Gallicagram: un outil de lexicométrie pour la recherche

Benoît de Courson (Max Planck Institute for the Study of Crime), b.decourson@csl.mpg.de
&
Benjamin Azoulay (ENS Paris-Saclay), benjamin.azoulay@ens-paris-saclay.fr

MAX PLANCK INSTITUTE
FOR THE STUDY OF
CRIME, SECURITY AND LAW

ENS | PSL

Introduction

Les textes sont des “**fossiles** des mentalités”

Un large volume de textes se prête à un traitement **quantitatif** et **automatisé**

Le plus naturel, c’est de mesurer la **fréquence d’un mot/syntagme**

…et si possible son évolution au cours du **temps**

C’est ce que fait *Gallicagram*, en particulier sur le **corpus de Presse de *Gallica***


L'AURORE

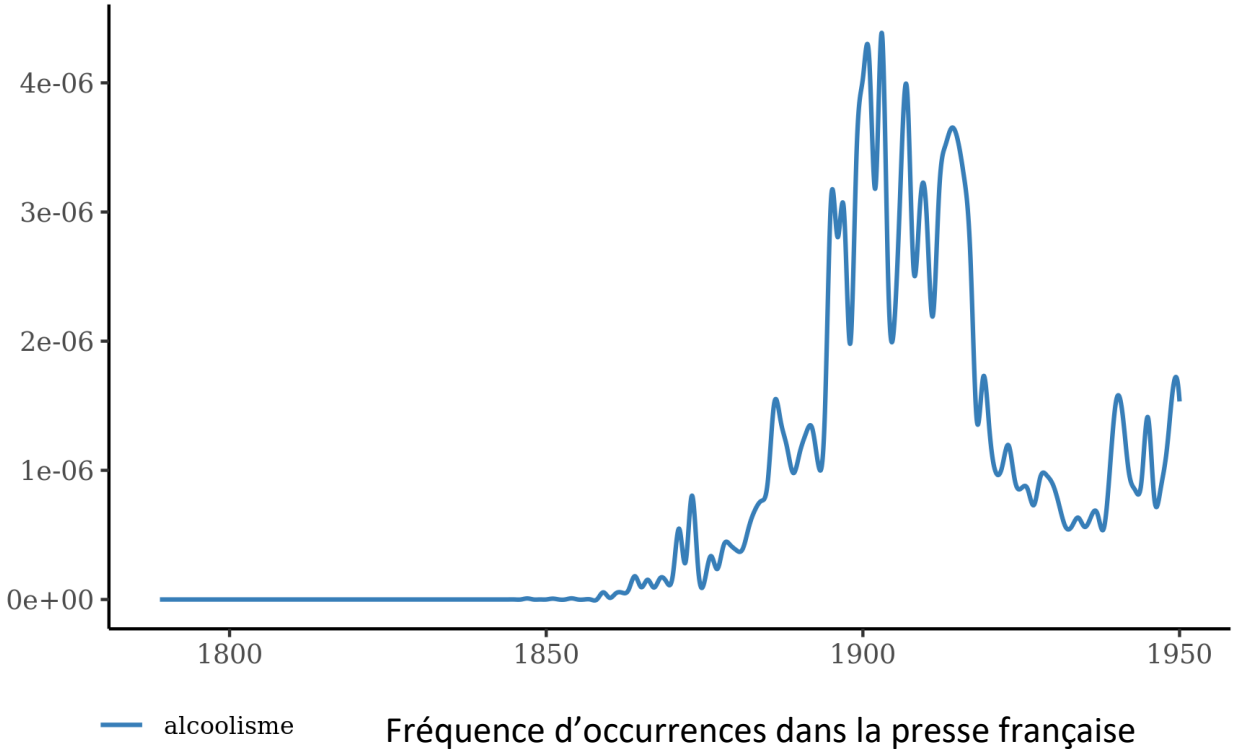
Littéraire, Artistique, Sociale

J'Accuse...!

LETTRE AU PRÉSIDENT DE LA RÉPUBLIQUE

Par ÉMILE ZOLA

Gallicagram en deux mots



Fréquence d'occurrences dans la presse française

L’évolution du projet

Initialement, un simple **compteur de résultats** sur *Gallica.*

Le moteur de **“recherche avancée”** permet de cibler le corpus de presse, précieux pour l’historien

Problème : une mesure **lente** et **grossière** (“Par document”)

“Moissonnage” des documents de Gallica en texte brut :

- 300 000 monographies (~livres)
- 3 millions de périodiques (~numéros de presse)

Sur ces deux corpus, extraction de la fréquence des mots et syntagmes

Idem sur les archives du ***Monde*** (3 millions d’articles)

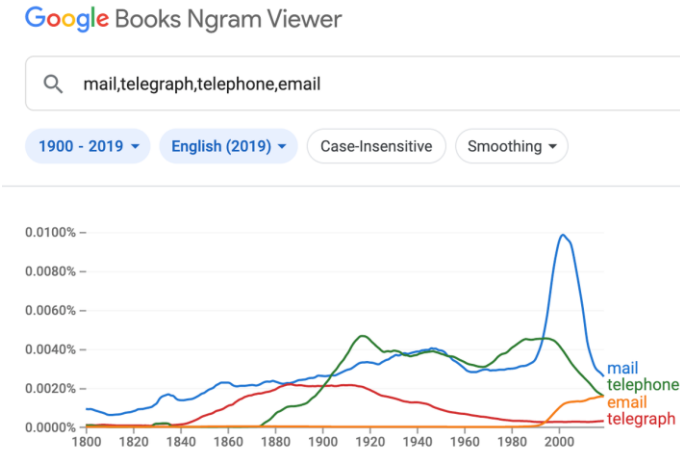
Ngram Viewer et les chercheurs, un « rendez-vous manqué » ?

En 2010, Google a lancé triomphalement Ngram Viewer (**4% des livres imprimés**)

Les chercheurs en humanités l'ont fort peu adopté

En histoire contemporaine:

- Une seule mention dans *XXe Siècle, Revue d'Histoire*
- Deux dans les *Annales*, sans pour autant l'utiliser
- Aucune dans la *Histoire & Mesure*



L’étude la plus citée sur *Ngram Viewer*

Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
24(9) 1722–1731

DOI: 10.1177/0956797613479387
pss.sagepub.com
SAGE

The Changing Psychology of Culture From 1800 Through 2000

Patricia M. Greenfield
Department of Psychology, University of California, Los Angeles

Abstract

The Google Books Ngram Viewer allows researchers to quantify culture across centuries by searching millions of books. This tool was used to test theory-based predictions about implications of an urbanizing population for the psychology of culture. Adaptation to rural environments prioritizes social obligation and duty, giving to other people, social belonging, religion in everyday life, authority relations, and physical activity. Adaptation to urban environments requires more individualistic and materialistic values; such adaptation prioritizes choice, personal possessions, and child-centered socialization in order to foster the development of psychological mindedness and the unique self. The Google Ngram Viewer generated relative frequencies of words indexing these values from the years 1800 to 2000 in American English books. As urban populations increased and rural populations declined, word frequencies moved in the predicted directions. Books published in the United Kingdom replicated this pattern. The analysis established long-term relationships between ecological change and cultural change, as predicted by the theory of social change and human development (Greenfield, 2009).

Ngram Viewer : les raisons de la colère

Impossible de restreindre le corpus

Opacité du corpus

Impossible de savoir dans quel **contexte** un mot a été utilisé :

- Qui l'a employé ? Dans quel ouvrage ?
- Dans quel sens ?
- Dans quel co-texte ?

Beaucoup **d'artefacts**

…et impossible de vérifier son analyse

Pas **d'API**, pas d'accès au **volume annuel** du corpus…

Une polémique récente

Historical language records reveal a surge of cognitive distortions in recent decades

Johan Bollen, Marijn ten Thij, Fritz Breithaupt, +3, and Marten Scheffer Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved June 15, 2021 (received for review February 1, 2021)

July 23, 2021 | 118 (30) e2102061118 | https://doi.org/10.1073/pnas.2102061118

VIEW RELATED CONTENT +

153,053 6

Significance

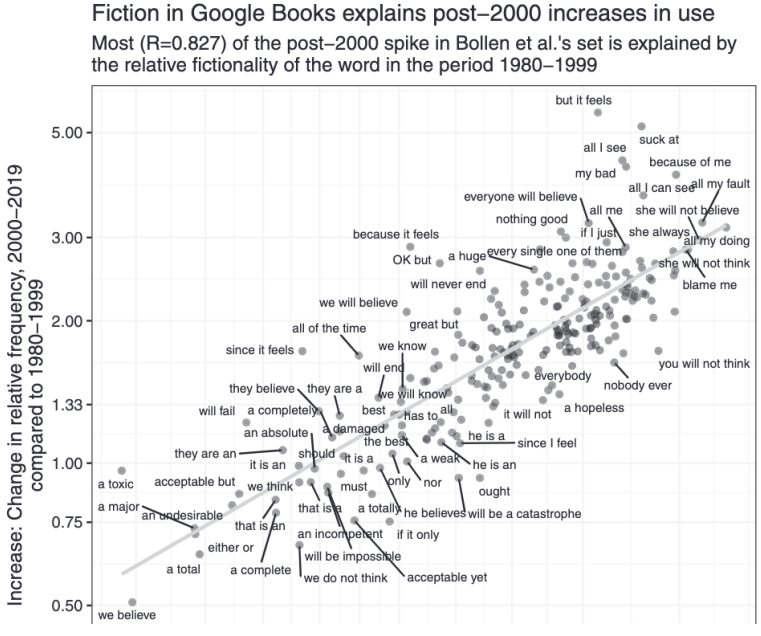
Can entire societies become more or less depressed over time? Here, we look for the historical traces of cognitive distortions, thinking patterns that are strongly associated with internalizing disorders such as depression and anxiety, in millions of books published over the course of the last two centuries in English, Spanish, and German. We find a pronounced "hockey stick" pattern: Over the past two decades the textual analogs of cognitive distortions surged well above historical levels, including those of World War I and II, after declining or stabilizing for most of the 20th century. Our results point to the possibility that recent socioeconomic changes, new technology, and social media are associated with a surge of cognitive distortions.

Uncontrolled corpus composition drives an apparent surge in cognitive distortions

Benjamin Schmidt[a,1], Steven T. Piantadosi[b], and Kyle Mahowald[c]

Bollen et al. (1) present an exciting interdisciplinary combination of clinical psychology and corpus linguistics. They find that phrases chosen by cognitive-behavioral experts to reflect negative thoughts—"cognitive distortions"—have sharply increased in frequency since the 1980s in three Google Book datasets.

Unfortunately, their work faces a foundational lim itation: For the reported patterns to be meaningful corpus frequencies must actually reflect cognitive



Est-ce la taille qui compte ?

Ngram Viewer a un corpus **9 fois plus gros** que *Gallica*

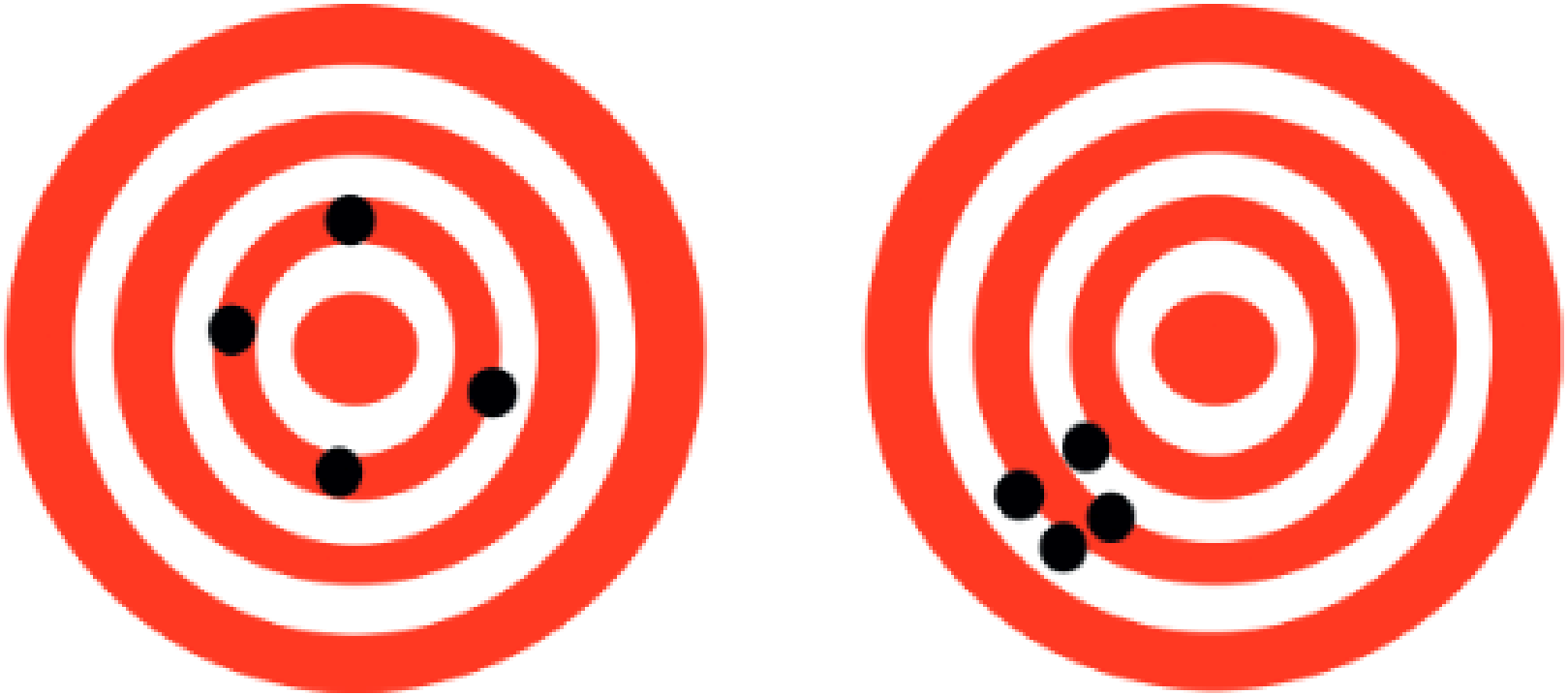
D'où une marge d'erreur en moyenne 3 fois plus faibles

MAIS du fait de la **loi des grands nombres**, ajouter des textes devient progressivement inutile (et évidemment dangereux)

La quantité **ne compense pas** la qualité et la maîtrisabilité du corpus

Meng's *big data paradox*: "once we take into account the **data quality**, the effective sample size of a "Big Data" set can be vanishingly small"

“It’s better to be roughly right than precisely wrong”
(J.-M. Keynes)



C’est l’heure du duel

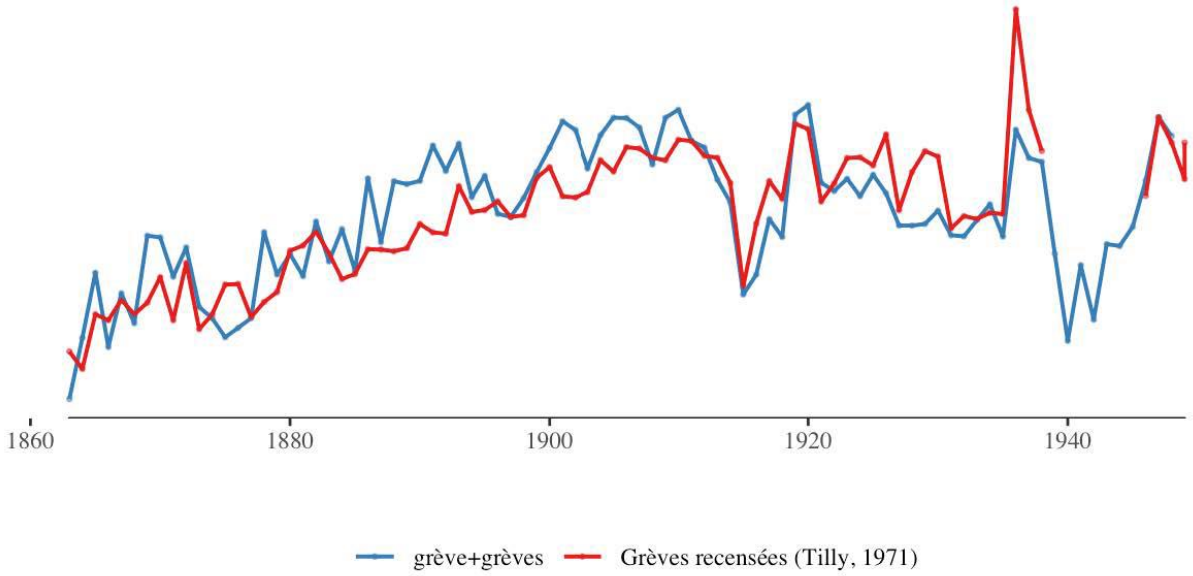
Gallica distingue la **presse** et les **livres**

A priori, un corpus de presse devrait permettre d’être plus près des **événements**

Pour éprouver la validité empirique, on peut comparer les séries temporelles de Gallicagram à des **séries statistiques** connues

Exemple : mesurons le nombre de résultats de la recherche « **grève(s)** »

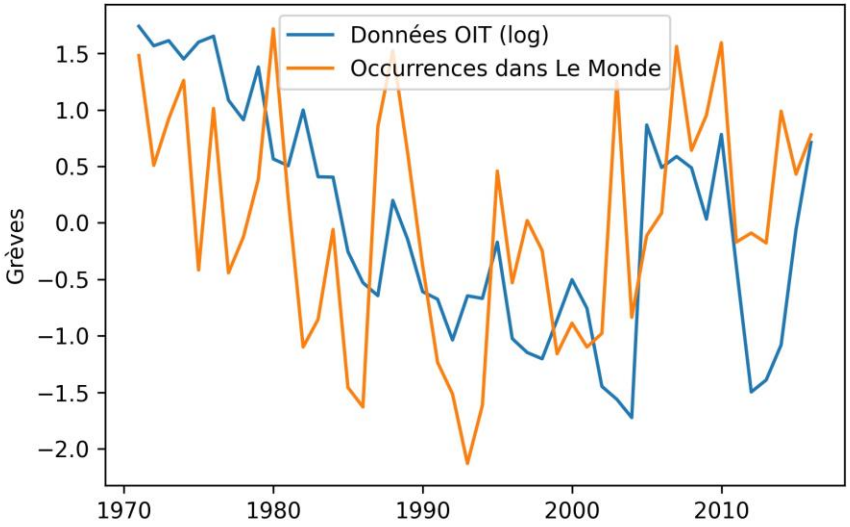
Un corpus restreint, un meilleur outil ?



Corrélation : r = .82

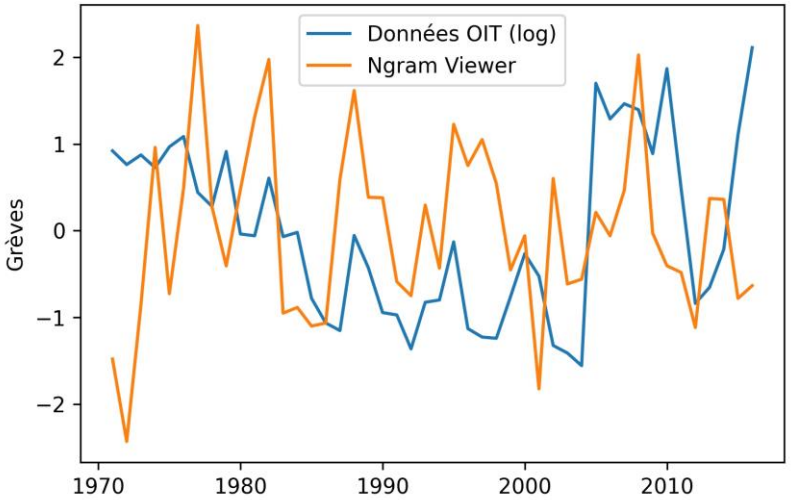
Ngram Viewer : r = .69

Le Monde (~20M mots/an) vs *Ngram Viewer* (~10Mds)

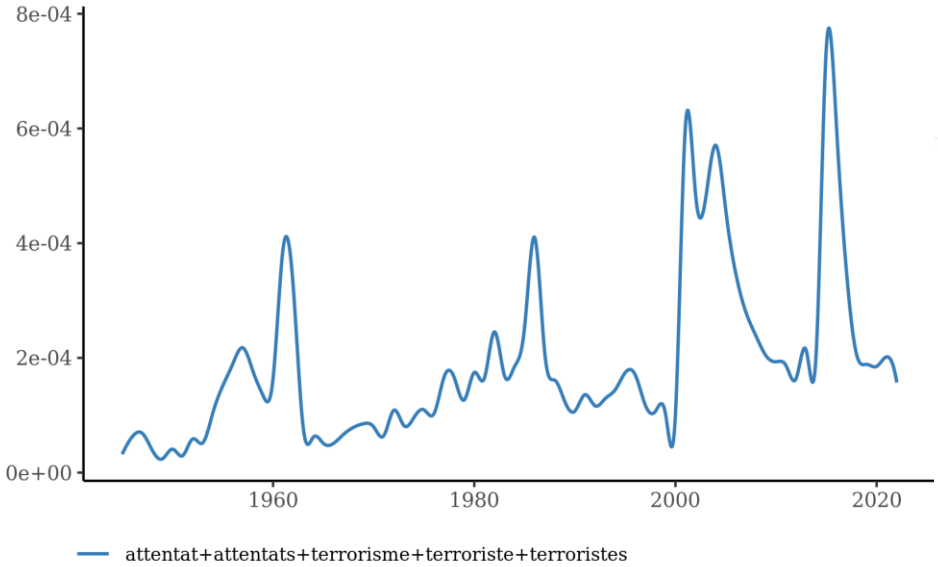


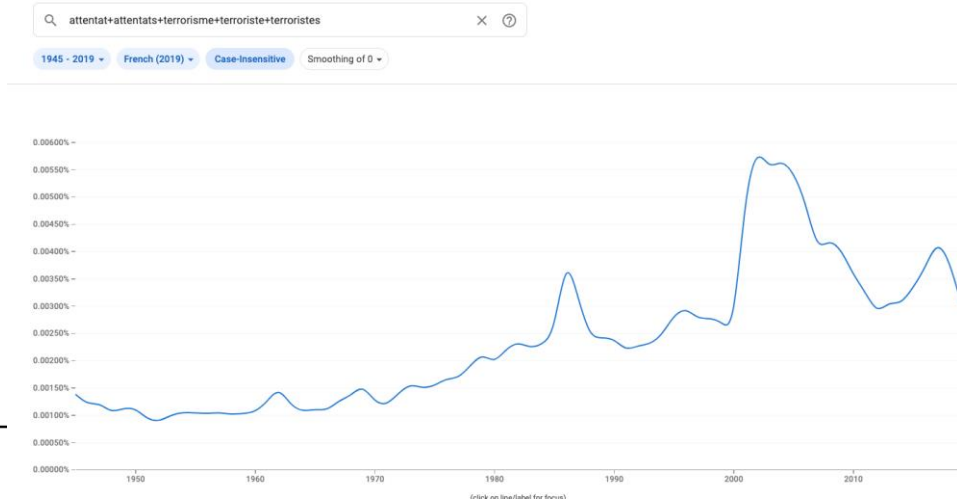
Let's **detrend**…

- *Le* Monde: r=.52
- *Ngram Viewer:* r=.02



Un corpus de presse est plus réactif



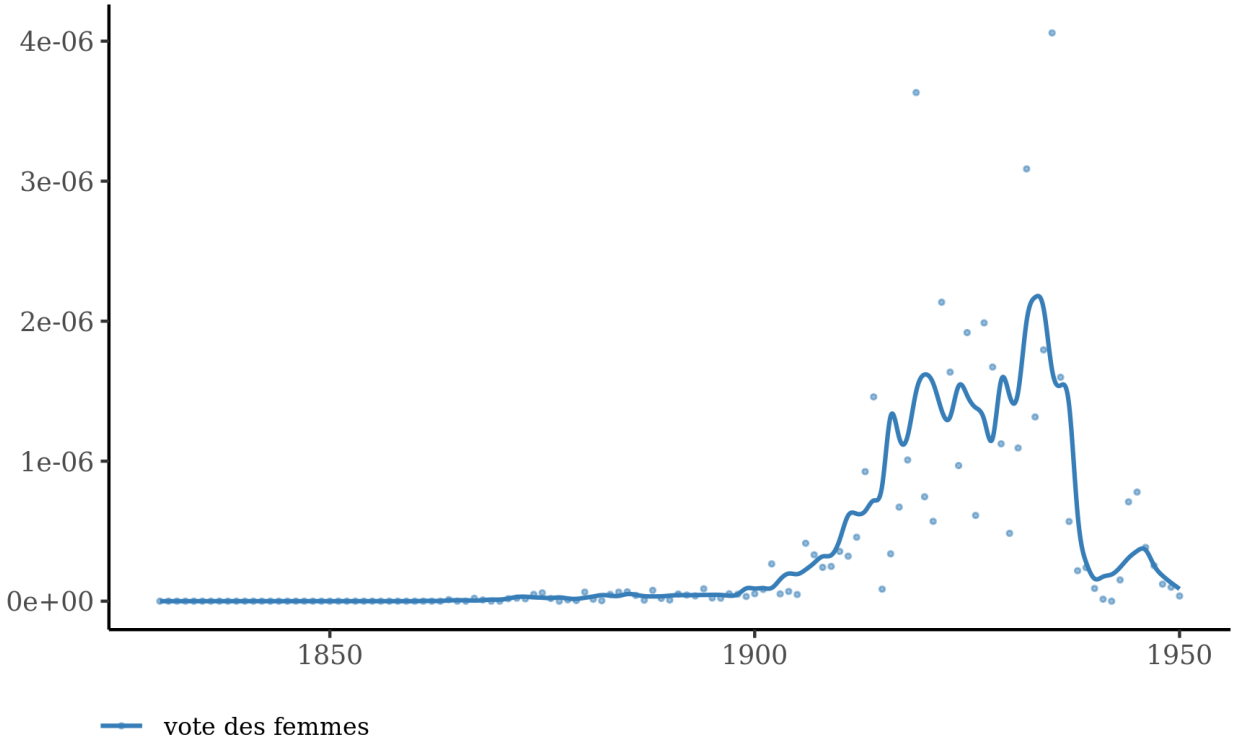


Ngram Viewer n'est pas adapté au court-terme (et n'a pas de résolution mensuelle)

Prise en main de Gallicagram

Florilège d’usages possibles

Les tendances : mesurer la force d'une idée



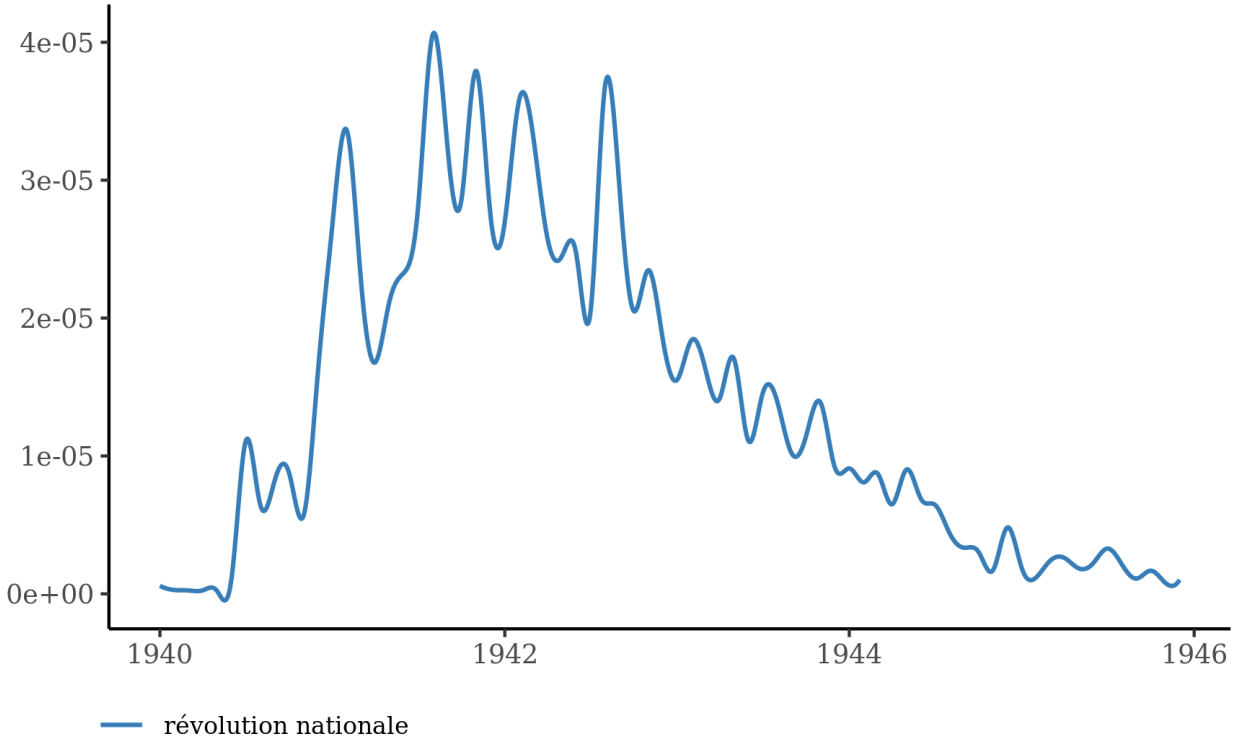
Première occurrence de « vote des femmes »

Lors de cette élection, il s'est présenté un fait étrange. On 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****. […]*

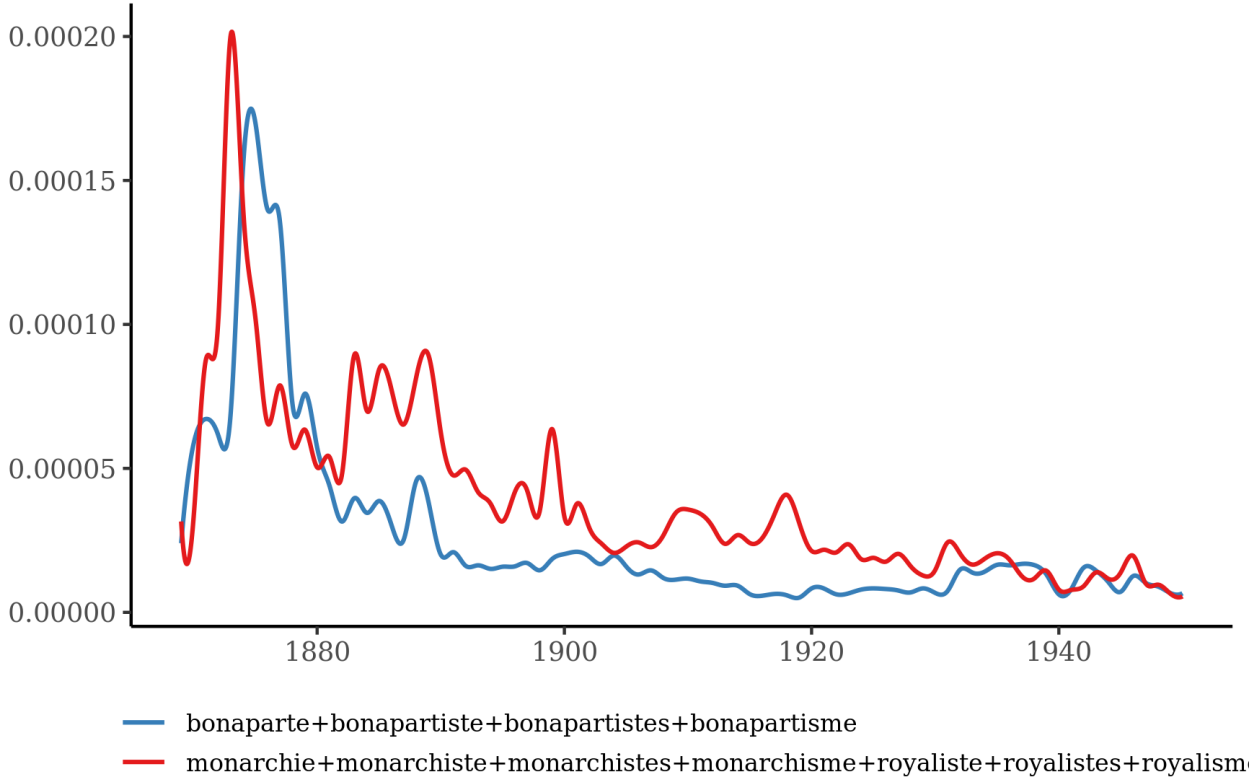
*La conclusion serait qu'****il faut convier les femmes à la vie politique : mais en ceci cependant ne nous hâtons pas trop.***

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

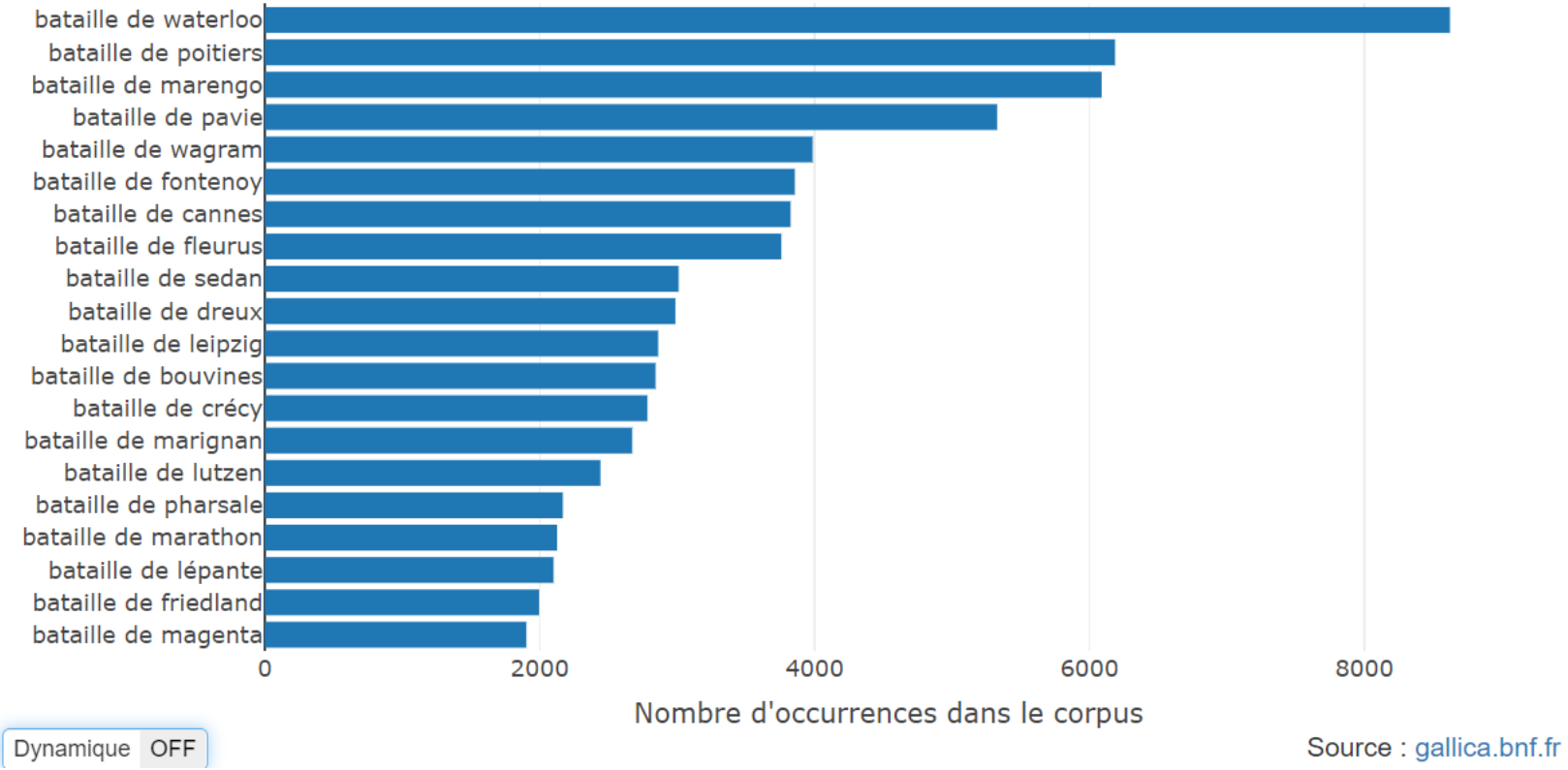
La Révolution nationale s’enraye



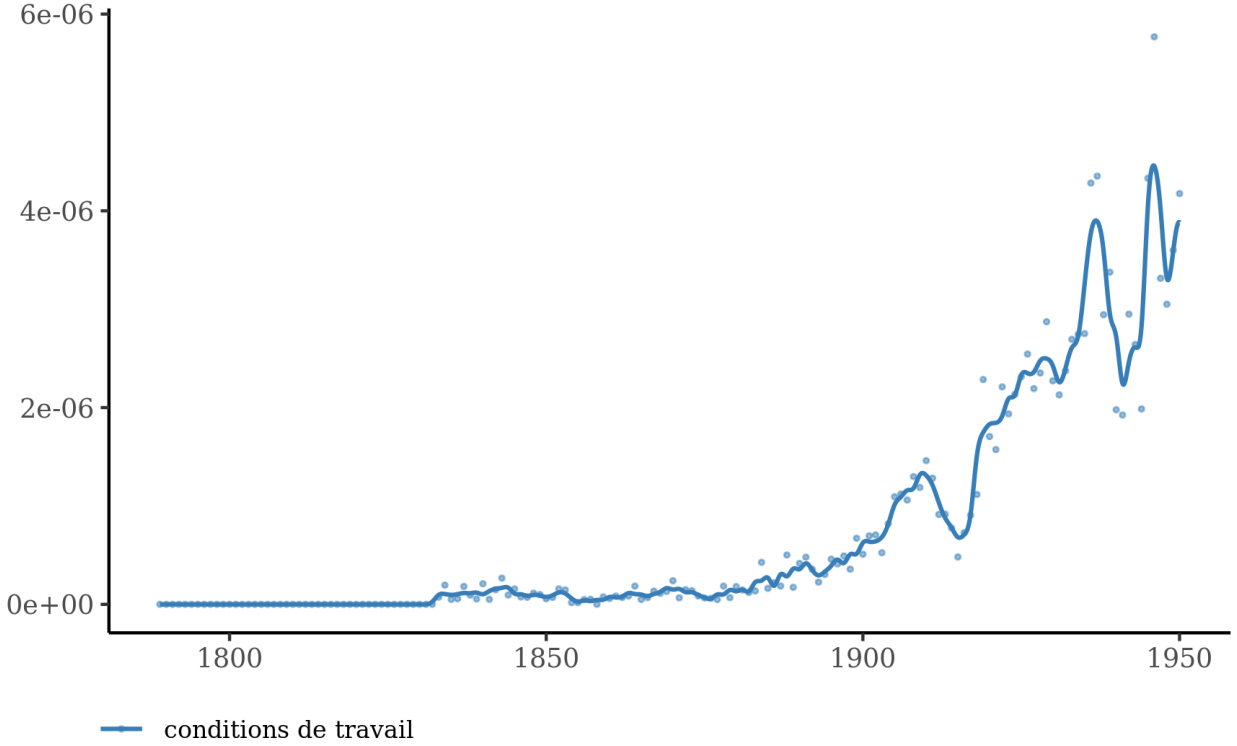
Mesurer la force d’une idée au cours du temps



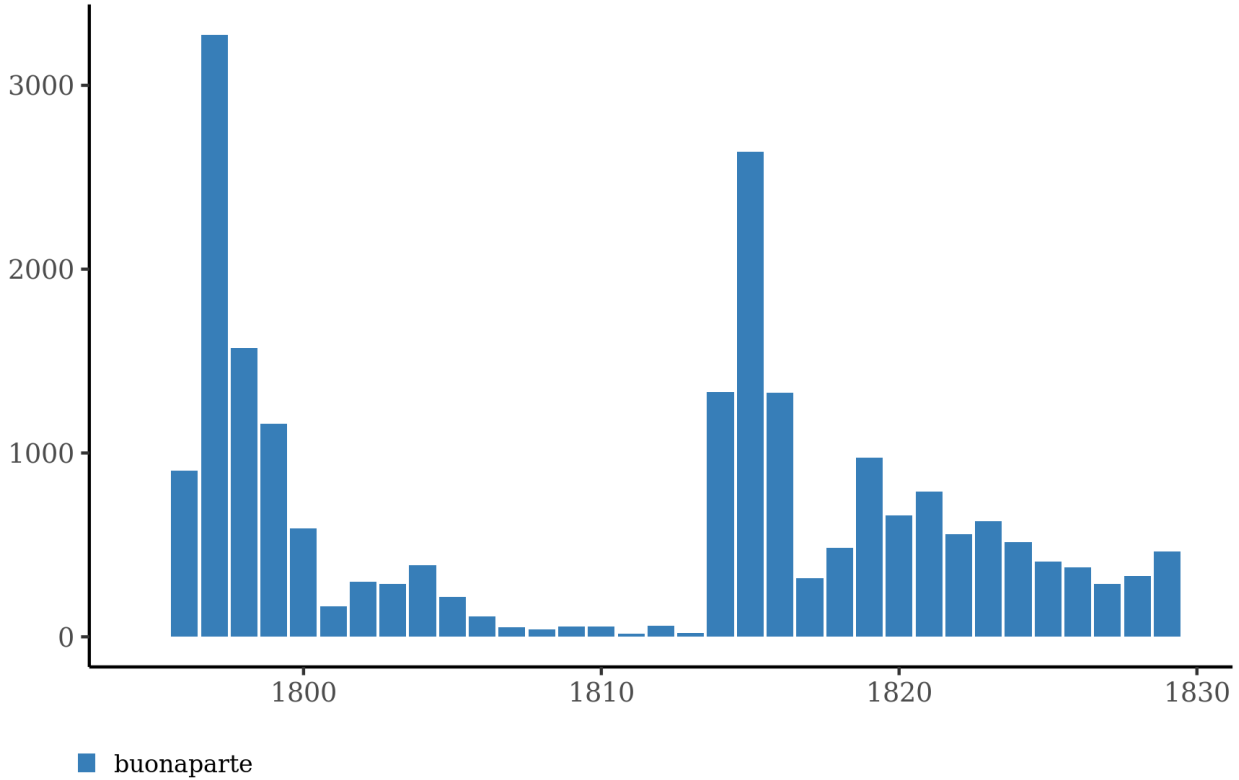
Les batailles qui sont passées à la postérité (corpus de livres)



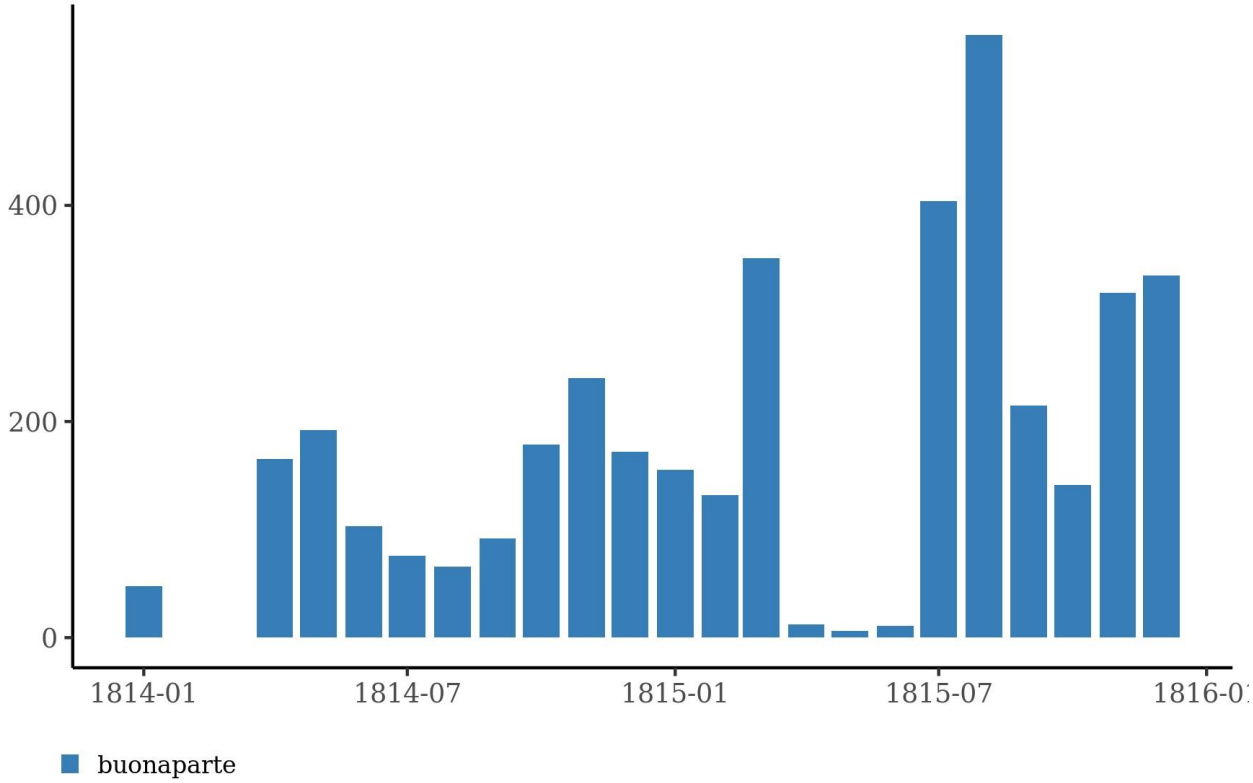
Mesurer une préoccupation



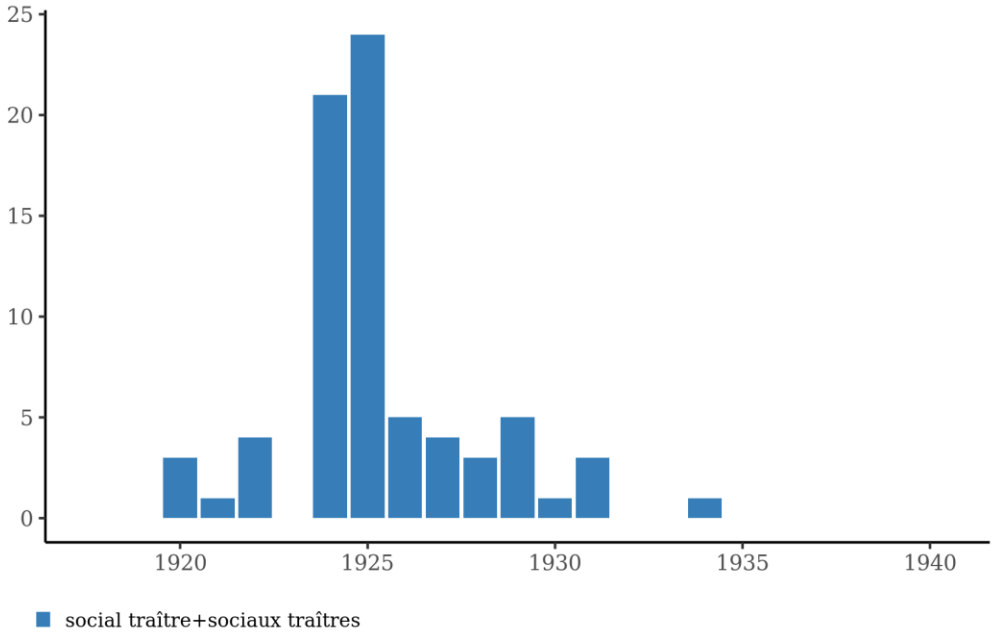
Mettre au jour la censure



Mettre au jour la censure : les Cent-Jours

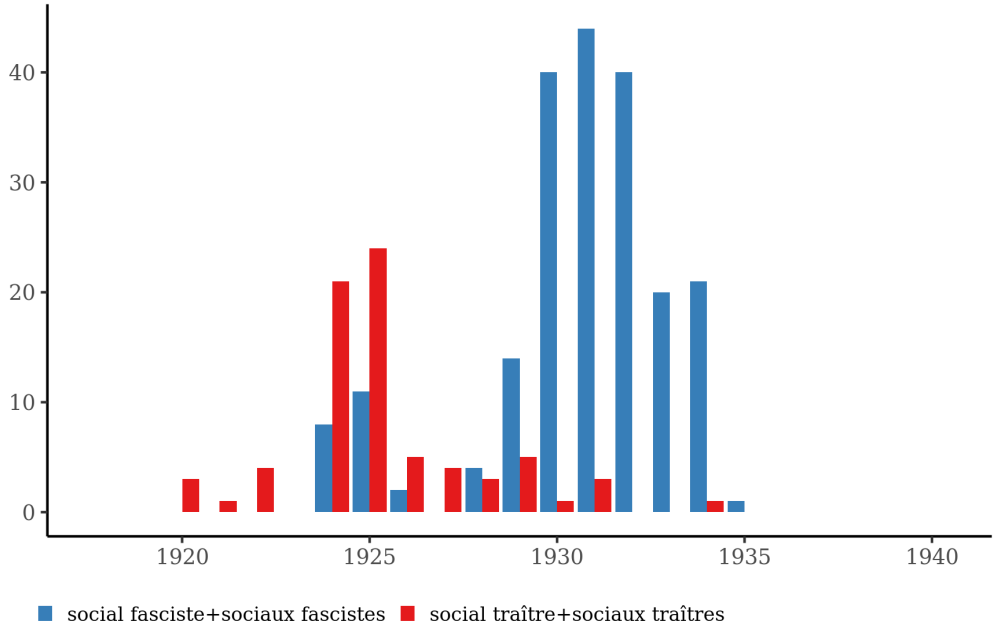


L’étude d’une source particulière



Occurrences dans *L'Humanité*

L’étude d’une source particulière



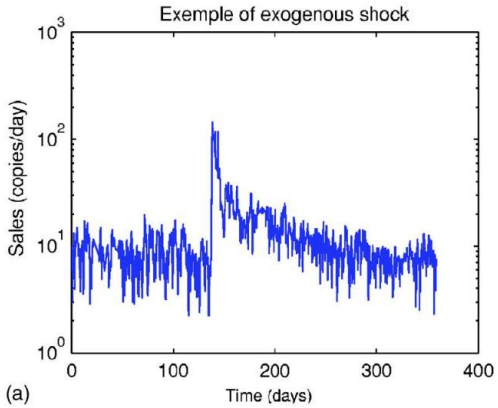
Occurrences dans *L'Humanité*

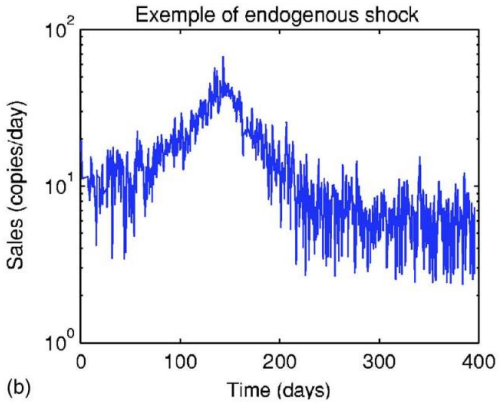
L’étude de la phénoménologie des courbes

On interprète en général les courbes en termes de pics et de creux

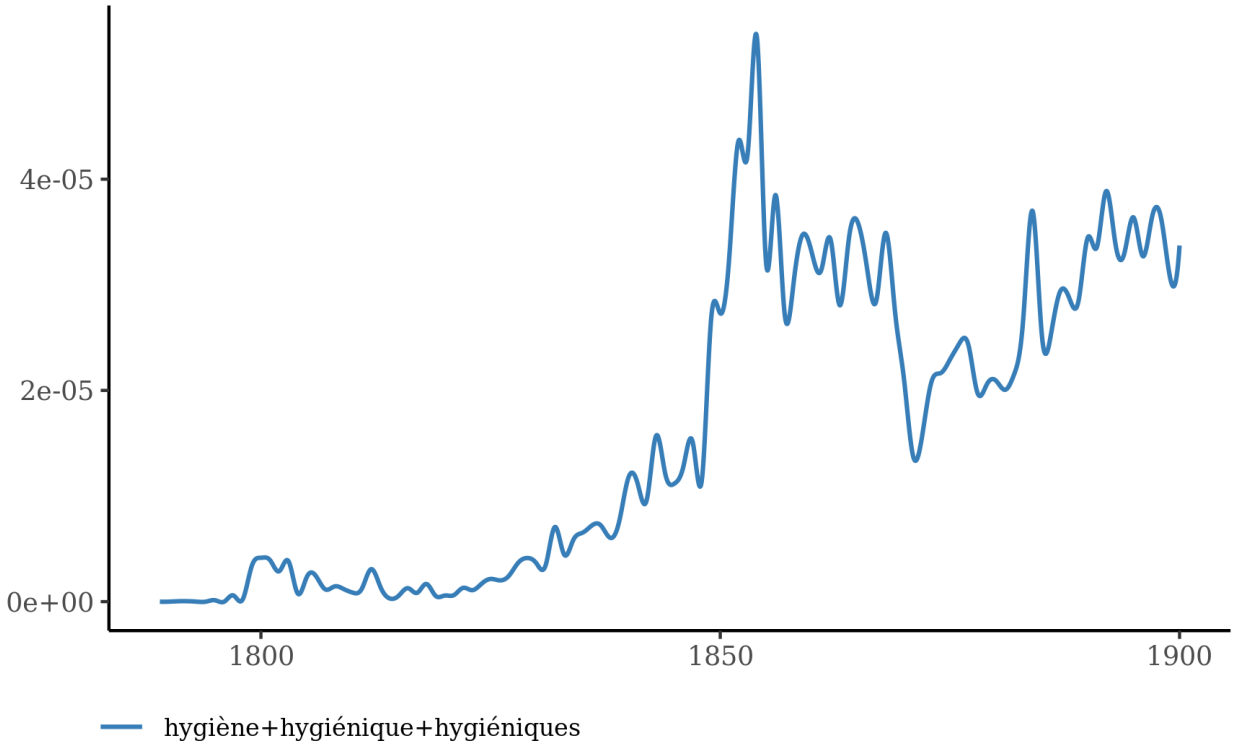
La façon dont une courbe augmente est aussi informative sur le phénomène étudié

On peut distinguer un choc exogène d’un emballement endogène grâce au profil des courbes (Deschâtres et Sornette, 2005)

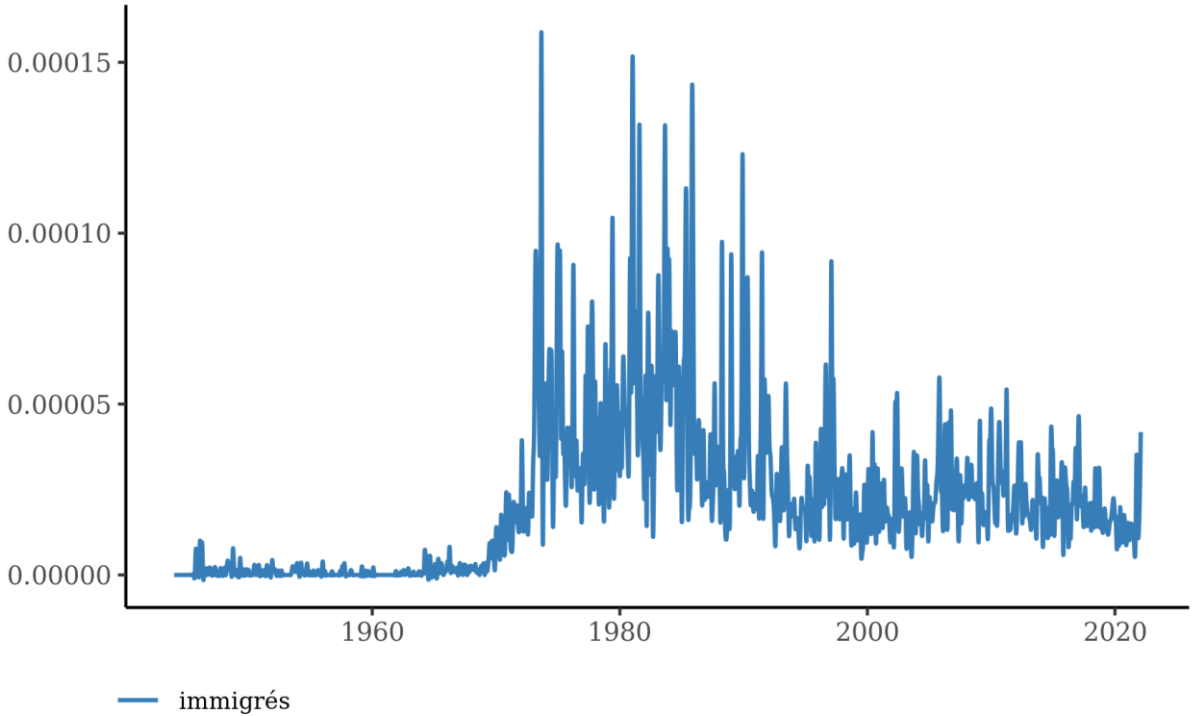




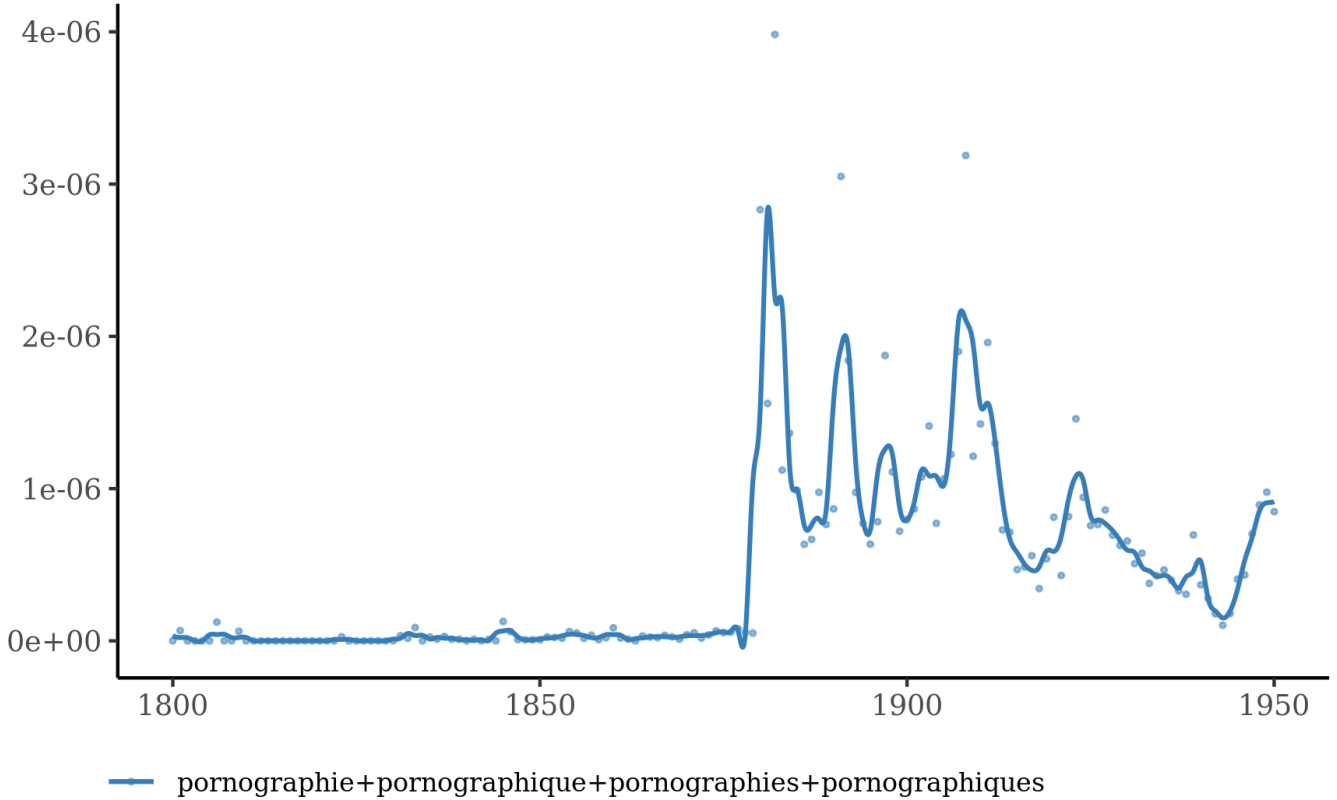
Un exemple d’emballement endogène



Un exemple d'emballement endogène (corpus *Le Monde*)



Un exemple de choc exogène



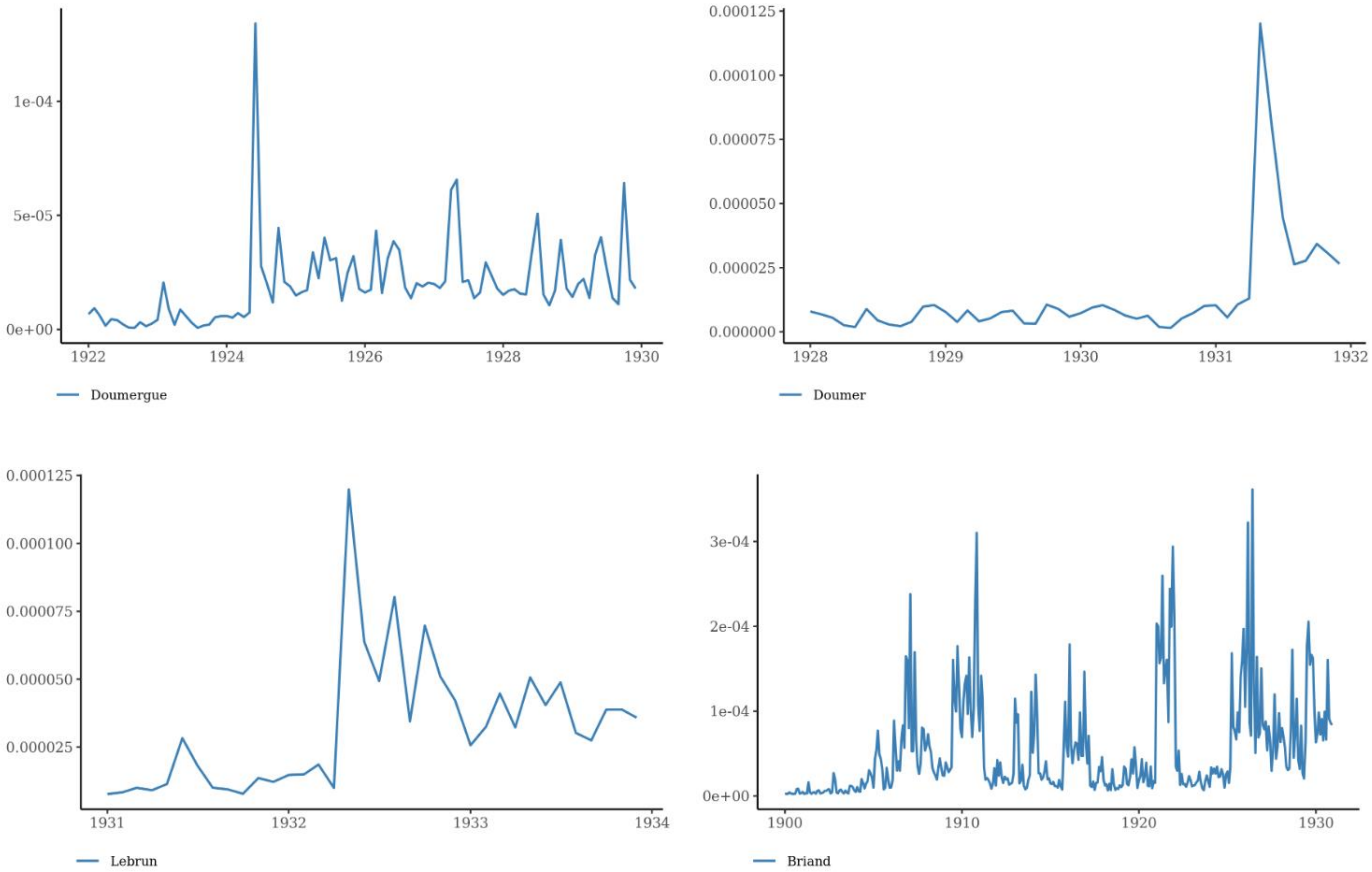


FIGURE 8 – Emballements exogènes dans les mentions de présidents de la République et du Conseil (recherche par n-gramme ; résolution mensuelle).

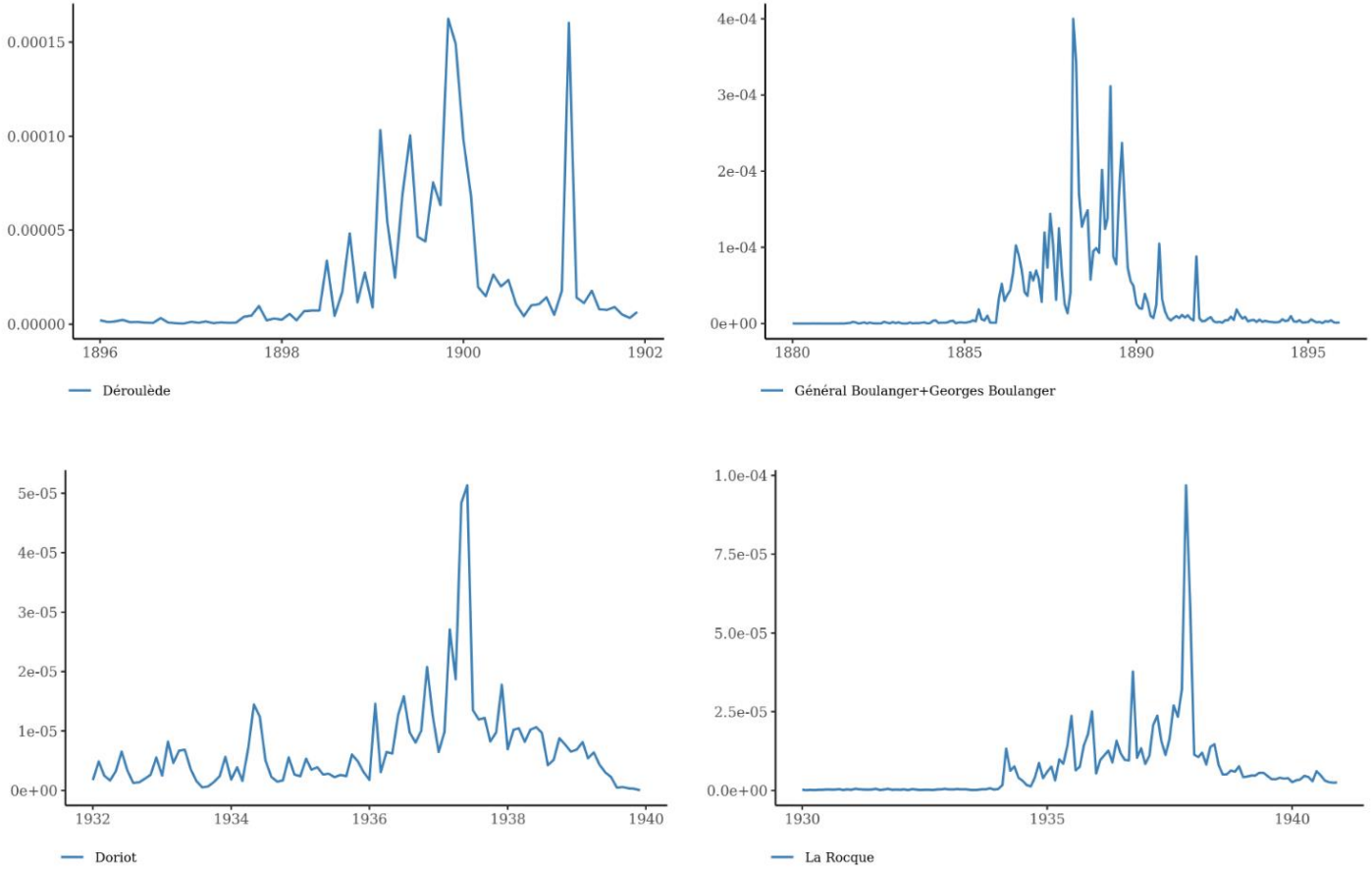


FIGURE 9 – Emballements endogènes dans les mentions de figures politiques d'extrême droite (recherche par n-gramme ; résolution mensuelle).

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 informatiques 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