

# GOOFRE version 2

Etienne Brunet, Laurent Vanni

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

## Abstract

The amount of data contained within Google Books has doubled over the last two years and now exceeds 500 billion words. A new treatment of the data has included a re-examination of scanned images, offering a more accurate recognition of the text. In addition, for the first time, included texts have been subjected to deambiguation and lemmatisation. Finally, the website Culturomics has made tools available that facilitate its accessibility. It seemed interesting, therefore, to develop a new expertise and to create a new database, complete with all the necessary statistical tools, available online or locally, for exploiting such large corpora.

## Résumé

Les données de Google Books ont doublé en deux ans, en franchissant le cap des 500 milliards de mots. Un nouveau traitement a repris les images scannées pour en proposer une lecture plus fidèle. Et pour la première fois les textes enregistrés ont bénéficié de la désambiguïsation et de la lemmatisation. Enfin le site Culturomics a fourni les outils nécessaires pour en assurer la diffusion. Il convenait donc de procéder à une nouvelle expertise et de créer une nouvelle base, pourvue de tout l'appareillage statistique qu'exige, en réseau ou en local, l'exploitation des grands corpus.

**Mots-clés :** Google Books, Culturomics, Big Data, expertise et exploitation statistique, croisement du temps et de l'espace

Avec une périodicité de deux ans, les JADT ne peuvent suivre d'assez près l'actualité. De même que les Jeux Olympiques ont doublé leur rythme en proposant les Jeux d'hiver, on pourrait imaginer pareillement des JA d'Hiver en alternance avec les JADT. Et pourtant en obligeant les auteurs – et plus cruellement encore les organisateurs – à publier les Actes le jour même de l'ouverture des débats, les statuts de l'entreprise ont assuré à nos publications une réactivité rare, que n'ont pas certains Congrès traditionnels dont les Actes suivent, des années après, le passage de la comète, quand tête et queue ont disparu, et presque le souvenir de l'événement<sup>1</sup>. On me dira peut-être que le destin du monde n'est pas suspendu à nos observations et qu'en matière littéraire ou linguistique, tout au moins, le rythme des révolutions n'a rien de pressant. Aucun séisme ne semble menaçant, sinon quelque manifeste annonçant un nouveau « nouveau roman » ou une réformette de l'orthographe dont l'explosion attendue fera le bruit d'un pet de nonne.

Cependant l'entreprise de *Google Books* dont nous avons rendu compte lors des JADT de 2012 a eu l'impact d'un ébranlement planétaire. Faire main basse à l'aide de scanners sur les livres du monde entier, ne pouvait pas ne pas déclencher une réaction de défense économique, culturelle et presque religieuse. Et chacun de défendre ses eaux territoriales contre les filets de l'envahisseur. Des projets nationaux ou européens ont été bâtis pour relever le défi, sans empêcher la croissance de la Tour de Babel (que certains appellent la Tour de Babil). Or en

---

<sup>1</sup> Ainsi ai-je dû attendre sept ans la parution d'une communication, où j'avais introduit imprudemment le mot « nouveau » dans le titre.

deux ans la hauteur a doublé, pour le français comme pour les autres langues<sup>2</sup>. En soi cette taille devenue plus épaisse pourrait ne pas changer les profils ni rendre caduques les analyses antérieures. Mais d'une part le gonflement des données n'a pas été homogène : certaines périodes primitivement dépourvues ont pris de l'embonpoint, et le déséquilibre entre les premières et les dernières périodes a été partiellement corrigé<sup>3</sup>. D'autre part les textes déjà disponibles dans la version de 2009 ont été repris à la base et soumis à des traitements améliorés ou radicalement nouveaux, si bien que le corpus de 2012 n'est superposable au précédent ni dans sa composition, ni dans son exploitation. Et l'analyse de ce projet gigantesque doit être reprise, sans a priori. Pour ceux qui abordent cette question pour la première fois, il convient d'expliquer que les millions de livres dépouillés sont repérables et consultables sur le réseau, à travers la grille de *Google Books* qui filtre plus ou moins les données, selon les contraintes du *copyright*. Or cette démarche classique dont l'entrée est un mot et la sortie un contexte peut être accompagnée d'une interrogation portant sur les fréquences. L'entrée là encore est un mot (ou un ensemble de mots) mais la sortie délivre un graphe reproduisant l'évolution des fréquences du mot considéré entre deux dates. Un site particulier, *Culturomics*, est dédié à cette exploitation statistique du corpus<sup>4</sup>. Consultons-le avec le couple de substantifs *homme-femme* qui vient souvent en tête dans les listes de fréquence et que les auteurs de *Culturomics* ont mis en avant, en tant que test de validité (le moins qu'on puisse attendre de la statistique est qu'elle confirme l'évidence).

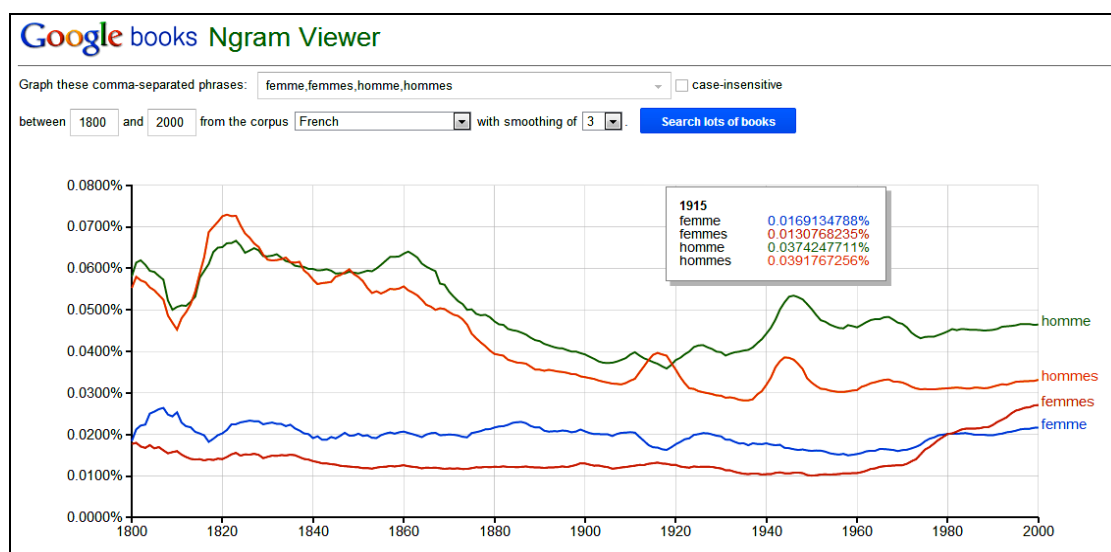


Figure 1. Interrogation de la base *Culturomics*. Le couple homme/femme de 1800 à 2000<sup>5</sup>

<sup>2</sup> On en est à 89 milliards pour le français, 349 pour l'anglais (où plusieurs variétés peuvent être isolées), 53 pour l'allemand, 67 pour l'espagnol, et 33 pour l'italien, nouveau venu. Ces chiffres correspondent aux données téléchargeables. Ils sont supérieurs dans la table 1 de l'article publié dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, p.170. Trois autres sont disponibles, dont nous ne dirons rien faute de connaissances et de clavier : le russe, le chinois et l'hébreu.

<sup>3</sup> La correction a été plus nette pour le français que pour les autres langues.

<sup>4</sup> Le parallélisme entre les données de Google Books et les corpus de *Culturomics* diverge au fil du temps. Les premières sont évolutives et s'enrichissent chaque jour. Les seconds sont statiques et représentent un état figé des premières, à une date donnée. La version 2012 est ainsi une mise à jour de la version 2009. D'autres mises à jour sont prévues dans les années à venir.

<sup>5</sup> Une fois complétées les zones de remplissage (les mots recherchés, les dates de début et de fin, le corpus choisi, et l'option de lissage), la chaîne envoyée sur le réseau est ici :

On peut s'arrêter avec le curseur à n'importe quelle année pour interpréter les à-coups de l'histoire et constater par exemple le sursaut des *hommes* en temps de guerre. L'année 1915, isolée sur le graphique, montre le détail des relevés (il s'agit de la part, en pourcentage, que prennent cette année-là les mots considérés, dans le vocabulaire de l'année). Mais l'enseignement majeur est l'orientation des courbes. L'écart entre *homme* et *femme* qui était de 1 à 3 au début de la chronologie se réduit d'année en année, au point que les *femmes* ont quasiment rejoint les *hommes* en 2000. Il s'agit là d'un gain dans l'expression et la communication, sans doute aussi dans la conscience. En réalité la présence grandissante du féminisme dans le discours n'est pas la preuve, mais tout au plus l'annonce espérée de l'égalité des sexes. On peut parcourir sur le même sujet les autres corpus et voir si les mêmes tendances s'y observent, ce qu'on tentera dans la suite de cet exposé.

## 1. Améliorations et lemmatisation

On pourrait croire que l'interrogation de *Culturomics* n'a pas varié depuis la première version et que seule l'assise des données s'est élargie. En réalité il existe un mode d'emploi dit avancé<sup>6</sup> qui change radicalement le point de vue et permet d'accéder non plus seulement aux formes brutes mais à des formes annotées et pourvues d'un code grammatical.

**1.1.** Contrairement aux corpus de l'édition 2009, les données de l'édition 2012 ont bénéficié de la lemmatisation. La séparation des homographes<sup>7</sup> est mise en œuvre et l'on obtient deux séries séparées pour la forme "le" article et la forme "le" pronom personnel. Il suffit d'ajouter la catégorie au mot recherché, par exemple *le\_DET* ou *le\_PRON*. Une dizaine de suffixes est disponible pour désigner les codes grammaticaux. Grosso modo ils se retrouvent inchangés d'une langue à l'autre. Noter que ces codes peuvent être soumis directement à l'interrogation, soit seuls, soit en combinaison avec d'autres codes ou des mots individuels. On peut ainsi dresser la courbe récapitulative de tous les verbes cumulés (*VERB*), ou de la construction préposition+déterminant+ substantif (*ADP\_DET\_NOUN*) ou de tous les noms qualifiés de "vieux" (*vieux\_ADJ\_NOUN + NOUN\_vieux\_ADJ*). On admirera la puissance du programme dans la figure 2 qui cumule plus de 300 milliards de mots du corpus anglais et répartit les parties du discours dans le temps, de 1800 à 2000. Jamais jusqu'ici la progression du nom et de l'adjectif et le déclin variable de toutes les autres catégories n'avait été observés à une telle échelle<sup>8</sup>.

---

[https://books.google.com/ngrams/interactive\\_chart?content=femme%2Cfemmes%2Chomme%2Chommes&year\\_start=1800&year\\_end=2000&corpus=19&smoothing=3](https://books.google.com/ngrams/interactive_chart?content=femme%2Cfemmes%2Chomme%2Chommes&year_start=1800&year_end=2000&corpus=19&smoothing=3)

<sup>6</sup> Les multiples façons d'interroger la base *Culturomics* sont détaillées à l'adresse <http://books.google.com/ngrams/info>

<sup>7</sup> La lemmatisation a deux phases: la première analyse les mots et distingue les homographes; la seconde regroupe les formes qui appartiennent au même lemme. Cette dernière fonction est assurée par le suffixe *\_INF*, attaché à l'entrée du dictionnaire, par exemple *aller\_INF*.

<sup>8</sup> La pente paraît faible et presque imperceptible pour certaines catégories, parce que les écarts sont traduits en pourcentages. Mais mesurés en termes probabilistes les variations sont très considérables et très significatives.

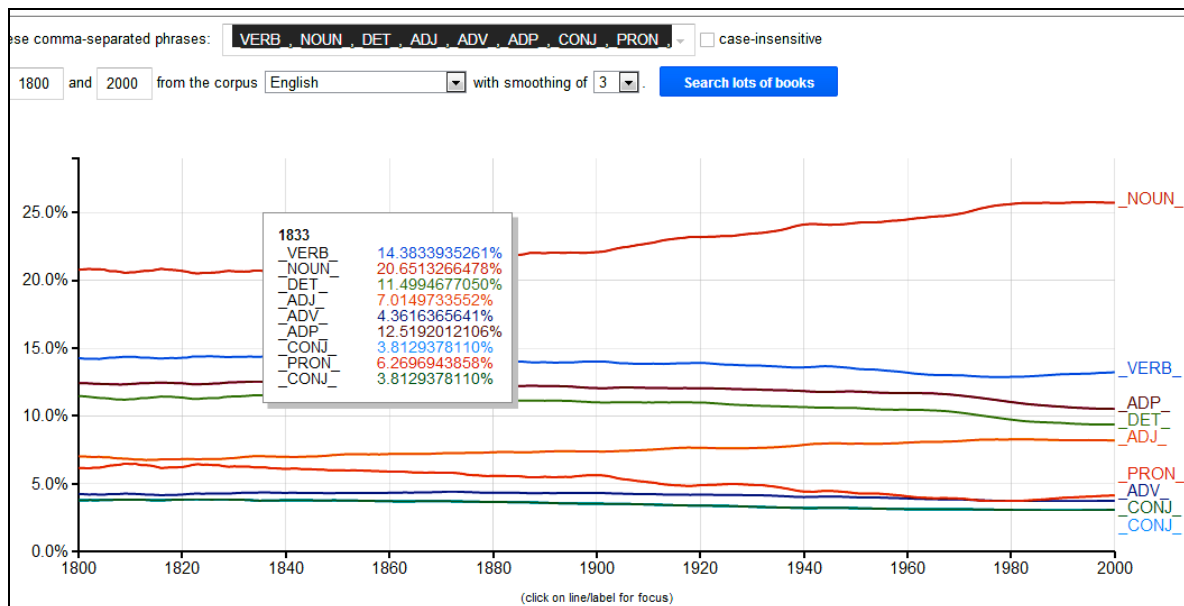


Figure 2. L'évolution des parties du discours dans le corpus anglais.

**1.2.** Aux codes grammaticaux s'ajoutent quelques symboles définissant la place des mots dans la phrase (START\_ ou \_END) ou le rapport des mots entre eux ( $\Rightarrow$ ,  $\Leftarrow$ , ROOT). Naturellement l'interrogation peut porter non seulement sur un mot unique, lemmatisé ou non, mais aussi sur une chaîne de plusieurs mots (de 1 à 5), chacun d'entre eux pouvant admettre des filtres. En certains cas cependant, la mention ajoutée de codes grammaticaux réduit la portée à 3 mots seulement.

**1.3.** Certaines manipulations numériques sont possibles, pour stipuler : 1° le regroupement de plusieurs mots dans une seule requête (signe +), 2° le rapport proportionnel d'un mot à l'autre par la soustraction (signe -) ou le quotient (signe /), 3° ou le rééquilibrage de deux mots (signe \* appliqué au moins fréquent). Ces signes du métalangage sont activés, si besoin est, par les parenthèses ( ) ou désactivés par les crochets [ ].

**1.4.** Les progrès ne résident pas seulement dans l'annotation des textes, ni dans la sophistication de l'exploitation. Ils viennent aussi des étapes initiales qui ont été reprises à partir des images scannées. Une lecture optique améliorée a permis de reconnaître les *s* longs des éditions anciennes ou du moins de corriger par quelque moyen les erreurs systématiques que provoquait cette graphie. De même la segmentation (ou *tokenization*) qui primitivement ne s'appuyait que sur les blancs a introduit un séparateur à la fin des phrases, en empêchant les ngrams de transgresser cette barrière. Inversement la frontière de la page a été abolie. Il faut applaudir à ces changements, tout en regrettant que ces décisions de bon sens n'aient pas été prises du premier coup.

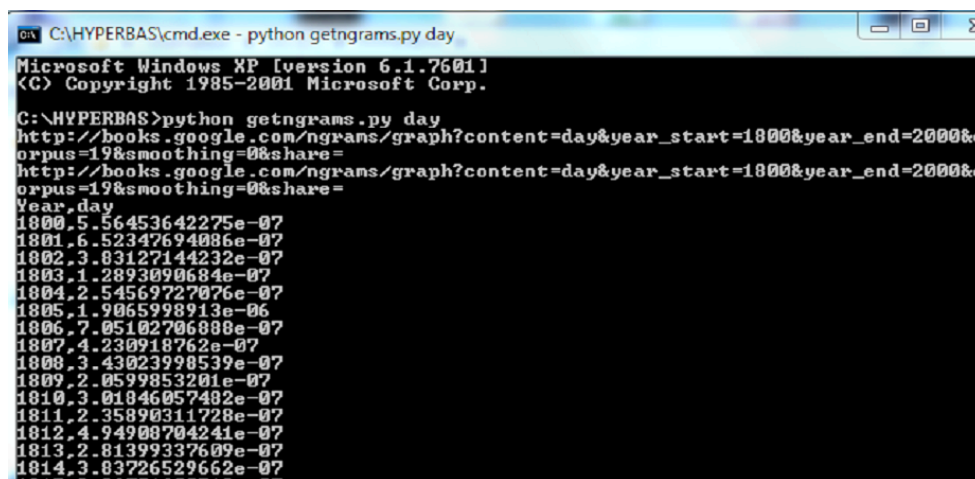
## 2. Accès plus souple et plus puissant

Un handicap pourtant empêchait encore le libre développement de la base Culturomics : en obtenant une courbe opaque au lieu d'une série de nombres, on se heurtait à un *terminus ad quem*, qui interrompait la chaîne des traitements ultérieurs. On pouvait d'autant plus se sentir frustré qu'on sentait en filigrane les données attendues : en promenant la souris sur une année

particulière, on voyait les pourcentages précis surgir dans une fenêtre éphémère<sup>9</sup>, sans qu'on puisse les retenir et les enregistrer.

**2.1.** Les auteurs de *Culturomics* ont donc proposé une API téléchargeable<sup>10</sup> qui pour un mot donné distribue les 201 pourcentages observés le long de la chronologie. Ce court programme, écrit en langage Python, peut être facilement modifié<sup>11</sup> et intégré dans une chaîne de traitement. La fenêtre ci-dessous montre le résultat obtenu quand le programme est lancé avec ses paramètres explicites ou par défaut, un seul étant exigé : le mot cherché (ici *day*).

En réalité la réponse de *Google Books* (enregistrée dans un fichier portant le nom du mot demandé) est une liste de 201 pourcentages, chacun représentant une année de 1800 à 2000 dans le corpus considéré. Pour transformer cette fréquence relative en fréquence absolue et permettre ainsi le calcul de l'écart réduit, on a fait intervenir l'étendue de chaque année dans ce corpus (dont les données ont été puisées dans des fichiers récapitulatifs de *Culturomics*). Ainsi, sachant que l'étendue de l'année 2000 est de 1.182.754.941 mots dans le corpus français de 2012, on reçoit de l'API la distribution par année d'un mot proposé, par exemple le mot "amour" où l'on relève la proportion 0,00013133 pour la même année 2000. La fréquence réelle est donc de  $1182754941 * 0.00013133 = 155331$  pour le mot *amour* dans l'année 2000. En réalité pour des raisons de lisibilité on a renoncé au détail menu des 201 années, en groupant les données par série de trois, soit 67 tranches. Une seconde option est proposée, en 12 tranches, qui répartit au mieux la partition, en déplaçant les jalons chronologiques de façon à égaliser ou tout au moins harmoniser le poids de chaque sous-ensemble. Il se trouve en effet que dans la composition du corpus les livres des époques anciennes sont beaucoup moins représentés que les livres modernes, parce qu'ils sont moins nombreux et moins disponibles dans les bibliothèques.



```

C:\HYPERBAS\cmd.exe - python getngrams.py day
Microsoft Windows XP [version 6.1.7601]
(C) Copyright 1985-2001 Microsoft Corp.
C:\HYPERBAS>python getngrams.py day
http://books.google.com/ngrams/graph?content=day&year_start=1800&year_end=2000&c
orpus=19&smoothing=0&share=
http://books.google.com/ngrams/graph?content=day&year_start=1800&year_end=2000&c
orpus=19&smoothing=0&share=
Year, day
1800, 5.56453642275e-07
1801, 6.52347694086e-07
1802, 3.83127144232e-07
1803, 1.2893090684e-07
1804, 2.54569727076e-07
1805, 1.9065998913e-06
1806, 7.05102706888e-07
1807, 4.230918762e-07
1808, 3.43023998539e-07
1809, 2.0599853201e-07
1810, 3.01846057482e-07
1811, 2.35890311728e-07
1812, 4.94908704241e-07
1813, 2.81399337609e-07
1814, 3.83726529662e-07
1815, 3.83726529662e-07

```

Figure 3. Résultat de l'API *getngrams.py*

<sup>9</sup> La figure 1 montre ainsi un zoom partiel opéré sur l'année 1915, et la figure 2 sur l'année 1833.

<sup>10</sup> Il ne s'agit pas toutefois d'une API véritable et certifiée, mais de la captation du message retourné par *Culturomics* en réponse à toute interrogation. Les éléments qui servent à établir les coordonnées des points de la courbe figurent dans cette réponse et sont saisis au passage. Ce détournement reste fragile et subordonné à la stabilité du dialogue serveur-client. Un tel changement est intervenu et a rendu inopérant le premier programme (*GetNgrams.py*) distribué par les auteurs de *Culturomics*. Laurent Vanni, du laboratoire BCL, s'est chargé des rectifications nécessaires et pourra assurer le maintien de ce service.

<sup>11</sup> Une légère retouche, due aussi à Laurent Vanni, convertit dans le codage ANSI traditionnel les caractères accentués de l'Unicode.

En reprenant l'exemple du couple *homme/femme*, on dispose ainsi (dans la figure 4) des éléments dont se repaît habituellement la lexicométrie (des fréquences absolues réparties dans un tableau avec des lignes correspondant aux mots et des colonnes réservés aux textes ou aux périodes). Dès lors toutes les transformations sont possibles qui mènent aux histogrammes (d'une ligne ou d'une colonne) et aux analyses factorielles ou arborées.

	I801	I804	I807	I810	I813	I816	I819	I822	I825	I828	I831	I834	I837	I840	I843	I846	I849	I852	I855	I858	I861	I864	I867																													
homme	211020	242257	188494	170801	160131	225357	479274	578344	686386	628959	438095	567516	605324	673060	699216	833459	558164	649523	855930	1037959	1130068	1148824	1036121	705116	526965	582597	621989	609151	595430	620484	579230	589015																				
hommes	615949	659736	655433	642825	580191	539105	299030	259434	395704	459704	467652	467613	439762	435262	330090	199130	313885	530711	489124	502932	573148	728330	852237	1032488	1252545	942754	854109	872633	901090	960708	948291	1048204																				
femme	1116027	1256394	1571180	1434246	198594	220071	177818	163450	148390	247693	533790	637751	717493	637799	422161	573436	619969	627284	660830	840584	562310	612919	742567	915700	984130	965148	888846	629018	491950	488818	528693	501556	505524	515661	486988	518533																
femmes	527709	562916	582367	559923	482048	455270	314306	289530	333719	339524	332795	354041	320291	293345	235936	148153	238311	339055	318982	333912	376650	476961	597512	705076	859570	673801	609657	617020	620406	663488	650270	721617	796227	876258	1121899	35494019	hommes															
	62461	100704	93187	80224	71603	68009	139488	199138	241535	231406	155594	200945	226543	201298	231805	278094	194622	206684	269965	344108	358847	353936	370559	259544	203581	235898	282901	291875	300986	331730	273486	308892	318239	344368	354148	340424	323160	261087	135561	116914	190216	219082	230752	231805	204493	185304	143556	72850	92896	160883		
	166925	170135	186139	230180	300901	352804	424260	342892	369068	398119	401887	427352	420290	454754	504499	562510	738236	17546277	femme	63944	63289	60144	53638	45490	49822	107298	122328	169325	141627	99481	141293	152679	144374	156404	180283	110319	136848	161067	210323	220983	217689	214864	150082	122598	138473	159550	160236	163920	171373	169787	171066	195091
	213797	222394	202705	185752	166025	103846	96999	123670	134787	136690	146207	120586	109015	87116	41060	66406	100707	103800	113997	133097	162938	209713	272207	322526	274111	320398	381197	413811	458225	441060	508210	604373	708224	907964	13218801	femmes																

Figure 4. La constitution d'un tableau de fréquences réelles (ou absolues)

La constitution de tels tableaux n'est pas des plus rapides, chaque ligne faisant l'objet d'une interrogation particulière, lancée sur le réseau. Il faut prévoir une ou deux secondes pour chacune. Mais aucune limitation dans les paramètres de la recherche, laquelle peut mélanger impunément les formes brutes, les formes codées, les catégories prises dans leur ensemble, les contraintes sur l'environnement prosodique, grammatical ou sémantique. On peut mêler dans le même ensemble des mots simples (ou unigrams) ou des expressions complexes (de 2, 3, 4 ou 5 mots). On peut enfin glisser dans le même tableau des observations issues de corpus différents, pourvu que la périodicité reste constante et que les jalons chronologiques ne soient pas déplacés<sup>12</sup>.

2.2. Ainsi aux quatre représentants français du tableau 4, ajoutons leurs homologues anglais, allemands, espagnols et italiens, en réunissant dans un même total les femmes de toutes les nationalités et symétriquement le collège international des hommes. L'évolution que l'on constate est sans appel. Si la masse des hommes est le double de celle des femmes (221 millions d'occurrences contre 548), leur supériorité s'amenuise au fil des ans. Les femmes, sept fois moins nombreuses au départ, voient diminuer leur handicap, au point de dépasser les

<sup>12</sup> La loi hypergéométrique habituellement utilisée en lexicométrie est inutilisable à l'échelle du milliard. On a donc eu recours à la loi normale et au calcul classique de l'écart réduit. Noter que les courbes et les analyses factorielles ou arborées prennent appui sur ces écarts, qui font toujours référence à la totalité du corpus considéré. On s'abstiendra donc de considérer le tableau des fréquences comme un tableau de contingence, qui se suffirait à lui-même et dont les totaux marginaux permettraient d'établir les effectifs théoriques et les mesures du CHI2. Un tel calcul pourrait se légitimer (on l'a même facilité pour l'analyse factorielle), mais on a préféré, pour plus de généralité et de stabilité, considérer toujours le **corpus entier comme la référence interne** pour les partitions chronologiques dont il est la somme. Dans le cas de corpus différents traités en même temps, cette pondération est indispensable, sans quoi les résultats ne seraient que le reflet de la taille des corpus et sous-corpus.

hommes dans la dernière tranche (près de 10 millions contre 8). Si l'on superpose les deux courbes comme dans la figure 5, l'inversion des tendances est manifeste.

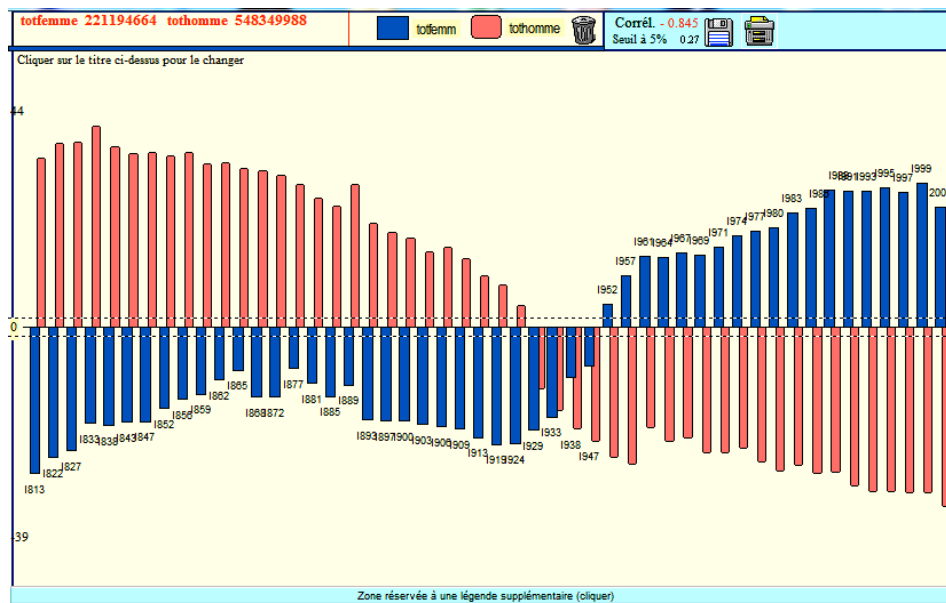


Figure 5. Hommes et femmes dans cinq langues occidentales. Evolution inverse.

En maintenant les 20 lignes et les 50 colonnes du tableau, on peut recourir à l'analyse arborée. S'ajoutant à tant de publications relevant des *gender studies*, la figure 6 apporte une confirmation intéressante, puisqu'elle rend compte de l'opposition *homme/femme* dans différentes langues<sup>13</sup>. On a ainsi l'occasion rare de croiser le temps et l'espace et d'observer si les frontières nationales empêchent un mouvement de propager ses ondes dans le monde entier.

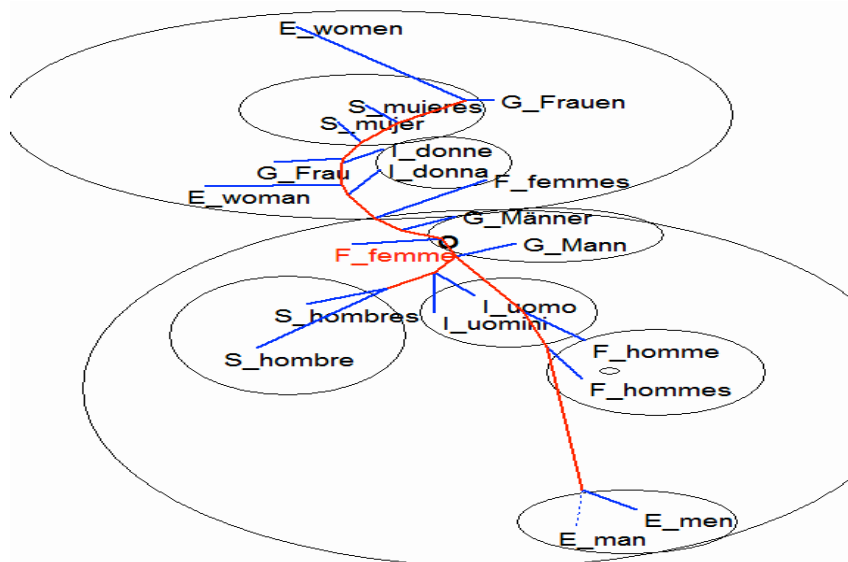


Figure 6. Le couple homme/femme dans cinq langues occidentales. Analyse arborée portant sur 20 mots et 762 millions d'occurrences.

<sup>13</sup> Il peut se faire qu'un même mot, par exemple un nom propre- soit commun à plusieurs corpus. On verra alors la popularité du personnage ou du toponyme évoluer dans la géographie comme dans l'histoire. En de tels cas un symbole initial est utile pour la distinction des langues (*F*= french, *E*= english, *G*= german, *S* = spanish, *I* = italian).

On pouvait penser que les langues imposeraient leurs frontières. Il reste certes une solidarité nationale surtout chez les hommes, le singulier et le pluriel se reconnaissant comme compatriotes et se serrant la main. Mais ce n'est là qu'un critère secondaire. Sachant que de façon générale, sur deux siècles, le pluriel tend à s'effacer devant le singulier (ce qu'on observe dans les déterminants, notamment dans le rapport de *le+la+l'* versus *les*), on pouvait imaginer que cette distinction du nombre se retrouverait dans des mots si courants. Il n'en est rien. Ni le nombre, ni la langue n'imposent leur domination. C'est le genre ou plus précisément le sexe qui fait la loi, et qui fait s'affronter deux camps irréconciliables : les hommes font bloc en bas du graphique, et les femmes en haut. On a la situation radicale de Sodome et Gomorrhe même si un mot franchit effrontément la ligne de démarcation : la *femme* française qui lorgne du côté des hommes. Cela tient peut-être à l'ambiguïté du mot *femme* qui réunit deux statuts que l'anglais distingue avec *woman* et *wife*. En tant que *wife* la *femme* française a moins de raison de s'opposer à la gent masculine. C'est l'occasion de souligner que la quantité ne dissipe pas toutes les incertitudes. Même dans un cas apparemment simple, il est dangereux de franchir le pont entre deux langues : les équivalences qu'on établit entre termes symétriques brutalisent toujours peu ou prou la réalité sémantique. En face d'un tel tableau, l'analyse factorielle (figure 7) propose quelques compléments. Cette méthode a l'avantage de mettre ensemble les lignes et les colonnes, et d'interpréter les unes en s'aidant des autres. Du côté des colonnes, les choses sont claires : la chronologie règne sans partage. Toutes les tranches de 1800 à 1933 sont à droite, toutes les autres à gauche. Du côté des lignes, comme dans l'analyse précédente, les choix sont tranchés et les camps retranchés, les femmes à gauche, les hommes à droite. Ici aussi la femme française franchit la frontière et confirme sa propension à l'indépendance, sinon à la trahison.

Le premier facteur, qui accapare 66% de la variance place les hommes dans les tranches éloignées et les femmes dans les tranches récentes et dans les deux cas il distingue le singulier et le pluriel, ce dernier occupant la position extrême. L'opposition semble ainsi moins forte entre l'*homme* et la *femme* qu'entre les *hommes* et les *femmes*. Les débats sur la place de l'homme et de la femme dans la société peuvent certes utiliser la valeur généralisante du singulier, mais le plus souvent le constat des inégalités se fait à l'aide de classes collectives : les *ouvriers*, les *travailleurs*, les *riches*, les *femmes*.

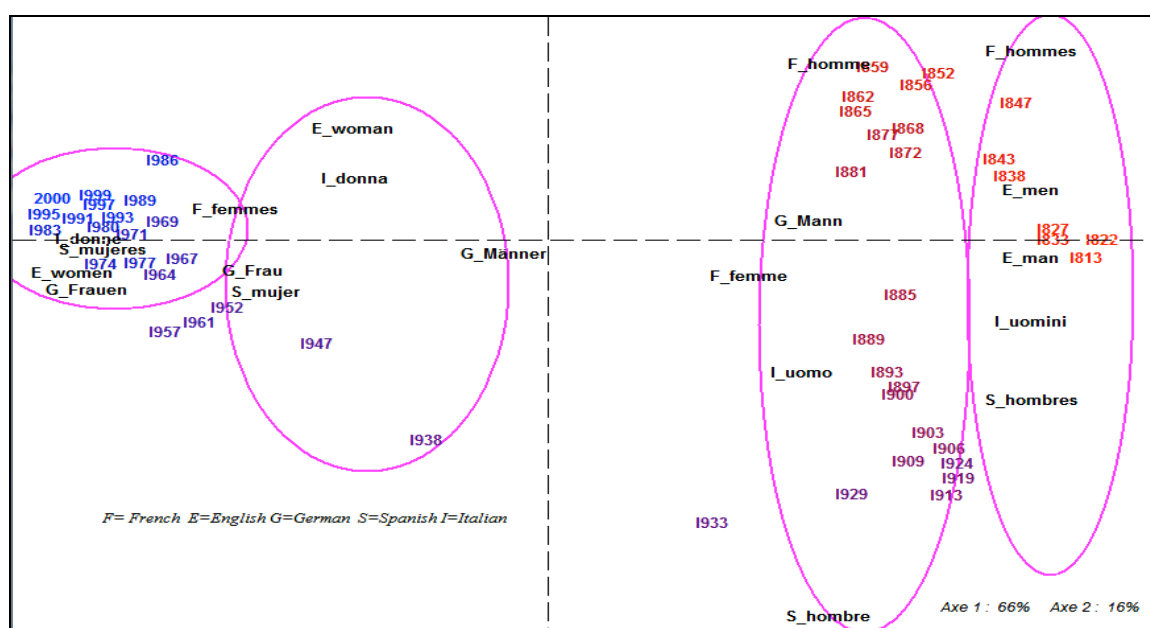


Figure 7. Analyse factorielle du même couple homme/femme



**2.3.** - L'interrogation croisée de corpus différents est plus robuste quand, échappant aux approximations de la traduction, on propose à l'analyse des éléments stables qui ne changent pas d'une langue à l'autre. C'est le cas des ponctuations, des toponymes, des noms de personnes et à moindre degré des parties du discours<sup>14</sup>. Tantôt c'est la convergence qu'on observe, tantôt les particularismes nationaux.

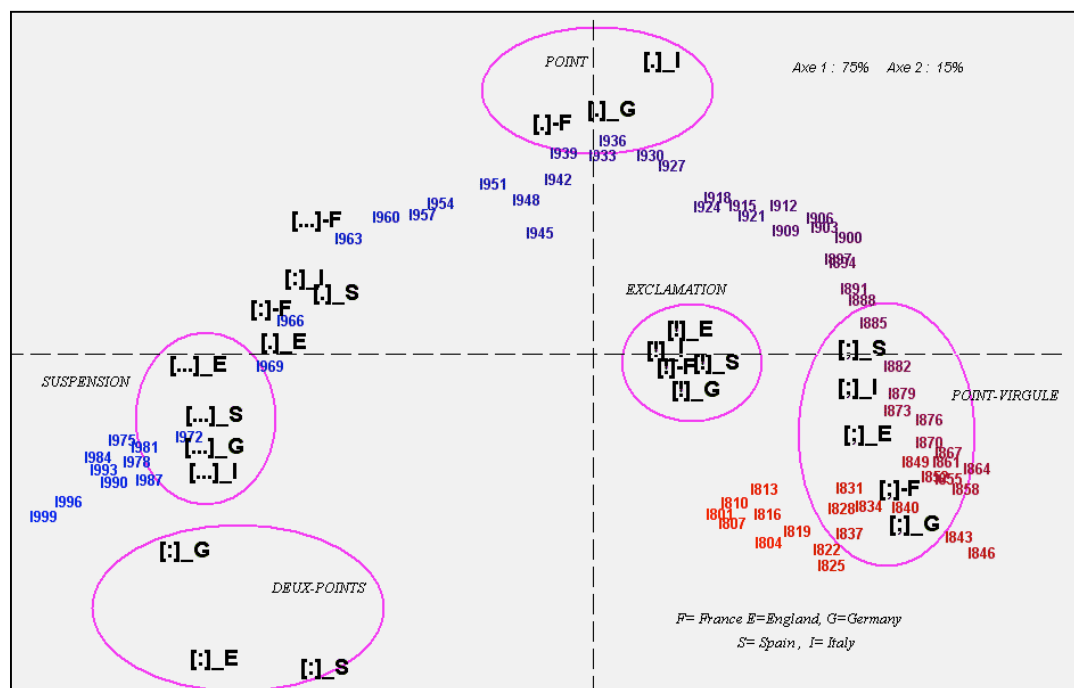


Figure 8. Analyse factorielle des signes de ponctuation

L'exemple des signes de ponctuation est typique du premier cas, d'un mouvement uniforme, semblable à la dérive des continents. Partout l'obsolescence frappe le point-virgule, qui figure pourtant parmi les plus anciens signes du système. Même déclin des signes où l'expression se fait plus intense et plus directe (exclamation et interrogation<sup>15</sup>). Ces signes voisinent avec les tranches éloignées dans le temps qu'on retrouve à droite de la figure 8. A l'opposé, le point, les deux points et les points de suspension s'orientent à gauche, en accord avec les tranches les plus récentes.

A quoi peut-on rattacher cette sorte de glaciation universelle qui se contente d'une expression plus neutre, plus froide, plus tournée vers le constat que vers l'émotion ? Sans doute moins à l'évolution des langues qu'à un changement dans la composition des corpus. Les publications les plus récentes, qui sont aussi les plus nombreuses, n'ont pas été soumises au tri de l'histoire : c'est le tout-venant de l'édition, où pullulent les ouvrages d'information, les traités techniques et les sujets les plus divers. Les livres plus anciens ont survécu à l'oubli et à la perte parce que, leur intérêt se maintenant, des rééditions ont eu lieu qui ont augmenté

<sup>14</sup> Le même jeu de huit étiquettes grammaticales a servi pour tous les corpus : NOUN, VERB, PRON, ADJ, ADV, DET, CONJ et ADP (=préposition + postposition). Il n'est pas certain qu'il convienne de façon homogène à toutes les langues.

<sup>15</sup> La figure 8 ne fait pas mention du point d'interrogation, écarté par prudence, à cause de la spécificité de son emploi en espagnol. En réalité la distorsion ne se produit pas et le point d'interrogation est rangé sans ambages à côté du point d'exclamation. En revanche la virgule n'a pas pu prendre place dans l'enquête : comme ce signe appartient au métalangage du moteur de recherche, il échappe à toute investigation, de même que les guillemets.

leur chance de survie. C'est là le privilège des œuvres littéraires, rarement le cas des publications techniques, que le progrès condamne très vite.

**2.4.** L'orthographe des toponymes comme celle des personnages historiques varie peu en passant les frontières. Avec quelques ajustements<sup>16</sup>, on peut mesurer la notoriété variable qu'un nom propre, de lieu ou de personne, peut avoir dans un pays particulier et à une époque déterminée. Prenons l'exemple des cinq capitales européennes correspondant aux cinq langues disponibles. Les mêmes noms, dans les mêmes divisions du temps, sont successivement recherchés dans les cinq corpus établis en 2012. L'analyse factorielle (figure 9) souligne un déplacement de l'intérêt historique qui tend à abandonner les cités du sud (Rome, Madrid et Paris) pour s'attacher à celles du Nord (Berlin et Londres). Mais le chauvinisme tend à contrarier cette tendance quand la capitale du pays est en cause : *Paris* échappe au déclin en France, et *Rome* en Italie (les points I\_Roma et F\_Paris s'orientent à gauche du graphique, comme E\_London et G\_Berlin, là où se concentrent les tranches contemporaines).

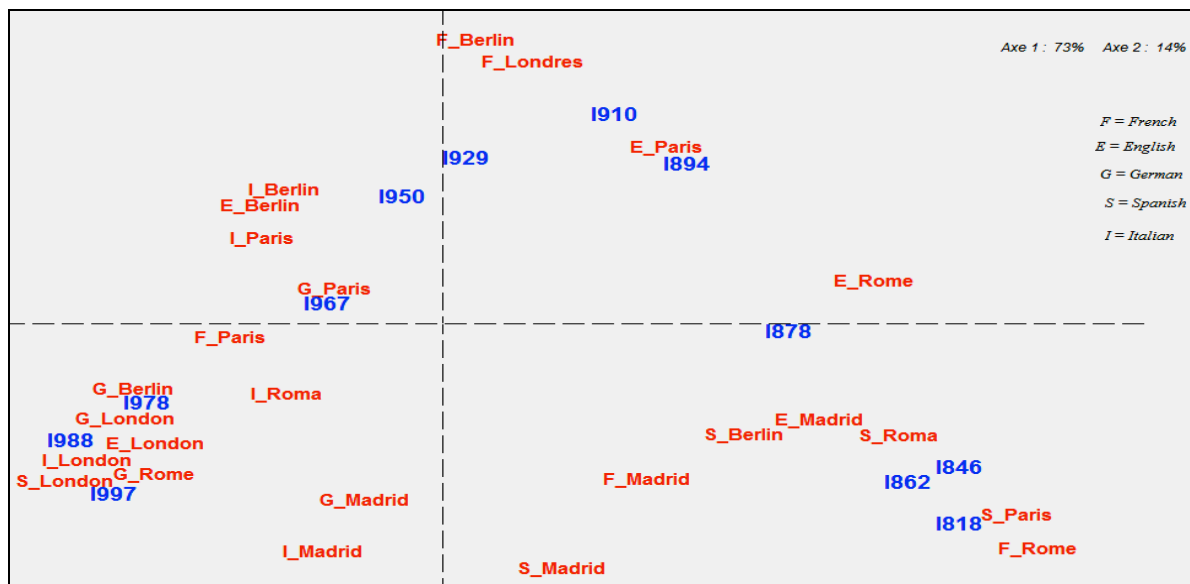


Figure 9. Analyse factorielle de cinq capitales dans cinq langues européennes

**2.5.** La tentation est grande d'aller plus avant et de ne pas se contenter de sondages, même au travers de mots ou signes très fréquents. Envisageons l'ensemble des mots et l'ensemble des corpus occidentaux, soit près de 600 milliards d'observations élémentaires. L'objet d'étude reprend et développe l'étude des parties du discours, déjà abordée dans la figure 2 consacrée au corpus anglais. Or ce corpus n'a rien de spécifique : les mêmes tendances et les mêmes accidents se retrouvent dans les autres langues : une tendance nette qui s'oriente en faveur des catégories nominales (noms et adjectifs) au détriment du verbe et de ses acolytes (adverbes, pronoms, conjonctions).

Il semble que des phénomènes de simplification soient à l'œuvre dans la syntaxe comme dans la ponctuation. Non seulement la phrase se dépouille de ses constructions lourdes qui tournent autour du verbe (conjonctions et pronoms relatifs notamment) mais aussi le langage paraît faire l'économie de l'attirail léger qui accompagne habituellement le nom (déterminants

<sup>16</sup> Ainsi *Rome* se dit *Roma* en italien et *London Londres* en français. Mais ni *Paris*, ni *Berlin*, ni *Madrid* n'admettent de variantes nationales.

et prépositions). Des mesures complémentaires semblent le prouver quand deux constructions sont en concurrence, par exemple *Prép+Nom* préféré à *Prép+Déf+Nom*, ou *Verb+Nom* plutôt que *Verb+Déf+Nom*. Si les prépositions hésitent entre la montée (en espagnol), la stabilité (en allemand) et la descente (partout ailleurs), la décroissance des déterminants est générale, alors que partout les substantifs et les adjectifs sont en progression.

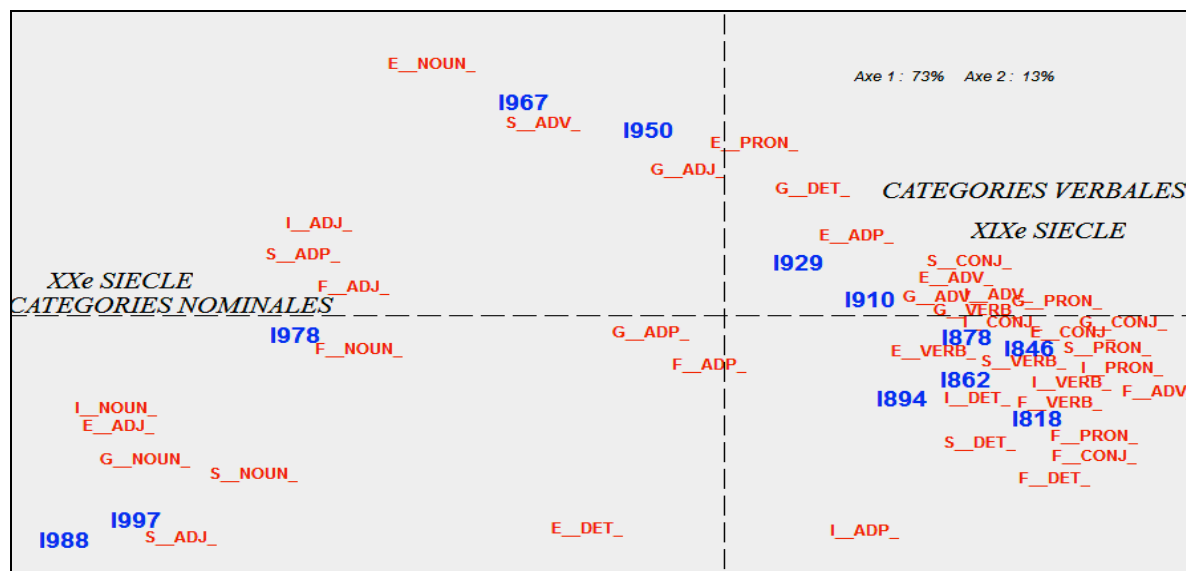


Figure 10. Analyse factorielle des parties du discours (cinq corpus, 600 milliards de mots)  
(E=english, F=french, G=german, S=spanish, I=italian)

Jamais jusqu'ici on n'avait observé cette dislocation du groupe nominal, le substantif marchant seul en tête, suivi à distance de l'adjectif et, plus loin, des prépositions et des déterminants qui peinent à le suivre dans sa progression. Mais là encore le jugement reste perplexe : s'agit-il d'un mouvement de fond, de nature proprement linguistique ? Ne serait-ce pas plutôt un simple artefact prévisible : une conséquence de la loi des genres ? Si la collection des données s'est faite au hasard des opportunités, ne risque-t-on de rencontrer l'utilitaire et le bric-à-brac dans la proximité immédiate, et des objets, plus rares, d'une valeur supérieure, dans l'héritage ancien ? On aurait alors l'opposition bien connue entre le littéraire et l'utilitaire, entre l'expression et l'information, qui se traduit par un dosage différent du verbe et du nom.

### 3. Doutes et vertiges

Du haut de la Tour de Babel, avec 600 milliards de mots entassés sous les pieds, voilà que surgissent le vertige et le doute. Les nuages qui enveloppent le sommet cachent les fondations. Comment s'assurer que l'édifice résistera à la malédiction, à la zizanie linguistique qui a frappé la Tour de la Bible ?

**3.1.** Les auteurs de *Culturomics* ont joué honnêtement le jeu de la transparence en permettant à tout un chacun de vérifier l'exactitude des chiffres. Il ne s'agit pas seulement de contrôler les renvois au texte : faute de temps et de patience, un sondage de cet ordre ne peut s'exercer que sur une frange infinitésimale des données, et si d'aventure la référence annoncée se révélait fautive ou manquante, aucune conclusion générale ne pourrait être tirée de ce manquement particulier. Un contrôle véritable ne peut être fait qu'en reprenant les calculs à la base, ce qui suppose qu'on ait accès aux données brutes. Certes on ne peut remonter le processus jusqu'aux images scannées et il faut se contenter des relevés et comptes qui en ont

été tirés – et qui sont téléchargeables. Il ne faut pas trop s’effrayer du nombre et de la taille des fichiers à transférer, si l’on se satisfait des unigrams (ou mots individuels). Ici on se félicite d’un progrès notable par rapport aux données de 2009 : les fichiers sont classés selon l’initiale des mots. Reste pour chaque lettre à procéder aux opérations lourdes de tri, de compactage et de regroupement dans une seule et même base. Les auteurs de *Culturomics* vont même jusqu’à recommander cette duplication, afin de développer l’exploitation de leurs données tout en soulageant leur serveur.

Même en concentrant les données, en neutralisant la distinction majuscules/minuscules, en réduisant 200 années en 12 tranches chronologiques, en écartant les mots rares qui ont beaucoup de chances de n’être que des erreurs de lecture<sup>17</sup>, on aboutit, avec des chiffres seuls et sans aucun texte, à une base de 300 millions d’octets, grosse de 1,5 million d’entrées. Bien sûr le but avoué de cette coûteuse opération n’est pas seulement de contrôler les données, mais surtout d’en permettre une exploitation facile et immédiate, sans les pesanteurs et les lenteurs liées au réseau. On a surtout eu le souci de s’affranchir des simples pourcentages et de retrouver les chiffres ou effectifs absolus, qui seuls permettent le plein déploiement de la statistique et l’application des méthodes multidimensionnelles. On verra plus loin un aperçu de l’exploitation de cette base, quand l’expertise des données aura été faite.

Une discordance initiale nous inquiète déjà: les chiffres que nous relevons dans les données téléchargées ne correspondent pas exactement à ceux qu’on obtient par le réseau. Reprenons l’exemple du mot *homme* dont le total s’élevait à 43 millions (précisément 43 424 969) dans le tableau 4, extrait de *Culturomics*. On en compte presque deux millions de plus (exactement 45191302) dans les données transférées. Il faut en conclure que les comptes définitifs fixés en 2012 ont été sujets à retouches et que la taille de chaque année du corpus a été calculée après le rejet des rebuts. Certes le profil chronologique du mot se subit pas de graves perturbations, mais l’approximation diminue un peu le crédit qu’on prête aux relevés.

**3.2.** Quant au volume des rebuts, nous ne pouvons en faire qu’une estimation, puisqu’ils ont été caviardés dans une purge préalable aux relevés disponibles. Mais on en aura une idée à partir de l’extrait de la figure 11 qui recense toutes les variétés retrouvées quand l’interrogation porte sur le mot *été*. Avec deux accents dans l’espace de trois lettres, les avatars orthographiques se multiplient à l’infini et une centaine d’avortons lexicaux sont nés de cette prolifération désordonnée.

Le lecteur optique peut être responsable d’un tiers de ces erreurs : chacune des deux voyelles du mot pouvant admettre six interprétations, il y a donc 36 combinaisons possibles dont aucune n’a été négligée<sup>18</sup>. Les deux tiers restants viennent de la lemmatisation. Il y a d’abord un doublage mécanique puisque chaque forme doit pouvoir être interrogée sous deux entrées: avec et sans codage grammatical. Reste le dernier tiers imputable aux mauvais choix du *parser*. En principe il n’y a que deux options possibles pour *été*. Ou bien on a affaire à la saison estivale ou bien c’est le verbe *être* au participe passé. Or c’est bien ce que l’on constate quand l’orthographe est correcte : sans mention de code, le mot a 129 257 681 emplois dont

<sup>17</sup> On a placé la barre à 100 occurrences, largement au dessus de celle de *Culturomics*, qui est de 30. La taille du corpus s’en est trouvée réduite à 70 milliards de mots dans le domaine français.

<sup>18</sup> Encore doit-on supposer correcte la lecture des trois lettres, indépendamment des accents. En réalité beaucoup des *e* reconnus sont des faux, des *a* ou des *o* déguisés. Et inversement beaucoup de vrais étés se sont perdus dans le dédale de l’alphabet.

2 71 786 avec l'étiquette nominale et 127 085 852 comme verbe. Il ne manque qu'une brouille au total : à peine 43 occurrences.

249614 ete	54731 été_verb	2171786	1310 ète_noun	9955 été
130 ete_	1485 été_x	été_noun	131 ète_pron	114 été_adj
5601 ete_adj	415 etè	127085852	366 ète_verb	354 été_adv
95 ete_adp	99 etè_adv	été_verb	40 ète_x	102 été_det
41892 ete_adv	119 etè_noun	962 etè	981 etè	355 etè_noun
62 ete_conj	91 etè_verb	53 etè_det	51 etè_noun	8954 etè_verb
16211 ete_det	75 etè	41 etè_noun	923 etè_verb	42 etè
99993	26 etè_adv	823 etè_verb	1000 etè	120 etè
ete_noun	66 etè	345 etè	67 etè_adv	39 etè_adv
2853 ete_pron	64022 etè	48 etè_det	135 etè_det	39 etè_verb
78 ete_prt	3960 etè_adj	39 etè_noun	751 etè_verb	234 etè
79049 ete_verb	12181 etè_adv	212 etè_verb	6259 etè	46 etè_adv
3271 ete_x	8065 etè_det	773 etè	90 etè_adj	75 etè_det
126767 etè	2103 etè_noun	81 etè_adv	297 etè_adv	71 etè_noun
123 etè_	276 etè_pron	109 etè_det	769 etè_det	614 etè
111 etè_adp	37184 etè_verb	36 etè_noun	1511 etè_noun	39 etè_det
20076 etè_adv	58 etè_x	515 etè_verb	148 etè_pron	531 etè_verb
48605	129257681 etè	2482 etè	41 etè_prt	109 etè
etè_noun	444 etè_	459 etè_adv	3224 etè_verb	44 etè_adv
1493 etè_pron	410 etè_verb	131 etè_det	66 etè_x	

Figure 11. Les avatars du mot été

Voilà semble-t-il de quoi rassurer le linguiste. Mais cette exactitude providentielle peut le troubler, quand il observe la panique du lemmatiseur confronté aux formes inconnues et distribuant les codes à l'aveuglette. Dans le tableau 11, on retrouve le jeu complet des codes disponibles, mis à part celui des conjonctions. Comment concilier deux comportements aussi différents de la machine : un désarroi irrémédiable dans les situations confuses et un découpage au laser dans les contextes réputés clairs. Le premier est facile à comprendre : quand un mot est mal saisi ou mal interprété, le lemmatiseur perd ses repères et tombe dans une erreur, qui à son tour en génère une seconde. Et pour peu que le lecteur optique se trompe de nouveau, on entre dans un labyrinthe inextricable où le fil est perdu. Le second traitement est probablement dû à des décisions autoritaires qui obligent le choix dans une alternative simpliste qui est souvent celle du tout ou rien. Faut-il accorder crédit à la bonne foi du premier traitement qui se trompe souvent mais de façon aléatoire, ou à l'autorité péremptoire du second qui peut conduire à l'erreur systémique ?

**3.3.** Pour en décider portons-nous à la fin de l'alphabet, là où les lexicographes, en fin de chantier, relâchent leur attention. Et observons le mot « *ver* », avec l'orthographe correcte. L'analyse ne semble pas avoir été supervisée puisqu'on y trouve beaucoup de codes fantaisistes que ne permet pas la grammaire française, non plus que l'espagnole qui donne un autre sens au même mot.

*758187 ver, 42188 ver\_adj, 135 ver\_adp, 526 ver\_adv, 9446 ver\_det, 528603 ver\_noun, 1503 ver\_pron, 3506 ver\_prt, 88034 ver\_verb, 88034 ver\_verb, 83771 ver\_x,*

En revanche le mot « *vers* » a reçu un traitement expéditif qui verse toutes les occurrences ( il y en a 38 millions) sur le compte de la préposition en oubliant les poètes qui font des *vers* et les morts qui en font d'autres. On saisit là la preuve d'un double traitement. Le singulier *ver* s'est prêté innocemment à l'automate et s'est trouvé bizarrement tronçonné, avec tout de même un ratio de bonnes réponses de 2 sur 3 (528603 sur 758187). Le pluriel s'est trouvé

assujetti à une décision automatique, sans égard au contexte. Or de telles décisions arbitraires frappent surtout les mots fréquents. On a observé tous les mots dont la fréquence dépasse le million. Il s'en est trouvé plusieurs milliers qui ont un code unique – ce qui est facile à repérer puisque le mot sans code et le mot avec code ont le même total. Il est impossible que dans ce lot énorme un lemmatiseur loyal et fiable n'ait pas repéré – à tort ou à raison – des homographes. L'examen du détail montre qu'on a souvent négligé les formes verbales qui coïncident avec un substantif (*aides, amende, arme, armée, armes, attaques, avantage, balance, barre, cause, crainte, charges, chasse, classe*, etc.). En d'autres cas, le lemmatiseur a été abusé par les pièges du français, quand un mot très commun en cache un autre auquel on ne songe pas, ce qui est le cas de *vers*, mais aussi de *but, bois, bout, car, cours, voies, vins, vit, vives*. Ce constat empêche qu'en s'aidant de *Culturomics* on puisse s'engager très sûrement dans l'analyse des parties du discours et de la structure des langues. Mais l'intérêt linguistique n'est pas lié tout entier au codage grammatical. L'évolution des mots (graphies ou lemmes) est en soi une donnée capitale d'où l'on peut extraire l'histoire des idées, des mœurs et des peuples.

## Conclusion

Pouvait-on faire autrement ? On peut en douter, vu la double contrainte d'une masse gigantesque à traiter en un temps limité. Les données de *Frantext* ont certes une fiabilité supérieure, due à une saisie manuelle et à des contrôles réitérés pendant plus de quarante ans. Avant que l'informatique ait été appelée à traiter les données des équipes de linguistes y avaient balisé le terrain et inventorié les difficultés et les remèdes. Et le but initial – la fabrication d'un dictionnaire – était clairement linguistique. L'entreprise de *Google Books* n'est que documentaire mais son ambition est sans limites. Là où *Frantext* se contentait de millions, *Culturomics* brasse des milliards. On peut espérer qu'un jour la qualité accompagnera la quantité. Mais en deux ans que de chemin parcouru vers l'un et l'autre objectif. On en donnera un seul exemple, tiré du verbe *être*. Maintenant que les formes verbales ont été désambiguïsées, même grossièrement, une seconde suffit pour réunir dans un même tableau les graphies qui appartiennent au modèle, même les formes homographes comme *été, être* ou *étais*. Malgré sa nécessité dans tout discours et sa présence presque dans chaque phrase, ce verbe dont la fréquence culmine à 1,4 milliard est en déclin régulier (corrélation chronologique = -0,96), comme la plupart des verbes. Survivent cependant les formes impersonnelles du participe et de l'infinitif (*étant, été* et *être*) et celles du présent (*suis, es, est, sommes, sont*). Tout le reste est rejeté dans les époques éloignées et accompagne le XIX<sup>e</sup> siècle. A la simplification de la ponctuation et de la syntaxe, s'ajoute donc celle de la conjugaison<sup>19</sup>.

Des milliers de graphes semblables peuvent être extraits de la base GOOFRE2 que nous offrons au public dans sa virginité première<sup>20</sup>. Qu'on en profite sans attendre. L'espérance de vie de tels produits est courte : une version nouvelle de *Culturomics* est en préparation.

<sup>19</sup> Et celle de l'orthographe. Qu'il y ait ou non des réformes officielles, l'usage anticipe sur la loi. Ainsi l'analyse de 350 milliards de lettres montre que l'accent circonflexe tend à disparaître, même dans les cas où sa suppression n'est pas légalement envisagée.

<sup>20</sup> Il est trop tard pour en détailler la structure et en expliquer le mode d'emploi. Précisons seulement que cette base permet trois modes d'exploitation : 1 - obtention de courbes sur Internet (comme dans les figures 1 et 2), 2 obtention de données numériques sur le réseau avec exploitation locale (figures 3 à 10), 3 – exploitation entièrement locale du corpus français.

## Bibliographie

- Bohannon J. (2011). « Google Books, and the future of Culturomics, <http://www.terceracultura.net/tc/wp-content/uploads/2011/01/culturomics.pdf>
- Bohannon J. (2011). “The Science Hall of Fame”, *The Gonzo scientist, Science*, 14 jan. 2011, <http://www.sciencemag.org/content/331/6014/143.3.f>
- Brunet E. (2012). « Au fond du GOOFRE, un gisement de 44 milliards de mots », *JADT 2012*, pp. 7-21. La base GOOFRE est téléchargeable: <http://logometrie.unice.fr/pages/bases> et <http://ancilla.unice.fr/GOOFRE.EXE>.
- Delahaye J.P. (2013). *Nicolas Gauvrit, Culturomics. Le numérique et la culture*, Odile Jacob, 224 pages.
- Lieberman E., Michel J.B., Jackson J., Tang T. & Nowak M.A. (2007). Quantifying the evolutionary dynamics of language, *Nature*, Nature publishing Group, p. 713-716.
- Michel J.B. & al. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books, *Science*, déc. 2010.

