Avec l'aimable autorisation du Y-12 National Security Complex

7. LE SITE SECRET CLINTON ENGINEER WORKS, dans l'Est du Tennessee, où fut construit le complexe électromagnétique Y-12. Fin 1945, plus de 1 100 calutrons étaient en service sur le site.

Avec l'aimable autorisation de la Milwaukee County Historical Society

8. À LA *ALLIS-CHALMERS MANUFACTURING COMPANY*, dans le Wisconsin, les bobines d'argent étaient déroulées, soudées, puis enroulées autour des corps de bobine en acier des boîtiers d'aimants.

Après la guerre, de nombreux calutrons sont rééquipés avec des bobinages en cuivre, mais pas les calutrons de l'usine pilote. Ceux-là fonctionnent jusqu'en 1974 (avec 60 tonnes d'argent dans les bobinages de leurs aimants jusqu'en 1970) pour séparer des isotopes autres que ceux de l'uranium, dont certains sont utilisés dans le réacteur à graphite du tout proche Laboratoire national d'Oak Ridge. L'installation crée des marqueurs radioactifs pour certains examens médicaux. D'autres calutrons développés pour le projet Manhattan séparent des isotopes stables jusqu'en 1998 ; ensuite, l'arrivée de sources meilleur marché d'isotopes impose la fin de l'exploitation. La fermeture de ces installations est une des raisons de l'actuelle pénurie d'isotopes médicaux aux États-Unis.

Les deux visages du projet Manhattan

Aujourd'hui, les installations Y-12 d'Oak Ridge sont toujours en service en tant que Complexe de la sécurité nationale du ministère de l'Énergie, sous contrat avec la *Babcock & Wilcox Company*. Elles ont été récemment reconverties en site de récupération et de stockage de matériaux nucléaires. Un rapport datant de 2010 de la *Nuclear Posture Review*, qui définit la stratégie nucléaire américaine pour les prochaines années, prône également la construction d'une nouvelle installation de traitement de l'uranium sur le site Y-12, qui pourrait être opérationnelle en 2021.

Souhaitons qu'Oak Ridge, de même que les deux autres sites du projet Manhattan – Los Alamos, au Nouveau-Mexique, et Hanford, dans l'État de Washington – resteront accessibles pour aider les générations futures à explorer et comprendre le projet. Los Alamos centralisait les recherches scientifiques du projet, et Hanford mettait en œuvre une autre stratégie pour obtenir un matériau fissile : la synthèse, dans des réacteurs géants, de plutonium à partir d'uranium.

Les historiens débattront sans fin pour savoir si le projet Manhattan a vraiment contribué à abréger la Seconde Guerre mondiale ou permis d'éviter une invasion sanglante du Japon. Le fait d'accélérer la fin de la guerre valait-il le coût en victimes civiles et les dégâts en termes d'image et d'autorité morale de l'Amérique ? Quoi qu'il en soit, le projet Manhattan a été la plus grande entreprise mondiale de « Big Science ». Il a influencé des événements à court et à long terme comme aucune autre initiative scientifique de l'histoire. Le projet et son programme *Silver* sont des modèles de ce que la science et l'ingénierie, conduites par des professionnels compétents, peuvent accomplir quand le besoin se fait pressant.

L'AUTEUR

Cameron REED est professeur de physique au *Alma College* à Alma dans le Michigan, aux États-Unis.

Article publié avec l'aimable autorisation de *American Scientist*.

✔ BIBLIOGRAPHIE

C. Reed, **The Physics of the Manhattan Project,** Springer, 2010.

C. Reed, **Bullion to B-Fields : the Silver Program of the Manhattan Project**, *Michigan Academician*, vol. 39, n° 3, pp. 205-212, 2009.

Le site Web du **Complexe de sécurité nationale Y-12** propose de nombreuses archives : www.y12.doe.gov

LOGIQUE & CALCUL

La culturomique

L'étonnant corpus de textes créé récemment par une équipe internationale de chercheurs dévoile des phénomènes linguistiques insoupçonnés.

Jean-Paul DELAHAYE

De nombreuses informations, concernant le monde physique, culturel ou social sont trop coûteuses à collecter : il n'est pas envisageable par exemple de savoir combien de personnes connaissent le nom de Kurt Gödel et le sujet de ses recherches. Alors que pour évaluer le nombre de personnes qui connaissent une personnalité très en vue, nous pouvons opérer par sondage, pour une personnalité scientifique à la notoriété moindre, il faudrait, pour un résultat fiable, un échantillon de plusieurs dizaines de milliers de personnes… ce qui revient trop cher. On a pallié aujourd'hui ce problème en utilisant un énorme corpus de données comportant cinq millions de livres !

Sonder ou tout prendre en compte ?

Les sondages donnent accès à des informations globales sur de grands ensembles impossibles à explorer de manière exhaustive : groupes humains, populations animales, objets sortant d'une chaîne de production, etc. Malheureusement, les sondages ne voient pas les faibles fréquences.

Pour des statistiques fines, il nous faut procéder autrement et nous savons que les élections sont le seul moyen pour mesurer l'importance des petits partis. Sans les approches exhaustives ou presque exhaustives, de nombreux faits restent invisibles ou impossibles à mesurer. L'informatique élargit le champ des variables évaluables : elle offre des capacités de stockage massif d'informations, elle réalise des collectes de faits, de chiffres et d'images à grande échelle et les automatise. Elle donne aussi des moyens de traitement numériques ultrarapides pour l'exploration des gigantesques bases de données. Plusieurs domaines de recherche sont nés ou ont été renouvelés. Ainsi, la fouille de données *(data mining)* conçoit des méthodes pour extraire l'information pertinente des masses inouïes de données enregistrées sur les supports numériques et la technologie des moteurs de recherche sur Internet se fonde sur des bases d'informations stockées sur les disques durs de centaines de milliers d'ordinateurs des *datacenters*, information mise à jour en permanence, puis manipulée instantanément pour répondre aux milliards de requêtes des utilisateurs que nous sommes tous.

De nombreuses disciplines scientifiques (toutes ?) sont concernées. Citons la météorologie où les réseaux de collectes d'information et les outils de calcul sont perfectionnés sans relâche ; la géographie et la démographie, qui disposent de tableaux chiffrés de plus en plus variés et nombreux ; l'économie, elle aussi soumise à une multiplication des informations mises à sa disposition ; la biologie et la génétique, avec ces centres de données de séquences qui sont devenus un outil quotidien des chercheurs ; les mathématiques avec, par exemple, la base de suites numériques de Neil Sloane.

La linguistique utilise depuis longtemps les outils informatiques. La constitution de corpus de textes donne une vision objective des langues parlées ou écrites. La réunion de tous les textes d'un auteur permet l'étude statistique du vocabulaire d'un écrivain et de ses habitudes stylistiques.

Les conclusions sont parfois étonnantes : Dominique Labbé, après analyse des œuvres de Molière et de Corneille, prétend que les 16 œuvres principales de Molière ont été écrites par Corneille, thèse qui, sans faire l'unanimité, s'appuie sur des arguments sérieux.

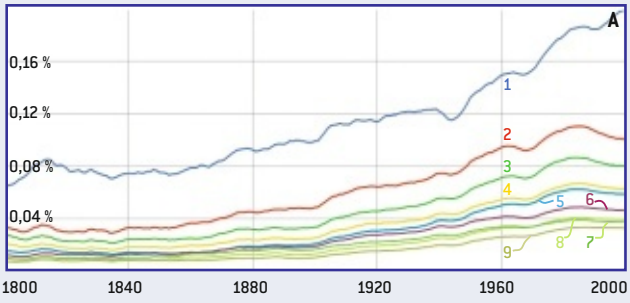
Cinq millions de livres

Un pas vient d'être franchi dans la taille des corpus linguistiques auquel la science informatique donne accès : une base de 5 195 769 livres a été constituée.

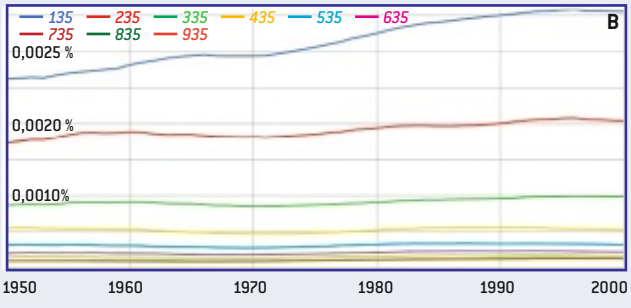
Ces livres, qui proviennent des bibliothèques de 40 universités, ont été numérisés par la Société *Google* ; ils représentent quatre pour cent de la totalité des livres jamais publiés. Les ouvrages ont été sélectionnés parmi 12 millions de livres numérisés en prenant en compte la qualité de la numérisation et des informations générales, les « métadonnées », dont on dispose sur eux. Le corpus ainsi constitué comporte 500 milliards de mots. C'est énorme : en lisant un livre chaque jour pendant 50 ans, vous n'en aurez parcouru que 18 000, soit 0,36 pour cent du corpus constitué. Les textes mis bout à bout, à raison de dix symboles par centimètre,

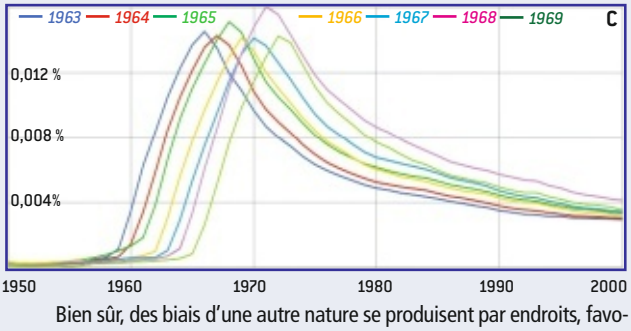
1. Des raffinements insoupçonnés de la loi de Benford

La loi de Benford est étonnante : si nous prenons des nombres au hasard dans une série en comportant beaucoup, comme les populations des villes, les longueurs des fleuves, les surfaces des pays, les cours de la Bourse, etc., les nombres commençant par le chiffre 1 sont plus nombreux que les nombres commençant par le chiffre 2, eux-mêmes plus nombreux que ceux commençant par 3, etc. Plusieurs tentatives d'explication du phénomène ont été proposées, pas toujours convaincantes.

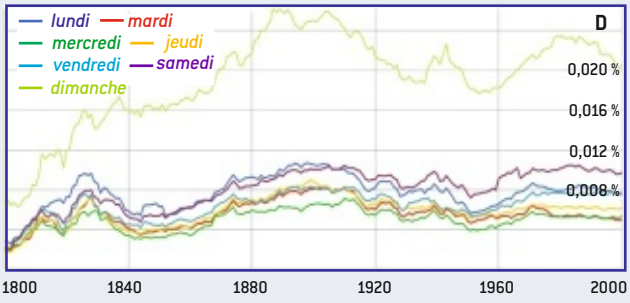


Une façon de tester cette loi est de comparer les fréquences d'usages des signes isolés 1, 2, 3, 4, 5, 6, 7, 8, 9, ce que permet le système des *n-grammes* (groupes de *n* signes) associé à la base de cinq millions de livres. La loi est bien vérifiée. La pente montante de toutes les courbes provient sans doute de l'augmentation de la proportion des livres techniques dans l'ensemble des livres publiés *(A)*. Plus frappant peut-être, la loi de Benford implique que 135 doit être plus souvent utilisé que 235, lui-même plus utilisé que 335, etc. C'est vrai *(B)*.





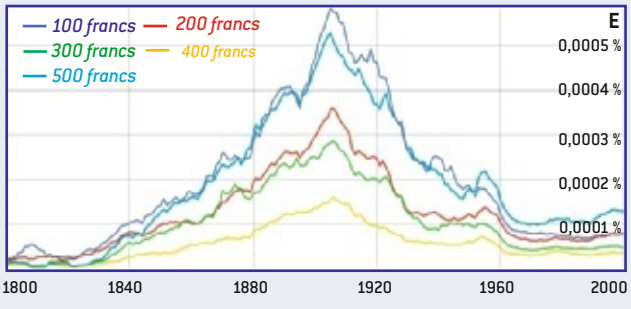
Bien sûr, des biais d'une autre nature se produisent par endroits, favorisant un nombre particulier et perturbant la loi de Benford générale au moins pendant quelques années. Les nombres ronds sont favorisés : 15 est plus utilisé que 14 ou 16, et 150 est aussi plus utilisé que 140 ou 160. Sans surprise encore, quand nous nous approchons de 1965, le nombre 1965 est de



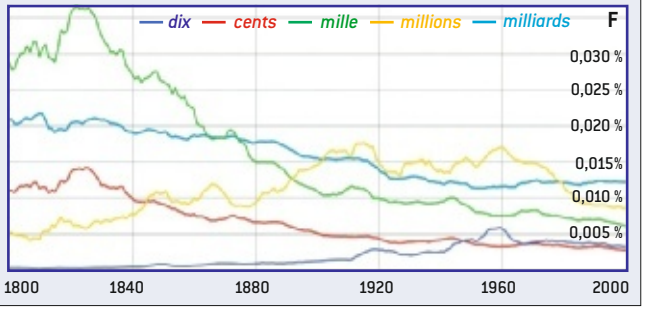
plus en plus cité. Du coup, c'est trois ou quatre ans après 1965 que 1965 atteint son pic d'utilisation. Ce décalage temporel n'est pas l'explication de tout ce qu'on observe et, par exemple, l'année 1968 est celle qui va le plus haut parmi les années comprises entre 1963 et 1969 dans les livres en français *(C)*, alors qu'un examen similaire montrerait que ce n'est pas le cas en anglais : les raffinements de la loi de Benford semblent sans limites.

Dans un ordre d'idées proche, l'étude de la fréquence d'utilisation des noms des jours de la semaine en français présente une structure imprévue, mais pas inexplicable *(D)*. *Dimanche* domine ; jusqu'en 1910 environ *lundi* est en deuxième position, mais il se fait dépasser par *samedi* aujourd'hui bien installé en second. Viennent alors *vendredi* et *jeudi*, puis *mardi* et *mercredi* qui se disputent la dernière place.

Les courbes pour *100 francs*, *200 francs*, *300 francs*, *400 francs* et *500 francs* présentent toutes un étonnant maximum vers 1910 *(E)*. Pourquoi ?



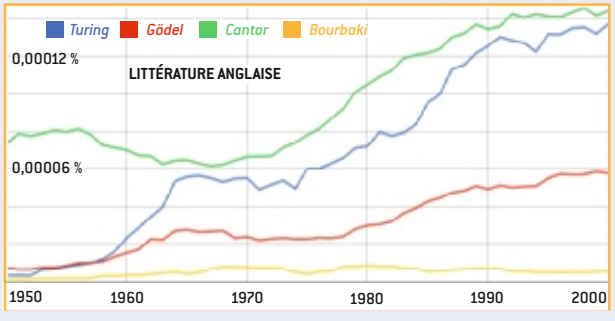
Autre mystère : est-ce parce que nous devenons plus riches, est-ce un effet de l'inflation, ou alors faut-il chercher d'autres causes encore pour expliquer que (dans la course entre *dix*, *cents*, *mille*, *millions*, *milliards*) *milliard* vient en 1980 de passer en tête devant *million* qui lui-même avait battu *mille* vers 1900 *(F)* ?

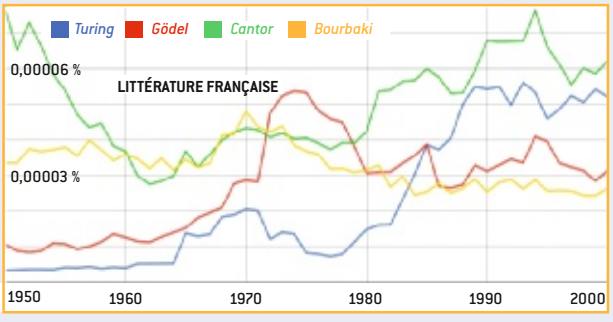


2. La popularité des scientifiques

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)*.

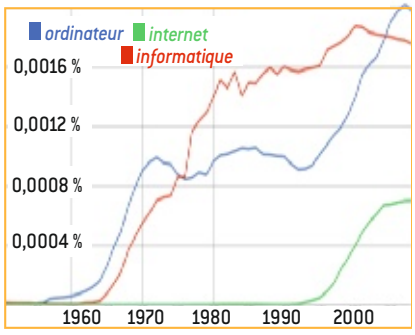
Tout internaute peut mener ses propres recherches. Essayez : dès que vous aurez commencé à produire quelques courbes, vous serez pris au jeu et de nouvelles questions vous viendront à l'esprit.

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.

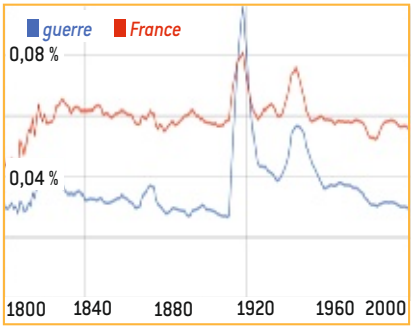
Le corpus ne prend pas en compte journaux, revues, tracs publicitaires et autres imprimés comme les notices, formulaires administratifs, affiches, etc. La date de publication retenue pour les livres du corpus est celle que donne l'éditeur et donc un livre peut apparaître plusieurs fois avec des dates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

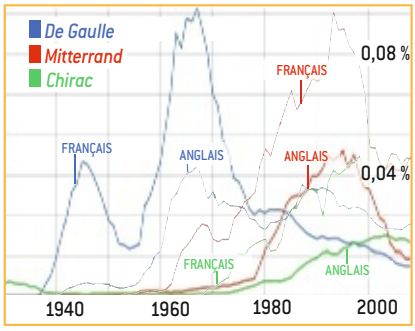


Le corpus des 500 milliards de mots est un outil idéal pour mesurer la notoriété des personnalités intellectuelles, politiques ou autres, y compris pour la suivre année après année sur les deux derniers siècles.

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

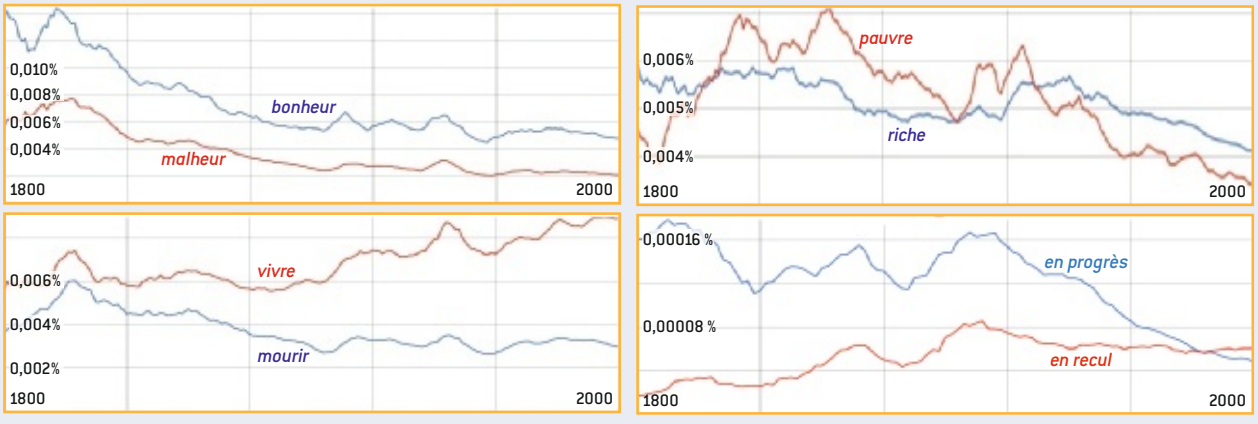
3. Le biais de positivité

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/

ART & SCIENCE

Le cerveau caché de Michel-Ange

Dans deux panneaux du plafond de la chapelle Sixtine, la Création d'Adam *et la* Séparation de la lumière et des ténèbres, *Michel-Ange aurait dissimulé des planches anatomiques du système nerveux.*

Loïc MANGIN

En 1505, Michel-Ange (1475-1564) quitte Florence et rejoint Rome à la demande du nouveau pape Jules II (1443-1513). Ce dernier confie au sculpteur le soin de lui concevoir un tombeau monumental dans la basilique Saint-Pierre : le mausolée ne sera jamais achevé… Après un exil de quelques mois à Bologne, Michel-Ange s'attelle à un autre projet, le plafond de la chapelle Sixtine qu'il peindra entre 1508 et 1512. Le projet fut proposé à l'artiste à l'instigation de l'architecte Bramante qui était persuadé de l'échec de son rival. Mal lui en prit, le plafond est un des plus célèbres chefs-d'œuvre de la Renaissance italienne !

Le plafond, de 40 mètres de longueur et 14 de largeur, est situé à 20 mètres de hauteur. Il est constitué de différentes surfaces en demi-lunes, en triangles, en rectangles… où Michel-Ange devait, selon son contrat, représenter les 12 apôtres. L'artiste vit plus grand et proposa d'y peindre des épisodes de la Genèse *(voir la figure a)*, de la *Séparation de la lumière et des ténèbres* jusqu'à l'*Ivresse de Noé* en passant par la *Création d'Adam*. Les murs latéraux étaient déjà ornés de fresques où le Pérugin, Botticelli et d'autres ont raconté les vies du Christ et de Moïse.

Intéressons-nous à la *Séparation*. Les historiens ont longtemps été intrigués par les irrégularités du cou de Dieu dans ce panneau *(voir les figures b et c)* et par le traitement de la lumière dans cette région du corps. Le personnage, Dieu, est éclairé en diagonale, du bas à gauche vers le haut à droite, mais le cou semble illuminé par une lumière venant de la droite. Est-ce une erreur ? Ian Suk et Rafael Tamargo, experts en neuroanatomie à l'École de médecine de l'Université Johns Hopkins, à Baltimore, y voient plutôt le signe d'un message caché…

D'abord, ils ont vérifié qu'aucun autre cou peint par Michel-Ange n'est aussi étrange et « grumeleux ». Ensuite, ils ont montré que les détails du cou se superposent à l'anatomie d'un cerveau humain vu du dessous *(voir la figure d)*. De fait, la correspondance est étonnante.

Les détails du cou de Dieu se superposent à l'anatomie du cerveau humain.

Les deux neurologues vont plus loin et discernent la moelle épinière dans les replis du drapé qui traversent le torse *(voir la figure e)*. De plus, au niveau de l'abdomen, un pli en Y semble se prolonger par deux protubérances sphériques situées sous le thorax *(voir la figure f)* : ce schéma coïncide précisément avec un dessin du nerf optique et des yeux, effectué par Léonard de Vinci en 1487. Or les deux artistes sont contemporains et s'apprécient mutuellement.

En 1990, Frank Meshberger avait déjà remarqué que, dans la *Création d'Adam*, le rideau entourant les anges et Dieu reproduit une coupe transversale de cerveau humain *(voir les figures g et h)*. Précisons que d'autres y virent un rein, un organe qui fit beaucoup souffrir Michel-Ange…

De même que nous voyons un visage dans le moindre nuage, ces ressemblances trahissent peut-être la trop grande habitude qu'ont les deux Américains des planches neuroanatomiques. Pourtant, Michel-Ange disposait des connaissances requises pour « cacher » un cerveau. En 1493, il offre au prêtre de la basilique Santa Maria del Santo Spirito, à Florence, un crucifix en bois peint. En échange, les autorités ecclésiastiques l'autorisent à examiner, et à disséquer, le corps de défunts en provenance de l'hôpital du couvent. Plus tard, il illustrera un traité d'anatomie de son ami, le chirurgien Realdo Colombo.

Si la ressemblance n'est pas fortuite, que signifie cette planche anatomique cachée dans la *Séparation ?* Selon I. Suk et R. Tamargo, Michel-Ange aurait cherché à rehausser le sens de ce panneau. Le neurologue américain Douglas Fields s'interroge : n'y aurait-il pas là une illustration de la lutte entre la religion et la science ? Ou bien Michel-Ange postule-t-il que l'intelligence – et l'organe qui en est le support, comme Hippocrate le pensait déjà – suffit au croyant pour communiquer avec Dieu, rendant l'Église inutile ? On comprend alors les précautions prises pour dissimuler la planche anatomique ! ■

I. Suk et R. Tamargo, **Concealed neuroanatomy in Michelangelo's *Separation of light from darkness* in the Sistine Chapel**, *Neurosurgery,* vol. 66(5), pp. 851-861, 2010.