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

Étienne Brunet

BCL (CNRS), Université de Nice, brunet@unice.fr

**Abstract**

In December 2010, an article appeared in *Science* that reported on a Google enterprise of Pharaonic proportions: the identification, digitalization and indexing of "all" printed texts in the world since the beginning of printing. Only for the French part, the figures are incredible: 500.000 full books, a million different words, and 44 billion tokens. For every word in the French data base, and for every lexical configuration between 1 and 5 words, the chronological curve can be obtained on the Internet.

One may critique the result of this huge achievement without trying to hide its weaknesses: a great number of spelling errors and human mistakes justify the implementation of a more relevant and efficient tool to treat the same data, to make it available to lexical statistical tools and to multidimensional analysis.

A new database, named *Goofre*, which has been created without losing any of the 44 billion tokens of the original corpus, offers an independent and synthetic exploration of data.

**Résumé**

En décembre 2010, un article a paru dans *Science*, qui rendait compte d'une entreprise pharaonique de Google : le recensement, la saisie et l'indexation des textes imprimés dans le monde depuis l'origine de l'imprimerie. En s'en tenant au domaine français les chiffres brutalisent l'imagination: 500 000 volumes complets, 1 million de mots différents, 44 milliards d'occurrences. Pour chacun des mots de la base française, et chacune des configurations lexicales de 1 à 5 mots, la courbe chronologique peut être obtenue sur Internet.

On fera la critique de cette réalisation sans cacher ses points faibles : beaucoup d'erreurs de lecture, mais aussi des défauts de conception et de traitement ont paru justifier une mise en œuvre plus pertinente des mêmes données, afin de les ouvrir aux outils de la lexicométrie et particulièrement aux analyses multidimensionnelles.

Une nouvelle base, du nom de *Goofre*, a été créée qui sans rien perdre des 44 milliards de mots du corpus original, en offre une exploitation autonome et synthétique.

**Mots-clés :** Google Books, Culturomics, langue française, lexicométrie, Frantext, logiciel

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: 500000 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 clausule 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[1].

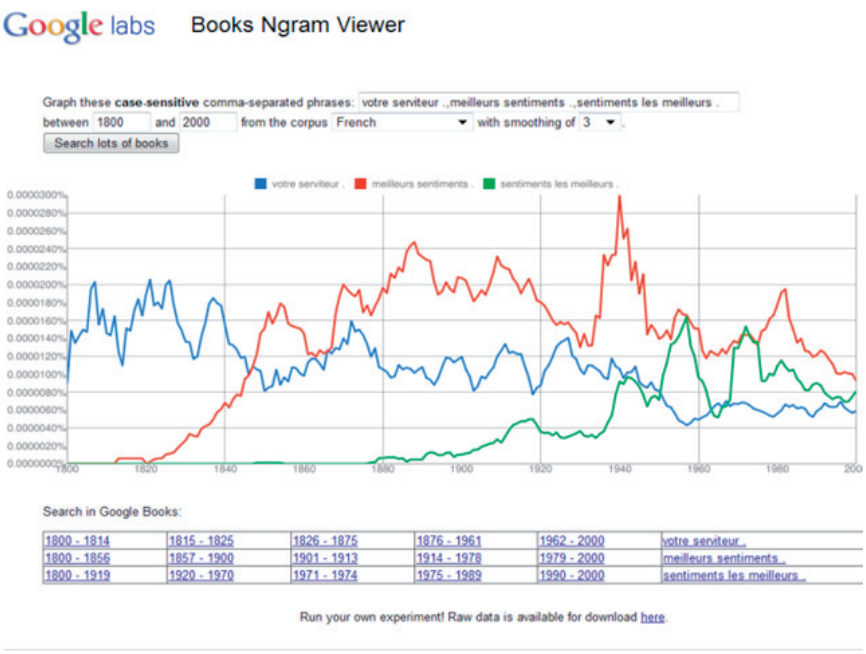



*Figure 1. Exemple d'interrogation de la base française du site Google Books*

**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 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

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.

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'inter-lettrage (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. 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.

| 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 | eſt |
| 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 | auroit |
| 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 |

| N° | écart | corpus | texte | mot |
|---|---|---|---|---|
| 1 | -2849.72 | 198979236 | 7623121 | ( |
| 1 | -2657.55 | 198105625 | 8369449 | ) |
| 1 | -2393.80 | 1704496965 | 138467818 | . |
| 1 | -1634.89 | 46880704 | 1282552 | # |
| 1 | -1326.75 | 26640774 | 583770 | p |
| 1 | -1228.76 | 17569507 | 195984 | of |
| 1 | -1124.06 | 388992902 | 31715169 | du |
| 1 | -1009.83 | 10491185 | 58803 | % |
| 1 | -986.65 | 1133874201 | 101803646 | la |
| 1 | -973.99 | 10852189 | 112916 | and |
| 1 | -948.58 | 853756041 | 75848608 | l' |
| 1 | -917.35 | 19673865 | 725601 | politique |
| 1 | -869.13 | 156450064 | 12173121 | : |
| 1 | -834.78 | 10993735 | 258202 | oeuvre |
| 1 | -828.98 | 6757103 | 23503 | économique |
| 1 | -773.91 | 368147941 | 31843216 | est |
| 1 | -769.33 | 15828400 | 647200 | / |
| 1 | -768.14 | 8399794 | 164064 | [ |
| 1 | -765.79 | 22140750 | 1107382 | in |
| 1 | -761.55 | 6368222 | 54658 | parfois |
| 1 | -757.64 | 8612869 | 185878 | façon |
| 1 | -741.65 | 8179891 | 173725 | ] |
| 1 | -730.88 | 575694802 | 51488995 | des |
| 1 | -730.26 | 6064364 | 61512 | groupe |
| 1 | -729.89 | 9193134 | 246160 | enfants |
| 1 | -729.13 | 6161350 | 67635 | éléments |
| 1 | -721.69 | 5951336 | 61629 | problème |

*Tableau 2. Les spécificités de la première période (1800-1840)*
*Excédents à gauche déficits à droite*

**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. 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 bévue : 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 s'ils restent boiteux ou monstrueux. Ainsi on peut obtenir la courbe de *tojours, toujurs* ou de *cpendant, cepndant, cependont* . 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 incommodités 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

| Tranche | Année moyenne | Taille (en milliards de mots) |
|---|---|---|
| 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.

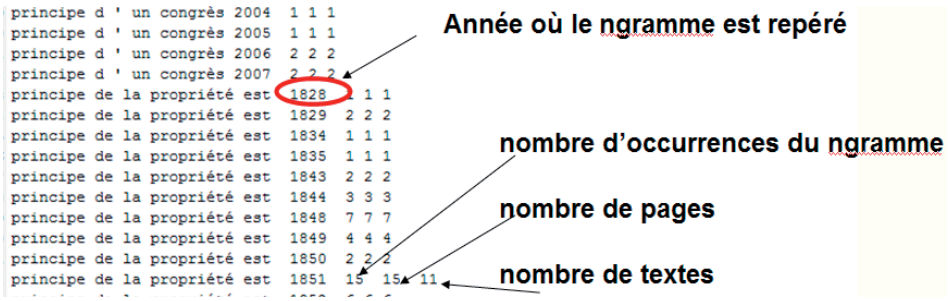


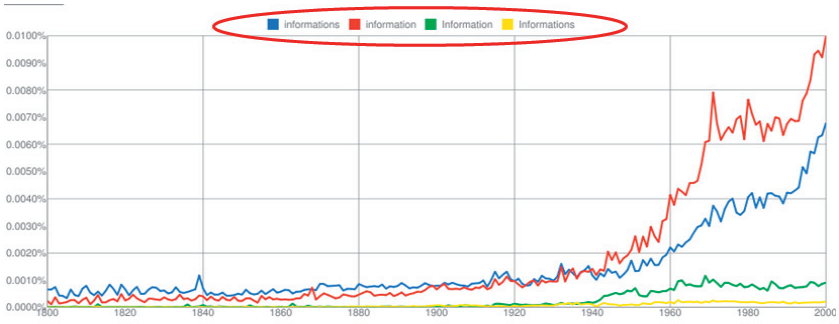

*Tableau 3. Extrait des données brutes obtenues de Google Books (ngrammes de longueur 5) (on observera dans la première ligne le traitement inapproprié de l'apostrophe, considérée comme un mot et non comme le résultat de l'élision)*

```
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
```

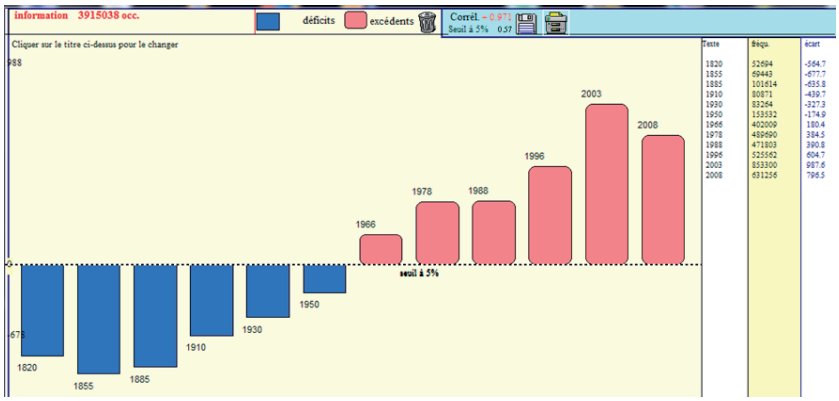
*Tableau 4. Extrait des données traitées par GOOFRE (unigrammes) (les trois premiers nombres cumulent ceux de la base pour l'ensemble du corpus, et les 12 derniers détaillent la répartition des occurrences dans la chronologie)*

**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. 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 boursicoteur ne risque guère de faire fortune.




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




*Graphique 6. L'histogramme de l'information dans Goofre (toutes formes réunies)*

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 *être* et près d'un milliard d'occurrences, et même de confronter le verbe *avoir* au verbe *être*, 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[2] (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.

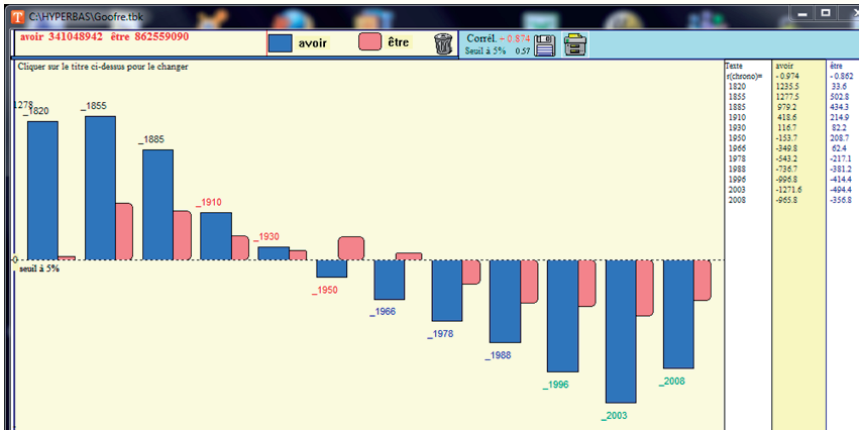



*Figure 7. Être et avoir dans la base Goofre (plus d'un milliard d'occurrences)*

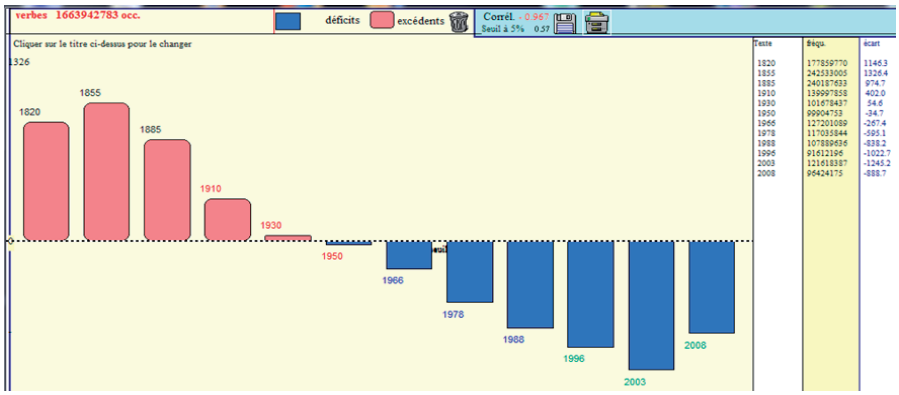



*Figure 8. Pouvoir, savoir, vouloir, falloir, aller, faire dans la base Goofre (1,7 milliard )*

## 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

2 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 *marché* dans la catégorie du verbe ou du substantif. Cet embarras obère l'exploitation de la base *Goofre*, mais aussi celle de *Thief*, dont les données sont antérieures à la lemmatisation de *Frantext*.

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**.

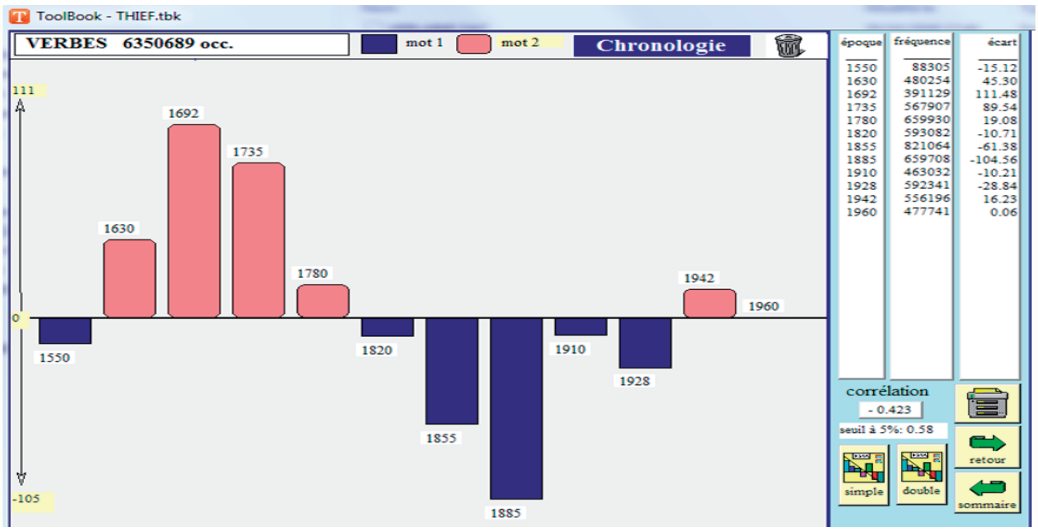



*Figure 9. Le verbe dans Frantext.de 1500 à 1980 (échantillon de 6,3 millions d'occurrences)*

**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 11.

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. 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.

| Google | Frantex |
|---|---|
| of, the, and, in, économique, to, der, production, des, notamment, du, système, la, de, par, produits, politique, les, population, analyse, sociale, niveau, loi, fonction, cas, période, enseignement, rapport, organisation, conditions, ou, gouvernement, nationale, commission, sont, éléments, culture, également, administration, société, origine, principe, droit, région, France, thé, pays, membres, générale, article, partie, étude, activité, différents, valeur, nombreux, pratique, fonctions, méthode, rapports, française, industrie, époque, unité, action, relations, travaux, permet, travail, certains, propriété, art, sud, an, groupe, recherche, parties, entre, produit, terme, études, 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, là, n', un, sa, yeux, son, cela, avez, jamais, amour, avait, vos, femme, fille, quelque, voix, main, où, sans, comme, père, mère, jeune, trop, petit, chose, ses, quoi, point, saint, enfin, eût, voir, mains, nous, s', peu, porte, faire, petite, dieu, tant, cependant, heures, quel, comment, qui, jours, mille, gens, parler, maison, peine, car, âme, 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. |

*Tableau 10.Les spécificités de la base Google…et de la base Frantext*

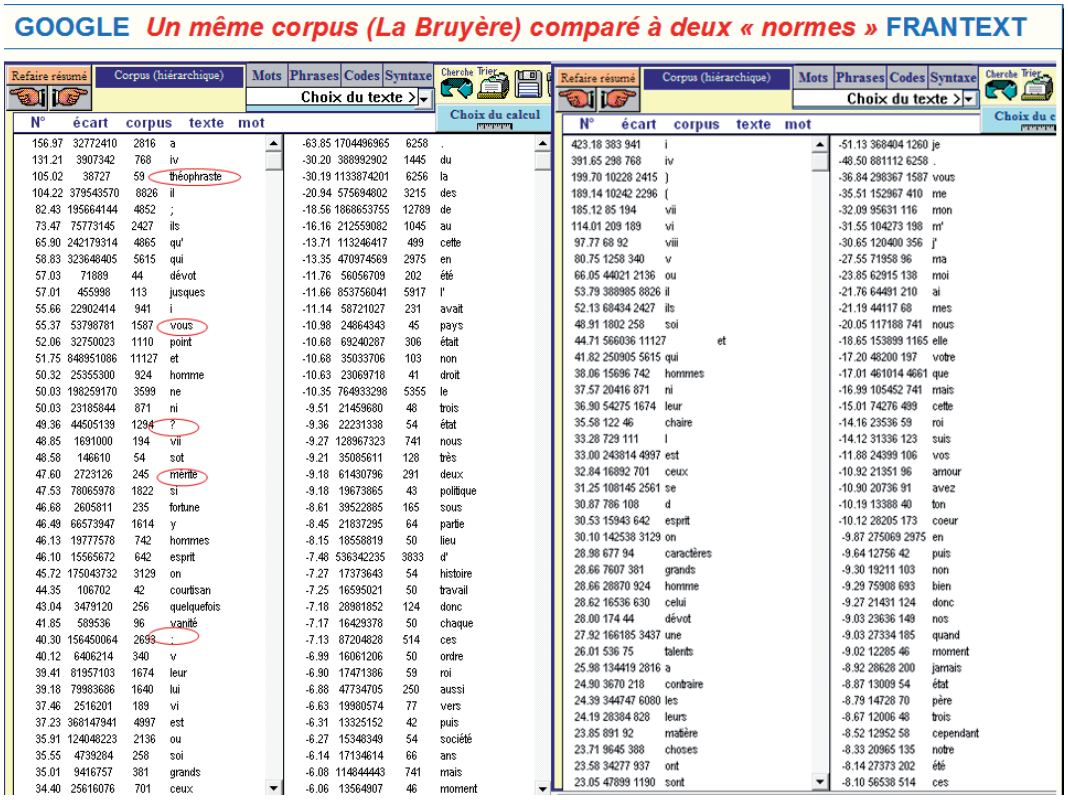



*Tableau 11. Comparaison de deux corpus de référence : Google et Frantext pour le calcul des spécificités positives et négatives d'un même corpus (La Bruyère)*

**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, européenne, femmes*), aux canaux de l'information (*information, informations, site, référence, recherche, éditions, film, réseaux*) et surtout aux méthodes scientifiques (*méthode, processus, approche, relation, fonction, stratégie, modèle, objectif, compétences, université, équipe, évaluation*). 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.

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

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

## 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'accentuant 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 ?

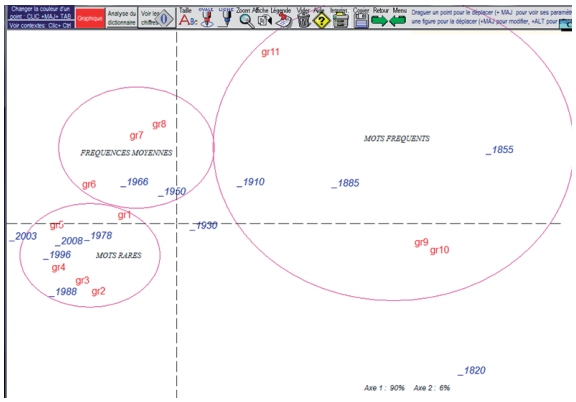
1 - Quelle que soit la taille d'un corpus, la langue reste inaccessible. « Il n'y a pas de probabilité en langue », disait Maurice Tournier.

2 - Les gros corpus ne sont pas nécessairement homogènes. La chronologie peut changer leur contenu.

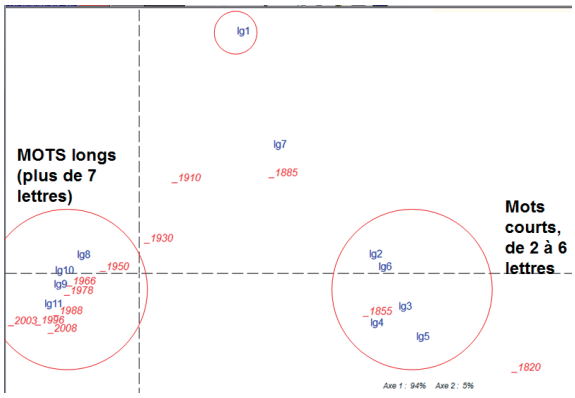
3 - La loi des grands nombres n'efface pas complètement les fautes de conception, les bévues des machines et les erreurs de traitement. Un corpus restreint mais plus pur vaut mieux qu'un terril énorme, mais dégradé.

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.

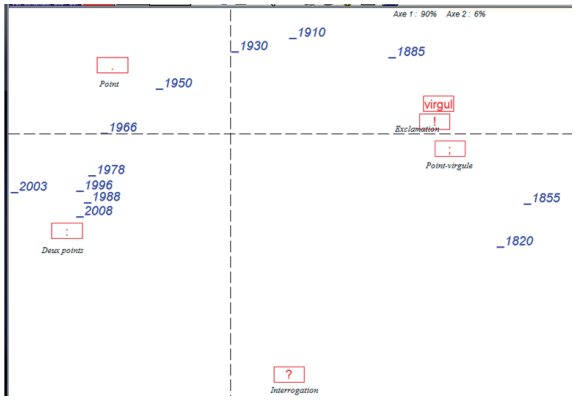
5 - La base *Google* gagne en tant que corpus ce qu'elle perd en tant que norme. Il y a des tas de choses à découvrir dans ce tas de mots. Concernant les faits langagiers. Et concernant les faits de civilisation et d'histoire.



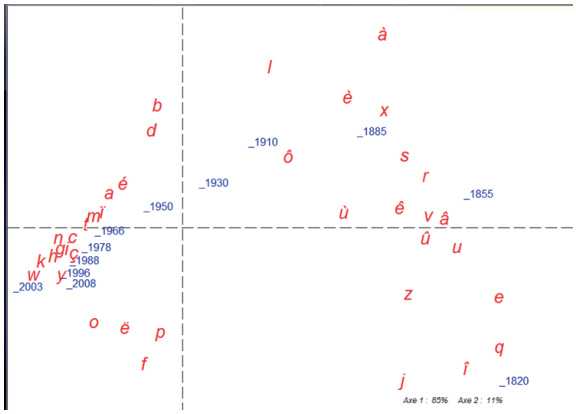

*L'inflation lexicale: Le vocabulaire français se spécialise et se diversifie*




*Les mots français s'allongent*




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




*169 milliards de lettres. Le déclin du circonflexe*

*Figure 13. Quelques analyses factorielles*

## 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.

Étienne Brunet, *Le vocabulaire français de 1789 à nos jours*, 3 vol., Slatkine-Champion, Genève-Paris, 1981.

Étienne Brunet, Ce que disent les chiffres, in Chaurand (sous la direction de) *Nouvelle Histoire de la langue française*, Seuil, Paris, 1999, p. 673-727.