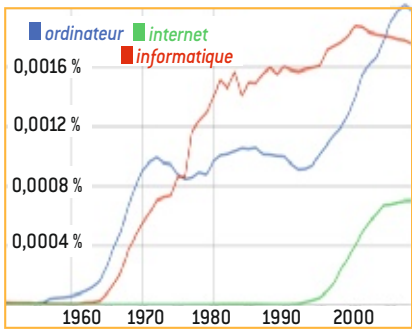
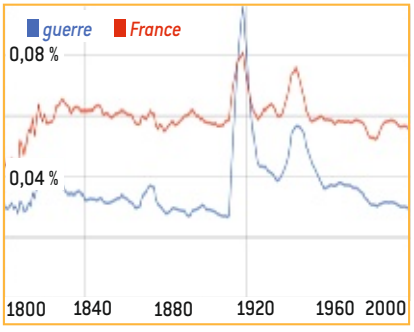
Le corpus ne prend pas en compte journaux, revues, tracs publicitaires et autres imprimés comme les notices, formulaires administratifs, affiches, etc. La date de publication retenue pour les livres du corpus est celle que donne l'éditeur et donc un livre peut apparaître plusieurs fois avec des dates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

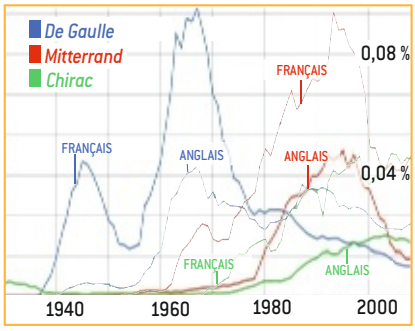


Le corpus des 500 milliards de mots est un outil idéal pour mesurer la notoriété des personnalités intellectuelles, politiques ou autres, y compris pour la suivre année après année sur les deux derniers siècles.

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

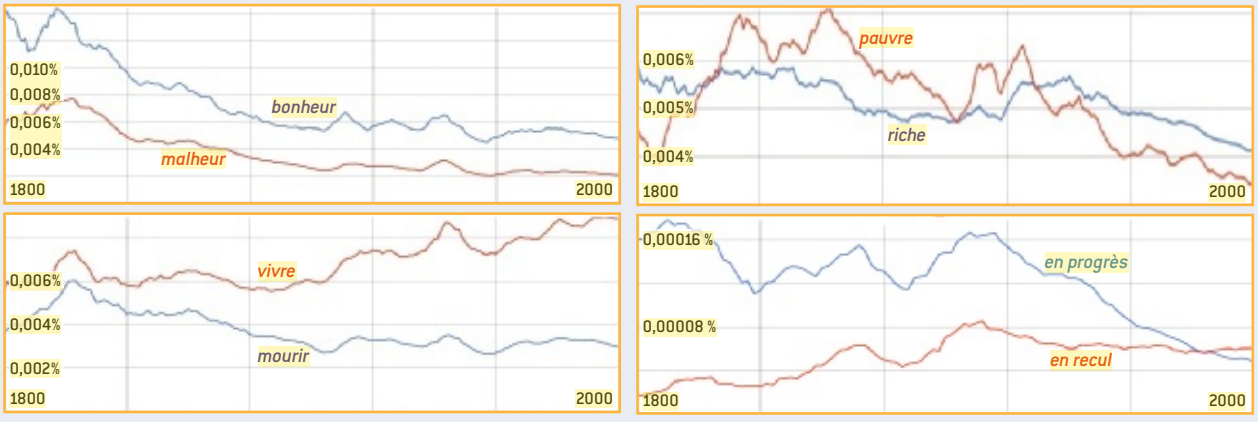
3. Le biais de positivité

La préférence pour le terme positif d'un couple de mots opposés est remarquable et étonnante par sa stabilité et sa régularité. Les courbes montrent ainsi ce biais pour les couples *positif/négatif, vivre/mourir, content/mécontent, beau/laid, grand/petit, rire/pleurer, mieux/pire, plus/moins, bien/mal, bonheur/malheur, succès/échec, réussir/échouer* et *bon/mauvais*.

Les fréquences d'utilisation des mots *bonheur* et *malheur* diminuent légèrement au cours des ans, mais *bonheur* domine de manière constante. Le mot *vivre* bat de plus en plus nettement *mourir*, trace sans doute du fait souvent noté que nos sociétés acceptent de moins en moins l'idée de la mort et donc l'évoquent aussi de moins en moins.

Le biais de positivité semble dans bien des cas de plus en plus fort : pendant le XIXe siècle, le mot *pauvre* domine sur le mot *riche*, mais depuis 1920, il a été doublé. À chacun de formuler ses explications en évoquant le changement des critères de la réussite sociale, le recul du sentiment de compassion promu par la religion, ou d'autres facteurs encore.

Conformément au biais de positivité attendu, l'expression *en progrès* domine largement sur *en recul*. Cependant, sa fréquence perd proportionnellement entre 1914 et 1920 (un effet dépressif de la guerre ?) et surtout perd nettement depuis 1960, au point que depuis 1990, *en recul* est passé en tête. Notre époque aurait-elle perdu confiance en elle ?



moitié du XIXe siècle : il faudrait trier à la main chaque mention, tâche impossible étant donné la taille du corpus.

Une des conclusions de ce classement de notoriété est que les mathématiques, seules, ne sont pas un moyen recommandé pour gagner une bonne place dans le panthéon scientifique. Parmi les mathématiciens, on notera *John von Neumann* (mathématicien, mais aussi physicien) qui obtient 137 millidarwins, *Henri Poincaré* 108, *Kurt Gödel* 40, *Felix Klein* 39, *Georg Cantor* 29, *Benoît Mandelbrot* 19, *Émile Borel* 12, *René Thom* 10. Le plus prolifique de tous les mathématiciens, *Paul Erdös*, ne s'étant occupé que de mathématiques au plus haut niveau sans penser à un public plus large, obtient un score de quatre millidarwins seulement.

Ce contraste entre réussite scientifique et notoriété mesurée par le grand corpus est très net pour les lauréats du prix Nobel, dont certains ne dépassent pas un millidarwin. Dans le haut du tableau, ils ne sont même pas majoritaires. Pour gagner des millidarwins, J. Bohannon conseille de participer à de nombreuses controverses et d'écrire des livres visant des audiences aussi larges que possible.

Préférence pour les termes positifs

Le *biais de positivité* identifié en psychologie est mis en évidence avec une étonnante netteté *(voir l'encadré ci-dessus)*. Ce biais est la tendance qu'ont les sujets humains à préférer le terme positif entre les deux termes opposés d'un couple d'expressions. On le met classiquement en évidence en demandant à un ensemble de sujets de choisir très rapidement au hasard *oui/non, amour/haine, mort/vie*, etc. Non seulement une préférence claire pour le terme positif se manifeste systématiquement, mais, de plus, Nicolas Gauvrit a établi que, chez les personnes déprimées, ce biais était atténué et constituait donc une mesure de la gravité d'une dépression.

Parmi les exceptions à cette règle figure le couple *oui/non*, mais cette exception ne signifie sans doute rien puisque l'usage du mot *non* dans l'expression de la négation (« c'est un non-dit », « ceci est non significatif », « c'est un non-sens », etc.) ne vient pas en symétrie du mot *oui* pour les cas positifs (« c'est dit », « ceci est significatif », « cela est sensé »).

L'exception du couple *guerre/paix* (*guerre* domine *paix*) est notable. Sans surprise, l'utilisation du mot *guerre* triple dans la période 1914-1920. Le mot *paix,* lui, augmente seulement un peu durant cette période tout en restant en second derrière *guerre*. Là encore l'explication est sans doute que la paix est l'état « normal » qu'on ne nomme pas quand on en bénéficie, contrairement à la guerre qu'on subit et dont on parle.

Plus amusant encore, l'usage des séries ordonnées de mots montre de curieuses régularités souvent insoupçonnées et d'une remarquable stabilité dans le temps.

On trouve ainsi de nouvelles formes de la loi de Benford *(voir l'encadré 1)*. Le mot *un* est non seulement un chiffre et un adjectif numéral, c'est aussi un article. Pas étonnant donc qu'il domine sans la moindre équivoque tous les autres chiffres *deux, trois,..., neuf*. En revanche, on reste émerveillé de la régularité des courbes pour *deux, trois,... neuf* et de leurs positions relatives.

Une autre curiosité numérique attend une explication. Si on compare les taux d'usage de *100 francs, 200 francs, 300 francs, 400 francs, 500 francs,* on trouve sans surprise que viennent en tête *100 francs* et *500 francs* (ce sont des chiffres ronds) suivis dans l'ordre de *200 francs, 300 francs, 400 francs*.

Cela est conforme à ce que la loi de Benford laisse attendre, en la corrigeant du fait que les chiffres ronds sont poussés en avant par les usages commerciaux qui dérangent un peu l'ordre naturel. La surprise vient de la croissance générale de tous les taux de citations (pour les cinq expressions) jusqu'en 1910, suivie d'une décroissance générale, elle aussi régulière. S'est-on plus particulièrement intéressé à l'argent vers 1910 ? Pourquoi donc ? Ou alors, l'explication est-elle liée à un changement dans les usages lexicaux, éditoriaux ou typographiques ? Étrange.

Une démarche qui a ses limites

Aussi impressionnantes que soient les quantités de données prises en compte pour produire les courbes, cela ne suffit pas à en faire un domaine scientifique nouveau. D'une part, la base reste quand même particulière (absence des journaux et revues). D'autre part, l'impossibilité d'accéder au corpus lui-même interdit souvent d'étudier les raisons de ce qu'on découvre. Par exemple, il faudrait disposer de statistiques sur la composition du corpus pour comprendre d'où vient l'augmentation de l'usage des chiffres isolés, où de celles des sommes comme *100 francs* vers 1910. La vue donnée par les *n-grammes* est certes intéressante, mais mille questions se posent dont les réponses ne peuvent provenir que de tests menés directement et spécifiquement sur les cinq millions de livres numérisés, ou sur des sous-ensembles soigneusement constitués, dont il faudrait pouvoir disposer.

Notons encore que d'autres bases numériques seraient intéressantes à exploiter pour le linguiste, le sociologue ou l'historien, et que comme celle du grand corpus de livres, elles sont aujourd'hui interdites d'accès, et seraient encore plus difficiles à manipuler du fait de leur volume.

C'est le cas de la colossale base d'informations, qui enchanterait un sociologue, constituée par les milliards de fichiers des *datacenters* de *Facebook* dont on sait qu'elle concerne 500 millions d'inscrits et contient 30 milliards de photos. Les ingénieurs de *Facebook* indiquent recevoir 12 téraoctets d'informations nouvelles chaque jour, c'est-à-dire trois fois plus que le corpus géant dont nous avons parlé. Cela donne un total plusieurs centaines de fois plus gros que le corpus géant de livres, déjà difficile à manipuler. Le nombre de 30 000 téraoctets (3×10^{15} octets) a été cité pour le total des données de *Facebook*. Les données collectées par les moteurs de recherche *Google, Bing,* etc., ou de *Gmail,* seraient aussi intéressantes à explorer.

Le début de cet article mentionnait la difficulté d'évaluer certaines variables du monde social, économique ou politique que les sondages ne permettent pas de connaître. Les masses d'informations que certaines entreprises sur Internet constituent aujourd'hui sont de nouveaux objets du monde. On aurait pu croire qu'étant stockés sous format numérique, on les exploiterait facilement, contrairement aux données du monde réel non numérisées et difficiles à parcourir, car éparpillées. Ce n'est pas vrai. Outre les impossibilités liées aux droits de la propriété intellectuelle et aux règles de la protection de la vie privée qui limitent (heureusement !) l'accès des chercheurs aux bases de données des firmes sur Internet, les masses constituées sont aujourd'hui si volumineuses que, techniquement, elles ne sont exploitables que très partiellement et sans doute uniquement... par sondage.

L'AUTEUR

Jean-Paul DELAHAYE est professeur à l'Université de Lille et chercheur au Laboratoire d'informatique fondamentale de Lille (LIFL).

BIBLIOGRAPHIE

J. Evans et J. Foster, **Metaknowledge**, *Science*, vol. 331, pp. 721-725, 2011.

J.-B. Michel *et al.*, **Quantitative analysis of culture using millions of digitized books**, *Science*, vol. 331, pp. 176-182, 2010 (http://mfi.uchicago.edu/publications/papers/Science_Culturomics.pdf) ; **Supporting online material :** http://dericbownds.net/uploaded_images/Michel.SOM.pdf

J. Bohannon, **Google opens books to new cultural studies**, *Science*, vol. 330, p. 1600, 2010.

J. Bohannon, **The science hall of fame**, *Science*, vol. 331, p. 143, 2011 : http://www.sciencemag.org/content/331/6014/143.3.full

N. Gauvrit et J.-P. Delahaye, **Scatter and regularity implies Benford's law**, dans H. Zenil (ed.), *Randomness through complexity*, World Scientific, 2011.

N. Gauvrit, **Dépressivité et biais de positivité**, *Cahiers Romans de Sciences Cognitives*, vol.3, n° 3, pp. 63-71, 2009.

SUR LE WEB

Site Internet de la culturomique : http://www.culturomics.org/

Page permettant d'engendrer des courbes : http://ngrams.googlelabs.com/