CORPUS

Corpus

24 | 2023
Les corpus numériques pour la didactique des langues : de la formation des enseignants à l'élaboration de dispositifs d'apprentissage

Gallicagram : les archives de presse sous les rotatives de la statistique textuelle

Gallicagram*: applying textual statistics to press archives*

Benoît de Courson, Benjamin Azoulay, Clara de Courson, Laurent Vanni et Étienne Brunet

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/corpus/7944
DOI : 10.4000/corpus.7944
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Benoît de Courson, Benjamin Azoulay, Clara de Courson, Laurent Vanni et Étienne Brunet, « *Gallicagram* : les archives de presse sous les rotatives de la statistique textuelle », *Corpus* [En ligne], 24 | 2023, mis en ligne le 15 janvier 2023, consulté le 14 février 2023. URL : http://journals.openedition.org/corpus/7944 ; DOI : https://doi.org/10.4000/corpus.7944

Ce document a été généré automatiquement le 14 février 2023.

Gallicagram : les archives de presse sous les rotatives de la statistique textuelle

Gallicagram*: applying textual statistics to press archives*

Benoît de Courson, Benjamin Azoulay, Clara de Courson, Laurent Vanni et Étienne Brunet

Introduction

1 Au cœur de l'hiver pandémique 2021, deux étudiants désœuvrés ont fabriqué un petit robot informatique[1]. Construit sur une poignée de lignes de code, l'ancêtre de *Gallicagram* passait en revue les documents contenant par exemple le mot *fourchette* dans *Gallica*[2] et dénombrait les résultats recensés par le site de la Bibliothèque nationale de France, année par année. Il les assortissait d'une courbe, reflétant supposément l'importance culturelle de la fourchette en France au cours du temps. Un principe enfantin, mais applicable à n'importe quel mot, concept, entité ou personnage ; une fruste machine à remonter le temps, plaisante à manipuler, parfois instructive. On le sait, *Gallica* est pour une bonne partie des chercheurs français un terrain de jeu quotidien. Un outil fondé sur un corpus aussi fiable et familier nous a d'entrée de jeu semblé rassurer cette population souvent rétive au *big data*, voire à toute forme de quantification. Mieux, la qualité des métadonnées et le caractère ouvert des données de *Gallica* s'avéraient répondre en partie aux réserves des chercheurs à l'encontre de *Ngram Viewer*[3], l'outil dominant en la matière, fondé sur le corpus de *Google Books* (Michel *et al.* 2011).

2 Encouragés par le bon accueil que nous ont fait chercheurs et bibliothèques, et avec le soutien logistique de l'ENS Paris-Saclay, nous nous sommes décidés à aller au bout des choses, et c'est ainsi que ce projet de confinement a dérapé. Nous avons téléchargé l'ensemble des documents océrisés dans *Gallica*, soit environ 300 000 livres et 3 millions de périodiques, et avons ordonné à la machine d'y compter chaque mot. En est issue

une énorme base de données, dénombrant les occurrences de chaque mot ou groupe de mots, respectivement dans les livres ou dans la presse française.

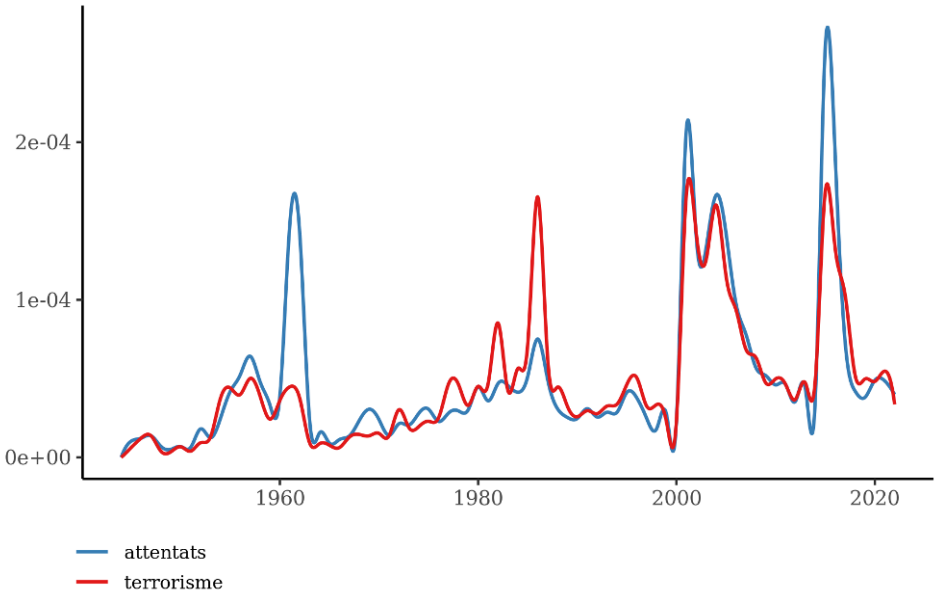
3 Mais que faire du nombre de fourchettes en 1864 ? Pour notre petit robot, c'était une révolution : non seulement nous passions de l'échelle du document à celle, plus fine, du mot, mais nous nous émancipions du pesant site de *Gallica* pour interroger directement notre propre base de données – qui a l'avantage de répondre en un instant.

4 Le droit de la propriété intellectuelle limite les ressources de *Gallica* à partir des années 1950 ; c'est pour pallier ce manque que nous avons appliqué la même méthode aux archives du quotidien *Le Monde*, constituées de plus de 3 millions d'articles impeccablement numérisés et océrisés, couvrant la période 1944-2022. À ce stade, les vastes bases de données obtenues ne nous servaient qu'à tracer (Figure 1) des courbes temporelles, usage le plus évident mais certainement pas le seul. Perdues sur le plateau de Saclay, nos fourchettes attendaient davantage.

5 Nous avons pris langue avec d'éminents chercheurs pour leur faire part de notre aventure ; Damon Mayaffre nous a adressés à ses comparses du laboratoire « Bases, Corpus, Langage » de l'Université de Nice. Leur logiciel *Hyperbase* explorait depuis longtemps de vastes corpus avec les méthodes de la statistique textuelle (Brunet 2012). Étienne Brunet, son créateur, avait d'ailleurs estimé quelques années plus tôt que *Gallica*, territoire *a priori* naturel pour le lexicomètre francophone, ne se prêtait pas au traitement statistique (Brunet 2016 : 3). Il lui reprochait notamment sa taille, plus faible que *Google Books* – point sur lequel nous divergeons, nous le verrons. Il ne s'en est pas moins enthousiasmé pour « un instrument beaucoup mieux documenté, beaucoup plus souple et bien plus fiable que *Ngram Viewer* »[4]. À sa demande, nous avons facilité l'accès externe aux données *via* une API. Ni une ni deux, avec l'aide de Laurent Vanni, Étienne Brunet programmait l'intégration de *Gallicagram* à *Hyperbase : Gallibase* était né[5].

6 Cet article présente, dans un premier temps, les données de *Gallicagram*, leurs mérites, leurs limites et l'originalité du projet. Dans une seconde partie, Étienne Brunet explique comment *Gallibase* permet d'en tirer profit par l'analyse statistique d'un grand nombre de mots.

Figure 1. Fréquence des mots *attentats* et *terrorisme* dans les archives du *Monde* (1944-2022), où l'on retrouve les grandes vagues terroristes en France, ainsi qu'une tendance générale à la hausse. La principale divergence entre les courbes est intéressante : au moment précis où ils sont le fait de l'OAS (1961-1962), *Le Monde* parle bien davantage d'*attentats* que de *terrorisme* – contrairement à la période 1954-1960, où ils sont imputables au FLN. Les tracés respectifs des courbes témoignent donc d'une appréciation différenciée des attentats, tendant à atténuer la criminalisation des acteurs français en regard des acteurs algériens



1. Gallicagram

Au commencement étaient les corpus

7 Il n'y a pas de lexicométrie sans corpus, et c'est la qualité de celui-ci qui détermine la fiabilité de l'analyse. Pour notre entreprise, le corpus se devait de répondre à deux critères : (i) posséder un certain échelonnement temporel, afin de mesurer l'évolution de l'usage d'un mot au fil du temps et (ii) comprendre un volume suffisant de mots pour pouvoir détecter un signal et non seulement du « bruit ». En sa qualité de « bibliothèque virtuelle de l'honnête homme »[6], *Gallica* répond bien à ces deux critères. En mars 2021, date de notre grande moisson, 300 000 « monographies » et 3 millions de « périodiques »[7] y avaient déjà été intégrés. Quoique très inégalement distribués au cours du temps[8], ces documents composent un corpus exploitable entre 1600 et 1940 pour les livres, et entre 1789 et 1950 pour la presse[9].

8 En termes de masse de données, ces corpus se situent à mi-chemin entre *Frantext* et *Google Books* ; *Gallicagram* permet, nous semble-t-il, de naviguer entre deux écueils. *Frantext* contient en tout et pour tout 5 000 textes, essentiellement échantillonnés en fonction du canon littéraire. *Frantext* a toutefois l'avantage de ne comprendre qu'une édition de chaque œuvre là où *Gallica* en agglomère une dizaine pour les plus fameuses (*Les Lettres persanes*, *Les Liaisons dangereuses*…). Le premier est particulièrement adapté à des « relevés » sur auteurs, *a fortiori* s'agissant d'écrivains ne posant pas de difficultés éditoriales particulières – ainsi de Rousseau ou de Marivaux ; l'entreprise s'avère plus hasardeuse pour Diderot ou de Prévost dont les œuvres, pour canoniques qu'elles soient, ne sont que très partiellement représentées sur Frantext[10]. La qualité des

données engrangées par *Frantext* est proportionnée aux dimensions réduites de la base : outre une lemmatisation impeccable et une indexation très détaillée des documents par genres textuels (romans, écrits personnels, formes brèves fictionnelles, récits de voyage, policier, contes, pamphlets...), *Frantext* permet la recherche par cooccurrence[11]. Mais pour une approche proprement lexicométrique, le faible volume de textes gêne l'analyse : les mesures de fréquences sont inévitablement bruitées, donc imprécises et instables en diachronie. Si certains[12] utilisent tout de même *Frantext* à des fins lexicométriques, c'est en général sur des termes relativement fréquents et à une échelle temporelle assez grossière. On peut donc soupçonner que ce choix est contraint. En se limitant au canon littéraire, *Frantext* laisse par ailleurs de côté la masse des écrits non passés à la postérité, mais qui n'en renseignent pas moins sur les usages scripturaux d'une période donnée. Dans ce corpus restreint, on trouvera évidemment *La Princesse de Clèves*, parangon de la nouvelle historique et galante ; mais il sera impossible de procéder à une analyse sérielle du genre, en replaçant *La Princesse* aux côtés de la riche production contemporaine du *Mercure galant*.

9 Quant à *Google Books*, *Gallica* s'en distingue d'abord par la qualité de ses métadonnées (temporelles notamment), renseignées manuellement par des documentalistes. Elle permet de dater correctement les occurrences, tandis que l'on estime que 10 % des documents de *Google Books* comportent une erreur de datation (James & Weiss 2012). Avantage décisif, *Gallica* distingue les documents par type, ce qui permet à *Gallicagram* de traiter à part les « livres » et la « presse ». Cette distinction, pour schématique qu'elle soit, garantit au corpus une composition raisonnablement stable au cours du temps. *Ngram Viewer* exploite au contraire un bloc de textes où figurent tout ensemble des romans, des essais, des journaux, de la littérature scientifique, et jusqu'à des recettes de cuisine (du moins on le présume[13]). Le problème est particulièrement criant sur la période contemporaine : « Les publications les plus récentes, qui sont aussi les plus nombreuses, n'ont pas été soumises au tri de l'histoire : c'est le tout-venant de l'édition, où pullulent les ouvrages d'information, les traités techniques et les sujets les plus divers » (Brunet 2015 : 13). En conséquence, une grande partie des évolutions révélées par ses courbes ne sont rien d'autre que des artefacts, issus d'une variation dans la composition du corpus[14]. Enfin, le classement thématique de Dewey permet de saisir l'évolution de la structure du corpus et donc apprécier si l'évolution d'un mot est due à une augmentation d'un certain type de documents ou à une véritable émergence. Par exemple, la part du corpus de livres liés à la technique se contracte considérablement à compter des années vingt. Si la fréquence du mot *machine* s'effondre à partir de la première guerre mondiale, c'est donc vraisemblablement à cause de cet effet de structure.

10 Les documents de presse ont aussi l'avantage d'enregistrer l'actualité immédiate. Nos corpus périodiques sont en conséquence très réactifs aux événements, là où les livres souffrent nécessairement d'une latence, due au temps de composition et au processus éditorial. Il en résulte des courbes nettement plus accidentées dans les corpus de presse, presque toujours échancrées au moment des guerres, et souvent bouleversées lors des changements de régime. Cette réactivité se mesure aisément : dans *Le Monde*, la fréquence du mot « terrorisme » est multipliée par 7 entre 2000 et 2001. Dans *Ngram Viewer*, elle ne bondit que d'un facteur 2,4 et culmine en fait en 2002, signe d'un délai entre les événements et leur transcription. Cette synchronicité légitime la résolution

mensuelle des corpus de presse et s'avère précieuse pour *Gallibase*, qui cherche notamment à représenter le degré auquel un mot est propre à une période donnée.

Que mesure-t-on au juste ?

11 La lexicométrie donne rapidement le vertige. On se surprend à confondre les courbes produites par l'outil avec des courbes sondagières, mesurant la prégnance d'un thème ou la force d'un courant d'opinion. Il faut alors se rappeler du principe – enfantin – de l'outil : compter des mots. *Gallicagram* ne mesure pas directement l'opinion publique, non plus que la popularité d'un personnage ; le logiciel se contente d'enregistrer la fréquence d'un terme dans un corpus donné. Mais le point de mire de la requête est nécessairement d'un autre ordre. C'est là qu'intervient le chercheur, qui entend déterminer la cause de chaque fluctuation apparaissant à l'écran. Il convient avant tout d'apprécier si cette évolution traduit une véritable variation de son usage par les scripteurs, ou simplement une transformation de la composition du corpus. *Gallica* s'y prête mieux que *Ngram Viewer* : la composition du corpus est, on l'a vu, raisonnablement stable du fait de la distinction entre livres et presse, là où *Ngram Viewer* les agglomère dans des proportions variables et opaques. La fluctuation peut aussi être causée par un artefact (voir *infra*, la « poudre *Coza* »). Si la fluctuation s'avère réelle, elle est imputable à différents processus : évolution linguistique, fait divers[15], tendance culturelle proprement dite… Encore faut-il donner au chercheur les moyens de son interprétation. *Gallicagram* inclut à cette fin, dans l'onglet « Distributions » de l'application, une description aussi fine que possible du corpus, fondée sur les métadonnées de *Gallica*. Concernant *Le Monde*, une catégorisation des articles en rubriques par *topic modelling* est en cours, et permettra d'anticiper et de corriger certains effets de structure.

Transparence et accès au corpus

12 Bien davantage, l'outil s'efforce de donner à l'utilisateur un accès aux documents exploités. *Gallica* étant un corpus ouvert et bien référencé, il est possible de renvoyer le chercheur d'une occurrence au document qui l'énonce. Dans le logiciel, un clic sur un point du graphique renvoie vers la recherche correspondante dans *Gallica*. Ainsi quand la courbe du mot *grève* tressaille en août 1888, on peut cliquer sur le pic et afficher les documents correspondants avec les occurrences dans leur contexte : ils révèlent aussitôt une grève des terrassiers, déclenchée le mois précédent. Cet accès immédiat au corpus – qui fait cruellement défaut à *Ngram Viewer* – permet aussi d'identifier les erreurs induites par le logiciel de reconnaissance optique des caractères (OCR). Noyées dans une masse de résultats, ces erreurs ne modifient d'ordinaire pas significativement les mesures. Mais pour peu que l'on s'intéresse à des syntagmes rares, ou aux premières occurrences d'une expression, cette vérification « à la source » est essentielle[16]. Elle permet surtout de repérer les fausses pistes, comme le pic d'*ivrognerie* en 1907, dû à une publicité massive pour la « poudre *Coza* »[17]. Dans le cas du *Monde*, cet accès est théoriquement possible mais limité en pratique par deux obstacles. D'une part, le moteur de recherche interne du *Monde*, qui confond la date de publication de l'article avec sa date de mise à jour. D'autre part, les droits d'auteur, qui réservent la plupart des articles aux abonnées du journal. Sauf à accepter que notre outil soit muet à partir de 1950, nous devons à l'heure qu'il est nous en accommoder.

13 Plus généralement, *Gallicagram* a fait vœu de transparence, à rebours de l'opacité de *Ngram Viewer* qui dérange tant les chercheurs. Il s'agit là d'un parti pris éthique aussi bien que scientifique. Le *big data* implique toujours un acte de foi dans les données et dans les méthodes. Pour sonder l'immensité, le chercheur renonce à une vision exhaustive de ses sources et confie les traitements à une machine, dont la mémoire est certes très supérieure à la sienne, mais qui n'a pas le sens commun. C'est peut-être la raison de l'inconfort des chercheurs en sciences sociales avec cet exercice, acte fainéant et anxiogène pour le chercheur, qui perd l'entière maîtrise de sa recherche. Loin d'une fascination béate, *Gallicagram* est fondé sur une confiance limitée dans nos propres analyses et appelle constamment l'utilisateur à la vigilance[18]. Conscient que le *big data* peut à tout moment s'égarer et induire le chercheur en erreur, *Gallicagram* cherche à lui donner les moyens d'effectuer des vérifications minimales, en particulier en évaluant la structure du corpus et en se penchant sur un échantillon d'occurrences.

Un méga-corpus, à quoi bon ?

14 Lors de son lancement, les créateurs de *Ngram Viewer* mettaient en avant, comme principal atout du logiciel, le volume de textes traités. Plus de 5 millions de livres, soit 4 % des ouvrages jamais imprimés. Convenons-en : aucune bibliothèque, physique ou numérique, ne peut rivaliser avec cette masse de textes. Est-ce à dire que la qualité de l'outil lexicométrique se mesure au volume de son corpus ? Rien de moins sûr. D'un point de vue statistique, la lexicométrie s'apparente à un sondage : on cherche à estimer la fréquence d'un mot sur une période donnée et dans un contexte de production déterminé. Cette estimation se heurte à deux obstacles : le bruit et le biais. Tout comme la marge d'erreur d'un sondage se réduit lorsqu'on augmente le nombre de participants, le bruit aura tendance à s'effacer à mesure que l'on ajoute des textes au corpus. On observe de fait, sur *Ngram Viewer* comme sur *Gallicagram*, des courbes très erratiques aux périodes de vache maigre du corpus, et beaucoup plus stables aux périodes d'abondance. Mais le biais, lui, ne décroît pas avec le volume – sauf à atteindre l'exhaustivité, évidemment chimérique. Il aurait même tendance à augmenter : en lexicométrie, ce biais est avant tout dû à la construction du corpus, et l'on devient nécessairement moins scrupuleux lorsqu'on cherche à accumuler le plus de textes possibles. S'ensuivent de piètres métadonnées, on l'a vu, et un corpus fourre-tout, peu homogène au cours du temps. D'un point de vue statistique, constituer un corpus gigantesque revient donc à s'acharner sur le bruit en oubliant le biais, c'est-à-dire à tirer avec une précision astronomique sur un point qui n'est pas le cœur de cible[19]. Il ne viendrait pas à l'idée d'un sondeur politique d'interroger cent mille personnes tout en sacrifiant sa rigueur : c'est pourtant le choix de *Google*. Sans doute faut-il y voir un biais cognitif consistant à jauger un outil en proportion du volume de données sur lequel il est assis.

15 Quelle serait alors la juste taille d'un corpus ? Les textes ont des rendements très décroissants : au fur et à mesure que l'on ajoute des textes, le bruit cesse de fausser la perspective et la loi des grands nombres fait son office. Il faudrait idéalement s'en tenir là, à un corpus de taille raisonnable, et résister à la tentation d'ajouter toujours des textes au détriment de la structure du corpus. *Google Books* se rêve sans doute en nouvelle bibliothèque d'Alexandrie, mais la lexicométrie n'en a nul besoin. Entendons-

nous : notre entreprise nécessite un large corpus. Mais une fois une masse minimale assurée[20], il convient de se concentrer sur les biais.

16 *Gallica* est sans doute trop grossier et il conviendrait d'en proposer des versions filtrées, en isolant par exemple les fictions narratives[21]. En attendant, il nous semble que son usage est déjà un net progrès par rapport à *Ngram Viewer*. Pour éprouver cette supériorité, nous avons montré ailleurs (Azoulay & Courson 2021) que la fréquence du mot *grève* dans notre corpus de presse prédisait mieux que *Ngram Viewer* le nombre réel de grèves (Figure 2a). Le corpus du *Monde* est une illustration encore plus frappante de cet argument : avec « seulement » 1,4 milliard de mots, il représente 1,5 % du corpus de *Ngram Viewer*. Et de même, il rend compte du climat social mieux que *Ngram Viewer*, qui manque les mouvements de 1947-1948, 1963 et 1968, puis s'emballe à partir de 2010. Le contrôle corpus, crucial pour les comparaisons diachroniques, triomphe ici sur la quantité.

Figure 2a. Les occurrences de grèves dans *Gallicagram* : comparaison de la fréquence de *grève(s)* dans le corpus de presse avec le nombre de grèves recensées dans Tilly & Shorter 1973

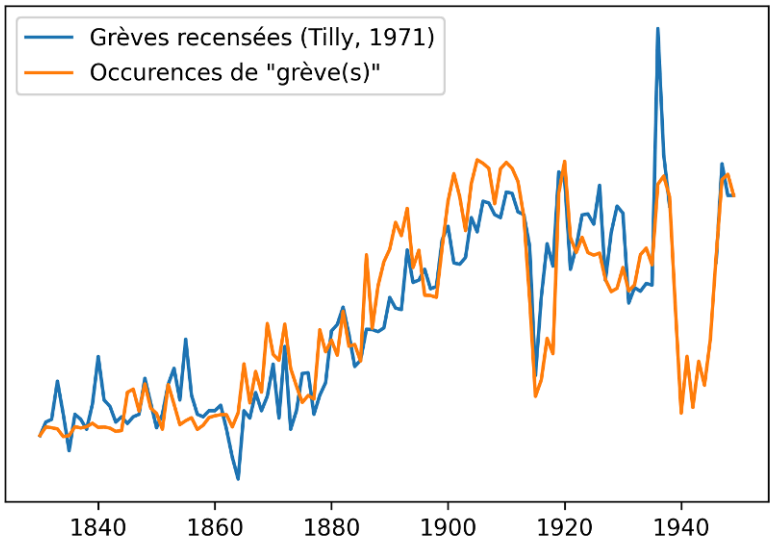
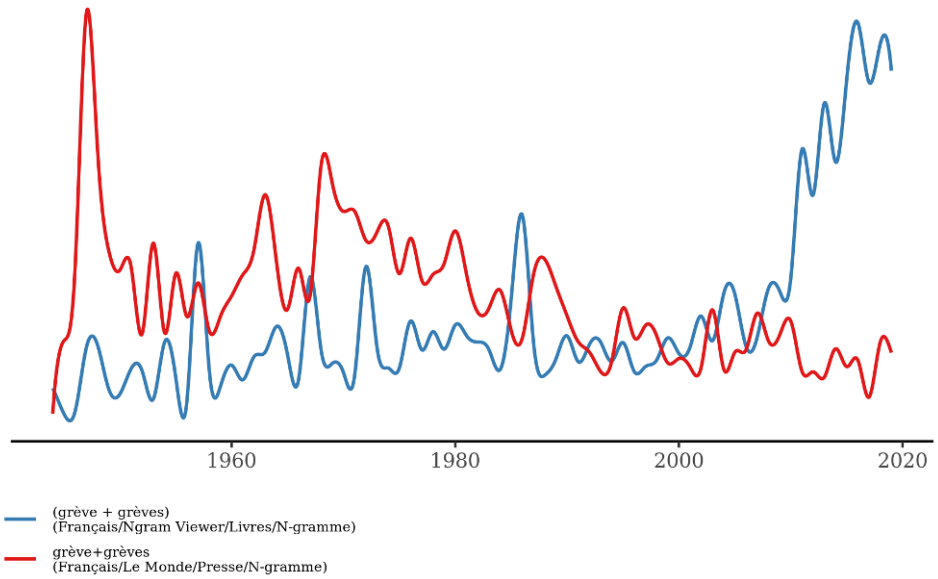


Figure 2b. Les occurrences de grèves dans *Gallicagram* : comparaison de la fréquence de *grève(s)* dans *Le Monde* et dans *Ngram Viewer*



Gare aux additions : les synonymes, ensemble ou séparément ?

17 Il est rare qu'une idée ou un fait social n'ait eu qu'une seule dénomination dans la langue. Ainsi de l'alcoolisme, désigné sous l'Ancien Régime comme « ivrognerie » ; l'émergence occurrentielle du premier vers 1860 ne suffit pas à conclure à celle d'une nouvelle problématique sociétale. Il convient de ne pas chercher un mot isolément, mais d'y adjoindre ses synonymes. *Gallicagram* fait alors face à un dilemme : mêler les mots et additionner leurs fréquences (Figure 3a), ou bien les juxtaposer et présenter une courbe par mot (Figure 3b). Sommer les fréquences revient implicitement à supposer que ces mots ont le même sens, or la synonymie est rarement exacte[22]. *Alcoolisme* désigne certes le même comportement qu' *ivrognerie*, mais le terme recatégorise le phénomène sous l'angle de la pathologie et non plus du vice, congédiant la connotation morale inhérente au désignant *ivrognerie*. La substitution d'un substantif à l'autre est donc signifiante, et la gommer induit une perte d'information. On évite cet écueil en les gardant séparés (Figure 3b), mais le graphique perd alors en clarté. Entend-on ajouter d'autres synonymes – le CNRTL en propose une dizaine ? Peine perdue : au-delà de trois mots, nos graphiques deviennent illisibles. On peut néanmoins changer de mode de visualisation, en utilisant par exemple les graphiques en « bulles » (Figure 4), capables d'accommoder jusqu'à une dizaine de mots.

Figure 3a. Fréquences des mots « alcoolisme » et « ivrognerie » additionnés, dans la presse numérisée de *Gallica*, correspondant à la requête « alcoolisme+ivrognerie » dans l'application

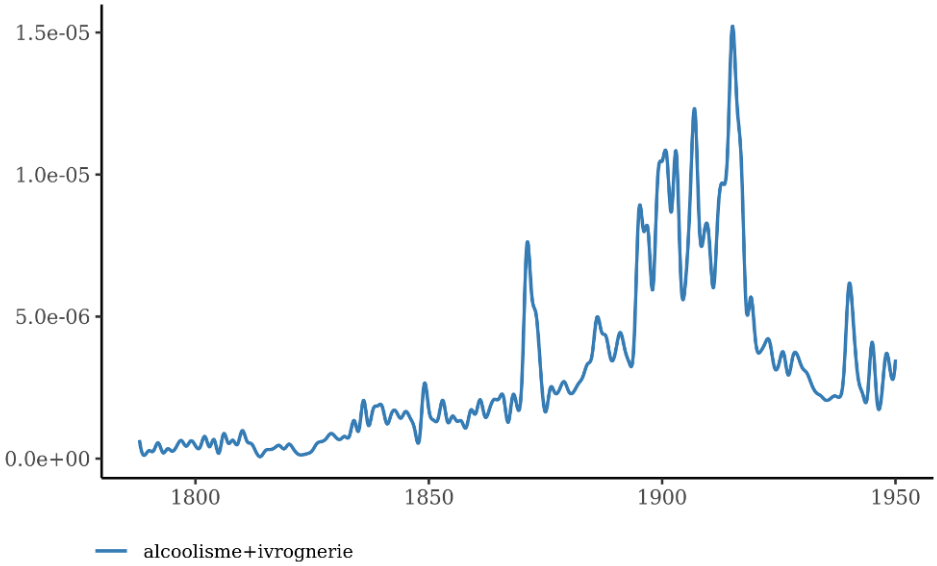


Figure 3b. Fréquences des mots « alcoolisme » et « ivrognerie » gardés séparément, dans la presse numérisée de *Gallica*, correspondant à la requête « alcoolisme&ivrognerie » dans l'application

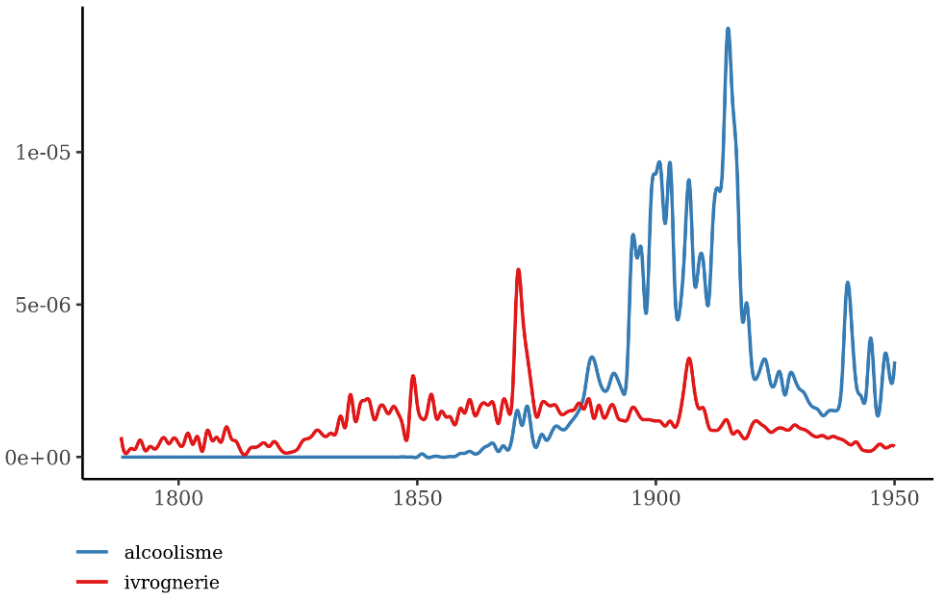
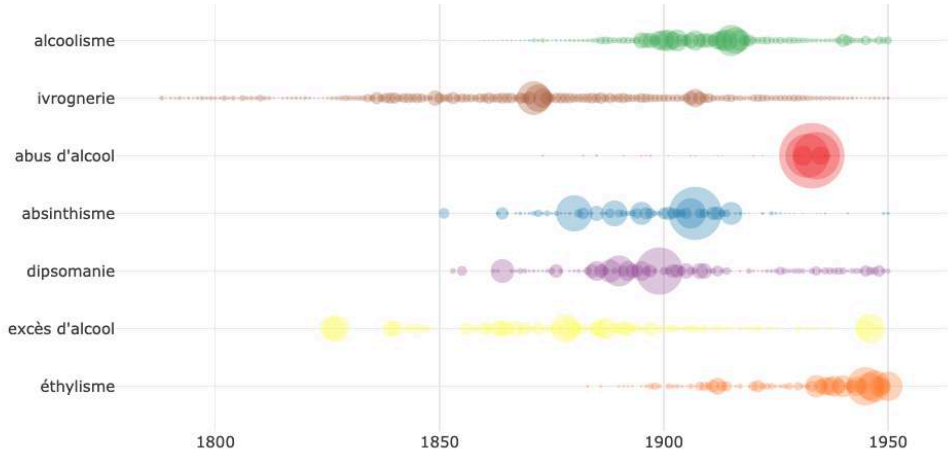


Figure 4. Graphique en bulles des synonymes d'*alcoolisme* sur la presse de *Gallica*. L'option « rééchelonnement » (z-score) est ici activée, pour comparer des mots aux fréquences très différentes. À noter qu'*abus d'alcool* est gonflé au milieu des années 1930 par une publicité pour l'Aspro, une forme de l'aspirine



18 On aimerait pourtant aller au-delà et étudier de larges familles de mots, y repérer les mots qui co-évoluent et les années qui se ressemblent, pour établir par exemple une « périodisation endogène » (Guaresi, Mayaffre & Vanni 2021). Nos visualisations ne se prêtent pas à une telle entreprise. Nous nous heurtons tout simplement aux limites de l'œil humain : il faudrait se situer dans un espace à *n* dimensions pour représenter tout ce tableau de données, or un écran d'ordinateur n'en possède que deux. La ruse de la couleur (Figure 3b) ou la segmentation de l'écran (Figure 4) ne font que repousser le problème – l'impasse est réelle. Pour avancer, il nous faut réduire le nombre de dimensions d'une manière ou d'une autre. Par chance, notre nuage de points en *n* dimensions possède en général une certaine structure. D'une part, il est sous-tendu par la chronologie : les années qui se suivent tendent à se ressembler du point de vue du vocabulaire employé. D'autre part, des mots d'un même lexique tendent à évoluer de manière parallèle – ainsi *guerre* et *armée*, dont les courbes corrèlent à hauteur de 50 % dans le corpus de presse, entre 1789 et 1950. Autrement dit, la fréquence de *guerre* permet largement de deviner la fréquence d'*armée*. Ces deux phénomènes rendent possible une « réduction de la dimensionnalité » : en exploitant la redondance du tableau, on peut chercher des axes qui sous-tendent nos données, et ratatiner ainsi notre nuage en perdant le moins d'informations possible. C'est une façon de comprendre les analyses factorielles qu'opère *Gallibase* et par là même leur grand intérêt : elles permettent de représenter au mieux en deux dimensions un tableau qui en possède *n*. Elle accommode ainsi *n* mots à un écran d'ordinateur, là où *Gallicagram* était impuissant ou condamné à la douloureuse addition.

2. Gallibase

Des courbes aux tableaux

19 Les jeunes auteurs qui ont écrit, sur les bancs de l'École, les lignes qui précèdent, cèdent ici la place à un vétéran d'une discipline bizarre qu'on appelle lexicométrie ou textométrie ou stylométrie ou encore logométrie. Le flottement qui enveloppe sa désignation n'incite guère à lui faire confiance. Elle vise pourtant à introduire la

mesure et la précision dans le domaine linguistique et littéraire et, à travers les mots, dans l'ensemble des sciences humaines. La statistique n'est jamais la démarche première. Elle vient après, quand les données ont été amassées, sans son concours, sans même qu'on songe initialement à y avoir recours. Les grandes entreprises documentaires comme le *TLF*, la *BnF*, ou *Google Books* se soucient avant tout d'engranger les textes et les mots pour les protéger de l'oubli, assurer leur permanence et leur sécurité, et frayer un chemin pour leur consultation. Elles ont bénéficié des techniques puissantes que l'informatique leur offre pour enregistrer, organiser, et diffuser les données. Or les mêmes calculateurs leur permettent aussi de calculer. Non sans réticence, on s'est contenté au début de la plus simple démarche arithmétique : l'addition. *Gallica* indique la fréquence du mot sur lequel porte l'interrogation, au moins pour avertir l'utilisateur de l'étendue du champ qui s'ouvre devant lui. S'il s'intéresse au *topinambour*, il apprend ainsi qu'une récolte de 7 199 rhizomes s'offre à lui. *Google Books* présente un dénombrement semblable, à l'échelle supérieure, avec 64 600 occurrences pour le même mot au singulier (encore faut-il faire une demande expresse en sollicitant le bouton idoine). Quant au *TLF* il fut le premier des dictionnaires à indiquer, au bas de chaque article, la fréquence du mot considéré dans l'ensemble de son corpus et aussi le pourcentage dans les sous-ensembles, ce qui suscite une comparaison et amorce déjà une démarche statistique.

20 La fréquence absolue d'un mot et même de tous les mots d'un texte ne permet en aucune façon de le rapprocher d'un autre texte. Encore faut-il accorder la fréquence d'un mot à la taille du texte. La façon naturelle qui vient à l'esprit, c'est d'établir un rapport entre l'une et l'autre et d'appeler le résultat pourcentage ou fréquence relative. Il a fallu du temps aux entreprises documentaires pour en arriver là. Si le *TLF* a pu devancer les autres sur ce point, c'est parce qu'il avait un objectif précis, le dictionnaire du XX^{e} siècle, et un corpus limité et daté qui n'a plus varié pendant la rédaction du dictionnaire. Il disposait ainsi d'un dénominateur constant. Cette suspension manquait à *Google Books* qui accumulait les textes dans d'énormes silos numériques, jamais stabilisés. Et c'est par une décision volontaire et extérieure[23] que le flot des entrées a été arrêté provisoirement en 2009. Mais le flot n'ayant pas suspendu sa progression, il a fallu ajuster les compteurs en 2012, puis en 2019. Et c'est autant de bases différentes, incluses l'une dans l'autre comme les poupées russes, qui s'offrent à l'utilisation. La *BnF* a suivi le même chemin et on peut prévoir que pareil débordement guette le présent projet de *Gallicagram*. Mais la jeunesse et l'énergie de ses auteurs laissent espérer qu'ils sauront s'adapter.

21 À peine ont-ils livré la première version de ce projet qu'ils songent à augmenter ses fonctions statistiques et à soumettre les données à des méthodes plus synthétiques qui traitent les tableaux et non plus les séries simples. Leur ambition va donc plus loin que le projet *Ngram Viewer* qui depuis douze ans propose le même logiciel à base de courbes chronologiques. Or si rien n'est plus clair qu'une courbe individuelle, rien n'est plus opaque que la multiplication des courbes, qu'elles soient enchevêtrées sur un même plan ou échelonnées dans l'espace. Et l'on s'enfonce dans l'obscurité dès que croît le nombre de mots représentés. Si certaines questions restent ponctuelles, comme la première attestation de l'expression « vote des femmes », la plupart des réponses attendues d'une base documentaire font intervenir un groupe souvent imprécis de mots ou expressions censés être en relation avec la question. La tentation est grande

alors de cumuler le tout en une seule distribution globale. Mais on écrase alors les différences et les nuances, en faisant taire les minorités.

Les deux approches de Gallibase

22 Suivant en cela l'exemple de *Ngram Viewer*, les auteurs de *Gallicagram* ont voulu fixer des pierres en attente, laissant à d'autres le soin de prolonger le mur et d'exploiter plus avant les données. Mais plus généreux que leurs prédécesseurs, ils proposent un fichier où toutes les informations utiles sont rassemblées : soit, pour chaque mot de la série, non seulement les pourcentages mais aussi les fréquences absolues et la taille du sous-ensemble, si bien qu'on peut vérifier ces pourcentages[24]. On peut aussi contrôler dans la source même la présence du mot, le lien à *Gallica* ayant été conservé, même s'il en coûte quelques centaines d'octets pour chaque citation (ils ont été tronqués sous la rubrique « url » dans l'exemple ci-dessous). Le détail des métadonnées reproduit le choix fait par l'utilisateur au moment de l'interrogation : les cinq dernières se répètent d'une ligne à l'autre et n'ont pas d'incidence sur le tableau, qui est construit en croisant les variables « *date* » (la colonne) et « *mot* » (la ligne) et en inscrivant la variable « *count* » (la fréquence absolue) à l'intersection. Reste la variable « *base* » qu'il faut enregistrer dans le total marginal des colonnes et qui servira à établir les probabilités.

```
"date","count","base","ratio","mot","url","resolution","corpus","langue","bibli","search_mode"
"1788/01",426,608519,0.000700060310360071,"jour","https://gallica.bnf.fr/.....","Mois","Pre
sse","Français","Gallica","N-gramme"
```

23 Ce formatage paraît un peu lourd et redondant. Mais il fallait tenir compte de la variété des disciplines et d'une très grande multiplicité de sources nationales, étrangères, régionales, chacune répondant à sa façon et à son rythme aux sollicitations. Cependant, là où *Gallicagram* se trouvait seul en cause, avec la pleine maîtrise des données, un processus plus rapide a été ménagé. Dans cette procédure il n'est plus besoin d'interroger le site en mode conversationnel, et de remplir une à une les cases du protocole. Un automate sous forme d'API délivre immédiatement la réponse pourvu que la requête qu'il reçoit ait le format requis. Il reste néanmoins à l'utilisateur le soin de respecter ce format et de préciser ce qu'il veut, mais la tâche est beaucoup plus simple que précédemment : la requête Python ci-dessous n'a plus que cinq paramètres à fournir, dont certains, fonctionnant par défaut, peuvent être omis : 1 - le mot ou expression recherché, 2 - le corpus intéressé, 3 - la date de départ, 4 - la date d'arrivée et 5 - la mesure du temps : année ou mois.

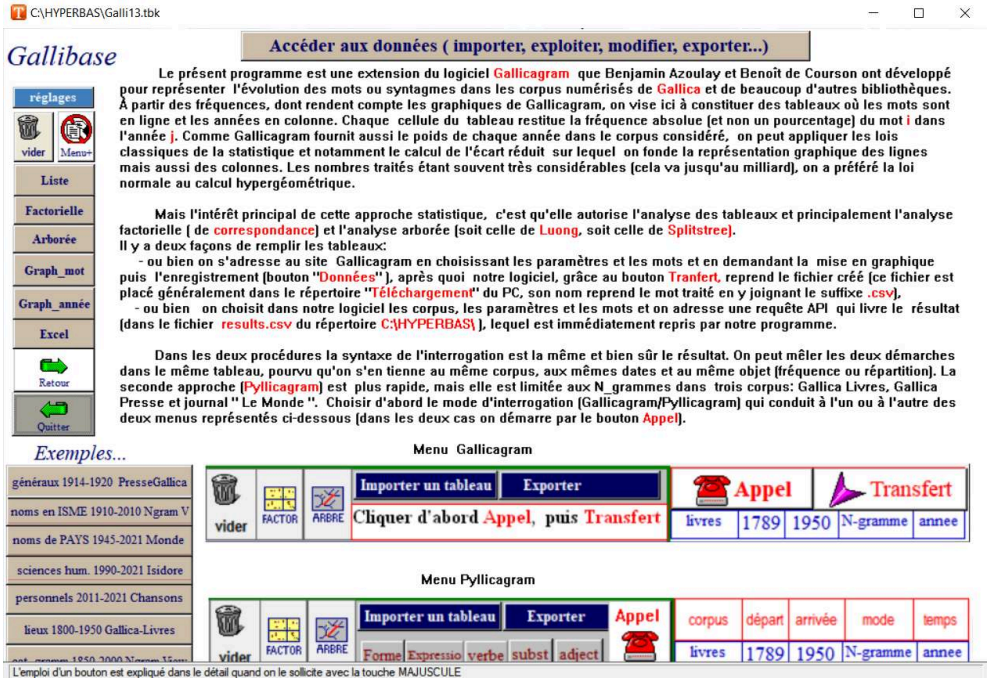
```
C:\HYPERBAS\python26\python.exe C:\HYPERBAS\python26\pyllicagram.py "droit des animaux" -c lemonde -d 2007 -f 2018 -r mois
```

24 Pareillement le résultat renvoyé par l'automate (dans un fichier qui porte toujours le même nom : *results.csv*) est d'une grande lisibilité. Encore pourrait-il être plus sobre en supprimant le ratio, dont le logiciel n'a pas besoin.

gram	*annee*	*mois*	*jour*	*n*	*total*	*ratio*
droit des animaux	2007	06	30	1	1450024	6.89643757621e-07
droit des animaux	2007	07	31	0	1167213	0.0
droit des animaux	2007	08	31	0	1011914	0.0
droit des animaux	2007	09	30	3	1329240	2.25692877133e-06

25 En réalité l'utilisateur n'a pas à connaître ces détails techniques s'il se sert du logiciel *GALLIBASE*, dont voici la porte d'entrée :

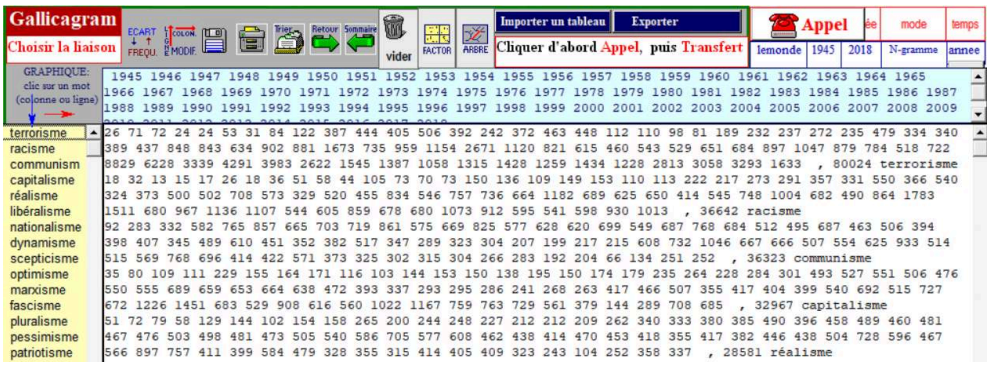
Figure 5. Fenêtre de démarrage de *Gallibase*



26 Cet écran initial (Figure 5) s'emploie à fournir les instructions élémentaires, les adresses des pages principales et le lien vers quelques exemples variés. On y insiste surtout sur la nécessité de choisir une des deux procédures qu'on vient d'évoquer, soit l'accès au site de *Gallicagram*, soit l'envoi direct d'une requête à l'API *Pyllicagram*. Dans les deux cas, on est conduit à la même page, où le choix peut encore être inversé grâce au bouton « *Choisir la liaison* ». Pour décider entre la consultation du site de *Gallicagram* ou l'utilisation de l'API *Pyllicagram*, il faut savoir que la première option convient dans tous les cas et que la seconde est réservée à trois corpus : *Gallica Livres*, *Gallica Presse* et journal *Le Monde.* Ces trois corpus sont heureusement les meilleurs, étant plus sûrs (surtout *Le Monde*), plus étendus, et plus rapides. On peut passer d'une procédure à l'autre dans la même séance, à condition que les paramètres de sélection soient rigoureusement les mêmes. Tant que ces paramètres ne varient pas, les mots ou expressions peuvent être ajoutés à la série en cours par une nouvelle requête. La suppression est également possible : il suffit de cliquer sur le mot indésirable. On a prévu aussi les regroupements partiels ou la totalisation entière. Dès que le tableau a plus de trois colonnes et plus de trois lignes, il devient exploitable. Mais les leçons qu'on en attend seront plus riches si son assise est plus large. En particulier la largeur d'un écran n'étant pas infinie, on ne doit pas outrepasser 200 colonnes. En cas de dépassement, les tranches chronologiques regrouperont 2, 3, n colonnes. La ressource de l' « ascenseur » permet d'étendre autant qu'on veut le nombre de lignes. Mais là aussi on risque de rencontrer des limites dans la taille du tableau et de toute façon la lecture des résultats de l'analyse est embrouillée quand trop de points s'y trouvent et s'y recouvrent[25].

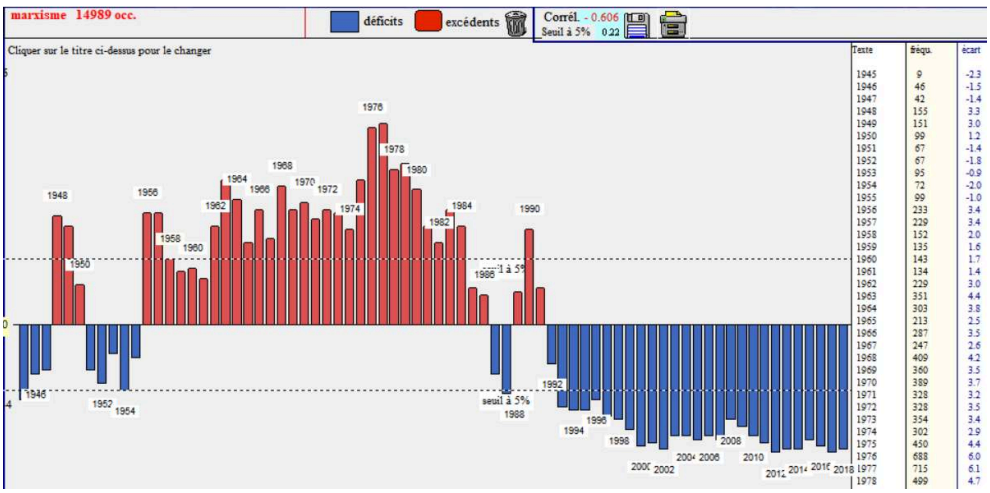
27 L'image ci-dessous (Figure 6) ne représente qu'une vue partielle de l'écran, lequel ne peut contenir lui-même qu'un tiers d'un tableau réunissant les plus fréquents des mots en *isme*[26]. L'enquête est menée dans le journal *Le Monde* de 1945 à 2018. Le champ exploré recouvre près d'1,4 milliard de mots, soit plus de 20 millions par an. C'est, raconté par une certaine presse, le récit de la bataille idéologique qui s'est livrée en France et dans le monde entre la seconde guerre mondiale et la guerre de l'Ukraine.

Figure 6. Tableau de données dans *Gallibase*. Ici, les occurrences d'une liste de mots en « isme »



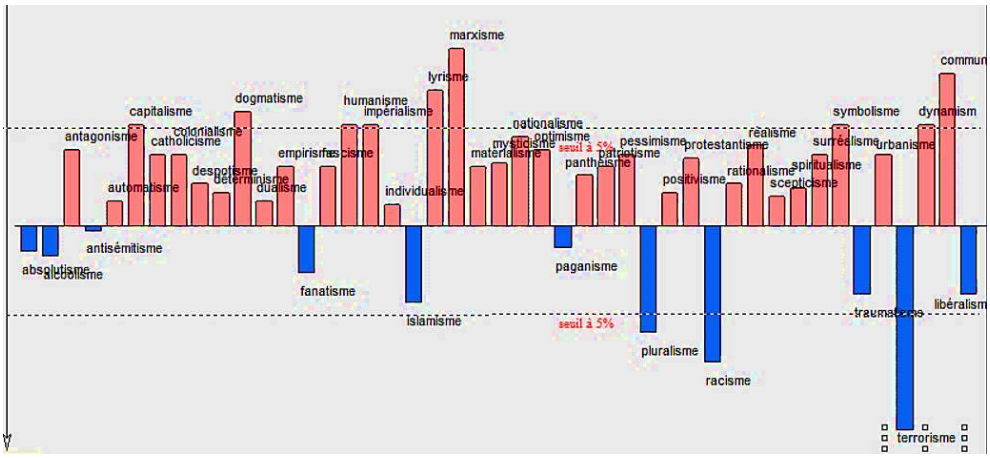
28 Mais avant de déployer la carte d'état-major, où évoluent les bataillons, utilisons la longue-vue pour fixer quelques positions. Par exemple le *marxisme* (Figure 7), longtemps dominant, s'écroule avec la chute du mur de Berlin[27].

Figure 7. L'évolution du mot *marxisme* dans Le Monde de 1945 à 2018. Affichage dans *Gallibase*



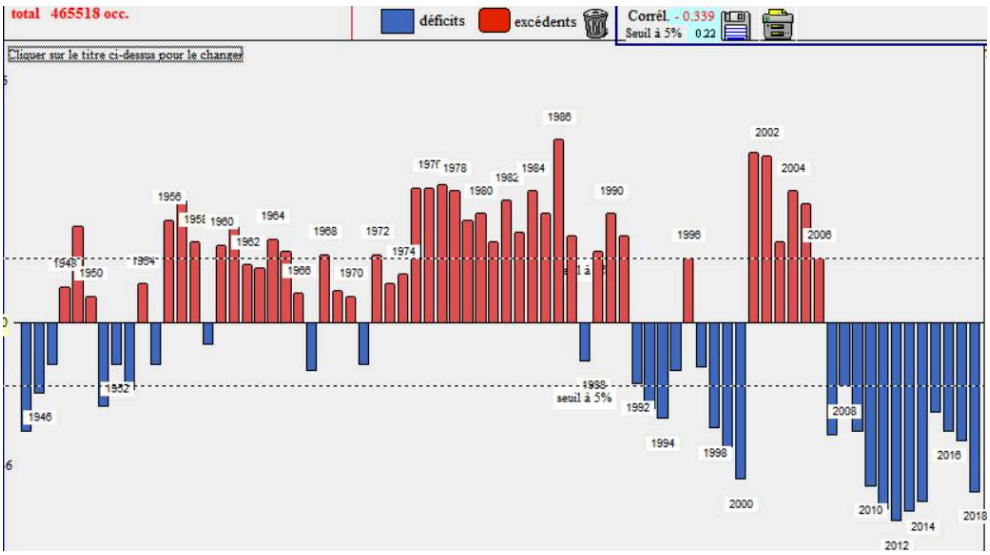
29 Dans un tableau voué à l'exploration statistique, lignes et colonnes ont le même statut et sont en principe réversibles[28]. Cela autorise à faire un zoom sur une colonne. Nul ne sera surpris de voir la mêlée fort agitée en 1968, *marxisme* et *communisme* combattant au premier rang (Figure 8).

Figure 8. La fréquence des mots en « isme » en 1968 dans *Gallibase*



30 Ce n'est pourtant pas en 68 que culmine le débat idéologique. Lorsqu'on réunit dans une même courbe cumulative toutes les séries particulières, on obtient comme un cardiogramme gardant la trace des éruptions politiques et morales du siècle. La première poussée de fièvre coïncide avec la guerre d'Algérie et la crise de Cuba, puis de 1973 à 1992 le débat s'envenime de crise en crise (pétrole, Viet Nam, guerre du Golfe, mur de Berlin, Bosnie). La guerre des mots s'apaise un moment pour reprendre avec les attentats du 11 septembre 2001 et la guerre d'Irak. Puis le feu semble s'éteindre jusqu'à la dernière année prise en compte, avant le Covid et l'Ukraine[29].

Figure 9. Courbe cumulative de 44 mots en « isme » (un demi-million d'occurrences) dans *Gallibase*

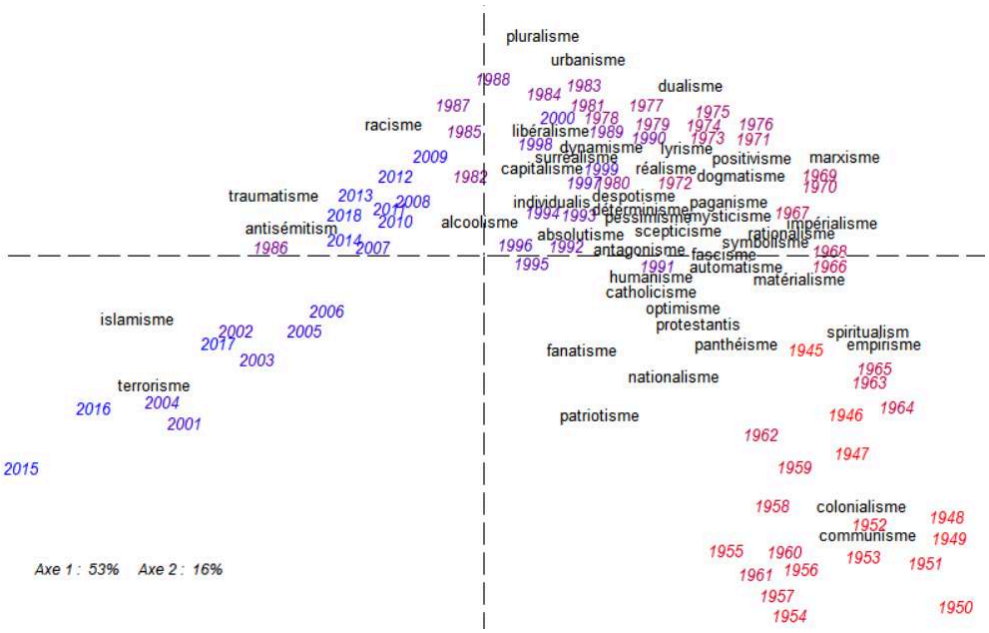


Analyse factorielle

31 On en vient enfin à l'analyse factorielle qui n'offre pas une somme ou un résumé, comme le graphique précédent, mais une synthèse où lignes et colonnes sont invitées à prendre place sur un plan selon qu'elles partagent des goûts communs (ou des répulsions partagées) avec les autres invités. On n'a jamais pratiqué cette méthode pour répartir les places à l'assemblée nationale non plus qu'à l'ONU car le calcul est trop compliqué et chaque représentant se trouverait interdit sur le seuil. Ce n'est pas le lieu

ici d'en développer les tenants et aboutissants. Un exemple suffira à illustrer son principe et ses vertus. On choisira l'analyse factorielle de correspondance, de Jean-Paul Benzécri, parce que sa pratique est très répandue en France et qu'elle convient admirablement à la statistique linguistique. Les fréquences constituent en effet un matériau facile à amasser, à isoler, à trier, et à manipuler sans grande dépense[30]. L'analyse qui suit repose sur près de 500 000 unités lexicales dans une urne qui en contient 1,4 milliard. Jamais un laboratoire médical n'a disposé d'autant de données. Nul besoin de traitement préalable ou d'affinage spécifique : la méthode préfère les fréquences brutes[31], car elle peut alors assembler ou diviser les lignes ou les colonnes sans changer le résultat global. Et surtout elle réunit dans la même figure les unes et les autres en donnant un sens à leur voisinage ou à leur éloignement.

Figure 10. Analyse factorielle de correspondance pour 44 mots en « isme » sur la période 1945-2018 dans *Gallibase*. Données issues de *Gallicagram* corpus *Le Monde*

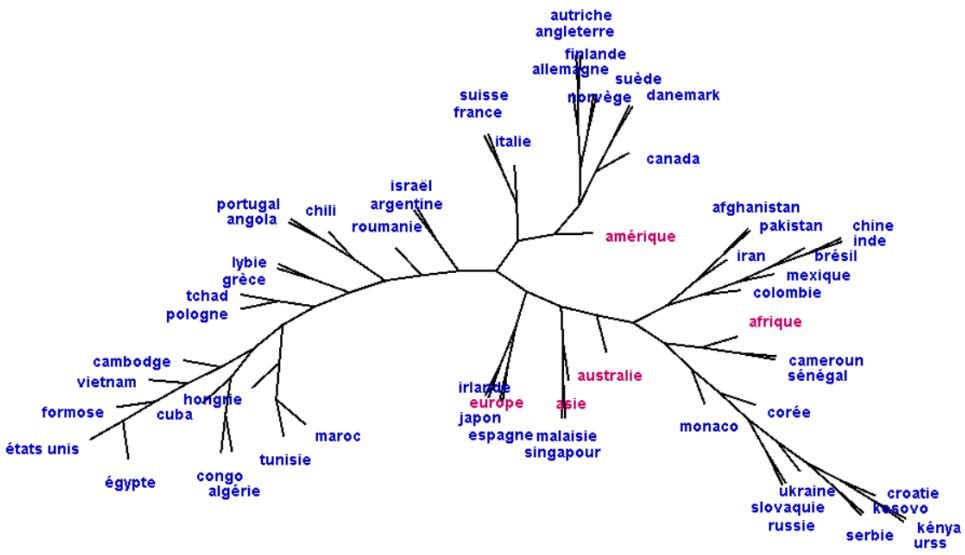


32 Ce qui frappe d'entrée de jeu (Figure 10) c'est le mouvement en croissant qui ordonne les années de la droite vers la gauche. On a ajouté la couleur après coup pour que l'œil distingue la suite des années, à travers un camaïeu qui passe progressivement du rouge au bleu. Sur cet arc viennent se greffer les mots selon la période où ils sont le plus souvent employés. *Nationalisme, patriotisme* et aussi *communisme* sortent grandis de la guerre au moment où le *colonialisme* doit affronter les indépendances. Suit une période dominée par les pôles opposés de la pensée religieuse et philosophique (*catholicisme, protestantisme, panthéisme, paganisme, spiritualisme, humanisme, matérialisme, rationalisme scepticisme, positivisme, déterminisme, empirisme*). Avant de franchir l'axe vertical et le passage au 21e siècle les dernières années du 20e se préoccupent plutôt des questions sociales et économiques (*marxisme, libéralisme, capitalisme, pluralisme*). La boucle se ferme à gauche avec une violence plus nettement politique (*racisme, terrorisme, islamisme, antisémitisme*) qui caractérise la dernière décennie.

Analyse arborée

33 On conviendra que les notions abstraites ne se prêtent pas toujours à des distinctions franches et que l'analyse factorielle ne peut pas savoir ce que chacun met dans des mots dont le contour est flou. Il en va autrement lorsque l'analyse porte sur des objets précis dont les frontières, même contestées, sont nettes, par exemple les noms de pays. Quand le nom d'une nation est évoqué dans la presse, c'est généralement parce qu'il s'y passe quelque chose et que l'événement mérite une mention. On peut donc lire l'histoire dans le miroir de la géographie. Reprenons notre corpus du Monde, cette fois en y ajoutant les dernières années. La carte ci-dessous que dessine l'analyse arborée (Figure 11) relève les points chauds du globe dont la presse fait état. Elle donne l'impression d'une carte des volcans, comme si les mouvements humains obéissaient à une tectonique des plaques. Il faut comprendre que l'automate n'a aucune connaissance géographique. S'il met ensemble certains pays comme le Maroc, l'Égypte, l'Algérie, la Tunisie, la Lybie dans la branche gauche de la figure, c'est parce que des événements s'y sont produits en même temps, sans doute par contagion. L'automate les découvre voisins dans le temps sans savoir qu'ils le sont dans l'espace. D'autres lignes de fracture apparaissent sur la même branche, mêlant les continents : l'Europe (Hongrie, Pologne, et Roumanie), l'Amérique du Sud (Argentine, Chili), l'Afrique (Tchad, Angola), l'Asie (Cambodge, Vietnam, Formose). Cet agrégat cesse de paraître incohérent quand on aperçoit les États-Unis au bout de la chaîne[32]. À l'autre bout de l'axe se dresse l'URSS, et les pays voisins (Russie, Ukraine, Serbie, Slovaquie, Kosovo). De ce côté aussi se trouvent les lieux explosifs où les armes parlent (Afghanistan, Pakistan, Iran, Corée, Kenya, Colombie) et les pays émergents qui attendent leur tour (Brésil, Mexique, Inde, Chine). Quant à l'Europe elle est à l'écart, à mi-chemin des deux extrêmes, avec un embranchement qui sépare le sud et le nord. L'activité volcanique y est faible sinon quelque temps en Irlande et en Espagne.

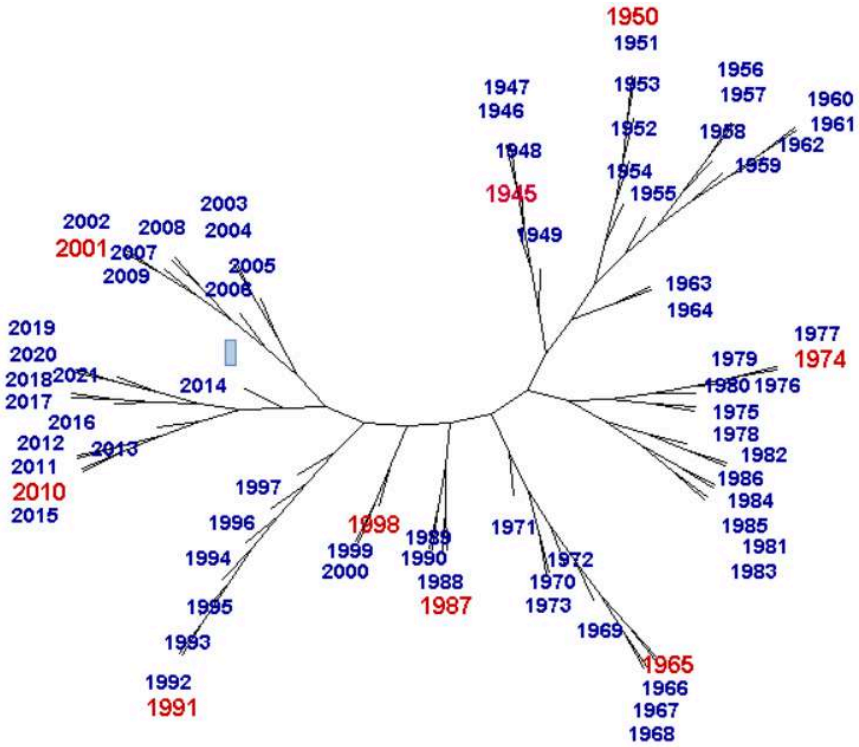
Figure 11. Analyse arborée proposée par *Gallibase* des noms de pays dans *Le Monde* (1945-2021)



34 Alors que l'analyse factorielle explicite en même temps les lignes et les colonnes, l'analyse arborée agit séparément. Le graphique précédent (Figure 11) rendait compte des lignes, c'est-à-dire des mots relevés dans le discours, celui qui suit (Figure 12) est

relatif aux colonnes, aux années où les mots ont été relevés. On voit bien qu'un lien de proximité unit les années consécutives : tout événement qui dure déborde d'une année sur l'autre. Et l'on attend une chaîne régulièrement ordonnée de 1945 à 2021. Or l'orientation générale est bien celle qu'on attend et chaque année tend la main à sa voisine chronologique. Mais il y a des ruptures, des nœuds où pendent des sous-chaînes. On laisse à un historien véritable le soin de comprendre et d'expliquer cette périodisation de l'histoire proposée par le calcul[33].

Figure 12. Analyse arborée proposée par *Gallibase* des dates issues de l'analyse des noms de pays dans *Le Monde* (1945-2021)



Bases et bibliothèques

35 Ce n'est pas par hasard que nous avons choisi nos deux exemples dans le corpus du journal *Le Monde*. Car il couvre la période la plus récente, celle que le corpus de presse de *Gallica*, embarrassé par le copyright, ne peut explorer pleinement. Rien n'empêche pourtant d'enregistrer les textes contemporains, s'il s'agit de distribuer des informations statistiques. Mais les responsables de la *BnF*, sachant que la diffusion du texte, même partielle, ne pourrait être pleinement légale, ont préféré réserver les lourds investissements de la saisie et du traitement là où le champ était libre. Les corpus *Monde* et *Gallica Presse* sont heureusement complémentaires et peuvent satisfaire les historiens modernes aussi bien que les contemporains, sans parler des sociologues et des linguistes à qui ces données ouvrent enfin de belles perspectives. On peut aussi considérer que les corpus *Gallica Livres* et *Ngram Viewer* se complètent utilement : on trouvera dans la seconde les données récentes qui manquent dans la première. Et même dans les périodes qui leur sont communes, ils ne font pas doublon, leur catalogue

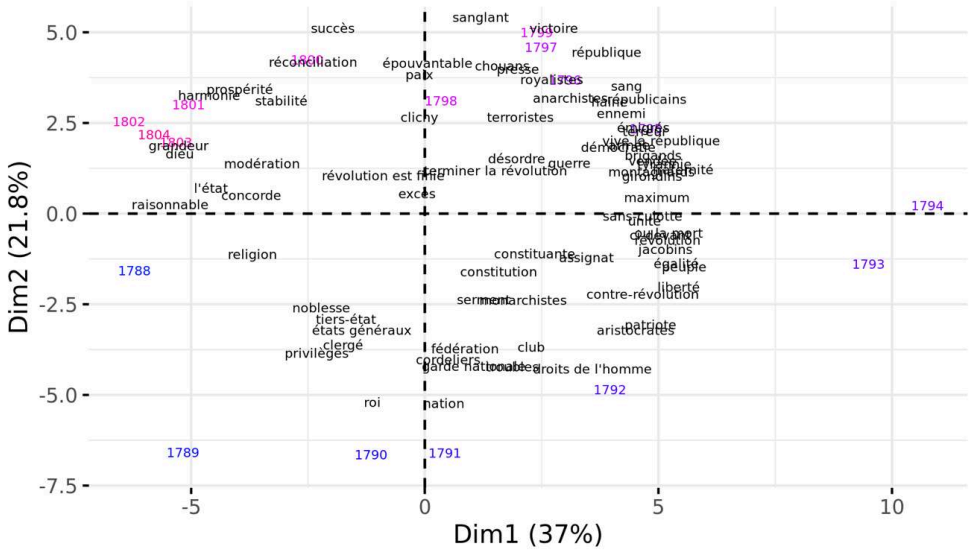
est différent et l'on peut trouver dans le plus petit des exemples qui ne sont pas dans le grand. Au reste *Gallicagram* offre un passage gracieux à son concurrent[34].

36 *Gallicagram* offre aussi une visibilité nouvelle et un débouché à une pléiade de bibliothèques dont les services étaient jusqu'ici méconnus. Certaines viennent de la francité belge, québécoise ou suisse, beaucoup d'autres de la province. Leur contenu est varié : presse locale, recherche scientifique, archives audiovisuelles et même chansons. Il est parfois difficile de recevoir et de comprendre leur message car ils ne sont pas indexés au niveau des N-grammes, c'est-à-dire des chaînes de 1 à 3 mots, et l'on n'a souvent qu'un faible repérage au niveau des pages ou des documents. Les recherches sont alors beaucoup plus lentes, par balayage des textes, et les résultats statistiques restent moins précis et moins sûrs[35]. Certains corpus (surtout de presse) ont une granularité fine, qui tient compte des mois, des semaines ou des jours. Le nombre de colonnes peut alors dépasser les limites raisonnables d'un tableau. Dans une telle situation le programme opère des regroupements de 2, 3, n colonnes, afin d'assurer la lisibilité des graphiques et de ne pas encombrer exagérément l'espace dans les représentations factorielles ou arborées. Ceux qui s'intéressent à ces bases spécialisées seront heureux qu'on leur épargne le voyage et les contraintes de la consultation sur place. Ils auront la patience d'attendre quelques minutes avant de recevoir le fichier attendu. Ces sources lointaines sont souvent mal documentées et *Gallicagram* ne fournit pas toujours *a priori* la distribution de leur catalogue dans le temps. Quelques essais seront nécessaires pour en prendre la mesure.

Conclusion

37 Concluons à l'unisson. L'intégration de *Gallibase* à *Gallicagram* est en cours. À l'heure qu'il est, les analyses factorielles viennent d'être mises en ligne[36]. Terminons par un exemple (Figure 13) sur la période révolutionnaire, moment où *Gallica* commence à s'étoffer en matière de périodiques, à la fois du fait de l'essor de la presse imprimée et des soins déployés par la BnF pour numériser les archives de cette période turbulente. En se penchant sur quelques mots spécifiques à la période révolutionnaire[37], plusieurs impressions se dégagent. Sur la disposition des années, d'abord, qui suit ici une boucle et non une parabole comme dans l'exemple précédent[38]. Elle indique que le profil lexical du Consulat est proche de celui de 1788, à tout le moins dans le référentiel des mots choisis. Elle donne à lire la Révolution non pas comme une évolution unidirectionnelle, mais comme un cycle, voire comme la perturbation d'un système qui tend finalement à revenir à l'équilibre. La trajectoire atteint son point extrême en 1794 – l'année de Thermidor –, dont l'algorithme fait un point de bascule, amorce du retour à l'ordre. L'isolement des années 1793-1794 est également saillant : l'exceptionnalité politique se répercute dans un lexique propre. Cette boucle n'est d'ailleurs pas parfaite : l'année 1799 semble celle d'un bref retour en arrière – probable écho de la « Revanche des conseils » et des espoirs néo-jacobins qui précèdent directement le 18 Brumaire.

Figure 13. Analyse factorielle d'un glossaire de la Révolution sur la presse de *Gallica*, entre 1788 et 1804. 2,2 millions d'occurrences sont exploitées. On a ici préféré l'Analyse en Composantes Principales (ACP) à l'Analyse Factorielle des Correspondances (AFC)



38 Les mots, eux, suivent à peu près la boucle des années ; ils offrent un déroulé des événements, des États généraux à la proclamation de la *concorde* et de l'*harmonie*, en passant par les lois du *maximum*. Ils offrent aussi une vue en coupe de ce que le Consulat et 1788 ont en partage : le lexique religieux, l'*État*, la *modération* et le *raisonnable*. Certains mots se superposent, ce qui nous renseigne sur la concomitance de leur usage, mais rend parfois la lecture difficile. En ligne, les *zooms* permis par *Gallicagram* et surtout le balayage des mots avec la souris pour les afficher[39] pallient toutefois ce défaut. On peut ainsi analyser l'amas de mots situé à l'extrême-droite du graphe, correspondant en fait à l'extrême-gauche du discours politique – l'orientation du graphe est arbitraire, rappelons-le –, avec *égalité*, *peuple*, *ci-devant* et l'expression même de la radicalité : « N + *ou la mort* ». Le lexique devient dysphorique lors du Directoire, malgré les *victoires* militaires. On y lit l'inquiétude face aux ennemis des deux bords, *anarchistes* et *royalistes* : toute une « écriture de l'horreur » (Ritz 2017), de l'*épouvantable* et du *sanglant*. On retrouve le caractère rétrospectif de la qualification *Terreur*, placée en 1795. *Excès* et le syntagme *terminer la révolution* sont situés à mi-chemin entre 1798 et 1790-1791, car utilisés pendant ces deux périodes. On peut y lire une recherche désespérée du juste milieu et d'une stabilisation, qui caractérise la Constituante comme le Directoire. Enfin, le Consulat est marqué par un registre euphorique, vraisemblablement imputable à la censure : en 1800, *la révolution est finie*[40], et dès lors on célèbre l'*harmonie* et la *prospérité*. L'isolement des années 1801-1804 suggère d'ailleurs que la censure empêche même de parler de la Révolution, ce qui explique aussi la structure en boucle. Cela fait écho à *L'histoire de France* publiée par Jacques Anquetil en 1805, où « les années 1789-1792 n'occupent qu'une place infime dans une histoire de France beaucoup plus vaste. La part la plus importante de la Révolution n'est pas traitée, comme s'il n'y avait rien entre la fin du règne de Louis XVI et le début de celui de Napoléon I[er]. » (Ritz 2018 : 9).

39 Cette rapide exploration numérique illustre la belle synergie entre les outils d'*Hyperbase* et les données de *Gallicagram*. Là où *Ngram Viewer* n'est utilisable que sur le

temps long (trop de hasards intervenant dans la rédaction des livres et leur publication), les corpus de presse de *Gallicagram* permettent des études sur périodes courtes. On aurait même pu utiliser la résolution mensuelle si l'on ne craignait pas de surcharger les graphiques. On se heurte une nouvelle fois aux limites de l'œil humain, mais on aura, entre temps, repoussé considérablement la frontière. La possibilité de sonder *n* mots simultanément permet une étude systématique incluant tous les mots possiblement pertinents. Cela limite la possibilité de « picorage », le chercheur choisissant, consciemment ou non, de ne garder que le mot qui valide, peut-être par hasard, son hypothèse.

40 Mais l'expérience montre – et c'est ici Étienne Brunet qui reprend la plume – que la statistique fait peur aux populations littéraires. Les courbes sont acceptées mais les tableaux provoquent incompréhension et réticence. Depuis douze ans le site *Ngram Viewer* ne propose rien d'autre que des courbes et *Frantext* a même cessé d'en produire. Notre logiciel *THIEF* (Truchement Hypertexte pour l'Interrogation et l'Exploitation de Frantext) s'est trouvé orphelin et ne fonctionne plus qu'en mode local. De la même façon, la base *GOOFRE2.EXE* qui exploite les données de *Google Books* est assujettie à l'API de *Ngram Viewer*. Cet API n'est pas toujours disponible et il arrive que son format change, imposant une conversion à ceux qui suivent.

41 Les outils statistiques proposés dans cette présentation sont d'une grande simplicité quand on les compare à l'extrême sophistication des méthodes qui fouillent internet et exploitent les données souvent futiles mais utiles qu'on y trouve. La *BnF*, héritière du *Trésor de la Langue Française*[41], offre un profit supérieur aux moteurs de recherche : « un trésor est caché dedans ».

BIBLIOGRAPHIE

Azoulay B. & de Courson B. (2021). « Gallicagram : Un Outil de Lexicométrie Pour La Recherche », SocArXiv.

Brunet É. (1981). « Le vocabulaire français de 1789 à nos jours d'après les données du *Trésor de la Langue Française* », 3 tomes. Genève-Paris : Slatkine-Champion.

Brunet É. (2012). « Au fond du GOOFRE, un gisement de 44 milliards de mots », 11e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012), Liège, Belgique, 7-21. https://hal.science/hal-01790505.

Brunet É. (2015). « La répétition dans la phrase. Étude statistique », *Semen. Revue de sémio-linguistique des textes et discours* 38.

Feltgen Q., Fagard B. & Nadal J.-P. (2017). « Frequency Patterns of Semantic Change : Corpus-Based Evidence of a near-Critical Dynamics in Language Change », *Royal Society Open Science.*

Guaresi M., Mayaffre D. & Vanni L. (2021). « Entre rupture et continuité, le discours du PCF (1920-2020) », *Histoire & mesure* 36(2) : 125-162. https://doi.org/10.4000/histoiremesure.14904.

Guiraud P. (1959). *Problèmes et méthodes de la statistique linguistique*. Dordrecht : D. Reidel.

Héran F. (2015). « Les mots de la démographie des origines à nos jours : une exploration numérique », *Population* 70(3) : 525-566.

James R. & Weiss A. (2012). « An Assessment of Google Books' Metadata », *Journal of Library Metadata* 12(1) : 15-22.

Langlais P.-C., Camps J.-B., Baumard N. & Morin O. (2022). « From Roland to Conan : First Results on the Corpus of French Literary Fictions (1050-1920) », *in Digital Humanities 2022* (*DH2022*).

Meng X.-L. (2018). « Statistical Paradises and Paradoxes in Big Data (I) : Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election », *The Annals of Applied Statistics* 12 : 685-726.

Michel J.-B., Kui Shen Y., Presser Aiden A., Veres A., Gray M. K., The Google Books Team, Pickett J. P. *et al.* (2011). « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science* 331(6014) : 176-182.

Ritz O. (2017). « Le 18 Fructidor de Louis-Sébastien Mercier », *Orages, Littérature et culture (1760-1830)* 16 : 71.

Ritz O. (2018). « Finir la Révolution par l'éloge », *Exercices de rhétorique* 11.

Schmidt B., Piantadosi S. T. & Mahowald K. (2021). « Uncontrolled Corpus Composition Drives an Apparent Surge in Cognitive Distortions », *Proceedings of the National Academy of Sciences* 118(45).

Tilly C. & Shorter E. (1973). « Les vagues de grèves en France, 1890-1968 », *Annales* 28(4) : 857-887.

NOTES

1. *Gallicagram* est utilisable directement en ligne : https://shiny.ens-paris-saclay.fr/app/gallicagram.
2. *Gallica* est le site internet rassemblant les archives numérisées de la Bibliothèque nationale de France.
3. Pour une parfaite introduction aux problèmes de *Ngram Viewer*, voir : Héran, 2015.
4. Correspondance personnelle.
5. *Gallibase*, qui fonctionne sur Windows, est téléchargeable à cette adresse : http://ancilla.unice.fr/pages/bases/. Ses fonctionnalités seront progressivement intégrées à *Gallicagram*.
6. https://gallica.bnf.fr/edit/und/a-propos
7. Nous nous sommes restreints aux documents en langue française, océrisés avec une précision estimée à plus de 50 %. Les équipes de *Gallica* ont mis en ligne notre corpus de livres à cette adresse : https://api.bnf.fr/fr/node/222.
8. Ces distributions sont disponibles sur le site, ou encore ici : https://github.com/regicid/pyllicagram.
9. Évidemment, cette fenêtre temporelle, à l'intérieur de laquelle les analyses issues du corpus nous paraissent pertinentes, est indicative. Ces dates correspondent aux périodes où nous avons obtenu des résultats probants. L'outil est particulièrement fiable au XIX^{e} siècle, période sur laquelle *Gallica* a concentré ses efforts d'océrisation. Au XVII^{e} siècle, le corpus « Livres » ne donne un résultat interprétable que sur des termes suffisamment courants.
10. Pour Prévost, le corpus est de plus largement composé non de ses propres romans, mais des traductions de ceux de Richardson, que les relevés monographiques ne permettent pas de discriminer.
11. Les recherches par cooccurrence sont également prises en charge par *Gallicagram*, mais en sélectionnant le mode de recherche « Par document » au lieu de « Par n-gramme ». Ce type de

requête utilise l'API de Gallica, et compte donc le nombre de documents où figurent cette cooccurrence et non le nombre des occurrences effectives.

12. Voir par exemple Feltgen, Fagard & Nadal 2017 et, auparavant, Brunet 1981.

13. Il faut préciser que *Ngram Viewer* n'exploite pas la totalité de *Google Books* mais un sous-corpus filtré, dont la composition n'a pas été révélée par Google. On ne peut donc que spéculer sur sa composition.

14. Pour un exemple récent : Schmidt, Piantadosi & Mahowald 2021.

15. La courbe du mot *divorce* jaillit par exemple en 1888, lors de la séparation très médiatisée du couple royal de Serbie.

16. Les premières détections du mot *décolonisation* en 1915 se révèlent ainsi une mauvaise reconnaissance du mot *décoloration*. La première occurrence de *vote des femmes* en 1848 est par contre fort instructive.

17. La majorité des occurrences proviennent de la reproduction de la phrase « L'ivrognerie n'existe plus, la poudre *Coza* l'a vaincue ».

18. C'est par exemple la raison pour laquelle nous avons inclus des intervalles de confiance rudimentaires, qui révèlent les périodes de vache maigre du corpus, où les fréquences sont très bruitées.

19. Cet argument est analogue au *big data paradox*, introduit dans Meng 2018. Pour une présentation plus accessible : https://towardsdatascience.com/n-is-the-enemy-c72cc1ba683b.

20. Il faut préciser par souci d'honnêteté que cette « masse minimale » où le bruit s'estompe est impossible à définir en général : elle dépend de la précision désirée et de la fréquence du mot recherché, un mot rare étant plus difficile à suivre (le coefficient de variation de l'estimateur s'écrit $(1-p)/\sqrt{(np)}$, où n est le nombre de mots et p la fréquence réelle). En revanche, on peut affirmer avec certitude que *Ngram Viewer* et *Gallica* la dépassent en général, pour des niveaux de précision raisonnables et des mots répandus. Prenons un exemple. Si notre mot a une fréquence « réelle » de 0,01 % (un mot comme « carotte » se situe dans cet ordre de grandeur), si l'on dispose d'un million de mots et si l'on suppose notre estimateur non biaisé, alors on a 95 % de chances de tomber juste à 2 % près. Notre corpus de presse possède plus d'un million d'occurrences par mois à partir de 1789, et jusqu'à 100 millions par mois au début du XX^{e} siècle.

21. Nous songeons à réaliser ce sous-corpus, en s'appuyant sur Langlais *et al.* 2022.

22. « [I]l y a des synonymes lexicaux, mais rarement au sens strict du mot » (Josette Rey-Debove, « La synonymie ou les échanges de signes comme fondement de la sémantique », *Langages* 31(128), 1997, 91-104, https://doi.org/10.3406/lgge.1997.2135).

23. C'est dans le milieu universitaire, à Harvard, qu'est né ce projet, à l'initiative de Jean-Baptiste Michel (*et al.*).

24. *Ngram Viewer* ne délivre que les fréquences relatives laissant à l'utilisateur le soin de reconstituer les effectifs absolus à partir des effectifs globaux de chaque année, lesquels sont enregistrés dans des fichiers à part. Cela est un frein majeur pour les utilisateurs peu aptes au traitement informatique. C'est pourquoi nous avions réalisé une base qui pour le français s'est chargée de réaliser tous les téléchargements nécessaires, non seulement pour le total de chaque année, de 1800 à 2000, mais aussi pour chacun des unigrammes (c'est-à-dire les mots individuels). Cette base, nommée *GOOFRE2.tbk*, fonctionne en mode local mais aussi *on line*, avec ou sans le recours à l'API de *Ngram Viewer*. Elle est incorporée au logiciel *HYPERBASE*, téléchargeable sur le site *ancilla.unice.fr*.

25. L'écran varie à peine quand on inverse la procédure : seuls changent les boutons du haut de l'écran dans la partie droite. Deux boutons suffisent dans le mode *Gallicagram* le premier, *APPEL*, pour se mettre en rapport avec le site et le second *TRANSFERT* pour désigner le fichier résultat que le site a renvoyé (généralement dans le répertoire « Téléchargement » de l'utilisateur). Dans le mode *Pyllicagram* les boutons en rouge sont découverts pour faire les choix correspondants. Le bouton APPEL s'y trouve aussi pour lancer la requête et récupérer le fichier des résultats lequel

est toujours déposé au même endroit (dans le répertoire *C:\HYPERBAS*) et sous le même nom (*results.csv*).

26. Les mots à qui le suffixe est étranger, comme *prisme*, ont été rejetés. Le mot *traumatisme* où le suffixe est présent aurait dû être éliminé, car il représente une réalité, loin de tout débat ou système de pensée.

27. Le graphique, qui s'abstient de tout lissage, est fondé sur l'écart réduit et relève de la loi normale. Contrairement aux pourcentages, l'écart réduit donne une assurance sur les observations. Lorsque sa valeur dépasse +2 (ou -2 en cas de déficit), le hasard n'a que 5 chances sur 100 de produire le résultat observé. Dans tous les graphiques cette zone incertaine autour de la moyenne est délimitée par des pointillés. On voit que le *marxisme* est au-delà de cette limite, soit dans l'excédent, avant 1988, soit dans le déficit, après 1992. En outre un coefficient de corrélation chronologique confirme la réalité de l'évolution : sa valeur -0,606 est très largement au-delà du seuil de 5 % (0,22 pour une série de 74 unités).

28. Ce n'est pas tout-à-fait vrai dans le cas présent parce que l'ordre a du sens dans la chronologie des colonnes – ce qui permet le calcul de corrélation – alors que le classement des lignes (alphabétique ou selon la fréquence) est indifférent et sans signification.

29. Rappelons que le corpus exploité ici est constitué d'articles d'un journal quotidien. Le lien que nous posons entre les débats et l'actualité serait moins étroit s'il s'agissait de livres.

30. Il y a soixante ans Pierre Guiraud disait déjà : « La linguistique est la science statistique type ; les statisticiens le savent bien ; les linguistes l'ignorent encore. » (Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, Reidel, Dordrecht, 1959).

31. Elle ne répugne pas à traiter les données pondérées, par exemple les fréquences relatives. Dans le cas des écarts réduits qui peuvent être négatifs, un traitement supplémentaire est nécessaire, assuré par notre logiciel, pour les transférer en zone positive, sans changer les distances.

32. On y voit aussi Israël.

33. On a mis en rouge la première année de chaque période.

34. Si l'on fait un décompte des initiatives françaises, soit 57 milliards de mots pour *Gallica presse*, 16 pour *Gallica livres* et 1,4 pour *Le Monde*, cela équilibre au total le corpus de *Google Books*, que l'on peut estimer à 90 milliards pour le domaine français.

35. Il ne s'agit plus alors de fréquence à proprement parler, mais de répartition. Les pourcentages et les calculs n'ont plus le même sens. Il faudra se méfier de l'interférence de la fréquence dans les relevés de répartition ; ainsi les mots d'usage courant peuvent se trouver dans tous les documents, avec un pourcentage de 100 %, sans qu'on puisse les distinguer les uns des autres.

36. On peut les activer sur *Gallicagram* en listant des mots séparés d'esperluettes et en remplaçant, au-dessus du graphe, « Courbes » pas « AFC » ou « ACP ».

37. Les auteurs remercient Olivier Ritz, qui fourmillait d'idées pour étoffer cette liste.

38. Les analyses factorielles sur corpus diachroniques disposent souvent les années en parabole, phénomène surnommé effet Guttman. Il révèle la prépondérance du facteur temporel, une évolution continue et cohérente du discours au fil du temps (Guaresi, Mayaffre & Vanni 2021).

39. Une version interactive identique au rendu du site est disponible ici : https://regicid.github.io/acp_r%C3%A9volution.html.

40. Cette expression est utilisée 51 fois seulement sur la période, essentiellement en 1800 (16 fois) et en 1791 (15 fois), d'où son positionnement sur le graphe.

41. Les milliers de textes qui avaient été saisis pour la rédaction du *TLF* ont été transmis à la *BnF* au moment où commençaient les travaux d'océrisation.

RÉSUMÉS

Gallicagram est un nouvel outil de lexicométrie, fondé notamment sur les archives océrisées de la Bibliothèque nationale de France et sur celles du journal *Le Monde* ; il dénombre dans le corpus choisi et pour une période donnée les occurrences d'un mot ou d'un syntagme, et offre différents modes de visualisation des données obtenues. Ce logiciel mérite à plusieurs titres d'être investi par les chercheurs : outre le volume des données qu'il exploite, suffisant pour fonder des analyses lexicométriques depuis le XVIIe siècle jusqu'à nos jours, *Gallicagram* a sur son concurrent immédiat, *Ngram Viewer*, l'avantage d'une transparence très supérieure et d'une structure plus constante au cours du temps. L'article présente *Gallibase*, son extension qui lui applique les outils de la statistique textuelle – en particulier les analyses factorielles et arborées. Il illustre son potentiel et insiste sur l'intérêt spécifique des corpus de presse, qui permettent des études sur périodes courtes.

Gallicagram is a lexicometry tool, based primarily on the archives of the French National Library and those of *Le Monde* newspaper. It counts the occurrences of a word and syntagma for a chosen corpus and a given period and offers several visualization options of the resulting data. For researchers, this software offers several assets: a large enough volume of data sufficient for lexicometric analysis from 1600 to present; transparency, which its competitor *Ngram Viewer* notably lacks; and a more constant structure throughout time. This article presents *Gallibase,* its extension which applies the tools of textual statistics, in particular factor analysis and tree clustering. It illustrates its potential and insists on the value of press corpora, which allows for the study of short periods.

INDEX

Keywords : Gallica, digitized collections, lexicometry, factor analysis
Mots-clés : Gallica, collections numérisée, lexicométrie, analyse factorielle

AUTEURS

BENOÎT DE COURSON

Max Planck Institute for the Study of Crime, Security and Law, Freiburg im Breisgau, Germany
Leiden University, Netherlands

BENJAMIN AZOULAY

ENS Paris-Saclay, Gif-sur-Yvette, France

CLARA DE COURSON

Sorbonne Université (EA 4509-STIH)

LAURENT VANNI

Laboratoire Bases, Corpus, Langage, CNRS, Université de Nice-Côte d'Azur

ÉTIENNE BRUNET

Laboratoire Bases, Corpus, Langage, CNRS, Université de Nice-Côte d'Azur