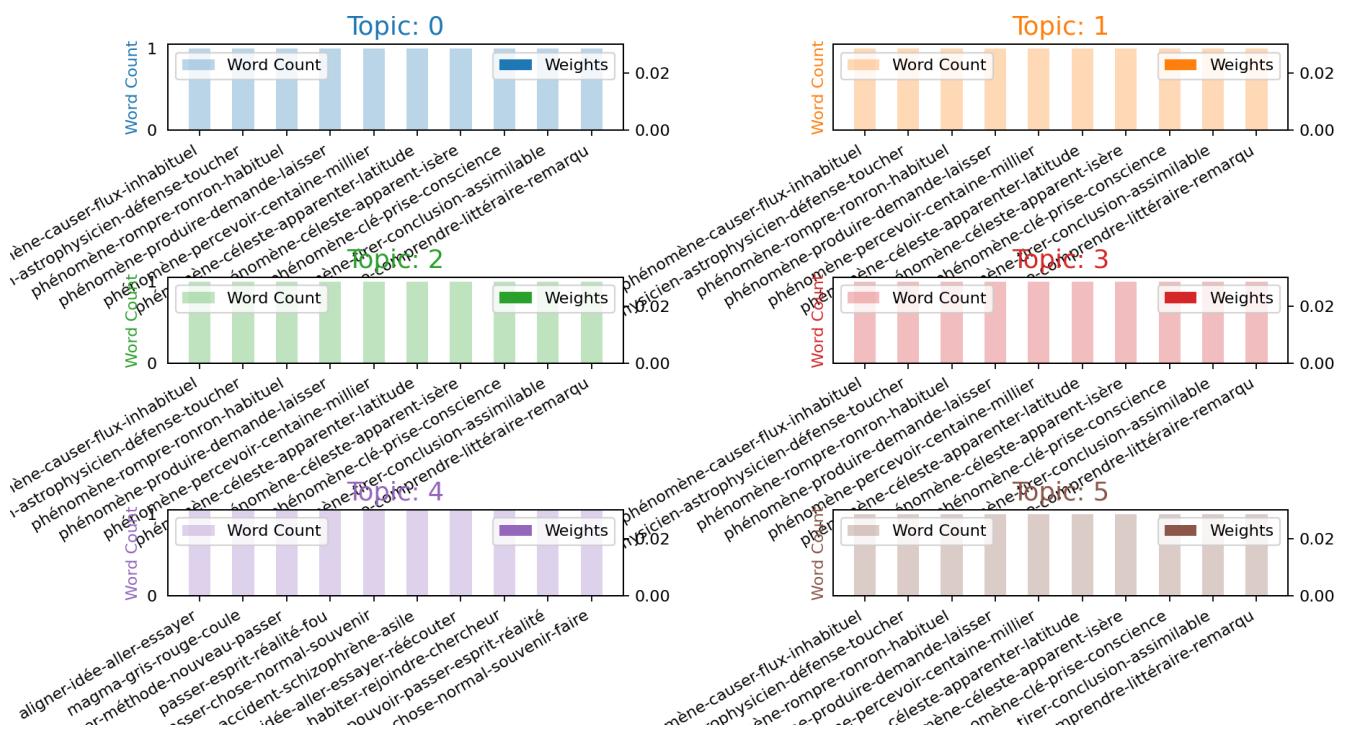


FIGURE 18 – Graphique-Model8-tetragrams-500passes

Topic 1

phénomène-clé-prise-conscience-incarnation
 phénomène-causer-flux-inhabituel-électron
 phénomène-produire-demande-laisser-imagination
 phénomène-percevoir-centaine-millier-témoin
 physicien-astrophysicien-défense-toucher-défense
 physique-comprendre-littéraire-remarqu-reentrer
 phénomène-céleste-apparent-isère-parler
 phénomène-surnaturel-mystification-titre-barrer
 phénomène-céleste-apparenter-latitude-farce
 physicien-astro-dépasser-tien-faire

Topic 2

physique-comprendre-littéraire-remarqu-reentrer
 phénomène-surnaturel-mystification-titre-barrer
 phénomène-causer-flux-inhabituel-électron
 phénomène-percevoir-centaine-millier-témoin
 phénomène-clé-prise-conscience-incarnation
 physicien-astrophysicien-défense-toucher-défense
 phénomène-produire-demande-laisser-imagination
 phénomène-céleste-apparenter-latitude-farce
 physicien-astro-dépasser-tien-faire
 phénomène-céleste-apparent-isère-parler

Topic 3

phénomène-céleste-apparenter-latitude-farce
 phénomène-percevoir-centaine-millier-témoin
 phénomène-causer-flux-inhabituel-électron
 phénomène-clé-prise-conscience-incarnation
 phénomène-céleste-apparent-isère-parler
 phénomène-produire-demande-laisser-imagination
 physicien-astrophysicien-défense-toucher-défense
 physicien-astro-dépasser-tien-faire
 phénomène-surnaturel-mystification-titre-barrer
 physique-comprendre-littéraire-remarqu-reentrer

Topic 4

phénomène-surnaturel-mystification-titre-barrer
 physique-comprendre-littéraire-remarqu-reentrer
 phénomène-percevoir-centaine-millier-témoin
 phénomène-causer-flux-inhabituel-électron
 physicien-astrophysicien-défense-toucher-défense
 phénomène-céleste-apparenter-latitude-farce
 phénomène-céleste-apparent-isère-parler
 phénomène-clé-prise-conscience-incarnation
 phénomène-produire-demande-laisser-imagination
 physicien-astro-dépasser-tien-faire

Topic 5

univers-désertique-contreire-habiter-rejoindre
 arrive-aligner-idée-aller-essayer
 méthode-nouveau-passé-chose-normal
 passer-chose-normal-souvenir-faire
 pouvoir-passé-esprit-réalité-fou
 magma-gris-rouge-coule-lave
passer-esprit-réalité-fou-choquer
 aligner-idée-aller-essayer-réécouter
 surréaliste-pouvoir-passé-esprit-réalité
 passer-univers-désertique-contreire-habiter

Topic 6

phénomène-céleste-apparent-isère-parler
 phénomène-céleste-apparenter-latitude-farce
 phénomène-surnaturel-mystification-titre-barrer
 phénomène-percevoir-centaine-millier-témoin
 physicien-astrophysicien-défense-toucher-défense
 physicien-astro-dépasser-tien-faire
 phénomène-causer-flux-inhabituel-électron
 phénomène-clé-prise-conscience-incarnation
 phénomène-produire-demande-laisser-imagination
 physique-comprendre-littéraire-remarqu-reentrer

FIGURE 19 –
 WordCloud-Model9-pentagrams-500passes

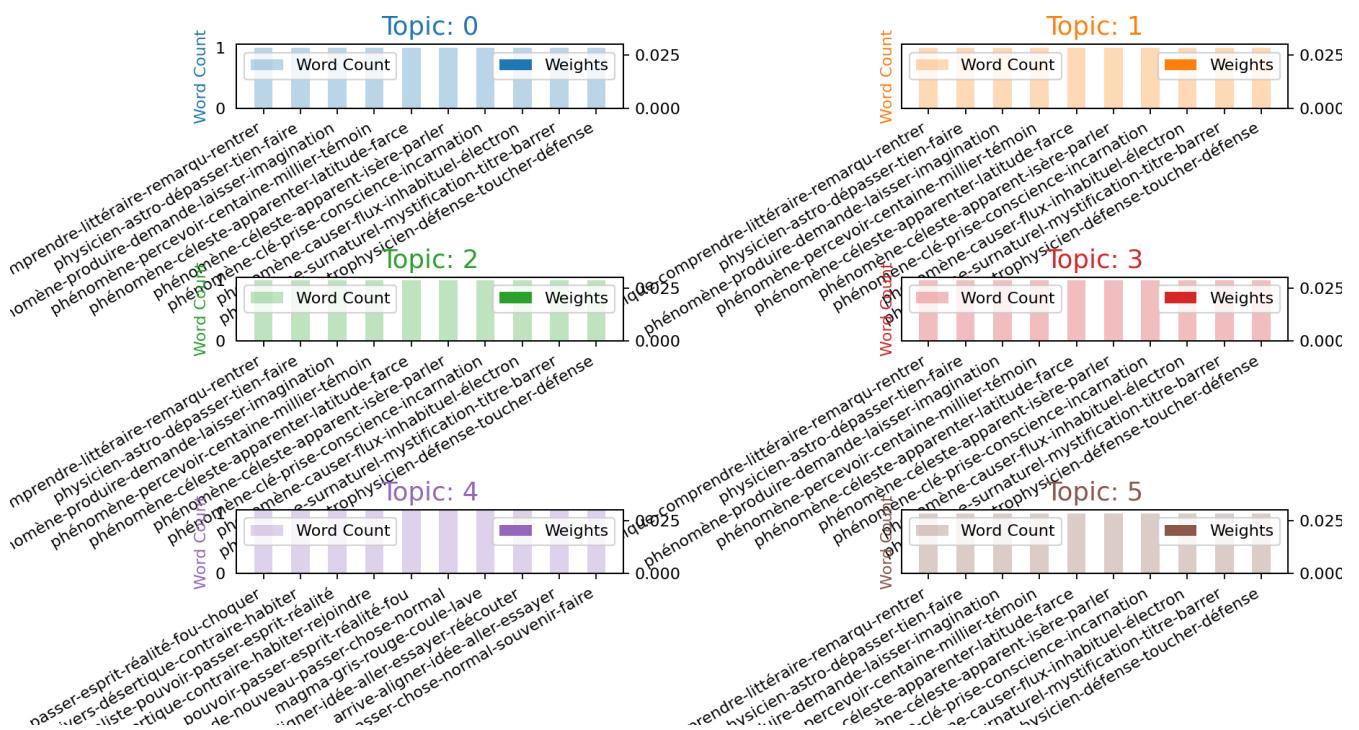


FIGURE 20 – Graphique-Model9-pentagrams-500passes



FIGURE 21 –
WordCloud-Model10-hexagrams-500passes

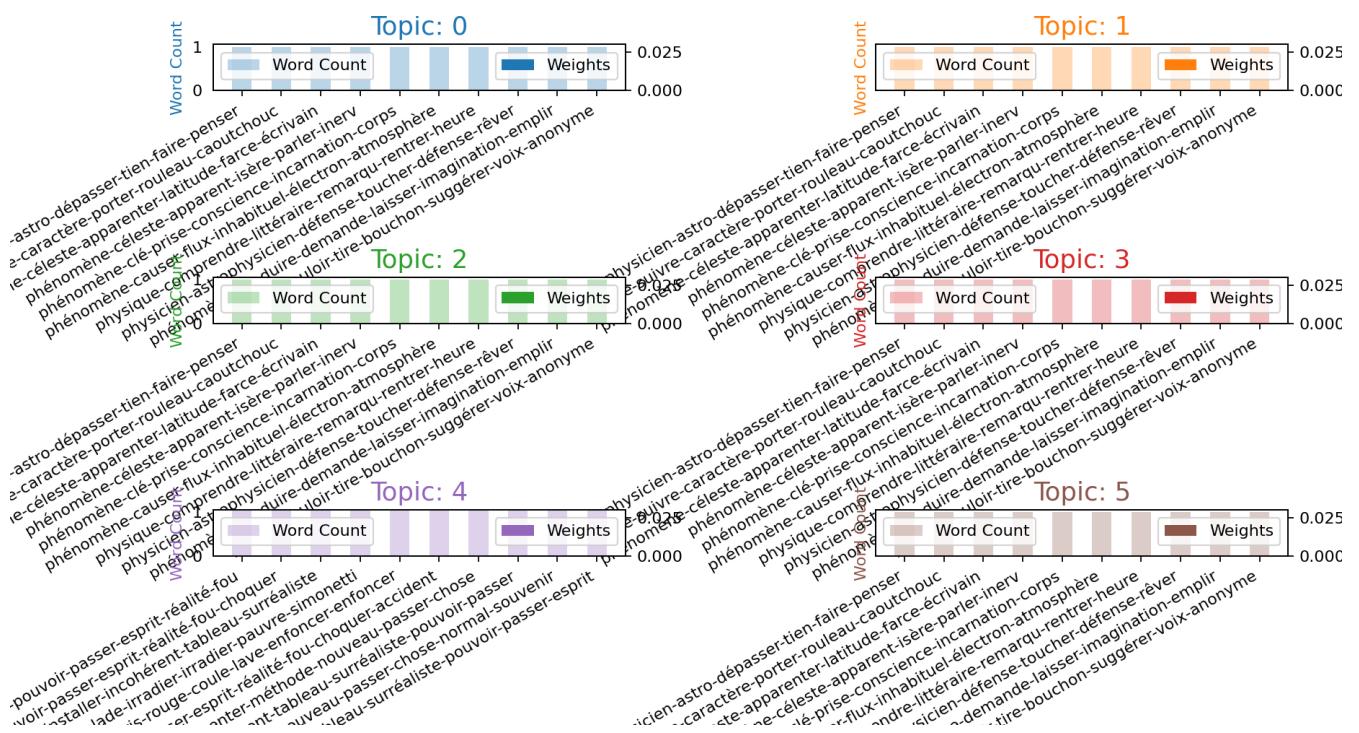


FIGURE 22 – Graphique-Model10-hexagrams-500passes

Topic 1

physicien-astro-dépasser-tien-faire-penser-moment
 physicien-astrophysicien-défense-toucher-défense-rêver-bavard
 phénomène-percevoir-centaine-millier-témoin-nuit-septembre
 phénomène-céleste-apparent-isère-parler-inerv-institut
 physique-comprendre-littéraire-remarqu-reentrer-heure-soir
 phénomène-causer-flux-inhabituel-électron-atmosphère-suite
 phénomène-clé-prise-conscience-incarnation-corps-vivant
 phénomène-céleste-apparenter-latitude-farce-écrivain-savoir
 phrase-vouloir-tire-bouchon-suggérer-voix-anonyme-manhattan
 phénomène-rompre-ronron-habituel-détriment-sacro-saint

Topic 3

phénomène-causer-flux-inhabituel-électron-atmosphère-suite
 phénomène-céleste-apparenter-latitude-farce-écrivain-savoir
 physicien-astrophysicien-défense-toucher-défense-rêver-bavard
 phénomène-céleste-apparent-isère-parler-inerv-institut
 phénomène-rompre-ronron-habituel-détriment-sacro-saint
 phrase-vouloir-tire-bouchon-suggérer-voix-anonyme-manhattan
 physique-comprendre-littéraire-remarqu-reentrer-heure-soir
 phénomène-clé-prise-conscience-incarnation-corps-vivant
 physicien-astro-dépasser-tien-faire-penser-moment
 phénomène-percevoir-centaine-millier-témoin-nuit-septembre

Topic 5

esprit-réalité-fou-choquer-accident-schizophrène-asile
 tableau-surréaliste-pouvoir-passé-esprit-réalité-fou
 surréaliste-pouvoir-passé-esprit-réalité-fou-choquer
 passer-univers-désertique-contraire-habiter-rejoindre-chercheur
 magma-gris-rouge-coule-lave-enfoncer-enfoncer
 expérimenter-méthode-nouveau-passé-chose-normal-souvenir
 réinstaller-incohérent-tableau-surréaliste-pouvoir-passé-esprit
 méthode-nouveau-passé-chose-normal-souvenir-faire
 sûr-expérimenter-méthode-nouveau-passé-chose-normal
 décor-familier-réinstaller-incohérent-tableau-surréaliste-pouvoir

Topic 2

physique-comprendre-littéraire-remarqu-reentrer-heure-soir
 phénomène-percevoir-centaine-millier-témoin-nuit-septembre
 physicien-astro-dépasser-tien-faire-penser-moment
 phénomène-rompre-ronron-habituel-détriment-sacro-saint
 phrase-vouloir-tire-bouchon-suggérer-voix-anonyme-manhattan

phénomène-clé-prise-conscience-incarnation-corps-vivant
 phénomène-causer-flux-inhabituel-électron-atmosphère-suite
 physicien-astrophysicien-défense-toucher-défense-rêver-bavard
 phénomène-céleste-apparenter-latitude-farce-écrivain-savoir
 phénomène-céleste-apparent-isère-parler-inerv-institut

Topic 4

physicien-astrophysicien-défense-toucher-défense-rêver-bavard
 phénomène-causer-flux-inhabituel-électron-atmosphère-suite
 physicien-astro-dépasser-tien-faire-penser-moment
 phénomène-clé-prise-conscience-incarnation-corps-vivant
 physique-comprendre-littéraire-remarqu-reentrer-heure-soir
 phénomène-céleste-apparent-isère-parler-inerv-institut
 phénomène-céleste-apparenter-latitude-farce-écrivain-savoir
 phénomène-percevoir-centaine-millier-témoin-nuit-septembre
 phrase-vouloir-tire-bouchon-suggérer-voix-anonyme-manhattan
 phénomène-rompre-ronron-habituel-détriment-sacro-saint

Topic 6

phénomène-clé-prise-conscience-incarnation-corps-vivant
 physique-comprendre-littéraire-remarqu-reentrer-heure-soir
 physicien-astrophysicien-défense-toucher-défense-rêver-bavard
 phénomène-céleste-apparent-isère-parler-inerv-institut
 phénomène-rompre-ronron-habituel-détriment-sacro-saint
 phénomène-percevoir-centaine-millier-témoin-nuit-septembre
 physicien-astro-dépasser-tien-faire-penser-moment
 phénomène-céleste-apparenter-latitude-farce-écrivain-savoir
 phrase-vouloir-tire-bouchon-suggérer-voix-anonyme-manhattan
 phénomène-causer-flux-inhabituel-électron-atmosphère-suite

FIGURE 23 –
 WordCloud-Model11-heptagrams-500passes

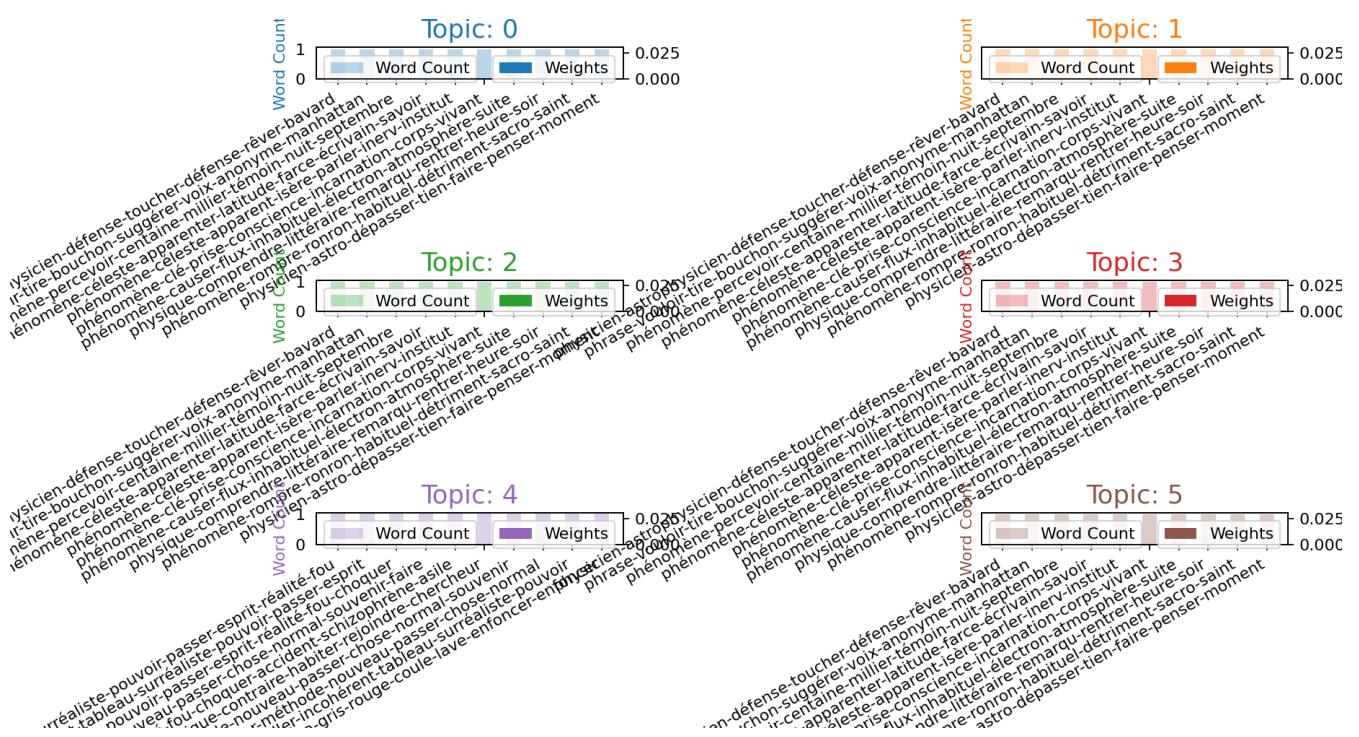


FIGURE 24 –
Graphique-Model11-heptagrams-500passes

Oublions dans un premier temps les questions liées à la forme. Nous reviendrons plus tard sur les problèmes de lisibilité que posent ces visualisations à partir d'un certain nombre de *topics*.

Mais faisons plutôt ici quelques observations sur les résultats des listes de n-grams. De manière intéressante, trois mots sont souvent au centre des n-grams les plus fréquents : « phrase », « physique » et « phénomène ». Ils sont ensuite associés à une multitude de combinaison de mots différents. Les termes « physique » et « phénomène » montrent l'omni-présence de thématiques liées à des questionnements scientifiques. Ils sont souvent associés à des concepts propres aux découvertes scientifiques qui marquent la période de production de ces œuvres : ainsi, on trouve par exemple les termes « électron » et « atmosphère ».

Le terme « physicien » apparaît également de manière récurrente dans l'ensemble des visualisations que nous venons de présenter. Il montre que les thématiques relatives à des questionnements scientifiques apparaissent de manière incarnée, souvent sous la figure d'un personnage central, dans les œuvres qui composent notre corpus.

Toutefois, il faut peut-être ici nuancer les observations que nous menons sur nos résultats. Comme on peut le voir sur les graphiques, l'importance relative des groupes de n-grams au sein de chaque *topic* est négligeable, et décroît à mesure que l'on augmente le nombre de grammes pris en compte. Cela ne semble pas anormal : plus les n-grams auxquels on s'intéresse comptent de mots, moins ils ont de chance d'être retrouvés en ces termes exacts ailleurs dans le texte. Cependant, cette observation invite à la prudence lorsqu'il s'agit de tirer des conclusions de ces résultats.

3.2 Intérêts et limites des résultats obtenus

Il nous reste désormais à conclure, et à mettre en perspective les résultats obtenus : quels enseignements peut-on tirer de ces travaux, pour l'histoire des représentations associées au futur d'une part, et d'autre part, d'un point de vue purement méthodologique ? Quels auront été les intérêts de ces recherches ? Quels en sont les limites, et quelles perspectives s'ouvrent à nous désormais ?

Un intérêt méthodologique avant tout

Tout d'abord, évoquons la question des résultats obtenus et de leur contribution à l'histoire des représentations associées à l'avenir. A partir des modèles obtenus et des différentes visualisations qu'on a présentées, on obtient quelques éléments thématiques intéressants pour sonder l'imaginaire projectionniste de la période étudiée. La présence relativement abondante du thème de la conquête et de l'exploration spatiales, en particulier, semble indiquer que la projection de la civilisation au-delà des frontières terrestres renvoie à un questionnement largement partagé dans la littérature d'anticipation contemporaine. De même pour les deux thématiques du voyage et de la famille, qui renvoient à des mutations socio-économiques propres à la période d'après-guerre, et imprègnent l'imaginaire social d'après-guerre.

En outre, la grande fréquence d'apparition des termes « physique » et « phénomène », ainsi que la figure omniprésente du physicien, témoignent d'une époque marquée par les transformations technologiques. Les représentations associées au futur en sont d'autant plus empreintes de questionnements scientifiques, questionnements qui, semble-t-il transcendent la littérature d'anticipation contemporaine.

On s'étonnera, en revanche, de l'absence de certaines représentations dont on aurait pensé qu'elles se retrouveraient très fréquemment dans l'imaginaire associé au futur. On pense notamment à la thématique de l'énergie nucléaire, et en particulier à l'idée d'une apocalypse nucléaire. Même si on retrouve le mot « radiation » dans le deuxième *topic*, on en compte relativement peu d'occurrences, et peu ou pas d'autres termes liés à cette thématique.

Ces observations permettent donc de mettre en perspective, grâce à des résultats obtenus de manière quantitative, ce que l'on sait des représentations futuristes de l'époque. Elles montrent tout l'intérêt de notre méthode pour sonder et appréhender l'imaginaire projectionniste d'une période.

Cependant, il faut bien reconnaître qu'il s'agit encore d'un matériau insuffisant pour bâtir véritablement une histoire des représentations associées à l'avenir. À ce stade, ces analyses peuvent constituer, au mieux, un outil complémentaire pour les historiens des

représentations. L'intérêt de cette étude, en termes de résultats, semble donc limité. Mais nous reviendrons plus tard sur les limites de cette étude. Car si les résultats de ces recherches ne constituent pas un matériau suffisant pour l'histoire des représentations, l'intérêt méthodologique, lui, est tout à fait remarquable.

D'abord, parcqu'il est certain que tout le questionnement préalable qui constitue le point de départ de cette enquête, et qui concerne les représentations associées au futur et les manières de les apprêhender, est d'intérêt pour quiconque s'intéresse à l'histoire des représentations. En particulier, toute la discussion sur l'existence d'un imaginaire social et de représentations associées au futur, et sur le lien entre cet imaginaire et la littérature d'anticipation, défriche bon nombre de questionnements auxquels les réponses sont loin d'être évidentes.

Ensuite, parce que l'utilisation du *topic modeling* ouvre de nombreuses perspectives d'application sur des corpus littéraires, même si elle ne permet pas encore de mener à des résultats aboutis dans le cadre de notre étude. Il s'agit, en effet, d'une méthode particulièrement intéressante pour explorer de vastes corpus et systématiser leur analyse. Or, puisque la quantité d'oeuvres littéraires croît exponentiellement, c'est précisément l'un des défis qui se posent aux chercheurs en histoire de la littérature. Nous proposons donc, avec cette exemple d'application, d'explorer une voie d'avenir en posant quelques bases essentielles.

Notre méthode ainsi que notre démarche, en ce qu'elles ouvrent de nombreuses perspectives pour la recherche en histoire de la littérature, possèdent donc un intérêt certain. Il faut cependant reconnaître qu'elles sont limitées dans les résultats qu'elles apportent sur l'objet même de notre étude, c'est-à-dire l'étude de l'histoire des représentations associées au futur.

Les limites de notre étude

Abordons donc ici la question des limites de nos recherches. D'abord, sur le plan de la forme, nous nous intéresserons à des aspects liés à la présentation et aux outils de visualisation de nos résultats. Ensuite, sur celui du fond, nous évoquerons les différentes conclusions auxquelles nous sommes parvenus, en soulevant la question des limites de nos résultats en termes d'apport à la recherche.

Sur le plan de la forme, donc, nous pouvons évoquer quelques problèmes de lisibilité sur certains de nos graphiques. En particulier, avec les analyses de n-grams, qui deviennent assez chargées dès lors qu'on observe des listes de plus de trois grammes. En outre, le format à respecter pour la bonne insertion de ces graphiques et nuages de mots au sein de notre mémoire nous oblige à limiter la taille de ces graphiques, n'arrangeant pas la question de leur lisibilité. On peut également faire un commentaire sur les visualisations

des résultats qu'on obtient avec LDAvis, pour lesquelles on a du se contenter d'intégrer une image et un lien. Pourtant, la page internet à laquelle on accède en obtenant ces résultats est nettement plus intéressante, en ce qu'elle est conçue de manière interactive et offre de belles perspectives d'exploration des *topics*.

Ces considérations triviales peuvent sembler assez anecdotiques, mais il est vraiment important d'offrir des visualisations particulièrement claires et parlantes dans le cadre de notre étude : elle qui se fonde déjà sur une méthode compliquée, peu évidente à comprendre pour les non-initiés, doit pouvoir prétendre à une présentation de résultats claire. Aussi, il est essentiel de bien comprendre les limites des possibilités de visualisation des *topic models* obtenus en résultats. Il faut par exemple relever que, pour des raisons de clarté, toutes ces visualisations se contentent de présenter les dix – pour nos wordclouds, graphiques - ou trente mots – pour LDAvis - les plus pertinents, mais qu'ils ne présentent pas l'ensemble des mots associés à chaque *topic*. Pour conclure brièvement sur cet aspect, il s'agit donc pour nous de rester conscients des limites de nos présentations de résultats et de rester ouverts à toutes nouvelles perspectives de visualisations des *topic models*.

De manière plus intéressante, évoquons maintenant la question plus délicate des limites de nos résultats sur le plan du fond. Il faut ici admettre que nos résultats sont assez limités dans l'apport qu'ils font à l'histoire des futurs envisagés, alors même que c'était là l'objectif premier de cette étude. Si on parvient à identifier quelques éléments thématiques structurants pour les représentations futuristes, si quelques termes et sujets intéressants se distinguent, il serait prétentieux d'en déduire que notre méthode permet d'explorer les imaginaires et les représentations de manière aboutie. Il nous faut donc ici aborder rétrospectivement la question de l'intérêt du choix du *topic modeling* comme méthode, qui bien que très intéressant dans le cadre de notre étude, ne permet pas d'aboutir à un matériau suffisant pour faire de l'histoire des futurs envisagés.

Peut-être qu'une des pistes pour comprendre l'échec de notre méthode réside en notre volonté de sonder le contenu de notre corpus à des fins interprétatives, et non à des fins typologiques. La plupart du temps, en effet, le *topic modeling* permet d'explorer de très vastes corpus sans intervention humaine à des fins de classification thématique du contenu : ainsi, les différents *topics* obtenus, ainsi que les termes qui les composent, n'ont que très peu d'intérêt *per se*. Ce qui importe avant tout, c'est de parvenir à classer le contenu en différentes catégories pertinentes, quand bien même la cohérence interne de chacune de ces catégories n'est pas optimale.

Au contraire, dans le cadre de notre démarche, c'est la cohérence de ces *topics*, la pertinence des termes qu'ils révèlent, qui nous intéressent, puisque ceux sont ces éléments qui sous-tendent l'exploration thématique de l'imaginaire social et des représentations. Si nous avions utilisé le *topic modeling* à des fins de classification des différents textes

de notre corpus, et *a fortiori*, afin de dresser une typologie des œuvres de science-fiction fondée sur leur contenu thématique, aurait-on obtenu des résultats plus pertinents ? C'est là une question qui mérite d'être posée, et une perspective intéressante pour la suite de nos recherches. Mais nous y reviendrons plus tard dans notre exposé. Retenons surtout ici que bien qu'elle offre des résultats pertinents dans le cadre de notre étude, notre méthode n'a pas les moyens de ses ambitions en termes d'exploration des représentations.

En fait, il semblerait surtout que l'utilisation exclusive du *topic modeling* soit une limite pour notre étude. Si cette méthode permet d'obtenir des résultats intéressants pour appréhender et sonder un vaste corpus comme le nôtre, est-ce pour autant suffisant pour explorer les imaginaires, et plus précisément les futurs envisagés d'une époque ? Ce qui est certain, c'est que l'analyse en *topic modeling* ne permet pas, seule, d'aboutir à des résultats suffisants pour servir de base à une histoire des représentations liées à l'avenir. C'est justement pour cette raison que nous terminerons notre étude en évoquant les méthodes complémentaires que nous pourrions envisager d'utiliser, ainsi que les différentes perspectives de recherche qui s'offrent à nous pour la suite.

Méthodes complémentaires et perspectives de recherche

Nous commencerons ici par évoquer les différents moyens qui s'offrent à nous pour enrichir notre méthode et poursuivre nos travaux sur l'histoire des futurs envisagés, avant d'évoquer plus en profondeur les perspectives de recherches qui pourraient être exploitées pour la suite. Quelles analyses complémentaires, quels outils pouvons nous envisager d'exploiter pour affiner notre appréhension des représentations liées à l'avenir ? Quelles perspectives s'offrent à nous pour la suite de nos recherches ?

Commençons par rappeler la possibilité d'utiliser d'autres techniques d'implémentation du *topic modeling*. On a expliqué que la LDA n'était pas la seule méthode existante pour la modélisation de sujets. On a évoqué brièvement les modèles de la LSA et la PLSA. Mais là encore, il nous appartient de suivre l'actualité des méthodes et usages en la matière, pour essayer de mettre en œuvre d'autres méthodes parallèlement à notre algorithme de LDA. On pourra ici se référer à un ouvrage plus récent (2012), et relativement complet, sur les méthodes de l'analyse statistiques de données textuelles². On consultera notamment les chapitres 4 et 5, respectivement consacrés au *clustering* de textes, et au *topic modeling*, qui nous renseignent précisément et clairement sur les différentes alternatives qui s'offrent à nous. On pense notamment à un autre type de modélisation de sujet automatique, la Non-Negative Matrix Factorisation (NMF), qui pourrait se révéler assez pertinent dans le cadre de notre analyse.

Dans un second temps, on peut envisager d'enrichir notre analyse en couplant notre

2. AGGARWAL C., ZHAI C. « Mining Text Data », Springer, 2012.

algorithme de LDA à d'autres méthodes d'analyses textuelles complémentaires. Dans le cadre d'un défi proposé annuellement par l'association EGC (Extraction et Gestion des Connaissances), un article paru en 2016 a attiré notre attention³ : l'équipe de recherche y montrait l'intérêt d'utiliser conjointement les techniques de topic modeling et d'extraction de motifs récurrents, pour mettre en évidence non seulement les mots partageant des relations thématiques, mais aussi leurs relations fréquentes, au niveau intra et inter thématiques. En confrontant leurs méthodes aux données du Défi, l'équipe a ainsi pu aboutir à une utilisation probante de ces méthodes couplées, et à une représentation intéressante, en bi-graphe, des relations entre les mots et les *topics*. On a donc là un bon moyen d'affiner notre analyse de topics, en s'intéressant aux associations récurrentes des mots, et donc à l'usage qui est fait de ces mots.

Analogiquement, on pourra également employer l'un des modules développés par un autre étudiant du master, Michel Capot, pour travailler sur les cooccurrences lexicales⁴. Le module porte le nom de *lexnet* et se fonde, à l'instar de notre module, sur l'utilisation de bibliothèques extérieures (NetworkX, pandas, scipy, matplotlib et numpy). À partir d'un fichier .tsv contenant un texte annoté, il permet d'obtenir différents graphes pondérés pour visualiser les cooccurrences lexicales au sein du texte. On note également la possibilité de faire certains calculs fondés sur les propriétés des graphes, notamment le taux d'intersection de différents champs lexicaux qui pourrait s'avérer intéressant sur notre corpus. Ce module permettrait d'explorer d'une autre manière la sémantique et le lexique de notre corpus de textes.

Il faut également évoquer la dimension temporelle, que nous avons ici volontairement écartée en nous concentrant sur une période relativement restreinte. Il s'agit pourtant là d'une des perspectives de recherche majeures pour la suite de nos travaux. Il s'agirait donc d'employer des méthodes qui permettent de sonder thématiquement un corpus composé d'oeuvres étalées sur une plus vaste période, et d'apprécier justement l'évolution sémantique des représentations et des images à travers le temps. Bien entendu, de premières analyses pourraient être réalisées très simplement, en découplant notre corpus en plusieurs sous-corpus pour chaque période et en comparant les résultats de chaque sous-corpus. Cette méthode reposeraient, certes, sur un découpage un peu arbitraire des différentes sous-périodes ; mais elle aurait le mérite de la clarté, et permettrait sans doute de faire ressortir et d'apprécier différentes évolutions historiques.

Mais une méthode plus élégante doit ici être évoquée, d'autant plus qu'elle semble de plus en plus prometteuse et que des chercheurs de nombreux champs la mettent à contribution, dans le cadre de travaux variés. Celle des *word embeddings* – plongement

3. ALIGON, Julien, GUILLET, Fabrice, BLANCHARD, Julien, PICAROUGNE, Fabien. « Analyse par Motifs Fréquents et Topic Modeling », paru dans le cadre du défi EGC, Laboratoire Informatique de Nantes Atlantique, 2016.

4. Pour consulter le github du module en question, voir <https://github.com/johnloque/lexnet>.

lexical en français, et en particulier de ce qu'on appelle les *diachronic word embeddings*, qui s'intéressent à l'évolution historique du lexique. On citera notamment l'article de référence de William Hamilton, intitulé *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*⁵. Cet article paru en 2016 a permis de poser bon nombre de jalons, et notamment deux lois statistiques de l'évolution sémantique, pour mieux comprendre et analyser l'évolution historique du lexique. Depuis, on assiste à une profusion de littérature sur le sujet, présentant de nouvelles méthodes et des travaux intéressants. Une thèse soutenue en 2021 à l'université Paris-Saclay⁶ rend bien compte de l'état de l'art en la matière, en même temps qu'elle propose une méthode innovante et des exemples d'application pratiques. Elle utilise notamment des méthodes d'apprentissage de plongements contextualisés à l'aide de modèles de langue pré-entraînés, comme BERT. Particulièrement probante dans le cadre des différents exemples d'applications présentés, cette méthode ouvre des perspectives prometteuses pour sonder l'évolution historique des représentations sémantiques liées à l'avenir dans un corpus de littérature d'anticipation.

Enfin, une dernière perspective de recherche, plus ambitieuse et lointaine, peut être mentionnée. On a vu que la science-fiction était un genre complexe, qui peinait à être défini, et que les topologies existantes des sous-genres de la SF n'étaient pas pleinement satisfaisantes. En ce sens, pourquoi ne pas également employer cette exploration thématique à des fins de catégorisation de notre objet ? On pourrait à terme, en constituant un corpus de textes suffisamment conséquent, envisager de proposer une classification thématique des sous-genres de la SF, en se fondant sur nos méthodes de *topic modeling*, ainsi que sur d'autres méthodes de clustering. Il faudrait évidemment un corpus de textes suffisamment important pour qu'il soit représentatif de la littérature SF dans son ensemble, et en ce sens, le choix même des textes à inclure dans le corpus impliquerait un effort épistémologique remarquable. Il s'agirait là d'un travail fastidieux, ne serait-ce que pour constituer un corpus conséquent. Cependant, ce projet pourrait se révéler prometteur. Il permettrait peut-être d'apporter une réponse statistique plus neutre à ce problème de catégorisation de la littérature de science-fiction, qui semble aujourd'hui avancer davantage grâce aux vues lucides et à l'érudition de certains chercheurs, que grâce à des méthodes stables de classification.

5. . HAMILTON William, LESKOVEC Jure, JURAFSKY Dan. « *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change* », arXiv e-prints, Computer Science Department, Stanford University, 2016.

6. MONTARIOL Syrielle, *Models of diachronic semantic changes using word embeddings. Document and Text Processing*. Université Paris-Saclay, 2021.

Conclusion

Ce mémoire propose donc d'explorer les *futurs envisagés*, objet d'étude original dont on comprend aisément l'intérêt pour la connaissance historique. En s'appuyant sur différents concepts existant, et notamment en mobilisant les travaux de l'histoire des représentations et de la sociocréditique, nous avons su donner une définition claire de la notion de *futurs envisagés*. Nous avons également mené une longue réflexion sur les modalités d'accès à ces futurs envisagés, c'est-à-dire aux représentations liées à l'avenir et au lien que ces imaginaires entretiennent avec la socialité ambiante, à travers la littérature en particulier. À partir de ces différents éléments, nous avons défini un support d'investigation, la littérature d'anticipation, qu'on a présenté longuement. Ainsi, nous avons finalement constitué un vaste corpus de littérature d'anticipation contemporaine, propice à l'étude des futurs envisagés. Puis, nous avons construit une méthode innovante, fondée sur l'utilisation d'un outil informatique novateur : le *topic modeling*. Nous avons donc mis au point une méthode pour explorer thématiquement notre corpus, et ainsi accéder aux représentations du futur dans notre corpus. En construisant un module dédié à l'utilisation du *topic modeling*, en créant un modèle adapté à notre corpus et en l'entraînant sur ces données, nous avons pu obtenir des résultats satisfaisants, présentés à l'aide de visualisations adaptées. En présentant et en analysant ces résultats, nous avons finalement pu rendre compte de l'intérêt de notre méthode, tout en évoquant franchement ses limites. Aussi, nous avons pu ouvrir la réflexion sur les différentes perspectives envisageables pour poursuivre ces travaux et mieux comprendre les différentes représentations associées au futur, et leur évolution historique. Pour conclure, ces travaux ont donc permis d'ouvrir la voie à l'étude de l'histoire des représentations liées à l'avenir, et des projections futuristes. En défrichant une discussion épistémologique nécessaire, et en montrant l'intérêt d'une telle étude, nous avons posé de précieuses bases pour la constitution de méthodes ultérieures. En outre, si notre méthode est encore lacunaire, l'idée d'une exploration thématique semble prometteuse, et il est certain qu'en la perfectionnant, et en l'enrichissant des outils complémentaires que nous avons proposé d'utiliser, nous parviendrons à constituer un matériau suffisant pour étudier l'histoire des futurs envisagés. Enfin, notre étude semble offrir de belles perspectives pour l'étude de la science-fiction, à des fins typologiques, et il serait intéressant d'explorer cette voie dans de futurs travaux.

Annexes

1. Bibliographie du Corpus

Collection Présence du futur de Denoël

- La sortie est au fond de l'espace, Jacques Sternberg, 1956
- Le Voyageur imprudent, René Barjavel, 1958
- La Mort de la Terre, J.-H. Rosny aîné, 1958
- Voyage au pays de la 4e dimension, Gaston de pawlowski, 1962
- Niourk, Stefan Wul, 1970
- Retour à la Terre 1, Jean-Pierre Andrevon, 1975
- La Grande Quincaillerie, Georges Soria, 1976
- Shéol, Jean-Pierre Fontana, 1976
- Le Désert du monde, Jean-Pierre Andrevon, 1977
- Paysages de mort, Jean-Pierre Andrevon, 1978
- Cinq solutions pour en finir, Dominique Douay, 1978
- Le dormeur s'éveillera-t-il ?, Philippe Curval, 1979
- Compagnons en terre étrangère 1, Jean-Pierre Andrevon, 1979
(Recueil de nouvelles)
- Compagnons en terre étrangère 2, Jean-Pierre Andrevon, 1980
(Recueil de nouvelles)
- La Guerre olympique, Pierre Pelot, 1980
- Vue en coupe d'une ville malade, Serge Brussolo, 1980 (Recueil de nouvelles)
- Neutron, Jean-Pierre Andrevon, 1981

- L’Odeur de la bête, Philippe Curval, 1981
- Damiers imaginaires, Emmanuel Jouanne, 1982
- Mourir au hasard, Pierre Pelot, 1982
- Portrait du diable en chapeau melon, Serge Brussolo, 1982
- Les Faiseurs d’orages, Jean-Pierre Hubert, 1984
- Un navire de nulle part, Antoine Volodine, 1986
- Autoportrait, Jean-Claude Dunyach, 1986
- Demain les puces, Patrice Duvic, 1986 (Recueil de nouvelles)
- Superfuturs, Philippe Curval, 1986 (Recueil de nouvelles)
- Cruautés, Emmanuel Jouanne, 1987
- Le Château d’encre, Serge Brussolo, 1988
- Script, Alain Darteville, 1989
- Fragments du rêve, Jean-Pierre Vernay, 1990 (Recueil de nouvelles)
- Guerre de rien, Jacques Barbéri, 1990

Collection Anticipation de Fleuve Noir

- Les Astres morts, Jean-Gaston Vandel, 1952
- L’Invasion de la Terre, Jimmy Guieu, 1952
- Alerte aux robots!, Jean-Gaston Vandel, 1952
- Hantise sur le monde, Jimmy Guieu, 1953
- Le Monde oublié, Jimmy Guieu, 1954
- Territoire robot, Jean-Gaston Vandel, 1954
- La Foudre anti-D, Jean-Gaston Vandel, 1956
- La Mort de la vie, Jimmy Guieu, 1957
- La Peur géante, Stefan Wul, 1957
- Cité Noë n°2, Jimmy Guieu, 1957
- Chocs en synthèse, Max-André Rayjean, 1958
- Ère cinquième, Max-André Rayjean, 1959

- Réaction déluge, F. Richard-Bessière, 1959
- Le Péril des hommes, Max-André Rayjean, 1960
- Survie, Peter Randa, 1960
- L'Ère des Biocybs, Jimmy Guieu, 1960
- Chirurgiens d'une planète, Gilles d'Argyre (alias Gérard Klein), 1960
- Les Éphémères, Peter Randa, 1962
- Les Portes de Thulé, Jimmy Guieu, 1964
- Je m'appelle... "tous", F. Richard-Bessière (alias François Richard et Henri Bessière), 1965
- Le Quatrième futur, Max-André Rayjean, 1967
- Le Sceptre du hasard, Gilles d'Argyre (alias Gérard Klein), 1968
- Solution de continuité, Doris Le May et Jean-Louis Le May, 1969
- Les Enfants de l'histoire, Kurt Steiner (alias André Ruellan), 1969
- Demain le froid, Doris Le May et Jean-Louis Le May, 1969
- Année 500.000, Daniel Piret, 1972
- L'Empreinte de Shark Ergan, Doris Le May et Jean-Louis Le May, 1973
- 1973... et la suite, F. Richard-Bessière (alias François Richard et Henri Bessière), 1973
- Les Bioniques d'Atria, Pierre Barbet, 1973
- Le Monde figé, Max-André Rayjean, 1973
- L'Iceberg rouge, Maurice Limat, 1974
- Projet Apocalypse, Georges Murcie, 1974
- Le Vieux et son implant, Paul Béra, 1975
- Pandémoniopolis, Gabriel Jan, 1975
- Vendredi, par exemple..., Pierre Suragne, 1975
- L'Autoroute sauvage, Gilles Thomas (alias Julia Verlanger), 1976
- Nuit d'émeute, Paul Béra, 1977
- Concentration 44, Gabriel Jan, 1978

- Les Maîtres verts, Gabriel Jan, 1978
- L'Ordre établi, Christopher Stork (alias collectif d'écrivains belges), 1979
- Et les hommes voulaient mourir, Dan Dastier, 1979
- Bactéries 3000, André Caroff, 1979

2. Dossier annexe

2.1 - Corpus

Ce dossier contient les différents ouvrages de notre corpus au format .txt, déjà standardisés selon le procédé exposé auparavant. Ceux sont ces fichiers qui constituent nos données brutes, avant la phase de *preprocessing*.

2.2 - Preprocessed Data

Ce dossier contient notre corpus après la phase de pré-traitement des données qu'on a expliquée – c'est-à-dire après tokenisation et lemmatisation. Outre le corpus lemmatisé, il contient également les différentes listes de n-grams qui serviront de base pour entraîner notre modèle.

2.3 - Sampled Data

Ce dossier contient les données de notre corpus, et des différentes listes de n-grams, après la phase d'échantillonnage.

2.4 – TopicModel

Ce dossier contient les différents modèles entraînés, à la fois sur le corpus et sur les listes de n-grams, en faisant varier les paramètres.

2.5 - Results

Ce dossier contient les différents résultats que nous avons obtenus – sous forme de figures, wordclouds, graphiques, etc. Ceux sont ces figures qui ont été présentées auparavant.

2.6 – Module TM

Ce dossier contient les scripts en python de notre module, divisé en différentes parties correspondant à chacune des étapes de notre méthode. Il contient également deux fichiers texte, readme.txt et requirements.txt, qui correspondent respectivement à une brève présentation du module et aux différents packages requis pour réemployer le module. Enfin, le module contient un script python TMAssessment.py, qui correspond au script utilisé dans le cadre de nos recherches, et qui peut servir d'exemple pour réemployer notre module.

Table des matières

Résumé	iii
Abstract	iv
Introduction	i
Bibliographie	vii
1 La littérature d'anticipation comme support d'investigation	1
1.1 Du questionnement d'historien	1
1.2 ...à la définition d'un support d'investigation adapté	11
1.3 Constituer un corpus de littérature d'anticipation	22
2 Construire une méthode computationnelle pour explorer notre corpus	29
2.1 Le <i>topic modeling</i>	30
2.2 Construction d'un module	35
2.3 Elaboration d'un modèle entraîné sur les données de notre corpus	40
3 Résultats et retours méthodologiques	47
3.1 Présentation des résultats	47
3.2 Intérêts et limites des résultats obtenus	69
Conclusion	75

Table des figures

1	Graphique de cohérence du modèle en fonction du nombre de topics	42
2	Wordcloud-Model5-6topics-3000passes (sans TF-IDF)	43
3	Wordcloud-Model4-6topics-5000passes (avec TF-IDF)	44
4	Wordcloud-Model12-6topics-5000passes (avec TF-IDF)	48
5	Wordcloud-Model12-6topics-5000passes (avec TF-IDF)	49
6	LDAvis-model12-notopicselected	51
7	LDAvis-model12-topic1selected	52
8	LDAvis-model12-topic2selected	52
9	LDAvis-model12-topic3selected	53
10	LDAvis-model12-topic4selected	53
11	LDAvis-model12-topic5selected	54
12	LDAvis-model12-topic6selected	54
13	WordCloud-Model6-bigrams-500passes	56
14	Graphique-Model6-bigrams-500passes	57
15	WordCloud-Model7-trigrams-500passes	58
16	Graphique-Model7-trigrams-500passes	59
17	WordCloud-Model8-tetragrams-500passes	60
18	Graphique-Model8-tetragrams-500passes	61
19	WordCloud-Model9-pentagrams-500passes	62
20	Graphique-Model9-pentagrams-500passes	63
21	WordCloud-Model10-hexagrams-500passes	64
22	Graphique-Model10-hexagrams-500passes	65
23	WordCloud-Model11-heptagrams-500passes	66
24	Graphique-Model11-heptagrams-500passes	67