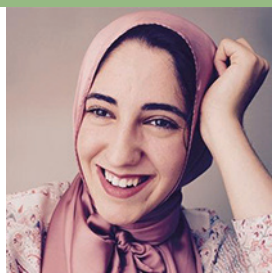
# Quantifier la littérature, qualifier la quantification :
## la culturomique et le N gramme de Google



Sanna ALAS
Étudiante diplômée – Université de Californie – Los Angeles

## Introduction

Des rouleaux de papyrus poussiéreux des tombes de l'ancienne Égypte, jusqu'aux romans à quatre sous des rayonnages de librairies d'aujourd'hui, l'écriture a été une partie intégrante de l'expression culturelle humaine. La presse à imprimer à caractères mobiles de Johannes Gutenberg a largement accéléré le processus, et depuis, les livres sont publiés et diffusés à des rythmes étonnants. En 2010, on estimait à 129 864 880 le nombre de livres ayant été publiés, ce qui serait équivalent à 4 milliards de pages et 3000 milliards de mots (Taycher 2010).

Dénombrer ces livres ne représente pourtant qu'une partie de la mission plus ample que s'est fixée Google, laquelle consiste à digitaliser tous les livres du monde, une entreprise ambitieuse qui a déjà produit une bibliothèque digitale toujours en croissance estimée aujourd'hui à 30 millions de livres (Darnton 2013) dans 480 langues différentes (Jackson, 2010). Selon les mots d'Eric Schmidt, qui était alors le président-directeur général de Google, ce projet vise à favoriser un « égalitarisme de la dispersion de l'information » grâce auquel chaque individu pourra accéder à des textes qui auparavant pouvaient rester inaccessibles et les compulser (Schmidt., 2005). Google n'est pas seul dans cette entreprise ; son effort pour digitaliser a été soutenu par d'innombrables bibliothèques, des groupes de publication et des institutions académiques (y compris dans ma propre Université de Californie) partout dans le monde. Des millions de textes qui avaient été oubliés sont ajoutés à cette nouvelle bibliothèque digitale.

## Lire à des distances différentes

Pour ceux qui étudient les choses de l'écrit, qu'ils soient étudiants en littérature, économistes, historiens ou linguistes, ce corpus a le potentiel de transformer leur façon de conduire leur recherche, ainsi que de soulever de nouvelles questions et de nouvelles méthodes d'analyse. Traditionnellement, l'analyse des textes a été dominée par une approche qualitative, procédant à une lecture rapprochée dans lequel un petit nombre de textes clefs sont analysés en grand détail. L'approche utilisée y est généralement inductive et exploratoire ; l'intention de l'auteur, le contexte historique, et les connotations de mots y jouent un grand rôle. Cette approche limite nécessairement la quantité de textes qui peuvent être analysés, et il en résulte un corpus réduit d'ouvrages canoniques sur lesquels la recherche se focalise. Cette approche qualitative est souvent considérée comme diamétralement opposée à une approche quantitative, dont le principe réside dans la déduction, le test d'hypothèses et l'expérimentation. Cette dernière méthode est généralement présentée comme scientifique, empirique et objective à cause de son usage de l'analyse numérique (Leeds-Hurwitz 1995).

Pourtant, avec la montée des données massives, beaucoup de gens testent des méthodes d'analyse quantitative dans des domaines qui ont été traditionnellement dominés par les méthodes qualitatives. Les chercheurs en littérature peuvent maintenant fouiller dans les milliers de livres qui pouvaient avoir été oubliés par ailleurs et poser des questions nouvelles au sujet de la culture ou du milieu littéraire d'une période. Ceci a engendré des tensions entre les tenants de la recherche qualitative et de la recherche quantitative. C'est pourquoi, étant donné les collections massives d'information littéraire qui émergent, on peut se demander quel est aujourd'hui le rôle de chacune de ces méthodes dans l'analyse des données ?

Selon la perspective de certains, comme Franco Moretti, les analyses quantitatives sont supérieures et devraient supplanter les analyses qualitatives. Le travail de ce dernier dans son laboratoire littéraire de Stanford en fait le champion de la lecture distante, dans laquelle les textes d'une période entière sont agrégés et analysés comme un tout, transcendant le point de vue jugé limité de la lecture rapprochée (Moretti, 2013). Parmi les études conduites par ce laboratoire on peut citer la différenciation quantitative des genres littéraires (Allison, 2011), la cartographie réticulaire des intrigues et des caractères (Moretti, 2011), ou encore les changements sémantiques dans le roman britannique (Heuser, 2012), toutes conduites au moyen de techniques de critique computationnelle. J'évoque ici Moretti non pour critiquer sa méthodologie, mais plutôt pour analyser sa rhétorique. Il présente un cas plutôt extrême dans sa façon de créer une fausse opposition entre recherches qualitatives et recherches quantitatives, alors qu'en fait ces méthodes s'appuient l'une sur l'autre. Ce cas révèle peut-être le sentiment inconscient que la recherche « scientifique » ou quantitative a plus de valeur, est plus précise et plus objective. Pour prendre un exemple moins extrême mais aussi significatif, le premier grand article publié à propos d'une recherche conduite en utilisant le corpus littéraire massif de Google Books a été publié non pas dans une revue consacrée aux humanités mais dans le journal *Science* (Michel, 2011). Ceci montre que la rhétorique scientifique a été, de façon générale, associée aux valeurs de la « précision » et de « l'objectivité ».

Dans le présent texte, j'espère déstabiliser le discours binaire qui entoure la recherche quantitative et la recherche qualitative en littérature et démontrer que ces deux méthodes ne sont pas opposées de façon inhérente. Jusque là j'ai cherché à fournir un arrière-plan de l'émergence d'une bibliothèque digitale, et des discussions critiques à son propos. Dans le reste de cette étude, je me focaliserai sur un corpus, le N-gramme de Google Books, et j'analyserai la manière dont il a été utilisé comme un supplément ou un remplacement de différentes méthodes de recherche à travers diverses disciplines.[1]

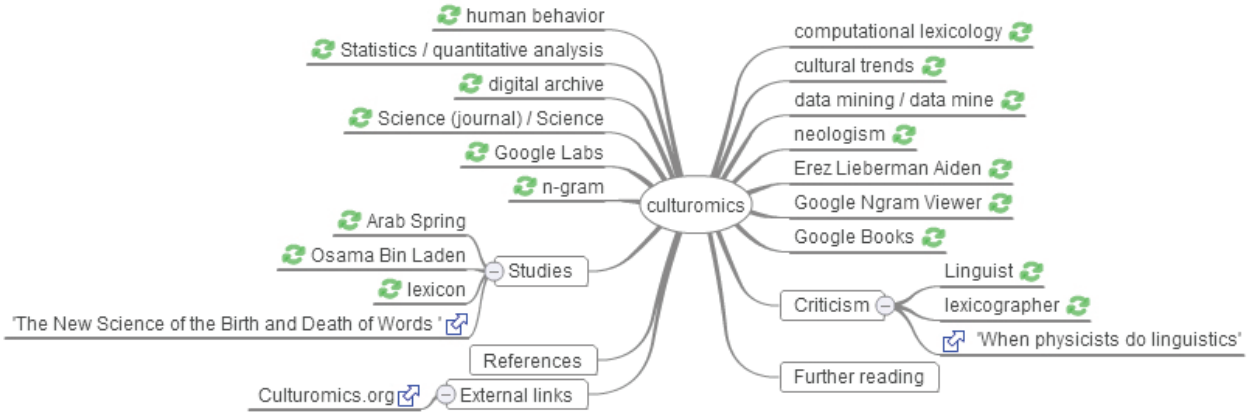



Figure 1. Carte Wikipedia représentant la « culturomique ». On y voit les différentes disciplines et champs de recherche qu'elle a influencés et ceux qui à leur tour l'ont influencée.

1. On le trouvera à cette adresse : https://books.google.com/ngrams

Le visualisateur N-gramme de Google Books fait partie d'un projet plus important appelé « culturomique » (figure 1). Fondée par Jean-Baptiste Michel,Yuan Kui Shen, Aviva Presser Aiden et d'autres de l'Observatoire culturel de Harvard, la culturomique cherche « à digitaliser et à analyser des données à propos de la culture sur de très grandes échelles : tous les livres, tous les journaux, tous les manuscrits, etc. » (FAQ Culturomics, 2015). Le mot a émergé comme un analogue de termes tels que « génomique », « protéomique » (à ne pas confondre avec la technique en microbiologie qui porte le même nom) et pointe vers la tendance actuelle pour le travail interdisciplinaire utilisant les nouvelles technologies et les méthodes d'interprétation quantitatives / computationnelles. Ce projet, qui tombe sous le parapluie des « humanités digitales », se présente lui-même comme un ajout aux formes traditionnelles de la recherche qualitative plutôt que comme un remplacement. Ce point de vue est incroyablement important parce qu'il subvertit la rhétorique de la supériorité pseudo scientifique et crée un aller-retour entre les méthodes quantitatives et qualitatives qui sont souvent considérées comme disparates. En analysant les types d'enquêtes conduites en utilisant le visualisateur N-gramme de Google Books, j'espère montrer comment l'émergence de la littérature comme une forme de données massives transforme les façons selon lesquelles les chercheurs analysent et comprennent des textes.

## Le N-gramme de Google Books : une brève discussion de la « datafication »

Avant de présenter quelques applications de ce corpus, il est essentiel d'expliquer que les pages sont d'abord scannées puis ensuite transformées en lettres et en mots « discrets » appartenant à un texte en ligne. Le premier processus, le scan, s'appelle la *digitalisation* et doit être distingué du second processus, la *datafication*. La digitalisation se réfère à la conversion d'une information analogique en un code binaire utilisé par les ordinateurs, alors que la « datafication » se réfère à la quantification des informations binaires pour l'analyse et la tabulation (Mayer-Schonberger, 2013, p. 78). Si des livres ont été digitalisés depuis des années, ce qui a changé récemment est que chaque caractère est maintenant «datafié» à travers le processus de la reconnaissance optique des caractères (ROC). Ce processus utilise une série d'algorithmes qui permet de convertir une photographie du texte en un texte digital dans lequel on peut faire une recherche. Ceci permet aux chercheurs de parcourir des pages d'information et de différencier par période de temps, genre, langage, année de publication, etc.

Le N-gramme de Google Books dépend entièrement de ce processus. Il identifie la fréquence relative de différents mots figurant dans 22 différents corpus divisés par langue (anglais, espagnol, chinois , russe, allemand, hébreu, français et italien), par genre (anglais vs fiction anglaise), ou par dialecte (anglais américain et anglais britannique) (Karch, 2015). Le N-gramme lui-même est un outil statistique utilisé en probabilités et en linguistique computationnelle pour trouver le nombre de fois, N, où un mot spécifié ou une phrase apparaît dans un corps de textes plus étendu. Le N-gramme peut non seulement compter le nombre d'apparitions d'une phrase mais il peut aussi fournir de l'information au sujet de la vraisemblance de son apparition dans le futur. À titre d'exemple, la figure 2 ci-dessous est un N-gramme (insensible à la casse) des mots « qualitative » et « quantitative » utilisant le corpus anglais pour la période 1800 à 2008.

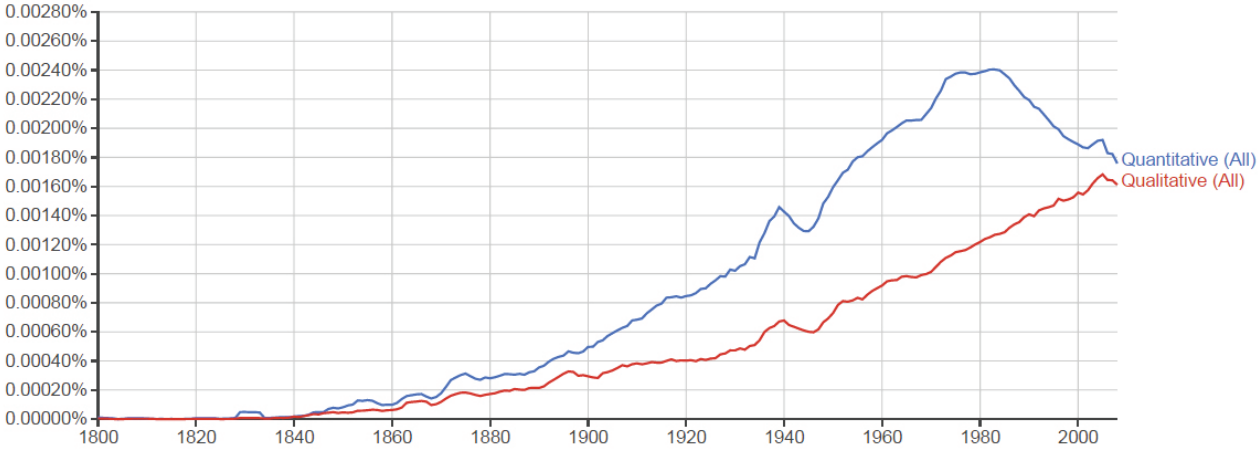



Figure 2. Ce N-gramme retrace les variations de l'usage des termes « quantitative » et « qualitative » dans le corpus anglais de Google Books. Les lois sur le copyright nous empêchent d'analyser les données brutes directement ; aussi, les données de fréquences fournissent une alternative pour chercher à travers la totalité de Google Books tout en protégeant les droits des éditeurs.

Bien que la collection de données soit gigantesque, il reste diverses limitations au visualisateur N-gramme de Google Books. L'une de ces limitations réside dans le processus ROC utilisé pour reconnaître les caractères, mis en défaut par le fait que les différents textes utilisent différentes polices et sont de qualité variable. De même, des erreurs dans les métadonnées d'un texte peuvent le dater faussement, et il en résulte des pics temporels factices qui ne représentent pas réellement la littérature d'une époque (FAQ Culturomics, 2015). De plus, seulement certaines langues sont représentées dans le corpus, et les textes de langues moins étudiées restent invisibles pour le N-gramme. Pour une recherche qui se propose de retracer les changements culturels à travers le temps, ces omissions risquent de réifier des interprétations hégémoniques de l'histoire, de la littérature et du langage, en continuant à ignorer des sources déjà marginalisées par les universitaires. Ceci étant dit, au fur et à mesure que la collection de livres de Google grandit et que les algorithmes qui dictent les processus de ce système sont développés, les ingénieurs du N-gramme espèrent le rendre de plus en plus précis.

## Littérature et données : une étude de cas en « culturomique »

L'Observatoire culturel de Harvard a été le premier groupe à faire un usage original du visualisateur N-gramme de Google Books. Il a produit plusieurs études fascinantes qui ont montré tout l'intérêt des puissants outils de la culturomique. Beaucoup de questions qu'ils ont traitées l'auraient été traditionnellement par les chercheurs des humanités – de l'évolution du langage à l'incidence de la renommée.

Plutôt que de présenter une vue d'ensemble de leurs études, je vais me focaliser sur une seule question, d'ailleurs exprimée dans le sous-titre d'un article : comment « détecter la censure et la suppression » (Michel, 2011, p. 181) ? Dans cette étude, la suppression et la censure sont quantifiées en comparant la fréquence d'apparition du nom de l'artiste juif Marc Chagall dans le corpus littéraire anglais et dans le corpus littéraire allemand pendant le régime nazi. Une recherche par N-gramme sur son nom a montré une décroissance drastique du nombre de fois où il a été mentionné entre 1936 et 1944 dans le corpus Allemand, une période qui est corrélée directement à la censure et à la suppression de l'Allemagne nazie.

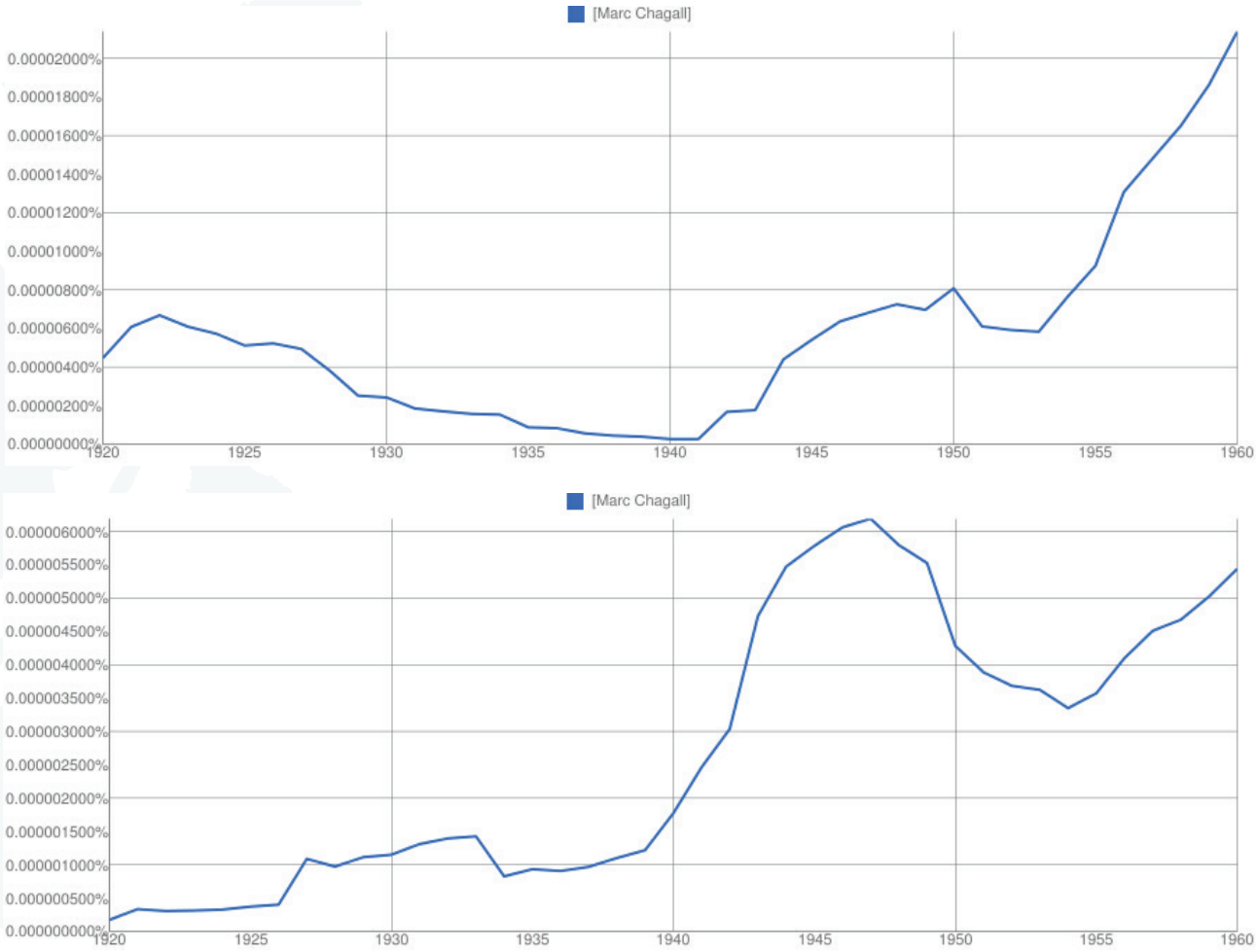



Figure 3. N-grammes de la requête « Marc Chagall » dans le corpus allemand (en haut) et dans le corpus anglais (en bas) entre les années 1920 et 1960

Cette étude est significative parce qu'elle montre que la censure subie par un individu est liée à un contexte socio-politique plus large. Cependant on doit prendre garde aux confusions, car le principe « corrélation n'est pas causalité » tient aussi lorsque les données sont littéraires.
Les chercheurs de Harvard ont conduit une étude similaire sur la censure durant cette même période où les travaux d'auteurs divers et d'artistes étaient bannis à cause de leur nature « indésirable ». En analysant la fréquence d'apparition des écrivains dans différentes catégories comme l'histoire, la littérature, la politique, la philosophie et les arts, ils ont pu déterminer une décroissance marquée de la production en philosophie, en politique, et en art (p. 181). Cette étude, à son tour, a permis de créer un « indice de suppression », grâce auquel les victimes de la censure peuvent être identifiées de novo.

Dans le cas de cette étude particulière, les notions de suppression et de censure qui ont été longtemps analysées comme des idées abstraites peuvent maintenant être quantifiées et mesurées. Imaginez les applications à long terme de telles métriques : il serait possible d'identifier les victimes de la censure ou de mesurer quantitativement le totalitarisme.
Cependant, de telles applications soulèvent aussi beaucoup de questions : pourquoi ces indices seraient-ils plus valides que les formes traditionnelles de preuves telles que les comptes-rendus d'archives, les comptes-rendus historiques ou les témoignages individuels ? L'utilisation de tels vastes ensembles de données fait implicitement l'hypothèse que les conclusions sont fondamentalement les moins faillibles et les plus concrètes, et par conséquent les plus convaincantes. Cependant, une telle supposition néglige les biais inhérents aux méthodes de recherche quantitative, dans lesquels des facteurs de confusion peuvent mener à des

conclusions fallacieuses. Au final, disons que les relations entre les différentes variables et les diverses corrélations identifiées ont un grand potentiel pour aider les chercheurs en sciences humaines et sociales à poser des questions différentes. Dans ce sens, il y a une rétroaction entre la recherche qualitative qui fournit un contexte et un cadre théorique et la recherche quantitative qui y applique ses méthodes.

## La littérature comme données : une étude de cas sur l'indice de misère littéraire

En utilisant le N-gramme de Google Books, des chercheurs ont montré que la littérature et les livres représentent en moyenne la situation économique de la décade précédente. Bien que ce point puisse sembler intuitif, il s'agit d'un important exemple de la manière selon laquelle des méthodes empiriques quantitatives peuvent être utilisées pour explorer des questions qui étaient traditionnellement dévolues aux méthodes qualitatives. Les chercheurs ont développé un « indice de misère littéraire» (IML) en utilisant l'analyse des données pour détecter les mots associés avec un état d'esprit particulier (Bentley 2014).

Trois méthodes ont été utilisées. La première était l'outil d'analyse textuel «Affect Net du Mot» (ANM), un outil digital qui génère des listes de mots synonymes et les associe avec des humeurs. Cette méthodologie a été modélisée à partir des six émotions de base (colère, dégoût, peur, joie, tristesse, surprise) identifiées par Strapparava *et al*. dans leur article de 2008 sur l'identification des émotions dans des textes.

La deuxième méthode utilisait une collection préexistante de mots présents dans l'outil «Recherche linguistique et comptage de mots » (RLCM) qui a été développé par le linguiste computationnel James W. Pennebaker et d'autres pour calculer le degré selon lequel un texte utilise des émotions positives ou négatives, des références à soi-même et d'autres dimensions du langage (Pennebaker, 2007).

La troisième et dernière méthode a été l'approche hédonométrique créée par Peter Dodds *et al*. dans leur recherche de 2011 évaluant le "contenu en bonheur" des comptes twitter. Cette étude a été conduite en utilisant un corpus de plus de 10 000 mots évalués du point de vue de l'émotion en utilisant l'outil "intelligence artificielle artificielle" du "Mechanical Turk" d'Amazon. Ces trois méthodes se situent donc dans une zone grise entre le qualitatif et le quantitatif, au sens où les humeurs que chaque méthode associait aux mots isolés l'ont été de façon subjective par des personnes, qu'il s'agisse de travailleurs anonymes sur le "Mechanical Turk" d'Amazon ou d'un ensemble de chercheurs. Cependant, tout comme dans le cas des conclusions tirées du visualisateur N-gramme de Google, l'accroissement de la taille de l'échantillon rend les résultats plus objectifs.

Pour la méthode ANM, l'indice de misère littéraire a été finalement calculé comme la différence entre les fréquences des mots caractérisés comme "de joie" et des mots caractérisés comme "de tristesse". Pour la méthode RLCM, l'indice a été calculé comme la différence entre les "mots de tristesse" et les "mots d'émotions positives". Enfin, pour la méthode hédonométrique, les mots associés avec le bonheur ont été corrélés à l'inverse de l'indice de misère économique, une hypothèse qui aurait besoin d'être validée. Les mots ainsi identifiés ont alors été entrés dans le visualisateur N-gramme de Google Books pour explorer le taux de croissance ou de décroissance de leur fréquence au cours du temps. Ce taux a ensuite été mis en graphique par rapport à un indice économique établi par ailleurs officiellement. Cet indice est une mesure du taux de chômage et du taux d'inflation. Il est communément utilisé comme un indicateur de la performance économique courante et future d'un pays.

Une corrélation positive a été trouvée, non seulement dans le corpus littéraire des États-Unis

mais aussi dans celui de l'Allemagne et de la Grande-Bretagne. Cette corrélation entre l'indice de misère littéraire et l'indice de misère économique est meilleure que pour n'importe quel autre indicateur d'humeur, reflétant la notion que l'économie a un effet profond sur l'état d'esprit et la production littéraire.

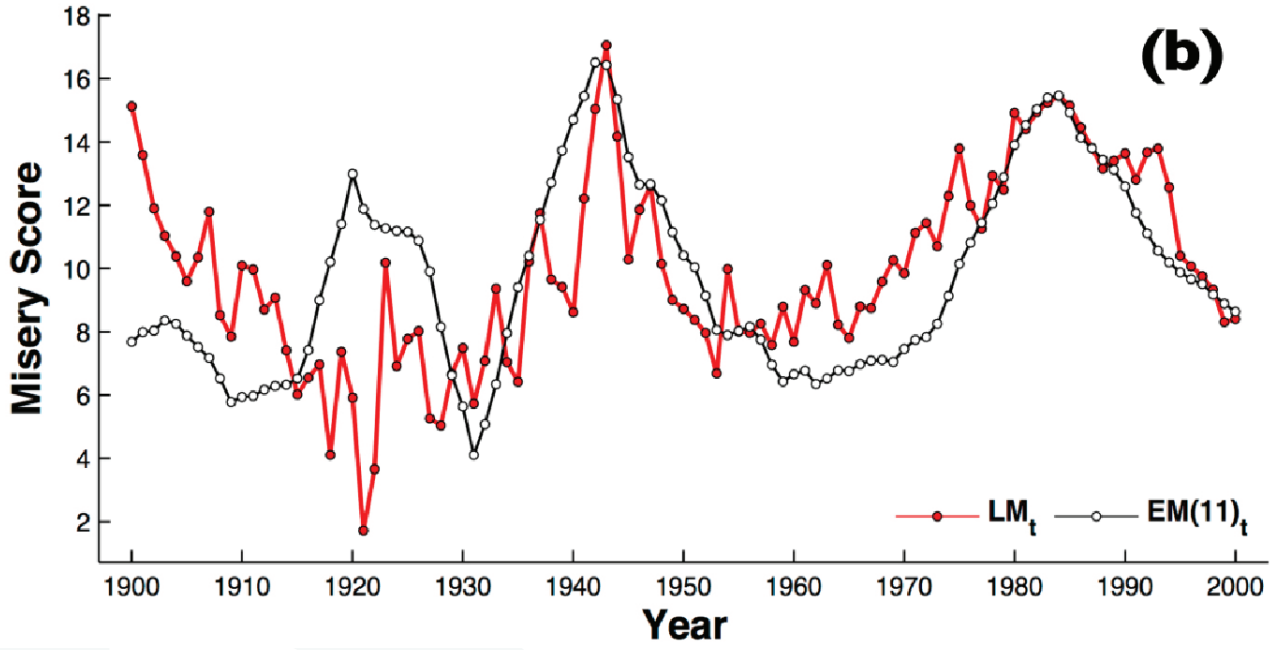



Figure 4. Comparaison entre, d'une part, l'indice de misère littéraire produit selon la méthode ANM (en rouge) pour tous les livres et, d'autre part, l'indice de misère économique aux États-Unis (en blanc). Les auteurs ont utilisé une moyenne mobile de l'indice économique sur 11 ans en arrière pour renforcer la corrélation entre les deux indices (Bentley, 2014).

Cet effet cependant peut être vu plus fortement dans le corpus onze ans après l'indice de misère économique. Cette période de 11 ans reflète la moyenne mobile qui a montré la plus grande corrélation avec l'indice de misère économique, plus qu'un simple retard ou que d'autres valeurs de la moyenne mobile. Ce calcul permet effectivement aux auteurs d'apercevoir des tendances plus importantes dans les données et de lisser les cycles de court terme qui semblent caractéristiques d'un corpus littéraire. Le graphique ci dessus présente des pics pendant des moments importants de l'histoire : la fin de la première guerre mondiale en 1918, les suites de la Grande Dépression vers 1938, et la crise de l'énergie en 1975. En fait, on a pu montrer que l'indice de misère littéraire avec la moyenne mobile sur 11 ans en arrière est plus robustement corrélé à l'indice de misère économique que les mesures économiques de l'inflation et du chômage qui sont les composantes de l'indice de misère économique lui-même.

C'est pourquoi cette étude a des conséquences fascinantes. Elle connecte deux formes disparates d'information, la production littéraire et la performance économique, d'une manière statistiquement mesurable. Elle montre tout le bénéfice qu'il peut y avoir à utiliser des sources de données non traditionnelles telles que la production littéraire et culturelle pour acquérir des informations "objectives" et des mesures quantitatives telles que des mesures de la performance économique. Elle change la conception de ce qui peut servir comme source de données dans la recherche quantitative ; les données massives peuvent être qualitatives si le corpus d'information est traité convenablement.

Ceci me conduit à retourner à notre point de discussion initial et à la rétroaction entre des méthodes de recherche quantitatives et qualitatives. L'article de Bentley est publié dans un

format qui est typique des articles scientifiques. Il présente une introduction qui fournit un survol du sujet en discussion et ensuite dessine une hypothèse de la manière suivante :

*Nous supposons que les auteurs de livres, qui sont à la fois des producteurs et des consommateurs de connaissances communes, sont informés par les conditions économiques du passé selon une certaine échelle de temps que l'on peut définir ; ils transmettent la connaissance commune non seulement factuellement mais aussi indirectement, à travers un contenu émotionnel. Par conséquent nous pouvons nous attendre à trouver une corrélation positive entre l'état d'esprit tel qu'il est exprimé dans les livres et les conditions économiques du passé récent. Vu que les livres prennent du temps à être écrits et publiés, nous nous attendons à ce que chaque année de publications fasse la moyenne des influences économiques du passé sur un nombre défini d'années (Bentley, 2014, p.1).*

Ce passage combine la méthodologie de la recherche quantitative, telle que l'indique l'usage des mots « nous supposons », « une échelle de temps que l'on peut définir », « corrélation positive » avec le savoir fondamental de la recherche qualitative dans ses descriptions de la position duale des auteurs comme producteurs et consommateurs de connaissances collective. Dans ce sens, l'information qualitative contextualise les méthodes quantitatives qui sont utilisées pour l'étudier.

Cette rétroaction est aussi observée dans la section Discussion de l'article, dans laquelle l'auteur cherche à aller au-delà de la corrélation et à rechercher la causalité. Il suggère plusieurs théories explicatives : l'une est celle de la construction de niches culturelles théoriques et l'assertion selon laquelle l'économie globale a été intégrée dans l'expérience émotionnelle de la société au 20e siècle (p.4) ; une seconde théorie repose sur « l'effet décade » selon lequel les souvenirs sont accumulés dans l'enfance et les individus se mettent à écrire à peu près 10 ans plus tard au sujet de leurs expériences (p.3). La première théorie est issue de la discussion des travaux en psychologie comportementale et évolutionniste de Steve Pinker sur les niches cognitives et sur les manières dont les humains ont évolué pour acquérir ces traits (Pinker, 2010). Cette théorie sert de pont entre la production culturelle et littéraire et le processus de l'évolution, déplaçant les discussions depuis un focus qualitatif sur le contenu (à partir des mots RLCM) vers un accent quantitatif sur la corrélation. De façon similaire, la seconde théorie formule des thèses sur le développement des enfants au moyen d'une étude longitudinale de 167 Californiens qui vivaient pendant la Grande Dépression (Elder, 1974) et cherche à prouver empiriquement à travers le processus de déduction ce que les méthodes quantitatives supposent en utilisant le processus d'induction : que les événements au cours de la vie influencent la mémoire dans le futur. Ce qui manque dans cette explication cependant, ce sont des preuves de l'argument implicite selon lequel tous les individus perçoivent les émotions de la même façon. Une autre faiblesse est la non-prise en compte du fait que, lorsqu'on mène des études rétroactives, il y a un grand potentiel de « biais de sagesse rétroactive » influençant la manière dont les données ont été perçues et présentées.

Cet usage du visualisateur N-gramme de Google Books montre que les méthodes qualitatives et quantitatives peuvent toutes les deux être utilisées simultanément dans un cadre de culturomique. On y trouve une rétroaction entre les méthodes quantitatives et qualitatives. Cependant, étant donné la nature scientifique de ce papier, les analyses quantitatives ont été privilégiées comme sources de preuves alors que les analyses qualitatives n'ont été utilisées que pour contextualiser l'argument. Ceci peut être un symptôme du rôle qu'y joue l'économie : comme la littérature est utilisée pour prédire (quoique rétroactivement) un indice économique, utiliser une preuve quantitative serait le choix logique. La discipline influence en définitive le type de questions que l'on pose et les méthodes que l'on choisit. Ici le processus de « datafication » est rendu clair puisque la littérature est découpée dans ses plus petits composants, dé-identifiée, et dé-contextualisée afin de tirer des conclusions empiriques au sujet de phénomènes sociaux, historiques et économiques. Dans cette étude, la littérature est devenue un autre fichier de

données qui, comme l'information statistique et démographique, est utilisé pour tirer des conclusions quantitatives.

## Conclusion : des spectres rhétoriques

Le corpus de Google Books, quoiqu'impressionnant par sa taille, n'est qu'une collection parmi d'innombrables autres. Chaque corpus a son propre ensemble de textes, sa propre méthode pour le décrire, et son propre ensemble d'outils pour analyser les données qu'il contient. De plus, ce corpus n'inclut pas le vaste éventail d'autres textes à travers lesquels la culture se manifeste : les dépêches d'information, les notes de musique, les noms d'enfants, etc. Cette pluralité démontre à quel point le cœur même de ces efforts pour quantifier la culture peut être subjectif. J'ai commencé cet article en discutant la tendance vers la digitalisation d'un corpus massif de production textuelle. Cela m'a conduit à une discussion sur la signification même de la lecture, et sur les fausses oppositions binaires qui existent autour de la recherche qualitative et quantitative en littérature. La notion de lecture rapprochée est ce que la plupart des chercheurs dans les humanités pratiquent, un processus qui demande un ensemble d'outils analytiques reposant sur le raisonnement inductif. Diamétralement opposée à cela est la notion de lecture distante dans laquelle les textes eux-mêmes sont réduits à leurs composants unitaires et analysés comme un collectif. La collectivisation des textes apparaît comme parallèle à la « datafication », dans laquelle les caractères d'un texte scanné sont transformés en une information significative qui peut alors être analysée. Ces processus ont produit différentes manières de s'attaquer au corps des données, l'un d'entre eux étant le visualisateur N-gramme de Google Books qui a été utilisé comme un outil à l'intérieur des humanités et à l'extérieur de celles-ci. Les études de cas qui ont été analysées reflètent les différentes méthodes de recherche et la manière dans laquelle les données sont comprises. Plutôt que de faire une opposition binaire entre les méthodes de recherche quantitative et qualitative, je soutiens que ces études quantitatives utilisent les données massives collectives pour établir une rétroaction entre des disciplines et des méthodes traditionnellement disparates.

Tout au long de la discussion cependant, il y a eu un spectre qui a semblé hanter la rhétorique de ces textes. La tendance à quantifier, à décrire les humanités en termes scientifiques, reflète une tension innée entre différentes méthodes de recherche et différentes disciplines. Il y a de nombreuses raisons à ce changement mais même dans ma propre rhétorique et dans mes propres analyses je me suis retrouvée à glisser vers des descriptions réductrices de ces méthodes qualitatives et quantitatives. Finalement on doit se demander si ce sont des paradigmes qui peuvent être réconciliés ou s'ils sont incommensurables. Un fait cependant est certain : la transformation de la littérature en « données massives » n'est pas un processus neutre mais un processus qui a la charge de faire apparaître de nouvelles questions, de nouvelles idées, de nouvelles controverses.

## Références


Allison, S. D., Heuser, R., Jockers, M. L., Moretti, F., & Witmore, M. (2011).Quantitative formalism: an experiment. Stanford Literary Lab.

Bentley, R. Alexander, et al. "Correction: Books Average Previous Decade of Economic Misery." PloS one 9.1 (2014).

Darnton, R. (2013). The National Digital Public Library Is Launched!.The New York Review of Books, April 25, 2013 issue.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PloS one, 6(12), e26752.

Elder GH (1974) Children of the Great Depression. Westview Press

FAQ - Culturomics. (n.d.). Retrieved March 21, 2015, from http://www.culturomics.org/Resources/faq

Jackson, J. (2010). Google: 129 million different books have been published.PC World, 7.

Karch, M. (n.d.). How to Use the Ngram Viewer - Google Books. Retrieved March 21, 2015, from http://google.about.com/od/n/a/Google-Books-Ngram-Viewer.htm

Heuser, R., & Le-Khac, L. (2012). A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method.

Leeds-Hurwitz, W. (Ed.). (1995). Social approaches to communication. Guilford Press.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

Michel, Jean-Baptiste, et al. "Quantitative analysis of culture using millions of digitized books." Science 331.6014 (2011): 176-182.

Moretti, F. (2011). Network theory, plot analysis. New Left Review.

Moretti, F. (2013). Distant reading. Verso Books.

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. Proceedings of the National Academy of Sciences, 107(Supplement 2), 8993-8999.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007 [LIWC manual]. Austin, TX: LIWC.net.

Schmidt, E. (2005). Books of revelation. Wall Street Journal, 18, A18.

Strapparava, C., & Mihalcea, R. (2008, March). Learning to identify emotions in text. In Proceedings of the 2008 ACM symposium on Applied computing (pp. 1556-1560). ACM.

Taycher, L. (2010). Books of the world, stand up and be counted! All 129,864,880 of you. Inside Google Books. Accessed December, 1, 2013.