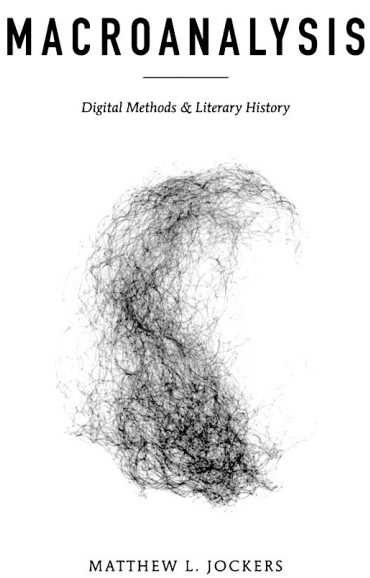
Franco Moretti : L'objet des humanités numériques, entre perspectives micro et macro

La visualisation de données pose problème: on y navigue entre le très petit et le très grand sans intermédiaire susceptible de permettre une lecture critique. Ce document est le compte-rendu de la conférence ***Micromégas: the very small, the very large and the object of digital humanities*** donnée par **Franco Moretti** le 15.12.2014 à l'Université de Lausanne, et a été mis en ligne le 16.12.2014.

ONLINE
GRANDJEAN Martin (2014), *Franco Moretti : L'objet des humanités numériques, entre perspectives micro et macro*, http://www.martingrandjean.ch/franco-moretti-very-small-very-large-digital-humanities/

Le problème de la visualisation de données

La couverture de Macroanalysis de Matthew Jockers illustre bien la problématique du "très petit" et du "très grand" : on y voit un réseau de personnages de roman. On ne peut toutefois pas y distinguer les noeuds, trop petits, et difficilement les arêtes tellement elles sont nombreuses. Une telle représentation est très riche en information. C'est d'ailleurs devenu une constante dans les humanités numériques que de présenter des résultats graphiques rendu très difficilement lisibles par le nombre très élevé d'éléments qu'ils contiennent. Ce constat est alimenté par de nombreux autres exemples de visualisations d'analyses littéraires à l'aide de nuages de points ou de diagrammes de Voronoï qui sont plus des moyens de poser des questions que de trouver des réponses.



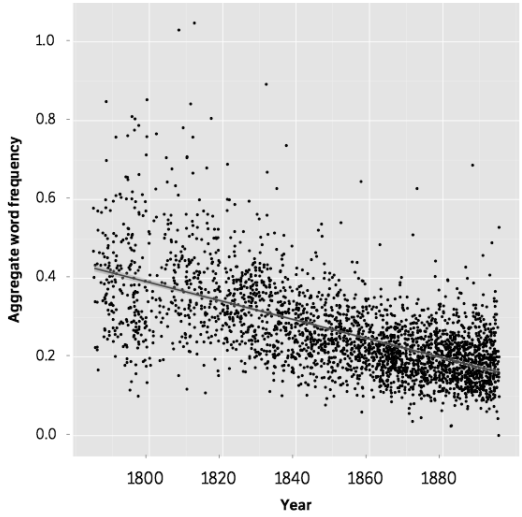
Jockers, M. (2013)
Macroanalysis

Ces images posent problème : elles nous font naviguer entre le très petit (l'unité du graphique, le point) et le très grand (l'image globale) sans médiateur pour faire le lien entre ces deux dimensions. Il nous faut un étage intermédiaire, un échelon qui permette l'acte de lecture. Les humanités numériques, en oblitérant ce que devait être la nouvelle analyse littéraire, ne favorisent pas la lecture.

Du “petit” au “grand”, ou l’inverse ?

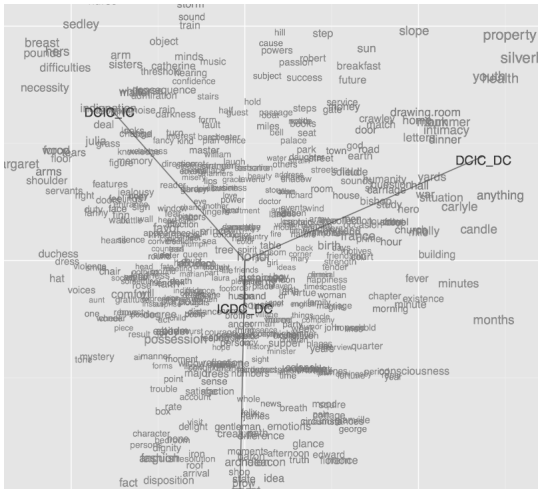
Il y a deux façons d’étudier des objets littéraires par le biais de telles abstractions visuelles :

1. Du petit au grand
Le grand, dans un nuage de points, c’est l’accumulation d’une multitude de petites unités, leur somme. En tant que tel, le grand n’a pas d’autre sens que la fréquence des petits. Il ne s’agit pas ici de critiquer cette approche qui propose de chercher la tendance globale derrière les données individuelles, mais de comprendre de quoi sont faites nos visualisations et nos analyses: celles-ci permettent de voir une tendance historique qui est contenue dans les points eux-mêmes.



Heuser and Le-Khac (2012) *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels : The Semantic Cohort Method*

2. Du grand au petit
C’est le cas lorsqu’on fait une analyse de fréquence de mots dans un texte. La symétrie de cette approche avec la précédente n’est qu’apparente. Un roman n’est pas réductible à la fréquence des prépositions qui le composent. Les petits détails montrent une partie de la structure. Ce n’est dès lors plus un graphique qui a pour but de montrer une explication causale, mais fonctionnelle. Les explications fonctionnelles sont reliées à l’échelle (ce qui n’est pas le cas dans le premier exemple). Le plus bas degré de l’échelle n’est pas toute l’histoire, mais son commencement seul.



Allison, Gemma, Heuser, Moretti, Tevel and Yamboliev (2013) *Style at the Scale of the Sentence*

Trouver le juste “milieu”

Quel objet peut-il connecter le petit et le grand ? Qu’y a-t-il entre les détails et le tout ? Essayer de le définir est un des chantiers dont les humanités numériques doivent se saisir. Aujourd’hui, notre cercle herméneutique est constitué d’aller-retour entre général et particulier, à l’image de ce que Spitzer,

s'inspirant de Dilthey décrit comme un "voyage entre détail et tout" à propos de son "cercle philologique"[1].

Franco Moretti distingue trois chemins possibles pour décrire ce "milieu":

1. Le "milieu" est un objet entièrement construit

Si le texte ne contient pas déjà, même partiellement, d'unité qui puisse servir d'intermédiaire entre le petit et le grand, le chercheur peut le construire. Dans le cadre de ses recherche sur le style familier dans la littérature américaine au XIXe siècle, Marissa Gemma[2] étudie les contractions de mots présentes dans son corpus. Le "milieu" entre les données issues de cette analyse des apostrophes et la globalité du texte doit être construit pour que l'analyse formelle puisse entraîner une analyse qualitative (en l'occurrence politique) des textes. Toutes les disciplines sont concernées par cette recherche d'un objet "construit" pour tirer une analyse substantielle à partir de la transformation de la connaissance formelle.

2. Le "milieu" est partiellement présent dans le texte et partiellement construit

C'est l'un des mots-clés les plus populaires dans le champs des humanités numériques : "*pattern*" (motif). Moretti propose l'exemple de l'analyse de phrases à deux propositions dans un corpus textuel. Dans le détail, on analyse ici quels mots apparaissent plus dans quels types de propositions, en fonction de leur place dans l'organisation de la phrase. On y trouve un *pattern* lorsqu'on se rend compte que des mots appartenant à un même groupe sémantique, par exemple les mots décrivant des émotions, se retrouvent tous dans la même "région" de la visualisation. Malgré le fait que pattern vienne de "patron" (en français), un modèle dont on fait les choses singulières, son sens actuel en anglais désigne un arrangement d'éléments. C'est donc un terme qui a, comme "révolution", subit un virage à 180°: de normatif il est devenu descriptif !

Mais qu'est-ce qui dit que l'on a trouvé un *pattern* ? Il s'agit d'une découverte subjective puisque découlant d'un travail de sélection empirique. Un pattern n'existe que lorsqu'un chercheur examine les données et construit ses hypothèses à partir de ses observations.

Un *pattern* est un élément qui se situe entre deux grandes forces : la distribution (large mais peu ordrée) et la structure (plus petite, mais plus ordrée). Il participe aux deux mondes en étant issu du monde désordonné de la distribution mais en évoquant déjà une certaine structure organisée. Cette tentative de trouver une logique dans le monde empirique a pour conséquence de rendre ces *patterns* nécessairement imparfaits et temporaires. Mais ce que l'on cherche bien souvent dans une vision utilitariste, c'est une clé d'explication. Dès lors, on ne peut pas s'arrêter aux patterns, parce que c'est

[1] Spitzer (1962) *Linguistics and Literary History* 19-20 et 25.

[2] Gemma, M. (2014) *Towards a Digital Study of the Colloquial Style in Nineteenth-Century American Literature.*

une promesse plus qu'un résultat: le processus de la connaissance n'est pas complet si on ne va pas en-deça. Aujourd'hui, la reconnaissance des *patterns* est une approche très populaire, mais l'étape décisive est de réussir à en tirer une structure.

3. Le "milieu" existe déjà dans le texte

Et si l'étape "milieu" déjà contenue dans le texte n'était pas tout simplement le paragraphe ? On ne parle pas en paragraphe, mais on écrit en paragraphes. C'est une unité très importante de l'articulation d'un texte, mais elle souffre d'une méconnaissance et d'une faible littérature à son sujet. A ce titre, les travaux d'Auerbach (en particulier *Die ernste Nachahmung des alltäglischen* cité par Moretti[3]) sont d'un grand intérêt pour montrer en quoi le paragraphe est un élément stylistique.

Dans le cadre d'une étude sur la taille des paragraphes (nombre de mots et de phrases) dans un corpus limité, Moretti montre que les courts paragraphes sont surreprésentés à cause des dialogues. Au-delà de cet aspect, il postule que le paragraphe est une unité de contenu. Il propose donc un *topic modelling* de ces paragraphes en se fixant pour hypothèse de trouver le thème de chaque paragraphe. Cette recherche est également l'occasion de questionner le rapport entre la longueur d'un paragraphe et l'expression de ce thème : Y a-t-il une longueur idéale pour présenter un sujet (en termes techniques, la longueur d'un paragraphe est-elle corrélée avec la présence de mots classés dans le même *topic* ?) ? La corrélation est la plus forte dans les paragraphes longs de 50 à 100 mots, il semblerait que ce soit la longueur la plus propice à la présentation d'un sujet unique dans le corpus étudié.

Etudiant le contenu de ces paragraphes, Moretti constate que les paragraphes de dialogues sont logiquement plutôt courts, alors que les paragraphes qui parlent des circonstances de la vie sont plus souvent longs. Cela entraîne une remise en question évidente: les paragraphes ne sont-ils monopolisés que par un seul sujet ? Les résultats obtenus par *topic modelling* permettent de tester cette hypothèse puisqu'on peut mesurer la proportion de mots de l'un ou l'autre *topic* dans chaque paragraphe: et cela fonctionne, même s'il est difficile d'expliquer pourquoi et comment… Finalement, les mots identifiés et regroupés par *topic modelling* parlent-ils vraiment d'un seul *topic* ?

Ce doute radical est à l'image de l'état de la recherche actuelle. Beaucoup d'études montrent des résultats intéressants, mais personne n'est encore capable de briser l'opacité de l'outil.

[3] Moretti, F. (2013) *The Bourgeois, between history and literature* 71-72.