La lecture distante : introduction et exemples d'application

Marie Puren

LA LECTURE DISTANTE

Introduction et exemples d'application

Marie Puren

20 novembre 2020

UVSQ | M1 ACPCI, ECMAH, EMAS, HCS, RCL

PLAN

De l'analyse qualitative à l'analyse quantitative des textes

- Analyse des textes dominée par une approche qualitative : lecture en détail d'un petit nombre de textes-clé
- Approche inductive (proposer, à partir de l'étude du cas particulier, une explication pour le cas général) et exploratoire
- Etude de l'intention de l'auteur,du contexte historique, et des connotations des mots
- Approche limitée : travail sur une petite quantité de textes, en général corpus réduit d'ouvrages dits "canoniques" sur lesquels la recherche se focalise

De l'analyse qualitative à l'analyse quantitative des textes

- Approche opposée à approche quantitative
- Approche quantitative repose sur la déduction, le test d'hypothèses et l'expérimentation
- Deuxième approche présentée comme scientifique, empirique et objective à cause de son usage de l'analyse informatique

UNE ÉTUDE QUI A FAIT DATE

- Seconde approche particulièrement mise en avant pas une étude publiée en 2011 dans la revue Science :
  - Jean-Baptiste Michel et al., ”Quantitative Analysis of Culture Using Millions of Digitized Books”, Science 14 Jan 2011 : Vol. 331, Issue 6014
- Etude menée par une équipe en partie issue de la firme Google
- Utilise un corpus de plusieurs millions de livres numérisés, Google Livres
- Revendique une histoire culturelle sans historiens, laissant les données ”brutes” parler d’elles-mêmes.

GOOGLE NGRAM VIEWER

- Utilisation de Google Ngram Viewer pour réaliser cette étude
- N-gramme : 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 corpus de textes
    - Google Ngram Viewer : maximum 5-grammes, donc 5 mots.
- Corpus Google Livres : divisés par langue, par genre (anglais vs fiction anglaise), ou par dialecte (anglais américain et anglais britannique)
- Visualiser la fréquence d'apparition de mots dans ce corpus

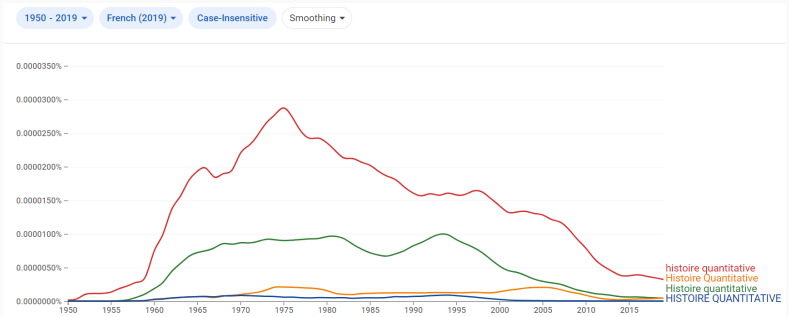


FIGURE : N-gramme de "histoire quantitative" dans le corpus français de 1950 à 2019

GOOGLE LIVRES

- Repose sur le corpus de Google Livres
- Pages des livres sont d'abord numérisées, puis transformées en lettres et en mots "discrets" (dénombrables : discontinus, séparés, distincts)
- Posiible de de parcourir des pages d'information et de les différencier par période, genre, langage, année de publication, etc.

LE PROJET CULTUROMICS

- Google Ngram Viewer fait fait partie d'un projet plus important appelé "Culturomics" (culturomique)
- But de ce projet fondé par les auteurs de l'article (Observatoire culturel de Harvard) : numériser et analyser des données à propos de la culture sur de très grandes échelles : tous les livres, tous les journaux, tous les manuscrits, etc.
- "Culturomique" analogue à "génomique" par exemple :
  - Génomique => ne pas se contenter de l'étude d'un seul gène, mais étudier l'ensemble du génome
  - Culturomique => ne pas se contenter de l'étude d'un seul texte, mais étudier l'ensemble des textes produits par l'humanité

UNE NOUVELLE TENDANCE POUR LIRE DES TEXTES ?

S’inscrit plus largement dans des travaux initiés dans les années 2000 par Franco Moretti sur la ”lecture distante”

LA LECTURE DISTANTE

UNE “RÉVOLUTION” POUR L’ÉTUDE DES TEXTES

- Franco Moretti, Graphs, Maps, Trees : Abstract Models for Literary History, Verso, 2005..
  Alors historien de la littérature au Stanford Literary Lab ; professeur à l’EPFL aujourd’hui.
- Abandonner la pratique traditionnelle, à savoir la ”close reading” ou ”lecture attentive” ou ”lecture proche”
- Travailler non pas sur des textes singuliers, mais sur de grandes bases de données de milliers de textes.

But = identifier des ”patterns”, des modèles ou motifs, au sein de corpus qui traversent les siècles et les frontières.

UNE “RÉVOLUTION” POUR L’ÉTUDE DES TEXTES

- Emprunte modèle de l’économie-monde, concept forgé par l’historien Fernand Braudel en 1949 puis développé en 1966 : territoire fonctionnant autour d’un centre économique autonome, dont l’influence s’étend jusqu’à sa périphérie (espace bien délimité : cela ne correspond pas à l’économie mondiale)
  - Méditerrannée au XVIème siècle
  - L’empire hispanique au XVIIème siècle
  - L’empire britannique au XIXème siècle
  - L’économie contrôlée par les Etats-Unis au XXème siècle
- Nouvelle ambition de la critique littéraire selon Moretti : viser la littérature-monde
- Toute oeuvre littéraire, même un chef d’oeuvre, doit être analysée en interaction avec :
  - les littératures étrangères de la même période
  - la production éditoriale (même les oeuvres oubliées) de son temps

UNE “RÉVOLUTION” POUR L’ÉTUDE DES TEXTES

- Impossible de lire par soi-même tout ce qui a été publié depuis l’invention de l’imprimerie, d’autant plus que ces textes sont dans des langues différentes
- Besoin de coopérer avec d’autres chercheurs(ses) et d’utiliser des résultats obtenus et des analyses réalisées par d’autres
- Dans les années 2000, quand Moretti commence à développer ce concept : pas mention de l’informatique à ce moment-là
- Recherche d’une méthode quantitative de terrain, favorisant la collaboration entre chercheurs et chercheuses
- Modèle de la science historique / archéologique : personne ne reprendrait totalement un dépouilement d’archives ferait à nouveau

Les arguments en faveur de la lecture distante (1)

Trois arguments principaux en faveur de la pratique de la lecture distante dans son livre de 2005, et dans Franco Moretti, Distant reading, Verso, 2013, un recueil de dix articles.

- Premier argument : les chercheurs et chercheuses ont déjà à disposition des corpus de textes numériques ou numérisés qu'ils peuvent utiliser.
- Exemples de corpus numérisés : HathiTrust, Gale, Google Livres, Europeana, Gallica, Internet Archive

Les arguments en faveur de la lecture distante (2)

- Deuxième argument : tendance à se concentrer sur une petite sélection de textes littéraires - le ”canon”.
  - Canon : ”Règle directrice appliquée dans un art, dans une discipline intellectuelle”
- Avec lecture distante : possible d’inclure aussi des textes littéraires aujourd’hui oubliés

L'abattoir de la littérature

- Constat de Moretti : seul 0.5% d'une production littéraire est vraiment étudiée => "abattoir" (slaughterhouse) de la littérature (cf. Franco Moretti, "The Slaughterhouse of Literature", In MLQ : Modern Language Quarterly, Volume 61, Number 1, Duke University Press, March 2000)
- Abattoir ? Histoire de la littérature ne conserverait que les oeuvres littéraires de meilleure qualité
- Mais ne permet pas de bien connaître le contexte littéraire dans lequel ces oeuvres ont été publiées : genres et thèmes à la mode par exemple

LES ARGUMENTS EN FAVEUR DE LA LECTURE DISTANTE (3)

Troisième argument : garantir une plus grande objectivité

> [...] what we really need is a little pact with the devil : we know how to read texts, now let's learn how not to read them. Distant reading : where distance, let me repeat it, is a condition of knowledge : it allows you to focus on units that are much smaller or much larger than the text : devices, themes, tropes—or genres and systems.

Franco Moretti, "Conjectures on world literature", New Left Review, 1, January-February 2000, https ://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature

Exemples d'application

LES GRAPHIQUES : LA LONGUEUR DES TITRES

- Franco MORETTI, “Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)”, Critical Inquiry, 36-1 (2009), p. 134-158.
- Etude de la longueur des titres de romans britanniques entre 1740-1850, et de leur évolution

LES GRAPHIQUES : LA LONGUEUR DES TITRES

- Etude quantitative sur l'évolution de la longueur des titres de ces romans
- Calculs "simples" sur le nombre de mots qui composent un titre sur la période, en utilisant les métadonnées décrivant ces ouvrages (métadonnées = auteur, titre, année... ce que l'on trouve dans un catalogue de bibliothèque)

LES GRAPHIQUES : LA LONGUEUR DES TITRES

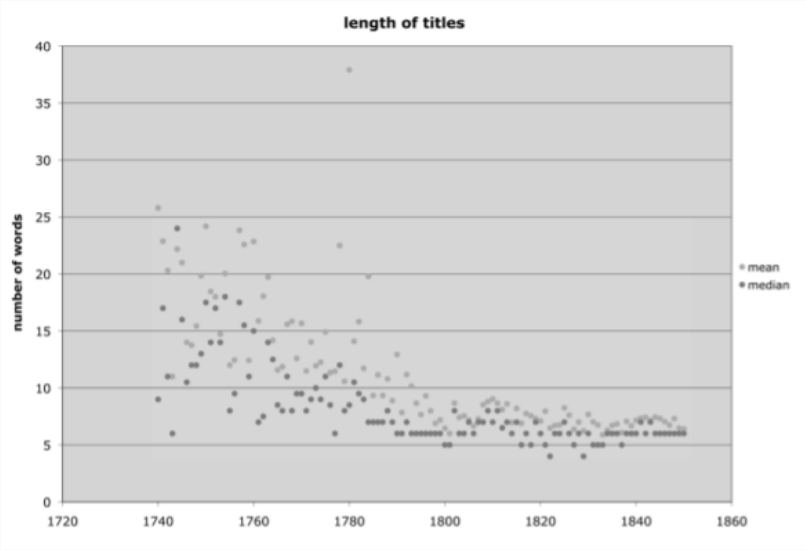


FIGURE : Longueur moyenne et médiane des titres

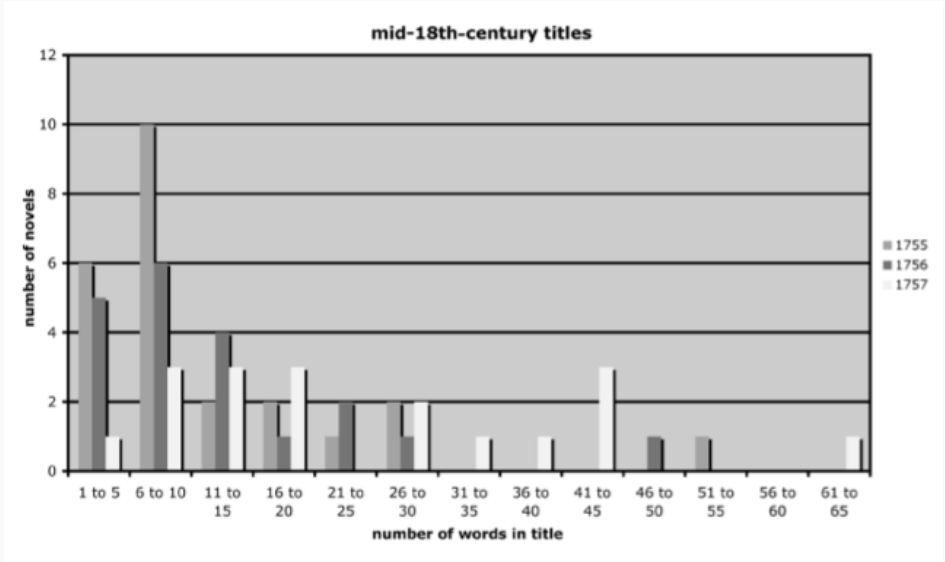


FIGURE : Longueur des titres par nombre de romans

LES GRAPHIQUES : LA LONGUEUR DES TITRES

Titres deviennent de plus en plus courts.

- Le marché des livres en Grande-Bretagne se développe considérablement entre la 2ème moitié du XVIIIème siècle et la 1ère moitié du XIXème siècle.
- Besoin de donner plus de visibilité immédiate aux romans publiés : titres courts plus faciles à retenir + mieux vaut un titre court quand il apparaît dans un catalogue ou une critique

Deviennent également de plus en plus similaires : on emploie de plus en plus les mêmes mots pour donner un titre à un roman.

- Compression du sens
- Développement de ”signaux” spéciaux pour placer les livres dans le bon créneau commercial (”marketing”)

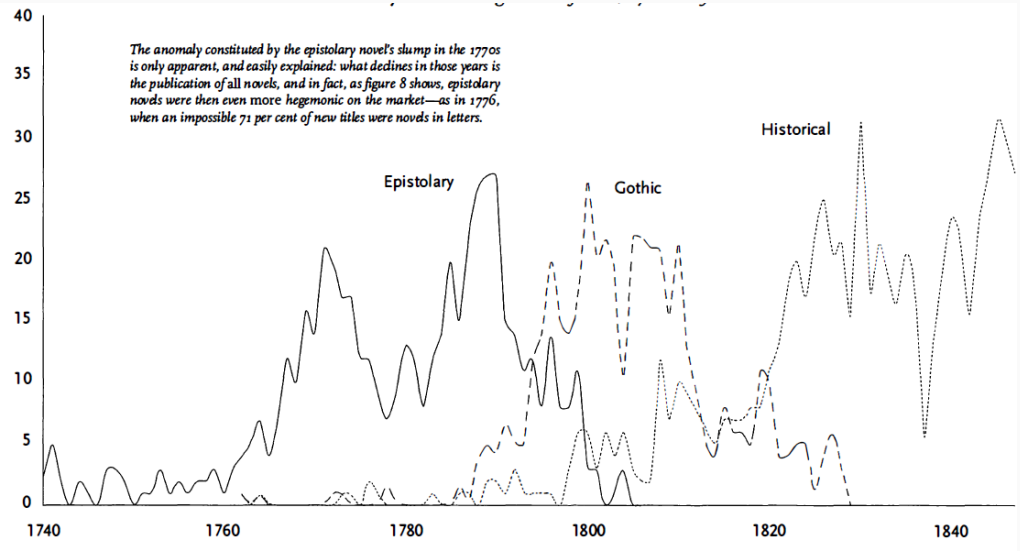
LES GRAPHIQUES : LA LONGUEUR DES TITRES

Conclusion méthodologique :

- Plaidoyer pour l'utilisation de la lecture distante. Ex. sur l'utilisation des déterminants définis et indéfinis dans les titres : "Here is a modest example of what quantitative stylistics could do : take those units of language that are so frequent that we hardly notice them, and show how powerfully they contribute to the construction of meaning". (p.156)

- Pas UNE méthode, mais **une manière d'aborder les textes**

LES GRAPHES : LES GENRES LITTÉRAIRES



Graphs, Maps and Trees, p.15

- Graphes qui dressent la carte des nouveaux romans historiques, gothiques et épistolaires, publiés chaque année pendant un siècle entre 1740 et 1850.
- Trois vagues de genres littéraires : la vague "épistolaire", puis la vague "gothique" et la vague "historique"

LES GRAPHES : LES RÉSEAUX

- Nombreux objets littéraires peuvent être utilisés pour faire des graphes
- Très souvent : graphes représentent des réseaux
- Quelques exemples : une correspondance épistolaire, les mots d'un livre (in pourrait relier les mots quand ils apparaissent dans la même phrase), des personnages de romans

LES GRAPHES : LES RÉSEAUX

- Franco Moretti, “Operationalizing” : or, the function of measurement in modern literary theory, in : Pamphlet, 6, Stanford Literary Lab, 2013.

- Approche empirique de la littérature avec la mesure de l’espace qu’un personnage occupe dans un texte
- Exemple de Phèdre de Racine :
  - Phèdre (personnage principal) prononce 29%,
  - Thésée prononce 14%,
  - Hyppolite prononce 21% de tous les mots.
- Compter simplement cela ne suffit pas pour déterminer le poids et la position d’un personnage dans un texte
- Moretti s’appuie sur la théorie des réseaux : la direction des relations est tout aussi importanet pour comprendre le poids d’un individu (sa place) dans un réseau - ici, réseau des personnages.

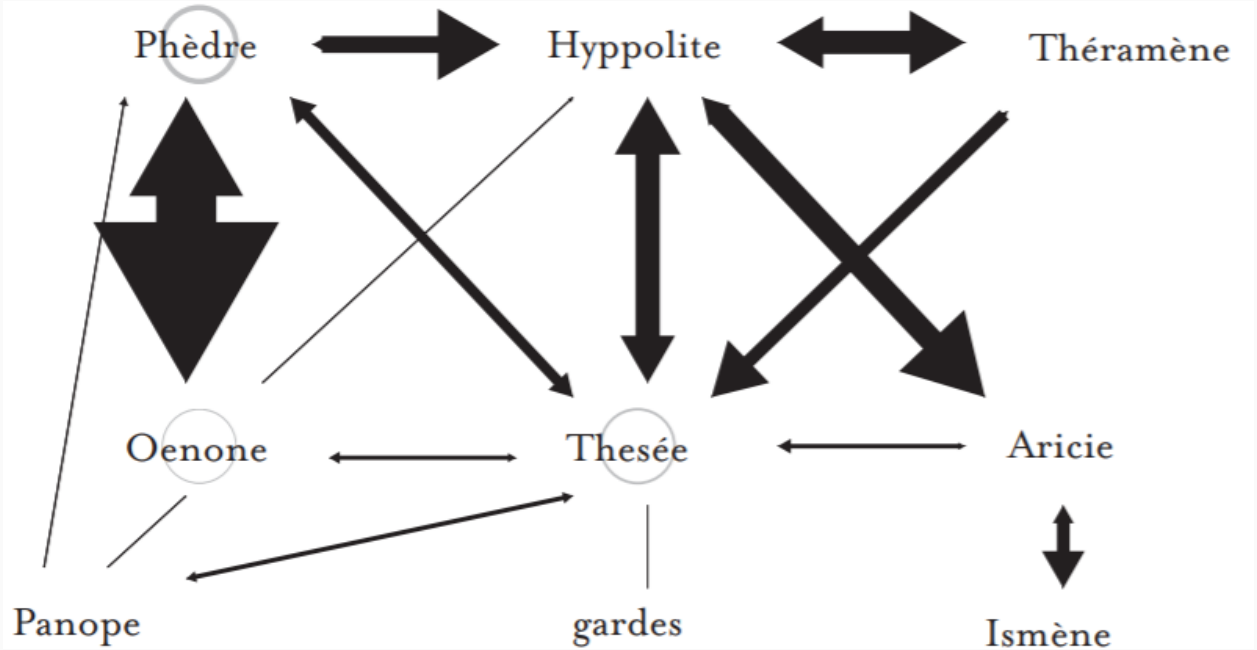
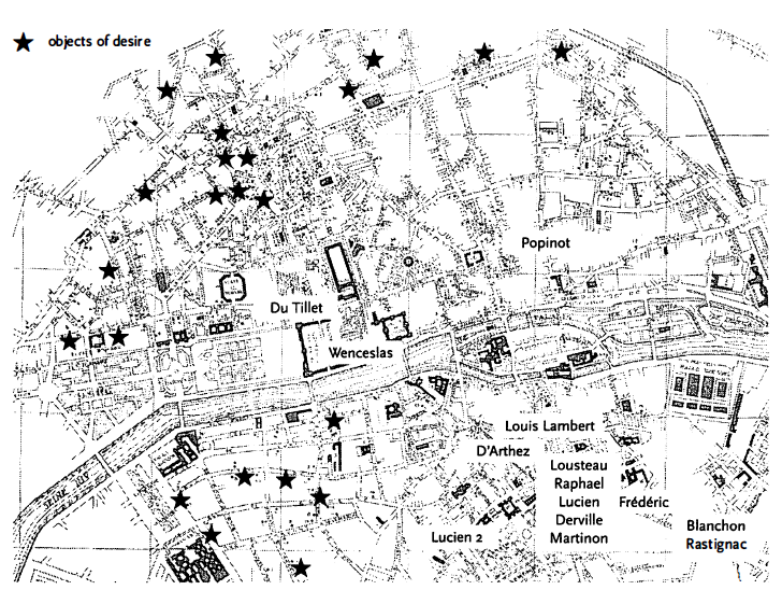


FIGURE : Direction et poids des relations entre les personnages

- Phèdre parle surtout à sa confidente Oenone, et moins avec son mari Thésée ou son amant Hyppolite.
- Réseat montre que le personnage principal du roman, Phèdre, n'est pas au centre des relations, mais c'est Thésée, son mari.

LES CARTES : LIEUX DE VIE DES PERSONNAGES LITTÉRAIRES



Graphs, Maps and Trees, p.55)

- Carte représentant les lieux où vivent :
  - les personnages masculins de romans (XIXe siècle) se passant à Paris (représentés par leurs noms)
  - Les personnages féminins qu'ils désirent (représentés par des étoiles).
- Hommes et femmes vivent dans des mondes sociaux différents.
  - Les hommes dans le monde artistique et intellectuel de la rive gauche.
  - Les femmes dans le monde commercial et plus riche de la rive droite, ou l'élite aristocratique du faubourg Saint-Germain.

L'ANALYSE DE CORPUS TEXTUELS

LA LECTURE DISTANTE : EST-CE SI NOUVEAU ?

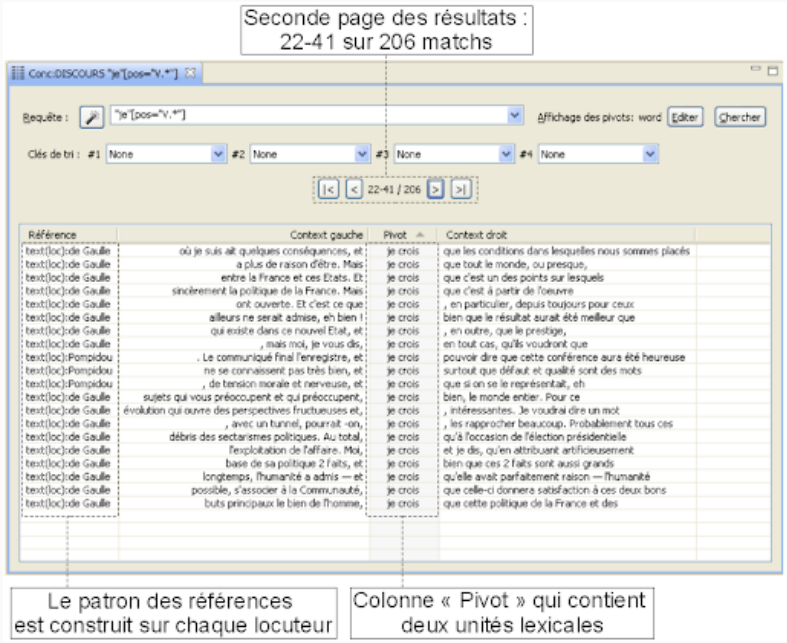
- Franco Moretti inscrit dans une tradition qui débute avant les années 2000
- Lire des corpus à distance, notamment avec l'aide de l'informatique : méthode pas si récente que cela
- Clairement formulée, de manière percutante et imagée, par Moretti
- La "lecture distante" est très liée à l'histoire des humanités numériques

LA LECTURE DISTANTE : EST-CE SI NOUVEAU ?

- Années soixante et soixante-dix : débuts officiels de la linguistique computationnelle
- Nouveaux outils en traitement automatique des langues (TAL) ou en linguistique de corpus
- Utilisation de corpus de référence pour une langue donnée : Henry Kučera et Nelson Francis créent le Brown University Standard Corpus of Present-Day American English, ou "Brown Corpus" (anglais américain) (fin des années soixante-dix)
- Indexer avec des ordinateurs tous les mots de la langue, dans tous les types de discours : base Frantext de l'ATILF pour le français (5410 textes littéraires, philosophiques, scientifiques et techniques, soit 254 millions de mots)

COMPTER DES MOTS

- Compter des mots : activité commune, depuis les années 1960, à des linguistes, des historiens, des politistes, des sociologues
- Objectifs :
  - Quantifier les données
  - Faire des calculs statistiques de façon à identifier des régularités statistiques dans les textes
  - Etablir des concordances : établir un tableau qui rassemble toutes les occurrences d'un terme en contexte



LEXICOMÉTRIE, TEXTOMÉTRIE…

- Accès à des outils informatiques, qui ne cessent de se multiplier, pour étude des textes
- Lexicométrie, rebaptisée dans les années quatre-vingt-dix “textométrie”.
- Logométrie, analyse automatique, statistique linguistique, statistique lexicale ou linguistique quantitative, statistique textuelle, voire analyse des données en linguistique.

LES LIMITES DE LA LECTURE DISTANTE

LE CORPUS ÉTUDIÉ : LES ABSENCES

- Touche à la question du corpus : pour réaliser ce type d'étude, nécessaire de travailler sur un corpus bien adapté
- C'est de la manière dont on a constitué le corpus, que va découler les résultats.

LE CORPUS ÉTUDIÉ : LES ABSENCES

Ian Milligan, ”Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997-2010”, in Canadian Historical Review, n° 94 (4), décembre 2013, p. 540-569.

- Analyse des mentions des journaux canadiens du XIXe siècle dans les articles de la Canadian Historical Review
- Grands journaux ”centraux” et anglophones, numérisés et disponibles pour les chercheurs : de plus en plus utilisés aux dépens des journaux francophones et/ou locaux
- Tendance à produire une vision centrale et anglo-saxonne de l’histoire canadienne.

LE CORPUS ÉTUDIÉ : LES ABSENCES

- Biais créé par la numérisation : ce qui est numérisé a été choisi en fonction de critères qui dépendent d'une évaluation scientifique, mais aussi d'un financement, d'une mode, d'une histoire particulière…
  - Certaines sources marginalisées par les universitaires (privilégier le canon)
  - Cf. Les accusations portées contre Google : vision anglo-saxonne du patrimoine mondial (numérisations menées dans le cadre de Google Livres)
- Certaines langues moins bien représentées que d'autres
- Or ce qui n'est pas en ligne, n'existe pas : Gabriela Ossenbach, "If it's not online, it doesn't exist", IJHE, n° 5(1), 2015, p. 80-82.

LE CORPUS ÉTUDIÉ : LES ERREURS

- Autre biais : les erreurs inhérentes à la numérisation
- Si la réussite de la reconnaissance automatique des caractères (OCR) a considérablement augmenté, elle n'est pas parfaite
- Difficulté à reconnaître certaines polices par exemple
- Reconnaissance de l'écriture manuscrite : encore à améliorer
- Numérisation de mauvaise qualité / mal réalisée => erreurs dans l'OCR

LES PROBLÈMES POSÉS PAR GOOGLE NGRAM VIEWER

Google Ngram viewer : bon outil pour lire à distance ?

- Impossible d'interpréter correctement ces graphiques sans avoir accès aux documents numérisés sur lesquels se basent ses résultats.
  - Tâche colossale, au vue de l'ampleur du corpus numérisé
  - Difficilement réalisable car beaucoup de documents sont encore sous droits et pas consultables.
- Ngram viewer : outil heuristique, destiné à poser des questions et à déceler des tendances.
- Erreurs dans les métadonnées décrivant un texte : datation fausse, dont peuvent résulter des pics temporels factices qui ne représentent pas réellement la littérature d'une époque

QUELLES ÉTUDES ET QUELLES MÉTHODES ?

LA PREMIÈRE GUERRE MONDIALE SUR TWITTER

Frédéric Clavert. Échos du centenaire sur le web et sur Twitter : Partage de liens et Grande Guerre. Revisiter la commémoration. Pratiques, usages et appropriation du Centenaire de la Grande Guerre., Sarah Gensburger ; Valérie Tesnière, Mar 2016, Nanterre, France.

- Frédéric Clavert : collecte des tweets évoquant la Grande Guerre ou concernant les commémorations de la Grande Guerre en utiliant des hashtags comme ”ww1”.
- Utilisation du logiciel 140dev. Exploite l’API de streaming de Twitter, qui permet d’obtenir jusqu’à 1% du flux de l’ensemble des tweets émis en temps réel.
    - API = Interface de programmation. Solution qui permet à de applications informatiques de communiquer entre elles et de s’échanger des données.
- Collecte de plusieurs millions de tweets

LA PREMIÈRE GUERRE MONDIALE SUR TWITTER

- Tweet et RT = création de liens entre les tweets
- Liens visualisés grâce à un logiciel de visualisation de réseaux (Gephi)
- Utilisation d'un logiciel d'analyse textuelle (Iramuteq)

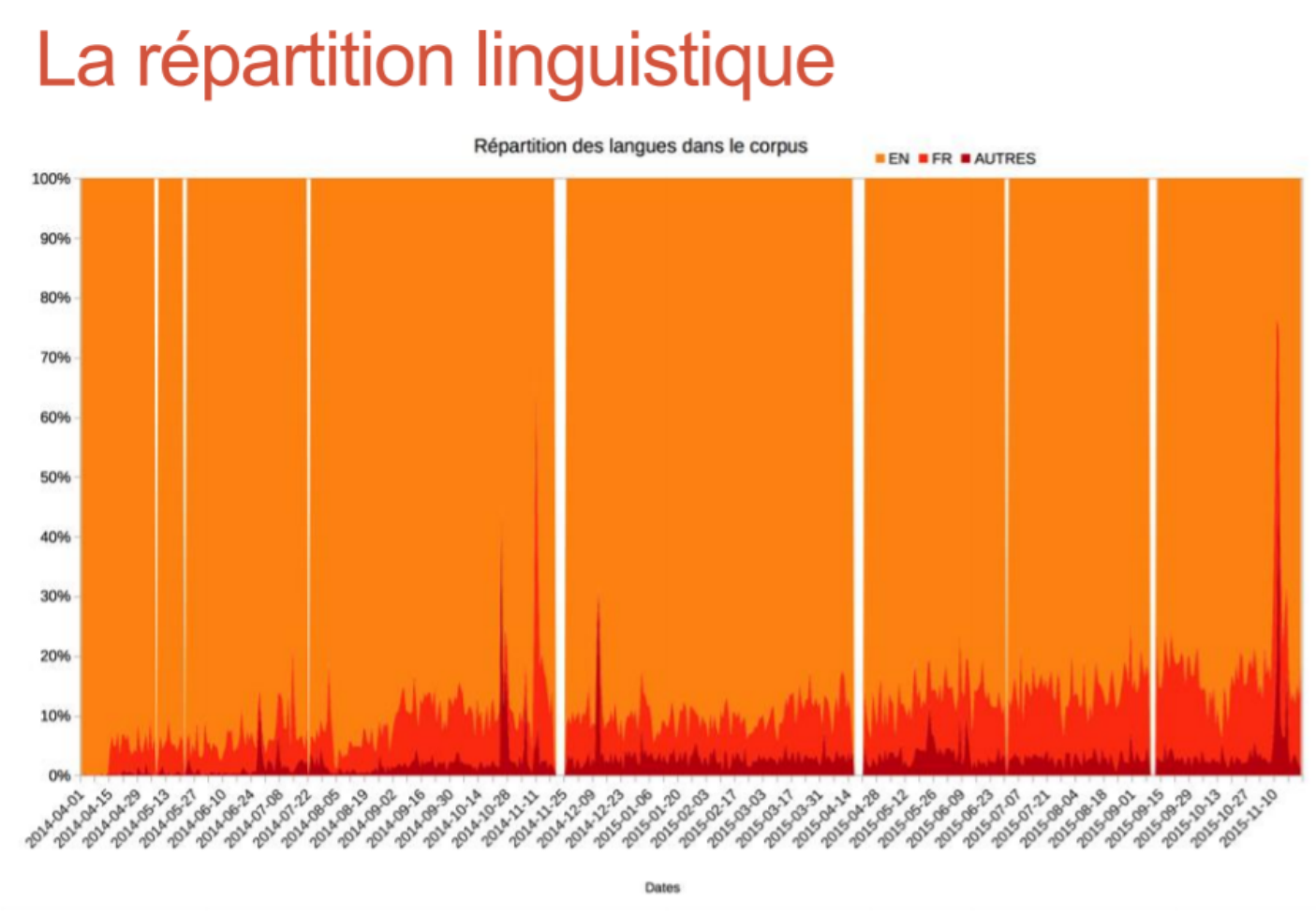


FIGURE : Les langues du corpus

Figure : Nuages de mots

LA PREMIÈRE GUERRE MONDIALE SUR TWITTER

- Appropriation du Centenaire par les utilisateurs de Twitter
- Base de données des Morts pour la France = nouveau lieu de mémoire en ligne ?

LE PEUPLE CONTRE LA DÉMOCRATIE

Bonin, From antagonist to protagonist : ‘Democracy’ and ‘people’ in British parliamentary debates, 1775–1885, Digital Scholarship in the Humanities, 2019

- Démocratie = terme central du discours politique actuel
- Défini comme ”le pouvoir du peuple”
- Ce terme a t-il toujours eu le même sens ?

LE PEUPLE CONTRE LA DÉMOCRATIE

- Analyse des débats parlementaires britanniques entre 1775 et 1885
- Utilise lecture distante et lecture proche :
  - lecture distante à l'aide des outils de la linguistique des corpus
  - examen plus approfondi de certains débats et acteurs clés

LE PEUPLE CONTRE LA DÉMOCRATIE

- Numérisation + logiciels d'analyse des textes => changement dans le paysage de l'histoire intellectuelle
- Vastes quantités de données à étudier : documents juridiques, coupures de journeaux... difficiles à étudier avec les méthodes classiques
- Jusqu'à présent, forte réticence de l'histoire de la pensée politique à prendre un "virage numérique"
- Corpus débats parlementaires = corpus remarquable pour les linguistes, les historiens et les théoriciens de la politique
    - Largement numérisés, corrigés, formatés et rendus accessibles aux chercheurs.
    - Donne un aperçu des utilisations plus banales des concepts.
    - Comparaisons entre pays ayant des institutions similaires, de manière diachronique et synchrone

LE PEUPLE CONTRE LA DÉMOCRATIE

- Taille écrasante du corpus
  - Corpus britannique de 1803 à 2005 : 1,6 milliard de mots.
  - Entre 1803 et 1900, les mots ”démocratie” et ses dérivés, ainsi que ”peuple” apparaissent plus de 4 300 et 174 000 fois, respectivement.
- Utilisation d’outils lexicométriques pour travailler sur ce corpus

LE PEUPLE CONTRE LA DÉMOCRATIE

Travail sur les co-occurrences :

- **Co-occurrence** : présence simultanée de mots ou groupes de mots dans un même extrait de taille restreinte, ce qui indique une proximité non seulement spatiale mais aussi sémantique.
- Différents logiciels peuvent calculer un score de co-occurrence entre deux termes d'une manière qui met l'accent sur la spécificité du terme co-occurrent.
- Ex. "la démocratie d'Athènes était turbulente" : les mots les plus proches de "démocratie" sont "le" et "de". Cependant, comme ces deux mots sont également extrêmement fréquents dans tout texte, ils auront un score de co-occurrence inférieur à celui de "Athènes" et "turbulent", mots moins présents ailleurs dans le corpus.
- En générant des listes et des visualisations des co-occurrences de "démocratie" ou de "peuple" => repérer les différents champs sémantiques dans lesquels ces termes ont été utilisés.

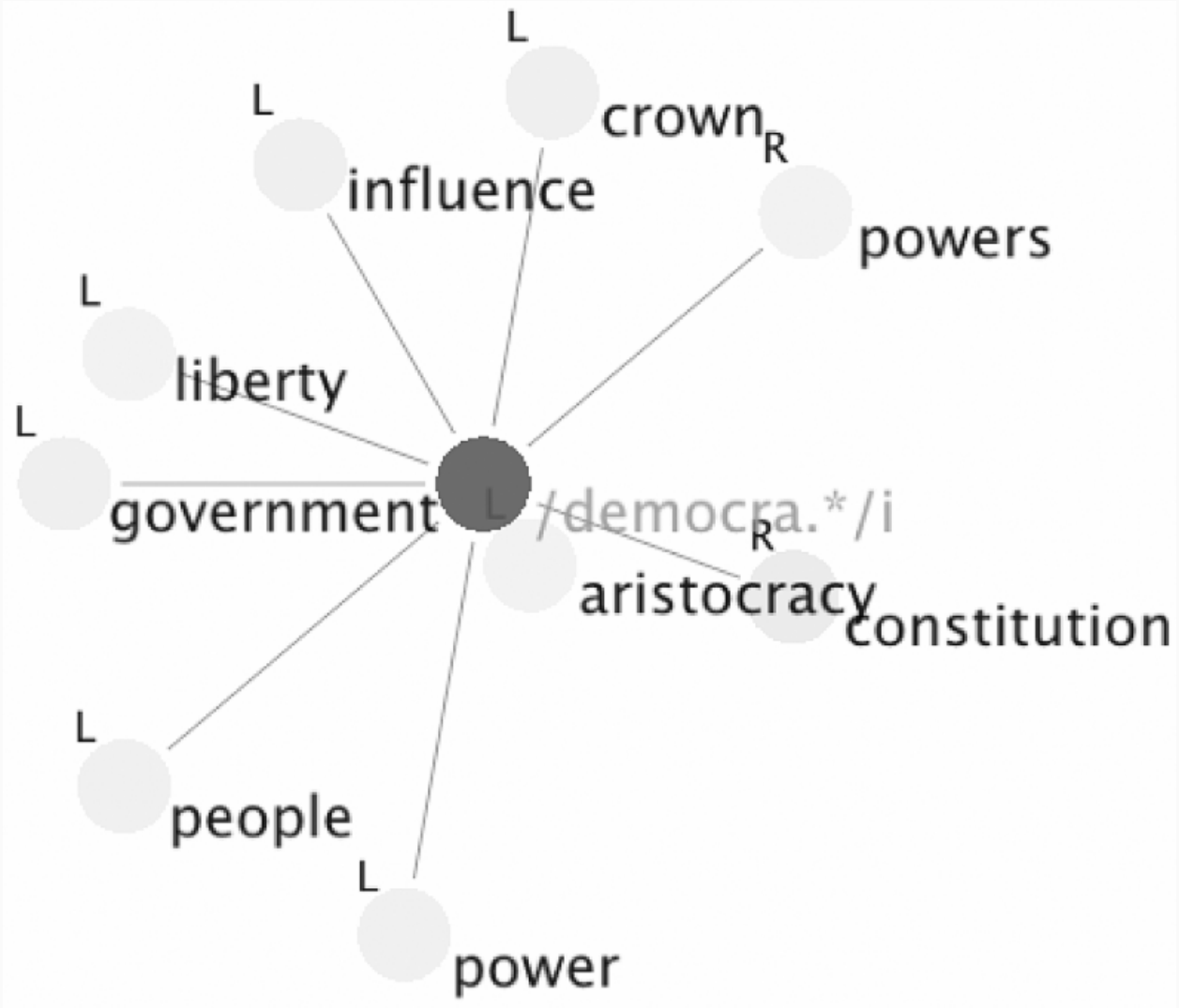


Figure : ‘Democra*’ dans les débats parlementaires, 1775–84

LE PEUPLE CONTRE LA DÉMOCRATIE

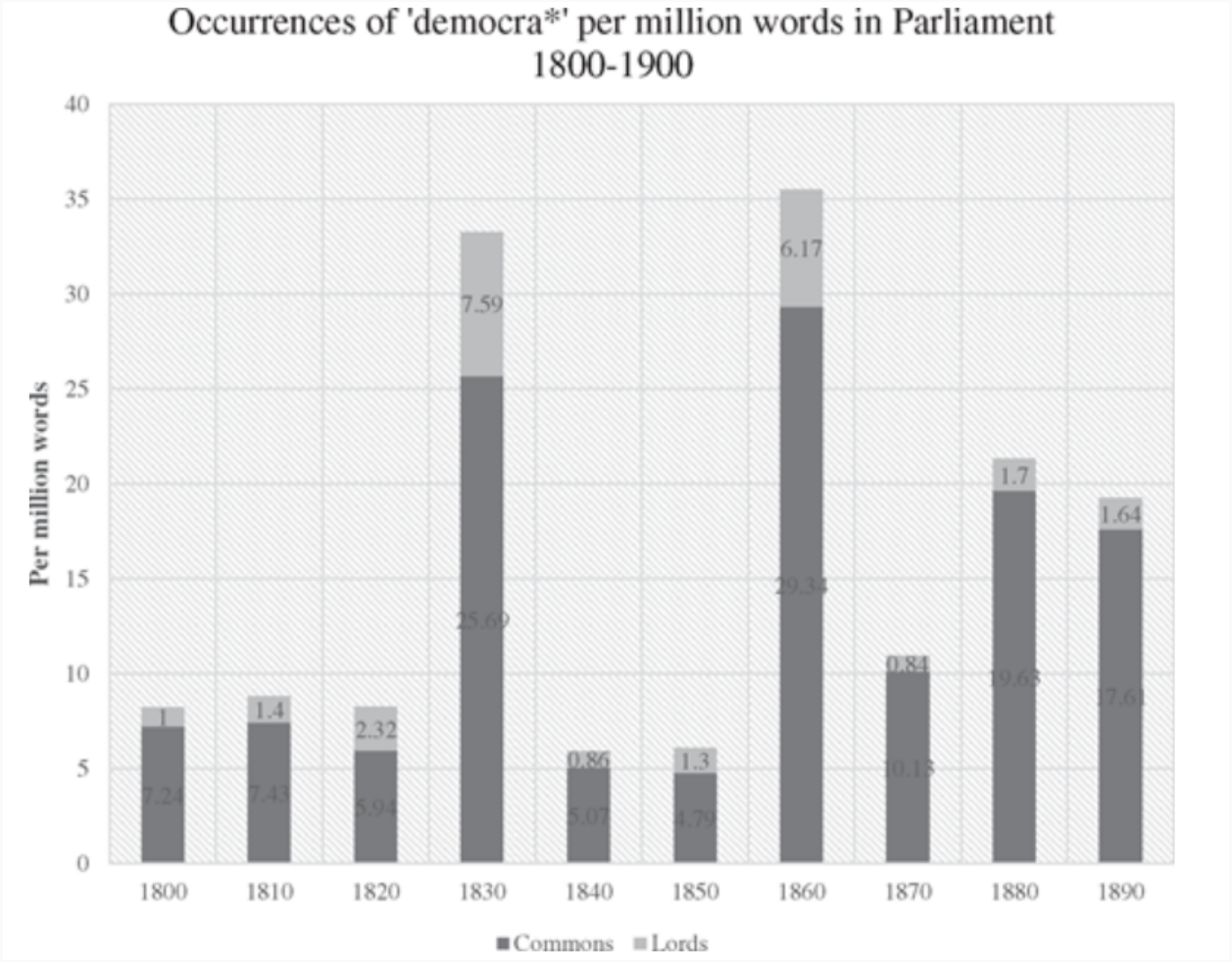


Figure : Occurrences de ‘democra*’ par million de mots par décennie

LE PEUPLE CONTRE LA DÉMOCRATIE

- Contrairement aux idées reçues, ”démocratie” et ”peuple” = opposés dans les discours des députés.
- Au XVIIIe siècle, la démocratie, au sens politique du terme,considérée comme une menace pour la constitution britannique.
- XIXe siècle, en Grande-Bretagne : ”démocratie” pas caractérisée par le pouvoir du ”peuple” mais par le pouvoir des classes inférieures, de la ”populace” et de la ”mob” (foule).
- Dans débats parlementaires : ”peuple” = un antagoniste de la ”démocratie”, et non son protagoniste.