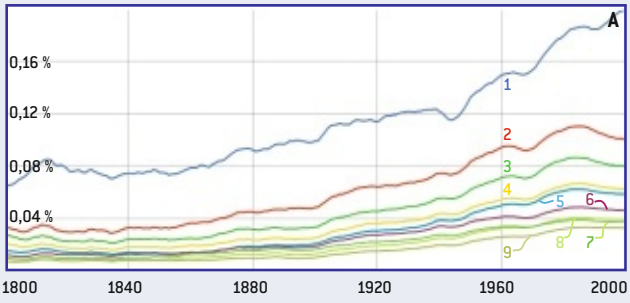
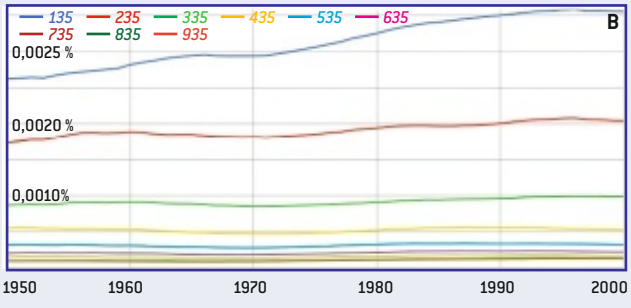
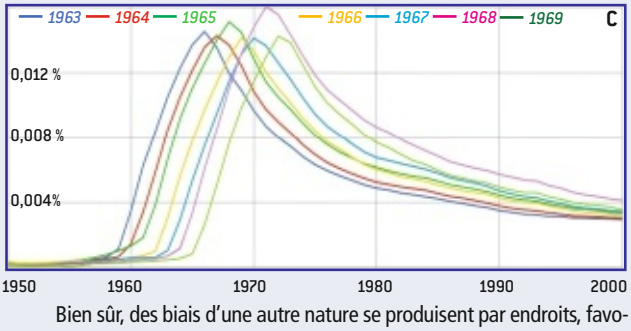
1. Des raffinements insoupçonnés de la loi de Benford

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.

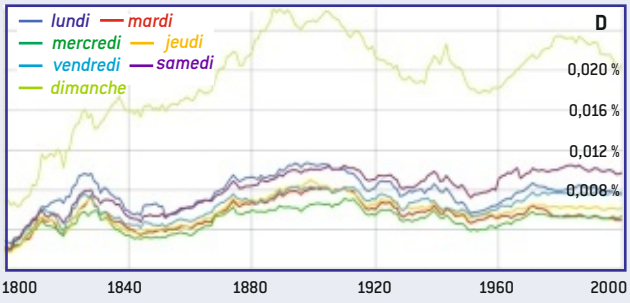


Une façon de tester cette loi est de comparer les fréquences d'usages des signes isolés 1, 2, 3, 4, 5, 6, 7, 8, 9, ce que permet le système des *n-grammes* (groupes de *n* signes) associé à la base de cinq millions de livres. La loi est bien vérifiée. La pente montante de toutes les courbes provient 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)*.





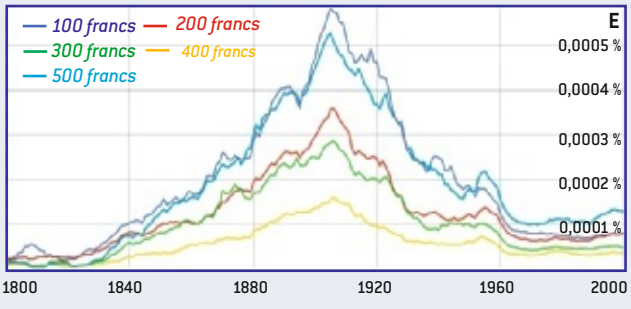
Bien sûr, des biais d'une autre nature se produisent par endroits, favorisant un nombre particulier et perturbant la loi de Benford générale au moins pendant quelques années. Les nombres ronds sont favorisés : 15 est plus utilisé que 14 ou 16, et 150 est aussi plus utilisé que 140 ou 160. Sans surprise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



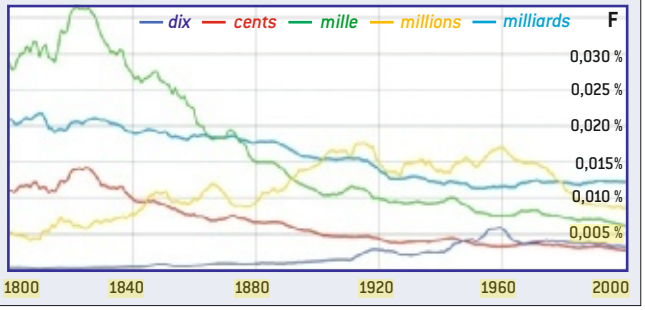
plus en plus cité. Du coup, c'est trois ou quatre ans après 1965 que 1965 atteint son pic d'utilisation. Ce décalage temporel n'est pas l'explication de tout ce qu'on observe et, par exemple, l'année 1968 est celle qui va le plus haut parmi les années comprises entre 1963 et 1969 dans les livres en français *(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 ?



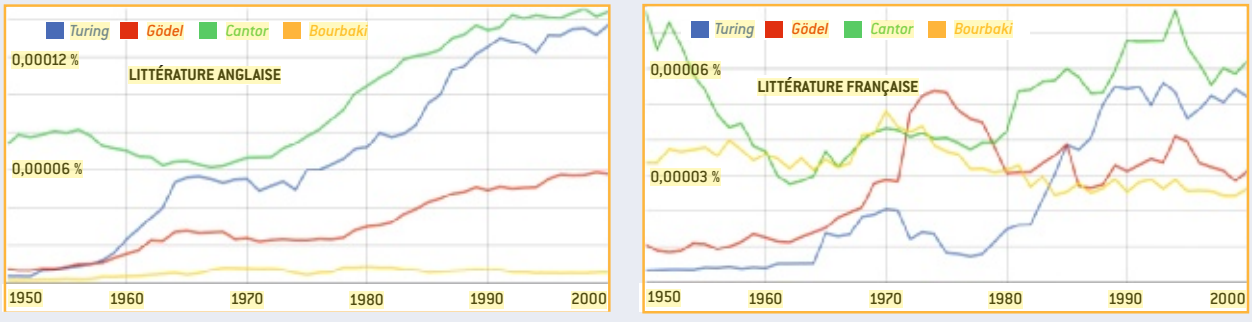
Autre mystère : est-ce parce que nous devenons plus riches, est-ce un effet de l'inflation, ou alors faut-il chercher d'autres causes encore pour expliquer que (dans la course entre *dix*, *cents*, *mille*, *millions*, *milliards*) *milliard* vient en 1980 de passer en tête devant *million* qui lui-même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