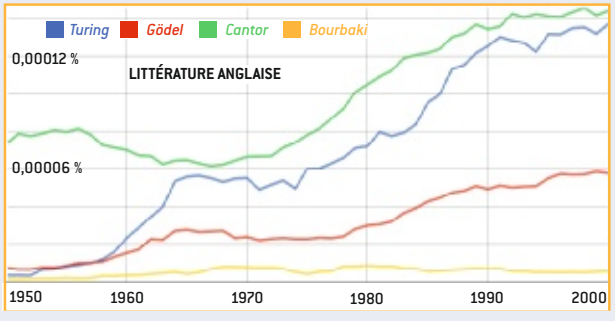
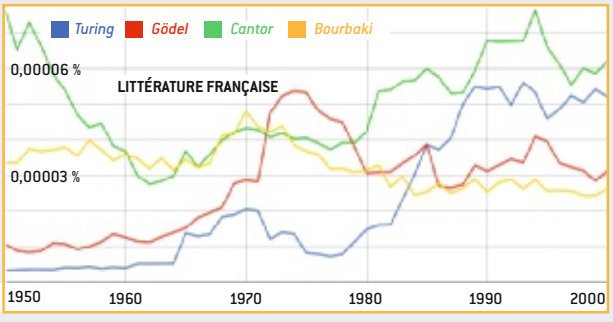
2. La popularité des scientifiques

L'évolution de la fréquence d'utilisation des noms des quatre mathématiciens Turing, Gödel, Cantor et Bourbaki entre 1950 et 2000 dans les livres écrits en anglais et dans les livres écrits en français montre que, dans le monde anglo-saxon, Cantor est de manière constante le préféré, suivi par Turing qui prend la deuxième place dès 1955. Dans les ouvrages en français, la situation est bien plus instable et Turing dépasse Gödel à partir de 1985. Malgré sa renommée internationale, Bourbaki est plus apprécié en français qu'en anglais. On remarquera aussi qu'en français, d'une façon générale, ces mathématiciens sont moins cités qu'en anglais : Turing par exemple atteint 0,00014 pour cent (14 occurrences pour 10 millions de mots) en anglais en 2000, alors qu'en français, il n'arrive à la même date qu'à 0,00006 pour cent (6 occurrences pour 10 millions de mots).

Sur la base des fréquences d'utilisation de leur nom dans le corpus de cinq millions de livres, John Bohannon a défini une unité de mesure de notoriété, le darwin (http://www.sciencemag.org/site/feature/misc/webfeat/gonzoscientist/episode14/index.xhtml) et a classé les principales figures de la science. Voici, mesurés en millidarwins, les 15 premiers scientifiques : Bertrand Russell (1 500), Charles Darwin (1 000), Albert Einstein (878), Lewis Caroll (479), Claude Bernard (429), Oliver Lodge (394), Julian Huxley (350), Karl Pearson (346), Niels Bohr (289), Graham Bell (274), Max Planck (256), Francis Galton (255), Robert Oppenheimer (252), Louis Pasteur ((237), Haïm Weizmann (236). Pour les mathématiciens, le classement est rendu étrange par la présence de physiciens mathématiciens ou d'écrivains : Bertrand Russell (1 500), Lewis Caroll (479), Karl Pearson (346), A. N. Whitehead (229), James Jeans (182), Norbert Wiener (163), John von Neumann (137), Henri Poincaré (108), Ronald Ross (98), James Clerk Maxwell (92), Stephen Hawkins (88), Simon Newcomb (82).





formeraient une ligne de cinq millions de kilomètres, plus de dix fois la distance de la Terre à la Lune !

Ce corpus contient 361 milliards de mots anglais, 45 milliards de mots français, 45 milliards de mots espagnols, 37 milliards de mots allemands, et quelques milliards de mots d'autres langues. Les plus vieux ouvrages pris en compte datent du XVI^{e} siècle. À partir de 1800, le corpus contient plus de 60 millions de mots par an, et à partir de 1900, il compte plus d'un milliard de mots par an.

Le projet de numérisation de *Google* inquiète les auteurs et encore plus les éditeurs qui craignent d'être dépossédés de leurs fonds. Toutefois, nul ne nie qu'un tel ensemble de livres présente un intérêt historique, linguistique, littéraire, social, scientifique et même philosophique..., pourvu qu'il soit consultable.

Jean-Baptiste Michel, de l'Université Harvard, et l'équipe réunie afin de constituer ce corpus proposent un nom, la *culturomique* (en anglais *culturomics*) pour la nouvelle discipline née de l'étude de ce corpus.

La mise à disposition directe du corpus, accompagné des informations spécifiques à chacun des livres qu'il inclut pour que chacun puisse y organiser les études statistiques qui l'intéressent, n'est hélas pas envisageable, du fait des règles et lois qui protègent la propriété intellectuelle. De plus, la taille informatique du corpus, environ quatre téraoctets (4×10^{12} octets), excède les capacités de stockage d'un ordinateur courant et encore plus les capacités des traitements qu'un utilisateur pourrait souhaiter mener.

L'accès n'est que partiel

Aussi, pour permettre l'exploitation des cinq millions de livres, les chercheurs ont opéré par avance une multitude de calculs : à défaut de mettre à la libre disposition tout le corpus, ils autorisent l'exploration de la fréquence des mots et groupes de mots, par langue et par année, ce qui est déjà d'un grand intérêt.

Pour chaque mot ou groupe de mots de moins de cinq mots apparaissant un minimum de fois, et pour chaque année entre 1800 et 2008, vous pouvez savoir quelle a été sa fréquence exacte d'utilisation dans le corpus. Vous pouvez même vous intéresser à plusieurs séquences de mots simultanément dont l'évolution de la fréquence d'utilisation est rendue comparable par l'affichage d'un graphique. Les courbes obtenues (des *n-grammes*) sont automatiquement engendrées par les programmes du site Internet du projet *(voir la bibliographie)*.

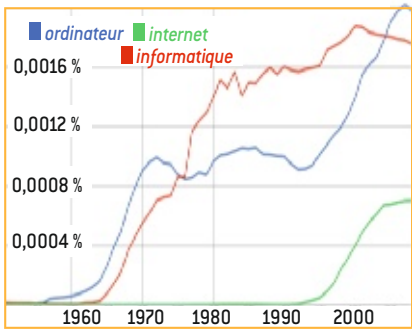
Tout internaute peut mener ses propres recherches. Essayez : dès que vous aurez commencé à produire quelques courbes, vous serez pris au jeu et de nouvelles questions vous viendront à l'esprit.

L'outil mis à la disposition de tous est donc un jouet linguistique que la suite de la rubrique va exploiter, mais c'est aussi un outil sérieux et plusieurs publications scientifiques ont déjà été réalisées en exploitant les *n-grammes* produits.

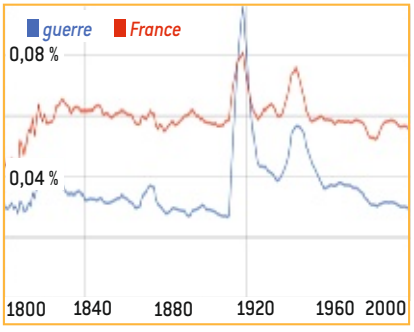
Le corpus ne prend pas en compte journaux, revues, tracs publicitaires et autres imprimés comme les notices, formulaires administratifs, affiches, etc. La date de publication retenue pour les livres du corpus est celle que donne l'éditeur et donc un livre peut apparaître plusieurs fois avec des dates différentes s'il a été réédité. Malgré le soin mis à l'élaboration du corpus, les erreurs de numérisation ne sont pas rares et introduisent des artefacts. Ainsi, le mot *internet* est utilisé avec une très faible fréquence même au XIX^{e} siècle. La raison est que l'abréviation d'usage courant *internat* pour international entraîne une erreur provenant de la confusion entre *a* et *e*. Le piège des homonymes doit aussi être pris en compte : un certain Chirac écrivit un code socialiste en 1886 !

Prévisible... et vrai

Bien des courbes obtenues ne font que confirmer des évidences que précise l'étude du mégatexte. Si, par exemple, on s'intéresse en français à l'usage des mots *ordinateur, informatique, internet* le long de la période 1950-2008, on obtient un graphe qui montre sans surprise que l'usage du mot *ordinateur* prend une importance mesurable à partir de 1960. À cette date, sa fréquence atteint 0,0001 pour cent, ce qui signifie que parmi tous les mots du corpus français pour l'année 1960, un mot sur un million environ est le mot *ordinateur*. Légèrement plus tard, vers 1964, le mot *informatique*, introduit en 1962 par Philippe Dreyfus, atteint un usage mesurable, puis dépasse le mot *ordinateur* en 1975. Le mot *internet* n'apparaît de manière significative que vers 1995 pour atteindre 0,0004 pour cent en 2000.



L'étude de la fréquence d'usage en français des mots *guerre* et *France* donne elle aussi un résultat sans surprise. Les deux mots ont des pics d'utilisation simultanés pendant les périodes 1914-1920 et 1940-1948. Le graphique nous apprend cependant que *guerre* est presque toujours un mot moins utilisé que *France*. Seule la période de la Première Guerre mondiale montre une inversion : malgré le nationalisme exacerbé des deux belligérants, le mot *guerre* passe devant le mot *France*, sans doute une conséquence des multiples difficultés matérielles de la guerre qui préoccupent les esprits.

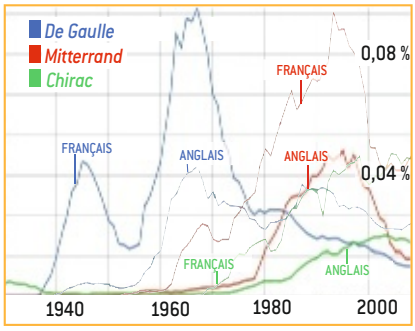


Le corpus des 500 milliards de mots est un outil idéal pour mesurer la notoriété des personnalités intellectuelles, politiques ou autres, y compris pour la suivre année après année sur les deux derniers siècles.

Mesures de notoriété

Il est par exemple amusant de comparer l'évolution de la notoriété de *De Gaulle, Mitterrand* et *Chirac* entre 1930 et 2008 dans la partie anglaise du corpus et dans la partie française. On remarque que *De Gaulle* dans la partie anglaise n'est dépassé par *Mitterrand* qu'à partir de 1984, alors que dans la partie française, le croisement se produit dès 1975. Dans la partie anglaise, *De Gaulle* domine globalement. Dans la partie française, c'est *Mitterrand*. Quant à *Chirac*, il a de la peine, même élu président, à égaler les deux autres. Dans la partie française, alors qu'il est président de la République depuis 1995, il ne dépasse *De Gaulle* qu'en 1998, et *Mitterrand* qu'en 2004 !

L'encadré 2 examine la notoriété des scientifiques et montre que, selon les



pays, les gagnants ne sont pas les mêmes. Avec le corpus géant des cinq millions de livres, John Bohannon, dans un article de la revue *Science* de janvier 2011, a concocté un panthéon objectif des scientifiques. Contrairement aux études fondées sur les indicateurs de citations qui ne prennent en compte que les publications « strictement scientifiques » et publiées dans des revues spécialisées, ce panthéon mesure l'impact sur la société des célébrités de la science. Les cinq hommes de science qui arrivent en tête sont : Bertrand Russell, Charles Darwin, Albert Einstein, Lewis Caroll, Claude Bernard. Bertrand Russell, le premier, est souvent mentionné à propos de ses prises de position politique sans qu'elles aient nécessairement un lien avec la science. On peut donc considérer que le gagnant est plutôt Charles Darwin. Son nom apparaît 148 429 fois dans les livres de la base publiés entre 1939 et 2000, livres qui sont 69 048 à le mentionner au moins une fois.

L'unité de mesure introduite pour réaliser les classements est le millidarwin, le millième du darwin. Un darwin est la fréquence annuelle moyenne de mentions que Charles Darwin obtient entre l'année où il a atteint 30 ans et l'année 2000. Un scientifique qui atteint un millième de cette fréquence moyenne entre l'année de ses 30 ans et l'année 2000 se voit donc attribué une notoriété de un millidarwin. Toutes sortes de difficultés méthodologiques se présentent pour calculer cette note de célébrité. Par exemple, le mathématicien collectif Bourbaki n'est pas classé, car il est le plus souvent mentionné sans son prénom et, que sans son prénom, il est alors confondu avec le général Bourbaki, dont la notoriété a été forte durant la seconde