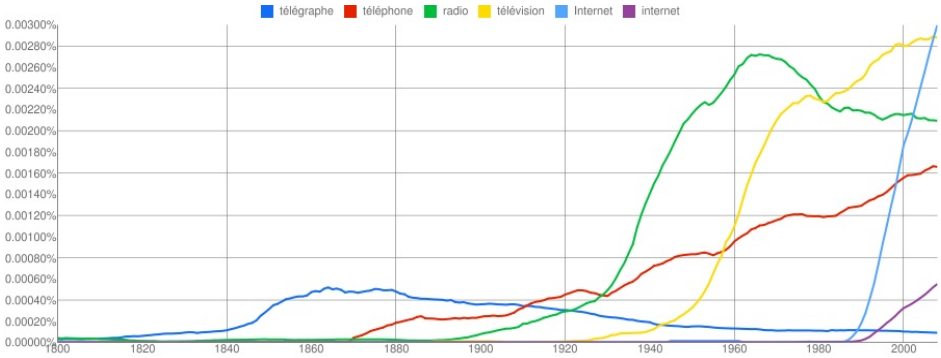
L'INTERPRÉTATION DES GRAPHIQUES PRODUITS PAR NGRAM VIEWER

LE 11 JANVIER 2011 ***PATRICK PECCATTE***

Extrêmement séduisant au premier abord, Ngram Viewer n'est, pour Patrick Peccatte, qu'"un outil heuristique qui permet plus de poser de nouvelles questions que d'apporter des réponses."

Ngram Viewer [en] est un nouvel outil mis en ligne **par Google le 16 décembre dernier** [en]. Il permet de visualiser sous forme de graphiques les fréquences d'apparition de suites de mots dans les livres numérisés depuis 2003 sur ***Google Books***. Ce projet a été initié en 2007 par un mathématicien et physicien américain, **Erez Lieberman Aiden** [en]. Il a été soutenu par **Google Labs** et développé par des chercheurs de Harvard, en particulier Jean-Baptiste Michel, jeune polytechnicien français.



L'application contient actuellement les mots extraits de plus de 5 millions d'ouvrages, ce qui correspond d'après les développeurs à 4% des livres jamais publiés. Les ouvrages les plus anciens utilisés dans le projet remontent aux XVI^{ème} siècle mais la très grande majorité sont postérieurs à 1800.

Il s'agit en fait d'un énorme lexique interrogeable contenant plus de 500 milliards de mots et organisé en sous-lexiques par langue : anglais (361 milliards de mots = Mm) [différencié en anglais américain et britannique], français (45 Mm), espagnol (45 Mm), allemand (37 Mm), russe (35 Mm), chinois (13 Mm) et hébreu (2 Mm).

Sans trop entrer dans les détails techniques, les lexiques sont des tables composées de **n-grammes**, c'est-à-dire des séquences de mots apparaissant dans les ouvrages numérisés. L'outil met ainsi en œuvre cinq catégories de tables : monogrammes (mots uniques), bigrammes (deux mots qui se suivent)… , jusqu'aux 5-grammes (cinq mots successifs). Il n'est donc pas possible de connaître à l'aide de Ngram Viewer les fréquences d'apparition du vers de Verlaine ***De la musique avant toute chose*** qui comporte six mots. Par contre, on trouvera les deux séquences de cinq mots chacune ***De la musique avant toute*** et ***la musique avant toute chose*** dont les courbes représentatives affichées par Ngram Viewer sont manifestement corrélées.

Les lexiques sont **mis à la disposition du public** [en] selon la licence Creative Commons et sous la forme de fichiers au format CSV. Bien que très volumineux, ils sont donc facilement lisibles et l'on devrait ainsi voir apparaître de nouvelles applications les utilisant. À titre d'exemple, une ligne du lexique 5-grammes français se présente ainsi :

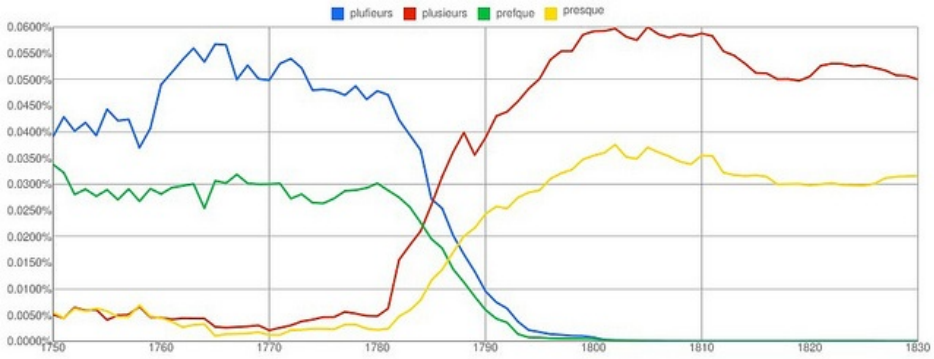
principes fondamentaux de la philosophie 1988 17 16 12

où la suite de mots *principes fondamentaux de la philosophie* est un 5-gramme, 1988 l'année de parution des livres analysés, 17 le nombre d'occurrences de la suite de mots dans l'ensemble des ouvrages de l'année en question, 16 le nombre de pages différentes et 12 le nombre de livres où la séquence apparaît. Aucune référence aux ouvrages analysés ne figure dans ces tables qui ne contiennent qu'une compilation de comptages d'occurrences.

Une masse de statistiques extrêmement sommaires et synthétiques

Ces différents sous-lexiques sont donc par construction totalement « autonomes », indépendants de Google Books. Il s'agit là manifestement d'un choix stratégique de Google qui aurait pu construire un outil beaucoup plus sophistiqué relié à sa base d'ouvrages numérisés. Le projet peut dès lors fonctionner sans qu'il soit nécessaire de mettre à la disposition des utilisateurs l'accès aux documents (initiative controversée comme on le sait). Mais ce choix comporte aussi un inconvénient majeur puisqu'il interdit de rechercher sur le voisinage plus éloigné des mots et empêche toute contextualisation des résultats (quel livre, quelle page, quel paragraphe contiennent telle suite de mots). L'utilisateur ne dispose que de statistiques extrêmement sommaires et synthétiques, mais il en voit énormément. On regrettera que les concepteurs n'aient pas facilité la tâche des analystes, ne serait-ce qu'en stockant dans chaque entrée de lexique les id Google Books des trois ouvrages qui contribuent le plus au nombre d'occurrences.

Les approximations de la reconnaissance de caractères (OCR) utilisée dans Google Books se retrouvent sur Ngram Viewer. Ainsi, la plupart des observateurs mentionnés dans la webographie sélective ci-dessous mettent en évidence l'évolution progressive de la graphie du **s long** – reconnu par l'OCR comme un *f* – vers la forme du *s* minuscule que nous connaissons actuellement.



De même, de nombreuses évolutions de graphies issues pour la plupart de diverses réformes de l'orthographe peuvent être visualisées très rapidement, et le résultat est souvent spectaculaire (exemples: ***mes parens, mes parents*** en français, ***quando, cuando*** en espagnol).

Mais on relève aussi de nombreuses erreurs d'OCR et surtout l'attribution de dates de publication erronées à des documents comme on peut le voir par exemple sur le mot ***Internet***. La réédition de certains ouvrages est certainement la cause d'un grand nombre de ces erreurs. Pour **Natalie Binder** [en], il se pourrait même à terme que l'intérêt principal de *Ngram Viewer* consiste à identifier rapidement les erreurs d'OCR et de dates sur Google Books !

L'aspect purement lexical du projet qui ne distingue pas les polysémies rend de nombreuses recherches pratiquement impossibles (essayez d'afficher la fréquence des noms de saison en français par exemple).

La culturonomique, un nouveau champ d'application de la lexicométrie

L'équipe de développement de Ngram Viewer a publié dans la revue *Science* un article intitulé ***Quantitative analysis of culture using millions of digitized books*** [pdf, en] qui introduit le terme *cultoromics* (*cultoronomique* en français) pour désigner un nouveau champ d'application de la lexicométrie. Les auteurs ont aussi lancé un site web **culturomics.org** [en]. Amalgame de *culture* et de ***genomics*** [en], domaine dans lequel plusieurs membres de l'équipe dont Erez Aiden ont travaillé, cette activité prétend en quelque sorte mettre en évidence des évolutions culturelles sur de longues périodes à travers l'analyse de fréquence portant sur de très vastes corpus de mots.

Dans leur remarquable billet ***Prodiges et vertiges de la lexicométrie*** sur le blog *Socioargu*, Francis Chateauraynaud et Josquin Debaz s'interrogent sur la pertinence de certaines recherches ignorant les évolutions du sens des mots sur de longues périodes et émettent de sérieuses réserves concernant l'ambition culturonomique. À tout le moins, la tentative manifeste de créer une nouvelle discipline en la nommant d'après un champ de recherche de la biologie et sur une seule référence de publication dans un journal scientifique semble assez immodeste et pose problème. Je renvoie sur ces questions méthodologiques et épistémologiques à l'article de *Socioargu* ainsi qu'à ceux de **Dan Cohen** [en], d'**Olivier Ertzscheid**, et à la discussion sur **Language Log** [en].

La mise en ligne de Ngram Viewer a provoqué une profusion d'exemples postés sur

différents sites ou blogs, très souvent sans aucun commentaires. Ils sont proposés sur un mode presque ludique, présentés sous un format antagonique (X *vs* Y), et comme si les courbes tracées suffisaient à mettre au jour de réels phénomènes linguistiques ou culturels. Quelques collections sont apparues (**clic**, **clic**, **clic**, **clic**, **clic** [en]) et il existe aussi une **extension pour Chrome** [en] permettant de donner directement la courbe de fréquences d'une entrée de Wikipedia en anglais.

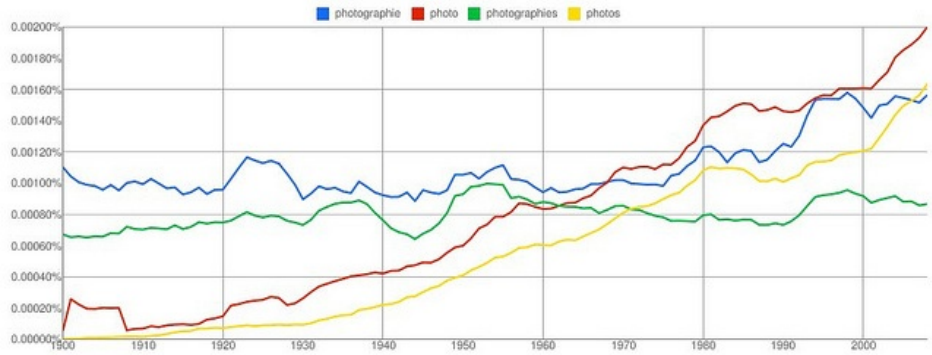
Délicate et difficile interprétation

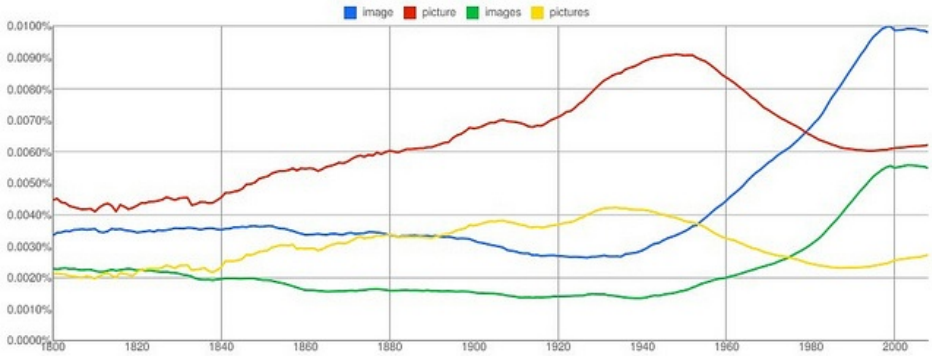
La facilité d'usage ne masque pas cependant le fait que l'interprétation de la plupart de ces graphiques est totalement impossible sans plonger dans l'analyse des documents numérisés sur Google Books. Or cette tâche est non seulement d'une ampleur colossale pour le moindre exemple de visualisation mais elle est tout simplement irréalisable en ligne puisque les documents sous copyright ne sont pas consultables. Les cas intéressants sur le plan « culturel » pour lesquels une interprétation probante peut être réalisée montrent des corrélations avec des événements historiques majeurs comme les deux guerres mondiales. C'est d'ailleurs l'un des exemples proposés par les auteurs de l'article de *Science*.

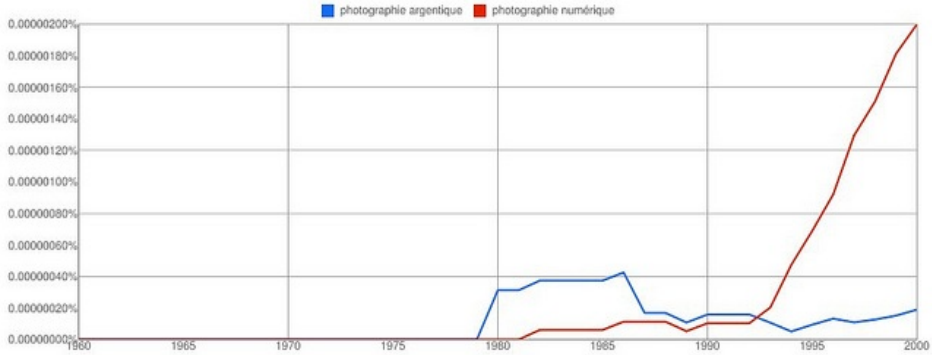
En l'absence de possibilité de vérification des hypothèses que l'on peut être amené à formuler sur une visualisation, l'utilisateur est laissé seul avec ses propres connaissances et intuitions en face du phénomène ou de l'artefact repéré. Comme le signalent les auteurs de l'article de *Socioargu* mentionné, cela signifie que l'investigateur doit d'abord *« disposer d'une culture générale suffisante pour comprendre le positionnement relatif des mots dans le temps »*.

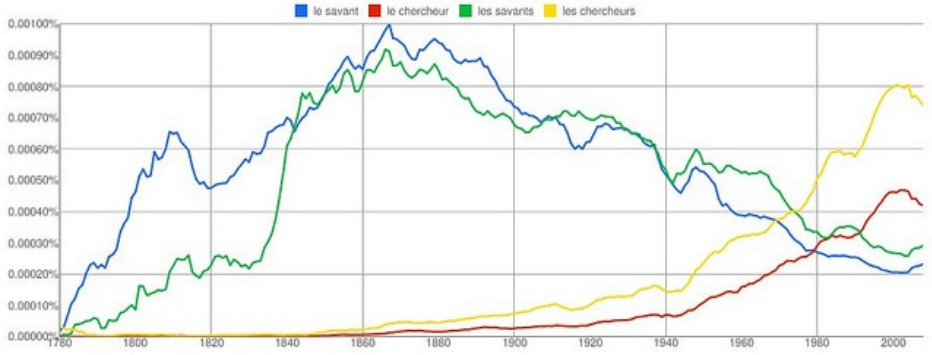
Ngram Viewer doit en fait être considéré comme un outil heuristique qui permet plus de poser de nouvelles questions que d'apporter des réponses. Pour commencer à dépasser le stade du jeu avec *Ngram Viewer*, il serait intéressant de mettre en commun les efforts de groupes de spécialistes intéressés par un sujet en ouvrant des espaces de discussions sur des visualisations, créer en somme une véritable activité de travail collaboratif à partir des graphiques produits permettant de documenter et approfondir les résultats. Un début d'interprétation de ces vastes mais très sommaires lexiques pourrait alors être envisagé et ouvrir des champs de réflexion nouveaux pour les *digital humanities*.

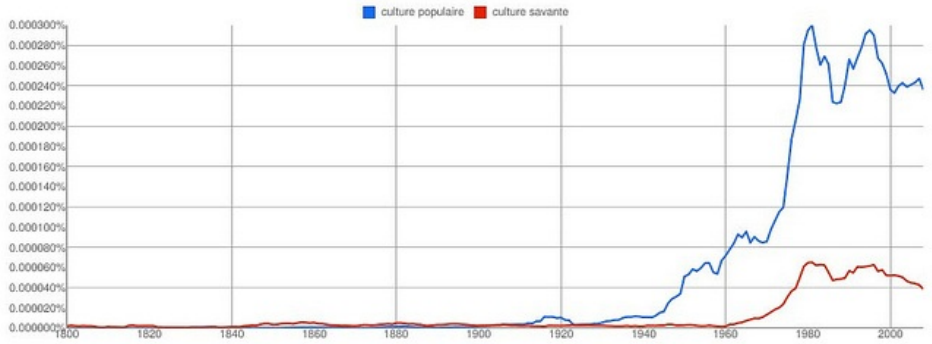
Pour terminer, voici quelques exemples de résultats en relation avec des questions diverses abordées sur *Culture Visuelle*.

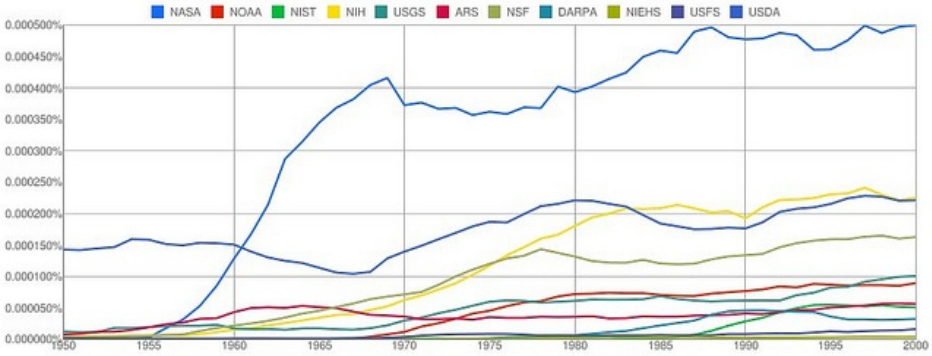


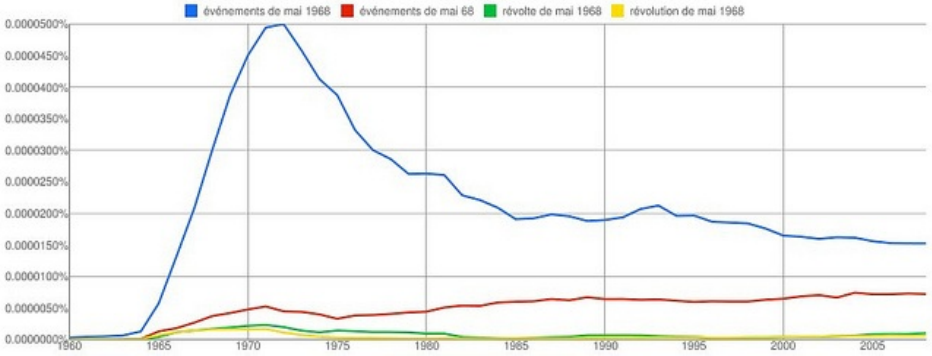


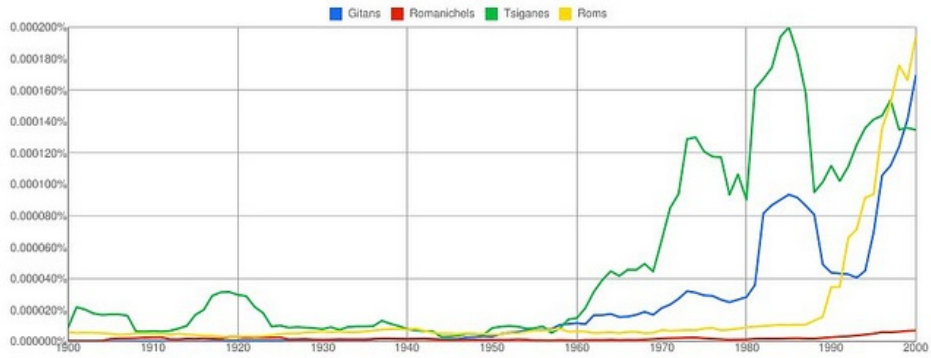

photographie argentique
photographie numérique
0.00000200%
0.00000180%
0.00000160%
0.00000140%
0.00000120%
0.00000100%
0.00000080%
0.00000060%
0.00000040%
0.00000020%
0.00000000%
1960
1965
1970
1975
1980
1985
1990
1995
2000

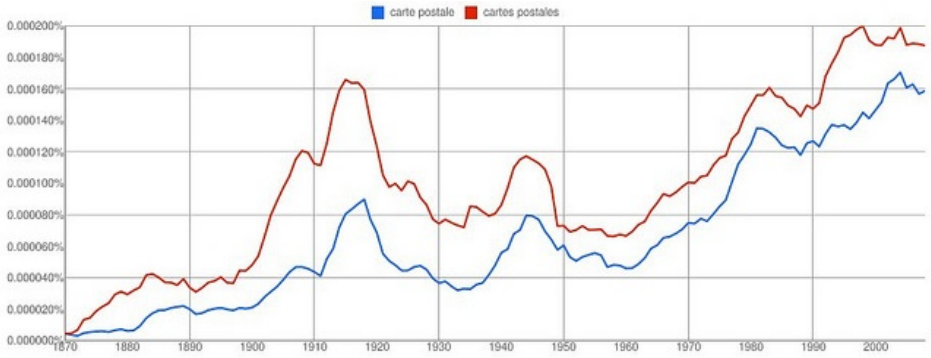

le savant
le chercheur
les savants
les chercheurs
0.00100%
0.00090%
0.00080%
0.00070%
0.00060%
0.00050%
0.00040%
0.00030%
0.00020%
0.00010%
0.00000%
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000


culture populaire
culture savante
0.000300%
0.000280%
0.000260%
0.000240%
0.000220%
0.000200%
0.000180%
0.000160%
0.000140%
0.000120%
0.000100%
0.000080%
0.000060%
0.000040%
0.000020%
0.000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000


NASA
NOAA
NIST
NIH
USGS
ARS
NSF
DARPA
NIEHS
USFS
USDA
0.000500%
0.000450%
0.000400%
0.000350%
0.000300%
0.000250%
0.000200%
0.000150%
0.000100%
0.000050%
0.000000%
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000


événements de mai 1968
événements de mai 68
révolte de mai 1968
révolution de mai 1968
0.0000500%
0.0000450%
0.0000400%
0.0000350%
0.0000300%
0.0000250%
0.0000200%
0.0000150%
0.0000100%
0.0000050%
0.0000000%
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005





Webographie sélective

En anglais

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. *Quantitative Analysis of Culture Using Millions of Digitized Books* . Science (Published online ahead of print: 12/16/2010). Disponible sur **librarian.net** [format PDF]
Google Opens Books to New Cultural Studies , John Bohannon (*Science,* 17, décembre 2010) [pdf]
Google Books Ngrams and the number of words for "snow" , Natalia Cecire (17 décembre 2010)
Google's word engine isn't ready for prime time / The problem with Google's thin description / Fixing Google's word engine, Natalie Binder (17-21 décembre 2010)
Initial Thoughts on the Google Books Ngram Viewer and Datasets , Dan Cohen (19 décembre 2010)
On "culturomics" and "ngrams", Language Log, 23 décembre 2010

En français

Google: Le plus grand corpus linguistique de tous les temps, Jean Véronis (16 décembre 2010)
Culturonomics : juste une question de corpus ? , Olivier Ertzscheid (16 décembre 2010)
Google Ngram viewer : un extraordinaire corpus mais…, Rémi Mathis (20 décembre 2010)
Prodiges et vertiges de la lexicométrie, Francis Chateauraynaud et Josquin Debaz (23 décembre 2010)
Culturomics. Google met la culture à portée de tous… ou corporifie la culture humaine ?, Corinne Dangas (28 décembre 2010)
Google labs Books Ngram Viewer : un nouvel outil pour les historiens ?, Emilien Ruiz (29 décembre 2010)

—

Billet **initialement publié sur Déjà vu**, un blog de Culture Visuelle

Image CC Flickr **Oberazzi**

LUCHO
le 11 janvier 2011 - 19:24 • SIGNALER UN ABUS - PERMALINK

http://ngrams.googlelabs.com/graph?content=sexe%2C+d%C3%A9pression&year_start=1800&year_end=2000&corpus=7&smoothing=3

Pour partage / sans comentaire

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

LAURENT BLAISE
le 11 janvier 2011 - 22:20 • SIGNALER UN ABUS - PERMALINK

Il n'empêche que cet outil permet de donner une vision intéressante de l'évolution des concepts et des effets de mode. Exemple : PNB et PIB http://ngrams.googlelabs.com/graph?content=PNB,PIB&year_start=1800&year_end=2010&corpus=7&smoothing=3

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

1 ping

Les tweets qui mentionnent L'interprétation des graphiques produits par Ngram Viewer » Article » OWNI, Digital Journalism -- Topsy.com le 11 janvier 2011 - 17:29

[...] Ce billet était mentionné sur Twitter par damien douani, Frédéric Clavert et des autres. Frédéric Clavert a dit: @ppeccatte C'est toi qui a écrit ça: http://j.mp/gK0p8m Il est très bon, cet article. [...]