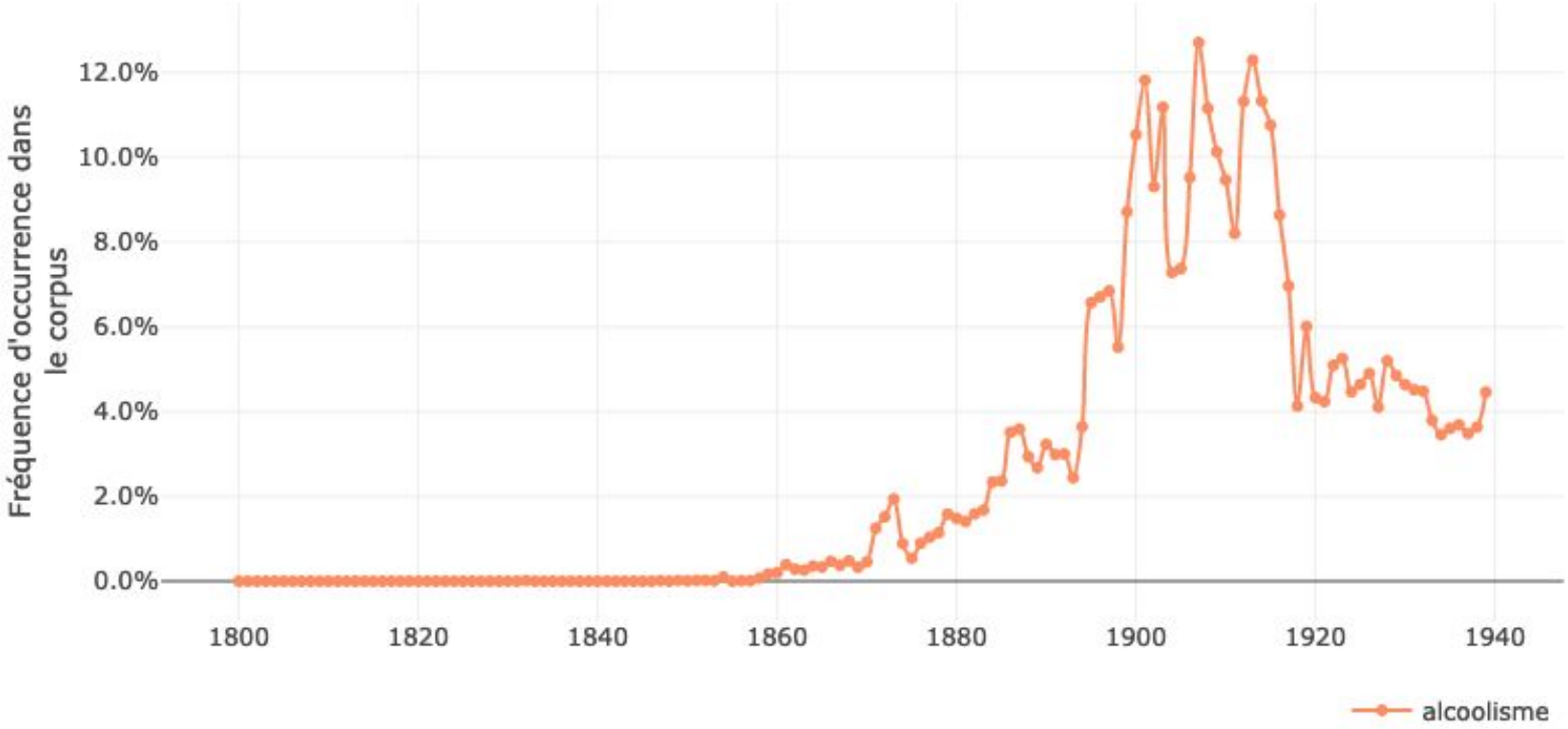
Gallicagram, un outil de lexicographie

Le *big data* sans ses gros sabots ?

Les textes comme fossiles des mentalités

- Pour savoir comment vivaient nos ancêtres, on recourt aux fossiles
- Les idées, elles, laissent moins de traces
- Pour faire une histoire des mentalités, les textes sont les meilleurs fossiles disponibles
- Une nuance : les textes renseignent davantage sur les auteurs que sur les lecteurs

Les textes comme fossiles des mentalités

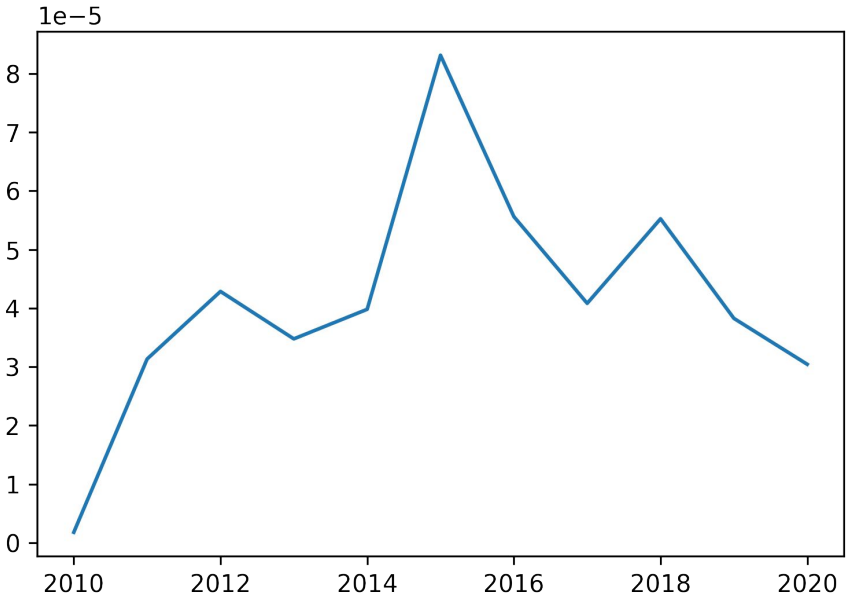


Ngram viewer et les chercheurs, un “rendez-vous manqué” ?

- En 2010, Google a lancé *Ngram viewer*, basé sur Google Books (4% des livres imprimés depuis Gutenberg)
- Dans *Science*, l’équipe a annoncé triomphalement la naissance d’une “*new science*”, les “*culturomics*”
- 11 ans plus tard, quel bilan ?

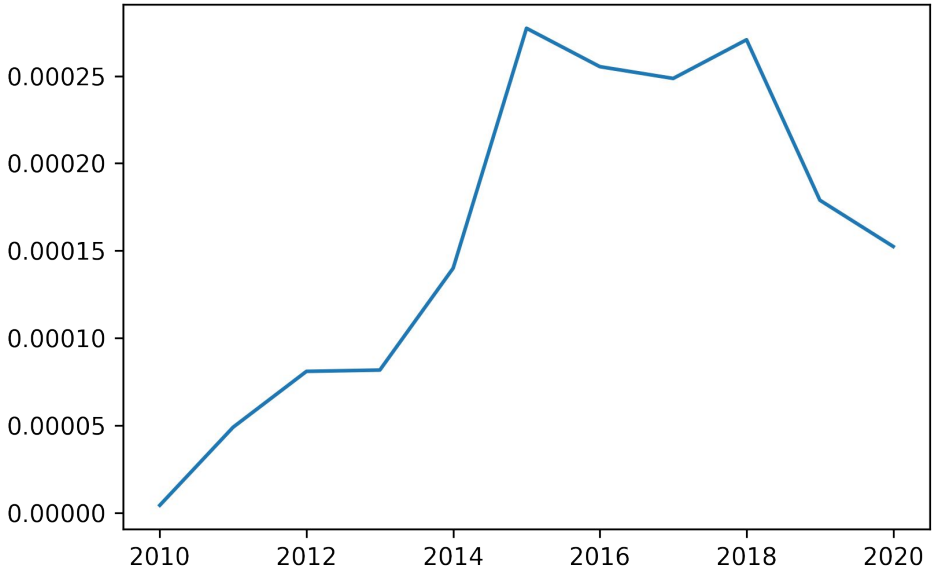
Ngramisons les culturomics

- Google Scholar permet de mesurer l’importance du champ
- Problème : culturomics désigne aussi une technique de microbiologie
- A droite, le nombre résultats pour la recherche “culturomics -microbiote -microbiome -microbial -bacterial”, divisé par le nombre total d’articles



Ngramisons les culturomics

- Google Scholar permet de mesurer l'importance du champ
- Problème : culturomics désigne aussi une technique de microbiologie
- A droite, le nombre résultats pour la recherche "culturomics -microbiote -microbiome -microbial -bacterial", divisé par le nombre total d'articles
- "ngram viewer" donne une forme similaire



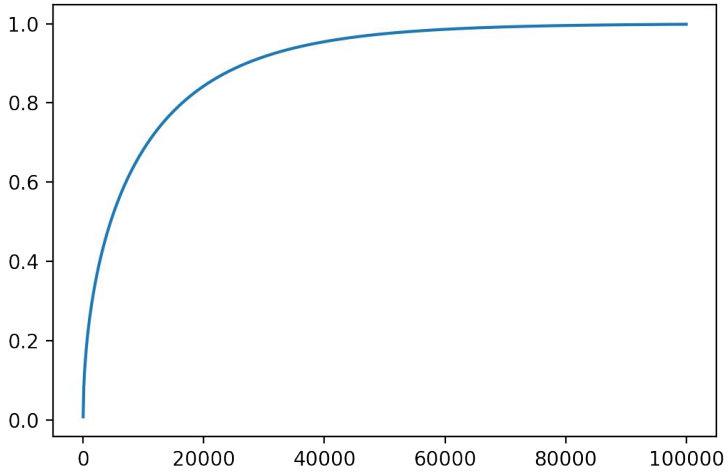
Ngram viewer et les chercheurs, un “rendez-vous manqué”

- Malgré une délicieuse interface, l’utilisation de Ngram viewer en sciences sociales n’a pas pris
- Les études l’utilisant sont *très* macroscopiques : l’article le plus remarqué est The Changing Psychology of Culture From 1800 Through 2000 (Greenfield, 2013)
- Deux raisons :
  - Ngram viewer suscite la méfiance à cause de :
    - Son manque de transparence
    - L’absence de métadonnées
  - Le corpus Google Books ne peut être contrôlé ou restreint

Un gigantesque *dataset*, à quoi bon ?

- On cherche à estimer θ, l'importance culturelle d'un syntagme
- Si on a n données, le pourcentage d'erreur suit la loi suivante :

$$\left| \frac{\hat{\theta} - \theta}{\theta} \right| =\approx \mathcal{N}(\theta, \frac{1}{\sqrt{n}})$$

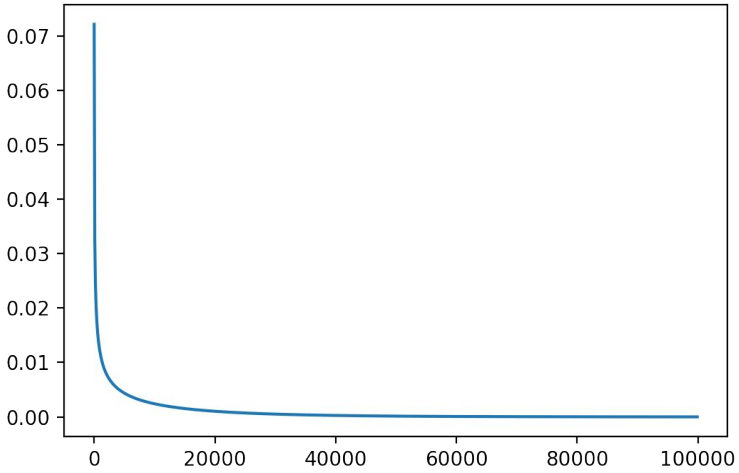


- Autrement dit, la probabilité d'être juste à 1% près se comporte comme ceci :

Un gigantesque *dataset*, à quoi bon ?

- On cherche à estimer θ, l'importance culturelle d'un syntagme
- Si on a n données, le pourcentage d'erreur suit la loi suivante :

$$\left| \frac{\hat{\theta} - \theta}{\theta} \right| =\approx \mathcal{N}(\theta, \frac{1}{\sqrt{n}})$$



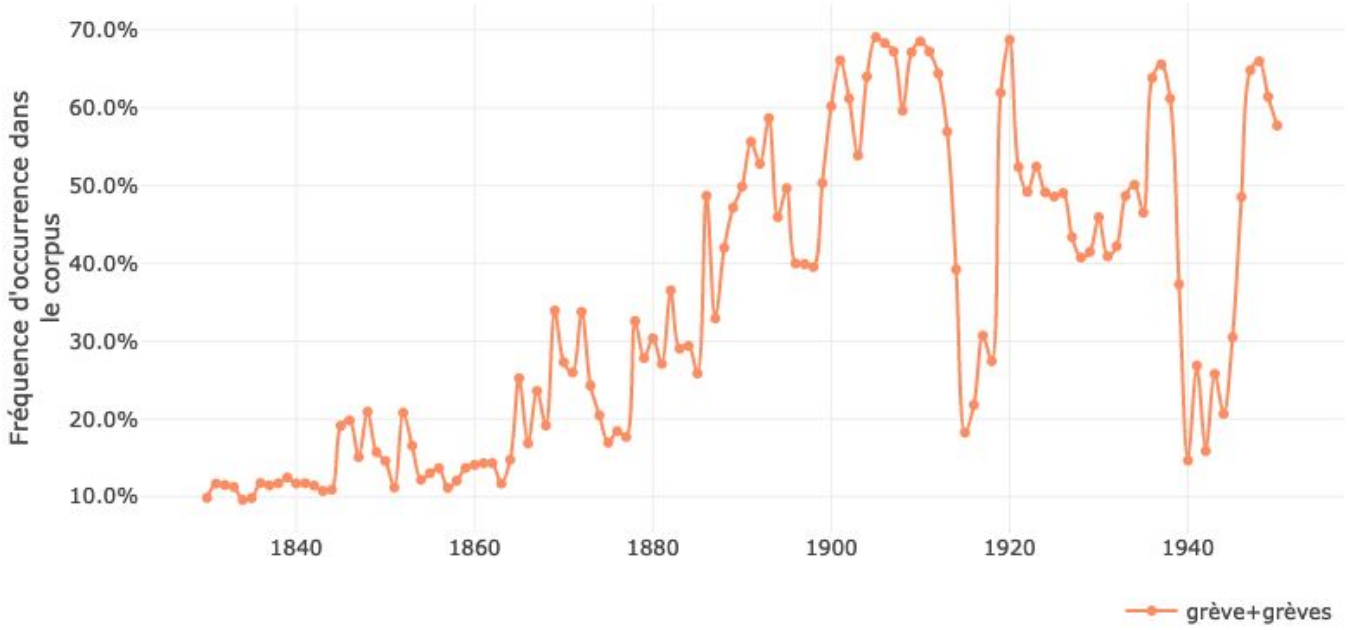
- Autrement dit, la probabilité d'être juste à 1% près se comporte comme ceci :
- Et le "gain marginal" d'une donnée supplémentaire, comme ceci :

Un gigantesque *dataset*, à quoi bon ?

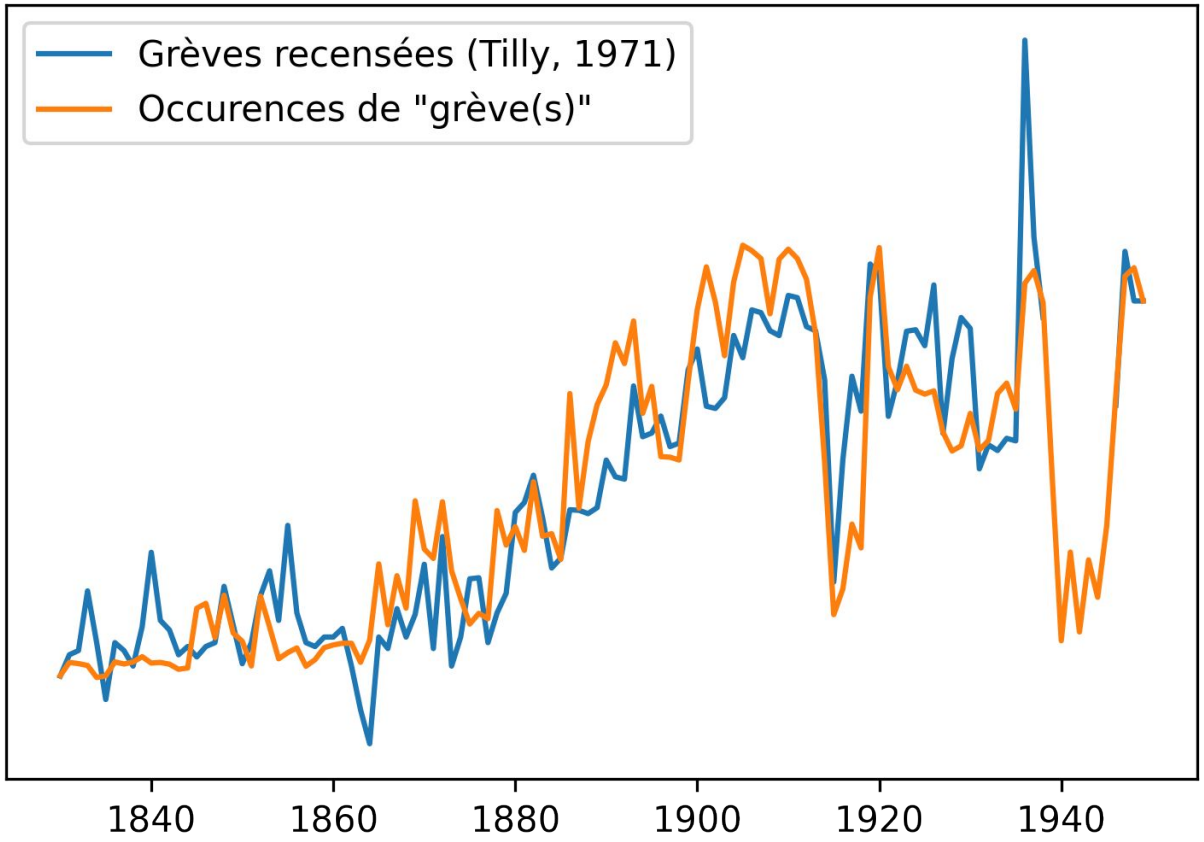
- Le volume de données est donc à rendements décroissants, et tendant rapidement vers 0
- Google a choisi d'offrir un "bloc" de données empêche d'affiner l'analyse
- Un corpus transparent et contrôlable comme Gallica peut donc offrir des résultats plus pertinents, même s'il est nettement plus réduit
- Par exemple, en se restreignant à la presse, on devrait être plus proche des événements
- Essayons avec un type d'événements aisément quantifiables : les grèves

Un corpus restreint, des résultats plus fiables ?

On mesure le nombre de résultats de la recherche “grève” ou “grèves”,



Un corpus restreint, des résultats plus fiables ?



Corrélation :
r = .90

Pour ngram viewer :
r = .82

Prise en main de Gallicagram

Les fonctions de base du logiciel

- La syntaxe de recherche
  - L’insensibilité à la casse
  - La recherche ET
  - La recherche OU
- La délimitation des bornes chronologiques
- La résolution du graphe
- Les modes de recherche
- L’accès au corpus sous-jacent
- Les options interactives dans le graphique
  - Le jeu des étiquettes
  - Le zoom
  - La sélection des courbes
  - Le lissage des courbes
- Le téléchargement des données et du graphe

De nombreux corpus à explorer

- Les corpus de Gallica
  - Corpus massifs
  - Corpus restreints
  - Corpus personnalisés
- Le cas de Ngram Viewer/Google Books
- Les autres corpus francophones
- Les corpus en langues étrangères

Les fonctions avancées de Gallicagram

- La distribution chronologique des documents constitutifs de la base de données
- Le rééchelonnement des résultats
- La comparaison de deux syntagmes
- La comparaison des résultats entre les modes de recherche
- L’étude des corrélations

Dépasser les limites de Ngram Viewer

Dépasser les limites de Ngram Viewer

Conditions de validité des observations macro :

- L’accès à la totalité du corpus
- La maîtrise amont et aval du corpus
- La diversité des méthodes d’analyse

Certaines critiques, pourtant rédhibitoires, à l’égard de l’outil de Google n’ont pas encore été formulées dans la littérature

Indétermination totale du corpus utilisé par Ngram

La liste exhaustive des documents contenus dans le corpus exploité par Ngram n’est pas disponible :

- Impossible de savoir quels documents de la bibliothèque Google books ont été éliminés lors du filtrage (qualité d’océrisation, métadonnées lacunaires)
- Impossible de savoir quels documents étaient numérisés lors de la constitution du corpus de Ngram Viewer (date de numérisation)

L’accès au corpus de livres de Google est très limité

Renvois aléatoires à Google Books : 1, 2

Restrictions d’accès aux documents sous droits

Liste de résultats de recherche extrêmement fragmentaire dans Google Books :

- Illustration : 1
- Impossible de Gallicagramiser Google Books : inexactitude du nombre de résultats affiché + indétermination des objets comptés (documents, pages, occurrences)

L’accès à la totalité du corpus

Rupture radicale entre l’échelle microscopique et l’échelle macroscopique

Gallicagram assure un accès à la totalité du corpus sous-jacent :

- Tester et éliminer des hypothèses
- S’assurer de la qualité de l’OCR et donc de la représentativité des résultats

La maîtrise amont et aval du corpus

Le dialogue des échelles micro/macro ne suffit pas à la maîtrise du corpus pour accréditer les observations macroscopiques :

- Maîtrise amont : délimitation des corpus
- Maîtrise aval : analyse de la structure des corpus

Délimiter le corpus : éviter les effets de structure

Gallicagram se restreint aux :

- Textes rédigés dans une langue prédéfinie
- Textes océrisés, avec un taux de succès supérieur à 50%
- A des textes de même nature contrairement à Ngram Viewer : 0, 1, 2, 3

Le corpus est personnalisable :

- Corpus thématiques (presse religieuse, presse féministe…)
- Corpus sur-mesure
- Bibliothèques régionales

Gallicagram calcule les indicateurs en valeur relative, pour compenser les variations du volume de la base

Analyser la structure des corpus : évaluer les effets de structure

Un volume minimum de documents pour valider une observation (1789-1944 pour la presse)

Gallicagram bénéficie de la grande qualité des métadonnées associées aux documents de Gallica ≠ Ngram Viewer/Google Books. Cela permet de répondre à deux questions essentielles :

- parmi quoi cherche-t-on ?
  - Des onglets dédiés dans l'application à la description des deux corpus massifs de Gallica
- quelle est la composition des résultats ?
  - Une application complémentaire dédiée à l'analyse des résultats de recherche dans la presse de Gallica : Gallicapresse

La diversité des méthodes d'analyse

$$V(s,i) = \frac{\text{nombre de documents incluant } s \text{ pour la période } i}{\text{nombre de documents dans le corpus pour la période } i} = \frac{\sum_{j=1}^{n_d} \mathbb{1}_{s \in d_j \bigcap y(d_j)=i}}{\sum_{j=1}^{n_d} \mathbb{1}_{y(d_j)=i}}$$

$$P(s,i) = \frac{\text{nombre de pages incluant } s \text{ pour la période } i}{\text{nombre de pages dans le corpus pour la période } i} = \frac{\sum_{j=1}^{n_p} \mathbb{1}_{s \in p_j \bigcap y(p_j)=i}}{\sum_{j=1}^{n_p} \mathbb{1}_{y(p_j)=i}}$$

$$M(s,i) = \frac{\text{nombre d'occurrences de } s \text{ pour la période } i}{\text{nombre de n-grammes dans le corpus pour la période } i} = \frac{\sum_{j=1}^{n_m} \mathbb{1}_{s \in m_j \bigcap y(m_j)=i}}{\sum_{j=1}^{n_m} \mathbb{1}_{y(m_j)=i}}$$

La diversité des méthodes d’analyse

Le mode de recherche est généralement imposé par l’architecture des moteurs de recherche

Pour Gallica : plusieurs modes de recherche permettront des comparaisons internes

- La pertinence de chaque mode de mesure est fonction de la recherche effectuée
- Privilégier la recherche au document dans le corpus de presse
- Essayer la recherche par page dans les corpus de presse restreints pour gonfler le dénominateur et réduire le bruit tout en tirant profit de la mise en page thématique des journaux
- Rechercher par page dans le corpus de livres sauf pour les termes rares (<5% d’occurrence dans la recherche par document)
- Comparer les résultats entre les modes de recherche et entre les corpus
- Utiliser les outils avancés pour tester les observations

Les données exploitées et les traitements effectués

La récupération des données

Extraction depuis les API de recherche des différentes bibliothèques ou extraction robotisée depuis leur site internet grand public (recherche exacte) : 1, 2, 3

Dans certains cas le volume de la base n'est pas divulgué, il faut l'estimer à partir du nombre d'apparition à chaque période d'une combinaison des mots les plus utilisés de la langue

Des sous-corpus que seul Gallicagram sait interroger : les corpus larges par titre de presse (>20 titres), les corpus de presse thématique

Les traitements effectués

Code source (open source)

Preprint : détaille pour chaque mode de recherche les données exploités et les traitements effectués

Limites de l’outil et précautions d’usage

Les limites de l’outil

- Pas d’instantanéité

Les limites dues à la base de données

- Lenteur du programme de numérisation
- Qualité d’océrisation
- Erreurs dans les métadonnées

Les limites des moteurs de recherche exploités

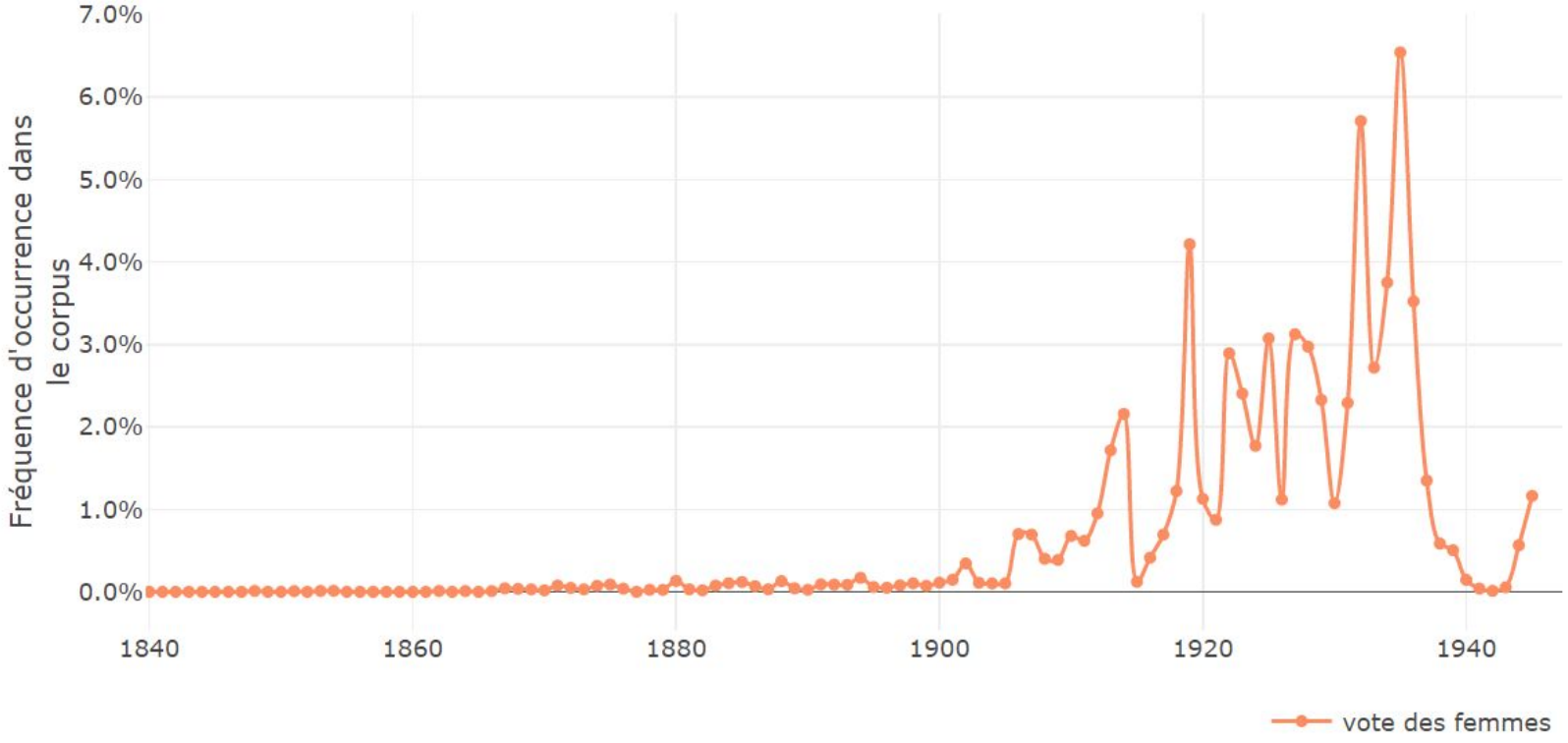
- La recherche par page appuyée sur l’API document de Gallica renvoie des résultats erronés
- Lenteur d’accès à certains sites (BNE) et manque d’API (BAnQ, KBR)
- Accès limité aux API (LOC)

Les limites interprétatives

- Auteurs ≠ lecteurs
- Pas d’information sur le tirage des documents
- Culture imprimée < culture

La diversité des usages possibles : illustration

Mesurer la force d’une idée : le vote des femmes



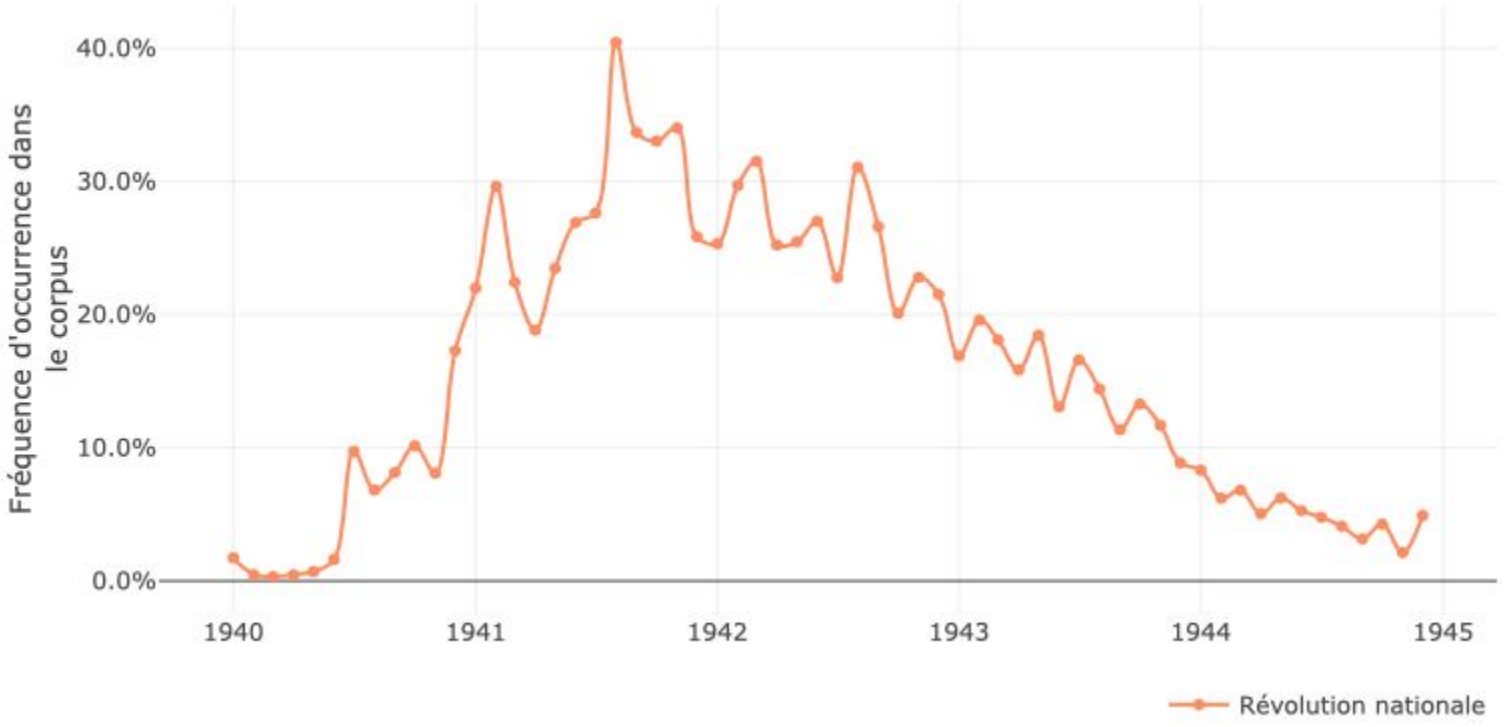
Première occurrence de “vote des femmes”

Lors de cette élection, il s’est présenté un fait étrange. Un a vu une femme se présenter pour voter. Un adjoint a protesté sur-le-champ. Pourquoi ne pas tenir compte du vote des femmes ? Il serait peu galant de les oublier. [...]

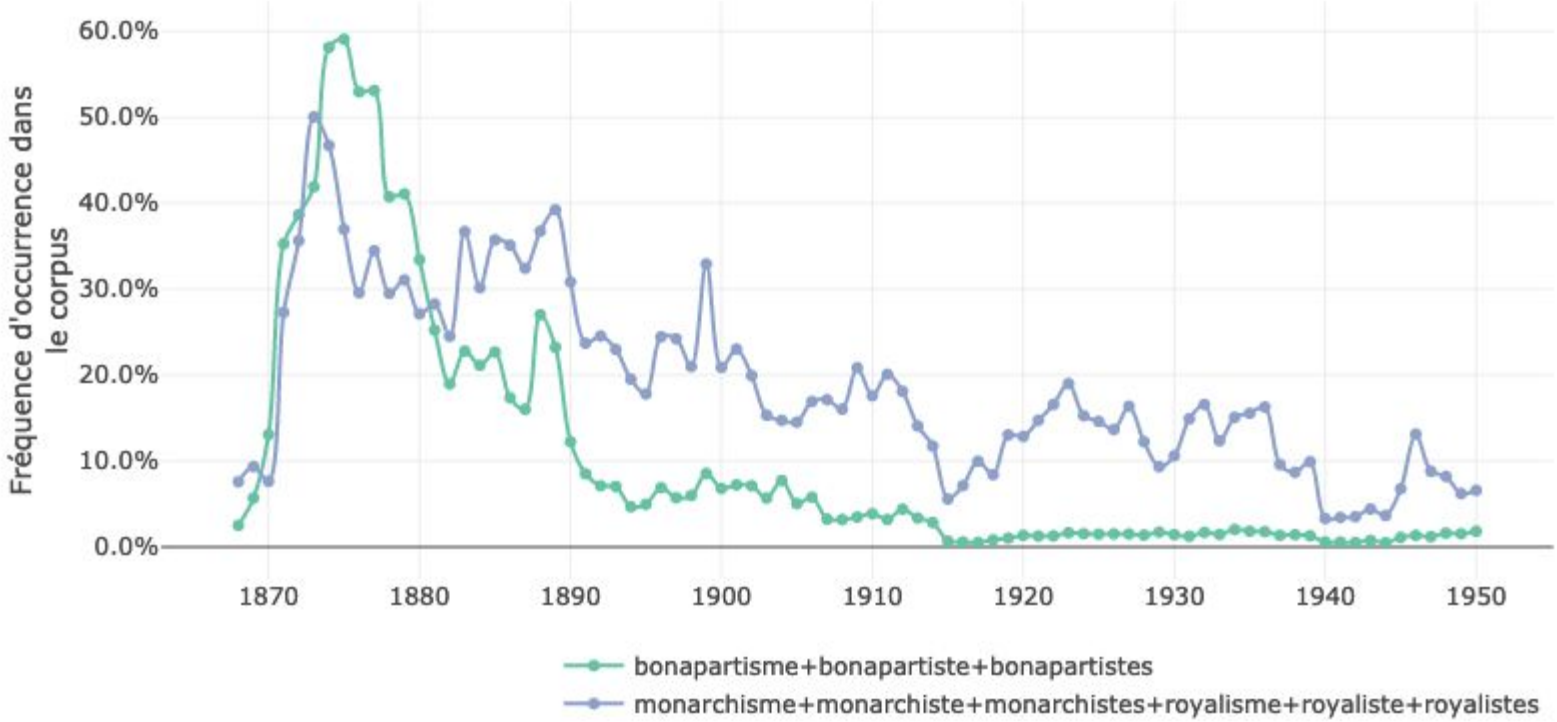
La conclusion serait qu’il faut convier les femmes à la vie politique : mais en ceci cependant ne nous hâtons pas trop ; les tribunes, les journaux, la paix, la guerre, le travail, tout cela regarde l'homme et l’homme seul. Dans l’ancienne Grèce, les Athéniennes ne s’occupaient que d’amour. A Rome, les femmes donnaient tout leur temps aux parfums, à la promenade et à la musique. Il ne faut pas que le Forum tue le boudoir.

-- *Journal de la Seine et Marne,* 17 juin 1848, p.3

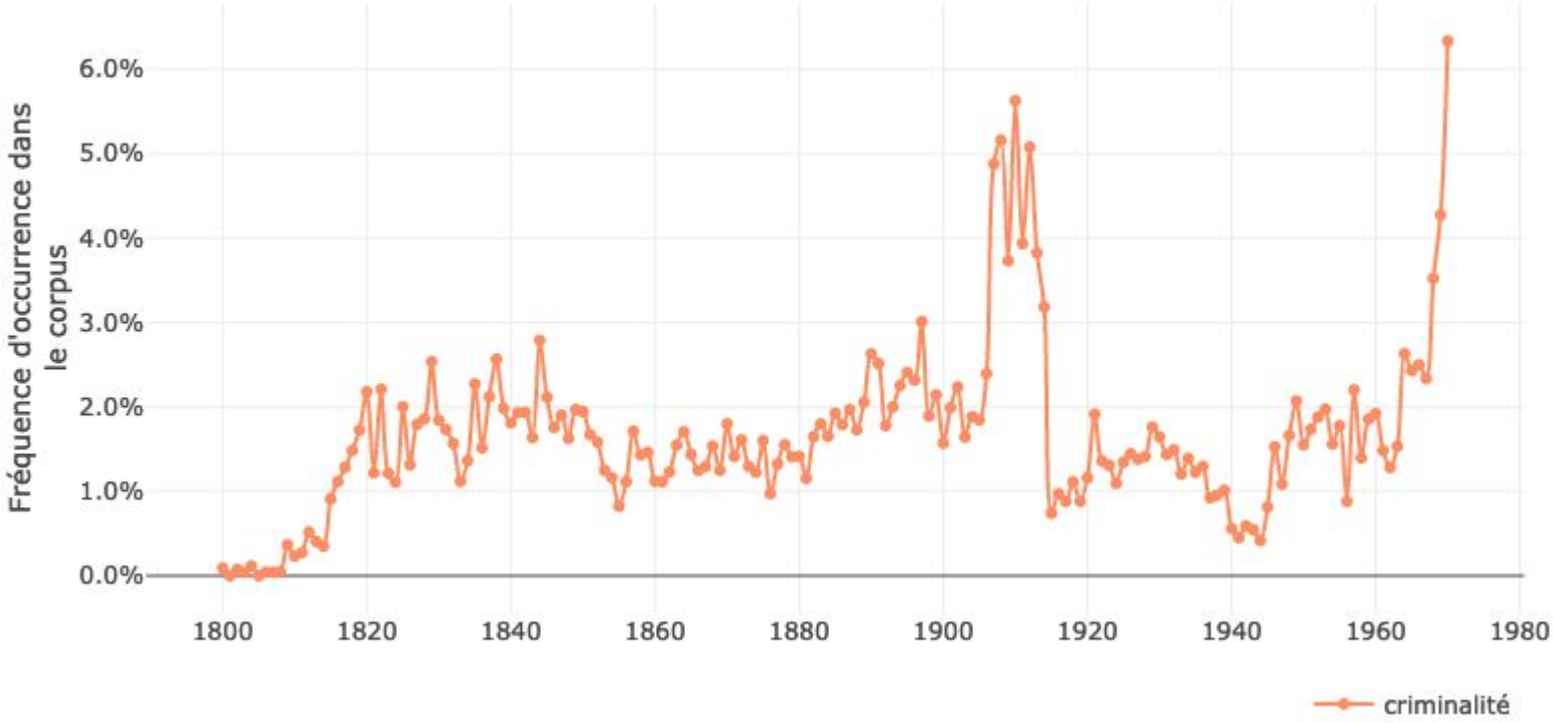
La Révolution nationale s’enraye



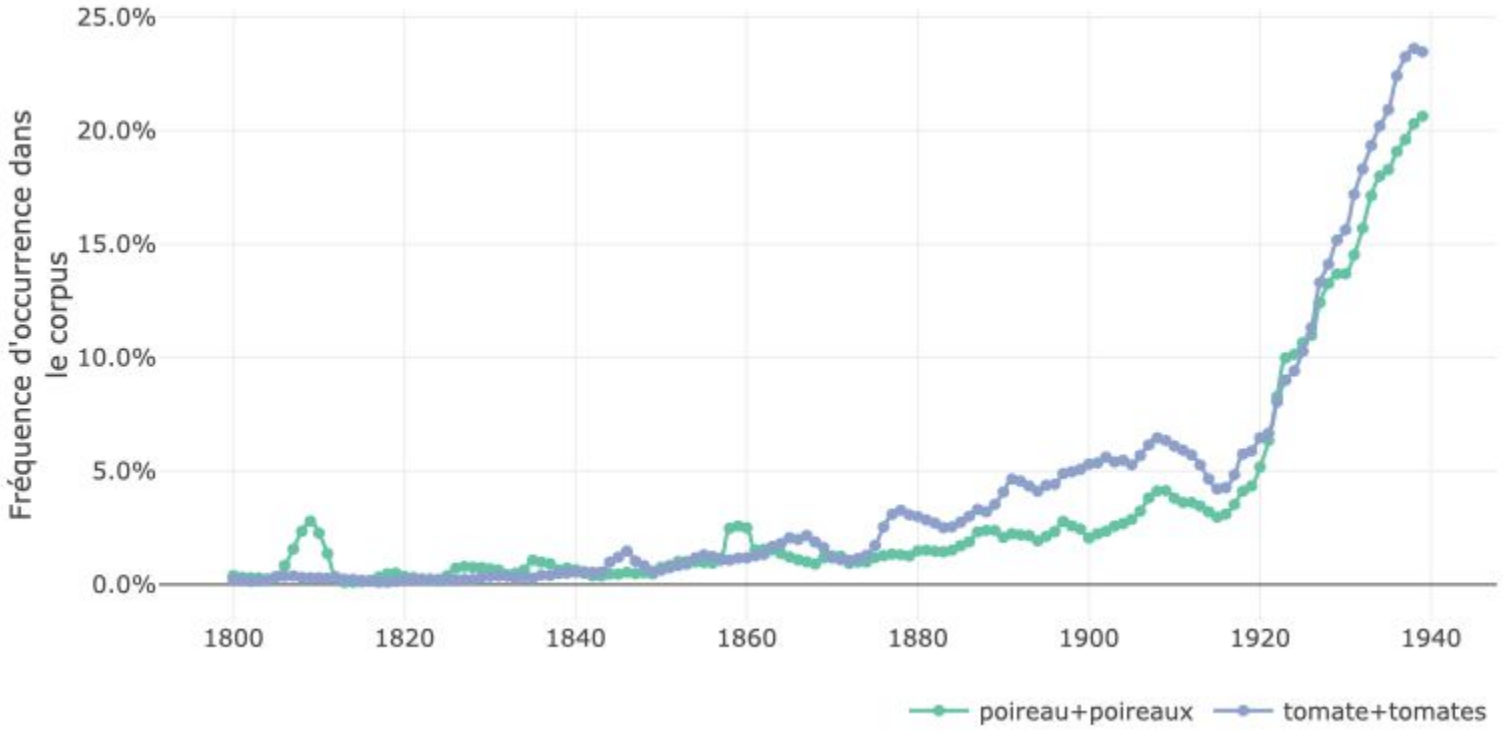
Mesurer la force d’une idée au cours du temps



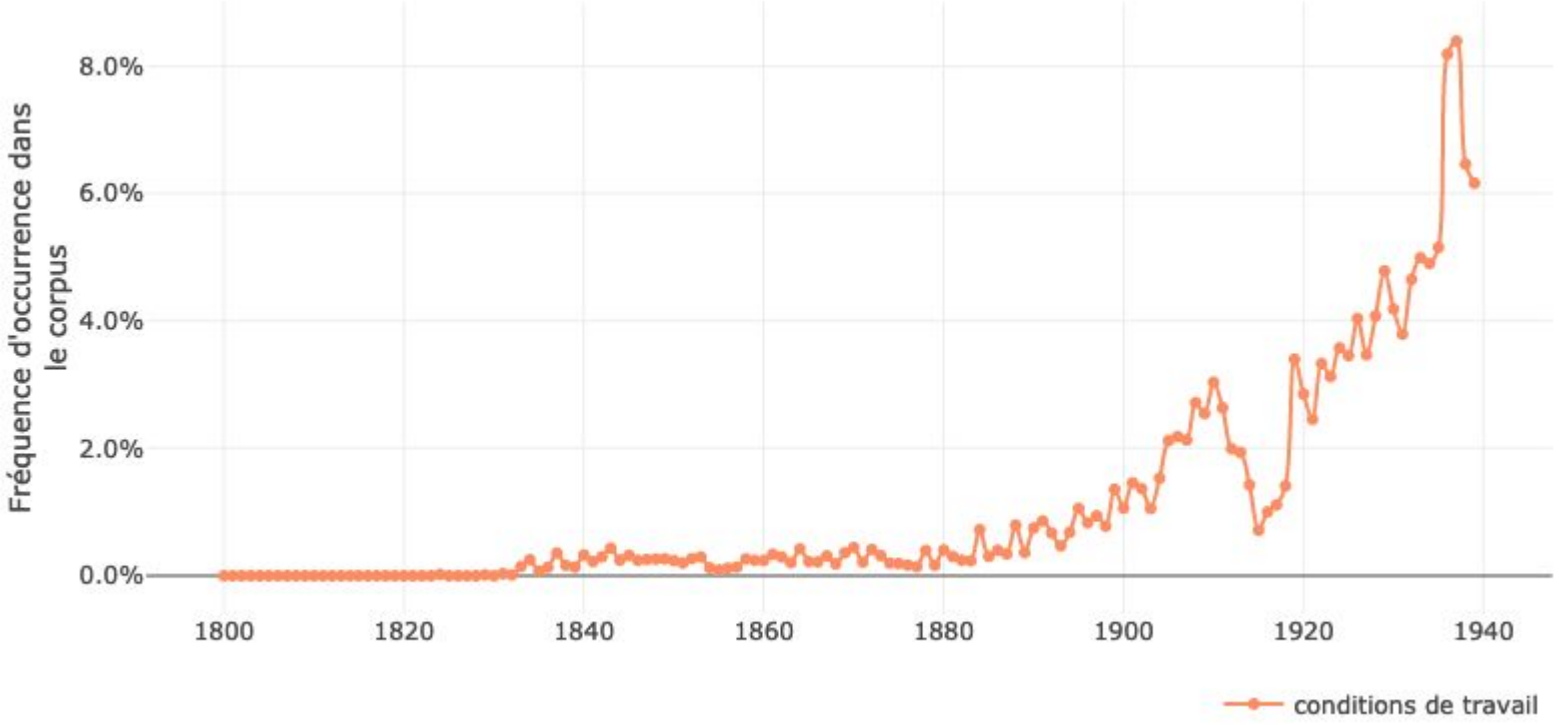
Mesurer l’insécurité ressentie



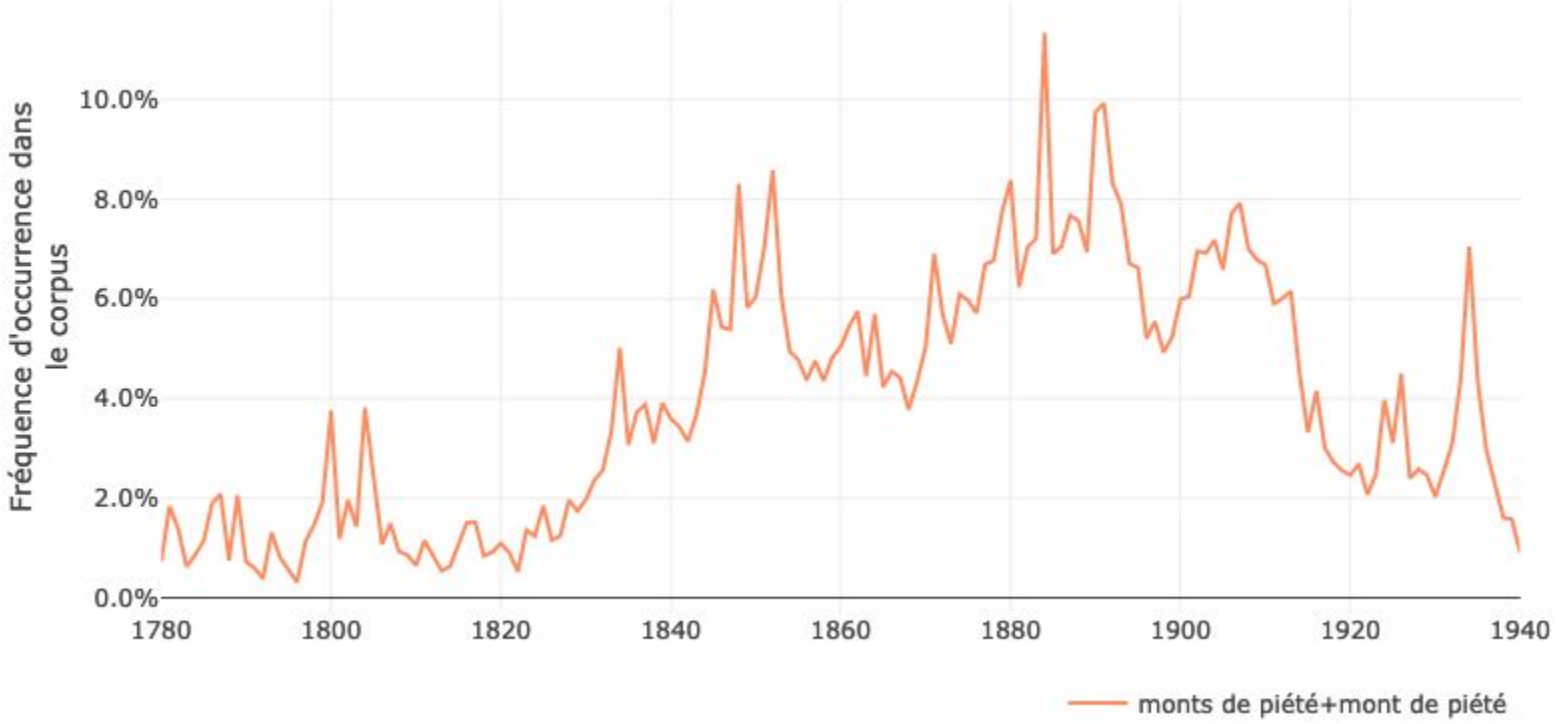
Mesurer l’importance d’un thème dans la presse



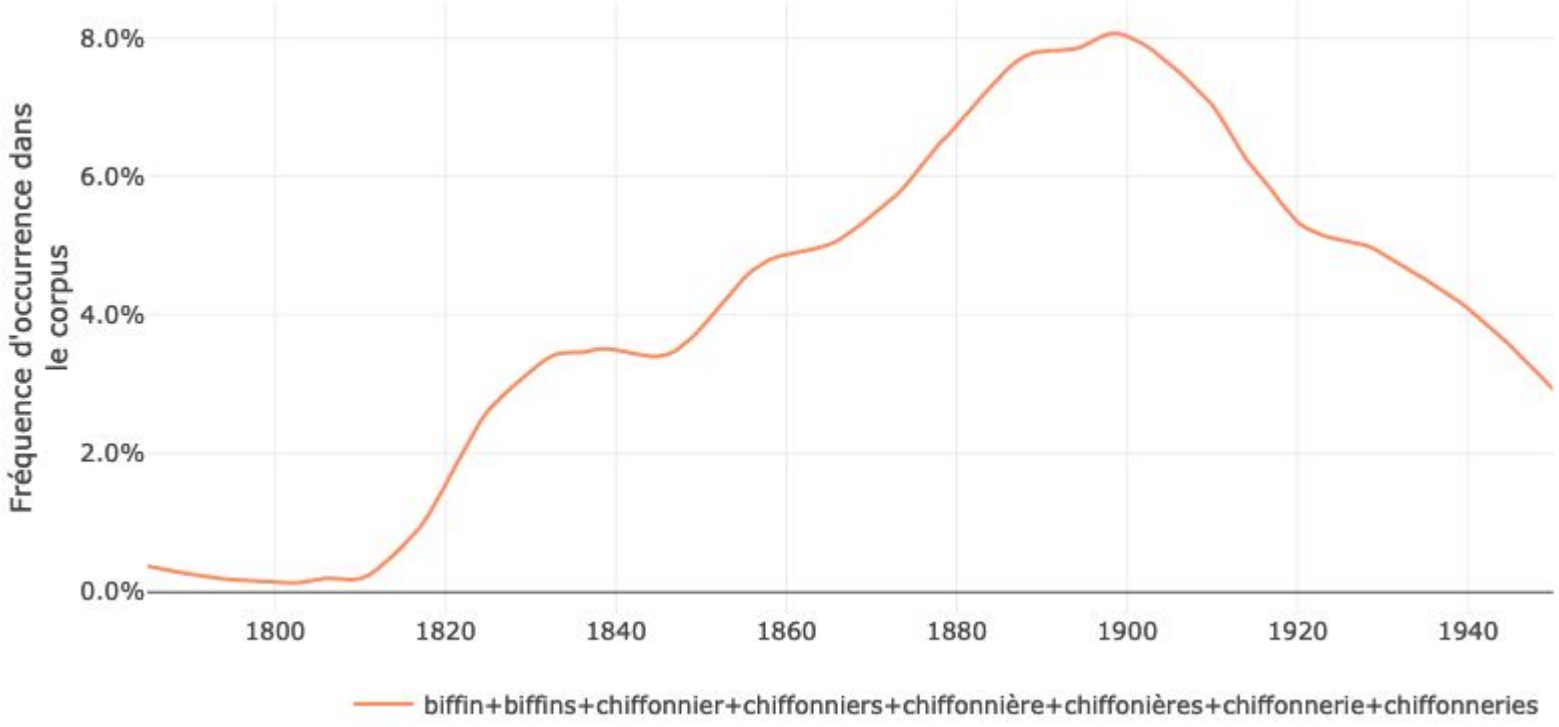
Mesurer une préoccupation



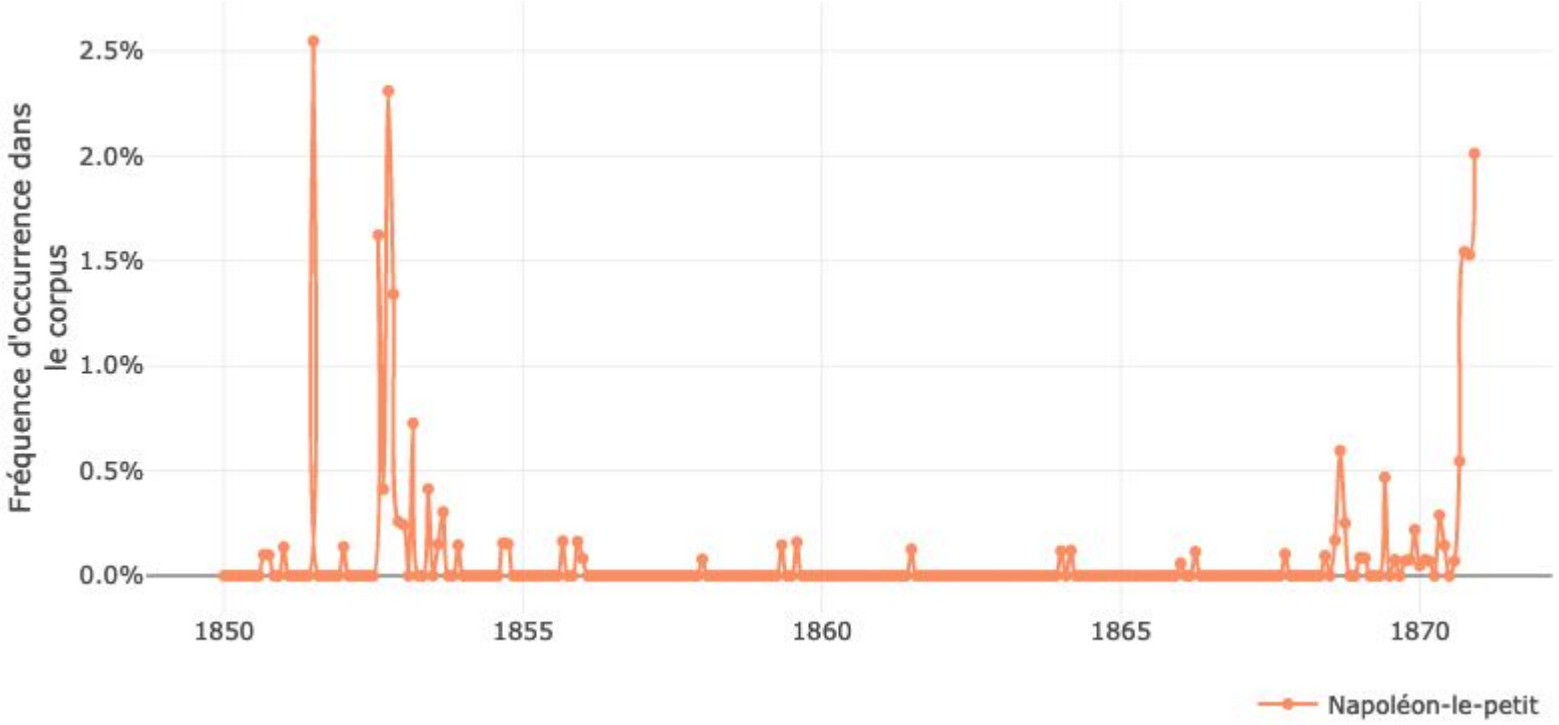
Dater un phénomène



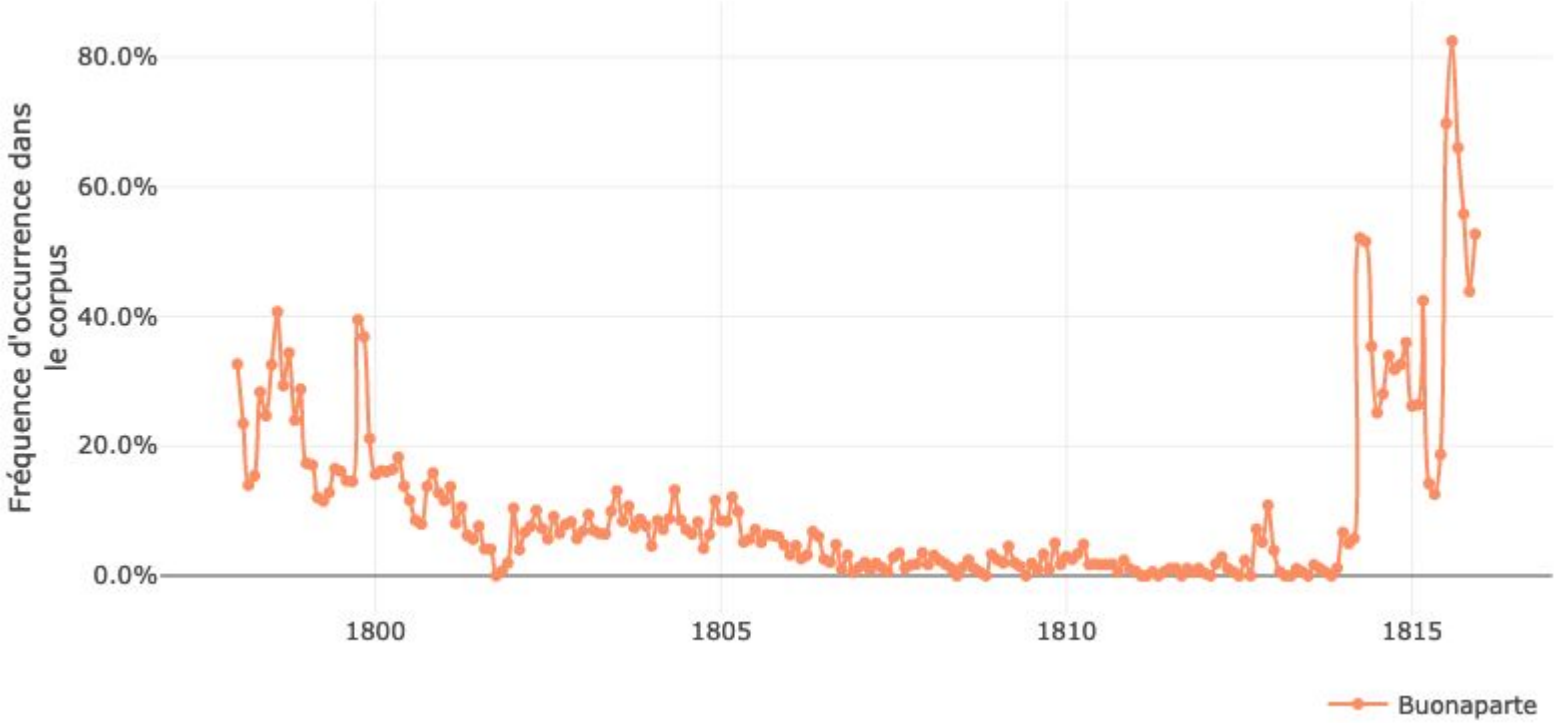
Dater le déclin des chiffonniers



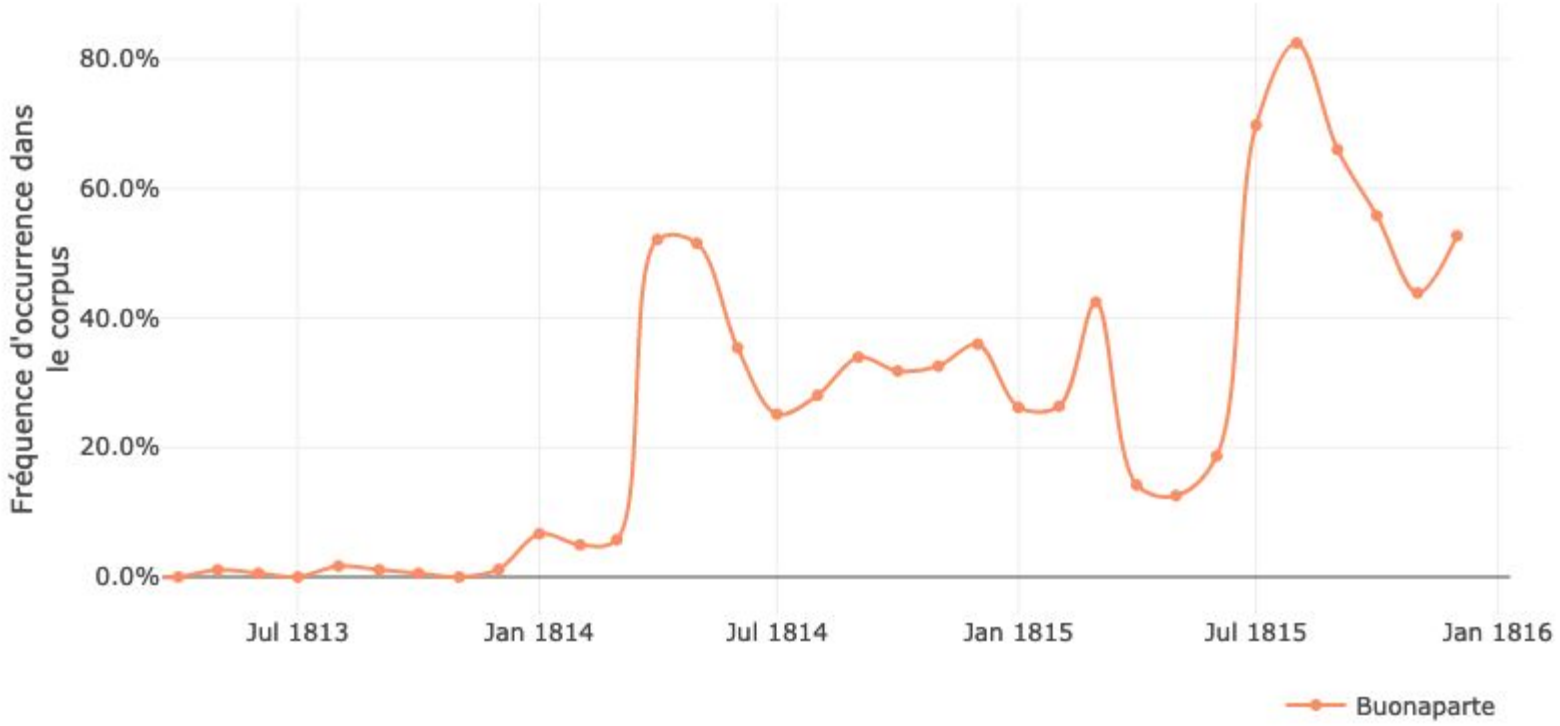
Mesurer la censure



Mesurer la censure



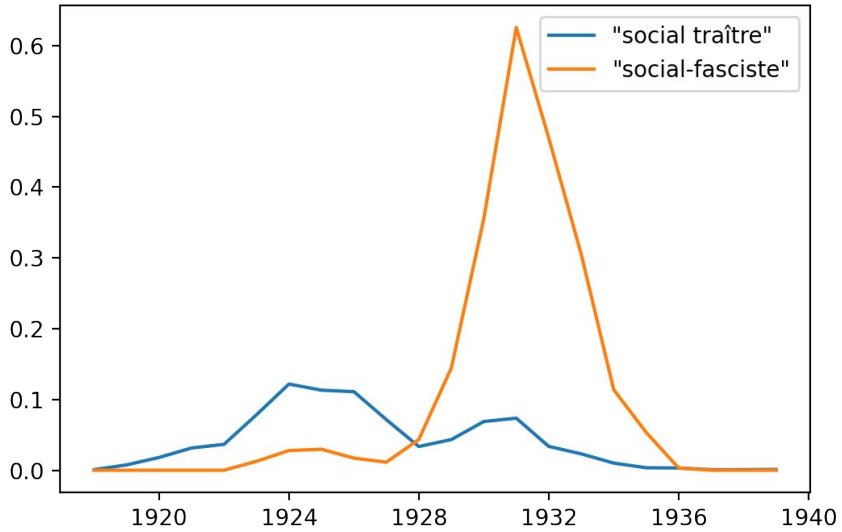
Mesurer la censure



Etude d’une source particulière

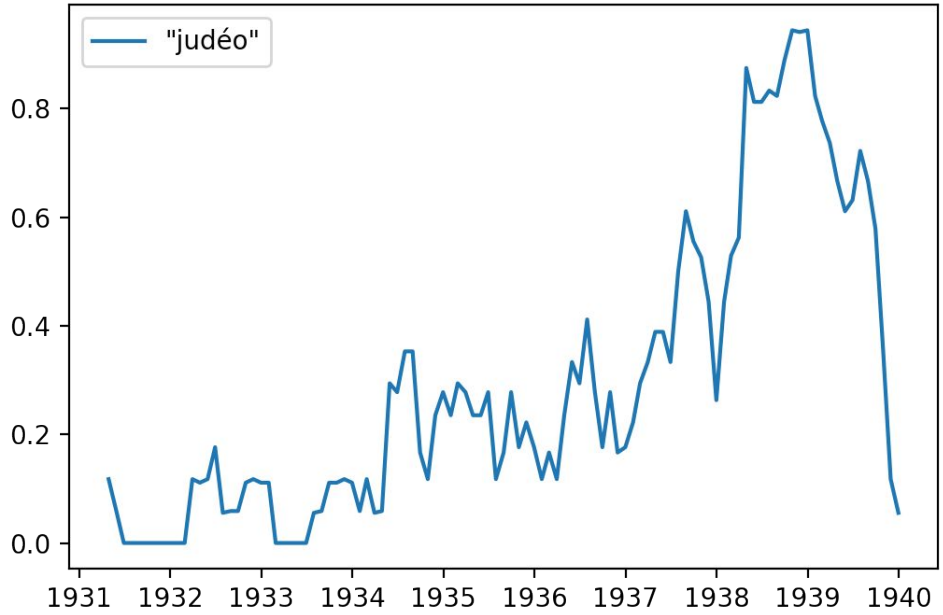
Occurences dans L'Humanité (moyenne glissante 3 ans)

Etude d’une source particulière



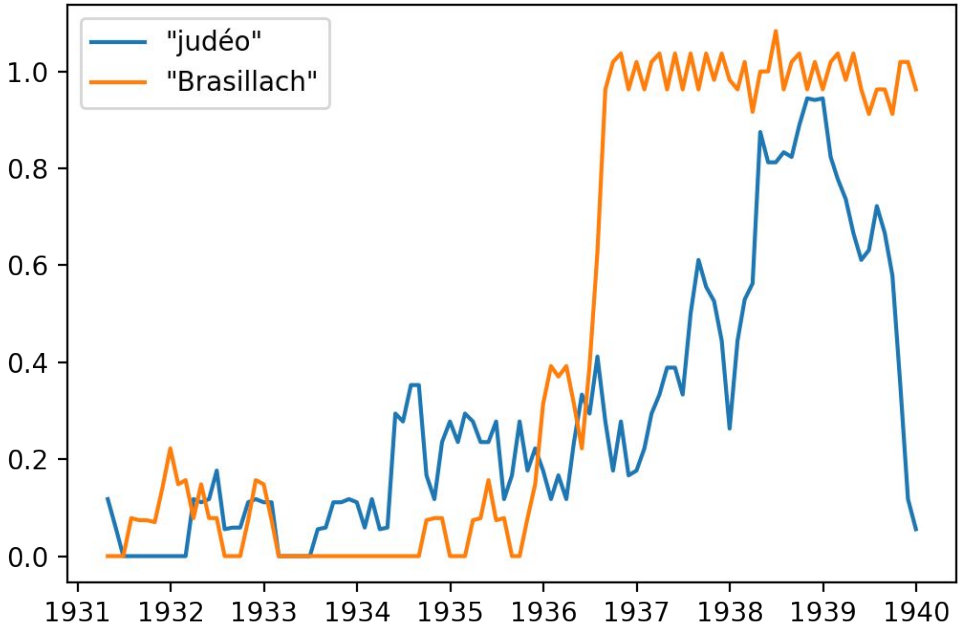
Occurences de "social-traître" dans L'Humanité (moyenne glissante 3 ans)

Etude d’une source particulière



Occurences dans *Je suis partout*

Etude d’une source particulière



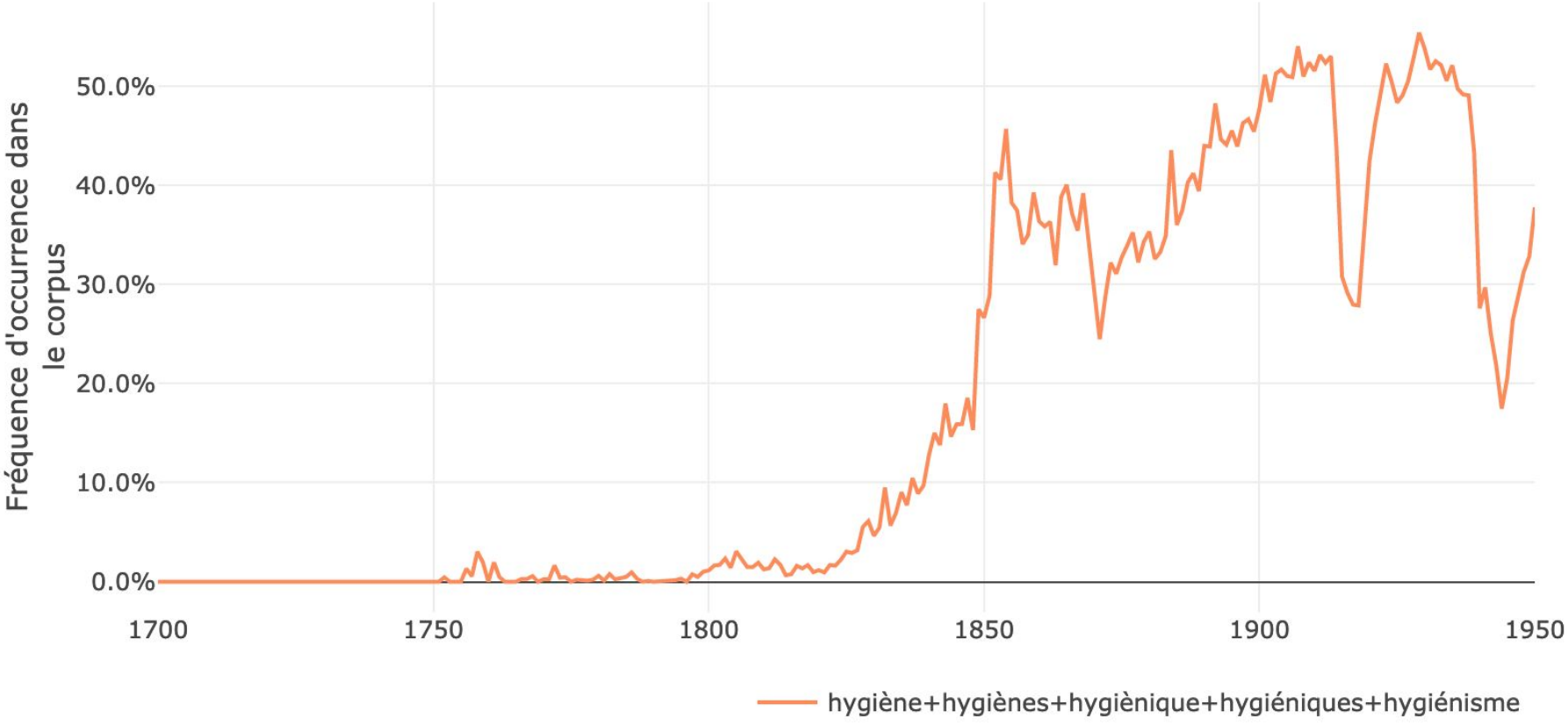
Occurences dans *Je suis partout*

Un chantier toujours en cours

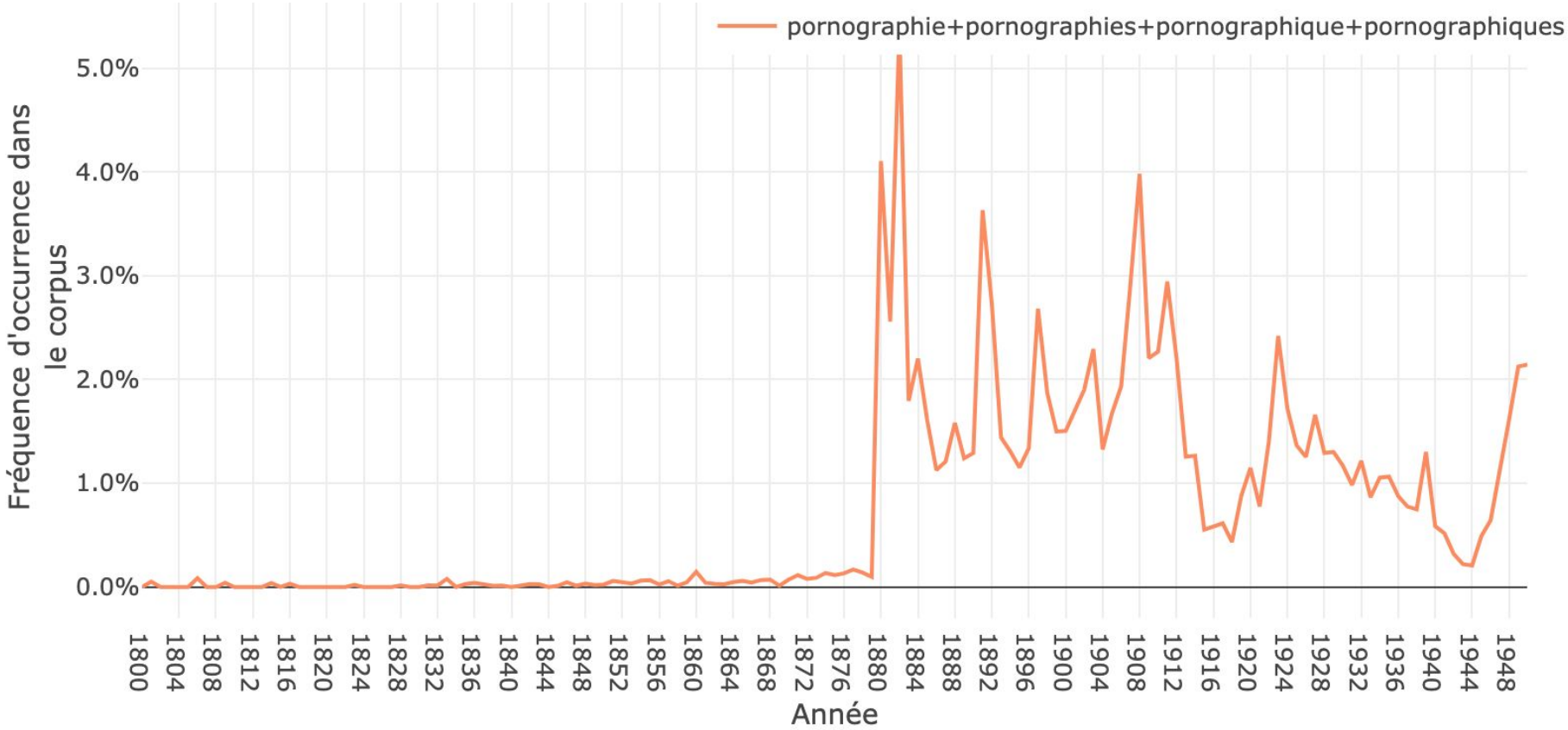
Les nouveautés à venir

- Un preprint dans les semaines à venir
- Toujours plus de bibliothèques (renforcer le corpus anglophone : Australie, N-Z)
- De nouveaux graphes décrivant les corpus
- La recherche par n-gramme dans les corpus de presse et de livres de Gallica
- Explorer les cooccurrences grâce à la recherche par proximité
- Gallicanet, un outil pour cartographier des réseaux
- Modéliser les phénomènes d'émergence et détecter les chocs exogènes

Emballement endogène



Choc exogène



Remerciements

- Un grand merci à la Bibliothèque nationale de France, à Gallica et à ses équipes. Tout particulièrement à Messieurs Jean-Philippe Moreux et Arnaud Laborderie qui nous soutiennent dans notre projet.
- Merci à l'Ecole Normale Supérieure de Paris-Saclay qui a mis ses infrastructures informatique à notre disposition et particulièrement à Monsieur Pascal Soullard qui nous a épaulés pour la mise en ligne du site internet.
- Merci à l'Ecole Normale Supérieure qui nous a fourni la puissance de calcul nécessaire à l'extraction et au traitement de vastes bases de données.

BnF

école normale supérieure paris—saclay

ENS | PSL