François Héran*

Les mots de la démographie des origines à nos jours : une exploration numérique

L'analyse textuelle consiste d'abord à analyser la fréquence des mots utilisés dans un texte ou un corpus de textes. Le programme Ngram Viewer en propose une version globalisante, en comptabilisant le nombre d'occurrences annuelles de mots ou de groupes de mots dans le corpus gigantesque rassemblé par le programme Google Books : plus de sept millions d'ouvrages publiés sur cinq siècles en huit langues. Quelles sont les forces et limites de cet outil inédit ? Que nous apprend-il sur la diffusion des mots, de ceux de la démographie en particulier, au cours du dernier siècle ? Comment l'usage des mots figurant dans les ouvrages du corpus est-il lié aux événements historiques ? Analysant les succès et les disparitions du vocabulaire de la discipline, François Héran *met en évidence le dépérissement des termes de la démographie classique et l'apparition de nouveaux thèmes et de nouvelles notions témoignant de l'élargissement disciplinaire des sciences de la population.*

Explorer les millions d'ouvrages imprimés depuis Gutenberg et obtenir instantanément un relevé lisible de la fréquence des mots employés sur une durée de plusieurs siècles, qui aurait pu envisager un tel exploit il y a seulement une génération ? Mis en ligne le 16 décembre 2010 et renouvelé en novembre 2013, le site Ngram Viewer puise dans la bibliothèque de Google Books pour explorer le lexique de 8 millions de volumes publiés depuis le XVI[e] siècle. L'application calcule instantanément la fréquence de mots ou de suites de mots dans huit langues (neuf si on distingue l'anglais d'Angleterre et des États-Unis). Elle ouvre la voie à un nombre infini de requêtes. L'idée d'en tirer parti pour étudier l'évolution du vocabulaire démographique a suscité deux tentatives, l'une publiée dans *Population et sociétés* (Héran, 2013), l'autre dans *Demographic Research* (Bijak *et al.*, 2014). Tout en soulignant les limites de Ngram Viewer, ces deux essais ont invité les démographes à l'utiliser au mieux.

* Directeur de recherche, Ined.

Correspondance : François Héran, Institut national d'études démographiques, 133 boulevard Davout, 75980 Paris Cedex 20, courriel : heran@ined.fr

 DOI : 10.3917/popu.1503.0525

On se propose ici d'approfondir l'exploration historique du vocabulaire de la démographie en insistant sur les temps d'innovation ou de rupture, de relance ou de recul. Quelle lumière cette chronologie jette-t-elle sur la place de la démographie dans l'ensemble des savoirs ? La réflexion initiée dans les deux articles cités portera une attention plus soutenue aux aspects méthodologiques, à commencer par des questions simples : comment Google Books sélectionne-t-il les documents à saisir et comment passe-t-on de ce corpus à celui de Ngram Viewer ? Le fait que les revues scientifiques soient exclues du corpus biaise-t-il les résultats ou apporte-t-il au contraire un éclairage sur les rapports entre démographie et société, démographie et culture ? Après une présentation attentive des potentialités et des limites de l'outil, déjà nourrie – mais pas exclusivement – d'exemples démographiques, on abordera une question sensible : est-il vrai que le poids de la démographie dans la culture écrite est en net recul ? Si oui, quelles en seraient les raisons et quels pourraient être les remèdes ?

I. Ngram Viewer, ou comment sonder un océan de mots

1. La marche presque irrésistible du programme Google Books

On ne peut produire ni commenter les courbes lexicales de Ngram Viewer sans comprendre au préalable sa construction et, en amont, celle de la source qu'il exploite, la bibliothèque numérique de Google Books. Des explications sont disponibles sur les sites officiels et certains blogs institutionnels. Les plus détaillées figurent dans l'annexe technique de l'article princeps sur Ngram Viewer, qui revient sur certaines des méthodes de Google (Michel *et al.*, 2010). Les manuels s'intéressent depuis peu à cette application (Hai-Jew, 2014). Mais ces éléments épars appellent une synthèse et, si possible, une synthèse critique.

Il convient d'abord de bien distinguer Google Books et Ngram Viewer. En octobre 2004, Google, le géant américain des moteurs de recherche, lance le programme Google Books qui ambitionne de numériser le plus d'ouvrages possible depuis les débuts de l'imprimerie, en partant des grandes bibliothèques des États-Unis et d'Europe *(Library project)*, puis en faisant de même avec les catalogues des éditeurs *(Partner program)*. Si Google affiche volontiers des objectifs de pure connaissance et d'ouverture du savoir, ses visées sont d'abord commerciales. C'est un atout majeur de stocker la masse des informations accumulées dans les bibliothèques de toutes les langues, ne serait-ce que pour nourrir les logiciels de traduction automatique. C'en est un autre d'intensifier ainsi la lecture en ligne et de capter les marques d'attention des lecteurs pour dresser leur profil de consommateur culturel – deux dimensions du projet qu'on laissera ici de côté.

L'objectif du programme Google Books a de quoi déconcerter le statisticien attaché aux règles de la représentativité. Il ne s'agit pas de numériser un échantillon aléatoire ou raisonné de la production d'ouvrages ou de périodiques mais, conformément à la philosophie des *big data*, de viser l'exhaustivité mondiale, ici la totalité des ouvrages conservés en bibliothèque depuis le XVI^e^ siècle, toutes langues réunies(1). La cible de Google Books inclut toutes sortes de documents : monographies, thèses, fiction, textes officiels... En revanche, les périodiques ont été écartés. De même, ont été laissés de côté les documents difficiles à scanner pour des raisons de format, de qualité de conservation ou de mise en page, c'est-à-dire les cartes, les affiches et, surtout, la presse. Il s'ensuit que la culture écrite emmagasinée par Google Books est plus savante que la culture commune (ce qui contribue par exemple à expliquer que le très journalistique « baby-boom » apparaisse tardivement dans le corpus, comme on le verra).

Pour accomplir ce programme titanesque de numérisation, il fallait avancer à marche forcée. Les opérateurs de Google ont déployé dans les grandes bibliothèques un système de saisie optique capable de traiter mille pages à l'heure sans aplatir les ouvrages (la courbure des pages étant automatiquement corrigée). En octobre 2010, cinq ans après le lancement de Google Books, la direction des opérations annonçait avoir saisi plus de 15 millions de volumes dans une centaine de pays en plus de 400 langues. En avril 2013, le chiffre avait déjà doublé : 30 millions de volumes, soit le quart des 130 millions qui auraient été imprimés et conservés par l'humanité depuis les débuts de l'imprimerie, à en croire les ingénieurs de Google(2). À ce rythme, la saisie exhaustive pourrait s'achever avant 2020.

Les gestionnaires des grandes bibliothèques américaines le reconnaissent : ils auraient mis des décennies à numériser leur fonds là où quelques années ont suffi à Google Books. Les mêmes professionnels ont observé néanmoins un net ralentissement de la numérisation dans les dernières années (Howard, 2012), apparemment lié au fait que Google Books s'attache désormais à combler des lacunes en privilégiant des fonds spécifiques (comme le fonds hispanique de l'Université du Texas). Il consacrerait plus de temps à éviter les doublons. Enfin, il mettrait désormais l'accent sur les bibliothèques européennes. Ces témoignages de bibliothécaires ne comblent qu'en partie la grande discrétion de la *major* américaine. Avare d'informations, elle les distille sur des blogs qu'elle ouvre et ferme à son gré, réduisant souvent la presse spécialisée à se contenter de conjectures.

(1) « Notre but ultime est de collaborer avec les éditeurs et les bibliothèques pour créer un catalogue virtuel exhaustif et facilement interrogeable de tous les ouvrages publiés dans toutes les langues, afin de permettre aux utilisateurs de découvrir de nouveaux livres et aux éditeurs de trouver de nouveaux lecteurs », https://books.google.com/googlebooks/library/index.html.

(2) La méthode de calcul a consisté à compiler et traiter les catalogues du monde entier. L'unité de compte est le « volume » unique identifié par l'ISBN, corrigé des doublons et délesté des cartes, affiches, microfilms et livres-audio. Au total et sans les périodiques, 135 millions de volumes auraient été imprimés et conservés depuis Gutenberg, 165 millions avec les périodiques (Taycher, 2010). Pour simplifier, on utilisera ici « ouvrage », « livre » ou « document » comme des équivalents acceptables du volume.

Qu'en est-il des fonds francophones dans cette vaste entreprise ? Une fraction importante de la production de langue française des XVII^e^ et XVIII^e^ siècles fut imprimée par des éditeurs-libraires hollandais, suisses ou anglais. D'où la forte présence des ouvrages de langue française de ces époques dans les fonds de la KB (la Bibliothèque royale des Pays-Bas), de la Bibliothèque universitaire de Lausanne ou de la Bodléienne à Oxford, toutes trois partenaires majeurs de Google Books en Europe. La Bibliothèque nationale de France (BNF), présidée de 2002 à 2007 par l'historien Jean-Noël Jeanneney, choisit de décliner l'offre de Google Books, au motif que l'Europe devait résister à la domination culturelle américaine (Jeanneney, 2005). La BNF a privilégié le programme Gallica, qui s'est fixé des normes de qualité élevées, dans le cadre de la bibliothèque numérique européenne Europeana, mais au prix d'une consultation en ligne laborieuse. La principale limite de Gallica pour notre propos est le privilège accordé à la saisie en « mode image » sur la saisie en « mode texte », fermant ainsi la voie à une analyse statistique du lexique. Depuis 2011, le Fonds pour la société numérique, *via* le Programme des investissements d'avenir, finance un vaste projet de numérisation des ouvrages de la BNF antérieurs à 1700, sans lien aucun avec le programme de Google Books.

Pour compenser le refus « gallican » de la BNF, Google Books n'a pas seulement puisé dans les fonds des bibliothèques américaines ou européennes, il a remporté en 2008 le marché de numérisation de la Bibliothèque municipale de Lyon, première bibliothèque municipale de France par le nombre de volumes (3,8 millions) et riche d'un fonds ancien (Colombet, 2008). Un autre accord a été conclu avec le catalogue en ligne de la librairie lyonnaise Decitre.

2. Ngram Viewer : aperçu sur les données, leur structure et leur potentiel

Le programme Ngram Viewer, pour sa part, intervient en aval de Google Books. Il a bénéficié d'un partenariat étroit avec les ingénieurs de Google, sans faire partie pour autant de ses applications commerciales. Conçu à Harvard par des chercheurs en informatique et traitement du langage, Ngram Viewer est hébergé par Google Books et reste d'accès libre[(3)]. Comme son nom l'indique, il produit des graphiques qui donnent à voir la fréquence relative de mots ou de suites de mots (les *n-grams*) imprimés au fil des siècles. Sur les 30 millions de documents numérisés par Google Books, Ngram Viewer en reprend 8,1 millions, répartis en huit langues (tableau 1). Le corpus francophone réunit près de 800 000 volumes, soit 102 milliards de mots, avec un nombre moyen de mots par volume particulièrement élevé, dû au caractère plus analytique de la syntaxe française.

Comment passe-t-on de 30 à 8,1 millions ? Les filtres successifs appliqués au corpus de Google Books sont succinctement décrits par les concepteurs de

(3) https://books.google.com/ngrams

Tableau 1. Dimensions du corpus de l'application Ngram Viewer dans sa version de 2012

Langue	Nombre de volumes numérisés	Nombre total de mots (« unigrammes »)	Nombre moyen de mots par volume
Anglais	4 541 627	468 491 999 592	103 155
Espagnol	854 649	83 967 471 303	98 248
Français	792 118	102 174 681 393	128 989
Allemand	657 991	64 784 628 286	98 458
Russe	591 310	67 137 666 353	113 541
Italien	305 763	40 288 810 817	131 765
Chinois	302 652	26 859 461 025	88 747
Hébreu	70 636	8 172 543 728	115 699
TOTAL	8 116 746	861 877 262 497	106 185

Note : Un mot ayant 50 occurrences dans le même volume compte 50 fois.
Source : Lin *et al.*, 2012.

Ngram Viewer (Michel *et al.*, 2010, complément en ligne, p. 7-8). Outre le choix de huit langues, ils ont pris en compte la possibilité de dater la publication (par concordance entre le document et les métadonnées de l'éditeur ou de la bibliothèque) et la qualité de la saisie optique (estimée par un algorithme spécial). Les difficultés de datation alléguées par les concepteurs les ont conduits à écarter la plupart des périodiques, y compris les revues scientifiques : un point majeur sur lequel on reviendra.

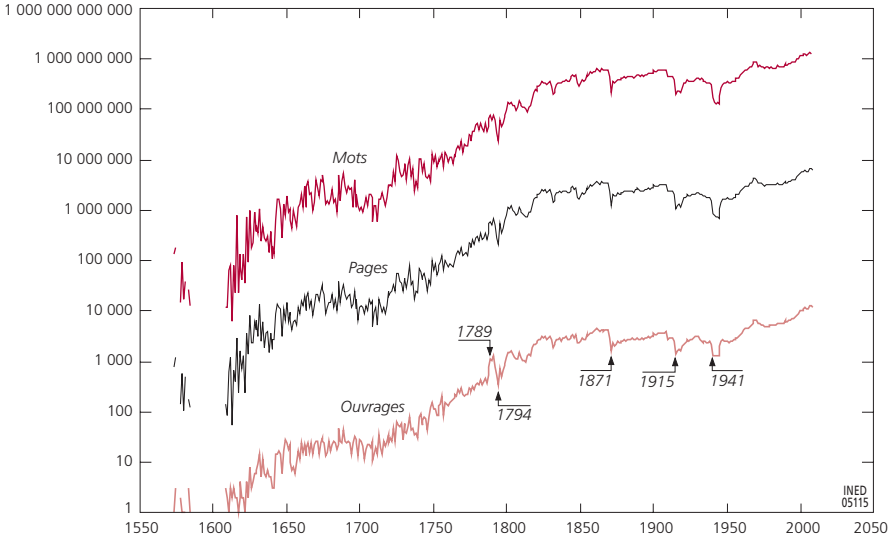
Couverture chronologique pour la France entre 1740 et 2008

La fenêtre temporelle de Ngram Viewer est d'une largeur inouïe, spécialement dans le corpus français. Ce dernier démarre en 1547, avec une couverture intermittente durant le XVI^e^ siècle, qui devient annuelle dès 1609 (figure 1). Le seuil des dix millions de mots traités chaque année en moyenne est franchi vers 1750[(4)]. À l'usage, l'année 1750 offre un bon point de départ aux requêtes sur Ngram Viewer : les effectifs réduisent sensiblement les aléas des courbes de fréquences, un demi-siècle avant le corpus de l'anglais américain.

Atout remarquable, le corpus français de Ngram Viewer inclut la Révolution. Comparées aux années 1786-1788, les années 1789-1791 comptent deux fois plus d'ouvrages et 45 % de mots en plus. Aux XIX^e^ et XX^e^ siècles, les courbes de production sont hachées par les épisodes révolutionnaires et les guerres, qui effacent à chaque fois les progrès des dernières décennies (figure 2). Le

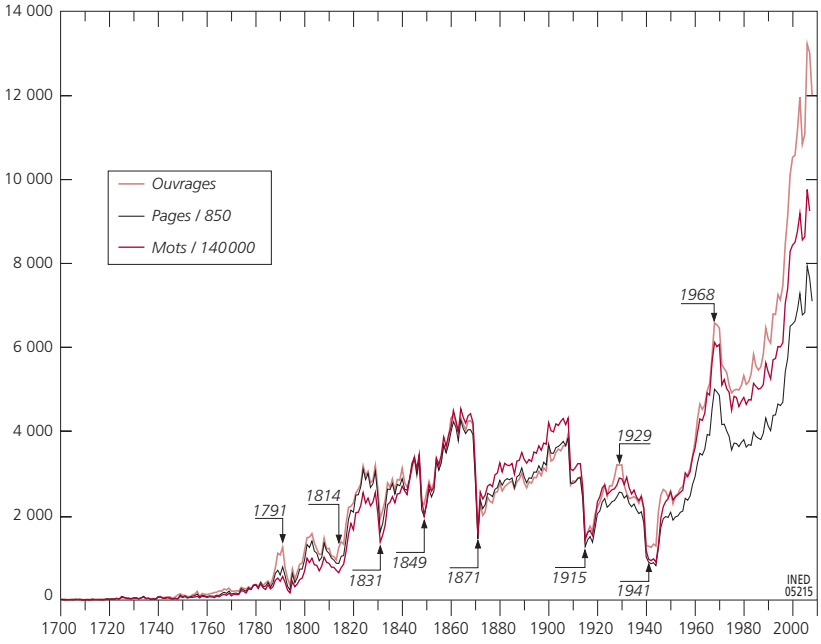
(4) Paraissent dans la décennie 1740-1749 la 3^e^ édition du dictionnaire de l'Académie française, la 5^e^ du *Dictionnaire de Trévoux* et la 8^e^ du Dictionnaire de Bayle (la plus lisible par sa typographie). La parution de *L'Encyclopédie* de Diderot et d'Alembert s'échelonne de 1751 à 1772.

Figure 1. Le corpus français de Ngram Viewer : données quantitatives (échelle logarithmique)



Source : Corpus francophone de Ngram Viewer.

Figure 2. Nombre d'ouvrages, de pages (divisé par 850) et de mots (divisé par 140 000) dans le corpus français de Ngram Viewer (échelle arithmétique)



Source : Corpus francophone de Ngram Viewer.

corpus atteint un sommet en 1968, sans que l'on sache si le creux qui suit marque l'épuisement d'une effervescence éditoriale ou si l'on entre dans une période mieux protégée par les droits d'auteurs. Quant à l'accélération de la hausse dans les années 1990 et 2000, elle pourrait être liée au fait que le corpus ajoute aux fonds de bibliothèques les catalogues d'éditeurs. Au total, la croissance du corpus français est telle que les ouvrages se distribuent en quatre parts égales sur les périodes 1547-1866, 1867-1936, 1937-1986 et 1987-2008, soit des durées de plus en plus courtes : 320 ans, 70 ans, 50 ans et 22 ans.

Le corpus s'interrompt pour l'instant en 2008. Cette clôture vaut pour les versions élaborées en 2009 et 2012 et devenues accessibles respectivement en 2010 et 2013. D'une version à l'autre, le corpus francophone a doublé de volume : 780 000 documents au lieu de 390 000. D'importantes corrections ont été apportées. Les suites de mots ne débordent plus les fins de phrase, tandis qu'à l'inverse ont été récupérées celles qui sont à cheval sur deux pages. Nombre d'erreurs de saisie optique ont été rectifiées, même si l'on trouve encore au XVII^e^ siècle des exemples épars de « natalité » issus d'une mauvaise lecture de… « nativité ». Le tracé allongé de la lettre *s* en position médiane reste souvent lu comme un *f* (une fois sur dix au cours du XVIII^e^ siècle, « Espagne » est interprété comme « Efpagne » et « peste » comme « pefte ») mais il suffit d'additionner les deux graphies pour résoudre le problème. Les césures en fin de ligne ne sont pas toutes corrigées (dans près de 30 % des cas, « exis » est encore un artefact produit par la césure non résorbée d'« exis-tence »). Défauts mineurs, qu'il ne faut pas confondre avec le respect légitime des orthographes anciennes[(5)].

La structure du corpus : des suites glissantes de un à cinq mots

Les inventeurs de Ngram Viewer appellent *gram* une chaîne de caractères bornée par des espaces. Une requête sur Ngram Viewer peut comprendre une suite de 1 à 5 *grams* : les *n-grams* sont donc des unigrammes, bigrammes, trigrammes, tétragrammes ou pentagrammes. Pour faire simple, on parlera ici d' « expressions » ou de « suites de mots ». Un corpus comprend les tables annuelles des suites possibles de 1 à 5 mots[(6)]. Chaque suite est suivie de l'année

(5) Jusqu'à la fin du XVIII^e^ siècle, on écrivait « & » pour « et », « hazard » pour « hasard ». Sous le Premier Empire, on se décida à remplacer « étoit » par « était », « François » par « Français ». Les éditeurs mirent une douzaine d'années à suivre les mutations adoptées en 1835 par l'Académie française : enfans > enfants, parens > parents, savans > savants, excédant > excédent, long-temps > longtemps, très-âgé > très âgé, etc. On ne saurait reprocher à Ngram Viewer de laisser en l'état les graphies anciennes : c'est à l'utilisateur d'adapter ses requêtes (par exemple, en additionnant les deux graphies possibles).

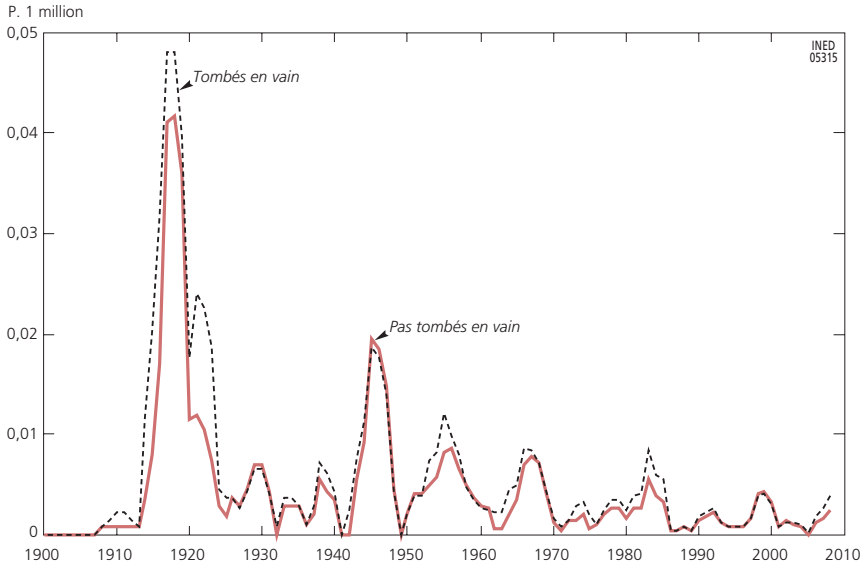
(6) Soit l'expression « espérance de vie en santé » figurant textuellement dans les documents numérisés. Ngram stocke chaque année, sur autant de tables séparées, les cinq mots qui la composent (« espérance », « de », « vie », « en » « santé »), les quatre bigrammes (« espérance de », « de vie », « vie en », « en santé »), les trois trigrammes (« espérance de vie », « de vie en », « vie en santé »), les deux tétragrammes (« espérance de vie en », « de vie en santé ») et l'unique pentagramme (l'expression complète), soit une somme cumulée de $S_5 = 15$ colonnes de données. Le lecteur peut s'intéresser aussi à des suites proches mais discontinues, comme « espérance de santé », mais elles ne figurent pas dans les mêmes passages.

de parution, du nombre d'occurrences dans les documents de l'année, du nombre de pages différentes concernées et du nombre d'ouvrages différents.

Grâce à cette structuration, il est possible d'interroger le corpus en insérant un joker (par convention, un astérisque) au début, à l'intérieur ou à la fin d'une expression de moins de cinq mots. On obtient ainsi les dix expressions les plus fréquentes à la fin de la période interrogée. Introduite fin 2013, cette innovation jette un éclairage précieux sur les associations de mots les plus fréquentes. Ainsi, la requête « âge au * » met en tête « âge au mariage » et « âge au décès » ; la première expression chute depuis les années 1980, alors que la seconde progresse, signe que la fin de vie a supplanté le mariage dans les préoccupations des démographes. Autre exemple : si la requête « interruption * de grossesse » détache sans surprise l'expression « interruption volontaire de grossesse », elle en révèle d'autres, moins fréquentes mais en progression : interruption « médicale », « thérapeutique » ou « involontaire » de grossesse.

Du fait de cette organisation des données, la fréquence d'une suite de *n* mots est calculée sur le nombre total de suites de « même longueur » dans les documents de l'année (par exemple, la fréquence d'un bigramme sur les bigrammes de l'année). Faut-il s'interdire pour autant de comparer des expressions d'inégale longueur ? En aucune façon, comme le montre par exemple la bonne superposition du trigramme « tombés en vain » et du tétragramme « pas tombés en vain », qui ressurgissent à chaque guerre (figure 3) ; elle signifie que, dans la majorité des cas, « tombés en vain » fait partie de la formule de dénégation

Figure 3. Exemple de comparaison entre une suite de trois mots et une suite de quatre mots



Note : Lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

« pas tombés en vain » (les exceptions pour certaines années pouvant s'analyser par une requête sans lissage). On peut calculer ainsi des ratios entre suites de longueur inégale, comme « espérance de vie / espérance », et découvrir ainsi qu'une fois sur quatre, quand on écrit le mot « espérance » en français, c'est au sein de l'expression « espérance de vie », alors que l'expression était à peine connue au début des années 1930 !

Barrière à l'entrée : la présence obligée dans 40 ouvrages différents

Pour éviter un accroissement exponentiel du corpus, les concepteurs de Ngram Viewer ont écarté les mots peu utilisés. On sait que la distribution des vocables dans un corpus suit une loi de Zipf : la langue écrite fait un usage intense d'un petit nombre de mots-outils (à commencer par les articles, auxiliaires ou prépositions) et un usage rare d'une foule de mots, dont les *hapax* (les mots à occurrence unique). Les concepteurs ont décidé d'écarter les mots rares en fixant un seuil élevé : pour figurer dans un corpus de Ngram Viewer, un mot doit apparaître, une année donnée, dans pas moins de 40 documents différents.

La contrepartie de cet allégement est la vitesse d'exécution. Une demande peut comprendre plusieurs requêtes (séparées par des virgules) et porter sur plusieurs corpus de langue (aisément différenciés par un code). Une fois la demande lancée, Ngram Viewer dessine en une seconde le graphique correspondant. Si l'on préfère redessiner soi-même ce dernier (comme c'est le cas dans le présent article), quelques minutes suffisent pour récupérer les chiffres correspondants[7]. Ngram Viewer arbitre au mieux entre la finesse de la recherche et la vitesse de la réponse. La lexicométrie y perd peut-être, mais la science sociale y gagne, car la période approximative où une notion entre dans l'usage réel est plus significative sociologiquement que la date exacte de sa prime apparition.

La hiérarchie des fréquences et le problème de la normalisation

L'ordre de grandeur des fréquences peut varier fortement d'un domaine à l'autre. Une requête contenant uniquement l'astérisque-joker révèle que les mots les plus utilisés dans la langue française imprimée sont « de », « la » et « et » ; ils représentent respectivement 4,3 %, 2,4 % et 1,8 % de tous les mots imprimés en français depuis deux siècles. Des vocables aussi courants que « vie », « mort » ou « âge » atteignent respectivement les fréquences 0,063 %, 0,025 % et 0,014 %, soit 63, 25 et 14 pour 100 000. Plus technique, le terme « démographie » descend à 0,0006 %, soit 6 pour un million, le même ordre de grandeur que « mortalité » (2 pour un million) ou « natalité » (4 pour un million)[8]. On le voit, la démographie a peut-être pignon sur rue mais elle n'est

(7) Les chiffres sont téléchargeables depuis la ligne « *var data* » du code-source de la page de requête. Il suffit de recopier cette ligne sur un tableur et de la redistribuer en autant de colonnes que d'années.

(8) Même ordre de grandeur que les *ppm*, ou « parties par million », utilisées dans les mesures de dilution.

pas au centre du monde. Des expressions plus spécialisées encore, comme « espérance de vie en santé », à la limite de la visibilité, ont une fréquence limitée à 1 pour 100 millions. Ce qui n'empêche pas qu'elles suivent dans leur ordre de grandeur une trajectoire révélatrice : « espérance de vie en santé » est monté en flèche entre 1990 et 2003. Plus que le niveau des fréquences calculées par Ngram Viewer, ce sont leurs variations au fil du temps qui retiennent l'attention, ainsi que les comparaisons de fréquences entre plusieurs expressions (voir encadré).

Cette remarque vaut spécialement pour la relation entre vocabulaire courant et vocabulaire savant et, au sein de ce dernier, entre vocabulaire générique et vocabulaire spécialisé. Dans les années 2000, par exemple, « démographie sociale » est 200 fois moins fréquent au sein des bigrammes que ne l'est « démographie » au sein des unigrammes. Sa courbe est marquée par un premier pic en 1957 et son apogée en 1983, suivi d'un déclin prononcé, puis d'un rebond à partir de 2004. Si heurtée soit-elle, la trajectoire de « démographie sociale » peut se comparer à celle du mot « démographie », qui culmine elle aussi dans les années 1980. Ngram Viewer permet ainsi des comparaisons sur plusieurs échelles, sans confondre les différents ordres de grandeur : chacun a son propre niveau de visibilité.

Les graphiques produits par Ngram Viewer ajustent automatiquement l'axe des fréquences à la valeur maximale observée dans la période choisie. Si l'on inclut dans une même requête « France », « Alsace », « Strasbourg » et « Bas-Rhin », la courbe « France » écrase les trois autres, car « Strasbourg » est 25 fois moins répandu que « France », « Alsace » 40 fois moins, « Bas-Rhin » 500 fois moins. Plusieurs solutions existent pour produire dans ce cas des courbes commensurables : retirer certaines d'entre elles, les multiplier par un facteur d'échelle, calculer des ratios entre courbes(9).

Sauf demande expresse d'un calcul de ratio, l'ordonnée des graphiques produits par Ngram Viewer est un pourcentage (toujours réduit dans le présent article en « parties pour million »). Mais quels sont les effectifs figurant au numérateur et au dénominateur ? La requête d'un *n-gram* (tel le trigramme « espérance de vie ») retient au numérateur toutes ses occurrences dans l'année, à condition qu'il soit présent dans au moins 40 documents différents. On pouvait penser que le même filtrage s'appliquerait au dénominateur, mais les développeurs de Ngram Viewer ont choisi d'inclure dans ce dernier tous les *n-grams* de l'année, y compris ceux qui n'ont pas franchi le seuil requis. Décision légitime, car c'est bien à l'ensemble de la production écrite qu'il faut rapporter les expressions analysées. Mais ce choix d'une « normalisation totale » a été critiqué au motif que les mots rares inclus dans la masse des documents récemment scannés apporteraient de plus en plus de « bruit » non significatif, ce qui, par contrecoup, réduirait artificiellement la proportion des expressions

(9) Les requêtes peuvent s'écrire « France, (Alsace*40) », « (France/40), Alsace », « Alsace/France » ou « Alsace / (France + Alsace) ».

Quelques propriétés utiles de Ngram Viewer

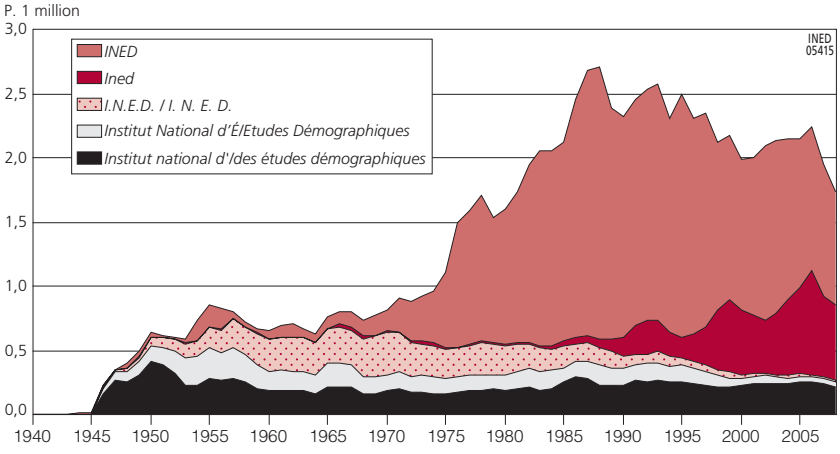
Lissage des séries annuelles

Les séries annuelles dessinées par Ngram Viewer sont marquées de fortes fluctuations. Une option de lissage par moyenne mobile est disponible. Un lissage d'ordre *n* ajoute à l'année de référence les *n* années qui précèdent et les *n* années qui suivent, soit une moyenne mobile sur 1+2*n* années. À l'usage, le lissage d'ordre 3 (moyenne mobile sur 7 ans) offre un bon compromis entre la précision temporelle et la lisibilité des tendances. Le lissage d'ordre 5 est parfois utile pour couvrir les siècles. À l'inverse, pour des événements bien datés (guerre, loi, création d'une institution), mieux vaut un lissage d'ordre 1, voire l'absence totale de lissage.

Distinctions de casse : banalisation des institutions et déclin des capitales

À la différence des moteurs de recherche classiques, Ngram Viewer est sensible à la graphie littérale et à la « casse » (capitales *versus* minuscules). Une option contraire (*case insensitive*) permet de différencier les variantes possibles et de les additionner, comme dans le cas de « PaCS », « PACS », « Pacs » et « pacs » ; elle échoue toutefois à identifier les acronymes ponctués ou les capitales non accentuées (« état » sous l'option *case insensitive* permet de retrouver « État » mais pas « Etat »). L'évolution des conventions typographiques appliquées à des institutions comme l'Ined (mais on aurait pu choisir aussi bien le CNRS, l'Inserm ou l'ONU) a une signification sociale que Ngram Viewer permet de suivre aisément (figure A). Plutôt rares dans l'après-guerre, les institutions se sont multipliées et banalisées, si bien que la graphie monumentale à la romaine (I.N.E.D.) a cédé progressivement la place à la graphie ordinaire d'un nom de famille (Ined). Encore dominante, la forme INED semble bien n'être plus qu'un maillon intermédiaire dans cette évolution : son temps est compté.

Figure A. Principales graphies de l'Institut national d'études démographiques

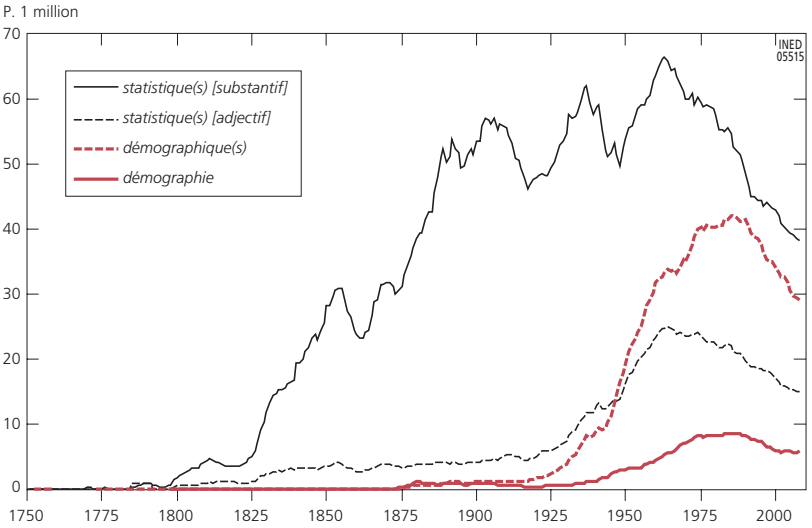


Note : Fréquence cumulée pour 1 million, lissage d'ordre 1.
Source : Corpus francophone de Ngram Viewer.

Substantifs et adjectifs : des destins séparés

Atout remarquable, les corpus de Ngram Viewer incluent des mots-clés produits par une analyse grammaticale automatique des parties du discours. Ils sont codés de façon transparente par des suffixes : _NOUN, _adj, _verb, _adv, _conj, etc. La requête « statistique_adj, statistique_NOUN » (figure B) révèle ainsi que l'adjectif « statistique » n'a réellement décollé qu'un siècle après le substantif (vers 1925 au lieu de 1825). Les associations de mots décelables par le joker sont éclairantes : « statistique(s) » a pris son essor comme adjectif en qualifiant des mots tels que « traitement », « analyse », « étude », « méthode », « enquêtes », « information » ou « Office », tous termes qui attestent une professionnalisation accrue.

Figure B. Distinguer substantifs et adjectifs : une application au mot « statistique(s) »

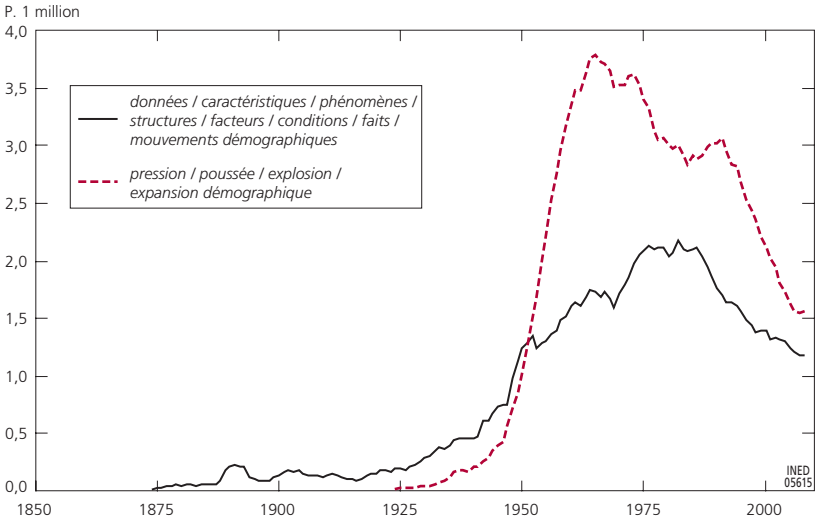


Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

« Démographique » : plus politique au singulier qu'au pluriel

Le respect littéral de la graphie par Ngram Viewer s'applique aussi aux distinctions de nombre. Alors que « démographiques » au pluriel est plutôt technique, « démographique » au singulier penche vers la sphère politique ou médiatique. La requête « * démographiques » révèle une association privilégiée avec des termes à connotation technique : « données », « phénomènes », « caractéristiques », « structures »… Au singulier, en revanche, « démographique » s'accole le plus souvent à des termes évaluant la hausse de la population, sur un mode qui peut être neutre (« croissance », « accroissement »…) ou normatif (« expansion », « pression », « poussée », « explosion »…), avec en filigrane la crainte de la surpopulation mondiale (figure C). Ceci ne doit pas faire oublier qu'avant l'usage de l'adjectif « démographique(s) », d'autres idéologies ont fleuri : la hantise de la « dépopulation » sous la II^{e} République (surtout entre 1890 et 1922) et celle de la « dénatalité » avant le baby-boom (Héran, 2013).

Figure C. Deux approches de la démographie : établir les faits, apprécier les évolutions



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

réellement usitées. Si cette hypothèse était avérée, elle pourrait remettre en cause les diagnostics de déclin portés sur de nombreuses activités scientifiques, dont celles de la démographie. On reviendra sur cette hypothèse dans la seconde partie de cet article, et ce sera pour la réfuter.

3. Ngram Viewer et ses critiques

Le programme Ngram Viewer a essuyé deux critiques principales de nature très différente. La première prend pour cible la prétention des concepteurs à fonder sur cet outil une nouvelle science, baptisée *culturomics* (Michel *et al.*, 2010). La seconde déplore l'impossibilité d'accéder au contexte complet des suites de mots numérisées.

La « culturomique », une nouvelle science ?

À en croire les concepteurs de Ngram Viewer (ou, du moins, certains d'entre eux), l'application inaugurerait une nouvelle science, à l'intersection des humanités numériques et des études culturelles. Elle révolutionnerait l'étude comparée des cultures. Les chercheurs français en sciences sociales n'ont pas manqué de souligner l'improvisation des déductions historiques, sociologiques ou culturelles qui ont accompagné les essais de « culturomique », tel cet ouvrage de vulgarisation publié par deux mathématiciens français d'après la version 2009 du corpus (Delahaye et Gauvrit, 2013). L'équivalent américain pour la version 2012 est le livre-manifeste des concepteurs de Ngram Viewer (Aiden

et Michel, 2013), plus à jour et plus fouillé, mais qui s'expose aux mêmes objections. L' « approche numérique de la culture » requiert une solide culture numérique (mathématique, informatique et linguistique), mais aussi de la culture tout court. La première ne peut compenser l'autre, sous peine de discréditer l'outil. Fort heureusement, chacun reste libre d'utiliser Ngram Viewer sans adhérer au projet « culturomique ». C'est déjà beaucoup d'avoir mis au point un instrument d'exploration aussi novateur et puissant. Nul besoin d'en faire le drapeau d'une nouvelle science.

Des mots sans contexte ?

Ngram Viewer a enthousiasmé des spécialistes du traitement automatique du langage, tel Jean Véronis, rédacteur d'un blog très visité, décédé en décembre 2013. Mais les historiens ou les sociologues férus de lexicographie ont multiplié les critiques. Quelques jours après le lancement de l'application, le rédacteur d'un blog pour historiens rendait déjà son verdict : « 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 » (Ruiz, 2010). La lexicométrie réclame pour chaque vocable une « concordance », à savoir la phrase complète où il figure, les phrases environnantes et les références précises de la citation (auteur, année, édition, page, type d'ouvrage). Ngram Viewer isolant les mots de leur contexte, les voies habituelles de l'interprétation s'en trouvent coupées, obligeant le lecteur à mobiliser une culture extérieure au corpus, avec ce défaut que le lien entre les mots et leurs ressorts se relâche : Ngram Viewer ne serait bon, somme toute, qu'à recouper les intuitions préexistantes (Chateauraynaud et Debaz, 2010 ; Peccate, 2011).

À ce type d'objections, les concepteurs répondent que le contrat passé avec Google impose l'anonymat des textes numérisés (Aiden et Michel, 2013). Ngram Viewer échappe ainsi aux poursuites des ayants droit[10]. Autre argument, le corpus lesté de toutes ces références aurait cessé d'être maniable, alors qu'il s'agit déjà du corpus le plus volumineux connu à ce jour. À quoi les lexicomètres rétorquent que le changement de taille ne justifie pas le sacrifice d'une exigence fondamentale à leurs yeux : « maîtriser le corpus » (Ruiz, 2010).

La critique est rude. Elle fait peu de cas d'une fonction très précieuse de Ngram Viewer : au pied de chaque graphique, l'application affiche en hypertexte une série de dates, qui correspondent aux pics et plateaux de chaque courbe. Il suffit de les pointer pour avoir accès aux ouvrages de la bibliothèque de

(10) Les ayants droit (auteurs dans l'aire anglophone, éditeurs dans l'aire francophone) ont porté plainte contre la saisie optique effectuée par Google sans leur autorisation. La justice américaine hésite entre l'*opting in* (Google peut numériser d'office tant que les ayants droit ne s'y opposent pas) et l'*opting out* (nécessité d'une autorisation préalable). Le problème se pose surtout pour la vaste zone grise des œuvres encore protégées mais retirées du commerce. Voir le résumé éclairant de Benhamou (2014, p. 90-92).

Google Books imprimés dans la même période[11]. Soit l'exemple d'« expert » et d'« expertise », un thème dont Chateauraynaud est un spécialiste reconnu. Selon lui, les courbes de ces deux termes restent mystérieuses tant qu'on ignore telle thèse de droit de 1934 ou l'ouvrage qu'il a lui-même publié sur le sujet en 1991 ; il faut examiner les liens de ces textes avec les textes antérieurs, savoir s'ils procèdent par citation, compilation, critique, récupération, etc. Et le critique de conclure que la connaissance fine du réseau intertextuel est indispensable pour interpréter les graphiques de Ngram Viewer. Or l'objection perd de sa force quand on découvre que les courbes de fréquence d' « expert » et d' « expertise » sont assorties de renvois à Google Books qui mettent précisément en exergue les deux ouvrages en question ! Il est donc très exagéré de soutenir que Ngram Viewer couperait « absolument » les mots de leur contexte.

En revendiquant la maîtrise complète du corpus pour chaque citation, la critique lexicographique ou « socio-informatique » se trompe d'échelle. Elle cherche l'individu singulier dans la fresque quantitative, un peu comme si le recensement de la population devait éclairer la situation de chaque habitant par son entourage singulier. Nul ne peut interpréter des macrodonnées comme s'il s'agissait de microdonnées. Les approches interactives et pragmatistes, si légitimes soient-elles par ailleurs, ne peuvent pas investir ce type de données.

La maîtrise du contenu s'impose pour des corpus de taille plus modeste. Mais, dans le cas présent, elle impliquerait à la fois de survoler de vastes étendues (plusieurs langues sur plusieurs siècles) et d'arpenter le terrain pas à pas. À la façon de l'archéologie aérienne, Ngram Viewer permettrait des vols de reconnaissance, mais ce serait ensuite au chercheur de terrain de vérifier au sol la réalité des structures soupçonnées de haut. Métaphore séduisante, utilisée dans la conclusion d'une précédente publication (Héran, 2013), mais qui se heurte à un problème d'échelle : l'archéologue peut visiter au sol quelques aires repérées du ciel, mais pas la totalité du pays. Aux piétons que nous sommes, Ngram Viewer a le mérite d'apporter une vision synoptique. Pourquoi nous priver du droit de monter au beffroi ?

4. Dérives et ruptures sémantiques

Un autre grief pèse sur Ngram Viewer : il postulerait l'invariance sémantique sur la longue durée. Or, nous rappelle-t-on, un mot comme « expert » peut changer de sens en deux siècles (Chateauraynaud et Debaz, 2010). Sans doute, mais qui pouvait en douter ? Les renvois de Ngram Viewer à Google Books confirment ces changements de sens. Ils mettent aussi sur la voie de découvertes qui n'ont rien d'intuitif. Trois exemples suffiront.

(11) Ces renvois ne livrent pas le contexte immédiat des expressions traitées par Ngram Viewer mais les titres des ouvrages scannés en amont par Google Books, avec un accès variable au texte. La fonction de recherche mobilisée ne suit pas les règles de Ngram Viewer mais celles de Google Search : confusion des capitales et des minuscules, chevauchement possible sur deux phrases, redressement automatique des graphies approximatives, etc.

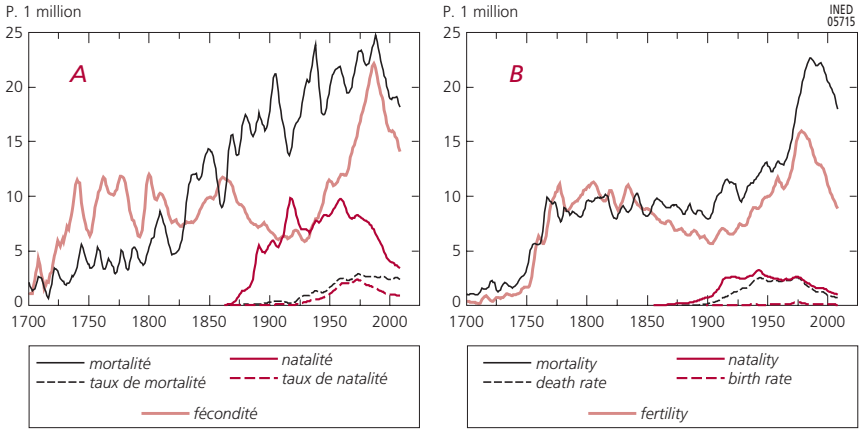
« Élite » existait au XVIIIe siècle mais Google Books révèle qu'il désignait rarement une élite sociale : c'était plutôt une anthologie, un florilège (par exemple, une « élite de poèmes », une « élite de bons mots »). Pour déceler ces ruptures de sens, le lecteur ne mobilise pas des références culturelles extérieures : c'est Ngram Viewer qui, sur ce point, enrichit sa culture.

Autre exemple, « données sensibles » était une notion philosophique touchant au dualisme du sensible et de l'intelligible (*sense data* et non pas *sensitive data*). Les philosophes continuent de l'entendre ainsi. Pour les sociologues, les démographes et les juristes, c'est tout autre chose : les « données sensibles » au sens de la loi Informatique et libertés de 1978 sont les données personnelles qui risquent de dévoiler des appartenances politiques, religieuses, ethniques, de santé, etc.

« Mortalité » / « natalité » : pourquoi ces faux jumeaux en français et pas en anglais ?

La notion de « mortalité » relève à l'origine du registre moral ou théologique : c'est le fait d'être mortel, « la sujétion à la mort ». Les références de Google Books rappellent qu'aux yeux des Pères de l'Église, le Christ incarné réunissait l'*æternitas* et la *mortalitas*. La *18e Provinciale* de Pascal évoque « l'homme sentant sa mortalité et son néant ». Certes, le sens pré-démographique de « mortalité » s'est développé dès le XVIIIe siècle (figure 4A), mais appliqué d'abord à la concentration des décès dans les temps d'épidémie ou chez les enfants en bas âge ; l'édition de 1738-1742 du *Dictionnaire de Trévoux* cite cette phrase aux accents bibliques : « la mortalité est sur les petits enfants », tel un fléau divin. Comme le calcul démographique ne contredit pas le fatalisme théologique, le même dictionnaire

Figure 4. « Natalité » *vs* « mortalité » en français, inconnu de l'anglais qui préfère associer *mortality* et *fertility*



Note : Fréquence pour 1 million, lissage pondéré d'ordre 5.
Source : Corpus francophone et anglophone de Ngram Viewer.

peut introduire à la suite la notion de « force de mortalité », mise en évidence à Londres par un certain « Jean Graunt » à partir des « billets de mortalité »…

Nul antécédent théologique, en revanche, au concept de « natalité ». Car l'homme est mortel, il n'est pas « natal ». Le mot « natalité » est une création savante tardive, qui n'accède à la visibilité dans le lexique français qu'à partir de 1862, avec, d'entrée de jeu, son acception démographique, parfaitement agnostique. Seul le temps qui passe nous fait croire aujourd'hui que « mortalité » et « natalité » sont des jumeaux.

Les Anglais, pour leur part, n'ont jamais cédé au démon de l'analogie, respectueux qu'ils sont du traitement dissymétrique de la naissance et de la mort par la théologie (figure 4B). C'est pourquoi, comme le savent les bons traducteurs, *mortality* n'a pas pour pendant *natality* mais *birth rate*, le « taux de naissance ». Les relevés de Ngram Viewer suggèrent toutefois que le concept le plus souvent couplé à *mortality* a longtemps été *fertility*, malgré une base de calcul différente. On laissera aux historiens de la démographie anglophone le soin de poursuivre cette comparaison.

5. Événements, perceptions, formulations

Quel rapport la courbe d'une expression dans Ngram Viewer entretient-elle avec le déroulement factuel des événements ? Il existe plusieurs cas de figure.

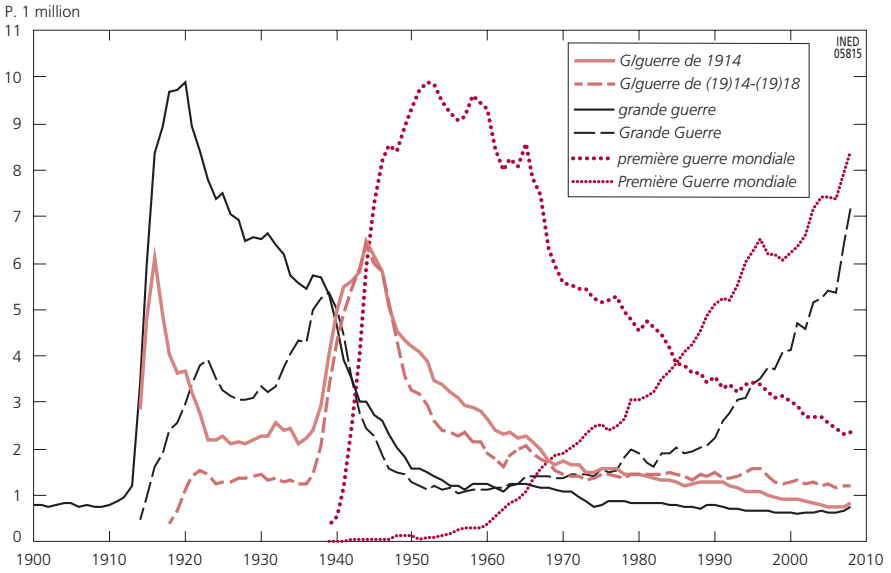
D'une guerre mondiale à l'autre, la noria des appellations

Face aux grands événements, le temps de réaction est court. Au XIX^e^ siècle, les documents épousent la courbe en dents de scie du choléra ou la série des Expositions universelles. Mais le lien peut être complexe entre l'événement et les représentations, comme le montrent les appellations successives de la Première Guerre mondiale depuis cent ans (figure 5).

Lorsqu'elle éclate en août 1914, la guerre est officiellement « Guerre de 1914 », tant l'état-major veut croire qu'elle s'achèvera avant Noël. Mais elle dure, et la population ravive un vieux surnom, « grande guerre », promu en « Grande Guerre » à mesure que le conflit s'amplifie. L'armistice suscite « guerre de 1914-1918 » ou de « 14-18 », au succès limité. Mais en 1940, tout est bouleversé : on relance les millésimes pour distinguer les deux conflits et voici « Grande Guerre » déclassé : ce n'était qu'une « première guerre mondiale » ! À compter des années 1960, le travail de la mémoire et de l'histoire apporte la consécration des capitales : « Première Guerre mondiale » progresse, puis « Grande Guerre », l'une tirant plutôt vers la macro-histoire, l'autre vers la guerre vécue. Le corpus s'interrompt avant les publications du centenaire (où « Grande Guerre » semble avoir triomphé). Mais, au total, le tracé très cohérent des courbes sur près d'un siècle atteste avec éclat la qualité des données.

La fortune des mots s'inscrit dans une histoire réflexive, ponctuée de révisions déchirantes (la dernière guerre n'est pas la dernière), de rebonds

Figure 5. Principales désignations et graphies de la Première Guerre mondiale en France



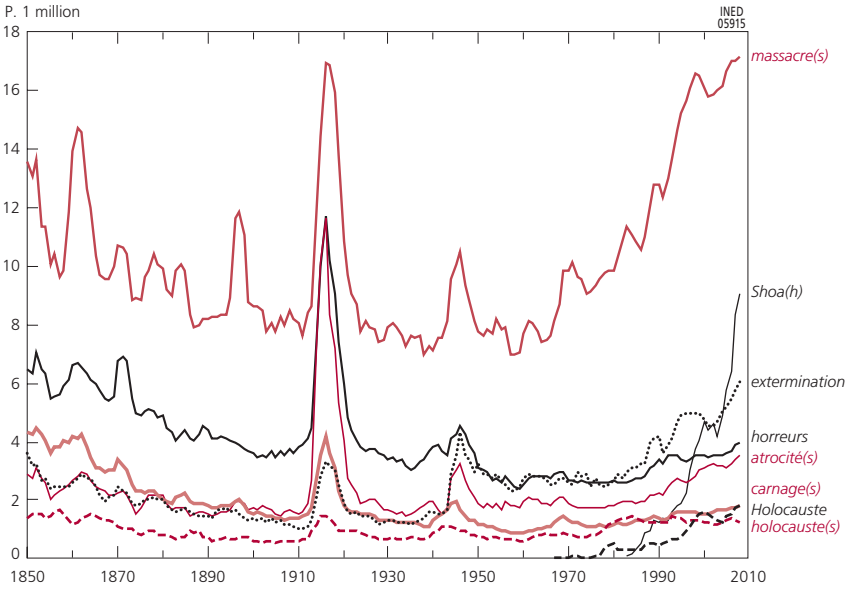
Note : Fréquence pour 1 million, lissage d'ordre 1.
Source : Corpus francophone de Ngram Viewer.

historiographiques et de retours mémoriels, lesquels peuvent revenir sur de longues périodes d'indifférence. Sur le coup, les Français furent plus sensibles aux atrocités de la Première Guerre mondiale qu'à celles de la Seconde, y compris à la Libération (figure 6). Il a fallu le travail des mémorialistes et des historiens pour dessiller les yeux des générations suivantes [(12)].

La noria des étiquettes peut avoir d'autre ressort que la simple suite des événements, comme en témoigne la désignation changeante des pays décolonisés (figure 7). Avec la poussée des indépendances, « pays sous-développés » est disqualifié. En 1952, Alfred Sauvy et Georges Balandier lancent simultanément « Tiers Monde », qui entérine la revendication de dignité des pays intéressés et connaît un succès planétaire, avant de prendre à son tour une teinte condescendante. D'autres appellations prennent le relais, sans faire écho cette fois à un événement moteur. Leur généralisation atteste surtout l'autorité croissante des organisations internationales sur le vocabulaire officiel, autorité désormais plus influente que celle des intellectuels.

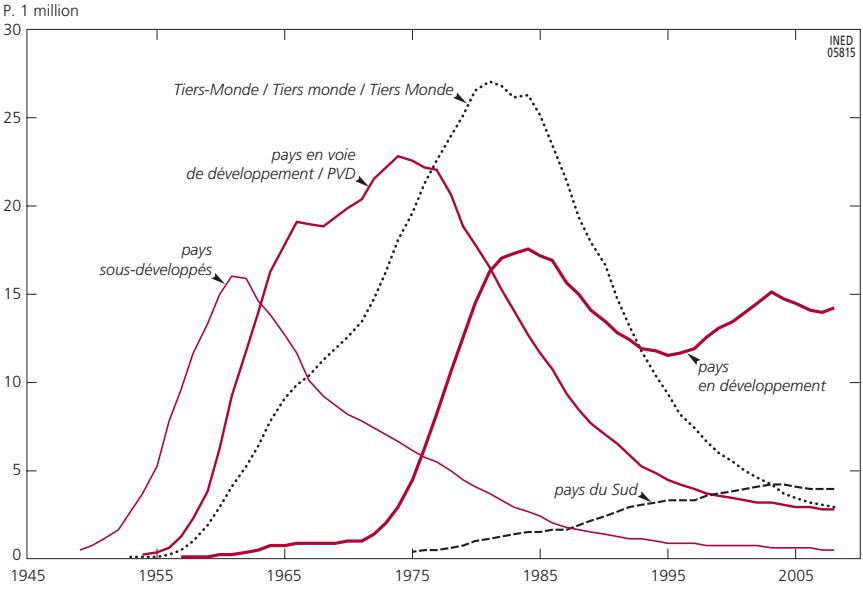
(12) Seul « extermination » fut plus employé dans les années 1945-1947 qu'en 1915-1918. En revanche, c'est surtout pendant la Grande Guerre qu'on trouve « holocauste(s) » dans les récits des témoins et des contemporains. Cette formule religieuse appartient au langage relevé et renvoie aux sacrifices complets du bétail dans la Bible. « Holocauste » au singulier et avec capitale, est tardif. C'est seulement à partir de 1978 qu'il se répand pour désigner l'extermination des juifs d'Europe par les nazis. Depuis le maître-film de Claude Lanzmann (1985), « Shoa(h) », repris d'un terme officiel israélien, est de rigueur en français, mais il est rare dans l'aire anglophone, où *Holocaust* reste de loin la formule la plus courante.

Figure 6. Les horreurs de la guerre : une vision asymétrique entre les deux guerres mondiales



Note : Fréquence pour 1 million, lissage d'ordre 1.
Source : Corpus francophone de Ngram Viewer.

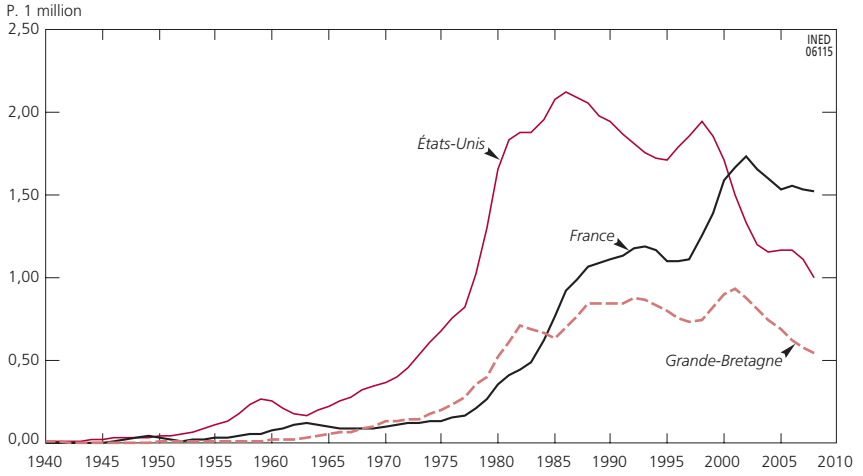
Figure 7. Les différentes désignations et graphies des pays anciennement colonisés



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

Le baby-boom offre un exemple de décalage chronologique entre réalités immédiates et perceptions (figure 8). Apparue aux États-Unis dès 1946, l'expression fait une timide apparition en 1949 dans le corpus français et reste discrète jusqu'à la fin des années 1970. Elle décolle en France quinze ans après les États-Unis, cinq ans après le Royaume-Uni. Elle dépasse la courbe américaine en 2000, avant d'entamer sa chute. Trajectoire très éloignée, on le voit, de la chronologie réelle du baby-boom, qu'on définisse cette dernière par le surcroît de naissances (1946-1974) ou par la hausse du taux de fécondité (1942-1965).

Figure 8. La perception tardive du « baby-boom » : États-Unis, France, Grande-Bretagne



Note: Fréquence pour 1 million, moyenne pondérée sur 5 ans (poids 1-2-3-2-1).
Source: Corpus francophone et anglophones de Ngram Viewer, toutes variantes graphiques réunies.

Pourquoi ce décalage ? On avancera ici quelques hypothèses. Un premier filtre est la réticence des scientifiques à user d'une expression qui avait le triple tort d'être d'importation américaine, de reposer à l'origine sur une métaphore boursière et de séduire les journalistes. Or la presse est exclue du corpus. Un second filtre tient aux réticences des démographes à admettre la réalité du baby-boom. Alfred Sauvy, le premier directeur de l'Ined, mit deux ou trois ans à reconnaître que la hausse spectaculaire de la natalité entamée en 1946 (200 000 naissances de plus dans cette seule année !) allait au-delà d'un rattrapage d'après-guerre (Lévy, 1990). Non sans malice, le biographe de Sauvy soupçonne à cette réticence une raison forte : si les Français redressaient d'eux-mêmes la natalité, quel besoin avait-on d'un institut de démographie ? Il fallut l'insistance des jeunes chercheurs de l'Ined pour convaincre Sauvy du caractère durable du baby-boom : le comportement des Français avait bel et bien changé. Menace pour l'Ined, le baby-boom devint son sauveur : les autorités demandèrent à ses experts d'anticiper le surpeuplement qui en résulterait dans les logements et les écoles.

Enfin, l'intérêt pour le baby-boom survit à l'événement en raison de ses effets à long terme sur le vieillissement de la population, alors même que sa contribution à la natalité n'est plus d'actualité. Certes, le baby-boom avait commencé par rajeunir la population, mais au fil du temps, on comprend qu'il contribue désormais à la vieillir, et cette prise de conscience tardive semble avoir soutenu sa courbe de diffusion. De nos jours, le baby-boomer vieillissant est à son tour intégré dans nos représentations et suscite moins d'intérêt.

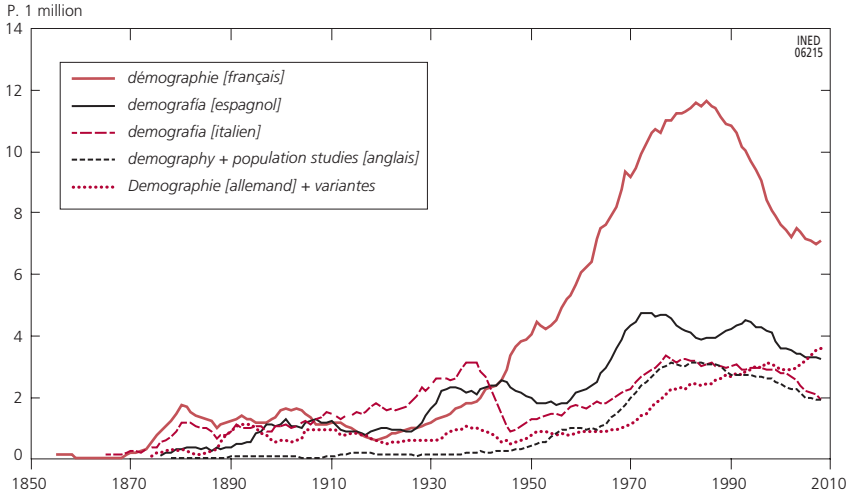
II. Essor et déclin de la démographie classique

Fort des instructions et précautions énumérées précédemment, déjà nourries d'exemples démographiques, on peut s'interroger plus avant sur l'évolution récente et probable de la démographie comme discipline. Comparée à la production écrite des autres nations couvertes par Ngram Viewer, celle de la France se signale par un intérêt précoce et soutenu pour les questions démographiques, mais qui semble avoir décliné dans les dernières décennies.

1. Les trois naissances de « démographie »

Le mot « démographie » en offre une première illustration (figure 9). Ngram Viewer repère quelques occurrences à la suite des *Éléments de statistique humaine*, ou *Démographie* d'Achille Guillard, l'inventeur du mot. Mais cette première

Figure 9. « Démographie » : un usage plus intense en France mais en net déclin depuis 30 ans



Note : Termes retenus en allemand : Demographie + Demografie + Bevölkerungswissenschaft + Bevölkerungsforschung. Fréquence pour 1 million, lissage d'ordre 3, toutes casses réunies.
Source : Corpus de Ngram Viewer en 5 langues (version 2012).

naissance reste sans lendemain. Il faut attendre la défaite de 1870 pour que la rivalité avec l'Allemagne pousse à relancer la « démographie », dont la visibilité reste encore modeste (elle repose sur quelques individus, dont le petit-fils de Guillard, Alphonse Bertillon). La troisième naissance, dans les années 1920, est la bonne, et c'est aussi la moins connue. Le label « démographie » ne cesse depuis lors de s'intensifier, sans césure visible pendant la Seconde Guerre mondiale, et ce jusqu'à l'apogée des années 1970 et 1980. Survient ensuite un net déclin, dont on reparlera.

Tardif si l'on retient sa date de naissance, l'usage de « démographie » reste précoce sur les scènes française et italienne. Les autres pays couverts par les corpus de Ngram Viewer ont ignoré la première et la deuxième naissance de la démographie (vers 1855 et vers 1872). Après la Seconde Guerre mondiale, la France est de loin le pays qui en fait l'usage le plus fréquent dans la culture écrite. Aucune expression alternative (telle *population studies*) ne comble l'écart avec la France dans l'aire anglophone, pas plus que les diverses désignations de la démographie en allemand. En France, l'intérêt pour la démographie a dépassé les cercles de la science sociale ou de la statistique nationale pour gagner la sphère publique.

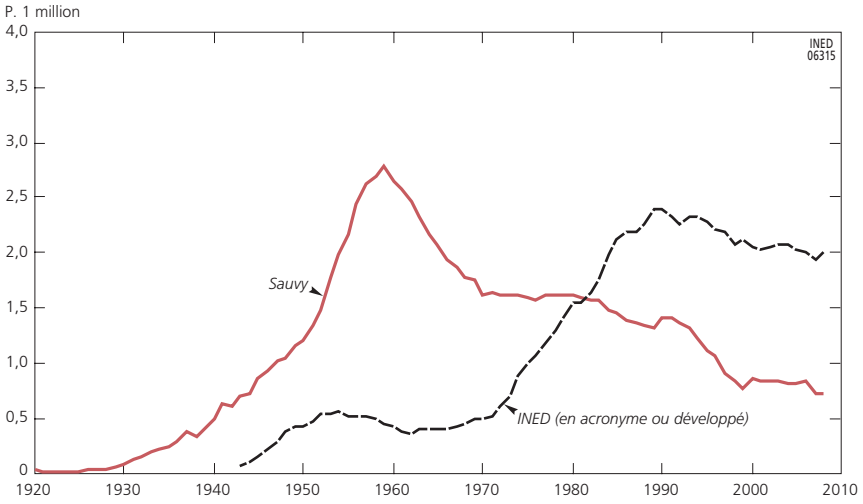
La création de l'Ined en octobre 1945 a sans doute contribué à ce résultat. L'idée d'asseoir la recherche démographique sur un institut national qui puisse informer la décision publique revient au professeur Robert Debré, qui l'avait proposée au Conseil national de la résistance (Rosental, 2003). Alfred Sauvy se voit confier la direction de l'Ined (alors I. N. E. D…), après avoir postulé sans succès à celle de l'Insee. Il dirige l'établissement jusqu'en 1962. Auteur de multiples essais, billettiste d'un grand quotidien, élu au Collège de France en 1959, Sauvy jouit d'une renommée qui éclipse celle de l'institut jusqu'au début des années 1970 (figure 10).

2. Le déclin du vocabulaire démographique : artefact ou réalité ?

Nombreuses sont les courbes liées au lexique des démographes qui se retournent dans les années 1980 ou 1990, suggérant que l'âge d'or de la démographie serait derrière nous, tant dans la sphère francophone que dans la sphère anglophone (figure 11). Parmi ces courbes en déclin, on compte des expressions aussi élémentaires que « démographie », « population », « natalité », « données démographiques », « croissance démographique », « remplacement des générations », « transition démographique » ou « démographie historique »… Le recul est encore plus net pour des formules techniques, comme « structure par âge » ou « répartition par âge », et leur équivalent anglais, *age structure* (figure 12).

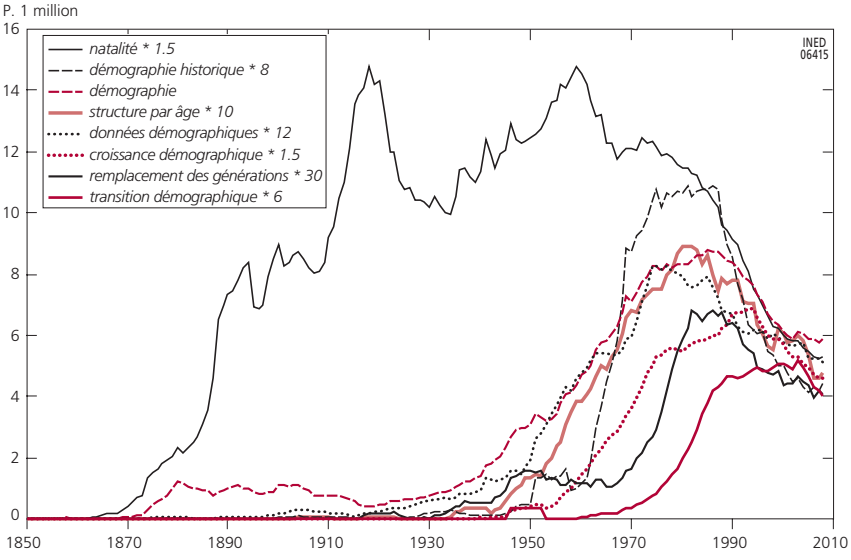
Ces retournements se sont confirmés depuis, ce qui donne à penser qu'un observateur disposant du même outil d'observation vers 1990 aurait pu jouer aisément les prophètes en se contentant de prolonger la tendance. Mais c'est une chose de percevoir le rétrécissement de la place relative de la démographie, c'en est une autre de pouvoir l'interpréter.

Figure 10. Alfred Sauvy et l'Ined : l'aura d'un homme, l'essor d'une institution



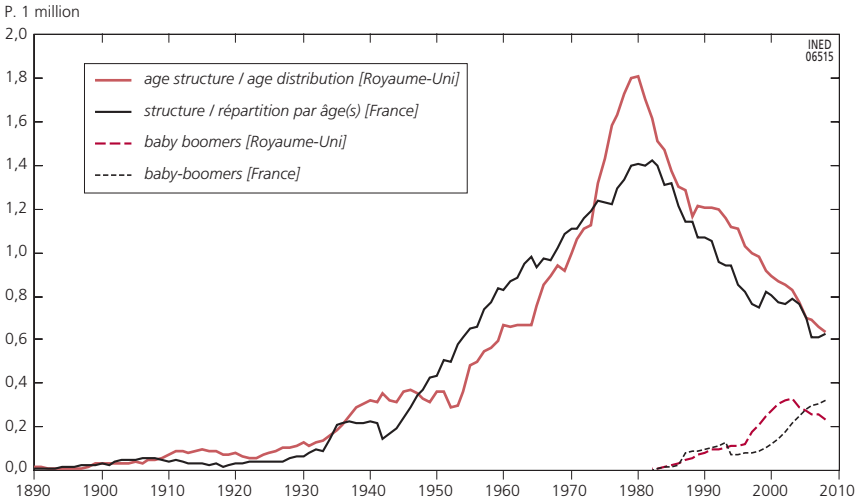
Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

Figure 11. Choix d'expressions démographiques en net recul dans les dernières décennies



Note : Pour faciliter la comparaison des courbes de fréquence avec celle du mot « démographie », les autres fréquences ont été multipliées par un facteur d'échelle indiqué dans la légende. Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

Figure 12. Structure par âge et baby-boomers en France et au Royaume-Uni



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone et anglophone britannique de Ngram Viewer.

S'agit-il d'un artefact ? Selon les auteurs d'une étude sur le vocabulaire anglais du climat, les enregistrements bruts de Google Books seraient de plus en plus perturbés par un « bruit » non significatif, lié à l'afflux de signes étrangers à la langue anglaise : « *growing numbers of characters, data, and other non-English 'noise'* » (Bentley *et al.*, 2012). De là un gonflement artificiel du corpus en fin de parcours, qu'il importerait de corriger en calant le calcul des fréquences non pas sur le nombre total de *n-grams* mais sur le nombre d'occurrences d'un mot bien anglais – tant qu'à faire, le plus fréquent de la langue anglaise : l'article défini *the*. C'est l'hypothèse à laquelle se réfèrent Bijak et ses collègues, intrigués de voir « démographie » et, plus encore, *demography* reculer en termes relatifs mais progresser en valeur absolue (Bijak *et al.*, 2014). Que faut-il en penser ?

Acerbi (2013) a testé le calage des fréquences sur l'article *the*, sans grand résultat : alors que la normalisation totale prônée par Ngram Viewer tend à ralentir la croissance du nombre de mots dans les dernières décennies, la normalisation par *the* l'accélère légèrement. L'écart reste minime, sans rapport avec le déclin récent des mots « démographie » ou *demography*. Mais faut-il s'étonner que l'hypothèse tourne court ? Elle revient à croire que les comptes seraient bons si l'on pouvait nettoyer le vocabulaire national des apports étrangers. L'illusion redouble quand on prend pour schibboleth le mot *the*. Comment peut-on ignorer que l'article défini est souvent omis en anglais et que cette pratique du *zero article* varie selon les lieux, les époques et même les personnes ? De 1945 à 2000, la fréquence de *the* dans Ngram Viewer a chuté de 11 % en anglais britannique et de 14 % en anglais américain, alors que les articles définis français (le, la, les, l') n'ont subi aucune baisse. Sur le choix entre deux expres-

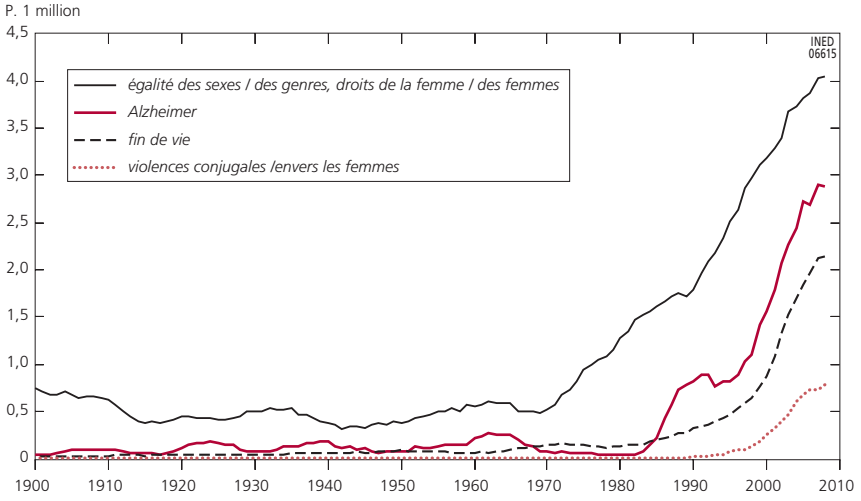
sions aussi banales que *in hospital* ou *in the hospital*, une requête révèle que, depuis les années 1880, les pratiques n'ont cessé de diverger : l'article *the* est désormais employé aux États-Unis dans 83 % des cas, contre 39 % au Royaume-Uni, alors que les taux étaient identiques un siècle plus tôt !

Il est donc vain de vouloir normaliser les fréquences sur Ngram Viewer en prenant un repère aussi sujet à variations que l'article *the*. Nulle montée artificielle d'un « bruit étranger » n'explique le doublement du nombre d'ouvrages de langue française dans Ngram Viewer entre 1990 et 2000. C'est un phénomène général. Les pays occidentaux impriment toujours plus de textes sur plus de sujets, sans compter le fait que les dernières productions numérisées viennent du catalogue des éditeurs et non plus seulement des bibliothèques.

Mais la meilleure réfutation à la théorie du « bruit étranger croissant » est le contraste observé entre la baisse des notions démographiques classiques et la hausse de nombreux thèmes à caractère social, civique ou éthique. Si l'artefact du bruit produisait les biais soupçonnés par Bentley ou Acerbi, il devrait affecter tous les lexiques, pas seulement celui de l'analyse démographique. Parmi les thèmes en pleine progression figurent la fin de vie, les discriminations, l'égalité des sexes, les violences intrafamiliales (figure 13), mais aussi et surtout la santé (figure 14).

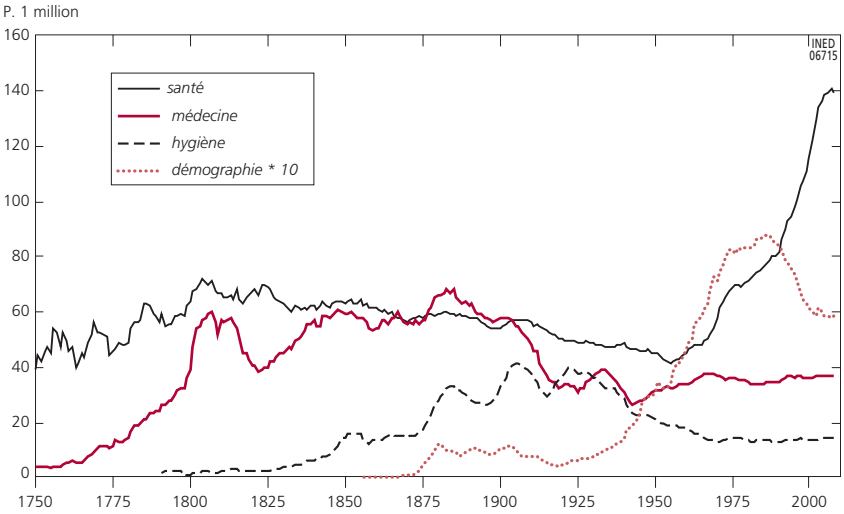
D'autres thèmes connaissent une progression fulgurante dans les écrits francophones : l'immigration, les questions identitaires ou religieuses, les « valeurs républicaines », le « lien social »… Or ces thèmes, s'ils ne relèvent pas des objets classiques de la démographie, sont aujourd'hui au cœur des préoccupations des démographes.

Figure 13. Quelques thèmes de préoccupation connexes à la démographie, en forte hausse dans les dernières décennies



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

Figure 14. Santé, médecine, hygiène, démographie : des priorités changeantes sur deux siècles et demi



Note : La fréquence du mot « démographie » a été multipliée par 10 pour faciliter la comparaison des profils. Fréquence pour 1 million, lissage d'ordre 3.

Source : Corpus francophone de Ngram Viewer.

3. Une mesure des liens science/société et non pas un suivi de la science de pointe

On dira que ces nouveaux centres d'intérêt débordent la sphère des revues scientifiques et concernent la société en général. La question du sort réservé aux revues scientifiques par Ngram Viewer est soulevée par Bijak et ses collègues (2014). Ils soupçonnent le corpus d'avoir intégré le vocabulaire des collections d'ouvrages de l'Ined sans faire de même pour les revues anglophones. De fait, on l'a vu, Ngram Viewer a bel et bien écarté les périodiques. Mais pourquoi ce biais se serait-il aggravé dans les dernières décennies ?

On rejoint ici la question de fond. Si l'hypothèse avancée revient à dire que le vocabulaire scientifique tend désormais à s'enfermer dans des publications à diffusion restreinte, à savoir des revues et non des collections d'ouvrages, ce phénomène appelle une autre interprétation : ce ne serait pas un biais de sélection mais l'indicateur d'un réel isolement. Car un vocabulaire scientifique devenu incapable de franchir le seuil de visibilité fixé par Ngram Viewer, y compris l'ordre de grandeur minimal (une proportion pour 100 millions), a toutes chances d'être déconnecté de la culture écrite générale. Or c'est là une propriété essentielle de Ngram Viewer : il n'a pas vocation à suivre les derniers progrès de la science mais à observer sa capacité à pénétrer la culture écrite (dans le jargon actuel, il contribue à une « mesure d'impact »). Du coup, il ne saurait documenter la compétition entre les chercheurs ou les écoles : il éclaire le rapport science/société. Bijak *et al.* espéraient de Ngram Viewer une

confirmation de la supériorité de certaines méthodes (analyse longitudinale, multiniveau, simulations, etc.). Mais, outre le fait que ces méthodes ne sont pas propres à la démographie, c'est utiliser Ngram Viewer à contre-emploi, avec un résultat forcément mitigé. Pour qui veut analyser l'évolution du contenu des revues scientifiques, il existe des outils bibliométriques plus adéquats.

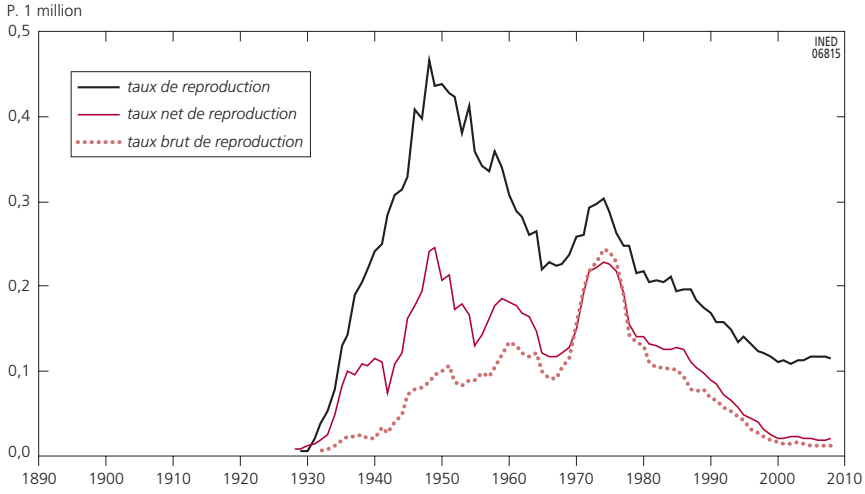
4. Recul général de l'analyse démographique, au profit d'analyses plus ciblées

On peut préciser à présent le diagnostic sur le recul des thématiques démographiques. Il est patent pour les expressions techniques définies dans les dictionnaires et les traités[13] (Pressat, 1979 ; Caselli *et al.*, 2001-2004 ; Meslé *et al.*, 2011). On le vérifie en reprenant une à une les grandes rubriques de la démographie, quitte à se contenter de commentaires succincts.

Fécondité

Dans le domaine de la fécondité, l' « indice synthétique de fécondité » est en chute libre et la « somme des naissances réduites » réduite au silence (courbes non reproduites ici). Seul l' « indicateur conjoncturel de fécondité » garde ses adeptes, peut-être parce que son intitulé dit bien sa fonction. Quant au « taux de reproduction », promu dans l'après-guerre, il n'a pas su se reproduire, malgré le sursaut des années 1970 (figure 15). Plus techniques encore, « taux net » et « taux brut de reproduction » sont mourants, car les effets perturbateurs de la mortalité avant l'âge de reproduction sont désormais négligeables.

Figure 15. Effacement des taux de reproduction

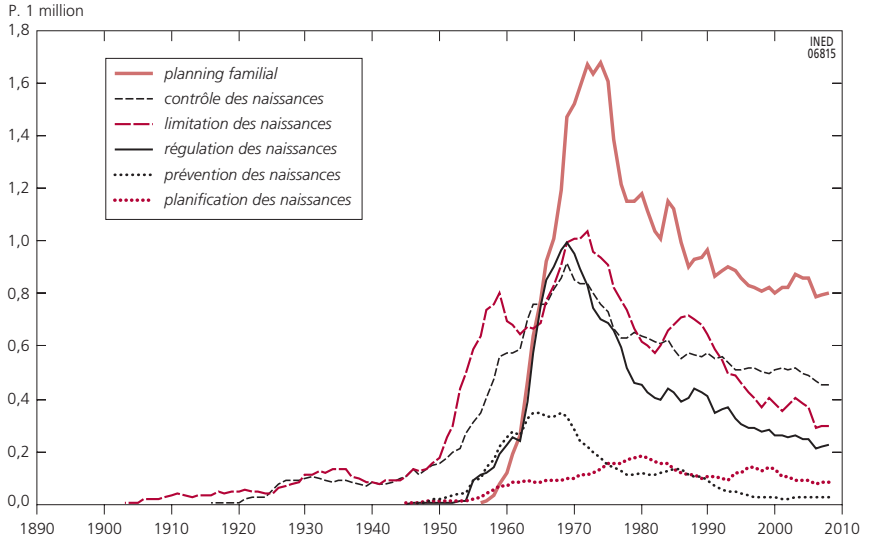


Note : Fréquence pour 1 million, lissage d'ordre 3, toutes casses réunies.
Source : Corpus francophone de Ngram Viewer.

(13) Le dictionnaire de Pressat a une volonté normative plus affichée que les deux autres publications.

L'intérêt pour la limitation des naissances et les méthodes contraceptives a fortement varié (figure 16). Retardée par l'essor initial du baby-boom mais relancée par les avancées techniques, la poussée militante devient spectaculaire dans les années 1960 et culmine au début des années 1970, avant de retomber lourdement. Les courbes semblent en passe de rejoindre un rythme de croisière. L'intérêt s'est déplacé vers la lutte contre la stérilité et l'infertilité, avec une nouvelle pointe pour l'assistance à la procréation (figure 17).

Figure 16. Limitation des naissances et planning familial : la lente érosion depuis les années militantes



Note : Fréquence pour 1 million, lissage d'ordre 3, toutes casses réunies.
Source : Corpus francophone de Ngram Viewer.

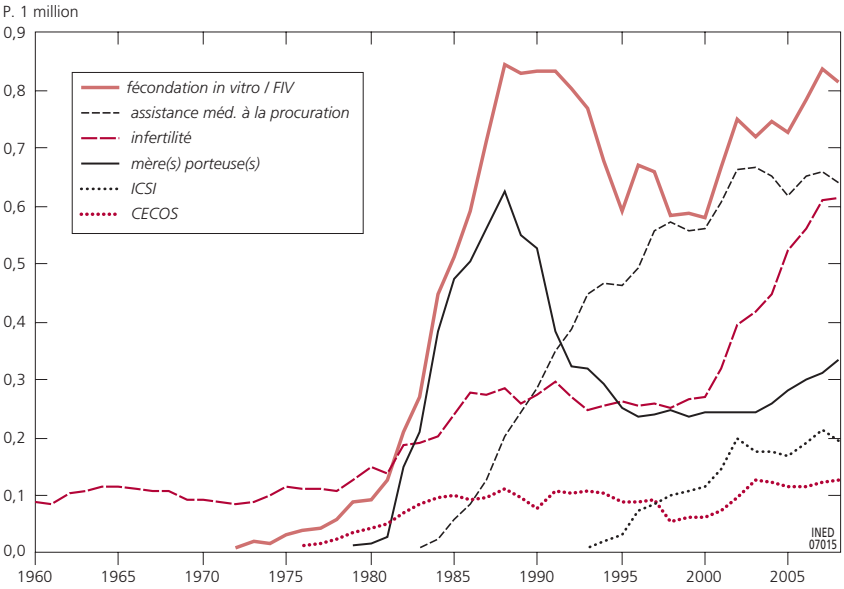
Démographie historique

Forgée par Louis Henry, portée par la vogue de l'« histoire sérielle », la « démographie historique » appartient désormais à l'histoire (figure 18). Au-delà des « registres paroissiaux » et de la « reconstitution des familles », une nouvelle génération d'historiens et d'économistes proches des démographes recourt désormais à une palette de sources et de méthodes dont l'approche lexicale peut difficilement cerner les contours proprement démographiques.

Mortalité

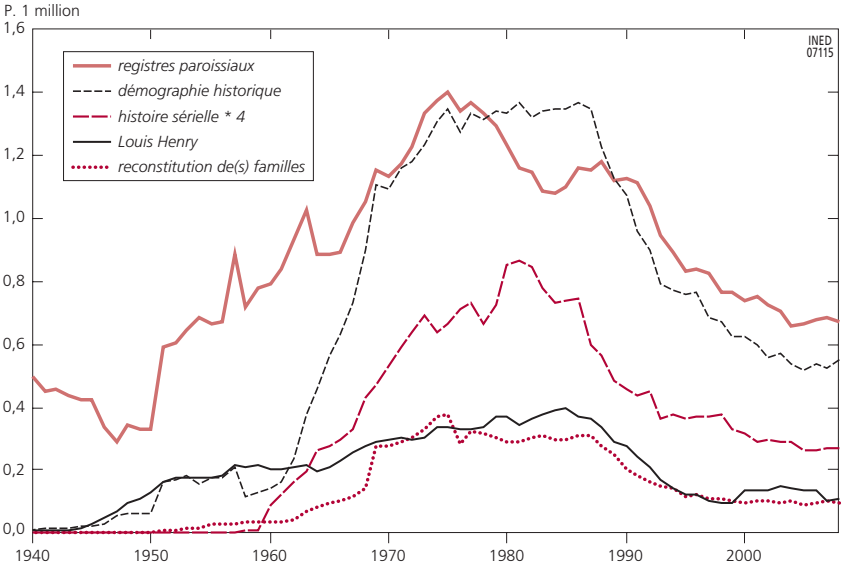
Alors que « taux de mortalité » se tasse depuis les années 1970, « espérance de vie » poursuit son envol depuis l'après-guerre (Héran, 2013). L'étude de la mortalité se porte mieux que celle de la fécondité. Mais les « causes de mort », devenues « causes de décès », objet d'une vieille émulation entre Français et Britanniques,

Figure 17. Infertilité et assistance médicale à la procréation



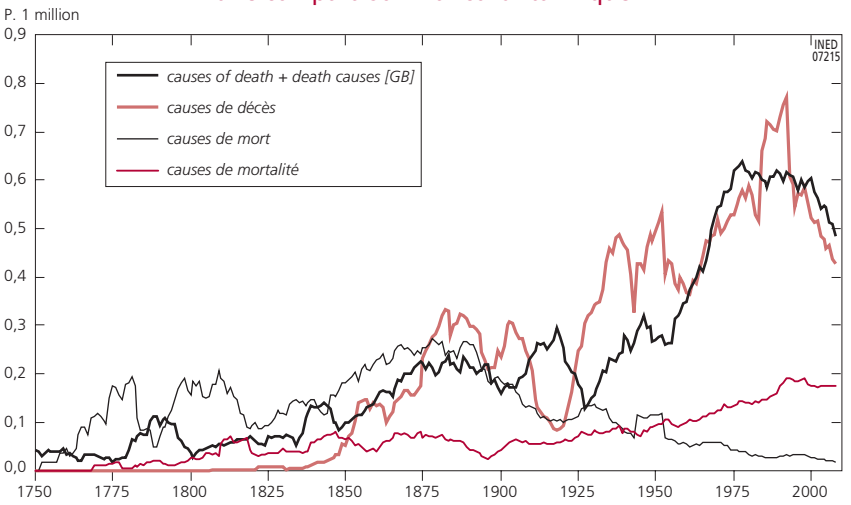
Note : Fréquence pour 1 million, lissage d'ordre 3. L'expression « assistance médicale à la procréation » inclut « procréation médicalement assistée » mais non « PMA » (qui signifie aussi « pays les moins avancés »);
CECOS = Centre d'étude et de conservation des œufs et du sperme;
ICSI = injection intra-cytoplasmique de spermatozoïde.
Source : Corpus francophone de Ngram Viewer.

Figure 18. Le vocabulaire de la démographie historique



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

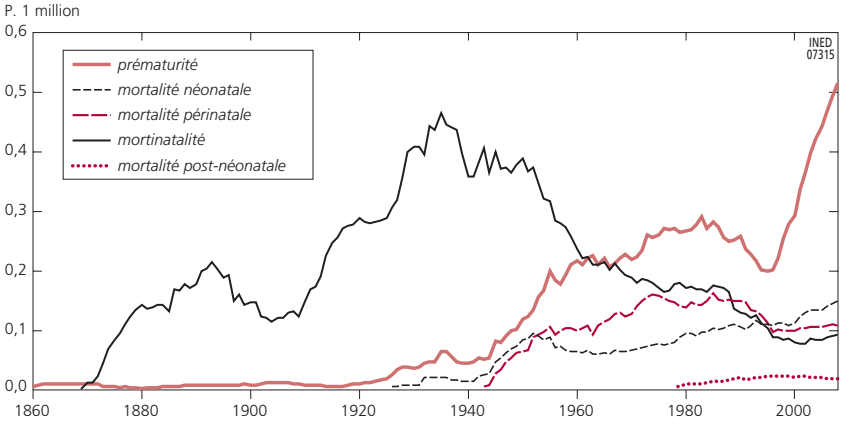
Figure 19. « Causes de décès » et termes alternatifs : une comparaison franco-britannique



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone et britannique de Ngram Viewer.

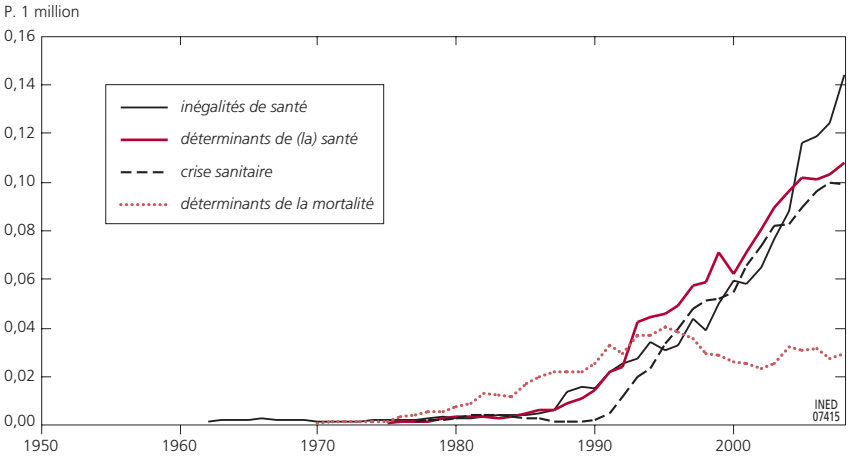
semblent plafonner (figure 19). Les composantes de la mortalité périnatale, en revanche, sont un sujet neuf (figure 20). Les « déterminants de la mortalité » intéressent moins que les « déterminants de la santé » (figure 21). Des thèmes comme « crise sanitaire » ou « inégalités de santé » progressent : ils débordent le cadre des études démographiques, poussant les spécialistes de la mortalité à se rapprocher d'analyses plus causales, en lien avec la santé publique et l'épidémiologie.

Figure 20. Nouvelles explorations de la mortalité infantile : la mortalité périnatale



Note : Fréquence pour 1 million, lissage d'ordre 5. Les variantes avec trait d'union (comme « néo-natale ») ont été prises en compte.
Source : Corpus francophone de Ngram Viewer.

Figure 21. Inégalités de santé, déterminants de santé : un nouveau défi pour les études de mortalité

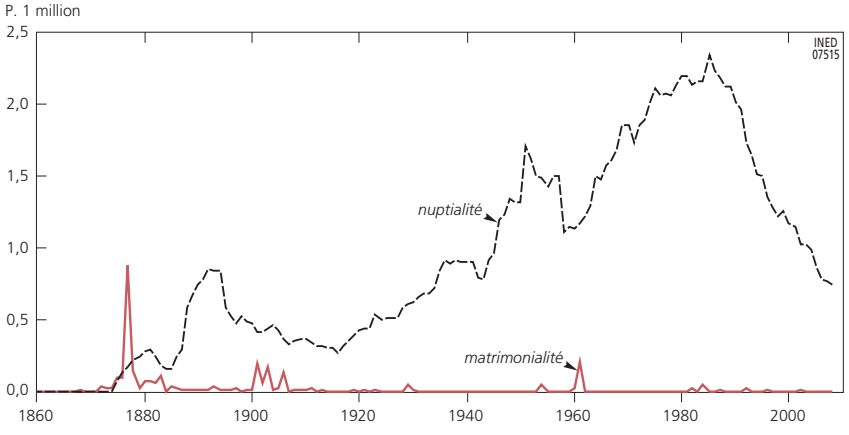


Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

Nuptialité : la fin du modèle matrimonial

La notion de « nuptialité » est antérieure à l'essor de la « démographie » comme discipline nommément désignée (figure 22). Elle supplante « matrimonialité » pour occuper une position de choix à la seconde naissance de la démographie, après la défaite de 1870. On ne doute pas alors que la nuptialité soit la clef de la reproduction, même si l'on soupçonne qu'elle dépend à son tour de la prospérité des campagnes et des prix agricoles. La Grande Guerre ayant ébranlé le marché matrimonial, l'intérêt pour la nuptialité remonte, en

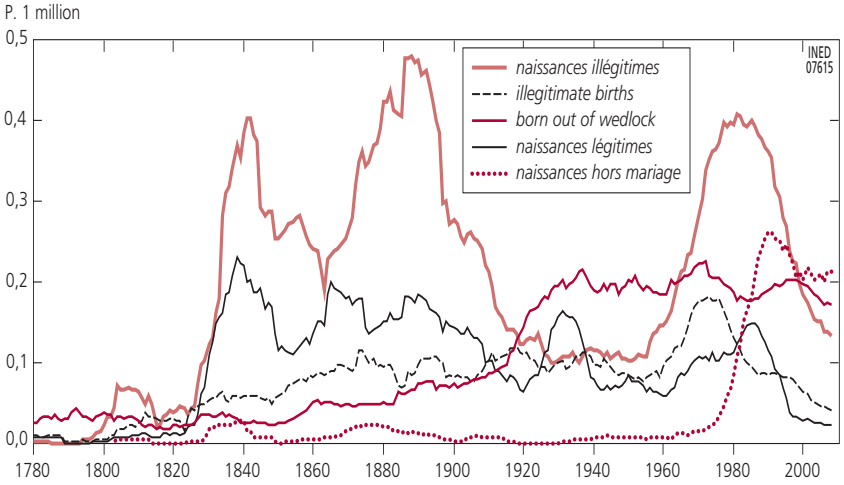
Figure 22. Essor et déclin de « nuptialité » depuis 1875



Note : Fréquence pour 1 million, lissage d'ordre 3, pour « nuptialité », pas de lissage pour « matrimonialité ».
Source : Corpus francophone de Ngram Viewer.

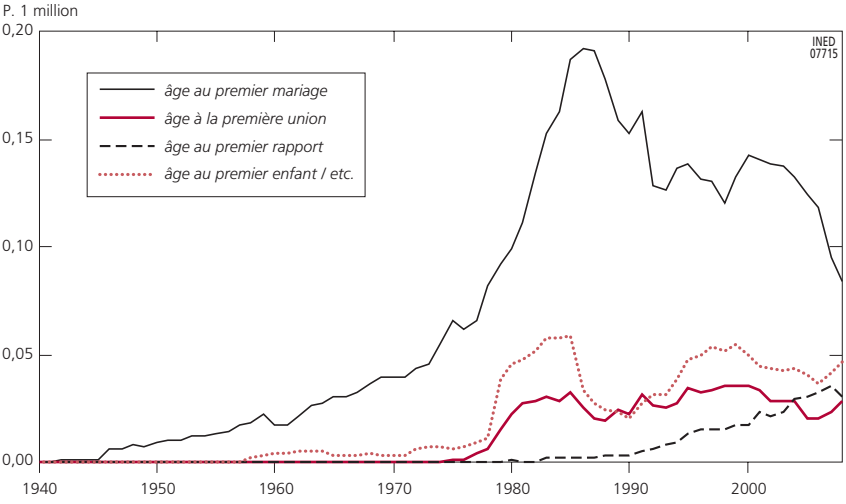
même temps que la hantise de la dénatalité et la redécouverte générale de la démographie : il progresse avec de fortes oscillations jusqu'aux années 1980. Commence alors la chute libre de « nuptialité », directement liée à la montée de la cohabitation. Les repères juridiques de l'analyse démographique vacillent : le calcul de la fécondité « par durée de mariage » devient obsolète, de même que la référence aux « naissances illégitimes » (figure 23). Progressent, en

Figure 23. Naissances hors mariage : la fin de l'illégitimité



Note : Fréquence pour 1 million, lissage d'ordre 5.
Source : Corpus francophone et anglophone de Ngram Viewer.

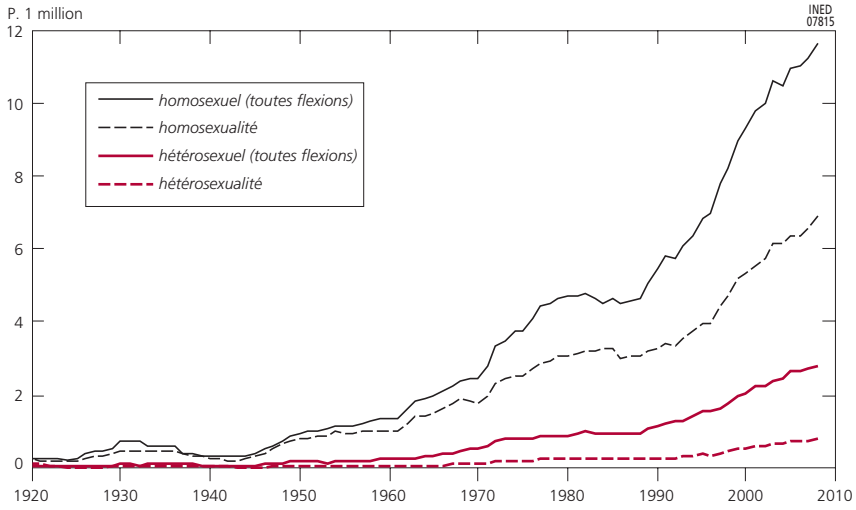
Figure 24. Passage à l'âge adulte : l'effacement du premier mariage



Note : Fréquence pour 1 million, lissage d'ordre 3. « âge au premier enfant / etc. » comprend également « âge à la première naissance », « âge à la première maternité » et « âge au premier accouchement ».
Source : Corpus francophone de Ngram Viewer.

revanche, les recherches sur le « passage à l'âge adulte » (comme en anglais *transition to adulthood*), dont les repères démographiques ne sont plus l'« âge au premier mariage » mais l'« âge au premier rapport », « à la première naissance » ou « au premier enfant » (figure 24).
À l'ébranlement de l'ordre matrimonial succède l'ébranlement de l'ordre sexuel. Le « Pacs » est âprement discuté en 1998, avant son adoption par le Parlement en novembre 1999. Le débat ne s'éteint pas avec la loi, sur fond d'un intérêt croissant pour la diversité des orientations sexuelles (figure 25). La clôture provisoire du corpus fin 2008 laisse de côté le débat sur le « mariage pour tous ».

Figure 25. Orientations sexuelles



Note : Fréquence pour 1 million, lissage d'ordre 3. L'option toutes flexions additionne les formes singulier/pluriel, féminin/masculin, substantif/adjectif.
Source : Corpus francophone de Ngram Viewer.

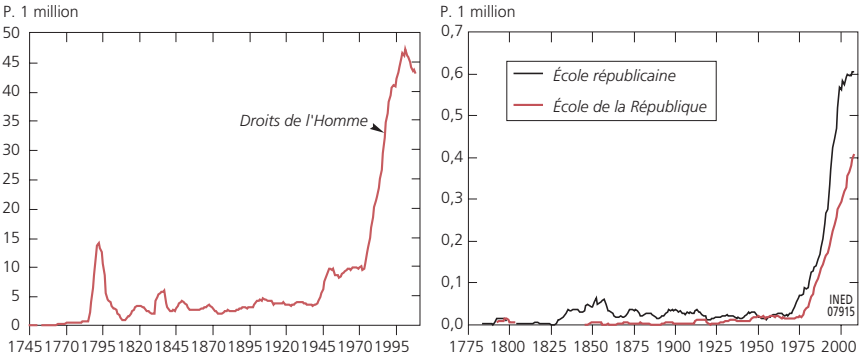
Immigration

La place grandissante des migrations internationales dans la production écrite des pays couverts par Ngram Viewer mériterait une étude en soi. On se contentera ici de brèves indications. Le vocabulaire technique des migrations (comme « solde migratoire ») suit le même parcours que celui des autres rubriques de la démographie : apogée dans les années 1970, déclin ensuite. Montent en flèche, en revanche, une série d'expressions agitées par le débat public. Elles touchent à l'ampleur des flux (« migration de masse », « vague migratoire »), à leur contrôle (« contrôle des frontières », « carte d'identité », « titre de séjour », « droit au séjour », « demandeurs d'asile »), au rapport des migrants à la religion (« communautarisme », « laïcité », « école de la République », « islamisme », « islam », etc.), à la question des origines (« origine étrangère », « identité de la France », « statistiques ethniques »), au modèle d'accueil et d'insertion (« intégration des immigrés »), aux valeurs de cohésion

sociale (« lien social », « État de droit », « valeurs communes », « devoir de mémoire »). La plupart de ces expressions prennent leur envol dans les années 1980 ou 1990, en lien direct avec l'intensification des débats politiques.

Tout cela était connu ou soupçonné. En revanche, Ngram Viewer jette une lumière crue sur les valeurs censées recréer de la cohésion sociale face au défaut d'intégration des migrants : « valeurs républicaines », « école républicaine », « laïcité », « droits des femmes », « droits de l'homme », « respect de la dignité », etc. On présente ces formules comme étant l'héritage de nos ancêtres, un trésor transmis continûment depuis la Révolution française ou la IIIe République. La réalité est tout autre : les générations précédentes les avaient entretenues à petit feu. Si ces valeurs nous sont revenues, c'est par l'extérieur, comme les « droits de l'Homme » après la défaite du nazisme, ou par un travail de création interne, comme le montre l'essor sans précédent d'expressions comme « École républicaine » ou « École de la République » (figure 26). Elles n'ont jamais été autant utilisées qu'à notre époque : proportionnellement vingt fois plus qu'au temps de Jules Ferry. Ces valeurs ne s'imposent pas d'elles-mêmes, il faut les réinventer. Les inculquer à ces « nouveaux venus » que sont les enfants d'une part, les migrants de l'autre (comme l'exige désormais la loi), implique une pédagogie plus innovante et exigeante qu'une simple leçon d'histoire.

Figure 26. Les valeurs de référence pour l'intégration, entre héritage, emprunt et invention

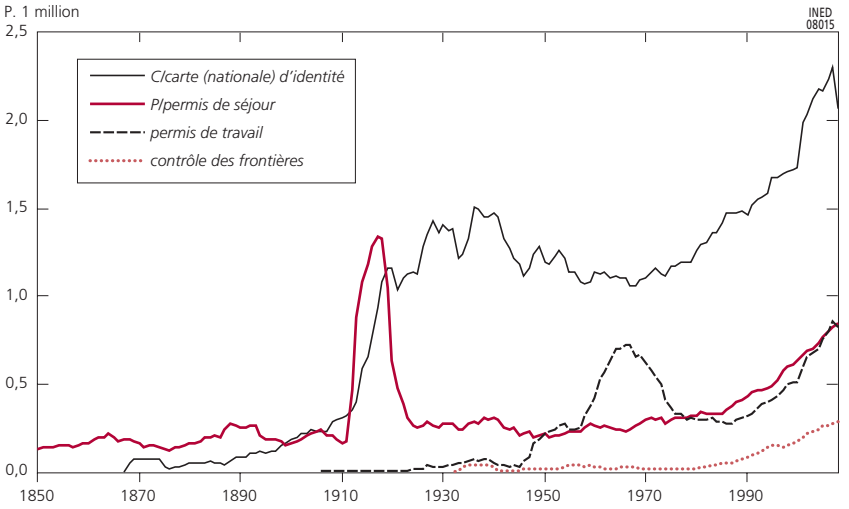


Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

L'histoire pèse bel et bien, mais pas toujours comme on l'imagine. Ngram Viewer souligne ainsi le rôle décisif de la Grande Guerre dans la mise en place du contrôle des identités, une mesure d'urgence qui s'est installée dans la durée (figure 27). Il confirme aussi que l'« assimilation » des étrangers ou des immigrés n'a jamais été un mot d'ordre puissant en France, contrairement à « intégration » (Héran, 2013). Elle apparaît à la veille de la Grande Guerre, lors de la crise des années 1930 et peu avant les débuts de la décolonisation, avec un écho très limité à chaque fois (figure 28). C'est seulement de façon rétrospective

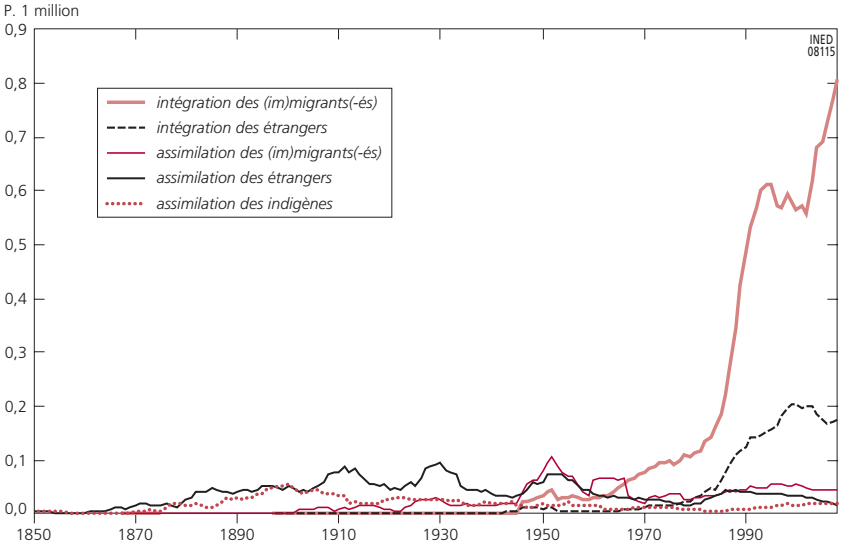
que l'on peut appliquer un modèle d'assimilation aux anciennes générations d'immigrés. Tout autre est la trajectoire du concept d'« intégration des immigrés ». Portée par les politiques nationales et européennes à partir des années 1980, elle continue de progresser. Dans les publications francophones,

Figure 27. Une innovation durable de la Guerre de 1914 : la carte d'identité



Note : Fréquence pour 1 million, lissage d'ordre 3. « C/carte » = addition de « Carte » et de « carte ».
Source : Corpus francophone de Ngram Viewer.

Figure 28. Le contraste entre « assimilation » et « intégration »



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus francophone de Ngram Viewer.

« intégration » pèse en 2008 vingt fois plus qu' « assimilation ». L'exploration du vocabulaire anglais (Héran, 2013) montre au contraire un usage à peu près interchangeable de *cultural assimilation* et *cultural integration*, facilité par l'acceptation des « identités à trait d'union » (ou *hyphenated*, comme *Korean-American*), y compris pour les « assimilés ».

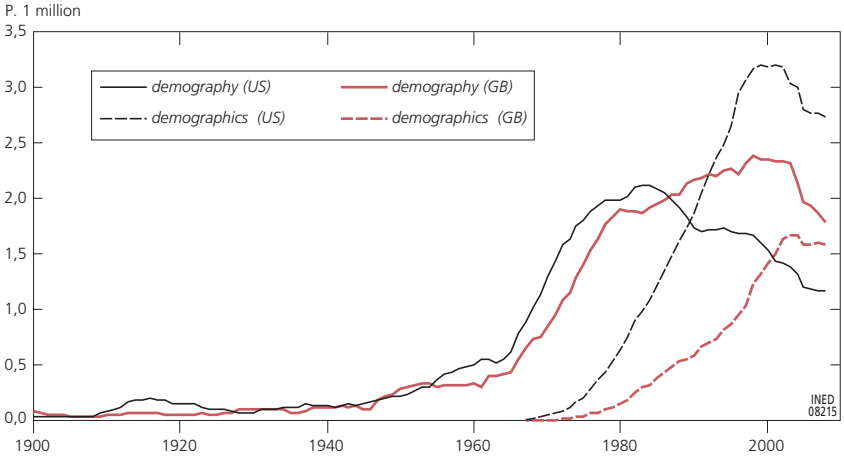
5. Des comparaisons internationales difficiles

La démographie est une science sociale à prétention universelle, dotée de longue date de thésaurus multilingues grâce aux efforts d'harmonisation des Nations unies. De son côté, Ngram Viewer permet d'explorer l'évolution du vocabulaire dans plusieurs langues à la fois. À l'expérience, pourtant, les correspondances de vocabulaire ne vont pas de soi.

Ainsi, l'expression « pyramide des âges », courante en français, est rare en anglais : *age pyramid* est inexistant, *population pyramid* peu usité. Les anglophones ignorent la métaphore architecturale : ils préfèrent *age structure* ou *age distribution*. Les démographes francophones parlent entre eux de « structures par âge(s) » ou « répartition par âge(s) », mais dans un registre plus technique. D'une langue à l'autre, les formules n'ont ni la même fréquence ni le même registre, ce qui complique les comparaisons.

Si « pyramide des âges » résiste au temps, les formules françaises sur la répartition par âge subissent depuis les années 1980 une chute aussi prononcée qu'en milieu anglophone, que ce soit au Royaume-Uni ou aux États-Unis (cf. *supra*, figure 12). Dans les trois pays, le vocabulaire de l'analyse démographique est en net recul depuis trois décennies. Aux États-Unis, ce déclin s'accompagne d'un basculement vers le marketing des générations : la

Figure 29. *Demographics* dépasse *demography* aux États-Unis et le talonne au Royaume-Uni



Note : Fréquence pour 1 million, lissage d'ordre 3.
Source : Corpus britannique et américain de Ngram Viewer.

catégorie-cible des *baby boomers* monte en flèche durant les années 1980 et 1990, au point que *demography* est désormais supplanté par *demographics*, l'activité de revente des données démographiques locales, elles-mêmes reliées au marché mondial des *big data*. Un mouvement analogue s'observe au Royaume-Uni, mais à un degré moindre (figure 29). La démographie française s'engagera-t-elle dans cette voie ? Peut-on imaginer que les *big data* puissent remplacer à terme l'état civil, le recensement et les enquêtes au même niveau de fiabilité ?

Conclusion : face au repli de la science démographique, le défi de s'ouvrir pour durer

La question du déclin possible de la démographie comme discipline n'est pas nouvelle : elle avait été posée il y a vingt ans par Jean-Claude Chasteland et Louis Roussel, qui, au soir de leur carrière, lancèrent une enquête en ligne sur le sujet (Chasteland et Roussel, 1997). Leurs conclusions restent valides. Centrés sur les concepts canoniques de la démographie, les relevés lexicaux attestent un recul certain, lié notamment à la « désinstitutionnalisation » des mœurs évoquée par Roussel. Maniés avec souplesse, ils disent autre chose : seule est menacée de repli une conception étroite de la démographie, enfermée dans les revues scientifiques et rechignant à étendre son périmètre aux disciplines connexes.

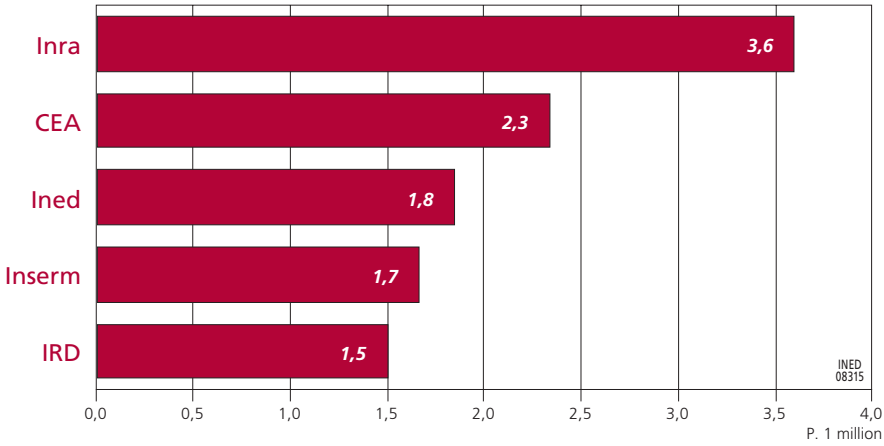
L'Ined joue un rôle dans ce processus d'ouverture. Il se signale par une forte présence dans la culture écrite explorée par Ngram Viewer, analogue à celle d'organismes de recherche de même ancienneté mais aux effectifs très supérieurs, comme l'Inserm, l'Inra, l'IRD ou le CEA (figure 30) – le CNRS étant hors concours par sa taille[(14)]. À quoi tient cette exceptionnelle visibilité ? On avancera trois hypothèses.

La première est que les autres organismes voient leur présence se réduire dans la culture écrite, parce qu'ils ont dû se tourner vers une production scientifique plus spécialisée, voire ésotérique, qui a du mal à percoler dans la société civile.

La seconde est que l'Ined, engagé lui-même dans ce mouvement, a su atténuer ses effets en pratiquant un traitement diversifié des questions de société. Pluridisciplinaire dès sa création (démographie, histoire, psychosociologie), l'institut a élargi son spectre (sociologie, économie, géographie, études de genre, santé publique), sans lâcher pour autant son « cœur de métier ». Ses enquêtes en témoignent. Menées en collaboration avec d'autres organismes et en associant de plus en plus les universitaires, elles restent ancrées sur la trilogie fécondité-mortalité-migrations, mais elles les relient désormais à des questions

(14) La figure est difficile à tracer, tant sont nombreuses les variantes graphiques des noms d'établissements (acronymes avec ou sans points, avec ou sans capitales, intitulés développés dépassant le seuil des cinq mots, etc.). C'est pourquoi on se limite ici aux dix dernières années, où les acronymes l'emportent.

Figure 30. Visibilité de l'Ined parmi quelques organismes de recherche français dans les dix dernières années du corpus : 1999-2008



Note : Fréquence pour 1 million, toutes casses réunies, moyenne des 10 années. Inra : Institut national de la recherche agronomique ; CEA : Commissariat à l'énergie atomique ; Inserm : Institut national de la santé et de la recherche médicale ; IRD : Institut de recherche pour le développement.
Source : Corpus francophone de Ngram Viewer.

de société vives ou sensibles : cohabitation sans mariage, devenir des enfants naturels, procréation assistée, interruption de grossesse, sexualité, excision, violences intrafamiliales, handicap, adoption, parcours des sans-abri, discriminations, décisions médicales en fin de vie (Héran, 2015).

La troisième hypothèse susceptible d'expliquer la visibilité de l'Ined est sa capacité d'adaptation, manifestée en trois phases successives depuis sa création.

Au commencement était le natalisme officiel, formellement inscrit dans les premiers statuts, mais débordé d'emblée par la créativité des pionniers recrutés par Sauvy, dont Louis Henry, Jean Bourgeois-Pichat, Pierre Depoid, Paul Vincent, auxquels s'ajoutèrent un pionnier des sondages comme Jean Stoetzel ou un historien comme Louis Chevalier.

Du milieu des années 1960 à la fin des années 1970, ces avancées furent érigées en canon, afin de consolider l'autonomie de la discipline et de faciliter son enseignement. L'analyse démographique (illustrée par les manuels de référence de Pressat), de même que la *formal demography* des pays anglophones, devint le socle de la science démographique.

Le troisième temps, inauguré dans les années 1980, fut marqué par l'abandon des objectifs natalistes, le relâchement des références à la « légitimité » des unions et des naissances, une volonté de renouer avec la statistique explicative des sciences économiques et sociales, la critique historique et sociologique des catégories, le développement d'enquêtes à la fois quantitatives et qualitatives sur des sujets de société, avec une attention accrue aux inégalités, aux discri-

minations et aux violences qui portent atteinte à la cohésion sociale – sans oublier la difficile mais nécessaire réalisation d'enquêtes de terrain dans les pays du Sud.

Cette histoire n'est pas propre à l'Ined. Les démographes qui œuvrent en France au sein de l'IRD ou du monde universitaire ont pris leur part dans ces évolutions. On en retrouve l'équivalent à l'étranger. À qui s'interroge sur l'avenir de la démographie, les données lexicales de Ngram Viewer apportent le recul historique et critique nécessaire. Elle suggère qu'à trop s'enfermer dans le système de publication propre aux « sciences dures », la démographie court le risque de se couper de la société et de la culture. On laissera au lecteur le soin de soumettre aux corpus anglophone et francophone de Ngram Viewer la formule *publish or perish* (en anglais dans le texte) : il se pourrait que ce mot d'ordre soit lui-même périssable s'il devait nous faire oublier qu'il y a une vie en dehors des palmarès bibliométriques. Pour renouer le lien si fragile avec le monde dans lequel nous vivons, qu'il soit national ou transnational, la démographie doit rester attentive aux questions de société et s'ouvrir aux disciplines connexes. C'est à ce prix sans doute qu'elle pourra traverser le siècle.

Références

Acerbi A., 2013, « Normalization biases in Google Ngram », *Wordpress.com*, https://acerbialberto.wordpress.com/2013/04/14/normalisation-biases-in-google-ngram

Aiden E., Michel J.-B, 2013, *Uncharted: Big Data as a lens on human culture*, New York, Riverhead Books / Penguin Books, 288 p.

Benhamou F., 2014, *Le Livre à l'heure du numérique*, Seuil, 221 p.

Bentley R. A, Garnett P., O'Brien M. J., Brock W. A., 2012, « Word diffusion and climate science », *Plus ONE*, doi: 10.1371/journal.pone.0047966

Bijak J., Courgeau D., Silverman E., Franck R., 2014, « Quantifying paradigm change in demography », *Demographic Research*, 30(32), p. 911-924.

Caselli G., Vallin J., Wunsch G. (dir.), 2001-2004, *Démographie. Analyse et synthèse*, Paris, Ined, 5 tomes, 552 p. + 454 p. + 478 p. + 226 p. + 468 p.

Chasteland J.-C., Roussel L. (dir.), 1997, *Les contours de la démographie au seuil du XXIe siècle*, Paris, Ined, 434 p.

Chateauraynaud F., Debaz J., 2010, « Prodiges et vertiges de la lexicométrie », *Socio-informatique et argumentation*, http://socioargu.hypotheses.org/1963

Colombet P., 2008, « A first for France: the City of Lyon and Google partner up to digitize books », *Google Book Search*, http://booksearch.blogspot.fr/2008/07/first-for-france-city-of-lyon-and.html

Delahaye J.-P., Gauvrit N., 2013, *Culturomics* : le numérique et la culture, Paris, Odile Jacob, 224 p.

Guillard A., 2013 [1855], *Éléments de statistique humaine ou Démographie comparée*, Paris, Ined, Classiques de l'économie et de la population, 408 p.

Hai-Jew S., 2014, « Querying Google Books Ngram Viewer's big data set corpuses to complement research », *in* Hai-Jew S. (ed.), *Enhancing Qualitative and Mixed Methods Research with Technology*, IGI Global, p. 514-555.

Héran F., 2013, « La démographie et son vocabulaire au fil des siècles : une exploration numérique », *Population et sociétés*, n° 505, 4 p.

Héran F., 2015, « La science par dérogation, ou comment l'enquête TeO a rempli sa mission », préface à Beauchemin C., Hamel C. et Simon P. (dir.), *Trajectoires et Origines. Enquête sur la diversité des populations en France*, Paris, Ined, Grandes enquêtes, à paraître.

Howard J., 2012, « Google begins to scale back its scanning of books from university libraries », *The Chronicle of Higher Education*, http://chronicle.com/article/Google-Begins-to-Scale-Back/131109

Jeanneney J.-N., 2005, *Quand Google défie l'Europe. Plaidoyer pour un sursaut*, Paris, Mille et une nuits, 113 p.

Lévy M.-L., 1990, *Alfred Sauvy, compagnon du siècle*, Paris, La Manufacture, 220 p.

Lin Y., Michel J.-B., Aiden E. L., Orwant J., Brockman W., Petrov S., 2012, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju (Rep. of Korea), Assoc. for Computational Linguistics, p. 169-174.

Meslé F., Toulemon L., Véron J. (dir.), 2011, *Dictionnaire de démographie et des sciences de la population,* Armand Colin, 528 p.

Michel J.-B., Shen Y. K., Presser A. A., Veres A., Grey M. K., Brockman W., The Google Books Team, 2010, « Quantitative analysis of culture using millions of digitized books », *Science*, 331(6014), p. 176-182 [*à* compléter par le "Supporting online material » : http://www.sciencemag.org/content/suppl/2010/12/16/science.1199644.DC1/Michel.SOM.revision.2.pdf].

Peccate P., 2011, « L'interprétation des graphiques produits par Ngram Viewer », *Déjà vu. Carnet de recherche visuel*, http://culturevisuelle.org/dejavu/469

Pressat R., 1979, *Dictionnaire de démographie*, Paris, Presses universitaires de France, 295 p.

Pressat R., 1980, « Le vocabulaire de la démographie », *Population*, 35(4-5), p. 849-859.

Rosental P.-A., 2003, *L'intelligence démographique. Sciences et politiques des populations en France (1930-1960),* Paris, Odile Jacob, 368 p.

Ruiz É., 2010, « Google Books Ngram Viewer : un nouvel outil pour les historiens ? », *La Boîte à outils des historiens*, http://www.boiteaoutils.info/2010/12/google-labs-books-ngram-viewer-un.html

Taycher L., 2010, « Books of the world, stand up and be counted! All 129,864,880 of you », http://booksearch.blogspot.fr/2010/08/books-of-world-stand-up-and-be-counted.html

François HÉRAN • **LES MOTS DE LA DÉMOGRAPHIE DES ORIGINES À NOS JOURS : UNE EXPLORATION NUMÉRIQUE**

Lancée fin 2010, l'application Ngram Viewer permet de suivre l'évolution du vocabulaire dans les millions d'ouvrages numérisés par Google Books, sur une période qui va du XVIII[e] siècle à nos jours pour le corpus francophone. L'article s'en saisit pour étudier la visibilité très changeante du vocabulaire démographique dans la culture écrite. La première partie examine la sélection et l'organisation des données dans Ngram Viewer. Elle relativise les critiques sur l'absence de contexte des suites de mots (ou *n-grams*). La seconde partie parcourt les rubriques de la démographie et montre que le déclin de la terminologie démographique depuis les années 1990 n'est pas un artefact. Sont touchés les concepts de l'analyse démographique liés au modèle matrimonial et les termes techniques désormais confinés dans les revues scientifiques, non couvertes par Ngram Viewer. Progressent en revanche les questions de société investies par les nouvelles générations de chercheurs : infertilité, mortalité périnatale, orientation sexuelle, nouvelles « transitions » à l'âge adulte, causes de décès, inégalités de santé, rapports de genre, intégration et discriminations, violences, systèmes de valeur. On en conclut que seule une démographie ouverte aux disciplines connexes peut retrouver sa visibilité d'antan et renouer le lien science/société aujourd'hui fragilisé.

François HÉRAN • **THE VOCABULARY OF DEMOGRAPHY, FROM ITS ORIGINS TO THE PRESENT DAY: A DIGITAL EXPLORATION**

Launched at the end of 2010, Ngram Viewer can be used to detect trends in word usage in the millions of documents digitized by Google Books, covering a period from the sixteenth century up to the present day (eighteenth century for the French corpus). This article exploits the capabilities of this new application to examine the changing visibility of demographic vocabulary in written culture. It begins by looking at how data are selected and organized in Ngram Viewer, and shows that the counting of word sequences (or *ngrams*) without reference to context – a shortcoming pointed up by critics – is not an insurmountable problem. It then focuses on the main themes of demography, showing that the decline in demographic terminology since the 1990s is not an artefact. This decline is most visible for the demographic concepts linked to the marriage model, and for technical terms now confined to scientific journals (not covered by Ngram Viewer). An upward trend is observed, on the other hand, for terms linked to the social questions attracting a new generation of researchers, such as infecundity, perinatal mortality, sexual orientation, new transitions to adulthood, causes of death, health inequalities, gender relations, integration and discrimination, violence, systems of values. This suggests that demography must broaden its horizons if it wishes to maintain its former visibility and restore the link between science and society that has become so fragile today.

François HÉRAN • **LAS PALABRAS DE LA DEMOGRAFÍA DESDE LOS ORÍGENES HASTA HOY: UNA EXPLORACIÓN DIGITAL**

Disponible desde finales de 2010, la aplicación Ngram Viewer permite seguir la evolución del vocabulario en los millones de obras numerizadas por Google Books, sobre un periodo que va desde el siglo XVIII hasta hoy para el corpus francés. Este artículo utiliza dicha aplicación para estudiar la visibilidad, muy cambiante, del vocabulario demográfico en la cultura escrita. La primera parte examina la selección y la organización de los datos en Ngram Viewer, y relativiza las críticas sobre la ausencia de contexto de las series de palabras (*n-grams*). La segunda parte recorre las rúbricas de la demografía y muestra que el declive de la terminología demográfica desde los años 1990 no es un artefacto. Son afectados los conceptos del análisis demográfico ligados al modelo matrimonial y los términos técnicos, confinados en las revistas científicas, no cubiertas por Ngram Viewer. Progresan en cambio los temas de sociedad abordados por las nuevas generaciones de investigadores: infertilidad, mortalidad perinatal, orientación sexual, nuevas "transiciones" a la edad adulta, causas de muerte, desigualdades de salud, relaciones de género, integración y discriminaciones, violencias, sistemas de valores. Se concluye que solo una demografía abierta a las disciplinas conexas puede recobrar su visibilidad de antaño y restablecer el lazo entre ciencia y sociedad, actualmente fragilizado.

Mots-clés : histoire de la démographie, vocabulaire de la démographie, données lexicales, outils numériques, rapport science/société, interdisciplinarité.

Keywords: History of demography, vocabulary of demography, lexical data, digital tools, science and society, interdisciplinarity.