



**HAL**  
open science

## Au fond du GOOFRE, un gisement de 44 milliards de mots

Étienne Brunet

► **To cite this version:**

Étienne Brunet. Au fond du GOOFRE, un gisement de 44 milliards de mots. 11ème Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012), Jun 2012, Liège, Belgique. pp.7-21. hal-01790505

**HAL Id: hal-01790505**

**<https://hal.science/hal-01790505>**

Submitted on 28 May 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## **Au fond du GOOFRE, un gisement de 44 milliards de mots**

Etienne Brunet, BCL(CNRS), Université de Nice Côte d'Azur

On connaît le projet pharaonique dans lequel Google s'est lancé il y a quelques années : procéder au recensement, à la saisie, à l'indexation et à la mise sur réseau des textes imprimés dans le monde depuis l'origine de l'imprimerie. Cette sorte de Bibliothèque d'Alexandrie reconstituée a pris l'allure d'une tour de Babel moderne, non seulement parce que les fondations sont gigantesques mais aussi parce que, sans attendre l'achèvement, la zizanie lézarde la construction, comme si, sous le coup de la colère divine, on ne parlait plus la même langue. Étant privé, dominateur et américain, le projet a suscité l'hostilité de beaucoup d'états, de bibliothèques, d'éditeurs, voire d'auteurs. Pourtant, des arrangements partiels se sont glissés dans les interstices du droit d'auteur, et, si la diffusion des textes est encore bridée par le copyright, cela n'a pas empêché les promoteurs de scanner les ouvrages à bras raccourcis sans se préoccuper des propriétaires. De la même façon les chemins, les paysages et les propriétés du monde entier ont subi le viol des satellites et des caméras baladeuses de Google, en n'épargnant guère que les visages et les plaques d'immatriculation. Les propos que nous tenons présentement dans cette salle sont peut-être déjà dans les tuyaux de *Google Books* ou le seront d'ici peu, comme les Actes précédents des JADT. Libre à chacun de s'en féliciter ou de s'en plaindre. J'imagine que beaucoup partageront ma neutralité, que favorise l'équilibre des avantages et des inconvénients. Certes les centimes sont perdus que les productions scientifiques pourraient espérer en droits d'auteur. Mais une once de lisibilité et de notoriété s'ajoute ainsi au patrimoine de chacun.

## 1. La base *Culturomics*. Étendue, fonctions et résultats

**a** – Dans le cas des statisticiens cependant, la neutralité peut être intéressée. Car la statistique a une soif dévorante de données et particulièrement des grandes masses de données qui bénéficient de la loi des grands nombres. Certes la partie dépouillée jusqu'ici, selon la présentation du projet faite en décembre 2010 dans *Science* (Michel *et al.*, 2010), ne représente qu'une faible partie des livres imprimés, et même de ceux qui ont survécu et qui attendent le scanner. Les auteurs de l'article avouent n'avoir pas dépassé 12 % de la masse à saisir. Quant à la frange réellement exploitée elle se limite à 4 %. Mais cette frange est grosse de **500 milliards de mots**. Même en s'en tenant au domaine français les chiffres brutalisent l'imagination : 500 000 volumes complets, un million de mots différents, 44 milliards d'occurrences. Jamais un dictionnaire de fréquences n'avait atteint de telles dimensions. *Frantext* culmine actuellement à 248 millions de mots. Les auteurs ne se privent pas de s'amuser avec les chiffres : la séquence des lettres traitées dépasse de mille fois celle du génome humain et, pour ceux qui n'ont jamais observé le génome de visu, ils évoquent la distance de la terre à la lune qu'il faudrait parcourir plus de dix fois, aller et retour, si l'on voulait écrire, bout à bout, sur une seule ligne, les textes recensés.

**b** – Mieux encore il ne s'agit pas d'un dictionnaire de fréquences qui ne distribuerait que des chiffres. Le texte même n'est pas hors d'atteinte. Certes quand un mot a des millions d'occurrences, personne n'aura la patience de consulter et de rapatrier les contextes correspondants. Mais cela reste possible pour les mots ou les expressions rares, même si les interdits du copyright limitent le contexte à deux ou trois lignes. Si l'on se satisfait d'un contexte court, de 1 à 5 mots consécutifs, la base donne des réponses immédiates qui prennent la forme d'une courbe chronologique ou de plusieurs courbes superposées. Nous choisirons pour exemple la clause finale de la correspondance française. À l'époque classique il était d'usage de se déclarer le serviteur de son correspondant, même s'il était de rang inférieur ou si on lui adressait des reproches ou des insultes. Cette locution cérémonieuse et un peu hypocrite tend à disparaître depuis la Révolution et la courbe est descendante qui dans le graphique 1 reproduit l'expression *VOTRE SERVITEUR* suivie d'un point. Deux formules plus modernes – mais pas nécessairement plus sincères – ont pris la relève et se font concurrence : *MEILLEURS SENTIMENTS*. et *SENTIMENTS LES*

MEILLEURS.. Quoique la seconde soit plus tardive, moins naturelle et moins heureuse, elle gagne du terrain et semble devoir l'emporter à terme<sup>1</sup>.

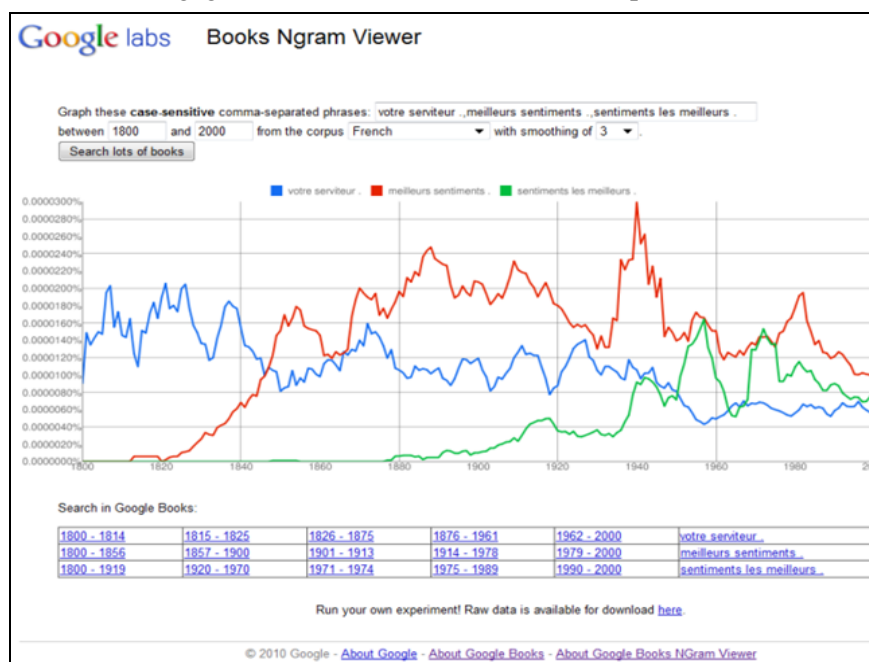


Figure 1. Exemple d'interrogation de la base française du site Google Books

c – Les auteurs de l'article multiplient les exemples qui tour à tour illustrent l'évolution de la langue, les changements de société ou les spécificités d'une époque ou d'un lieu. Quand les verbes anglais ont en concurrence des formes régulières et irrégulières, ils tendent à s'aligner sur le modèle standard, surtout de l'autre côté de l'Atlantique. La population des mots n'est pas soumise à la régulation que tentent d'imposer les dictionnaires. Les naissances l'emportent sur les disparitions, les unes et les autres échappant au contrôle des gardiens de la langue, dont les arrêts ont bien souvent une justification arbitraire. Et précisément la statistique à grande échelle pourrait les éclairer pour l'adoption d'un vocable que l'usage a confirmé ou l'abandon d'un autre que frappe l'oubli. Un conservatisme prudent peut être le rempart contre

1. L'adresse du serveur a changé. Ce n'est plus Google Labs mais Google Books <http://books.google.com/ngrams>. La base a reçu le nom de *Culturomics*. On peut choisir la langue du corpus exploré, le mot ou l'expression recherchée (ou plusieurs la virgule servant de séparateur) et le pas adopté pour le lissage de la courbe.

l'inflation lexicale. Mais s'il est trop ou trop peu réactif, il risque d'être débordé par le marché linguistique, qui accompagne l'histoire en marche et n'a d'autre loi que celle de l'offre et de la demande. Ainsi la « Grande Guerre » perd ses privilèges d'unicité quand se déclenche une seconde guerre mondiale. Elle n'est plus alors désignée que comme la première, et passe de l'absolu au relatif. De 1800 à 2008, les promoteurs de la base s'attachent surtout à voir dans les remous qui agitent les mots le reflet des mouvements ou des ruptures qui ébranlent la société.

Cela est sensible surtout dans les noms propres : certains noms peuvent être caviardés dans un pays parce qu'ils déplaisent au régime en place, comme celui de Chagall ou Picasso au pays des nazis, celui de Trotsky au temps de Staline ou la place Tiananmen dans la Chine d'aujourd'hui. Comme la base a été construite parallèlement dans sept langues différentes, et donc chez des populations distinctes, on peut croiser l'espace et le temps, suivre l'histoire contrastée des peuples et observer les courants et contre-courants qui se produisent sur tel ou tel point du globe terrestre. Pourtant, à partir d'un survol en orbite, les clapotis locaux se perdent dans la dérive d'ensemble qu'on observe dans les mots et les faits de société dont les mots portent témoignage. En deux siècles les mouvements se font de plus en plus rapides : les savants, les artistes, les figures de la société passent plus rapidement sur la scène. Ils arrivent plus jeunes à la notoriété, mais la gloire les abandonne plus vite. Est-ce dû seulement à la rapidité et à la puissance des communications ? N'y a-t-il pas aussi un appétit plus vorace de la nouveauté, une soif d'information agissant comme une drogue, l'agitation grandissante d'une société où la vie et l'ennui durent plus longtemps ?

## **2. La base Culturomics. Insuffisances**

**a** – De telles questions ne peuvent pas ne pas être évoquées quand on a devant soi, pour la première fois, un tel gisement, si riche et si peu exploité. Pour explorer le passé lointain de l'homme, et rendre compte de milliers d'années, on n'a que quelques milliers d'os épars sur la surface de la terre. L'histoire des deux derniers siècles, elle, dispose de 500 milliards de traces écrites laissées par l'homme, toutes datées et localisées. Reste à savoir si ces témoins sont fiables et si la qualité accompagne la quantité. On est donc amené à faire la critique de cette réalisation sans cacher ses points faibles : scannés à la chaîne et confiés à la lecture optique, sans intervention humaine préalable et sans correction postérieure, les textes regorgent d'**erreurs de lecture**, dues à la qualité

variable du papier, à l'encrage incertain, au traitement malheureux des indices, des exposants et des appels de notes, à l'élasticité de l'interlettrage (blancs non reconnus et mots collés). Toutes ces imperfections sont attendues d'un automate et on les retrouve dans la plupart des réalisations où la lecture échappe à l'œil humain. Mais on pouvait espérer plus d'attention portée aux difficultés supplémentaires que contenait le projet. Un apprentissage plus fin et moins expéditif aurait pu être appliqué aux polices inhabituelles des documents anciens.

En particulier l'automate n'a pas su reconnaître les **s longs** de l'ancienne typographie (confondus avec des **f**). Et cette seule négligence a des effets incalculables dont témoigne la liste des spécificités de la première période (tableau 2). Les éléments insolites qui se portent en tête de liste (**EFT**, **FE**, **FUR**, **FA**) sont des avatars mal repérés des formes régulières (**EST**, **SE**, **SUR** et **SA**) et cela jette la suspicion sur tous les mots où l'on trouve un **s** ou un **f**, c'est-à-dire un tiers du lexique français.

Tableau 2. Les spécificités de la première période (1800-1840)

N°	écart	corpus	texte	mot
1	4253.21	3140768	2555713	étoit
1	4152.23	3949728	2848331	avoit
1	3960.18	195664144	35784324	;
1	3120.05	1280050	1178085	eft
1	2630.89	242179314	36059703	qu'
1	2522.23	1071243	883501	étoient
1	2279.89	1142929	838962	avoient
1	2059.68	411269418	52964418	que
1	2059.05	883838	663952	enfans
1	2022.51	837551	634116	fe
1	1998.37	53798781	9668157	vous
1	1861.42	175043732	24583789	on
1	1858.26	692574	529092	tems
1	1835.28	17471386	4007321	roi
1	1784.60	88742448	13752541	je
1	1734.17	379543570	47459314	il
1	1725.60	903139	577683	seroit
1	1686.97	687651	484637	habitans
1	1670.47	679676	477372	auoit
1	1609.60	546573	408475	pouvoit
1	1542.44	75773145	11466126	ils
1	1446.42	22902414	4319099	i
1	1410.82	323648405	39448288	qui
1	1393.91	687083	412015	ame
1	1384.09	198259170	25339482	ne
1	1358.64	32750023	5543476	point
1	1328.25	6755990	1694430	prince

b – À ces défauts de la saisie mécanique, s'ajoutent ceux du **traitement informatique**. Passons sur le codage en unicode UTF8 qui

range les caractères accentués à la fin de l'alphabet<sup>2</sup>. Comme l'interrogation sur Internet ne classe pas les mots, cette difficulté passe inaperçue. Il n'en va pas de même avec les majuscules. Celles qui sont en début de phrase n'ont pas été neutralisées et les mots qui se trouvent à cette place sont traités comme des noms propres. Cela oblige l'utilisateur à proposer deux orthographes pour un même mot, afin d'être sûr de rassembler toutes les occurrences. S'il s'agit d'un nom, il devra doubler encore le nombre de formes afin d'obtenir le singulier et le pluriel. La tâche pour un adjectif et surtout pour un verbe se complique encore. Proposer ensemble toutes les formes conjuguées d'un verbe français va provoquer autant de courbes superposées au risque de rendre le graphique illisible. Enfin le traitement de l'apostrophe est catastrophique pour le français, le séparateur ayant été placé devant et non derrière l'apostrophe, si bien qu'on n'a plus le moyen de distinguer de *T* épenthétique de *VA-T-EN* et le *T*' pronom personnel de *JE T'AIME* (il en va de même pour *M'*, *S'*, *L'*, *C'*, *D'*, *N'*, *J'*, tous mots de très grande fréquence dont le total se compte en milliards d'occurrences). Un peu d'attention aux particularités du français et des langues latines aurait facilement corrigé cette bétise : il suffisait d'une ligne de code. L'on s'étonne qu'un français, Jean-Baptiste Michel, l'un des principaux signataires de l'article de *Science*, n'ait pas mieux surveillé l'application au français d'un programme fait pour l'anglais. Et l'on a tout lieu de s'inquiéter pour le traitement des autres langues. En comparaison les données de *Frantext*, quoique engrangées 50 ans plus tôt et soumises aux changements de la technologie, sont beaucoup plus fidèles. Elles ont été saisies au clavier et dûment contrôlées et corrigées après saisie. Elles sont aussi beaucoup plus riches d'information, un traitement proprement linguistique (distinction des homographes, regroupement des formes de la même vedette) s'étant ajouté aux résultats bruts de l'automate.

**c** – Quoique avouées partiellement dans l'article de présentation, les insuffisances du traitement peuvent n'être pas perçues par l'utilisateur, par suite d'un **camouflage** inconscient ou délibéré : Les avortons lexicaux, à qui il manque un bras ou une lettre, ont été éliminés, le seuil de fréquence pour survivre étant de 40 occurrences minimum. En revanche on a laissé leur chance à ceux qui ont franchi cette barre, même

---

2. Les bibliothèques logicielles traitant des chaînes de caractères UTF-8 (ou plus généralement Unicode) peuvent pour la plupart gérer l'ordre alphabétique de chaque langue en exploitant des normes élaborées par les différents pays. Les données qu'a fournies Google étaient triées, mais sans tenir compte de la langue.

s'ils restent boiteux ou monstrueux. Ainsi on peut obtenir la courbe de TOJOURS, TOUJURS ou de CEPNDANT, CEPNDANT, CEPNDONT. Cela prouve qu'un nettoyage ou une correction par comparaison avec la nomenclature officielle n'ont pas été faits. Le mode d'interrogation tend aussi à éviter les bévues. L'utilisateur ne cherche pas à poser des pièges. Les mots qu'il propose sont de bon aloi. Et les réponses qu'il obtient semblent l'être aussi. Mais si son doigt dérape sur le clavier, le mot tordu suscitera aussi des réponses, comme il arrive avec le moteur de recherche *Google* quand on tape des lettres au hasard. Quant au retour au texte, qui est possible et qui théoriquement permettrait le contrôle *a posteriori*, il est en réalité assez illusoire. Les renvois se chiffrent par millions pour un même mot et le dernier exemple ne peut être atteint qu'en passant séquentiellement par tous les autres. D'ailleurs les liens qui mènent au texte et à *Google Books* restent lâches ; les fréquences qu'on y lit en valeur absolue sont très approximatives et souvent mal accordées à celles qui sous-tendent les courbes. En réalité la distorsion vient d'un écart grandissant entre une base statistique, fixée et figée en 2009, et un agrégat énorme de textes ou de documents qui s'accroît exponentiellement dans *Google Books* et *Google*.

**d** – Enfin rien ne permet de connaître les **fréquences réelles**. On n'obtient que des points sur une courbe, issus d'une fréquence relative sous-jacente, et lissés par le procédé des moyennes mobiles. En conséquence aucune synthèse n'est possible. On est limité à une interrogation par **mots individuels**, sans même disposer du regroupement des formes qui appartiennent à la même entrée. Pas de listes, comme en propose *Frantext*. Pas de tableaux. Partant pas d'analyses multidimensionnelles, ni arborées, ni factorielles. Pas de spécificités. Pas de distance intertextuelle. Pas de synthèse sur l'évolution. On est devant un immense réservoir qui ne distribue son contenu que goutte à goutte, mot à mot. D'où l'idée d'installer des bassins de décantation et même une usine de traitement.

### 3. La base GOOFRE. Construction et avantages

**a** – On a donc complété *Culturomics* avec GOOFRE comme on a complété *Frantext* avec THIEF (*Tools for Helping Interrogation and Exploitation of Frantext*). Cela a été rendu possible par un geste généreux de Google : la livraison téléchargeable des fichiers sources. Leur volume est considérable. Pour le seul domaine français, la collection complète compte 1500 fichiers, chacun occupant 100 Mo. Dans un premier temps



nous nous sommes contenté des unigrammes (ou mots individuels). Il a fallu remédier aux inconvénients du codage unicode (UTF8) qui enregistre les caractères accentués sur deux octets et bouleverse l'ordre alphabétique, neutraliser les majuscules, et surtout **concentrer les données** pour qu'elles puissent être compatibles avec les ressources d'un ordinateur ordinaire. Au lieu d'occuper 200 lignes du fichier original, chaque mot est aligné sur une seule, où l'on trouve le cumul (en mots, pages et textes) et le détail de la répartition dans 12 tranches chronologiques.

Car on a cru devoir **redresser la perspective** : comme les données se présentent année par année, avec de grandes disparités, les dernières ayant 10 ou 100 fois plus de documents enregistrés que les premières, on a procédé à un équilibrage pour que les tranches aient un poids comparable :

<i>Tranche</i>	<i>Année moyenne</i>	<i>Taille (en milliards de mots)</i>
1800-1840	1820	4,3
1840-1870	1855	5,9
1870-1900	1885	5,9
1900-1920	1910	3,6
1920-1940	1930	2,7
1940-1960	1950	2,6
1960-1972	1966	3,4
1972-1983	1978	3,2
1983-1993	1988	3,1
1993-2000	1996	2,7
2000-2005	2003	3,6
2005-2008	2007	2,8

On comparera dans les tableaux 3 et 4 les données brutes qu'on obtient de *Google Books* en activant le téléchargement (lien sous *here* en bas à droite du graphique 1) à celles que fournit notre base GOOFRE après traitement.

**b** – Qu'avons-nous perdu au change ? Un peu de précision : le grain fin de l'année s'est un peu épaissi avec le rythme de la décennie. Qu'avons-nous gagné ? Beaucoup de place : une ligne contient ce qui était éparpillé sur 200. De la **sécurité** aussi : les fréquences relatives cachées derrière les points de la courbe cèdent la place à des écarts réduits fondés sur le schéma d'urne. Et l'urne est si énorme qu'elle éteint tout scrupule à l'égard de la loi normale. On a gagné encore de l'**autonomie** : tout en conservant les 44 milliards de mots de l'enquête initiale, la base GOOFRE en offre l'exploitation dans un fichier de

70 Mo, facile à installer sur un CD ou un disque dur, sans qu'il soit nécessaire de recourir à Internet.

**Tableau 3. Extrait des données brutes obtenues de Google Books (ngrammes de longueur 5)**

(On observera dans la première ligne le traitement inapproprié de l'apostrophe, considérée comme un mot et non comme le résultat l'élision)

principe d ' un congrès 2004	1 1 1				
principe d ' un congrès 2005	1 1 1				
principe d ' un congrès 2006	2 2 2				
principe d ' un congrès 2007	2 2 2				
principe de la propriété est	1828	1 1			
principe de la propriété est	1829	2 2 2			
principe de la propriété est	1834	1 1 1			
principe de la propriété est	1835	1 1 1			
principe de la propriété est	1843	2 2 2			
principe de la propriété est	1844	3 3 3			
principe de la propriété est	1848	7 7 7			
principe de la propriété est	1849	4 4 4			
principe de la propriété est	1850	2 2 2			
principe de la propriété est	1851	15	15	11	

Annotations dans l'image :

- Année où le ngramme est repéré (pointe vers la colonne des années)
- nombre d'occurrences du ngramme (pointe vers la colonne des occurrences)
- nombre de pages (pointe vers la colonne des pages)
- nombre de textes (pointe vers la colonne des textes)

**Tableau 4. Extrait des données traitées par GOOFRE (unigrammes)**

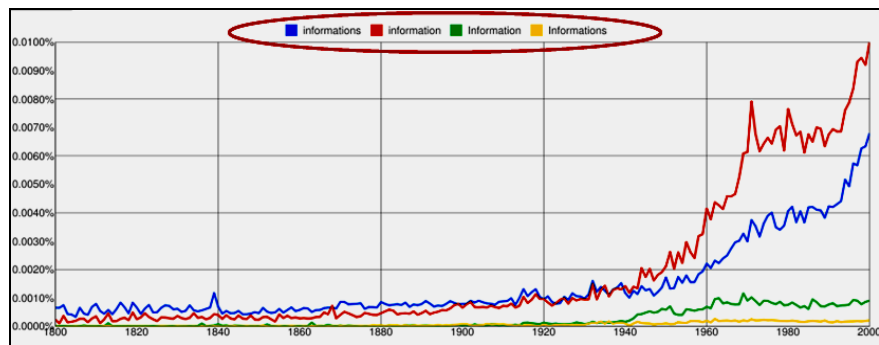
(Les trois premiers nombres cumulent ceux de la base pour l'ensemble du corpus, et les 12 derniers détaillent la répartition des occurrences dans la chronologie)

zénitale	316 260 118 , 152 23 89 5 19 10 4 6 5 0 1 2
zénithale	8993 6509 1896 , 399 2260 2101 1089 620 778 573 259 233 158 255 268
zénithales	4938 3811 1108 , 418 1662 1081 582 279 380 237 104 67 29 50 49
zéphire	676 646 540 , 210 162 92 47 34 19 21 14 27 10 15 25
zéphyr	6708 6321 4740 , 1732 1306 863 341 341 243 314 264 285 390 312 317

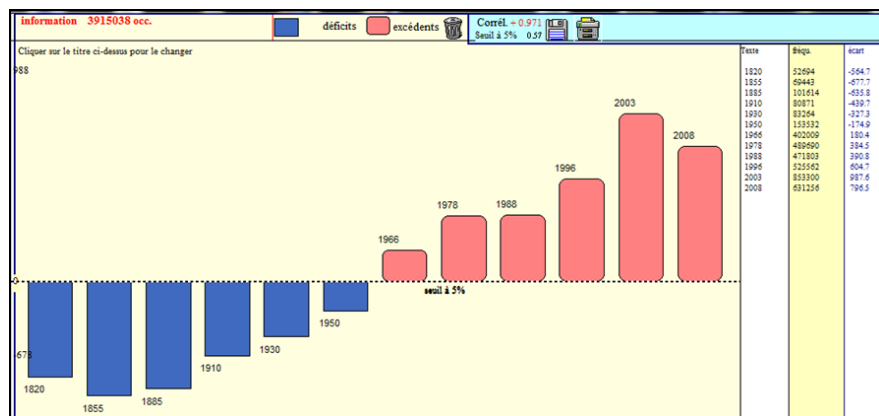
Mais l'avantage décisif vient de la **liberté de sélection** et de la **puissance des traitements**. La base ainsi concentrée est pourvue des différentes fonctions statistiques du logiciel HYPERBASE. Elle s'offre aux sélections diverses, aux regroupements, aux graphiques de toute sorte, aux calculs de spécificités et de distance intertextuelle, et se prête d'emblée aux méthodes multidimensionnelles. Les chercheurs en sociologie, en histoire ou en linguistique sont familiarisés aux outils modernes de l'exploitation statistique. Ils attendent autre chose que des courbes parcellaires, lissées, et détachés de l'ensemble, dont il est difficile de faire une synthèse. Avec seulement le cours de la bourse pour ses propres actions, un boursicotier ne risque guère de faire fortune.

S'il s'agit d'un mot unique, comme le mot INFORMATION (graphiques 5 et 6), les résultats sont bien sûr parallèles. Mais si l'interrogation porte sur un verbe, la base Google Books se trouve démunie ne pouvant guère représenter simultanément et de façon lisible la trentaine de formes qui se rattachent au paradigme. La base GOOFRE ne s'effraie pas de réunir sur la même photo près de cinquante formes du verbe ETRE et près d'un milliard d'occurrences, et même de confronter le verbe AVOIR au verbe ETRE, en constatant leur déclin concomitant (figure 7). Rien n'empêche

d'aller plus loin et d'étendre l'enquête à d'autres verbes : POUVOIR, SAVOIR, VOULOIR, FALLOIR, ALLER, FAIRE, lesquels invitent aux mêmes conclusions<sup>3</sup> (figure 8). Le verbe français tend à disparaître de la phrase française, ce qui est paradoxal, vu que, mis à part les phrases nominales, la phrase se construit, par définition, autour du verbe.



Graphique 5. Le profil des quatre formes de l'INFORMATION dans Google Books



Graphique 6. L'histogramme de l'INFORMATION dans GOFRE (toutes formes réunies)

3. Il est difficile d'aller jusqu'à l'exhaustivité dans la répartition des parties du discours, quand un corpus n'a pas été lemmatisé. Si l'on prend appui sur le seul dictionnaire des fréquences, en dehors du contexte, il devient impossible de distinguer les homographes et le mettre la forme MARCHE ou MARCHE dans la catégorie du verbe ou du substantif. Cet embarras obère l'exploitation de la base GOFRE, mais aussi celle de THIEF, dont les données sont antérieures à la lemmatisation de *Frantext*.

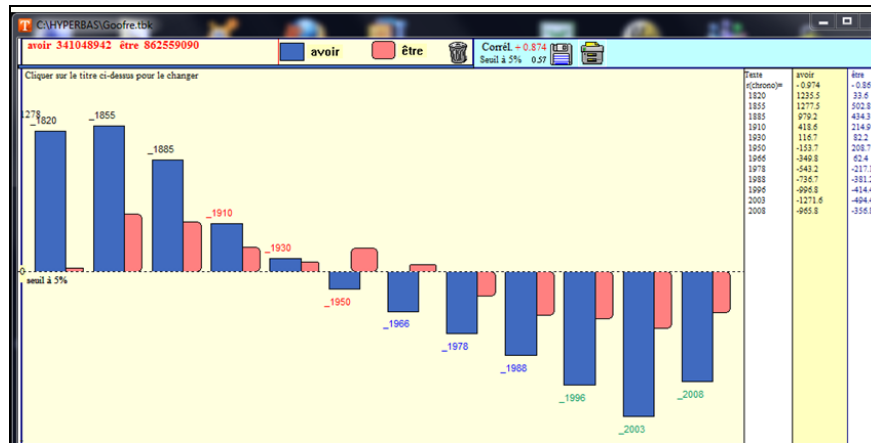


Figure 7. ETRE et AVOIR dans la base GOOFRE (plus d'un milliard d'occurrences)

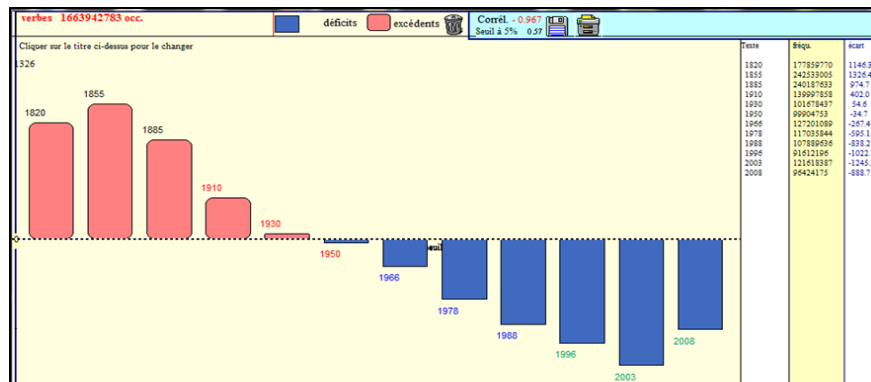


Figure 8. POUVOIR, SAVOIR, VOULOIR, FALLOIR, ALLER, FAIRE dans la base GOOFRE (1,7 milliard)

#### 4. Corpus de référence. Google ou Frantext ?

a – En réalité, les milliards d'observations engrangées ne doivent pas trop intimider la raison. Si étendu que soit le corpus, il a ses caractéristiques qui ne coïncident pas nécessairement avec d'autres corpus, qui eux aussi peuvent être représentatifs de la langue française. Ainsi la courbe du verbe dans *Frantext* (figure 9) indique bien une baisse sur le long terme : les siècles classiques lui donnent une part plus importante que l'époque moderne. Mais sur les deux derniers siècles – distance recouverte par *Google* – la tendance, d'abord à la baisse, se

redresse ensuite. On a tout lieu de penser que si les courbes divergent ainsi, c'est que les deux corpus n'ont pas le même contenu. Certes l'un et l'autre représentent l'écrit, qui fait moins usage du verbe que l'oral. Mais les textes recueillis dans *Frantext* représentent expressément la langue littéraire alors que les promoteurs de *Google* n'ont apparemment pas partagé cette préférence. Beaucoup des textes enregistrés dans la base *Google*, surtout les plus récents, sont entrés directement, sous la forme numérique où ils se trouvaient déjà. Et la production littéraire ne représente qu'une faible part dans ce flot de publications techniques, qui couvrent toutes les disciplines. La différence entre les deux corpus est dans l'opposition littéraire vs utilitaire. Or l'on a observé sur d'autres corpus, qui mêlaient les deux variétés, que le verbe se maintenait vivant dans l'écrit littéraire comme dans l'oral, alors que le langage utilitaire favorise le substantif, porteur de l'information. On a longtemps cru que la fracture principale était entre **l'écrit et l'oral**. Il semble plus judicieux maintenant d'opposer **le littéraire et l'utilitaire**. Face à la langue technique, qui véhicule moins d'émotion que d'information, l'oral et la littérature ne sont plus face à face, mais côte à côte, du côté de l'expressivité et du côté du **verbe**.

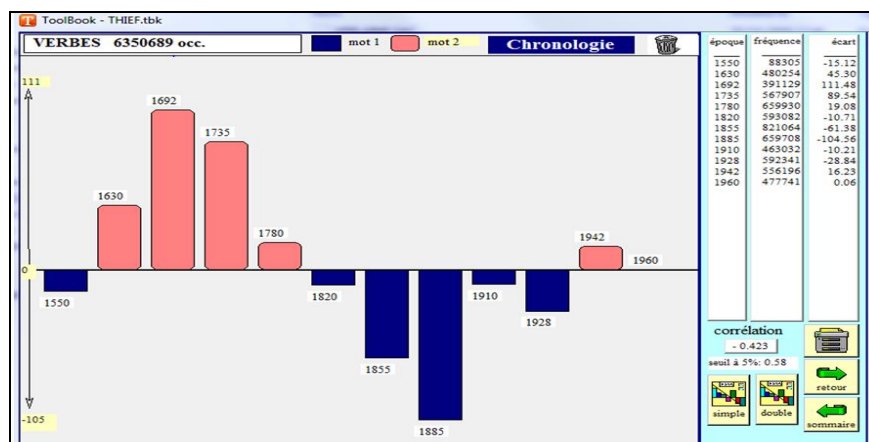


Figure 9. Le verbe dans *Frantext*.de 1500 à 1980 (échantillon de 6,3 millions d'occurrences)

b – La question se pose de constituer le corpus français de *Google* sinon comme norme, du moins comme toile de fond, chaque fois qu'on souhaite établir les spécificités d'un texte ou d'un corpus. Jusqu'ici ce rôle était tenu, faute de concurrent, par *Frantext*, avec l'avantage de choisir dans *Frantext* la période convenable. Quoique sa taille soit 200 fois plus petite, *Frantext* peut encore garder ses privilèges, dès qu'on a

affaire à des textes littéraires. Dans les autres cas le choix de *Google* paraît s'imposer. Normalement la comparaison se justifie mieux lorsque les deux termes de la comparaison appartiennent au même univers. Notre logiciel HYPERBASE a donc laissé le choix pour le français entre la norme littéraire de *Frantext*, et la norme plus technique de *Google*.

Mais le choix est moins crucial qu'on pourrait croire. L'expérience prouve que les **mêmes excédents** apparaissent, quel que soit le corpus de référence. Leur classement peut être différent mais rares sont les éléments d'une liste d'excédents qui ne sont pas dans l'autre. En revanche le vocabulaire « négatif » est très différent d'une norme à l'autre. Le phénomène peut s'expliquer par la **dissymétrie** des spécificités positives et négatives : les premières sont plus violentes et plus dispersées, les secondes plus molles et plus nombreuses. Car l'originalité est dans l'excès, non dans le manque ou l'indifférence. Ce qu'on appelle « vocabulaire négatif » n'est que l'image en miroir des excédents de la « norme ». Il caractérise la norme plus que le texte en question. On pourra vérifier cela dans l'exemple du corpus La Bruyère, comparé aux données de *Google*, puis de *Frantext* dans le tableau 10.

**Tableau 10. Comparaison de deux corpus de référence : *Google* (à gauche) et *Frantext* pour le calcul des spécificités positives et négatives d'un même corpus (La Bruyère)**

					Choix de calcul										
N°	écart	corpus	texte	mot			N°	écart	corpus	texte	mot				
156.97	32772410	2816	a		-63.85	1704498965	6258	.	423.18	383.941	i	-51.13	369404	1260	je
131.21	3907342	768	iv		-30.20	388992902	1445	du	391.65	298.768	iv	-48.50	881112	6258	.
105.02	38727	59	théophraste		-30.19	1133974201	6256	la	199.70	10228.2415		-36.84	286367	1587	vous
104.22	379543570	8828	il		-20.94	575894802	3215	des	189.14	10242.2296		-35.51	152967	410	me
82.43	195664144	4652	;		-18.56	1868653755	12789	de	185.12	85.194	vi	-32.09	95631	116	mon
73.47	75773145	2427	ils		-16.16	212559082	1045	au	114.01	209.189		-31.55	104273	198	ni
65.90	242179314	4865	qu'		-13.71	113246417	489	cette	97.77	69.92	vii	-30.85	120400	356	f'
59.63	323646405	5615	qui		-13.35	470374589	2975	en	80.75	1259.340	v	-27.55	71958	96	ma
57.03	71889	44	dévoit		-11.76	58926709	202	dés	66.05	44021.2136	ou	-23.85	62915	138	moi
57.01	455598	113	jusques		-11.66	853758041	5917	f'	53.79	389895.8826	il	-21.76	64481	210	ai
55.66	22902414	941	i		-11.14	58721027	231	avait	52.13	68434.2427	ils	-21.19	44117	68	mes
55.37	53798781	1587	vous		-10.98	24884943	45	pays	48.91	1802.258	soi	-20.05	117188	741	nous
52.06	32750023	1110	point		-10.68	69240287	306	était	44.71	566036.1127	et	-18.65	153899	1165	elle
51.75	448851086	11127	et		-10.68	35033706	103	non	41.82	250905.5615	qui	-17.20	48200	197	votre
50.32	25355300	924	homme		-10.63	23089718	41	droit	38.06	15696.742	hommes	-17.01	461014	4661	que
50.03	190259170	3599	ne		-10.35	764933288	5355	le	37.57	20416.871	ni	-16.99	105452	741	mais
50.03	23185844	871	ni		-9.51	21459680	48	trois	36.90	54275.1674	leur	-15.01	74276	489	cette
48.36	44505139	1294	?		-9.36	22231338	54	était	35.58	122.46	choix	-14.16	23536	59	roi
48.85	1691000	194	va		-9.27	128967323	741	nous	33.20	729.111	i	-14.12	31536	123	sus
48.58	146610	54	sot		-9.21	35065811	128	très	33.00	243814.4897	est	-11.80	24389	106	vos
47.80	2723128	245			-9.18	61430796	291	deux	32.84	16892.701	ceux	-10.92	21351	96	amour
47.53	78665978	1822	si		-9.18	19573895	43	politique	31.25	108145.2561	se	-10.90	20736	91	avez
46.68	2605811	235	fortune		-8.61	39522895	165	sous	30.87	786.108	d	-10.19	13388	40	ton
46.49	66573847	1614	y		-8.45	21637295	64	partie	30.53	15943.642	esprit	-10.12	28205	173	coeur
46.13	1977778	742	hommes		-8.15	16558819	50	lieu	30.10	142538.3128	on	-9.87	275069	2975	en
46.10	15565672	642	esprit		-7.48	536342235	3833	d'	28.98	677.94	caractères	-9.84	12756	42	puis
45.72	175043732	3128	on		-7.27	17373643	54	histoire	28.66	7607.381	grands	-9.30	19211	103	non
44.35	106702	42	courtsan		-7.25	16585021	50	travail	28.66	28870.924	homme	-9.29	75986	693	bien
43.04	3478120	256	quelquefois		-7.18	28981852	124	donc	28.62	16536.630	celui	-9.27	21451	124	donc
41.85	589536	96	vante		-7.17	16423878	50	chaque	28.00	174.44	dévoit	-9.03	23636	148	nos
40.30	156450064	2695	;		-7.13	87204828	514	ces	27.92	160185.3437	une	-8.03	27394	185	quand
40.12	6406214	340	v		-6.99	16061206	50	ordre	26.01	636.75	talents	-8.02	12295	46	moment
39.41	81957103	1674	leur		-6.90	17471586	59	roi	25.98	134419.2816	a	-8.92	28628	200	jamais
39.19	78983896	1640	lui		-6.90	47734705	250	aussi	24.90	3670.218	contraire	-8.87	13008	54	état
37.46	2516201	189	vi		-6.63	19890574	77	vers	24.39	344747.6080	les	-8.79	14728	70	père
37.23	368147841	4987	est		-6.91	13325152	42	puls	24.19	28384.828	leurs	-8.67	12006	48	trois
35.91	124048223	2136	ou		-6.27	15548348	54	société	23.85	891.92	maître	-8.52	12952	58	cependant
35.55	4739284	258	soi		-6.14	17134614	66	ans	23.71	9645.388	choses	-8.33	20965	135	notre
35.01	9416757	381	grands		-6.08	114844443	741	mais	23.58	94277.937	ont	-8.14	27373	202	été
34.40	25616076	701	ceux		-6.06	13564907	46	moment	23.05	47899.1190	sont	-8.10	56538	514	ces

La **confrontation directe** des deux normes est d'ailleurs faite dans la base GOOFRE. En se servant de *Frantext* comme toile de fond, on fait apparaître les spécificités de *Google* dans la liste positive et celles de *Frantext* dans la liste négative (tableau 11). La première est riche en anglicismes et en termes abstraits, scientifiques, économiques, politiques. Elle ne contient que deux verbes (SONT, PERMET), et quelques outils grammaticaux qui accompagnent volontiers l'abstraction : la préposition DE et l'article défini (surtout LA). La seconde, à l'opposé, est tournée vers le concret et les réalités quotidiennes de la maison, de la famille, de la vie. C'est un univers où les gens se rencontrent, se parlent, se touchent et s'aiment. L'abondance des pronoms de dialogue est telle qu'on pourrait croire avoir affaire à un corpus oral.

**Tableau 11. Les spécificités de la base *Google* et de la base *Frantext***

<i>Google</i>	<i>Frantext</i>
OF, THE, AND, IN, ECONOMIQUE, TO, DER, PRODUCTION, DES, NOTAMMENT, DU, SYSTEME, LA, DE, PAR, PRODUITS, POLITIQUE, LES, POPULATION, ANALYSE, SOCIALE, NIVEAU, LOI, FONCTION, CAS, PERIODE, ENSEIGNEMENT, RAPPORT, ORGANISATION, CONDITIONS, OU, GOUVERNEMENT, NATIONALE, COMMISSION, SONT, ELEMENTS, CULTURE, EGALEMENT, ADMINISTRATION, SOCIETE, ORIGINE, PRINCIPE, DROIT, REGION, FRANCE, THE, PAYS, MEMBRES, GENERALE, ARTICLE, PARTIE, ETUDE, ACTIVITE, DIFFERENTS, VALEUR, NOMBREUX, PRATIQUE, FONCTIONS, METHODE, RAPPORTS, FRANÇAISE, INDUSTRIE, EPOQUE, UNITE, ACTION, RELATIONS, TRAVAUX, PERMET, TRAVAIL, CERTAINS, PROPRIETE, ART, SUD, AN, GROUPE, RECHERCHE, PARTIES, ENTRE, PRODUIT, TERME, ETUDES, etc.	JE, VOUS, ME, SI, IL, MON, M', TU, MA, AI, NE, QUE, QUAND, QU', MAIS, DIT, PAS, TOUT, MES, SUIS, Y, VOTRE, CE, TE, LA, N', UN, SA, YEUX, SON, CELA, AVEZ, JAMAIS, AMOUR, AVAIT, VOS, FEMME, FILLE, QUELQUE, VOIX, MAIN, OU, SANS, COMME, PERE, MERE, JEUNE, TROP, PETIT, CHOSE, SES, QUOI, POINT, SAINT, ENFIN, EUT, VOIR, MAINS, NOUS, S', PEU, PORTE, FAIRE, PETITE, DIEU, TANT, CEPENDANT, HEURES, QUEL, COMMENT, QUI, JOURS, MILLE, GENS, PARLER, MAISON, PEINE, CAR, AME, DEVANT, TOUS, POURQUOI, MIEUX, AVOIR, MAINTENANT, FAISAIT, ENFANT, FOND, ET, MORT, CHAMBRE, DIRE, VIE, ILS, CHEZ, QUELLE, CHOSES, MONDE, BAS, ASSEZ, PERSONNE, GRAND, SEUL, VU, FEU, SANG, EUT, FOIS, PLUS, FAIT, DESSUS, FILS, etc.

c – Si la comparaison externe est toujours peu ou prou sujette à caution, il n'en va pas de même pour la **comparaison interne**. Un corpus normalement constitué peut légitimement servir de référence pour ses propres parties et jusqu'à ce jour la lexicométrie a reposé sur ce postulat. Choisissons la douzième et dernière période de notre base GOOFRE, celle qui va de 2005 à 2008 (tableau 12). On y reconnaît les

préoccupations du moment, l'attention portée à l'environnement (ENVIRONNEMENT, IMPACT, RISQUE, ESPACE), aux questions sociales et politiques (POLITIQUES, POLITIQUE, SOCIAL, SOCIALE, EUROPEENNE, FEMMES), aux canaux de l'information (INFORMATION, INFORMATIONS, SITE, REFERENCE, RECHERCHE, EDITIONS, FILM, RESEAUX) et surtout aux méthodes scientifiques (METHODE, PROCESSUS, APPROCHE, RELATION, FONCTION, STRATEGIE, MODELE, OBJECTIF, COMPETENCES, UNIVERSITE, EQUIPE, EVALUATION).

**Tableau 12. Spécificités de la période 2005-2008 dans Google**

<i>écart</i>	<i>corpus</i>	<i>tranche</i>	<i>forme</i>	<i>écart</i>	<i>corpus</i>	<i>tranche</i>	<i>forme</i>
1360.69	8179891	1464686	]	527.80	17569507	1649825	of
1336.79	15828400	2295772	/	526.93	4576484	563862	image
1325.12	8399794	1466152	[	520.77	984213	188013	stratégie
873.95	198979236	15586704	(	519.68	3498102	457856	modèle
844.80	156450064	12467961	:	519.63	2265989	333731	européenne
839.15	198105625	15405625	)	519.56	343412708	24065306	une
807.33	3439552	582021	processus	517.32	2020132	306756	approche
798.22	1844310	380525	contexte	513.89	2035855	307255	québec
773.91	10491185	1273764	%	513.37	562935	129366	impact
771.23	1339362	301980	acteurs	512.77	5416317	633080	espace
768.56	422179209	30547729	-	510.08	7190679	787771	partir
765.91	2079899	400432	gestion	506.70	6530440	728253	sociale
737.98	1267900	282471	environnement	504.17	24666152	2169697	paris
727.60	10852189	1269876	and	499.06	1015842	186691	press
698.13	19148108	1954774	the	498.85	861727	167227	éthique
662.56	698137	178914	compétences	498.27	5613564	642435	fonction
656.10	3851404	557030	cadre	498.24	1008615	185598	film
642.54	5342704	699455	recherche	497.43	1492702	242352	dimension
633.07	526985	145200	patients	497.18	5647444	644816	social
619.81	839902	191394	site	494.58	552210	124390	enseignants
616.17	26640774	2459236	p	486.77	12349436	1197520	années
615.36	2379975	381620	risque	484.34	1721111	263535	objectif
608.06	785238	180826	patient	484.28	727671	146584	gallimard
604.92	6231818	761758	mise	483.53	1205197	205444	intégration
602.88	5396632	682259	permet	478.17	2124817	304064	statut
594.97	1626446	287577	référence	473.71	3713158	457059	relation
594.02	5504585	687420	notamment	473.48	2528192	343171	al
592.44	2696106	407323	ça	470.19	760101	147862	presses
584.96	2482686	381393	activités	467.86	8336592	856129	tu
580.17	2351234	365268	information	464.47	828865	155360	outils
577.98	19673865	1868454	politique	463.24	8114710	834485	femmes
577.38	11851337	1233456	selon	460.67	792102	149902	réseaux
567.83	2187780	342825	identité	460.66	7883579	813498	lors
565.22	534528	134399	stratégies	460.33	47913858	3806283	entre
562.75	1094197	212500	évaluation	457.85	860890	157860	équipe
562.69	6722878	780374	politiques	456.62	3987610	474177	université
557.53	2426130	364845	pratiques	454.82	9428810	936347	jean
556.54	1816644	297500	voire	453.85	2120931	295041	accès
548.87	1563804	265988	informations	453.72	1163500	192724	éditions
541.65	6305879	729950	niveau	452.78	946376	167079	dossier

Le portrait-robot que dessine cette liste est très proche de celui que représentait le tableau 10, dévolu à l'ensemble du corpus, en sorte qu'on



peut voir dans la dernière période tout à la fois le résumé et la tendance du corpus entier, avec les mêmes caractéristiques : surabondance des abstractions, des analyses scientifiques, sociales, économiques ou politiques, prédominance de la catégorie nominale, anglicismes, ponctuation démonstrative.

## Conclusion

D'autres méthodes, et le notamment le coefficient de corrélation chronologique calculé pour tous les mots du corpus, confirment cette tendance à l'abstraction. Mais la place nous manque pour détailler tous les enseignements – le plus souvent des confirmations – qu'on peut tirer de la base GOOFRE et qu'illustre la figure 13 : sur l'allongement du mot, mais aussi le raccourcissement de la phrase ; sur la simplification de la ponctuation où le point tend à se substituer aux autres signes ; sur la spécialisation et la diversification du vocabulaire français ; sur l'évolution de l'orthographe et la disparition progressive du circonflexe, etc. Nous ne voulons insister que sur un point : la linéarité régulière de l'évolution. Quand on voit l'histoire de si haut, les remous et soubresauts se fondent dans le courant qui va en s'accroissant de 1800 à nos jours. Rares sont les retours, les repentirs, les hésitations ou les ruptures. Quel qu'en soit l'objet, les courbes, les analyses factorielles ou arborées ont une lisibilité radicale qu'on rencontre rarement à moins grande échelle.

Quelles leçons tirer du gigantisme ?

1. Quelle que soit la taille d'un corpus, la langue reste inaccessible. « Il n'y a pas de probabilité en langue », disait Maurice Tournier.
2. Les gros corpus ne sont pas nécessairement homogènes. La chronologie peut changer leur contenu.
3. La loi des grands nombres n'efface pas complètement les fautes de conception, les bévues des machines et les erreurs de traitement. Un corpus restreint mais plus pur vaut mieux qu'un terril énorme, mais dégradé.
4. Le gigantisme met en cause l'échelle des mesures et des outils. En particulier la loi hypergéométrique, inapplicable, cède ses droits à la loi normale. C'est la revanche de l'écart réduit.
5. La base *Google* gagne en tant que corpus ce qu'elle perd en tant que norme. Il y a des tas de choses à découvrir dans ce tas de mots. Concernant les faits langagiers. Et concernant les faits de civilisation et d'histoire.

## Références

- Erez Lieberman, J.B. Michel, Joe Jackson, Tina Tang & Martin A. Nowak, « Quantifying the evolutionary dynamics of language », *Nature*, Nature publishing Group, 2007, p. 713-716.
- J.B. Michel & al., « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, déc. 2010, [http://www.sciencemag.org/content/331/6014/176.full.html\\*related](http://www.sciencemag.org/content/331/6014/176.full.html*related).
- Étienne Brunet, *Le vocabulaire français de 1789 à nos jours*, 3 vol., Slatkine-Champion, Genève-Paris, 1981 (1981a).
- Étienne Brunet, « Ce que disent les chiffres », in Chaurand (dir.), *Nouvelle Histoire de la langue française*, Seuil, Paris, 1999, p. 673-727 (1999a).

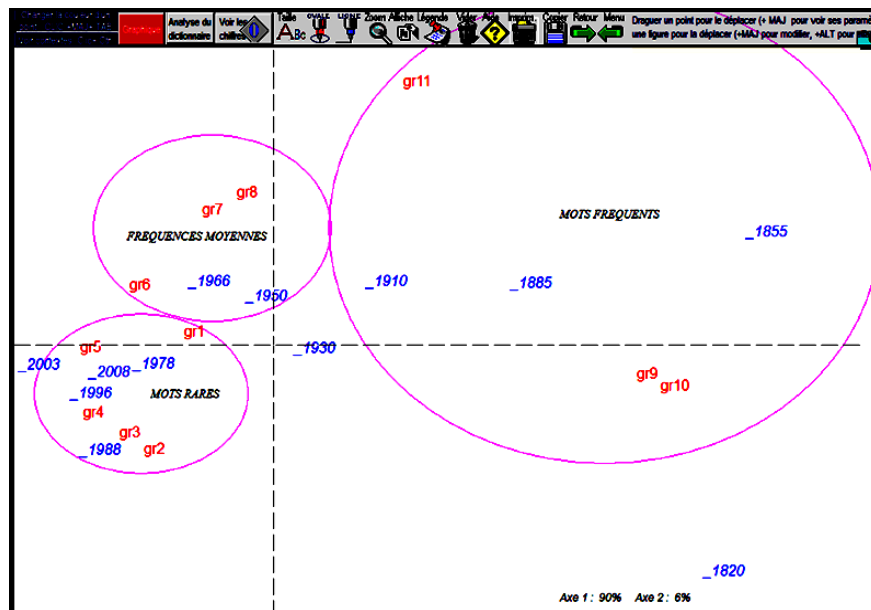


Figure 13a. L'inflation lexicale : Le vocabulaire français se spécialise et se diversifie

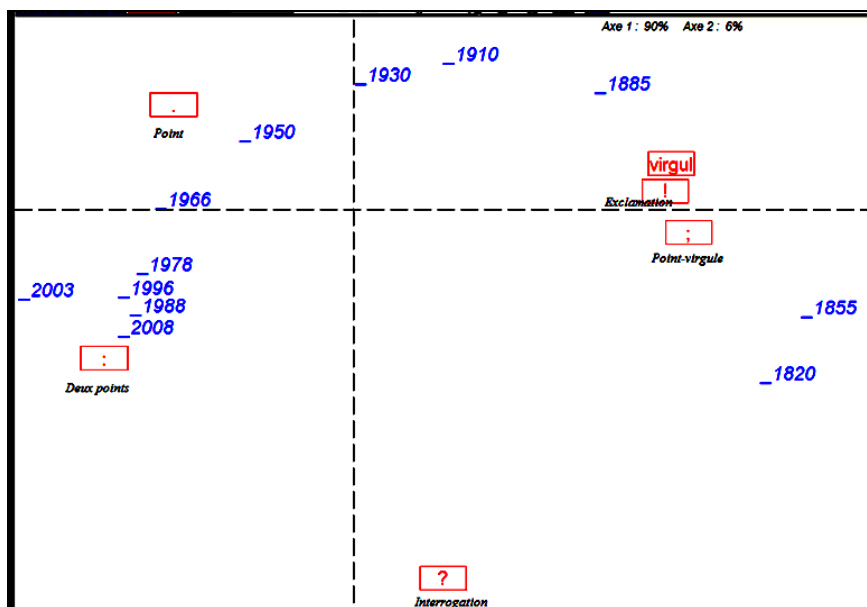


Figure 13b. La ponctuation se radicalise (point 1,7 milliard et deux-points 156 millions). La phrase se raccourcit

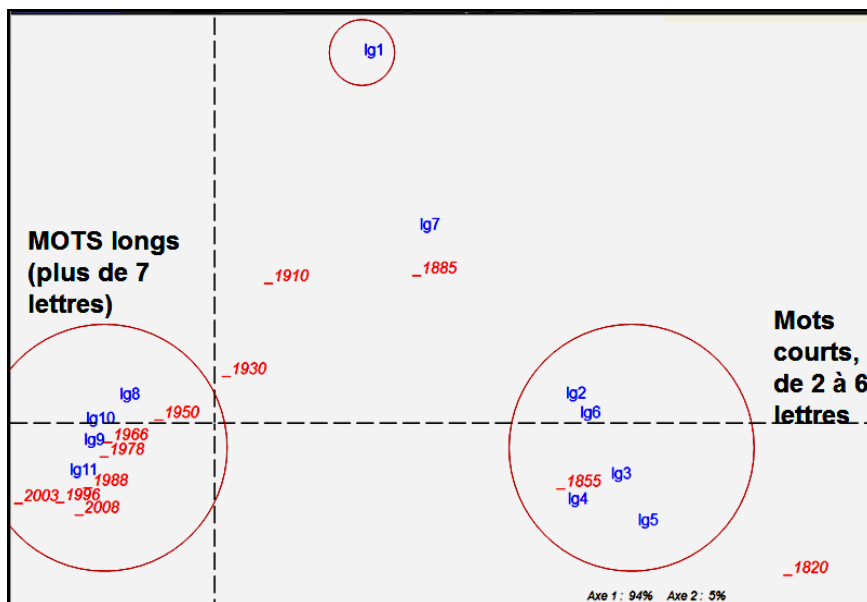


Figure 13c. Les mots français s'allongent

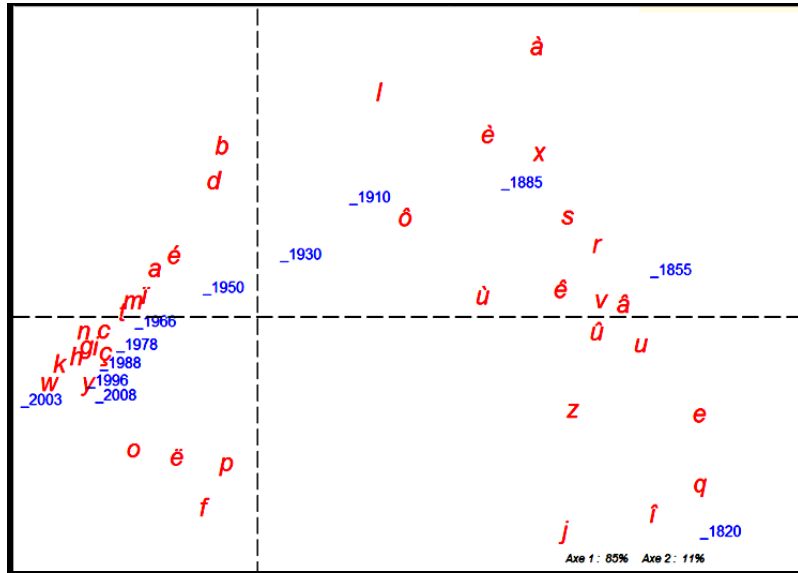


Figure 13d. 169 milliards de lettres. Le déclin du circonflexe