Mesures et savoirs : Quelles méthodes pour l'histoire culturelle à l'heure du big data ?

Marianne Reboul, Alexandre Gefen

Mesures et savoirs : quelles méthodes pour l'histoire culturelle à l'heure du big data ?

L'analyse quantitative de l'histoire culturelle a été ouverte par la mise à disposition de corpus de masse tel que celui de Google books (500 milliards de mots, 5 millions d'ouvrages, soit environ 4% de la littérature mondiale) et a été popularisé sous le nom de « culturonomics ». Elle s'ouvre désormais aux chercheurs, en promettant un accès profond aux faits culturels et à leurs évolutions qui affleurent à travers leurs traces textuelles dans les corpus textuelles numérisées. Encore faut-il pouvoir interroger ces corpus dont la taille et la nature posent des problèmes scientifiques nouveaux, leur dimension les rendant illisibles directement et mettant échec les méthodes de fouille et les outils traditionnels d'analyse statistique des données en imposant des méthodes statistiques nouvelles et le saut vers des formes d'intelligence visuelles originales. Dans le cadre d'un projet mené entre le Labex « Obvil » de Paris-Sorbonne et le Literary Lab de Stanford sur l'histoire de l'idée de littéraire (la définition de la littérature comme mot, comme concept et comme champ), et visant à produire une histoire *empirique* de la littérature, nous avons mené depuis deux ans des expériences de fouille d'un corpus de critique littéraire de 1618 titres, 140 millions de mots (dont plus de 50 000 occurrences du lemme « littérature ») de la fin de l'Ancien Régime à la Seconde Guerre mondiale. En présentant des exemples développés dans cette première expérimentation à grande échelle de mesure de l'histoire des idées, on présentera les méthodes de *text mining* contemporaines en essayant d'éprouver leur pertinence heuristique et de leur capacité à faire remonter des données signifiantes pour l'histoire et la théorie littéraire. On fera l'hypothèse que toute enquête quantitative sérieuse mobilise désormais non une échelle intermédiaire standard et immédiatement lisible, mais le maniement d'outils statistiques dont l'interprétation en sciences humaines pose des problèmes particuliers qui, paradoxalement, ne peuvent être résolus que par leur articulation étroite à du *close reading* et à des mesures fines.

1. L'intelligence des corpus

Loin d'être neutres et interrogeables de manière transparente, les corpus relèvent d'un travail de construction que les méthodes de fouilles de masse contemporaines propres au *big data*, supposément capables de dissimuler le « bruit » (erreurs de numérisation, textes surnuméraires, etc.) par l'effet de quantité ne rendent pas moins importants. Il paraît ainsi impossible de travailler autrement que pour obtenir des indications générales sur un corpus « plat » comme celui utilisé par Google Ngram Viewer (https://books.google.com/ngrams) soit l'ensemble des livres numérisés par Google books (6% de la littérature publiée dans le monde) où des traités de médecine voisinent des gazettes de modes. Les expériences de « culturonomique » (quantification des tendances culturelles) les plus simples et les plus spectaculaires montrent très vite ces enjeux. Si certains phénomènes sont détectables par des indices, par exemple certains faits culturels d'émergence : la renommée nouvelle d'un personnage historique est ainsi en général facile à détecter par son apparition dans le corpus de masse de Google, mais avec un décalage temporel et un manque de finesse évident qui ne rend pertinent de telles analyses que sur la très longue durée, constat vrai a fortiori si l'on interroge non un personnage mais un concept. Ainsi ce graphique extrait de Google Ngram (figure 1) renseignant sur le nombre d'occurrences des mots « Tocqueville » et de « démocratie » dans l'histoire moderne (la fréquence du mot dans le corpus découpé par année), qui donne une claire idée de la puissance de l'outil (rendre visible la période où s'est affirmée l'importance de Tocqueville, faire voir sur la très longue durée l'émergence de la démocratie comme objet de discours) mais aussi ses limites : il est difficile de savoir comment et dans quel champ a émergé la pensée politique de Tocqueville, de peser son influence intellectuelle par delà sa visibilité, de la corréler avec d'autres philosophes, il est impossible de savoir si c'est la démocratie comme concept ou comme pratique qui est bien l'objet des discours, de typer ces discours, etc. malgré les possibilités d'affiner les recherches (dans la figure 2, en demandant les adjectifs les plus

fréquemment employés en postposition au substantif « démocratie », on voit très facilement la mode puis le déclin du concept de « démocratie populaire » et la montée tardive de la « démocratie directe » par exemple).

Figure 1 : Google Ngram pour « Tocqueville » et « démocratie »

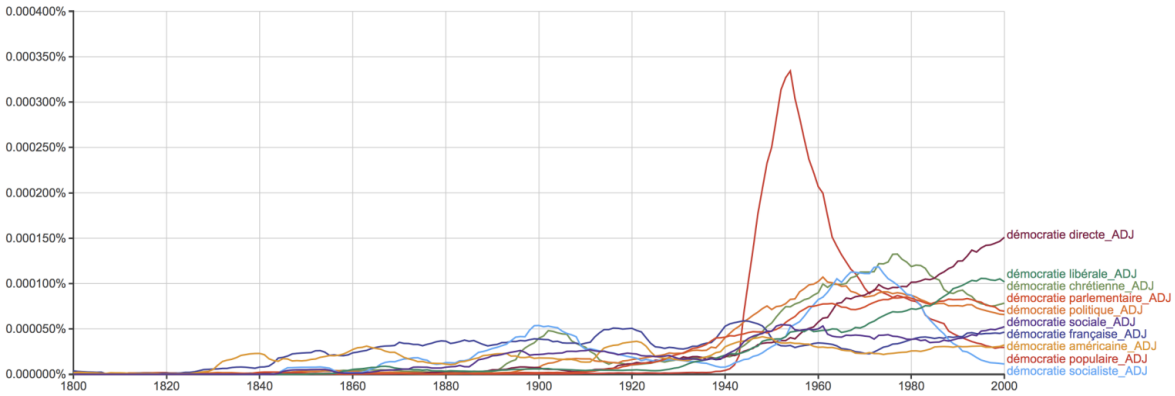

Figure 2 : Google Ngram pour adjectifs post-posés

L'interrogation, très impressionnante, du Google Ngram Viewer, ne saurait donc dispenser « de réfléchir aux questions d'échantillonnage, de représentativité, d'homogénéité du corpus, des conditions dans lesquelles les données sont produites, agrégées ou combinées », comme le rappelle Michel Wieviorka (Wieviorka 2013:26), en particulier lorsqu'il s'agit de fouiller des données ouvertes, mais dont les standards de constitution ne sont pas clairement connus – ce qui est précisément le cas de Google Books, qui pose notamment de considérables problèmes de représentativité, puisque construit à partir de la numérisation de bibliothèques essentiellement américaines avec des métadonnées souvent fautives[1]. Ainsi, pour l'analyse numérique du *long data* de l'histoire culturelle, une étape essentielle consiste à définir, constituer ou s'approprier son propre corpus : numériser ou agréger de textes, les convertir, les équiper en métadonnées (métadonnées bibliographiques ou, plus finement, utiles à leur classement), ou cerner les biais posés par un corpus pré-constitué. Ces opérations sont loin d'être transparentes, pour de multiples raisons. En premier lieu, les acteurs des humanités numériques dépendent dans une large mesure des politiques de numérisation de masse, celles des bibliothèques en particulier ; or ces dernières sont elles-mêmes contraintes par l'état physique des documents et par la possibilité d'équiper un nombre considérable de textes en métadonnées pertinentes : le problème est particulièrement sensible pour les périodiques, dont

[1] Voir les analyses précises de Frédéric Glorieux, https://resultats.hypotheses.org/92, les remarques de Ted Underwood (Underwood 2012), et les réflexions générales de Sarah Zhang (Zhang 2015).

l'exploitation productive suppose un découpage par articles et par auteurs. Il se pose aussi par exemple pour les corpus fournis par Gallica, le projet de numérisation de la BNF, qui sont ne sont pas équipés de métadonnées « genre » ni de métadonnées thématiques (en dehors du classement DEWEY rétrospectif pour les ouvrages anciens), rendant par exemple difficile d'isoler les essais critiques. Deux autres facteurs s'opposent à l'idée d'une analyse strictement empirique : d'une part, il est difficile d'identifier de manière automatique ou manuelle, au sein des textes, des informations d'indexation signifiantes (en particulier ce qu'on appelle les « entités nommées », noms, dates et lieux), alors même que ce balisage constitue un préalable nécessaire à certains types d'études (de réception, par exemple) ; d'autre part, pour les textes anciens, les informations de tirage et de diffusion sont rares, ce qui rend assez peu fiable la mesure de leur représentativité. En somme, dans l'anarchie des corpus numériques des bibliothèques, le rêve d'exhaustivité d'une bibliothèque de Babel est aussi éloigné de sa réalisation que l'avènement espéré par Tim Berners-Lee d'un web sémantique. Le travail sur le *big data* doit se faire donc par bricolage, à partir de corpus partiellement représentatifs et partiellement organisés, imposant des biais considérables à toute analyse quantitative et requérant des chercheurs une conscience méthodologique aguerrie.

Par ailleurs, ni la conversion numérique des textes, ni leur modélisation structurelle par des normes[2] visant à les abstraire selon des règles générales en séparant contenu textuel originel, présentation originelle (découpage textuel, typographie, mise en page, etc.) et présentation finale sur ordinateur ou tablettes, ne sont transparentes : convertir des récits en des bases de données, passer de la narration ou de la diction à de l'information, c'est identifier des noms et des concepts signifiants, transposer des typographies et des mises en pages, annoter un texte « à l'ancienne » ou avec des outils collaboratifs innovants, démembrer une structure en repérant l'énonciation éditoriale et les enjeux esthétiques des supports. Il s'agit de déterminer ce qui relève du texte ou du paratexte, d'identifier des structures, de choisir les normes orthographiques, en faisant à chaque fois le départ entre information signifiante et information négligeable : on choisira, par exemple, pour un texte classique de conserver ou non les « privilèges d'impression », les formes de ponctuation et de typographie d'époque, les informations spatiales de mise en forme des pages, les données concernant la matérialité du texte ; on considérera ou non les « tomes », « livres » ou « parties » comme relevant d'un même fichier, etc. S'il existe des outils de détection automatique des structures et de rapprochement des termes sémantiques par association à des bases de données externes, et s'il est envisageable de s'appuyer par exemple sur des cartes de mots-clés signifiants générés par des algorithmes, toutes ces opérations relèvent d'une lourde ingénierie des connaissances. Elles présupposent des modélisations plus ou moins normatives. Les processus de sélection et de structuration des entités des textes impliquent donc des sémantiques – si ce n'est des ontologies –, qu'elles soient explicites et gérées par outils des arbres ad-hoc, ou implicites et apportées par la tradition.

Ainsi, la standardisation et la modélisation égalisatrice par les normes ouvertes, « démocratiques » et générales du Web et la délégation de certaines opérations de sémantisation à des algorithmes de classement empiriques et transversaux ne sauraient faire oublier à quel point les sémantiques restent indissociables de savoirs disciplinaires et d'une tradition interprétative sans laquelle ni l'établissement ni l'interprétation des œuvres ne font sens. Il n'y a donc pas une œuvre numérisée mais des standards et des degrés d'encodage et de sémantisation : entre un texte numérisé en « texte pur » comme on en trouve encore sur le site du Projet Gutenberg, et un texte sur lequel le moindre détail aura été encodé, de la couleur de l'encre à la signification d'une référence permettant de lier un texte à un autre, c'est par un facteur cinquante qu'auront été multipliés le temps de travail, la taille du fichier et la densité informationnelle exploitable (rendant au passage le fichier final XML quasi illisible avant sa transformation pour affichage en HTML). Le degré de détail pouvant être atteint par une transcription numérique est donc extrêmement variable et la transcription dans la norme TEI génère donc sa propre logique (voir l'exemple donnée en figure 3), ses règles et ses débats, ses normes et ses bons usages, extension numérique de la philologie classique et des problématiques des éditions papier, en circonvenant des impossibilités propres à l'édition papier (pensons, par exemple, à la possibilité d'encoder deux transcriptions ou deux

[2] Notamment la Text Encoding Initiative (TEI), standard de fait en numérisation patrimoniale littéraire utilisant la norme XML (Burnard 2010).

interprétations d'un même passage) grâce au prix de la maîtrise d'un langage d'encodage, d'un travail philologique et herméneutique lourd, et du déploiement d'outils logiciels de transcription et d'affichage complexes.

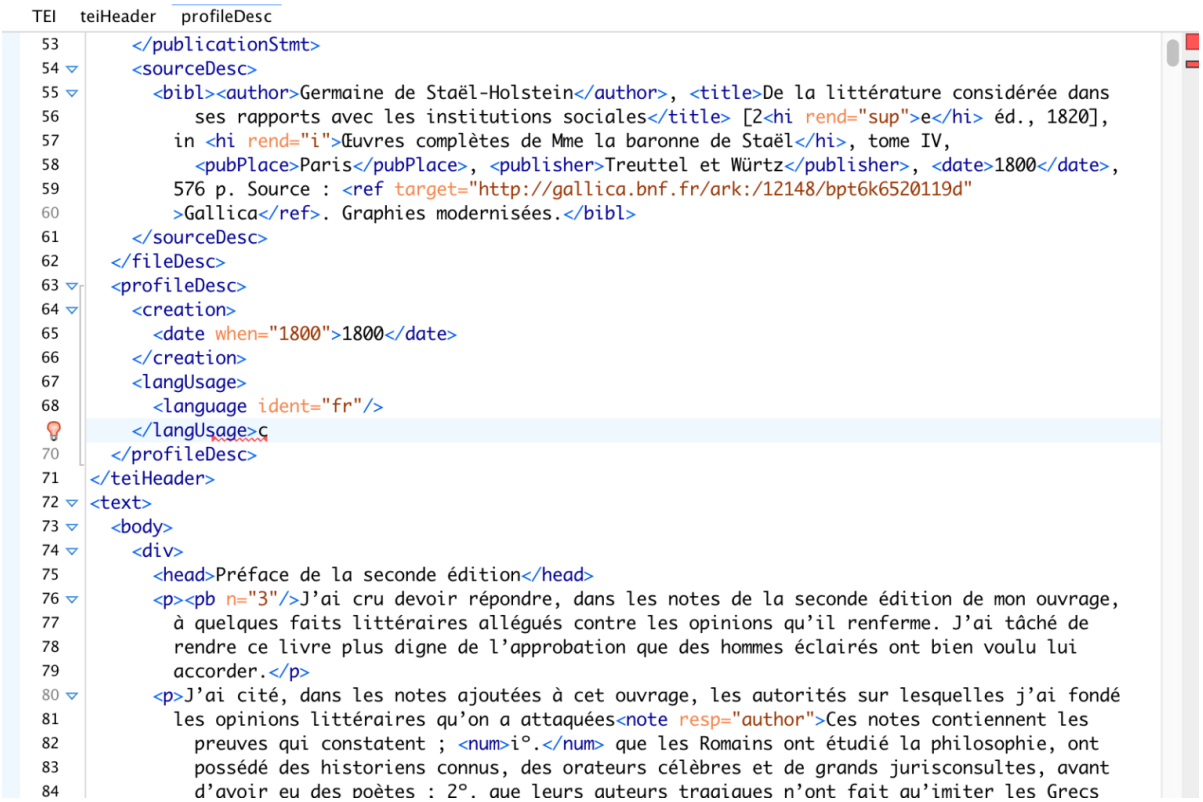

Figure 3 : Exemple de fichier aux normes TEI

L'« herméneutique intégrative » permise par le numérique ne considère donc le texte littéraire comme un objet informationnel ordinaire et empirique analysable par une machine qu'au prix d'une sélection a priori délicate (possibilité de l'agrégation dans un corpus d'une version particulière) et d'une décomposition structurelle et sémantique très complexe. Les méthodes permettant de faire de l'œuvre littéraire un ensemble de faits textuels ordinaires, examinables empiriquement dans leur connexion à d'autres faits textuels ou références, et susceptibles de comparaisons, supposent donc autant le recours à un savoir philologique et historique ancien avant toute expérimentation, et exposent donc tout savoir empirique aux aléas empiriques des conditions de sa constitution. Loin de dépasser le problème par la masse, l'apport des grands corpus faiblement édités et structurés, neutralisent assurément certains problèmes (des erreurs ponctuelles de graphies ou d'attribution), en en posant d'autres, certains biais statistiques étant difficilement détectables (comme des confusions systématiques dans des métadonnées par exemple). Pour donner des exemples plus précis propres au projet d'histoire culturelle qui nous occupe, l'intérêt de corpus large comme celui de Google Ngram se limite à des vérifications ou à des états généraux, faute de pouvoir être sûr des métadonnées et de la composition des corpus : débordant la critique littéraire, le corpus des Google books est un corpus de référence et de comparaison, mais c'est dans un corpus spécialisés et maîtrisés que nous avons cherché à cerner l'évolution du mot « littérature », ne serait que parce ce qu'il permettait de déployer des outils plus riches que le calcul fréquentiel de séries de mots permis par Google Ngram.

Ainsi par exemple de nos intuitions sur les périodes d'influence des philosophes allemands sur la critique française au XIXe siècle et sur les principaux passeurs de ces philosophies, graphes obtenu par la fouille fréquentielle du corpus critiques, qui ne trouvent sens que par rapport à un corpus maîtrisé de critique littéraire :

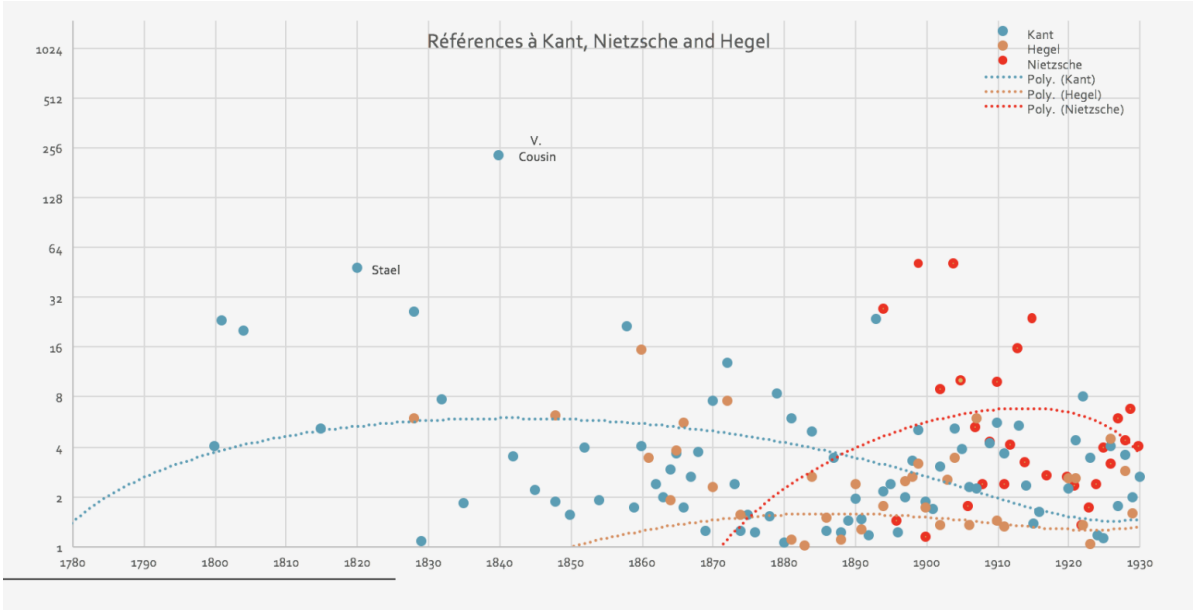

Figure 4 : Les auteurs de référence

2. Les méthodes de fouille textuelles

Reste à savoir quels méthodes de fouille textuelle il est possible d'utiliser à l'heure où nous écrivons. Face à des corpus dépassant plusieurs millions de mots et plusieurs milliers ou dizaines de milliers de textes par exemple, les outils communs de visualisation, type Voyant Tools (Sinclair & Rockwell 2014), ou de textométrie, type TXM (Heiden, Magué & Pincemin 2010), peuvent ne pas être efficaces, tout simplement parce qu'ils ne sont pas conçus pour les études de masse. Tout d'abord, le coût computationnel pour nombre de ce type d'outils est trop important, et d'autre part, s'ils sont capables de fournir les données, ils ne s'avèrent pas toujours capables de les trier. Obtenir des données statistiques n'est pas difficile. Le grand enjeu de l'analyse de masse est, d'une part, de vérifier la validité des données et résultats (qui peuvent très rapidement comporter des biais importants, minimes pour des micro-corpus, mais déterminants pour des macro-corpus), et, d'autre part, de faire le tri des résultats obtenus, de déterminer quels résultats peuvent être compris comme signifiants.

On ne peut pas réfléchir aux enjeux épistémologiques de la lecture des données en humanités numériques sans entrer dans le détail des outils disponibles, ce qui passe par l'explication d'un certain nombre de concepts clés de la fouille de textes.

Si le cerveau humain est capable, dans une moindre mesure, de déceler une structure fondamentale de portions textuelles (ne serait-ce que pour décrypter le sens d'une phrase, l'opération nécessitant un niveau minimum d'abstraction et de mémoire), il faut imaginer que la capacité, non seulement de stockage, mais d'analyse d'informations d'une machine est infiniment supérieure. D'autre part, une machine n'analyse pas un texte linéairement (du moins pas nécessairement) : le parcours des données peut être linéaire, mais il est bien souvent synchronique (il n'y a pas d'ordre de lecture des données pour une machine, sauf à le définir explicitement). Tout texte peut être alors considérablement enrichi, de façon durable, et chaque élément textuel recouvrir plusieurs informations. Lorsque ces enrichissements sont inscrits informatiquement, pour des éléments déterminés de chaque texte, on les appelle des métadonnées. L'analyse des corpus peut être brute, avec la simple considération des caractères, mais avec l'évolution des techniques d'analyses linguistiques, les fouilles de données ne se passent plus guère des métadonnées. Ces métadonnées sont généralement obtenues grâce à des outils d'analyse syntaxique et sémantique. Pour toute analyse de corpus de masse, il est donc nécessaire de passer par une phase importante de pré-traitement du corpus, qui se fait

généralement en trois étapes, à savoir la tokénisation, la lemmatisation et l'étiquetage syntaxique léger. C'est le cas pour notre corpus, entièrement enrichi, revu et disponible en XML.

La fouille de données part d'un principe simple : tous les éléments du corpus, depuis le caractère jusqu'à des séquences massives, sont des données quantifiables, dont l'enchaînement comporte une logique à déceler. Le présupposé est qu'il est possible, par l'analyse statistique des données (et par l'apprentissage machine), de déterminer des logiques (sémantiques, syntaxiques, etc) qui structurent les textes analysés, et qui ne sont pas visibles ou quantifiables manuellement. La machine fonctionne alors de manière simple, dans un premier temps, en déterminant des unités d'analyse. Ces fragments sont les noyaux des données, les ensembles minimaux pour l'analyse statistique. En déterminant l'unité minimale de l'analyse statistique, l'utilisateur fait déjà un choix essentiel qui va conditionner les résultats obtenus.

La première étape de l'immense majorité des analyses de fouilles de données est la « tokénisation », qui consiste à subdiviser les données en les considérant comme les atomes constituant du corps-texte. Ces atomes sont appelés des identificateurs, ou plus communément des « tokens » : il s'agit d'unités, le plus souvent lexicales (un mot, une phrase, une expression, une ponctuation, etc.). De nombreux problèmes sont susceptibles d'apparaître lors de la « tokénisation ». Il incombe donc à l'utilisateur soit d'avoir recours un « tokenizer » spécialement calibré pour ses besoins, soit d'en créer un lui-même, soit de prendre en compte, dans la suite de l'analyse, le bruit qui peut être occasionné. Dès les premières étapes de l'analyse automatique des textes peuvent donc paraître des biais, qui peuvent avoir des répercussions exponentielles sur les résultats obtenus. Dans notre cas, un des problèmes type que pose la « tokénisation » est celui des syntagmes : une expression comme « Les Belles Lettres » doit-elle être comprise comme un tout, ou comme un ensemble de « tokens » ? Nous avons choisi, pour cet exemple, de préserver les deux possibilités.

Il est dès lors possible de proposer les premières analyses statistiques. En effet, dès la constitution des unités, des « tokens », tout type d'analyse est possible, mais le français est une langue flexionnelle très irrégulière qui se prête mal à ce type de défrichement simple. À ce stade, par exemple, l'ordinateur ne peut déterminer que « eut » et « avoir » sont deux « tokens » différents à la valeur sémantique équivalente. Pour un très grand nombre d'expériences, il est plus recommandé de procéder à une lemmatisation, notamment pour les langues flexionnelles. Une alternative, généralement moins coûteuse computationnellement, mais aussi souvent moins fiable pour le français, est la racinisation : au lieu de déduire le lemme d'un « token », il est possible de le réduire à sa racine, par suppression flexionnelle. Pour le français, cette technique n'est guère efficace pour l'étude d'objets de masse, puisqu'elle ne permet pas d'associer, par exemple, deux verbes comme « étaient » et « sont ». Les lemmatiseurs fonctionnent majoritairement de deux manières : soit par définition de règles (il s'agit alors de lemmatiseurs de type « rule-based », fondés sur une grammaire à état fini, basée sur un tri de tableaux), soit par prédiction statistique (souvent déjà entraîné pour l'utilisateur). Le principal défaut de la technique « rule-based » est qu'elle réclame une grande quantité de données de base : pour que le résultat soit fiable, il faut que le lemmatiseur ait accès à un maximum de cas vérifiés. Autre point problématique, ce type de lemmatiseur est dépendant de la langue. Cependant, la massification des données vérifiées permet de résorber ce problème : il est possible d'avoir accès à des données sûres, à des dictionnaires riches, suffisamment complets pour obtenir des résultats fiables, et souvent plus exacts pour des études strictement linguistiques. C'est pourquoi nous avons opté pour cette technique[3]. La deuxième technique de prédiction statistique est la plus employée et la plus répandue, parce qu'elle s'adapte facilement à toutes les langues. Mais elle a pour contrepartie de ne pas être fiable au niveau de l'analyse fine, notamment pour le français, puisqu'elle se base sur des données de prédictions non vérifiées.

La dernière étape traditionnelle, préparatoire pour la fouille de données, très souvent employée, est l'étiquetage syntaxique. Grâce à l'étiquetage syntaxique léger (le plus fréquent,

[3] Le lemmatiseur (et étiqueteur) que nous avons programmé avec Frédéric Glorieux, et qui donne de bien meilleurs résultats pour le français que TreeTagger, est disponible en ligne, à cette adresse : https://github.com/oeuvres/Alix.

qui permet de déterminer la nature syntaxique des « tokens », et non leur fonction), il est possible, par exemple, très simplement, d'étudier statistiquement les patrons syntaxiques des corpus, ou de déterminer quelles sont les natures de mots les plus fréquemment associées à d'autres. La grande majorité des étiqueteurs les plus connus sont prévisionnels (TreeTagger (Schmid 1995), Stanford POStagger (Manning et al. 2014), etc.). Ils présentent les mêmes inconvénients que les lemmatiseurs : calibrés à la base pour l'anglais, et pour la fouille de données de masse, ils ne sont pas conçus pour l'analyse syntaxique spécifique du français (bien qu'ils puissent obéir à des modèles entraînés spécifiquement), et moins encore pour l'analyse de corpus littéraire. Là encore, pour l'enrichissement de notre corpus en XML, nous avons privilégié Alix, l'outil que nous avons développé spécifiquement pour le français.

Il s'agit là des éléments minimaux pour l'analyse automatique des corpus. Ces trois étapes sont très majoritairement les étapes centrales et communes de toutes les fouilles de données récentes. Il existe un autre type de prétraitement très répandu, nécessitant l'accès à une base de données fournie et vérifiée, celui de la reconnaissance des entités nommées. Si l'utilisateur a recours à une base de données importante, et que le niveau de complexité d'identification des textes est moyen, elle peut s'avérer être un élément très riche pour les analyses de fouille textuelle. La reconnaissance des entités nommées est aujourd'hui très répandue, essentiellement pour l'anglais, moins pour le français. Elle pose de nombreux problèmes lorsqu'il s'agit de textes littéraires, dont la complexité sémantique déroute bien souvent les logiciels. Dans notre cas par exemple, nombreux sont les noms qui sont à la fois des noms de personnages et des toponymes. Seule une implémentation manuelle permet d'obtenir des résultats entièrement satisfaisants.

3. Des méthodes statistiques au machine learning

La fouille de textes peut avoir deux objectifs principaux : elle vise à décrire un phénomène, à prouver son existence et à l'étayer à l'aide d'exemples concrets, ou encore elle tend à prédire un phénomène textuel tel qu'il devrait se produire selon la logique des données analysées.

Le premier aspect de la fouille de texte, à savoir l'analyse descriptive des corpus, est celui qui est majoritairement retenu aujourd'hui par les humanistes numériques. Il s'agit de discerner, dans un corpus défini, des phénomènes qui, dans bien des cas, n'auraient pu être identifiés manuellement, ou des hypothèses de lecture qui n'auraient pas pu être démontrés manuellement, et c'est ce type d'analyse que nous cherchons à mener. Les méthodes statistiques les plus simples peuvent d'abord être mises en œuvre : les analyses quantitatives sont les plus répandues, car elles sont les plus faciles à implémenter et les plus compréhensibles. Il est possible, par exemple, de mesurer quantitativement le vocabulaire d'un corpus selon certaines variables (temporelles, thématiques, par genre, etc.). Nombre d'études sont ainsi effectuées sur les changements de vocabulaire des auteurs sur la durée : Lancashire propose par exemple d'analyser quantitativement la richesse et l'appauvrissement du vocabulaire d'Agatha Christie, simplement en comptant le nombre de mots et en prenant en compte la fréquence de chacun. Autre exemple, Jockers propose une analyse simple de répartition d'un mot dans un corpus (« dispersion plot »), en montrant à quels endroits d'un corpus donné un mot défini (ou un ensemble de mots dans une fenêtre donnée) apparaît (« Achab » dans *Moby Dick*). Il est aussi possible, par simple comptage de n-grammes (les n-grammes sont des ensembles de *n* éléments contigus), de repérer les fréquences de patrons syntaxiques et de représenter graphiquement leur répartition dans un ensemble (Snow, Jurafsky & Ng 2005:4–5). Autant de méthodes d'analyse quantitative simples, peu coûteuses, et qui donnent des résultats immédiats indicatifs souvent convaincants, puisque comme le montre la figure 4, une simple quantification des fréquences des lemmes « classique » et « classicisme » permet de voir l'émergence de deux concepts clés dans le domaine de l'histoire littéraire aux XIXe siècle dans notre corpus spécialisé :

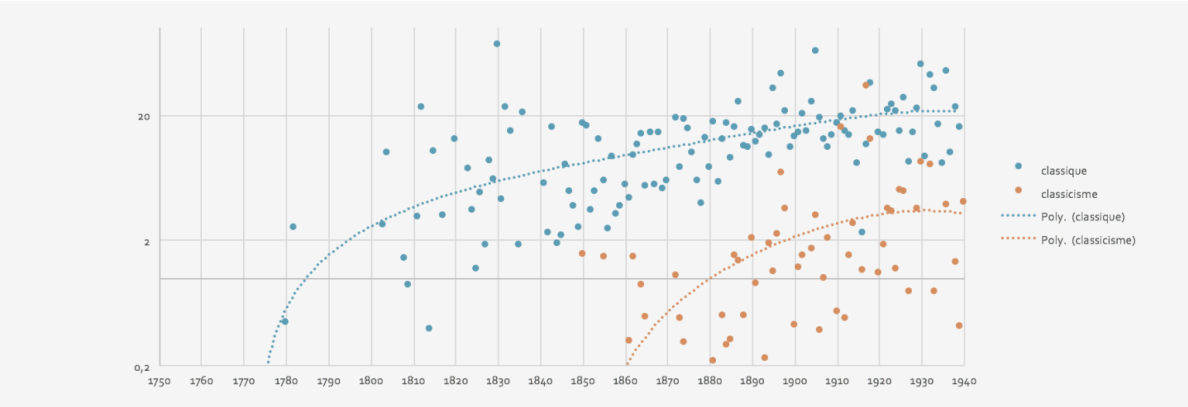

Figure 5 : La naissance d'un concept, « classique »

Mais ces méthodes posent essentiellement deux problèmes lors de l'analyse des corpus de masse : elles sont sujettes au bruit, qui peut, avec la masse, devenir significatif, et elles n'offrent pas de méthodes de tri fin des données. Si nous cherchons par exemple à savoir quels tokens apparaissent dans une fenêtre de cooccurrence sur un corpus réduit, il peut (à la limite) être aisé d'en analyser les données (si les termes cooccurrents recherchés ne sont pas trop fréquents). En revanche, si nous souhaitons, par exemple, rechercher des expressions sémantiquement équivalentes et dont les termes peuvent varier à moindre échelle, l'analyse statistique pure, sans autre filtre, est inopérante (comme c'est le cas dans TXM). Par exemple, si nous voulons considérer le poids de l'évocation de la production littéraire dans notre corpus, et que nous trouvons, d'une part, l'expression « ouvrage littéraire », et de l'autre « écrit littéraire », et si nous sentons que les termes « écrit » et « ouvrage » sont sémantiquement proches (sans dire pour autant qu'ils sont équivalents), l'analyse statistique pure de *n*-grammes est inopérante, puisque les deux termes ne sont pas graphiquement équivalents. Si équivalence il peut y avoir entre les deux termes, elle est d'ordre sémantique, et non formel. Or il est important pour notre analyse de pouvoir prendre en compte une potentielle équivalence sémantique, puisque les deux termes vont conditionner l'emploi, et donc le sens, du concept au cœur de notre étude, la littérature.

C'est dans ce cas de figure que la nécessité de l'apprentissage machine se fait sentir : la masse des données ne pose pas tant de problèmes du point de vue statistique, mais c'est bien plutôt la masse des résultats obtenus, sans tri, qui devient impossible à analyser sans autre entraînement machine. Même avec un corpus moyen, comme c'est notre cas, la masse des informations extraite serait trop longue à trier, et les résultats seraient analysés avec une trop grande marge d'erreur.

L'apprentissage machine est la capacité d'un système à améliorer ses performances en fonction de son environnement. Lorsque l'humain ne peut pas faire le tri des données qu'il obtient, lorsqu'il en a trop, il doit recourir à l'apprentissage machine, qui doit être capable d'effectuer le tri, voire mettre au jour de nouveaux résultats qui ne sont pas envisageables par la simple étude statistique. Généralement, l'apprentissage machine comporte au moins quatre éléments essentiels (outre les données à analyser) : un modèle (qui va servir à la machine à prédire ce qui devrait être son résultat, susceptible d'être modifié en fonction du résultat obtenu), un mode d'interaction avec l'environnement (une détermination, entre autres, des poids et valeurs à attribuer aux éléments pris en compte), une fonction de coût ou de pénalisation (lorsque les résultats attendus ne sont pas ceux obtenus, et lors de la modification dynamique du modèle), et un algorithme (ou des algorithmes la plupart du temps) d'adaptation du modèle. Autrement dit, il s'agit pour la machine, à partir d'un modèle donné et de conditions données, d'effectuer un choix parmi la masse de données. Il existe alors deux types d'apprentissage machine, celui dont le modèle a été calibré en fonction des besoins de l'utilisateur (avec un entraînement particulier, dit « supervised machine learning »), et celui dont le modèle est créé *ad hoc*, et qui n'a pas été spécifiquement entraîné. Les outils comme les vectoriseurs de mots sont traditionnellement basés sur un apprentissage machine supervisé, tandis que les techniques de « clustering » et de « topic modeling » font généralement (mais

pas toujours, pour le « topic modeling ») partie de la catégorie de l'apprentissage machine non-supervisé.

Là encore cependant, il serait illusoire de penser que la machine prend toutes les décisions de tri, y compris dans le cas d'un apprentissage non-supervisé. L'ensemble de la logique appliquée par l'ordinateur repose sur des présupposés algorithmiques précis (comme le fait de déterminer les mesures de similarité entre les nœuds, le nombre de « clusters », la validation des « clusters » etc.) Autant de phénomènes auxquels l'utilisateur doit être attentif, plus encore lorsqu'il s'agit de données de masse. L'avantage considérable du « clustering » est qu'il n'a pas besoin de modèle entraîné de base pour fonctionner : il n'est pas possible (du moins pas encore) d'avoir suffisamment de données catégorisées et vérifiées pour exclure toute possibilité d'erreur dans l'apprentissage machine. D'autre part, il est parfois simplement impossible de catégoriser les données. C'est ce qui rend, entre autres, l'analyse automatique des sentiments un exercice encore problématique et difficile à ce jour: dans le cas de l'analyse des phrases par exemple, deux phrases identiques peuvent être classées à la fois comme positives et négatives (dans le cas de l'ironie par exemple). C'est donc l'un des avantages importants de l'apprentissage machine non-supervisé que de ne pas reposer sur des données déjà étiquetées ou sur un modèle entraîné.

Le « clustering » consiste à demander à l'ordinateur de classer les données en des groupes déterminés. C'est grâce aux techniques de « clustering », par exemple, que les médias sociaux effectuent des rapprochements entre les sujets abordés (et prédisent d'autres sujets d'intérêt potentiels). Le « clustering » requiert simplement d'établir des unités de bases, sémantiques, stylistiques, voire relationnelles (Ardanuy & Sporleder 2014), et de demander à l'ordinateur quels sont les plus proches voisins, par des mesures de similarité. Le « clustering » de type « *k*-moyennes » est le plus répandu (étant donné un nombre *k* de clusters, il s'agit regrouper les clusters les plus proches de sorte à ce que la moyenne entre les moyennes entre les clusters soient aussi différentes les unes des autres que possible). Le « clustering » de type EM (« expectation maximisation »), de plus en plus utilisé, reprend les principes du « clustering » en « k-moyennes » (mais calcule la probabilité des candidats au « cluster » en fonction de leur distribution probable). Le « clustering » permet donc d'effectuer des associations que l'être humain ne peut pas voir directement, notamment lorsqu'il s'agit de regrouper des ensembles de données. Par exemple, nous avions effectué des expériences sur des profils de courbes par « clustering » chez Zola[4], permettant de déterminer sans supervision quels étaient les romans de Zola les plus proches entre eux, eu égard à la répartition de leur masse de mots. Il s'agit donc d'une technique d'apprentissage machine efficace, notamment dans le cas où le corpus est important, et où il est difficile de déceler immédiatement des liens logiques entre les éléments.

Mais l'inconvénient de ce type de technique, dans les usages littéraires qui peuvent en être faits, est que, s'il est possible de voir ce qui est rassemblé (de quoi sont composés les clusters), il est nettement moins évident de comprendre rapidement pourquoi. Les techniques de « topic modeling » (qu'il s'agisse de « topic modeling » supervisé ou non) fonctionnent à première vue, et seulement à première vue, sur les mêmes principes que le « clustering ». Le « topic modeling » consiste à déceler, dans un corpus donné, des « topics », c'est-à-dire des termes qui fonctionnent comme des nœuds déterminants sémantiquement, susceptibles d'être les thèmes principaux du corpus étudié. Le principe est simple : il s'agit de repérer les termes qui sont en cooccurrence (en pondérant la cooccurrence en fonction de la fréquence des termes), d'en relever les plus significatifs et de les grouper dans un nœud. Un « topic modeling » non supervisé donnerait donc une série de nœuds significatifs. Le « topic modeling » supervisé est de loin plus efficace : la supervision peut consister à fournir une simple liste de thèmes que l'ordinateur doit attribuer aux « clusters », mais plus souvent (ce qui est nettement plus efficace, mais qui réclame plus de données d'entraînement) il s'agit de donner des exemples types de documents dont on a déjà déterminé les sujets prédominants[5]. Globalement, les

[4] https://chapitres.hypotheses.org/458

[5] Un exemple de ce type d'analyse est le projet sur lequel travaille l'équipe du Literary Lab de Stanford : étant donné un micro-corpus entraîné (un corpus dont chaque paragraphe reçoit un degré de « suspense »),

techniques non-supervisées sont efficaces lorsque l'on ne connaît pas suffisamment la nature du corpus ou que l'on ne dispose pas de données d'entraînement machine. L'idéal reste donc de travailler sur une base déjà entraînée, comme sur des données distribuées dans des bases de données de type WordNet. Voici ce que donne (figure 5), l'analyse des « thématiques » de notre corpus, analyse d'emblée très variable notons-le car la technique est très sensible à des réglages compliqués qu'il n'est pas possible de détailler ici[6] :

- hommes France Dieu peuple liberté vie monde **politique** religion droit société philosophie
- saint fol cum Hist **manuscrit** quam Siècle sed ab Bibl Troyes Guillaume
- sens mots nature forme idées l'idée raison conscience vie exemple **science** c'est-à-dire
- vie littérature roman critique littéraire œuvre Hugo œuvres **France** poète française Victor
- coeur littérature dieu samedis histoire dramatique nom **poésie** ciel scène Ronsard terre
- chapitre lettres **édition** mémoires histoire France édit français française tome ibid
- vie cœur poète Dieu monde yeux âme **poésie** beauté l'amour terre ciel
- Paris soir maison ami nom femmes monsieur **monde** yeux pauvre matin mort
- Paris **langue** histoire française octobre France général français langue voltaire chapitre sieur
- comédie ouvrage **théâtre** ouvrages première poëte tragédie paroît nom fous musique mort
- théâtre comédie acte actes drame rôle scène dramatique **représentation** musique théâtre élève
- théâtre Molière scène Voltaire **goût** raison vrai comédie Racine tragédie Corneille monde

Figure 6 : Sur le corpus, topic modeling (LDA)

Apparaissent assez clairement l'importance de la critique théâtrale dans un corpus critique où la critique dramatique reste dominante (ce que nous savions déjà), l'importance des questionnement politiques et nationaux. Se révèlent les thématiques traditionnelles de la poésie (cœur, poète, dieu) au XIXe siècle ou encore un certains nombre de traits de vocabulaire typiques des discours critiques : on y reconnaît une certaine manière de parler, des auteurs canoniques, sans que l'on puisse faire véritablement d'une telle visualisation une preuve, faute de la maîtrise de la chaîne démonstrative qui autoriserait à faire de l'univers textuel qui nous est ici dépeint une véritable preuve, puisque le résultat obtenu dépend d'un modèle mathématique complexe et a été obtenu par les milliards d'opérations mathématiques que l'algorithme de topic modeling pour mesurer de manière probabiliste les regroupements de mots les plus nets. On voit donc la puissance pédagogique d'un outil capable de rendre visible des intuitions et des sentiments que l'esprit humain valide intuitivement, sans qu'il soit capable par lui-même de dresser la liste des mots statistiquement les plus caractéristiques du discours critique comme le fait l'ordinateur.

Dans le cas de notre étude, dans le cadre de notre collaboration avec le *Literary Lab* de Stanford sur l'histoire du concept de littérature, le « topic modeling » pose en fait plusieurs problèmes. Nous ne disposons pas de données pour un corpus entraîné. Si des ressources entraînées sont disponibles en ligne, elles ne sont pas adaptables à notre corpus, qui reste très spécifique (celui de la critique littéraire). D'autre part, malgré les apparences, notre corpus n'entre pas dans la catégorie des « macro-corpus ». Comme la plupart des bibliothèques implémentées pour le « topic modeling » fonctionnent sur le principe de l'échantillonnage, les résultats, d'une expérience à l'autre, peuvent être différents, parce que le corpus est réduit. Enfin, dans notre cas, le problème essentiel, du point de vue scientifique, est qu'il faudrait pouvoir (pour pallier le manque de données entraînées) donner des thèmes *a priori*, des concepts qui seraient susceptibles d'être les thèmes principaux du corpus. Or c'est précisément ce que nous voulons éviter, puisque nous souhaitons voir comment le concept de littérature est influencé par les concepts avec lesquels il interagit. Si ces concepts sont définis *a priori*, les résultats sont biaisés.

---

il s'agit de déterminer si un texte donné dans le corpus non-entraîné est susceptible de comporter du « suspense ».

[6] Voir la bibliographie de synthèse : http://mimno.infosci.cornell.edu/topics.html et par exemple (Posner 2012)

Le principe de distribution sémantique semble aller de soi : un mot est moins défini par l'ensemble des caractères qui le composent que par les mots qui l'entourent. Par exemple, si nous calculons la similarité de deux mots selon une distance de Levenshtein, c'est-à-dire la distance minimale de déplacements de caractères nécessaires pour arriver à un même résultat entre deux mots, « Athènes » et « Athéna » seront très proches, mais « Minerve » sera plus distant d'« Athéna » que ne l'est « Athènes », parce que les lettres communes dans un ordre proche sont peu nombreuses.

L'idée principale des principes de distribution sémantique est que le sens d'un mot vient de son usage plus que de sa forme[7]. En d'autres termes, le sens d'un mot est déterminé par ses cooccurrents. Ainsi, théoriquement, nous serions en mesure de dire que des mots qui ont des cooccurrents similaires tendent à avoir le même sens, ou tout du moins à être sémantiquement proches. Il est alors possible de créer, pour chaque mot, un vecteur de cooccurrences dans une fenêtre déterminée. Il est ensuite possible, une fois les vecteurs obtenus, de les représenter dans un espace vectoriel.

Dans notre cas d'étude, ce principe est déterminant. Nous souhaitons voir dans quelle mesure le concept de « littérature » évolue diachroniquement : il est donc nécessaire de voir comment non seulement le terme de « littérature », mais aussi ses équivalents distributionnels, évoluent au cours du temps. Le concept de « littérature » est alors compris comme un ensemble significatif de termes proches dans l'espace vectoriel, c'est-à-dire, sommairement, de mots qui sont déterminés par les mêmes cooccurrents.

Il faut désormais calculer la similarité des vecteurs. Une des mesures de similarités les plus connues est celle nommée Cosinus. Imaginons que le vecteur a soit constitué des indices de ses cooccurrents (a1, a2, a3... a*n*), et qu'il en soit de même pour le vecteur b. Il s'agit de calculer la similarité Cosinus entre le vecteur a et le vecteur b, en fonction de l'angle qu'ils forment dans l'espace vectoriel (nous mesurons l'angle et non l'ampleur du vecteur : si un vecteur pointait dans l'espace vectoriel très loin du dernier point d'un autre, leur angle pourrait tout de même être faible, et les mots qu'ils représentent pourraient donc tout de même être voisins). Nous obtiendrions donc, très schématiquement, ce type de représentation dans l'espace vectoriel (ici à une dimension) :

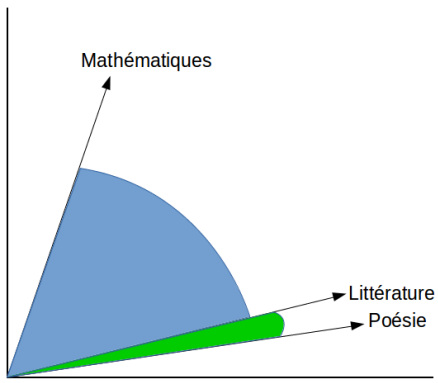

Figure 7 : Exemple d'espace vectoriel

Il s'agit ici d'une représentation très schématique, car il faut imaginer un espace vectoriel de *n* dimensions. Dans cet exemple, la similarité Cosinus mesure l'angle entre les vecteurs « Mathématiques », « Littérature » et « Poésie », et plus l'angle est réduit, plus la similarité des vecteurs est importante.

Prenons l'exemple des « siminymes » (c'est-à-dire les mots dont le vecteur distributionnel est similaire) du mot *romantisme* tels qu'obtenus par Alix :

Tableau 1

[7] Un des articles les plus connus sur le sujet, outre les travaux de Leonard Bloomfield, John Rupert Firth, ou Margaret Masterman, est celui de Zellig Harris (Harris 1954).

Mot recherché	romantisme
Cooccurrents	français:137, histoire:102, faire:101, classique:87, littérature:78, classicisme:77, siècle:69, romantique:67, dire:67, premier:64, apologétique:58, influence:53, allemand:52, voir:52, réalisme:50, pouvoir:50, chapitre:45, révolution:44, donner:43, littéraire:43, temps:42, France:41, poésie:41, théâtre:39, art:37, poète:35, devoir:35, Hugo:35, naturalisme:34
Siminymes	protestantisme, catholicisme, christianisme, paganisme, socialisme, naturalisme, symbolisme, journalisme, clergé, classicisme, réalisme, rationalisme, pape, Cid, Saint-Esprit, jansénisme, christ, Parnasse, despotisme, matérialisme, monde, sacerdoce, ciel, bouddhisme, Parthénon, divorce, midi, sénat, positivisme

Notons d'emblée que les cooccurrents obtenus de manière statistiques sans nécessiter de modèles complexes offrent d'emblée des résultats intéressants, même si relativement attendus qui permettent de vérifier par exemple que c'est à l'Allemagne que le concept de romantisme est associé par la critique, qu'Hugo constitue l'auteur le plus cité dans les analyses, que c'est bien au classicisme qu'est opposé le romantisme, certaines suppositions plus fines restant de l'ordre de l'intuition indirecte (l'importance du mot « français » relevant d'une construction d'un discours sur la littérature nationale dans la critique littéraire, largement académique). Pour ce qui est des vecteurs, l'impression est celle d'équivalences incompréhensibles, hormis celle avec des mots techniques équivalents (symbolisme, naturalisme). Mais si l'on considère les résultats de l'analyse vectorielles comme des équivalences de manière d'utiliser les mots, c'est la démonstration de l'emploi du romantisme comme une doctrine et même comme une religion, selon un constat qui aurait plu à Paul Bénichou, l'auteur du *Sacre de l'écrivain* : objet de discours historicisants dans le contexte de la construction d'une histoire littéraire nationale, le romantisme est perçu comme un culte et une foi autant que comme une simple école, pourrait-on conclure.

La performance de notre implémentation des vecteurs de mots peut encore être améliorée, notamment avec les nouvelles possibilités qu'offrent les réseaux neuronaux. C'est là une étape future essentielle du développement d'Alix. L'apport des réseaux neuronaux dans l'approche de la sémantique distributionnelle telle que nous la pratiquons est double. Tout d'abord l'intégration d'un réseau de neurones est moins dépendant des poids des fréquences, et moins sujet au bruit que les vecteurs linéaires. D'autre part, et c'est le grand apport des outils déjà implémentés comme Word2Vec (Mikolov et al. 2014) ou GloVe (Pennington, Socher & Manning 2014), les réseaux neuronaux se corrigent dynamiquement et s'entraînent. Sommairement, voici comment fonctionne un raison de neurones, lorsqu'il s'agit de traiter des ensembles textuels. Dans un réseau neuronal, chaque mot est assimilé à un vecteur de dimension *n*. Chaque mot est étudié de façon aléatoire et placé dans une matrice de vecteurs.

Le bénéfice essentiel des mots entraînés par apprentissage machine est leur capacité à se réformer (c'est à dire leur capacité à corriger leur score actuel à partir des résultats d'entraînement et à générer la probabilité d'une valeur attendue). Le modèle « skipgram » est une technique efficace, que GloVe a repris à Word2Vec. Le principe, sommairement, est le suivant : pour un couple de mots-vecteurs (c'est-à-dire de mots entourés de leur contexte-fenêtre), par exemple « espace, étoile », on calcule la probabilité du terme « espace » étant donné « étoile », en maximalisant le paramètre qui contrôle la probabilité. À cela Word2Vec ajoute le « negative sampling », à savoir qu'il calcule non seulement la probabilité qu'un couple existe, mais aussi la probabilité qu'il n'existe pas. GloVe, comme Alix, permet de mieux prendre en compte les répétitions et le poids des répétitions de cooccurrence que Word2Vec, ce qui est un point d'autant plus essentiel que notre corpus est relativement petit par rapport à ceux qui sont généralement analysés par Word2Vec (qui est davantage calibré pour analyser plusieurs centaines de gigas de données). Nous entendons donc, dans un futur proche, intégrer les vecteurs déjà produits par Alix à un système de réseau neuronal proche de celui de GloVe.

Ce type de technique a pour vertu d'offrir, au final, des résultats comportant beaucoup moins de bruit (d'erreurs) que ceux que nous obtenons avec de simples vecteurs linéaires. Il est donc possible, grâce à ce système, d'obtenir, sans données préalables nécessaires (sans dictionnaire, sans règles syntaxiques, et même sans corpus massif) une représentation sémantique de chaque mot de façon étonnamment juste (et plus encore si le paramétrage fin, comme la prise en compte de mots-outils, des lemmes, du taux de tokens en fenêtre glissante etc. est correctement calibré). Mais cela ne peut être qu'au prix d'une perte importante : l'intégration des réseaux neuronaux signifie que les résultats reposent sur la puissance de calcul de la machine, une puissance telle qu'un humain ne pourrait en une vie effectuer avec la même efficacité la même tâche, tâche que la machine effectue, elle, en quelques secondes. Il faut renoncer, même si les calculs et les résultats obtenus sont théoriquement démontrables, à vouloir discerner toutes les étapes du tri des données. C'est pour cette raison que le paramétrage de départ, et la bonne compréhension mathématique du système est un enjeu majeur dans l'implémentation des réseaux neuronaux.

Conclusion : des preuves numériques

Nous disposons ainsi d'une vraie gamme d'instruments heuristiques allant des mesures fermées (fréquentielles) aux analyses machine non dirigées (*topic modeling*) en passant par des mesure semi-ouvertes (collocation), typologie qui conduit à distinguer les quantifications simples, les recherches lemmatisées, les mesures de structures syntaxiques ou sémantiques (association mot et étiquette) et les analyses fondées sur les vecteurs de mots. Sans débattre de la pertinence relative de ces instruments, certains relevant de formes preuve directs ou indirects (qui varie selon que l'on cherche de phénomènes sémantiques ayant des échos textuels nets - présence d'un auteur, champs de référence- ou la détection par des analyses stochastiques de changements de paradigme qui seraient individuellement noyés dans le bruit des analyses fréquentielles), l'essentiel est de voir à quel point ils modifient notre approche traditionnelle de l'administration de la preuve en sciences humaines : avec la possibilité d'« opérationnaliser » (Moretti 2013) c'est-à-dire de vérifier des hypothèses théoriques ou historiques, les propositions des sciences humaines deviennent falsifiables - plus simplement dit, vérifiables. Les masses de données dans lesquelles est transcrite l'histoire culturelle permet de vérifier des hypothèses avancées par le savoir de l'érudition, mais autrement difficiles à établir car fondée sur une connaissance globale, une mémoire des œuvres, une synthèse intuitive, ardues à objectiver et donc à réfuter éventuellement. Derrière cette possibilité d'aligner l'épistémologie des sciences humaines sur celle des autre science (ambition récurrente on le notera à chaque changement de paradigme scientifique et que l'on pouvait sans doute aussi trouver dans la pensée positiviste de l'histoire littéraire ou dans le tournant linguistique et l'horizon constitué par la linguistique formelle) se jouent des questions autant institutionnelles que scientifiques (la visibilité et le sérieux des SHS). Le paradoxe est sans doute que les outils les plus élaborés et les plus mathématiques comme ceux de la sémantique distributionnelle des vecteurs et des topic models sont d'une complexité seulement accessibles par un très petit nombre de chercheurs et impliquent un nombre d'opération de calculs intermédiaire quantifiable en milliards. Assurément, pour contrebalancer l'opacité des boîtes noires numériques, ce domaine émergent qu'est l'histoire quantitative des idées ne peut être exploré que par la conjonction de l'esprit de géométrie et de l'esprit de finesse, la capacité à capturer et à représenter des faits mesurables massifs ne pouvant s'appuyant que sur une intuition cynégétique de terrain et une connaissance des corpus permettant la modélisation des évolutions supposées et la détection de phénomènes historiques inattendus.

Figure 8 : Le pouvoir heuristique d'un simple outil de visualisation en nuages de mots (Glorieux 2017)

Cette mathématisation de la preuve a d'autres effets par delà la tentation d'un nouveau positivisme : elle nous habitue à des formes de représentations inédites car, la lecture à distance (*distant reading*) de corpus constitués en *big data* par des cartes et graphes offre « une forme spécifique de savoir » (Moretti 2005:1) où les régularités sautent aux yeux de manière extrêmement pédagogiques et où les phénomènes minoritaires peuvent être détectés du regard, de manière bien différentes de la manière dont un érudit à l'ancienne remarquait et faisait ressortir une tendance littéraire méconnue ou un auteur négligé. A l'horizon, se dessinent peut-être encore d'autres évolutions épistémologiques profondes des humanités : on peut se pressentir qu'un usage empirique des humanités numériques pourrait se substituer à son usage théorique comme outil de vérification d'hypothèses abstraites. Un des aspects les plus dérangeants soulevé par les derniers développements de l'intelligence artificielle ces dernières années, dans son alliance avec les quantités considérables de données du big data, serait en effet de se passer de théorie au profit d'un apprentissage machine capable empirique de déduire par induction des lois générales ou en tout cas des régularités, perspectives empiriste qui rendrait obsolètes les propositions théoriques et qui ouvrirait même, puisque c'est la direction pour laquelle les apprentissages machines sont faits, des prédictions littéraires, faisant de l'histoire culturelle une science nomothétiques[8]. Après les méthodes statistiques avancées et l'émergence d'un savoir graphique qui prolongeaient différemment le travail historique en l'objectivant possiblement, une telle perspective rendrait transformerait encore plus radicalement le savoir culturelle, qui se réduirait à une manière habile de lancer une machine sur une piste et d'interpréter les résultats. Or si, de fait, les méthodes de vecteurs de mots arrivent déjà à se dispenser de toute représentation interne de la langue pour la comprendre ou en tout cas pour proposer des équivalences grammaticales ou sémantiques, l'ensemble des difficultés que nous avons rappelé, de la complexité de constitution d'un corpus jusqu'à la finesse nécessaire pour interpréter un résultat quelconque, font que ce type d'approche reste profondément dépendant de choix interprétatifs et de cadres théoriques. De la mesure au savoir, le chemin reste immense et nous restons seuls à pouvoir l'arpenter.

Références

Archer, Jodie & Matthew L. Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martin's Press.

Ardanuy, Mariona Coll & Caroline Sporleder. 2014. Structure-based Clustering of Novels. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)@ EACL*, 31–39.

[8] C'est déjà le cas avec le dernier essai de Matthew Jockers, qui prétend être capable de prédire la capacité d'un livre à devenir un best-sellers (Archer & Jockers 2016).

Burnard, Lou. 2010. TEI P5: Guidelines for Electronic Text Encoding and Interchange, 1.6. 0. *http://www. tei-c. org/release/doc/tei-p5-doc/en/Guidelines. pdf*.

Glorieux, Frédéric. 2017. Nuage de mots. http://obvil.lip6.fr/alix/wordcloud.jsp?&bibcode=proust_recherche&frantext=on.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2–3). 146–162.

Heiden, Serge, Jean-Philippe Magué & Bénédicte Pincemin. 2010. TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement. *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, vol. 2, 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard & David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *ACL (System Demonstrations)*, 55–60.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2014. *word2vec*. accessed 2014-04--15. https://code. google. com/p/word2vec.

Moretti, Franco. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.

Moretti, Franco. 2013. *Operationalizing': or, the Function of Measurement in Modern Literary Theory. Literary Lab Pamphlet, 6*. Stanford, CA: Stanford Literary Lab.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. *EMNLP*, vol. 14, 1532–1543.

Posner, Miriam. 2012. Very basic strategies for interpreting results from the Topic Modeling Tool. *Miriam Posner's Blog*.

Schmid, Helmut. 1995. Treetagger, a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 43. 28.

Sinclair, Stéfan & Geoffrey Rockwell. 2014. *Voyant tools*.

Snow, Rion, Daniel Jurafsky & Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 1297–1304.

Underwood, Ted. 2012. *The Stone and the Shell*.

Wieviorka, Michel. 2013. *L'impératif numérique ou La nouvelle ère des sciences humaines et sociales?* CNRS éd.

Zhang, Sarah. 2015. The Pitfalls of Using Google NGRAM to Study Language. *Wired http://www. google. com/s/www. wired. com/2015/10/pitfalls-of-studying-language-with-googlengram/amp*.

Marianne Reboul, Alexandre Gefen