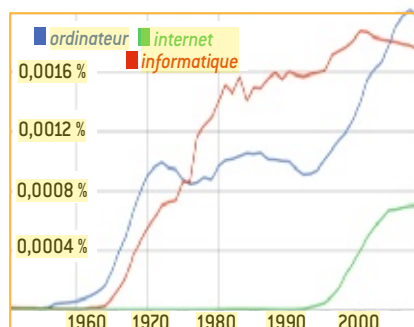


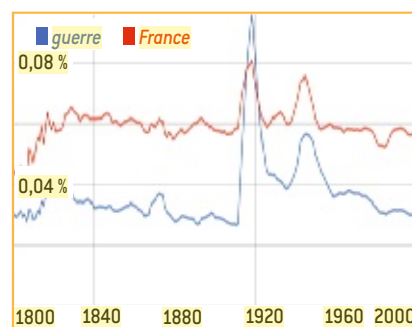
Le corpus ne prend pas en compte journaux, revues, tracts publicitaires et autres imprimés comme les notices, formulaires administratifs, affiches, etc. La date de publication retenue pour les livres du corpus est celle que donne l'éditeur et donc un livre peut apparaître plusieurs fois avec des dates 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<sup>e</sup> 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 pré-occupent les esprits.

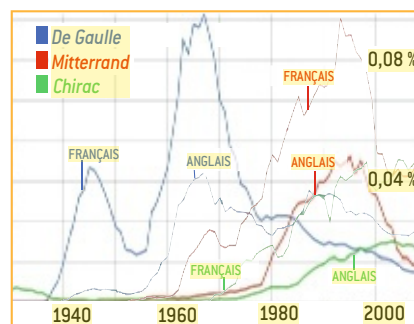


Le corpus des 500 milliards de mots est un outil idéal pour mesurer la notoriété des personnalités intellectuelles, politiques ou autres, y compris pour la suivre année après année sur les deux derniers siècles.

## 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 Carroll, 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