Gallicagram : un outil de lexicométrie pour la recherche

Benjamin Azoulay[1,†] et Benoît de Courson[2,†]

[1]Ecole Normale Supérieure Paris-Saclay, Gif sur Yvette
[2]Max Planck Institute for the Study of Crime, Security and Law, Freiburg im Breisgau, Allemagne
[†]Les deux auteurs ont également contribué à cet article,

Novembre 2021

Table des matières

Résumé

Gallicagram [1] est un outil de lexicométrie développé pour la recherche en sciences humaines et sociales. Il offre aux chercheurs un moyen de tester rigoureusement leurs hypothèses et de quantifier les effets observés en tirant profit de très vastes bases de données linguistiques, délimitées, accessibles et structurées. La maîtrise des corpus, l'accès aux documents exploités, le traitement distinct de la presse et des livres ainsi que les outils d'analyse intégrés au logiciel lui permettent d'échapper aux critiques portées contre Google Ngram Viewer, dont il s'inspire. Cet article présente le logiciel et illustre ses applications possibles, en particulier dans le champ de l'histoire contemporaine.

1. https://shiny.ens-paris-saclay.fr/app/gallicagram

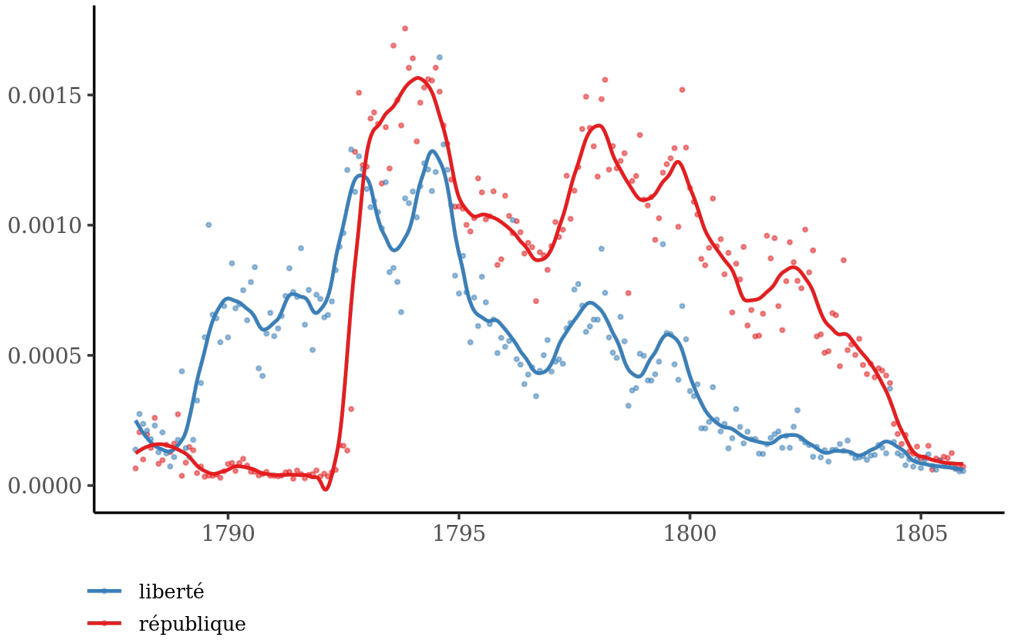

FIGURE 1 – Coévolution des syntagmes « république » et « liberté » dans le corpus de presse de Gallica (résolution mensuelle; recherche par n-gramme; lissage loess=1).

L'historien de demain sera programmeur ou ne sera plus.
Emmanuel Le Roy Ladurie, *Le territoire de l'historien*, 1973 p,11-14.

Introduction

Pour étudier la vie matérielle des générations passées, les archéologues recourent aux fossiles. Leurs vies intellectuelle ou culturelle laissent quant à elles des traces bien moins palpables. Pour écrire une histoire des mentalités, on se sert en général des textes, que l'on peut étudier comme les « fossiles des mentalités ». Les historiens les utilisent spontanément, mais en général de façon isolée et avec une approche qualitative. La discipline naissante des « humanités numériques » propose au contraire des méthodes comme la « lexicométrie», que l'on peut définir comme l'analyse quantitative et automatique de larges corpus de textes. Toutefois, cette approche reste aujourd'hui marginale en histoire.

En la matière, les chercheurs en sciences humaines et sociales ne manquent vraisemblablement pas de données, mais plutôt d'outils. Il est sans doute illusoire d'espérer que les historiens deviennent massivement « programmeurs », comme le souhaitait Emmanuel Le Roy Ladurie [2], mais cela n'interdit pas d'essayer de les doter des outils adéquats. En effet, grâce aux « interfaces utilisateur », une faible familiarité avec l'informatique suffit désormais pour utiliser des outils complexes, programmés par d'autres. Depuis 2011, « Google Ngram Viewer » permet ainsi de tracer en un instant la série temporelles de l'emploi d'un syntagme au cours du temps. Mais les graphiques ainsi produits sont souvent considérés avec scepticisme, voire incrédulité [3]. L'emploi de *Ngram Viewer* reste rare : les recherches « ngram viewer » et « Google ngram » dans Cairn.info indiquent que l'outil n'est évoqué que dans un seul article de la revue *Vingtième siècle* (*20 & 21*), et ce dans une note de bas de page évoquant une « rapide enquête » sur le site [4]. L'outil n'est guère plus usité dans les revues d'histoire quantitative : deux articles des *Annales* le citent (sans pour autant l'utiliser); aucun de la revue

2. Emmanuel LEROY LADURIE. *Le territoire de l'historien*. 1973, p. 11-14.
3. Pierre MOUNIER. « Les Humanités numériques, gadget ou progrès? » *Revue du Crieur* 7.2 (2017), p. 144-159.
4. Arnaud-Dominique HOUTE. « Le temps de l'auto-stop ». *20 21. Revue d'histoire* 148.4 (2020), p. 3-15.

Histoire & Mesure, sur un total de plus de 4500 articles publiés depuis 2011 par les trois revues citées. Cette méfiance s'explique, on le verra, par des limites réelles du logiciel, qui plaident pour le développement d'un outil plus crédible.

Gallicagram est un logiciel de lexicométrie disponible sur internet. Son accès est libre, tout comme son code source et les données sur lesquelles il repose. Son interface graphique présente les courbes d'évolution de l'usage des mots au cours du temps au sein des corpus numérisés par la Bibliothèque nationale de France (Gallica) et d'autres bibliothèques nationales et régionales. Il tente de remédier à certaines lacunes de Ngram Viewer et de fournir des fonctionnalités utiles aux historiens. En particulier, Gallicagram permet de tailler sur mesure son corpus, par exemple en n'exploitantque les archives d'un journal précis. Il est aussi possible de vérifier la pertinence des occurrences grâce à leur contexte et de tracer des courbes précises au mois près. Il se propose d'aider les chercheurs en sciences sociales à quantifier rapidement les phénomènes qu'ils étudient et à tester leurs hypothèses.

1 Éviter les écueils de Ngram Viewer

Google Ngram Viewer pose aux chercheurs en sciences humaines et sociales un certain nombre de problèmes méthodologiques qui restreignent considérablement leur capacité à proposer, à partir de cet outil, des interprétations valables [5]. Ces limites ont été largement documentées et discutées dans la littérature [6]. Si certains chercheurs ont cherché à proposer des palliatifs [7], certaines faiblesses du logiciel de Google, encore ignorées par la littérature, nous paraissent rédhibitoires. Gallicagram a été conçu pour répondre à ces enjeux essentiels d'interprétation en tirant le meilleur profit du libre accès aux corpus de Gallica.

1.1 Ngram Viewer : le pari de la quantité

Principal argument de vente de son logiciel, Google exploite environ 4% des livres publiés depuis l'aube de l'humanité. [8] Au-delà des vertiges que donne ce chiffre, il convient de s'interroger sur l'utilité d'une telle masse. Pour ce faire, on peut dresser une analogie entre la lexicométrie et la technique du sondage. Dans les deux cas, le but est d'estimer un paramètre, respectivement le taux d'utilisation d'un mot à une époque et dans un contexte donnés, et la part de la population qui souscrit à une opinion. Pour cela, le sondeur interroge des individus, jusqu'à obtenir une marge d'erreur raisonnablement faible. Pour un sondage d'intentions de vote, on n'interroge ainsi en général que quelques milliers de personnes. En interroger davantage réduirait bien sûr la marge d'erreur de l'estimation. Mais il paraît futile de la réduire au-delà de l'ordre du point de pourcentage, alors que des biais d'échantillonnage et de réponse séparent de toute façons l'estimation de la réalité qu'on cherche à cerner.
Il en va de même pour la lexicométrie, où l'on interroge des textes les uns après les autres pour y évaluer la présence ou l'absence d'une expression. Le volume du corpus joue le même rôle que la taille d'échantillon : ajouter des textes réduira la marge d'erreur de l'estimation, mais il convient de se demander ce que l'on mesure. Dans un monde idéal, le corpus serait représentatif des productions culturelles humaines et cette estimation correspondrait alors à l'importance culturelle d'un terme dans un espace linguistique. Avec un méga-corpus, une telle représentativité est illusoire. Dans Google Books, on a ainsi constaté une explosion de la part prise par littérature scientifique au cours du temps, dont témoigne l'explosion de la fréquence du syntagme « *et al.* » [9]. Ajouter des textes revient à viser toujours plus proche d'un point qui n'est de toute façon pas le coeur de la cible.

5. Francis Chateauraynaud et Josquin Debaz. *Prodiges et vertiges de la lexicométrie*. Sous la dir. d'Hypotheses. 2010. URL : https://socioargu.hypotheses.org/1963.

6. Emilien Ruiz. *Google labs Books Ngram Viewer : un nouvel outil pour les historiens ?* Sous la dir. de boiteaoutils.info. 2010. URL : https://boiteaoutils.info/2010/12/google-labs-books-ngram-viewer-un/ ; François Héran. « Les mots de la démographie des origines à nos jours : une exploration numérique ». *Population* Vol. 70.3 (2015), p. 525-566.

7. Nadja Younes. « State-of-the-art Research Using the Google Books Ngram Viewer : Improving the Method and Investigating Cultural Change » (2019).

8. Jean-Baptiste Michel et al. « Quantitative Analysis of Culture Using Millions of Digitized Books ». *Science* 331.6014 (jan. 2011), p. 176-182.

9. https://books.google.com/ngrams/graph?content=et+al&year_start=1800&year_end=2019&corpus=26&smoothing=3

Autrement dit, la pertinence de l'analyse lexicométrique n'est pas proportionnelle au volume du corpus. De toute façon, la loi des grands nombres indique que l'estimation converge, le « rendement marginal » d'un texte supplémentaire tend donc rapidement vers zéro. Un corpus aussi large est donc peu utile, et est par ailleurs handicapant à plusieurs titres. D'abord, la quantité est souvent concurrente à la qualité. Les métadonnées de Google Books, en particulier le titre et la date, sont établies par un algorithme et donc sujettes à caution, là où Gallica emploie des bibliothécaires. Une étude de 2012 estimait que 36% des documents de Google Books comportaient des métadonnées erronées [10]. Pour la date, donnée cruciale pour l'outil, le taux d'erreur serait autour de 10% [11].
Pour des études macroscopiques, la quantité peut compenser la qualité. Les données fautives peuvent être vues comme du bruit, ce qui importe peu lorsque l'on fait une étude sur une longue période ou sur un domaine très vaste. Il n'est donc pas surprenant que *Ngram Viewer* ait été si fécond pour étudier des questions très générales sur le très long terme. L'une des études les plus citées fondées sur l'outil analyse par exemple la montée des valeurs individualistes et matérialistes en Occident entre 1800 et 2000 [12]. Ainsi, on pourrait dire que *Ngram Viewer* est adapté à des études particulièrement macroscopiques – ses concepteurs parlent de « culturomics », contraction de « culture » et « genomics » – et non aux questions microscopiques qui occupent en général l'historien. Nous verrons que Gallicagram permet à l'inverse de restreindre le corpus à l'envi, par exemple aux seules archives de *l'Humanité*, ou à la presse résistante de la zone Sud.

1.2 Replacer les mots dans leur contexte : l'accès intégral au corpus et l'étude des cooccurences

Lors la mise en ligne de l'outil, l'historien quantitatif Emilien Ruiz s'interrogeait sur l'opportunité d'utiliser Ngram Viewer dans sa discipline [13] et déplorait qu'« absolument aucun accès au contexte n'est et ne sera jamais possible. Qui a écrit le terme recherché ? Dans quel sens le mot est-il employé ? Dans quel type d'ouvrage ? Autant de questions fondamentales qui restent en suspens » [14]. L'accès au corpus de Ngram Viewer est de toute façon impossible, pour des raisons légales. [15] Choisir un corpus ouvert comme Gallica permet d'une certaine façon de restaurer le lien entre les syntagmes et leurs contextes à travers deux options : l'accès aux textes et l'étude des cooccurences.

Gallicagram a été conçu pour faciliter l'accès aux textes sous-jacents des graphiques. Un simple clic sur un point d'une courbe renvoie à la recherche correspondante dans Gallica. Ngram Viewer propose quant à lui des renvois vers Google Books de façon non systématique [16]. Qui pis est, l'analyse n'est pas réalisée sur le corpus de Google Books mais sur une version largement filtrée, purgée selon les concepteurs des journaux et magazines, et des livres présentant des qualités d'océrisation trop faibles (deux tiers du corpus en 2011), sans que ces règles soient clairement exposées [17]. Il est impossible de savoir quels documents du corpus de Google Books ont été éliminés lors du filtrage. De même, la date de numérisation ne figurant pas dans Google Books, il est impossible de savoir quels documents ont été numérisés après la constitution des bases de données utilisées dans Ngram Viewer. Ainsi, de très nombreuses occurrences apparaissant dans Google Books ne sont pas prises en compte dans Ngram Viewer. Le corpus réel de Ngram Viewer demeure donc inconnu et Google Books restreint aux livres en français ne peut être tenu pour celui-ci. Il est par ailleurs impossible de consulter dans Google Books les ouvrages soumis aux droits d'auteur, pourtant pris en compte par Ngram Viewer. Le nombre de résultats annoncés par Google est, lui, souvent très supérieur au nombre de documents consultables [18]. Cette discordance des corpus exploités est problématique, puisqu'elle crée une rupture radicale entre l'échelle microscopique (le sens des mots dans leur contexte) et l'échelle macroscopique (l'évolution de l'usage des mots affichée dans

10. Ryan James et Andrew Weiss. « An assessment of Google Books' metadata ». *Journal of Library Metadata* 12.1 (2012), p. 15-22.
11. Pour s'en rendre compte, on peut chercher dans Google Books un néologisme avant son apparition, et l'on trouve invariablement des dizaines de documents mal datés ou mal oscérisés.
12. Patricia M Greenfield. « The changing psychology of culture from 1800 through 2000 ». *Psychological science* 24.9 (2013), p. 1722-1731.
13. Ruiz, *Google labs Books Ngram Viewer : un nouvel outil pour les historiens ?*
14. Ibid.
15. Erez Aiden et Jean-Baptiste Michel. *Uncharted : Big Data as a Lens on Human Culture*. New York : Riverhead Books, 2013.
16. Ngram Viewer ne propose pas toujours de renvoi vers le corpus Google Books. Par exemple, la recherche « Foch, 1914-1920 » dans Ngram Viewer ne propose pas d'ouvrir la recherche correspondante dans Google Books tandis que la recherche « Churchill 1933-1945 » figure plusieurs boutons de renvoi au corpus en bas de page.
17. Michel et al., « Quantitative Analysis of Culture Using Millions of Digitized Books ».
18. La recherche du terme « Clemenceau » dans le corpus français de Google Books pour l'année 1897 annonce « environ 1630 » résultats mais n'en présente que 63 étalés sur 4 pages. L'essentiel des résultats est donc inaccessible, bien qu'a priori non couverts par les droits d'auteur.

Ngram Viewer). Gallicagram a au contraire choisi la concordance des corpus : n'y est représenté que ce qui est accessible immédiatement.

Cet accès au corpus sous-jacent permet donc d'étudier le contexte d'usage des mots, qui peut varier considérablement selon les époques. Il rend possible le contrôle de la lecture effectuée par la reconnaissance optique des caractères (OCR) lorsque celle-ci est déterminante pour l'interprétation. Par exemple, une des plus anciennes occurrences du terme « décolonisation » repérée par Gallicagram résulte en réalité d'une mauvaise lecture du mot « décolorisation » [19].
Cela permet aussi le repérage des glissements sémantiques. Ces glissements constituent à la fois un objet d'étude et un obstacle pour l'historien. Pour étudier l'évolution sur le temps long d'une idée, une stabilité sémantique faciliterait certes l'analyse, mais l'évolution des termes employés constitue elle aussi un événement à analyser.
Enfin, l'accès aux textes constitue un garde-fou contre les difficultés qu'occasionnent souvent la polysémie et l'homonymie. Par construction, l'étude des « n-grammes » fragmente les phrases en petites unités, ce qui détache les mots de leur contexte. Une courbe mêlera donc en général des emplois très différents d'une même graphie, connus ou non du chercheur. Dans certains cas, l'augmentation ou la diminution de la fréquence d'un mot peut découler de l'apparition ou de la disparition d'un autre sens associé à ce mot. A ce sujet, l'accès aux textes permet de repérer ces diverses acceptions. Mais Gallicagram propose aussi d'effectuer une sorte de tri entre ces sens, à l'aide des recherches par « cooccurrence », i.e. la présence simultanée de deux mots à proximité dans un texte [20]. Grâce aux cooccurrences, on peut ainsi préciser l'acception d'un mot que l'on recherche. Par exemple, si l'on s'intéresse à l'apparition des grèves en France, il peut être judicieux d'exclure les occurrences où le mot « grève » est employé dans son sens littoral, ou en référence à la place parisienne. Pour ce faire, on pourra par exemple rechercher les cooccurrences du mot avec « travail », « usine », « atelier » ou « salarié » [21].

1.3 La maîtrise du corpus et la conservation de l'échelle macroscopique

S'il est nécessaire d'opérer des vérifications microscopiques ponctuelles pour s'assurer de l'interprétation d'une courbe, il est tout aussi souhaitable de conserver la vision panoramique offerte par l'échelle macroscopique, qui constitue pour certains l'intérêt véritable de la lexicométrie. C'est pourquoi Gallicagram donne une large place à la délimitation des corpus (maîtrise en amont) et à l'analyse de leur structure (maîtrise en aval). Ici, on peut se heurter à des effets de structure, dûs à une transformation de la composition du corpus au cours du temps [22]. Par exemple, une augmentation de la part des textes littéraires dans le corpus pourrait entraîner une augmentation de la fréquence de termes du champ lexical de l'amour, sans que l'usage de ces termes augmente parmi les seuls textes littéraires.

D'emblée, Gallicagram distingue le corpus de presse du corpus de livres. Ces textes sont de nature et de forme trop différentes pour être assimilés. De plus, notre mode de recherche « par document » renvoyant le nombre de documents où apparaît le syntagme étudié, la disproportion entre l'effectif des numéros de presse (3,3 millions) et celui des livres (330 000) dans la base aurait pour effet de noyer les livres. Google Ngram Viewer semble faire fi de cette distinction fondamentale en ce qui concerne le corpus français. Pour cette langue, en effet, le filtre « journaux » de Google Books [23] ne présente aucun résultat tandis que l'on retrouve de nombreux journaux compilés annuellement sous la rubrique « livres » [24]. Le filtre « magazines » du corpus français de Google Books [25] n'est guère plus satisfaisant. De cette absence d'indexation résulte donc aussi une surestimation du nombre de livres exploités par Google Ngram Viewer. Il est cependant impossible de l'estimer du fait de l'opacité du corpus retenu par les concepteurs de ce logiciel.

19. https://gallica.bnf.fr/ark:/12148/bpt6k11449886/f3.image.r=decolonisation
20. Cette option est disponible sur Gallicagram lorsqu'on utilise Gallica comme source et « Par document » comme mode de recherche.
21. Dans la syntaxe de Gallicagram, cela s'écrira (quelque peu laborieusement il faut l'admettre) « grève*travail+grève*usine+grève*atelier »
22. L'Insee définit ainsi l'effet de structure : « Lorsqu'une population est répartie en sous-populations, il peut arriver qu'une grandeur évolue dans un sens sur chaque sous-population et dans le sens contraire sur l'ensemble de la population. Ce paradoxe s'explique parce que les effectifs de certaines sous-populations augmentent alors que d'autres régressent. »
23. https://www.google.com/search?q=le&lr=lang_fr&biw=1280&bih=577&tbs=lr%3Alang_1fr%2Cbkt%3As&tbm=bks&ei=UnJ4YZG_EI-lUveMkHg&oq=le&gs_l
24. Par exemple, la recherche « Foch 1919 » dans Google Books avec le filtre « livres » présente les numéros du journal « L'Illustration ».
25. https://www.google.com/search?q=le&hl=fr&tbs=bkt:m,lr:lang_1fr&tbm=bks&source=lnt&lr=lang_fr&sa=X

Gallicagram propose d'autres options de délimitation des corpus. À l'intérieur du corpus de presse, l'utilisateur peut restreindre sa recherche à un nombre limité de titres, pour composer lui-même un corpus qu'il juge homogène. Il a, de même, le choix parmi des présélections thématiques ou géographiques de titres de presse (par exemple les principaux titres de la Résistance ou encore la presse d'un département spécifique). Il peut aussi importer dans l'application un « corpus personnalisé », créé sur Gallica grâce à la fonctionnalité « rapport de recherche ». Enfin, les résultats sont affichés non en valeurs absolues (nombre d'occurrences), mais en pourcentage du volume de la base, de sorte à neutraliser les variations du volume de documents disponibles.

*

Pour évaluer les possibles effets de structure de l'analyse, il est salutaire d'analyser la composition du corpus. Cette entreprise est rendue possible par la fiabilité et la qualité des métadonnées de Gallica, établies par les bibliothécaires de la BnF. L'absence d'une telle expertise chez Google Books constitue une source d'incertitude importante.

Le premier effet de structure à évaluer réside dans le volume du corpus exploité. Un volume insuffisant risque de nuire à la validité de l'interprétation. Par exemple, le corpus de presse semble utilisable sur la période 1789-1944. Malgré de fortes variations dans le volume de la base, au moins 2 000 numéros de presse sont disponibles pour chaque année sur cette période. Pour faciliter cette vérification qui doit être systématique, il est possible d'afficher la distribution chronologique des documents du corpus en dessous du graphique de fréquence généré dans Gallicagram.

L'utilisateur dispose pour chaque corpus global – presse et livres – d'un onglet dédié à l'analyse de leur structure. Les graphiques qui y sont représentés permettent d'asseoir ou d'amender les interprétations envisagées. Les données y sont figurées sous la forme de diagrammes chronologiques. L'onglet « corpus de presse » permet de visualiser la distribution chronologique des numéros, leur distribution géographique selon la ou les villes de publication des journaux, leur périodicité (quotidienne ou non), leur classification thématique (Dewey), la qualité de l'OCR, la date de numérisation, la bibliothèque d'origine du document, sa répartition entre Gallica et Retronews ou encore le nombre de pages par fascicule. L'onglet « corpus de livres » représente en outre le régime de droits d'auteur et le nombre de pages par ouvrage. S'y ajoute l'état de numérisation et d'océrisation de l'ensemble des livres de la Bibliothèque nationale de France.

Une autre application, « Gallicapresse », permet d'analyser la structure du résultat des recherches parmi le corpus de presse. L'utilisateur peut ainsi décrire l'évolution du corpus constitué par les occurrences trouvées (et non plus seulement la base) selon les catégories pré-citées titre de presse, la ville de publication, la périodicité et la classification thématique de Dewey. Les valeurs sont figurées en termes absolus ou relatifs à discrétion de l'utilisateur. Une cartographie est également disponible.

2 L'application et les corpus dont elle dépend

2.1 Le corpus de presse : une ouverture de perspectives en histoire politique

La recherche dans Google Ngram Viewer se fonde uniquement sur le corpus de livres de Google Books. L'entreprise californienne a décidé d'éliminer le corpus de presse, car il présentait une moindre qualité d'océrisation due à la complexité structurelle des documents numérisés (colonnes multiples, insertion d'images, titres, nombreuses césures en fin de ligne difficiles à distinguer du trait d'union, moindre état de conservation). Nous avons montré que dans le cas du corpus français, cette distinction ne semblait pas avoir été mise en oeuvre, ce qui renforce encore les incertitudes quant au contenu réel du corpus exploité par Ngram Viewer. Pour autant, le nombre de titres de presse français numérisés par Google semble assez limité [26]. Sous la présidence de Jean-Noël Jeanneney, fer de lance de la conversion de Gallica en projet de numérisation de masse en 2005, les équipes de la BnF ont su au contraire accentuer l'effort de numérisation des périodiques, promouvant une océrisation certes moins rapide, mais de meilleure qualité. [27]

26. La recherche « Alger » pour 1830 dans Google Books, date de la conquête française de l'Algérie, ne contient qu'un seul titre de presse, « Le défenseur de la monarchie et de la charte » et 26 revues contre 90 titres de presse ou de revues distincts dont 23 quotidiens pour la même recherche dans Gallica. Dans Google Books, les documents sans auteur sont généralement des journaux ou des revues, mais il est impossible de restreindre la recherche aux documents sans auteur. Il faut donc les dénombrer manuellement.

27. J. N. Jeanneney. *Quand Google défie l'Europe : Plaidoyer pour un sursaut*. 1001st edition. Paris : 1001 nuits, jan. 2005.

Cela ouvre de nouvelles perspectives pour l'histoire politique, avec la possibilité d'observer finement, dans un corpus de 3,3 millions de fascicules, les tendances, les évolutions et les évènements que révèle l'application. À ces trois optiques correspondent grossièrement trois options de visualisation dans le logiciel : le lissage loess, l'échelle annuelle et l'échelle mensuelle. Cette dernière option constitue une avancée par rapport à Ngram Viewer, qui ne représente les fréquences qu'à l'année près.

La dernière partie de cet article est consacrée à des études de cas illustrant la diversité des usages qu'il est possible de faire de Gallicagram en histoire politique contemporaine et la précision des analyses que cet outil permet de produire.

2.2 Plusieurs méthodes d'analyse pour recouper ses résultats

La multiplicité des méthodes d'analyse permet de recouper les résultats et de présenter des interprétations plus fiables.

Trois types de mesure sont ainsi disponibles : le dénombrement des documents figurant le syntagme recherché, celui des pages où apparaît la recherche et le comptage des occurrences du syntagme dans le corpus ciblé (*Appendice n°1 : Données et traitements*). Leur pertinence respective est fonction de la recherche effectuée, du corpus exploité et de l'hypothèse testée.

Ainsi, la recherche au document est plus pertinente que les autres types de mesure dans le corpus de presse car elle exploite des fascicules contenant peu de mots tout en évitant les artefacts liés à la répétition d'un même terme dans un même document. Ce type de mesures produit des résultats satisfaisants pour lorsque la fréquence est comprise environ entre 25% et 65%. Au-delà, l'amplitude des variations s'estompe, en deça le bruit statistique produit des fluctuations trop larges pour y voir clair.

Au contraire, la recherche préconisée dans le corpus de livres est la mesure des occurrences. En effet, par nature, les livres sont des documents volumineux contenant beaucoup de mots. Deux termes courants sont ainsi, aussi bien l'un que l'autre, susceptibles de figurer au moins une fois dans un grand nombre de livres. La courbe représentée figurera alors bien plus l'évolution de la structure du corpus que celle de l'usage effectif des mots. Pour l'étude de mots rares (en deçà de 5%) dans le corpus de livre, la recherche au document est plus adaptée.

Enfin, la recherche à la page s'avère particulièrement utile pour explorer des corpus de presse restreints. En effet, les titres de presse adoptent souvent une mise en page thématique, chaque page abordant des sujets prédéterminés. Les évolutions étudiées dans ce mode de recherche révèlent ainsi mieux certains effets de diffusion.

Néanmoins, l'utilisateur gagnera à expérimenter ces trois types de mesures pour une même recherche dans un même corpus et à en comparer les résultats.

Il lui est aussi possible de comparer les résultats d'un même type de mesures pour une même recherche dans différents corpus (presse, livres, presse restreinte à certains titres, corpus personnalisé).

Gallicagram permet la comparaison avec les résultats de Ngram Viewer, directement dans la fenêtre de l'application. L'utilisateur peut alors comparer le résultat d'une même recherche dans le corpus de livres de Gallica et dans le corpus de livres exploité par Ngram Viewer.

De plus, Gallicagram embarque un certain nombre d'outils, qui apparaissent lorsqu'on clique sur la boîte « option avancée » :

- La « matrice de corrélation » permet de mesurer à quel point les différentes recherches évoluent de façon synchrone. Nous recommandons cependant d'être prudent dans l'interprétation : il est facile d'obtenir une corrélation fallacieuse, par exemple parce qu'une large part des syntagmes sont sous-utilisés pendant les deux guerres mondiales, ce qui produit artificiellement des corrélations positives.
- La « différence de fréquence » affiche la soustraction d'une série temporelle par une autre.
- Le « rapport de fréquence » fait de même, avec un quotient.

— Le « Rééchelonnement des résultats » transforme les séries temporelles en un « z-scores », une méthode statistique répandue, consistant à diviser chaque série temporelle par son écart type afin de rendre comparables à l'oeil des courbes qui ne sont pas du même ordre de magnitude – ce qui est souvent le cas en linguistique conformément à loi de Zipf[28] : certains mots sont utilisés des milliers de fois plus souvent que d'autres.
— « Afficher toutes les données de la session » permet de comparer des séries issues de différents modes d'analyse, corpus ou langue. On peut en particulier utiliser cette option pour effectuer une comparaison inter-linguistique. Par exemple, la juxtaposition de la recherche « choléra » dans les journaux britanniques et français en 1831-1832 révèle l'arrivée successive du bacille dans les deux pays.

2.3 L'extension du logiciel vers d'autres bibliothèques et d'autres langues

Gallicagram a été spécialement développé pour les corpus de Gallica, mais son principe peut s'appliquer à toutes les bibliothèques numériques libres d'accès et de droits. C'est pourquoi nous avons intégré progressivement d'autres bibliothèques nationales et régionales francophones, puis en langues étrangères (allemand, anglais, néerlandais, espagnol). Nous avons enfin intégré des corpus de presse contemporains (*Le Monde* et *Le Figaro*) et des corpus scientifiques (Cairn et Isidore). Pour une liste, mieux vaut ici renvoyer le lecteur vers l'application plutôt que de dresser un long inventaire des corpus disponibles, qui ne serait de toute façon bientôt plus à jour, puisque de nouvelles bibliothèques sont régulièrement ajoutées. Grâce à l'option « Afficher toutes les données de la session », décrite plus haut, on peut effectuer des comparaisons intercorpus, et donc interculturelles ou interlinguistiques.

Tous les corpus proposés dans le logiciel sont clairement délimités selon les critères du genre et de la langue. L'accès aux sources accès est garanti : un clic sur un point du graphique ouvre la recherche correspondante dans la bibliothèque numérique choisie. Cependant, la maîtrise aval du corpus (description fine de sa structure) demeure limitée aux corpus de presse et de livres de Gallica. Pour les autres corpus, seule la distribution au cours du temps est accessible (onglet « Distributions » de Gallicagram). Elle permet de déterminer les périodes pour lesquelles une recherche dans ces corpus peut être considérée comme fiable.

2.4 Le mécanisme du logiciel

Gallicagram est un outil libre dont le code est accessible (*open source*) et qui exploite des données qui, pour l'essentiel, le sont aussi (*open data*). Cette transparence vise à donner aux chercheurs une compréhension fine du fonctionnement de l'application sans laquelle il peut être délicat d'interpréter les résultats du logiciel. Les données exploitées et les opérations effectuées par le logiciel varient selon le mode de recherche et le corpus choisis. Le détail, précisant pour chaque combinaison possible la délimitation du corpus, la formulation de la recherche et le calcul de l'indicateur est exposé en annexe de cet article (*Appendice n°1 : Données et traitements*).

Gallicagram propose une recherche par n-gramme dans les corpus de presse et de livres de Gallica. Ce mode de recherche interroge deux bases de données que nous avons créées à partir des océrisations proposées par Gallica. Le processus de construction de ces bases de données et la description de leur structure finale telle qu'elle est interrogée par le programme lors d'une recherche est présenté en annexe (*Appendice n°2 : Bases de données en n-grammes*).

2.5 Les précautions d'usage : des données imparfaites : des résultats à interpréter prudemment

L'utilisation de Gallicagram offre de nombreuses opportunités pour la recherche, mais elle doit tenir compte des limites de l'outil pour demeurer pertinente.

28. George Kingsley Zipf. « Relative frequency as a determinant of phonetic change ». *Harvard studies in classical philology* 40 (1929), p. 1-95.

Le programme de numérisation des documents de la BnF suit un rythme assez lent. Le processus d'OCR, qui transforme une image en texte interrogeable, est lui aussi peu rapide [29]. Ainsi, si le travail de numérisation et d'océrisation continuait à ce rythme et si le volume des collections restait figé au niveau actuel, il faudrait encore 240 années pour numériser la totalité des près de 6 millions de livres en français de la BnF. Le corpus de livres exploité par *Gallicagram* demeure donc modeste, comparé à celui de *Ngram Viewer*, qui compte environ dix fois plus d'ouvrages. Toutefois, comme on l'a vu plus haut, le volume du corpus ne détermine pas la valeur de l'analyse. Gallica et Google Books sont deux corpus de taille gigantesque, où l'on peut raisonnablement considérer que la loi des grands nombres s'applique, et que la taille de l'échantillon n'influe plus. La période la plus océrisée est le XIXe siècle avec des taux compris entre 11% et 20% d'océrisation dans le corpus de livres. Pour cette période, l'essentiel des documents numérisés a déjà été océrisé, alors que pour les siècles précédents, la majorité des documents numérisés reste en attente d'océrisation.

La qualité de l'océrisation peut aussi s'avérer problématique, d'autant qu'elle est très variable selon les périodes. Si la qualité de l'OCR semble satisfaisante pour les documents publiés après 1800 avec des taux de confiance dépassant en moyenne les 90% dans le corpus de livres, elle l'est beaucoup moins pour les périodes précédentes (77% en moyenne en 1660 pour ce même corpus). Cela est d'autant plus problématique que cet indicateur de confiance semble surestimé. L'utilisateur de Gallicagram gagnera donc à resserrer ses études sur les périodes postérieures à la Révolution française. Il peut cependant toujours accéder au corpus sous-jacent à sa recherche pour vérifier lui-même l'exactitude de l'OCR. Cela est essentiel et assez aisé pour la recherche de la date d'apparition d'un terme ou d'une expression. Dans ce cas là une erreur d'OCR sur une lettre peut falsifier l'hypothèse, comme dans la recherche « décolonisation », évoquée plus haut. Mais si les faux-positifs sont aisément identifiables, la recherche des faux négatifs est plus fastidieuse et c'est à l'utilisateur de deviner les termes proches avec lesquels l'OCR aurait pu confondre le mot recherché. Gallicagram est ainsi plus performant pour identifier les phases d'adoption d'un syntagme nouveau que le moment exact de son apparition. Par ailleurs, certaines confusions de l'OCR peuvent être anticipées. Il en va ainsi de la mauvaise lecture de l'ancienne typographie du « s » long que l'OCR lit souvent comme un « f » (par exemple « caffer » pour « casser »). De même, la mauvaise lecture de la ponctuation par l'OCR peut avoir créé des erreurs dans la reconnaissance de la délimitation des phrases et avoir ainsi altéré l'indexation des syntagmes contigus dans les bases de données en n-grammes. Certaines métadonnées utilisées dans la délimitation du corpus s'avèrent aussi erronées. Toutefois, ces métadonnées ont été établies par des bibliothécaires (hormis l'indice de qualité d'océrisation), les erreurs sont donc résiduelles.

*

Certaines limites tiennent à l'interprétation des courbes affichées par le logiciel. L'utilisateur doit garder à l'esprit que Gallicagram ne représente pas la notoriété d'un syntagme, mais seulement l'évolution de son usage dans un corpus de documents. L'audience d'une publication est par exemple ignorée, ce qui signifie qu'un journal national doté d'un lectorat considérable sera autant pondéré dans la courbe qu'une publication associative confidentielle. L'augmentation d'une valeur ne correspond pas non plus à l'augmentation du nombre de fois où un syntagme a été imprimé : le nombre d'exemplaires de chaque document reste lui aussi inconnu.
Par ailleurs, Gallicagram décrit l'emploi d'un syntagme par les auteurs d'une époque, et non sa réception par les lecteurs – ce qui peut évidemment se révéler problématique, mais constitue aussi un avantage lorsqu'on s'intéresse à la production des contenus culturels.

3 Études de cas

Gallicagram permet de représenter graphiquement l'évolution de l'emploi d'un syntagme au cours du temps. Son usage le plus immédiat est l'illustration de grandes tendances culturelles, telles que l'avènement d'une « civilisation des loisirs » au cours du XXe siècle [30]. Par exemple, la recherche des mots « vacances », « sport » et « loisir » dans la presse française (Fig. 2) présente bien une augmentation continue entre le début du XIXe siècle et le commencement de la

29. En 10 ans, le volume du corpus de livres océrisés en français avec un taux de fiabilité de la reconnaissance optique des caractère estimé supérieur à 50% a augmenté en moyenne de 23 000 livres par an pour passer de 137 000 en janvier 2011 à 370 000 livres en janvier 2021.
30. Joffre Dumazedier. *Vers une civilisation du loisir ?* Éditions du Seuil, 1962.

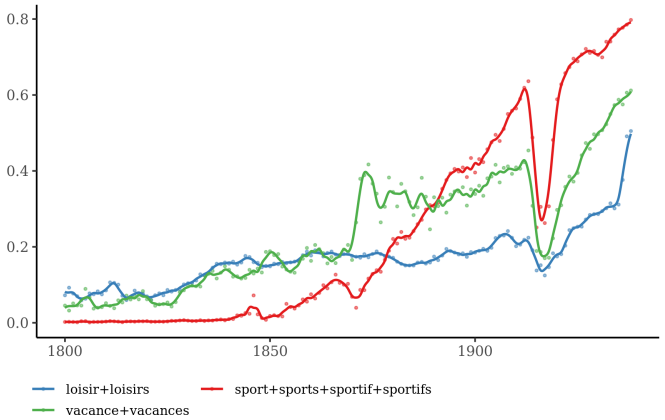

FIGURE 2 – Recherche des mots « vacances », « sport » et « loisir » et leur dérivés (Gallica-presse; recherche par document; lissage=2).

Seconde Guerre mondiale, si l'on excepte le premier conflit mondial. Toutefois, ce genre de résultat n'est ni surprenant, ni nouveau : on obtiendrait la même tendance avec un outil comme *Ngram Viewer*.

Les illustrations proposées ici permettent de de suggérer des applications scientifiques propres à Gallicagram. Le logiciel permet non seulement d'étudier les tendances de long-terme, mais aussi de s'approcher au plus près des évènements en tirant profit du corpus de presse et de la résolution mensuelle. Par ailleurs, il offre la possibilité de restreindre le corpus pour mener une étude à un niveau plus microscopique, à l'échelle d'un territoire, d'un thème, ou d'une source précise.

3.1 L'étude des événements

Lors du développement de Ngram Viewer, les concepteurs ont délibérément choisi d'en exclure la presse, réputée inadaptée à la reconnaissance automatique de caractères à cause de sa disposition en colonnes. Le corpus est donc essentiellement constitué de livres, écrits le plus souvent à distance des évènements. A l'inverse, Gallica dispose d'un corpus de presse considérable, qui s'avère précieux pour repérer les événements historiques dans les graphiques, vu la proximité temporelle entre leur survenue et la date de publication.

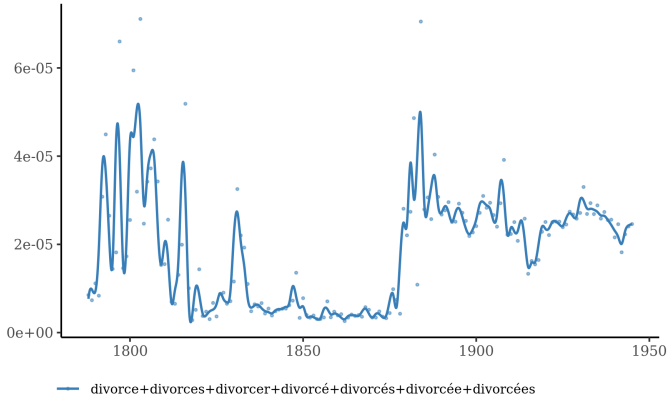

FIGURE 3 – Occurrences des dérivés du mot « divorce » dans le corpus presse (Gallica-presse; recherche par n-gramme; lissage=1).

La Fig. 3 permet par exemple de retrouver, voire de réévaluer, les moments où le thème du « divorce » a affleuré dans le débat public. On y observe, dans l'ordre, sa légalisation en 1792-1794, son interdiction par Louis XVIII en 1816, la tentative avortée de restauration avec le dépôt d'un projet de loi en 1831-1833, sa légalisation par la loi Naquet de 1884 et son assouplissement en 1908. Le reste de la courbe ouvre des pistes intéressantes : on devine un net recul du débat sur le divorce après 1833, et une réapparition du sujet qui s'amorce en 1876, année qui correspond au premier dépôt à la Chambre d'un projet de loi autorisant le divorce par Alfred Naquet.
On observe aussi des pics en 1797 et 1888 qui ne correspondent apparemment à aucun changement juridique. La possibilité d'accéder aux sources primaires dans le moteur de recherche de Gallica aide ici à déterminer leur signification. L'étude des sources primaires en 1797 [31] révèle un débat parlementaire véhément, que Grégoire Bigot qualifie de « tour de chauffe des anti-divorciaires » [32]. En pleine réaction thermidorienne, le Conseil des Cinq-Cents charge une commission spéciale d'étudier l'interdiction du divorce pour « incompatibilité d'humeur », avant que, toujours selon Grégoire Bigot [33], « le coup de barre "à gauche", consécutif au coup d'État de Fructidor an III, interdise d'aller plus loin dans la remise en cause des acquis de la Révolution ». Le pic de 1888 s'avère, quant à lui, dû à la sortie de la pièce de théâtre *Les surprises du divorce* en mars et au divorce du roi de Serbie en octobre, et donc peu interprétable. A l'inverse, la même requête sur Ngram Viewer [34] révèle peu de choses, avec des pics étonnants en 1807 et 1827, probablement dus à des recueils de documents juridiques datés de ces années là, quoique l'opacité du logiciel américain empêche de tirer toute conclusion définitive.

Pour éprouver quantitativement cette fidélité aux événements, on peut comparer les séries temporelles produites par Gallicagram à un type d'événement aisément quantifiable : les grèves. Les occurrences des termes « grève » et « grèves » suivent très fidèlement le nombre d'événements [35] effectivement recensées par Edward Shorter et Charles Tilly dans leur livre *Strikes in France* [36], qui a eu un écho durable chez les historiens quantitatifs français [37] (Fig. 4).
Plus formellement, les deux séries temporelles ont un très fort coefficient de corrélation (0,82). Ce résultat peut en partie s'expliquer par une variable confondante : la date. Les deux séries ont tendance à augmenter au fil du temps, il est donc prévisible qu'elles soient corrélées. Une corrélation partielle contrôlant cet effet donne une valeur à peine amoindrie de 0,75. Cette même analyse faite avec Ngram Viewer donne une valeur inférieure mais forte (0,69), qui suggère que le corpus français utilisé pour l'outil comporte, malgré les règles affichées par Google, une forte proportion de journaux [38].
On trouve donc un lien très fort entre un indicateur produit par lexicométrie et les comportements de l'époque – les données de Shorter & Tilly sont en effet fondées sur les statistiques du ministère du Travail. Autrement dit, les textes disent bien quelques choses du réel, ce qui peut sembler trivial mais vaut la peine d'être questionné. Par ailleurs, un chercheur s'intéressant à un objet non ou mal quantifié peut employer une série temporelle produite par Gallicagram en tant que première approximation. On pourrait par exemple l'utiliser pour estimer le nombre de « faillites » par an avant qu'il ne soit mesuré par l'Etat. Plus simplement, les recherches « krach » ou « épidémie » dans la presse donnent une forêt de pics pointus, qui permettent de répertorier les principaux krachs boursiers et épidémies sur deux siècles.

On peut aussi utiliser l'outil pour étayer la thèse du « déclin de la grève violente » à partir de la Première Guerre mondiale, proposée par ces mêmes auteurs [40]. Pour leur étude, ils ont constitué des séries manuellement, en parcourant les archives du *Temps* de 1890 à 1935, à chaque fois sur trois mois de l'année choisis aléatoirement, et en y recensant les grèves ayant entraîné soit des destructions matérielles étendues, soit plusieurs morts et blessés. Ils ont ainsi classé 98 grèves comme violentes, soit moins de 2 par an en moyenne. Aujourd'hui, cette démarche apparaît chronophage et

31. Disponibles ici, ou dans l'application, en cliquant sur le point correspondant dans le graphique.
32. Grégoire Bigot. « Impératifs politiques du droit privé : le divorce sur simple allégation d'incompatibilité d'humeur ou de caractère"(1792-1804) ». *Clio@ Themis* 3 (2010).
33. Ibid.
34. https://books.google.com/ngrams/graph?content=divorce&year_start=1789&year_end=1939&corpus=30&smoothing=0
35. Ces données sont disponibles à cette adresse.
36. Edward Shorter et Charles Tilly. « The shape of strikes in France, 1830–1960 ». *Comparative Studies in Society and History* 13.1 (1971), p. 60-86.
37. Michelle Perrot et Patrick Fridenson. « Charles Tilly et la France ». *Le Mouvement Social* 4 (2008), p. 143-145.
38. Avec le corpus de Livres de Gallica, qui est pour sa part constitué seulement de monographies, on obtient une série temporelle qui n'a plus grand chose à voir avec les valeurs de Shorter & Tilly et ne présente même pas de pic en 1936. C'est ce qu'on attendrait de Ngram Viewer si son corpus excluait réellement les périodiques. La distinction entre livres et journaux n'étant pas faite dans la version française de Google Books, la filtration du corpus pour Ngram Viewer reste mystérieuse.
40. Edward L Shorter et Charles Tilly. « Le déclin de la grève violente en France de 1890 à 1935 ». *Le mouvement social* (1971), p. 95-118.

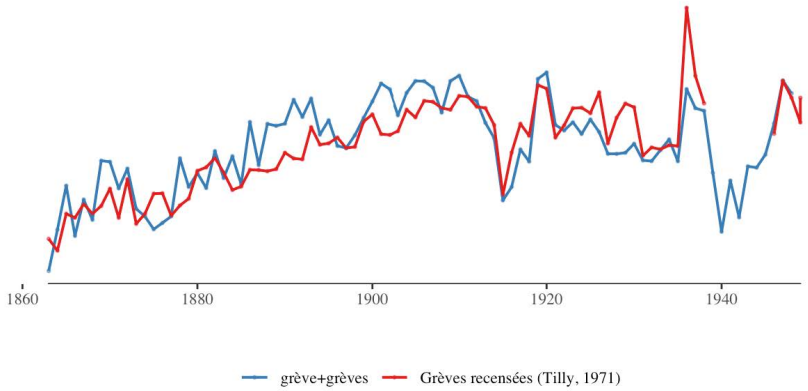

FIGURE 4 – Comparaison de la fréquence du mot « grève(s) » dans le corpus presse avec le nombre de grèves recensées en France par Shorter & Tilly, [39] échelles logarithmiques. Les données sont absentes entre 1939 et 1945.

assez peu exhaustive, même si elle évite de recenser le moindre faux-positif. Avec Gallicagram, on peut étendre l'étude à peu de frais, en utilisant des cooccurrences entre la grève et le lexique de la violence. On peut par exemple demander au logiciel de recenser pour chaque année les documents où le mot « grève » apparaît, au singulier ou au pluriel, à proximité (à moins de 10 mots de distance) des mots « violence(s) », « brutal(es) », « brutalité » ou « blessé(s)». On peut restreindre cette recherche aux seules archives du *Temps*, auquel cas on obtient un graphique assez similaire à la Figure I des auteurs. Certains points discordent, en particulier en 1919, où Shorter et Tilly ne recensent qu'une seule grève violente, et en 1920, où ils n'en rapportent aucune. Pourtant, respectivement 19 et 25 numéros du *Temps* figurent des cooccurrences entre « grève » et notre lexique de la violence. Pour trancher, on peut regarder le détail de ces documents, et il apparaît que malgré des faux positifs (évoquant par exemple des grèves allemandes), on trouve d'authentiques grèves violentes, comme le numéro du 16 mai, qui décrit l'assaut d'un poste de police par des grévistes pour libérer un camarade, et où « deux agents ont été blessés » [41]. Par ailleurs, ces années ont été marquées par la multiplication des grèves insurrectionnelles sur le modèle de la jeune révolution russe, par la création des Comités syndicalistes révolutionnaires au sein de la CGT et par des manifestations meurtrières chaque année lors du premier mai (deux morts en 1919, trois en 1920, des dizaines de blessés à chaque fois) [42]. Le choix des auteurs de n'étudier que trois mois par an explique sans doute qu'ils aient manqué les grèves violentes de ces deux années, qui semblent concentrées autour du moins de mai.

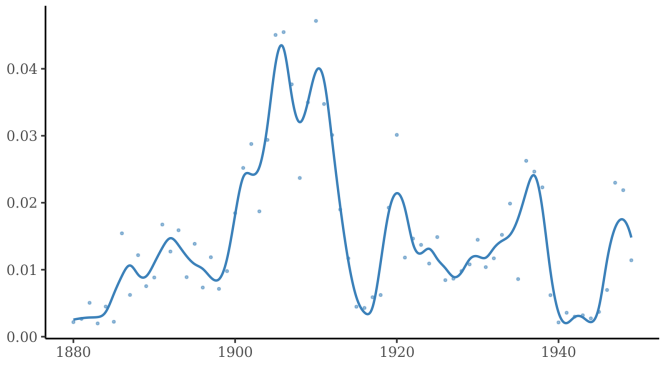

FIGURE 5 – Fréquence des cooccurences de « grève(s) » et des mots « violent(es)», « violence(s)», « brutal(es)», « brutalité », ou « blessé(es) » dans le corpus Presse, lissage=1

41. https://gallica.bnf.fr/ark:/12148/bpt6k244013x/texteBrut
42. Danielle TARTAKOWSKY. « Manifestations ouvrières et théories de la violence : 1919-1934 ». *Cultures & Conflits* 09-10 (1993).

L'étude peut enfin être élargie, à la fois dans le corpus et la chronologie. Si l'on prend par exemple le corpus de presse intégralement, on ne compte plus les grèves individuellement, puisqu'un même événement sera probablement évoqué en parallèle dans une multitude de journaux différents. Mais un large corpus a l'avantage d'être moins arbitraire, et neutralise en partie les effets des orientations éditoriales et les appréhensions différentes d'un même événement pour des journalistes de sensibilités diverses. De plus, le grand nombre de documents permet de réduire le bruit statistique et produit ainsi un résultat plus interprétable. Cet exemple illustre la liberté laissée à l'utilisateur entre une approche parfaitement contrôlée et une approche *big data* – un entre-deux restant possible, en incluant par exemple les principaux quotidiens.

En guise d'illustration, la Fig. 5 représente le nombre de ces cooccurrences dans la presse entière. Elle suggère, conformément à l'étude de Shorter et Tilly, une apogée de la grève violente à la Belle époque. Mais elle indique aussi une résurgence en 1919-1920, ainsi qu'au-delà de leurs bornes chronologiques, lors du Front Populaire (1936) et lors des grèves insurrectionnelles de 1947. Toutefois, ces deux dernières dates étant marquées par un nombre considérables de grèves, il est possible que les grèves violentes n'en constituent qu'une faible proportion.

3.2 Les études monographiques

Grâce à l'API de Gallica, l'utilisateur peut restreindre la recherche à une ou plusieurs sources précises [43]. Cela permet par exemple d'étudier l'évolution du ton ou des thèmes d'un journal en particulier. L'hebdomadaire « Je suis partout » présente ainsi à partir de 1932 – année qui correspond à l'arrivée de Robert Brasillach dans une rédaction auparavant politiquement diverse, et au soutien déclaré du journal à Benito Mussolini [44] – une explosion de la fréquence des syntagmes typiques de l'antisémitisme : « le(s) juif(s) » et le préfixe « judéo- » [45] (Fig. 6). S'ensuit une quasi-disparition de ces expressions à partir de la déclaration de guerre de septembre 1939, les journaux pro-fascistes étant suspects aux yeux du gouvernement. Puis, après 8 mois d'interdiction, le journal devenu collaborationniste reprend visiblement son ton antisémite.

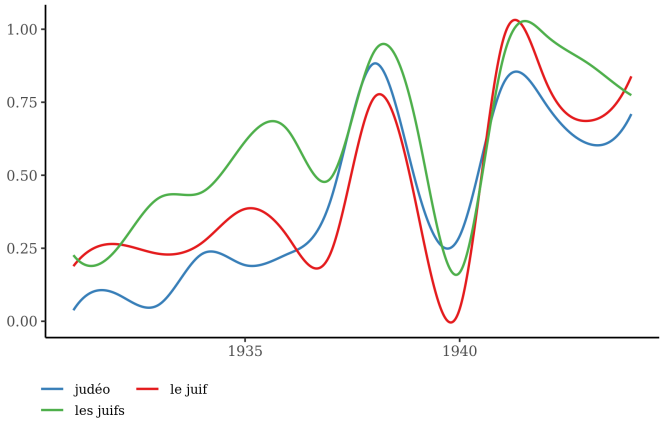

FIGURE 6 – Occurrence des vocables antisémites dans l'hebdomadaire « Je suis partout » (Gallica-presse; recherche par document).

43. Pour ce faire, il faut choisir le corpus « Recherche par titre de presse », et sélectionner dans le menu déroulant le ou les titres qu'on veut.
44. Pierre-Marie DIOUDONNAT. *«Je suis partout» : 1930-1944 : les maurrassiens devant la tentation fasciste*. La table ronde, 1973.
45. Ce préfixe est invariablement associé à un thème antisémite. On trouve surtout « judéo-marxiste », mais aussi « judéo-bolchevique », « judéo-germanique », « judéo-américain », et même « judéo-antifasciste »...

3.3 La disparition d'un syntagme : révélateur de la censure ?

Si l'outil a pour but primaire d'estimer la présence d'un syntagme à une période donnée, il est parfois plus intéressant de mettre au jour les absences, en particulier les disparitions brutales, que l'on peut interpréter comme la marque d'une censure politique [46]. A cet égard, les sobriquets donnés aux deux empereurs français sont révélateurs.
Étudions par exemple le syntagme « Buonaparte », nom de naissance de Napoléon I^{er}, utilisé par les royalistes pour ses sonorités étrangères, afin de souligner son manque de légitimité dynastique. On observe sa nette disparition entre le 18 brumaire et 1815 (Fig. 7a), cohérente avec l'hypothèse d'une censure de la presse par le pouvoir. Plus frappant, si l'on passe à la résolution par mois en se restreignant à la fin du règne (Fig. 7b, on peut appréhender, au mois près, la présence de Napoléon I^{er} au pouvoir : la forte augmentation des occurrences du terme « Buonaparte » commence véritablement en avril 1814, mois de sa première abdication, et s'interrompt pour les mois d'avril, mai et juin 1815, qui correspondent aux « Cent-jours ». La poignée d'occurrences de janvier 1814 [47] correspond principalement à des documents datés seulement à l'année près, référencés par défaut en janvier. Concernant Napoléon III, on observe substantiellement la même chose avec son surnom « Napoléon le petit », inventé par Victor Hugo (graphique non présenté ici). Cependant, la réapparition se fait ici deux ans avant la chute de l'Empereur. Cela peut s'expliquer par la libéralisation de la presse par la loi du 11 mai 1868, qui supprime l'autorisation préalable pour les journaux et allège ainsi la censure.

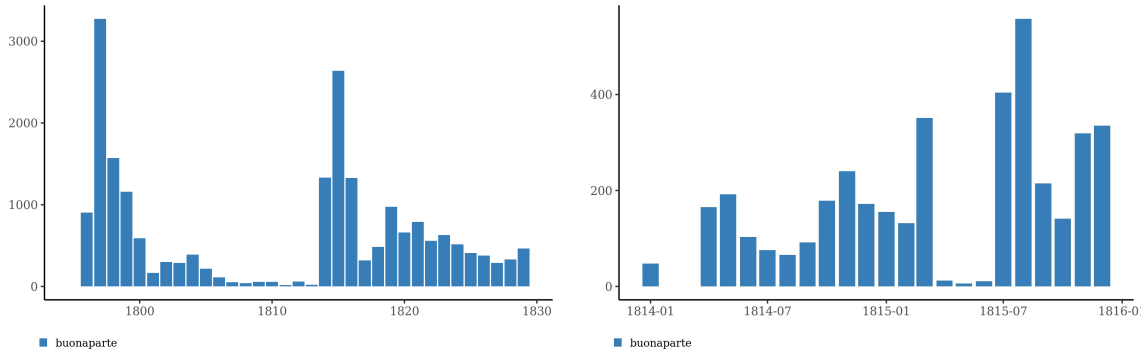

FIGURE 7 – Occurrences de « Buonaparte » dans la presse française, mode ngram, résolution année (gauche) et mois (droite).

3.4 Au-delà des tendances, l'étude de la phénoménologie des courbes

En lexicométrie, les séries temporelles sont le plus souvent analysées en termes d'augmentation et de diminution, ou encore de pics et de creux. Toutefois, la phénoménologie précise des courbes, – c'est à dire, non pas l'augmentation d'une courbe, mais la façon dont elle augmente – peut être instructive. En particulier, le profil d'une courbe ascendante nous renseigne sur l'origine de du phénomène de croissance : une courbe lisse, même exponentielle, suggère un emballement endogène (« par en-bas »), tandis qu'une augmentation brutale, discontinue et dépourvue de signes précurseurs, correspond plutôt à un choc exogène (un événement extérieur provoque la flambée du syntagme en question : voir Deschâtres et Sornette, 2004 [48] ou Deschâtres et Sornette, 2005 [49]).

On peut ainsi distinguer les carrières d'hommes politiques au profil « endogène », construites sur une mobilisation de l'opinion (par en-bas), des carrières « exogènes », où un individu est brusquement propulsé sur le devant de la scène. Les élections à la présidence de la République sous la IIIe République offrent, à ce titre, un exemple intéressant. A la

46. MICHEL et al., « Quantitative Analysis of Culture Using Millions of Digitized Books ».
47. Disponibles ici
48. Didier SORNETTE et al. « Endogenous versus exogenous shocks in complex networks : An empirical test using book sale rankings ». *Physical Review Letters* 93.22 (2004), p. 228701.
49. Fabrice DESCHATRES et Didier SORNETTE. « Dynamics of book sales : Endogenous versus exogenous shocks in complex networks ». *Physical Review E* 72.1 (2005), p. 016112.

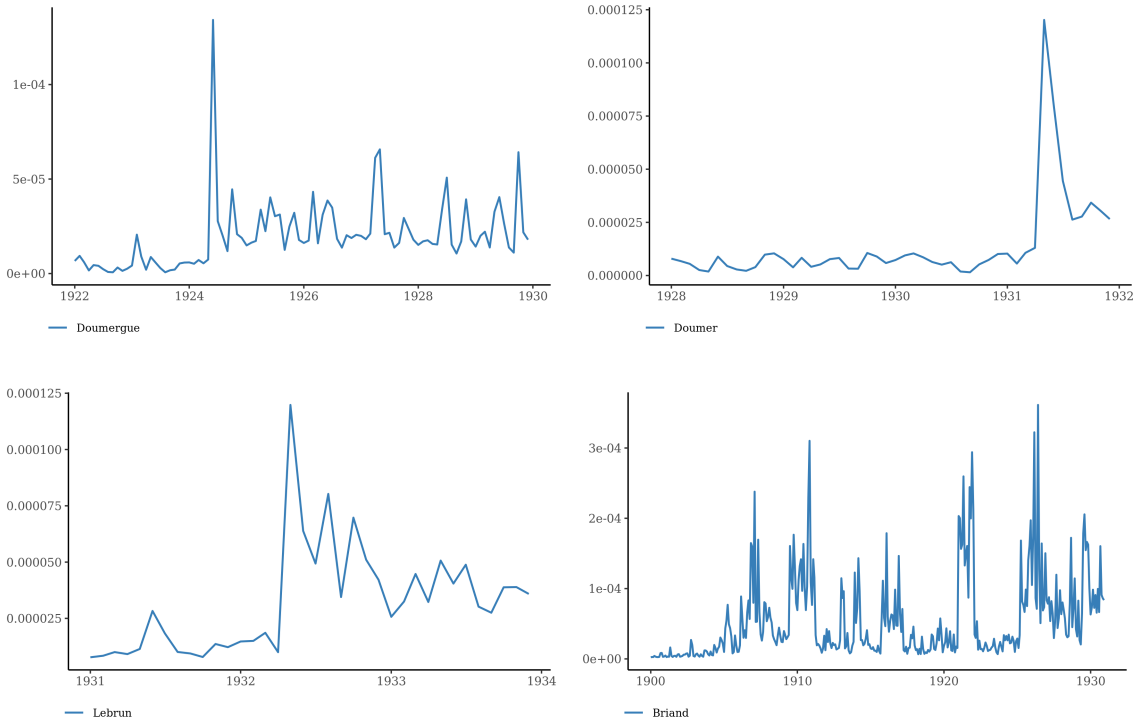

FIGURE 8 – Emballements exogènes dans les mentions de présidents de la République et du Conseil (recherche par n-gramme; résolution mensuelle).

suite de la Constitution Grévy, ce type d'élections amène généralement des « seconds couteaux » à la tête de l'Etat; Clemenceau affirmant même « voter pour le plus bête ». Dans Gallicagram, les mentions des trois derniers présidents de la République (Doumergue, Doumer et Lebrun), augmentent brusquement au mois de leur élection (Fig. 8). Celle de Pierre Laval suit le même profil en avril 1942, son retour à la tête du gouvernement étant imposé par l'occupant en dehors des logiques de popularité.

A l'inverse, les courbes de personnages que l'on pourrait qualifier par anachronie de « populistes » (le général Boulanger, Paul Déroulède, Jacques Doriot, le général de La Rocque) offrent des profils moins discontinus (Fig. 9). L'emballement est toujours précédé de frémissements, signes d'une dynamique endogène dans l'opinion publique. Il en va de même pour certains personnages très populaires pendant la période révolutionnaire, comme Marat et Danton (non montré ici). D'autres offrent des profils plus complexes : Aristide Briand (Fig. 8d) présente d'abord un emballement endogène vers 1904-1907, année où il devient rapporteur de la loi de Séparation des Eglises et de l'Etat avant d'entrer au gouvernement. Ensuite, on observe de multiples chocs exogènes, qui correspondent strictement à ses nominations à la présidence du Conseil et au Quai d'Orsay. On pourrait l'interpréter, quelque peu hâtivement, comme le signe d'une carrière d'abord consacrée à la défense d'idées radicales, puis à la conquête et à l'exercice du pouvoir. En d'autres termes, la transformation d'un homme de convictions en professionnel de la politique.

De façon plus spéculative, cette méthode peut aider à retracer la genèse d'un bouleversement politique comme une révolution : le profil de certains mots peut arguer pour une dynamique exogène, la révolution étant déclenchée par un événement extérieur, contingent, et donc imprévisible. Au contraire, un profil endogène pointerait vers une lente progression des idées dans l'opinion, qui aurait donc laissé des signes avant-coureurs. La Révolution de 1848, idéale du fait de l'abondance du corpus durant cette décennie, offre un tableau contrasté (Fig. 10). Si la plupart des mots présentent une discontinuité en février 1848 (en particulier « république », Fig. 10a), d'autres montrent des signes avant-coureurs frappants : « démocratie » (10b), « réformes » (10c), « corruption » (Fig. 10d). Les mots « misère » et « famine » (Fig.

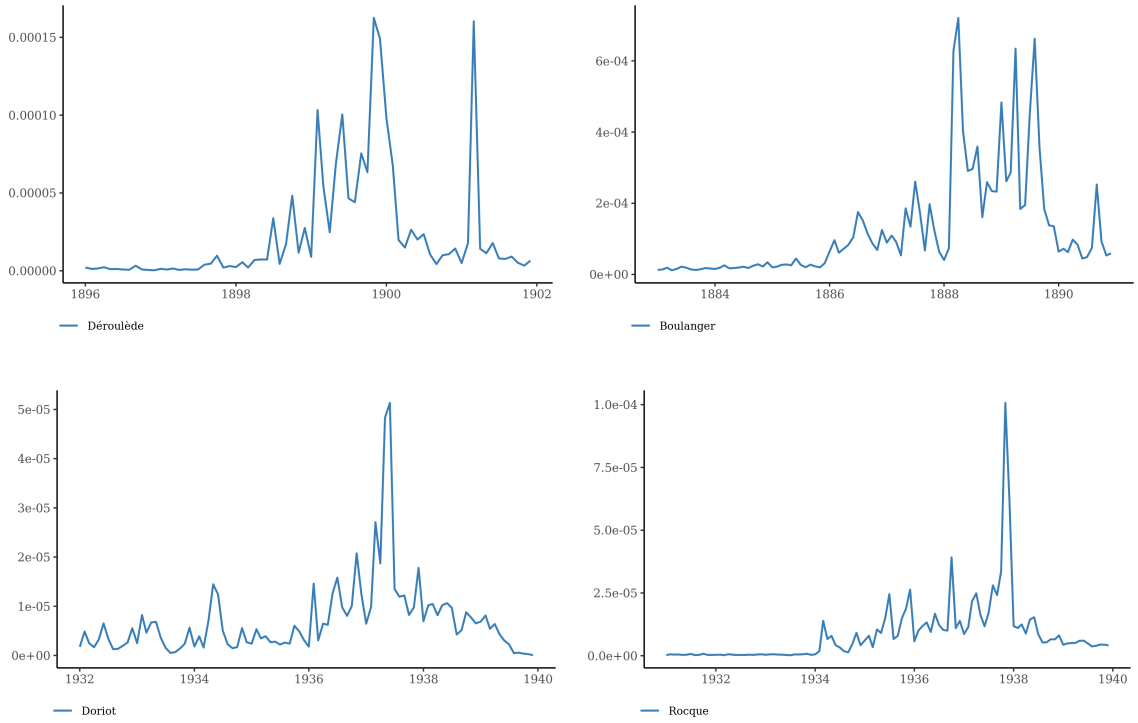

FIGURE 9 – Emballements endogènes dans les mentions de figures politiques d'extrême droite (recherche par n-gramme; résolution mensuelle).

10e) s'emballent entre 1844 et 1847 avant de passer au second plan pendant la révolution. Le mot « socialisme », quant à lui, présente un emballement nettement endogène à partir de février 1848 (Fig. 10f), comme si la révolution n'avait pas imposé le concept dans la culture française, mais seulement créé les conditions de sa diffusion dans la population.

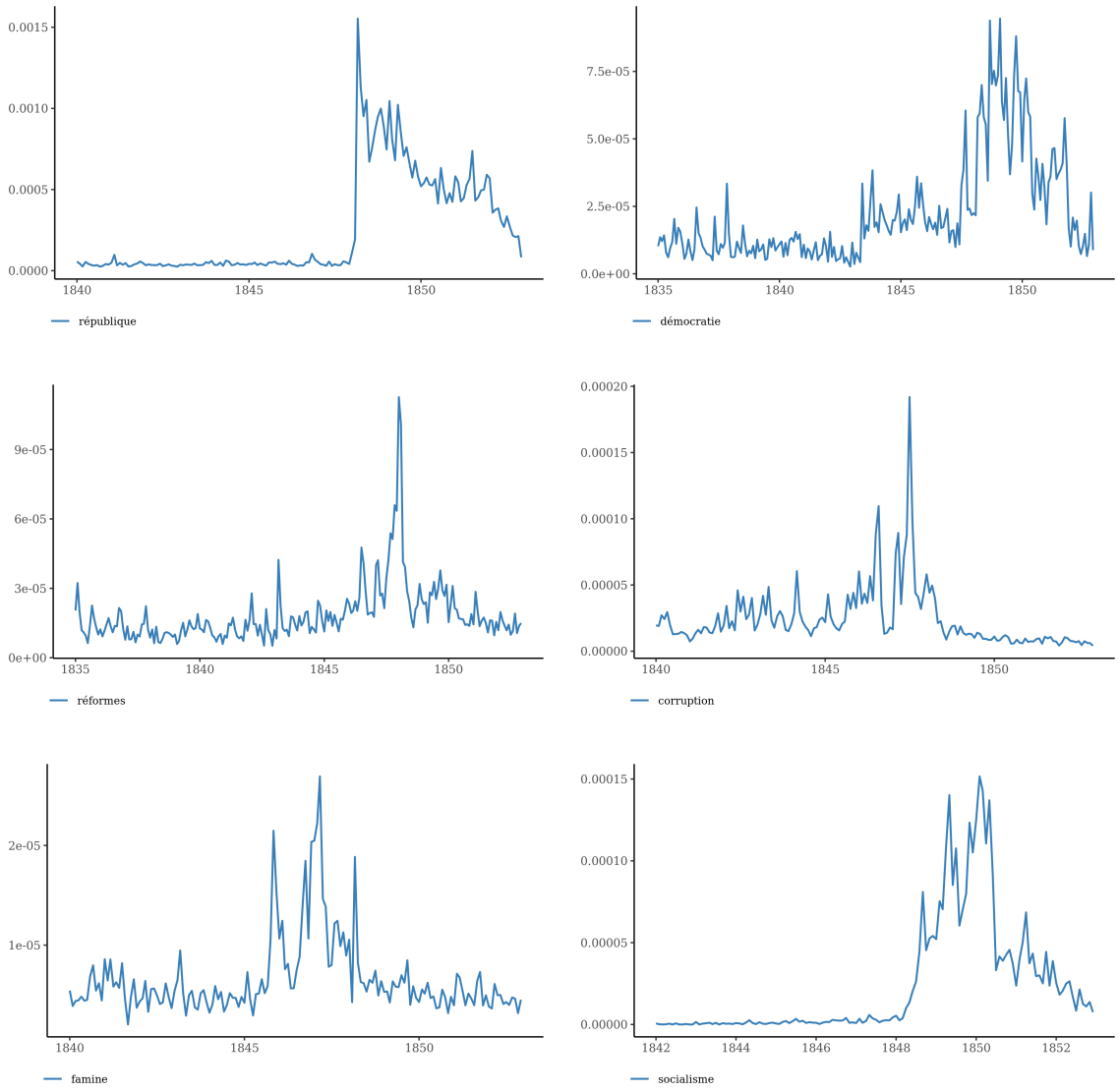

FIGURE 10 – Évolution de syntagmes autour de la Révolution de 1848 (recherche par n-gramme ; résolution mensuelle).

3.5 Gallicapresse : derrière les courbes, la structure des données

Un autre logiciel baptisé Gallicapresse [50], permet d'analyser la structure des données d'une recherche dans Gallicagram. Il décrit la composition des résultats de recherche selon 4 critères : le titre de presse, la périodicité des documents, la ville de publication des numéros de presse et le classement thématique des journaux dans la bibliothèque (classement de Dewey). Nous ne présenterons ici qu'une des options proposées par le logiciel.

L'étude géographique du boulangisme a déjà donné lieu à de nombreuses monographies [51] dont la plus connue demeure encore celle d'André Siegfried dans son *Tableau politique de la France de l'Ouest* (1913). L'historien qui entreprend une histoire du boulangisme à l'échelle nationale doit pour l'heure soit partir d'études locales, et donc nécessairement partielles, soit se référer aux résultats des élections législatives de 1889. Or celles-ci interviennent à un moment stratégique pour le camp républicain, alors même que le boulangisme est déjà en déroute, Boulanger s'étant alors exilé en Belgique. Les panoramas géographiques de la France boulangiste fondés sur les résultats des urnes sont donc peu satisfaisants pour saisir la dynamique même du phénomène. La carte du boulangisme générée par Gallicapresse (Fig. 11) offre une autre représentation de la diffusion géographique du mouvement, [52]. Elle confirme l'intensité du phénomène à Paris ainsi que dans sa région et fait émerger sa prégnance dans le Sud-Ouest de la France ainsi qu'en Auvergne et en Limousin. Au contraire, l'Est de la France, le Nord et la Bretagne semblent avoir largement échappé au phénomène.

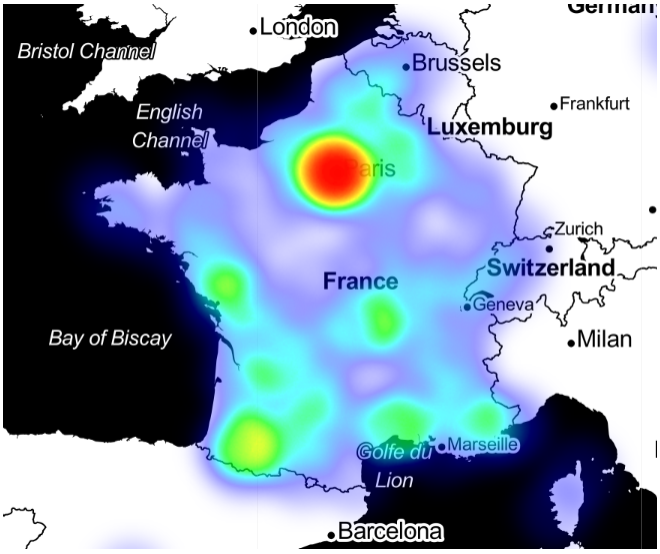

FIGURE 11 – Répartition géographique des mentions « Général Boulanger » dans la presse française, 1885-1889.

3.6 Gallicanet : un réseau social du passé

Gallicanet est un algorithme générant des graphes de réseau à partir de la proximité d'apparition d'une liste de mots dans un corpus numérique. A partir d'une liste de mots (ou de noms propres) entrée par l'utilisateur, le logiciel établit toutes les combinaisons possibles deux-à-deux et lance la recherche par proximité dans le corpus. Ce graphe est dit « orienté » puisque chaque cooccurrence est pondérée distinctement par la fréquence d'apparition de l'un des deux termes recherché. La notice précise la méthode adoptée et les calculs effectués.

Gallicanet peut être utilisé en histoire contemporaine pour reconstituer les cercles de sociabilité. A partir de la seule proximité des noms de la liste (distance en nombre de mots) dans les textes d'archive (presse de Gallica) ou dans les textes scientifiques (Cairn), le logiciel dessine des graphes de réseau qui peuvent permettre d'approcher la réalité sociale

50. Disponible ici : https://shiny.ens-paris-saclay.fr/app/gallicapresse
51. Jean GARRIGUES. *Le général Boulanger*. Editions Olivier Orban, 1991, p. 378.
52. Voir l'excellent site de Frédéric Salmon : geoelections.free.fr.

d'une époque. A partir d'une liste de 877 écrivains de l'entre-deux guerres [53], le logiciel a produit un graphe de réseau centré sur les grands acteurs du champ littéraire de l'époque : Paul Morand, Paul Valéry, François Mauriac, Georges Duhamel et André Gide. L'épaisseur des liens figure leur intensité (valeur de l'indicateur polarisé), tandis que le diamètre des noeuds représente l'importance de chaque personnage dans le champ (nombre de liens partant de ce noeud). A l'extrême Est du graphe, apparaît une poignée d'écrivains : Aragon, Eluard, Queneau, Desnos, Leiris, Plisnier ou encore René Char (Fig. 12). Il s'agit des auteurs d'avant-garde. Évoluant à distance des écrivains académiques dont ils méprisent à la fois le style et le conventionnalisme bourgeois, ils cultivent et revendiquent leur marginalité. Le logiciel, à partir d'une analyse purement quantitative, restitue parfaitement cette distance.

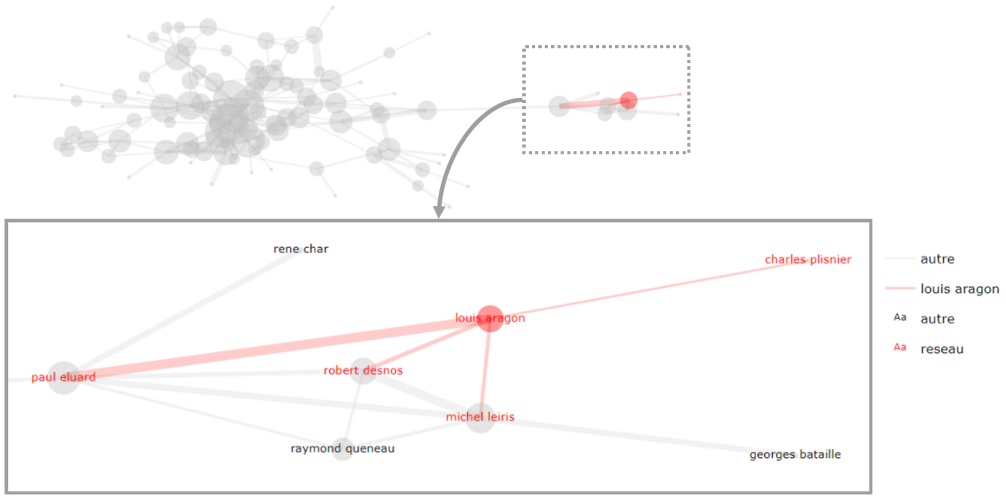

FIGURE 12 – Les avant-gardes (centré sur Louis Aragon) à l'extrême Est d'un réseau d'écrivains de l'entre-deux guerres généré par Gallicanet.

avec une liste de 570 personnages publics de l'Occupation tirée du Dictionnaire de la Collaboration de François Broche. [54] Si l'on centre le graphique sur Robert Brasillach (Nord du réseau), on retrouve autour de lui la rédaction de *Je suis partout* et à proximité immédiate (un noeud de distance), le tout-Paris collaborationniste organisé autour d'Otto Abetz, Fernand de Brinon et Marcel Déat (Fig. 13).

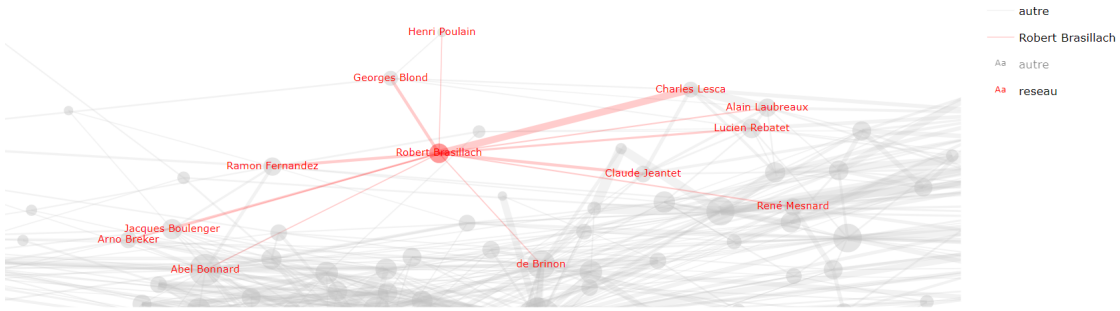

FIGURE 13 – La sociabilité de Robert Brasillach de 1940 à 1944, au nord de la cartographie sociale de l'Occupation générée par Gallicanet.

53. Données wikidata.org et data.bnf.fr. Ecrivains vivants en 1940 ayant publié au moins 6 ouvrages avant cette date.
54. François BROCHE. *Dictionnaire de la Collaboration*. Paris : Belin, oct. 2014.

Références

Aiden, Erez et Jean-Baptiste Michel. *Uncharted : Big Data as a Lens on Human Culture*. New York : Riverhead Books, 2013.

Bigot, Grégoire. « Impératifs politiques du droit privé : le divorce sur simple allégation d'incompatibilité d'humeur ou de caractère"(1792-1804) ». *Clio@ Themis* 3 (2010).

Broche, François. *Dictionnaire de la Collaboration*. Paris : Belin, oct. 2014.

Chateauraynaud, Francis et Josquin Debaz. *Prodiges et vertiges de la lexicométrie*. Sous la dir. d'Hypotheses. 2010. URL : https://socioargu.hypotheses.org/1963.

Deschatres, Fabrice et Didier Sornette. « Dynamics of book sales : Endogenous versus exogenous shocks in complex networks ». *Physical Review E* 72.1 (2005), p. 016112.

Dioudonnat, Pierre-Marie. *«Je suis partout» : 1930-1944 : les maurrassiens devant la tentation fasciste*. La table ronde, 1973.

Dumazedier, Joffre. *Vers une civilisation du loisir ?* Éditions du Seuil, 1962.

Garrigues, Jean. *Le général Boulanger*. Editions Olivier Orban, 1991, p. 378.

Greenfield, Patricia M. « The changing psychology of culture from 1800 through 2000 ». *Psychological science* 24.9 (2013), p. 1722-1731.

Héran, François. « Les mots de la démographie des origines à nos jours : une exploration numérique ». *Population* Vol. 70.3 (2015), p. 525-566.

Houte, Arnaud-Dominique. « Le temps de l'auto-stop ». *20 21. Revue d'histoire* 148.4 (2020), p. 3-15.

James, Ryan et Andrew Weiss. « An assessment of Google Books' metadata ». *Journal of Library Metadata* 12.1 (2012), p. 15-22.

Jeanneney, J. N. *Quand Google défie l'Europe : Plaidoyer pour un sursaut*. 1001st edition. Paris : 1001 nuits, jan. 2005.

Leroy Ladurie, Emmanuel. *Le territoire de l'historien*. 1973, p. 11-14.

Michel, Jean-Baptiste et al. « Quantitative Analysis of Culture Using Millions of Digitized Books ». *Science* 331.6014 (jan. 2011), p. 176-182.

Mounier, Pierre. « Les Humanités numériques, gadget ou progrès ? » *Revue du Crieur* 7.2 (2017), p. 144-159.

Perrot, Michelle et Patrick Fridenson. « Charles Tilly et la France ». *Le Mouvement Social* 4 (2008), p. 143-145.

Ruiz, Emilien. *Google labs Books Ngram Viewer : un nouvel outil pour les historiens ?* Sous la dir. de boiteaoutils.info. 2010. URL : https://boiteaoutils.info/2010/12/google-labs-books-ngram-viewer-un/.

Shorter, Edward et Charles Tilly. *Strikes in France 1830–1968*. 1St Edition. London, New York : Cambridge University Press, août 1974.

— « The shape of strikes in France, 1830–1960 ». *Comparative Studies in Society and History* 13.1 (1971), p. 60-86.

Shorter, Edward L et Charles Tilly. « Le déclin de la grève violente en France de 1890 à 1935 ». *Le mouvement social* (1971), p. 95-118.

Sornette, Didier et al. « Endogenous versus exogenous shocks in complex networks : An empirical test using book sale rankings ». *Physical Review Letters* 93.22 (2004), p. 228701.

Tartakowsky, Danielle. « Manifestations ouvrières et théories de la violence : 1919-1934 ». *Cultures & Conflits* 09-10 (1993).

Younes, Nadja. « State-of-the-art Research Using the Google Books Ngram Viewer : Improving the Method and Investigating Cultural Change » (2019).

Zipf, George Kingsley. « Relative frequency as a determinant of phonetic change ». *Harvard studies in classical philology* 40 (1929), p. 1-95.

Appendices

A Notice d'utilisation de Gallicagram

Ce document est la notice attachée à l'outil Gallicagram. Elle figure sur le site https://shiny.ens-paris-saclay.fr/app/gallicagram, sous l'onglet « Notice ».

— Gallicagram est un programme représentant graphiquement l'évolution au cours du temps de la fréquence d'apparition d'un ou plusieurs syntagmes dans les corpus numérisés de Gallica et de beaucoup d'autres bibliothèques.
— Développé par Benjamin Azoulay et Benoît de Courson, il est intégralement rédigé en langage R et présente une interface graphique interactive Shiny.
— Les données produites sont téléchargeables par l'utilisateur. Le code source de Gallicagram est libre d'accès et de droits.
— L'analyse de la structure des recherches dans le corpus de presse de Gallica peut être réalisée dans Gallicapresse.

Corpus

— Gallicagram a accès à de nombreuses bibliothèques. Quelle que soit la bibliothèque choisie, le corpus est circonscrit aux documents numérisés et océrisés et rédigés dans la langue choisie par l'utilisateur.
— Gallicagram permet aussi d'explorer des sous-corpus de Gallica totalement accessibles et bien documentés. De nombreux graphiques renseignant sur la structure des corpus de presse et de livres de Gallica sont présentés dans les onglets « Corpus de presse » et « Corpus de livres » de Gallicagram.

Options de recherche

— Gallicagram extrait des valeurs distinctes selon le mode de recherche sélectionné par l'utilisateur. En raison de l'architecture des moteurs de recherche propre à chaque bibliothèque, il est rare que plusieurs modes de recherche soient disponibles pour un même corpus. Les corpus de Gallica sont ceux qui présentent la plus grande diversité de modes de recherche.
— 4 modes de recherche sont proposés dans Gallicagram :
  — la recherche par document ($D(s,i)$) compte, pour chaque sous-période, le nombre de documents du corpus présentant au moins une occurrence du syntagme recherché ;
  — la recherche par page ($P(s,i)$) compte, pour chaque sous-période, le nombre de pages du corpus présentant au moins une occurrence du syntagme recherché ;
  — la recherche par article (article) compte, pour chaque sous-période, le nombre d'articles du corpus de presse présentant au moins une occurrence du syntagme recherché ;
  — la recherche par n-gramme($M(s,i)$) compte, pour chaque sous-période, le nombre total d'occurrences du syntagme recherché dans le corpus.

  Rapportés à leurs dénominateurs respectifs, cela donne trois indicateurs différents pour une sous-période i donnée (année ou mois). On notera s ce syntagme, $d_j, j \in \{1, ..., n_d\}$ les documents du corpus, $p_j, j \in \{1, ..., n_p\}$ ses pages et $m_j, j \in \{1, ..., n_m\}$ ses mots (n_d, n_p et n_m sont donc respectivement le nombre total de documents, pages et mots du corpus). On notera $y(\cdot)$ la fonction qui associe à une chaîne de caractères sa date. Les trois mesures sont ainsi définies :

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

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

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

— L'utilisateur peut choisir le corpus qu'il souhaite explorer.
— Il peut régler les bornes chronologiques de sa recherche.
— Pour certains corpus, il peut choisir la résolution (mensuelle ou annuelle) avec laquelle les résultats seront affichés.
— Pour chaque mode de recherche et chaque corpus, Gallicagram extrait aussi le volume de la base de données correspondante (nombre total de documents, de pages, d'articles, ou de n-grammes pour chaque sous-période).

Syntaxe de recherche

— L'utilisateur peut chercher un syntagme unique (ex. Clemenceau).
— Il peut aussi comparer les évolutions respectives de deux syntagmes concurrents en les séparant par une esperluette "&" (ex. Georges Clemenceau&Aristide Briand).
— Il peut effectuer une recherche conditionnelle de forme OU en utilisant le signe "+" (ex. juif+juive). Il s'agit d'un "ou" inclusif qui renverra tous les numéros contenant les termes séparés par un "+". La recherche dénombre des syntagmes exacts et isolés. Ainsi, entrer le mot "juif" ne permettra pas d'obtenir les résultats correspondant à son pluriel : "juifs". La recherche conditionnelle OU avec "+" permet d'intégrer ces résultats.
— Il peut enfin chercher des cooccurrences dans les corpus de Gallica à l'aide de l'opérateur "*" (ex. universel*nation). Une case apparait alors à côté du champ de la requête pour définir la distance maximale entre les termes recherchés (en nombre de mots).
— Ces trois options de recherche sont cumulables (ex. juif+juive+judéo&calviniste+huguenot+parpaillot ; ex. universel*nation+universel*patrie&étranger*ennemi). "&" est prioritaire sur "*" qui est prioritaire sur "+". Ainsi a*b+c*d&e = [(a*b)+(c*d)]&e
— La recherche n'est pas sensible à la casse (case insensitive).

Options de visualisation

— L'utilisateur peut :
  — isoler certaines recherches dans le visualiseur en cliquant sur la légende des courbes qu'il souhaite faire disparaître ;
  — effectuer des zooms sur le graphique et afficher la valeur précise de chaque point de la courbe en y positionnant la souris ;
  — afficher la distribution chronologique des documents composant la base de données sur la période qu'il a choisie ;
  — comparer toutes les recherches effectuées au cours de sa session à l'intérieur d'un seul graphique ;
  — accéder à la recherche correspondante (syntagme, corpus, sous-période) sur le site de la bibliothèque explorée afin d'accéder au corpus sous-jacent à la recherche ;
  — normaliser les valeurs ;
  — lisser les courbes affichées (type loess) ;
  — comparer l'évolution de deux syntagmes par soustraction ;
  — observer les corrélations entre les syntagmes recherchés ;
  — observer les corrélations pour un même terme entre les différents modes de recherche ;
  — télécharger les graphes et les données du visualiseur ainsi que les données de la totalité de la session.

Traitements

— Le traitement des données extraites consiste au calcul pour chaque sous-séquence temporelle de la fréquence d'apparition du terme défini par l'utilisateur. Cette fréquence est le rapport des deux variables extraites (le nombre de résultats et le volume de la base).
— Le graphique présente cette fréquence en ordonnées et le temps en abscisses selon l'échelle sélectionnée. La courbe qu'il affiche relie les points calculés par l'ordinateur.

Conception et précautions d'usage

— Toutes les informations nécessaires à la bonne utilisation de Gallicagram sont indiquées dans l'article de recherche associé à ce logiciel.

B Données et traitements

1. Corpus : Presse/Gallica
   (a) Mode de recherche : Par document
   - Délimitation du corpus : Il rassemble tous les documents en langue française de type « fascicule », datés, numérisés et océrisés avec une fiabilité supérieure ou égale à 50%, disponibles dans Gallica et consultables en ligne [55].
   - Formulation de la recherche : Gallicagram interroge l'API de recherche de Gallica pour chaque sous-période selon l'échelle choisie (mois ou année) sous la forme d'une recherche exacte à l'intérieur du corpus précédemment délimité. La recherche n'est pas sensible à la casse. La ponctuation est ignorée. Un syntagme n'a pas de limite de longueur. La comparaison des résultats pour plusieurs syntagmes s'effectue à l'aide de l'opérateur « & » placé entre chaque syntagme étudié. Le logiciel représente alors autant de courbes que de syntagmes séparés par des esperluettes. La recherche « OU » s'effectue à l'aide de l'opérateur « + » et dénombre chaque document du corpus figurant au moins l'un des syntagmes séparés par un « + ». La recherche par cooccurrences permet de chercher le nombre de documents figurant au moins une fois l'occurrence de deux syntagmes à une distance (en mots) inférieure ou égale à celle déterminée par l'utilisateur. Elle s'effectue à l'aide de l'opérateur « * ». L'ordre de priorité des opérateurs est le suivant : « & » est prioritaire sur « * », qui est prioritaire sur « + ».
   - Calcul de l'indicateur : Un indicateur est calculé pour chaque sous-période (mois ou année) et chaque syntagme ou groupe de syntagmes séparé par l'opérateur « & ». Le détail du calcul effectué ($V(s,i)$) figure en section **2.2**. Le nombre de documents peuplant le corpus scruté pour chaque sous période (dénominateur) est une constante. Il est stocké hors ligne et mis à jour régulièrement. Cela permet de réduire considérablement la durée de téléchargement des données.

   (b) Mode de recherche : Par page
   - Délimitation du corpus : Il s'agit du corpus de presse tel que précédemment défini en *1.(a)*. Son volume en nombre de pages océrisées à chaque sous période (mois et année) est enregistré hors-ligne. Gallica ne renseigne pas le volume (en nombre de pages) de chaque numéro de presse. Pour l'estimer de façon extrêmement fiable, nous avons interrogé l'API document de Gallica pour chaque numéro de presse avec la requête « espace » : « ». Ainsi, toutes les pages océrisées sont dénombrées.
   - Formulation de la recherche : Gallicagram constitue d'abord un rapport de recherche grâce à l'API de recherche de Gallica. Y figure l'adresse url de chaque document présentant au moins une occurrence du terme recherché. Pour des raisons de puissance machine et de temps de traitement, seules les recherches présentant moins de 5 000 résultats et un seul syntagme sont acceptées. Les modes « OU », « ET » ainsi que la recherche par cooccurrences sont indisponibles. Gallicagram utilise ensuite l'API document de Gallica pour extraire le nombre de pages où figure au moins une fois le syntagme recherché et ce pour chaque document listé dans le rapport de recherche généré.
   - Calcul de l'indicateur : Ce nombre est sommé pour chaque sous-période, puis rapporté au volume total du corpus (en nombre de pages océrisées) à chaque sous-période selon la formule indiquée en **2.2** ($P(s,i)$).

   (c) Mode de recherche : Par n-gramme
   - Délimitation du corpus : Pour constituer ce corpus, nous avons extrait tous les numéros de presse au format texte accessibles par l'API texte brut de Gallica et dans RetroNews. La base de données prétraitée est entièrement hors-ligne. Elle contient uniquement les documents mis en ligne avant le premier mars 2021.
   - Formulation de la recherche : L'utilisateur peut effectuer une recherche sur un syntagme composé de un à cinq membres (du monogramme au pentagramme). Gallicagram récupère alors le nombre d'occurrences du n-gramme à chaque sous-période (mois ou année). Ici, la fonction « OU » somme les fréquences retournées pour chaque n-gramme. La fonction « ET » affiche simultanément des recherches distinctes. La recherche par cooccurrences n'est pas disponible.

55. Lien vers ce corpus dans Gallica

— Calcul de l'indicateur : Le nombre d'occurrences du n-gramme à chaque sous-période est comparé à la somme des occurrences de tous les n-grammes référencés pour cette sous-période selon le calcul indiqué en **2.2** ($M(s,i)$). Par exemple, la fréquence du monogramme « Clemenceau » en 1914 correspond au nombre d'occurrences de ce terme dans le corpus en 1914 rapporté au nombre total de monogrammes détectés (*i.e.* de mots) dans le corpus de presse en 1914 (volume du corpus). Dans le cas où l'utilisateur voudrait cumuler les résultats de deux syntagmes de longueur différente (par exemple « Georges Clemenceau + Joffre »), les fréquences d'apparition des deux syntagmes dans leurs bases respectives sont additionnées. Enfin, les n-grammes figurant moins de trois fois dans le corpus d'une année donnée ne sont pas enregistrés et leur valeur retournée dans graphe est 0.

2. Corpus : Livres/Gallica
   (a) Mode de recherche : Par document
   — Délimitation du corpus : Il rassemble tous les documents en langue française de type « monographie », datés, numérisés et océrisés avec une fiabilité supérieure ou égale à 50%, disponibles dans Gallica et consultables en ligne [56]. Pour faire concorder ce corpus avec celui utilisé lors d'une recherche par n-gramme, l'utilisateur peut le restreindre aux ouvrages libres de droits et mis en ligne avant le premier mars 2021. Cela est utile pour comparer les résultats de recherche dans les deux modes, mais réduit le corpus d'environ 20%.
   — Formulation de la recherche : Gallicagram interroge l'API de recherche de Gallica pour chaque sous-période selon l'échelle choisie (mois ou année) sous la forme d'une recherche exacte à l'intérieur du corpus précédemment délimité et selon la syntaxe décrite en *1.(a)*. Le programme récupère, pour chaque sous-période, le nombre de documents du corpus où figure au moins une occurrence du syntagme recherché.
   — Calcul de l'indicateur : Ce nombre est rapporté au nombre de total de documents dans le corpus à chaque sous-période. Le détail du calcul effectué ($V(s,i)$) figure en **2.2**. Comme pour la recherche par document dans le corpus de presse *1.(a)*, il est stocké hors-ligne et mis à jour régulièrement afin de réduire le temps de téléchargement lors de l'exécution de l'application.

   (b) Mode de recherche : Par page
   — Délimitation du corpus : Il s'agit du corpus de livres tel que précédemment défini en *2.(a)*. Son volume en nombre de pages océrisées pour chaque année est enregistré hors-ligne. Pour les livres, l'indication du nombre de pages de chaque document est présente dans les résultats de l'API de recherche de Gallica sous la balise <dc :format>.
   — Formulation de la recherche et calcul de l'indicateur : La procédure et la restriction des options de recherche sont similaires à celles décrites en *1.(b)*. La formule utilisée est détaillée en **2.2** ($P(s,i)$).

   (c) Mode de recherche : Par n-gramme
   — Délimitation du corpus : Pour constituer ce corpus, nous avons extrait tous les livres au format texte accessibles par l'API texte brut de Gallica. Les ouvrages encore sous droits d'auteur n'y sont pas disponibles. Le corpus sous-jacent dans ce mode de recherche est donc constitué du corpus de livres précédemment défini restreint aux livres libres de droits. La base de données prétraitée est entièrement hors-ligne. Elle contient uniquement les documents mis en ligne avant le premier mars 2021.
   — Formulation de la recherche et calcul de l'indicateur : La procédure est similaire à celle décrite en *1.(c)*. La formule utilisée est détaillée en **2.2** ($M(s,i)$).

3. Corpus : Titre de presse/Gallica

   Ici, le mode de recherche par n-gramme n'est pas disponible car les extractions et les traitements nécessaires sont très longs et ne peuvent être automatisés. Il est donc impossible de les effectuer à discrétion sur des données choisies par l'utilisateur.

   (a) Mode de recherche : Par document

56. Lien vers ce corpus dans Gallica

— Délimitation du corpus : L'utilisateur sélectionne autant de titres de presse ou de revues qu'il le souhaite parmi la liste des 12 450 titres en langue française, numérisés et océrisés avec une fiabilité supérieure ou égale à 50%, disponibles dans Gallica et consultables en ligne [57], qui lui est proposée. L'ensemble des numéros de presse et des revues répertorié sous les titres sélectionnés à chaque sous-période constitue le corpus scruté par Gallicagram pour ce mode de recherche. L'utilisateur peut aussi sélectionner dans Gallicagram des listes de titres de presse élaborées par des documentalistes de Gallica selon des critères thématiques ou géographiques.
— Formulation de la recherche et calcul de l'indicateur : Gallicagram interroge l'API de recherche de Gallica dans ce corpus restreint pour chaque sous période, extrayant le nombre de résultats (numérateur) ainsi que le nombre de numéros de presse dans ce sous-corpus (dénominateur) pour chaque sous-période. La formule utilisée est détaillée en **2.2** ($V(s,i)$). La syntaxe et les options de recherche sont les mêmes que celles de la recherche par document dans le corpus de presse, détaillée en *1.(a)*.

(b) Mode de recherche : Par page
— Délimitation du corpus : Il s'agit du corpus de numéros de presse restreint par l'utilisateur tel que précédemment défini en *3.(a)*. Le volume du corpus en nombre de pages océrisées pour la sous-période est extrait grâce à l'API document de Gallica lors de l'exécution du programme. Il n'est pas stocké hors ligne car il varie selon le corpus défini par l'utilisateur.
— Formulation de la recherche et calcul de l'indicateur : La procédure et la restriction des options de recherche sont similaires à celles décrites en *1.(b)*. La formule utilisée est détaillée en **2.2** ($P(s,i)$).

4. Corpus : Personnalisé/Gallica

Ici, le mode de recherche par n-gramme n'est pas disponible car les extractions et les traitements nécessaires sont très longs et ne peuvent être automatisés. Il est donc impossible de les effectuer à discrétion sur des données choisies par l'utilisateur.

(a) Mode de recherche : Par document
— Délimitation du corpus : Un corpus personnalisé prend la forme d'une liste de documents rassemblée par l'utilisateur sur le site de Gallica et exportée au format « .csv » sous forme d'un « rapport de recherche ». Sa composition est donc à discrétion du chercheur. Il doit veiller à ne sélectionner que des documents océrisés dits « disponibles en mode texte » dans le filtre de recherche de Gallica.Pour des raisons de temps de traitement, seuls les rapports de recherche présentant moins de 5 000 lignes sont acceptés.
— Formulation de la recherche : Gallicagram utilise l'API document de Gallica pour déterminer, pour chaque document listé dans le rapport de recherche, s'il figure au moins une fois le syntagme recherché. En raison des limites de cette API, la recherche est restreinte à un seul syntagme et les modes « OU », « ET » ainsi que la recherche par cooccurrences sont indisponibles.
— Calcul de l'indicateur : Les résultats de la recherche sont agrégés par sous-période (mois ou année) et comparés au volume du corpus à chaque sous-période selon la formule figurant en **2.2** ($V(s,i)$).

(b) Mode de recherche : Par page
— Délimitation du corpus : Il s'agit du corpus personnalisé par l'utilisateur tel que précédemment défini en *4.(a)*.
— Formulation de la recherche : La syntaxe et les restrictions de la recherche sont similaires à celles définies en *4.(a)*.
— Calcul de l'indicateur : Gallicagram interroge ici deux fois l'API documents de Gallica pour chaque document. La première fois avec la recherche définie par l'utilisateur pour récupérer le nombre de pages figurant au moins une fois le syntagme recherché (numérateur). La deuxième fois avec le caractère « espace » (« ») afin de dénombrer le total des pages océrisées pour ce document (dénomninateur). Les résultats sont ensuite agrégés par sous-périodes et l'indicateur calculé selon la formule figurant en **2.2** ($P(s,i)$).

57. Le nombre de titres de presse et de revues indiqué correspond à l'état d'océrisation des fonds au 1er mars 2021.

5. Corpus : Livres Ngram Viewer/Google Books
   (a) Mode de recherche : Par document
      — Délimitation du corpus : Le corpus exploité par Ngram Viewer et les nombreuses zones d'ombre qui entourent sa composition ont été décrits en **1.**
      — Formulation de la recherche : La syntaxe et les options de recherche fonctionnent ici selon le modèle défini en *1.(a)*. Seule l'échelle d'affichage annuelle est disponible. Le nombre de documents composant le corpus à chaque sous-période (dénominateur) est stocké hors-ligne. Il provient du fichier « total_counts » pour le corpus français en version 3 (17 février 2020) mis à disposition sur le site de Google Ngram Viewer. Le nombre de documents figurant au moins une fois le syntagme recherché est extrait de nos serveurs où nous avons reconstitué la base de données de Google Ngram Viewer.
      — Calcul de l'indicateur : Il résulte de la comparaison pour chaque année de la comparaison des deux dénombrements précédents selon la formule indiquée en **2.2** ($V(s,i)$).
   (b) Mode de recherche : Par n-gramme
      — Délimitation du corpus : Il s'agit du corpus exploité par Google Ngram Viewer tel que défini en *5.(a)*.
      — Formulation de la recherche : La syntaxe et les options de recherche fonctionnent ici selon le modèle définit en *5.(a)*. Ici, la fonction de recherche « OU » est autorisée, mais elle réalise la somme des fréquences correspondant aux syntagmes séparés par l'opérateur « + ». La recherche « OU » est ici sensible à la casse. La recherche « ET » demeure insensible à la casse. Ainsi, la recherche « maroc&Alger+Algérie » est équivalente à « Maroc&Alger+Algérie » mais pas à « maroc&alger+algérie ».
      — Calcul de l'indicateur : L'indicateur est celui fourni par l'API « JSON » Ngram Viewer [58]. Il calcule, pour chaque année, le rapport entre le nombre de n-grammes correspondant au syntagme et le volume du corpus en nombre de n-grammes, selon la formule indiquée en **2.2** ($M(s,i)$).

Pour que les chercheurs puissent enregistrer les résultats de leur recherche, Gallicagram permet de télécharger le graphique et les données correspondantes lors de chaque exécution. Il donne aussi accès à l'historique des données générées lors d'une session de travail. Cela permet de réaliser des traitements *ex post* sur les données, en dehors de l'application [59].

58. Pour « Joffre », 1914-1920 par exemple.

59. Les données numériques sont au format « .csv » encodés en « UTF-8 » et séparés par une virgule. Les données graphiques générées avec « Plotly » sont au format « .html ». Une capture des graphes peut être enregistrée au format « .png » à l'aide d'une fonction de téléchargement embarquée dans l'interface de Plotly

C Bases de données en n-grammes

La première base est fondée sur le corpus de livres de Gallica. Elle repose sur tous les livres numérisés et océrisés avant le premier mars 2021, libres de droits et accessibles en ligne. Nous avons extrait le texte de chaque ouvrage au moyen de l'API texte brut de Gallica. Chaque page web correspondante a été enregistrée au format « .html » au nom de son identifiant (ark). Pour chaque année, le texte de tous les ouvrages correspondant a été agrégé dans un seul fichier « .txt » et nettoyé de sorte à ne conserver que les caractères latins (suppression des autres caractères, des chiffres et de la ponctuation à l'exception du point et de l'apostrophe). Chaque fichier annuel contient donc l'agrégation de tous les textes du corpus publiés cette année-là. Le texte agrégé est ensuite divisé en phrases. Pour chaque phrase, toutes les combinaisons des n-grammes de taille comprise entre 1 et 5 sont déterminés (monogramme, bigramme, trigramme, tetragramme, pentagramme). Les n-grammes identiques sont ensuite dénombrés. Seuls les n-grammes apparaissant strictement plus de deux fois dans le corpus de l'année sont référencés dans la base de données SQL où figure leur occurrence.

La seconde base est fondée sur le corpus de presse de Gallica. Elle repose sur tous les numéros de presse et de revue numérisés et océrisés avant le premier mars 2021, libres de droits et librement accessibles sur le site de Gallica ou de Retronews. Nous avons extrait le texte de chaque numéro enregistré dans Gallica au moyen de l'API texte brut, tandis que le texte des numéros enregistrés dans Retronews a été extrait à l'aide d'un robot simulant un usage humain, ainsi capable d'accéder aux données « JavaScript ». Les fichiers « .html » ou « .txt » produits sont enregistrés au nom de l'ark référencé par Gallica. Pour chaque mois, le texte de tous les numéros correspondant a été agrégé dans un seul fichier « .txt ». Chaque fichier mensuel contient donc l'agrégation de tous les textes du corpus publiés durant cette période.
Ces fichiers ont été traités selon la même méthode que pour le corpus de livres. Il en résulte ainsi, pour chaque mois, une série de cinq matrices. Pour une recherche à l'échelle annuelle, les valeurs mensuelles sont sommées par année.