GOOFRE VERSION 2 : VOIR ET TRAITER 600 MILLIARDS DE MOTS

Etienne Brunet[*]
Laurent Vanni[**]

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. Statistique textuelle. Vocabulaire français.

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 deux ans la hauteur a doublé, pour le français comme pour les autres langues[1]. 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 dégarnies ont pris de l'embonpoint, et le déséquilibre entre les premières et les dernières périodes a été partiellement corrigé[2]. D'autre part les textes déjà disponibles dans la version de 2009 ont été repris à la base et soumis à des traitements améliorés

[*] Université de Nice Sophia Antipolis, Nice, France. E-mail: prof.etiennebrunet@gmail.com
[**] Université de Nice Sophia Antipolis, Nice, France. E-mail: l.vanni5070@yahoo.com

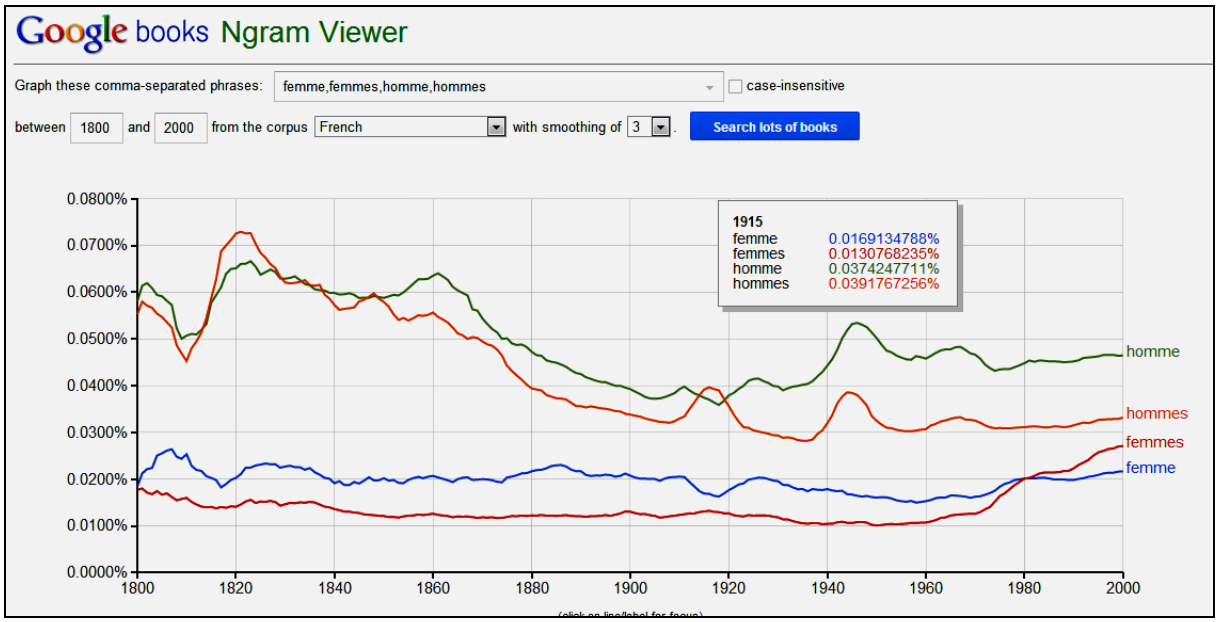
[1] On en est à 89 milliards pour le français, 349 pour l'anglais (où plusieurs variétés peuvent être isolées), 53 pour l'allemand, 67 pour l'espagnol, et 33 pour l'italien, nouveau venu. Ces chiffres correspondent aux données téléchargeables. Ils sont supérieurs dans la table 1 de l'article publié dans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

[2] La correction a été plus nette pour le français que pour les autres langues.

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[3]. Consultons-le sur le couple de substantifs *homme-femme* qui vient souvent en tête dans les listes de fréquence.

Fig. 1 - Interrogation de la base Culturomics. Le couple *homme/femme* de 1800 à 2000[4]



[3] 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.

[4] Une fois complétées les zones de remplissage (les mots cherchés, les dates de début et de fin, le corpus choisi, et l'option de lissage), la chaîne envoyée sur le réseau est ici : *https://books.google.com/ngrams/interactive_chart?content=femme%2Cfemmes%2Chomme%2Chommes&year_start=1800&year_end=2000&corpus=19&smoothing=3*

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.

Améliorations et lemmatisation

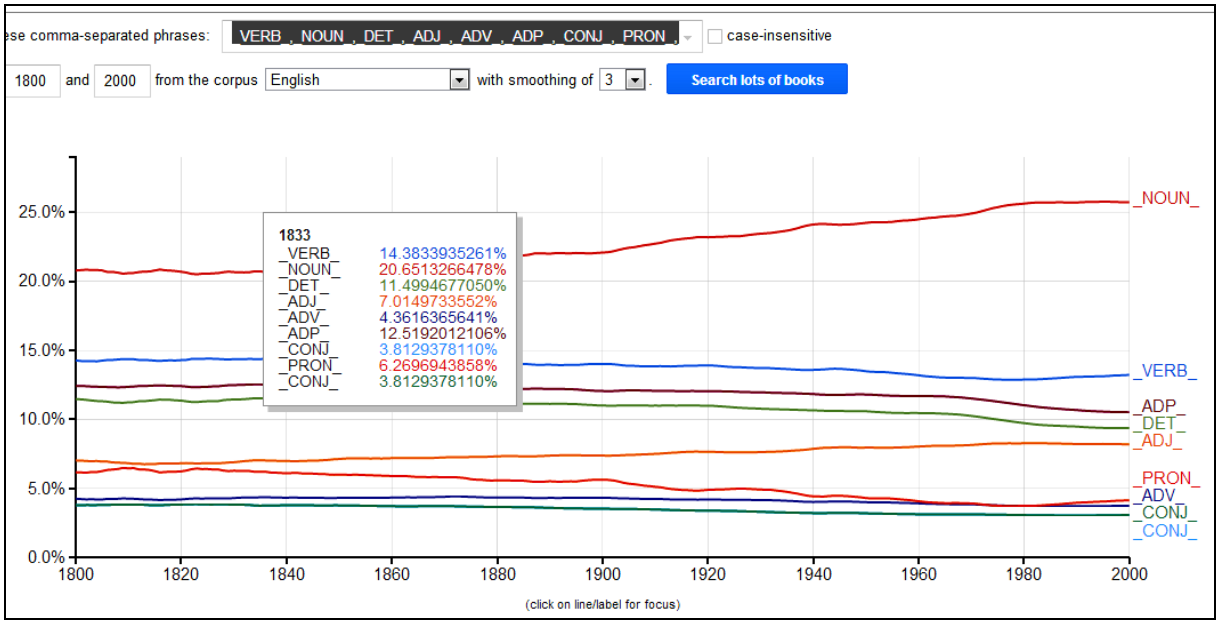
On pourrait croire que l'interrogation de *Culturomics* n'a pas varié depuis la première version et que seule l'assise des données s'est élargie. En réalité il existe un mode d'emploi dit avancé[5] qui change radicalement le point de vue et permet d'accéder non plus seulement aux formes brutes mais à des formes annotées et pourvues d'un code grammatical.

1 - Contrairement aux corpus de l'édition 2009, les données de l'édition 2012 ont bénéficié de la lemmatisation. Celle-ci n'est pas complète et ne permet pas d'obtenir d'un coup toutes les formes par exemple, du verbe "marcher". Mais si le regroupement des formes n'est pas prévu, la séparation des homographes 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

[5] Les multiples façons d'interroger la base Culturomics sont détaillées à l'adresse http://books.google.com/ngrams/info

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[6].

Fig. 2 - L’évolution des parties du discours dans le corpus anglais.



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 (=>, <= , _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

[6] 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.

cependant, la mention ajoutée de codes grammaticaux réduit la portée à 3 mots seulement.

3 - Certaines manipulations numériques sont possibles, pour stipuler
- le regroupement de plusieurs mots dans une seule requête (signe +),
- le rapport proportionnel d'un mot à l'autre par la soustraction (signe -) ou le quotient (signe /)
- 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 [ ].

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 les décisions de bon sens n'aient pas été prises du premier coup.

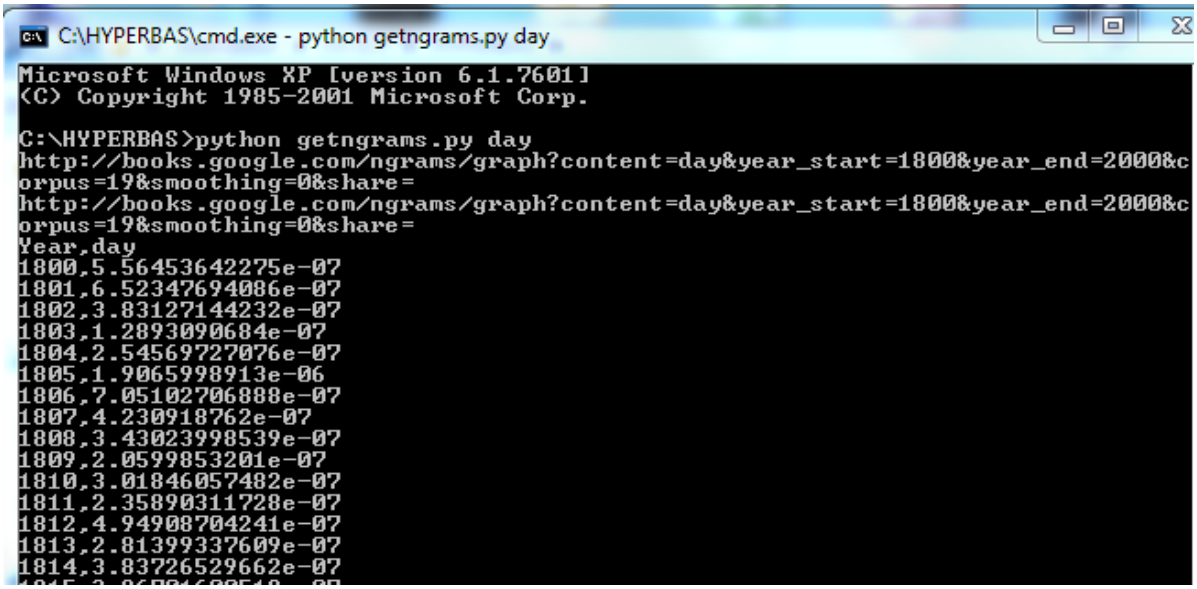
Accès plus souple et plus puissant

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[7], sans qu'on puisse les retenir et les enregistrer.

[7] La figure 1 montre ainsi un zoom partiel opéré sur l'année 1915, et la figure 2 sur l'année 1833.

1 - Les auteurs de *Culturomics* ont donc proposé une API téléchargeable[8] 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[9] 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*).

Fig. 3 - Résultat de l'API getngrams.py



En réalité la réponse de Google Books 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

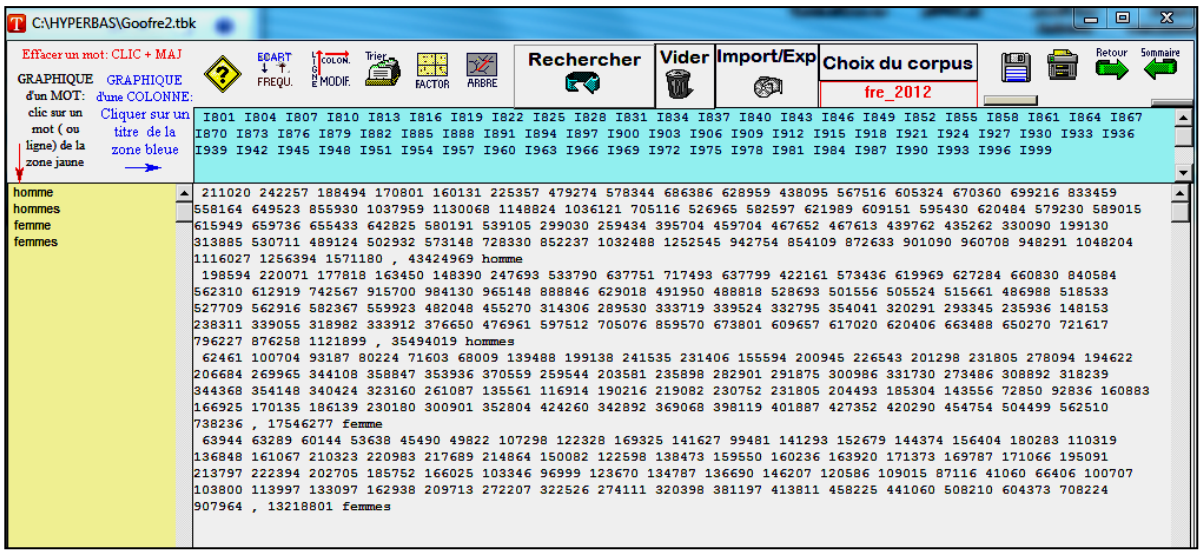
[8] Il ne s'agit pas toutefois d'une API véritable et certifiée, mais de la captation du message retourné par *Culturomics* en réponse à toute interrogation. Les éléments qui servent à établir les coordonnées des points de la courbe figurent dans cette réponse et sont saisis au passage. Ce détournement reste fragile et subordonné à la stabilité du dialogue serveur-client. Un tel changement est intervenu récemment 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.

[9] Une légère retouche, due aussi à Laurent Vanni, convertit dans le codage ANSI traditionnel les caractères accentués de l'unicode.

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 les regroupant par tranches équilibrées. Cette option affine et répartit au mieux la partition, en déplaçant les jalons chronologiques de façon à égaliser ou tout au moins harmoniser le poids de chaque tranche. 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.

En reprenant l'exemple du couple *homme/femme* représenté dans la figure 1, on dispose ainsi 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.

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



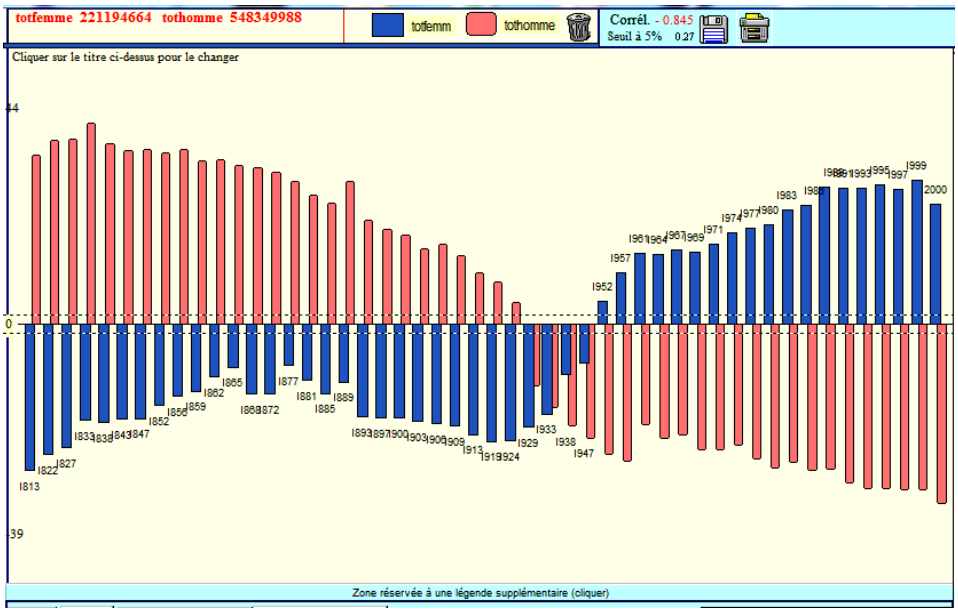
La constitution de tels tableaux n'est pas des plus rapides, chaque ligne faisant l'objet d'une interrogation particulière, lancée sur le réseau. Il faut prévoir une ou deux secondes pour chacune. Mais aucune limitation dans les paramètres de la

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[10].

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 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.

[10] 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.

Fig. 5 - *Hommes* et *femmes* dans cinq langues occidentales. Evolution inverse.



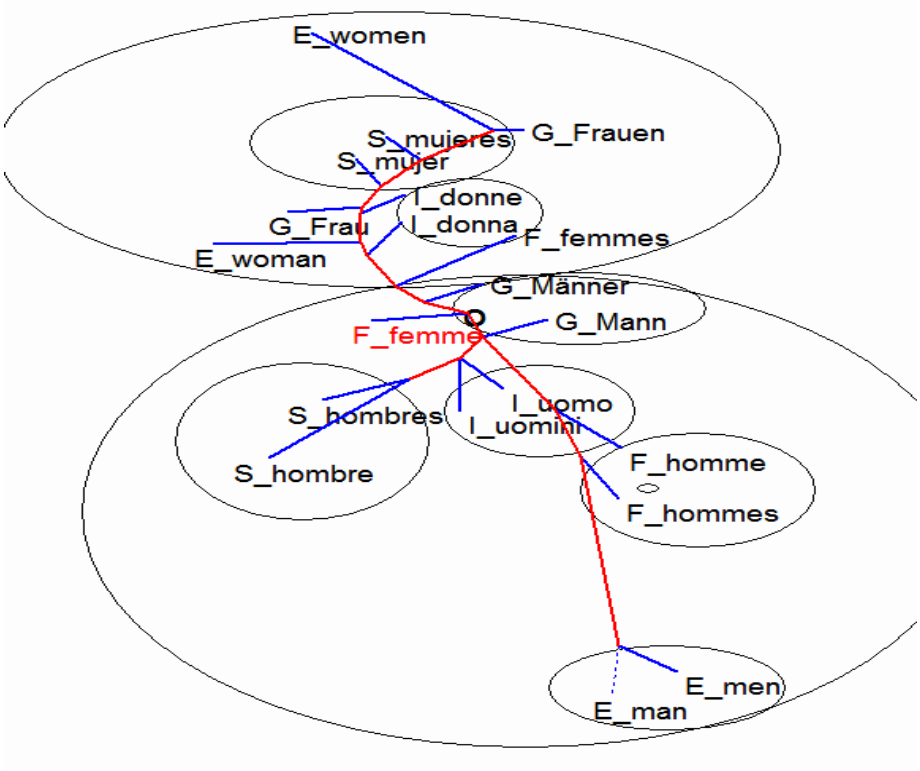
En maintenant les 20 lignes et les 50 colonnes du tableau, on peut recourir à l'analyse arborée. S'ajoutant à tant de publications relevant des *gender studies,* la figure 6 apporte une confirmation intéressante, puisqu'elle rend compte de l'opposition *homme/femme* dans différentes langues[11]. On a 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.

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

[11] 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).

haut. On a la situation radicale de Sodome et Gomorrhe même si un mot s'approche effrontément de 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.

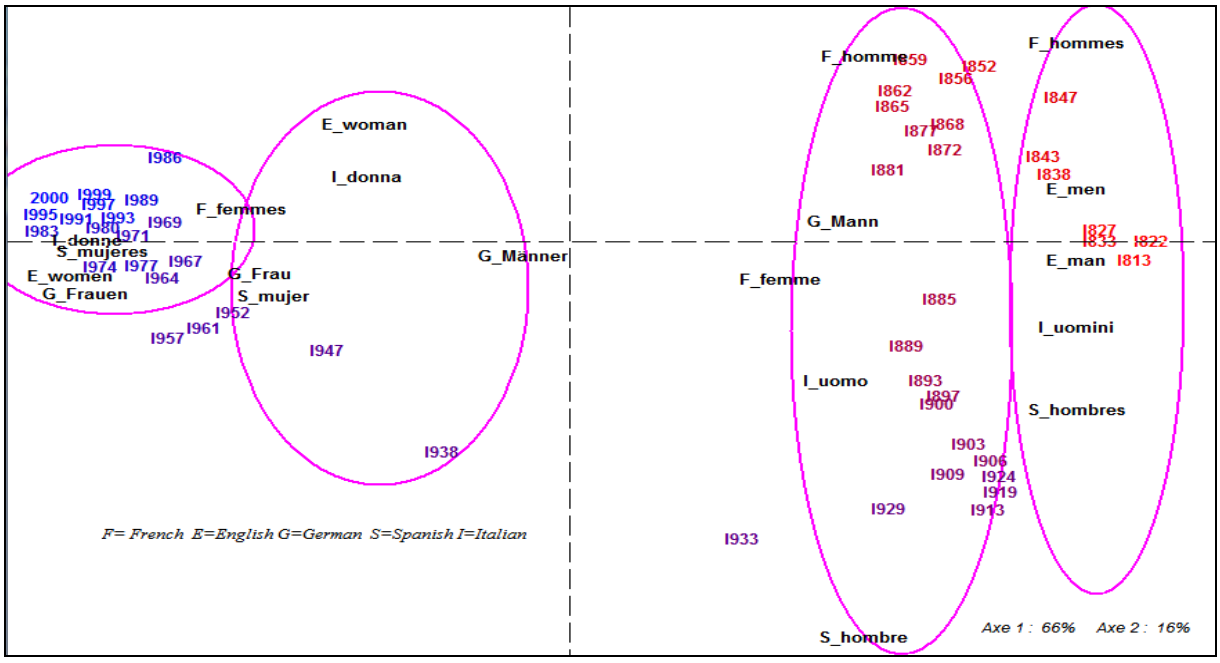
Fig. 6 - Le couple *homme/femme* dans cinq langues occidentales
Analyse arborée portant sur 20 mots et 762 millions d'occurrences.



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 joue avec 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*.

Fig. 7 - Analyse factorielle du même couple *homme/femme*



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[12]. Tantôt c'est la convergence qu'on observe, tantôt les particularismes nationaux.

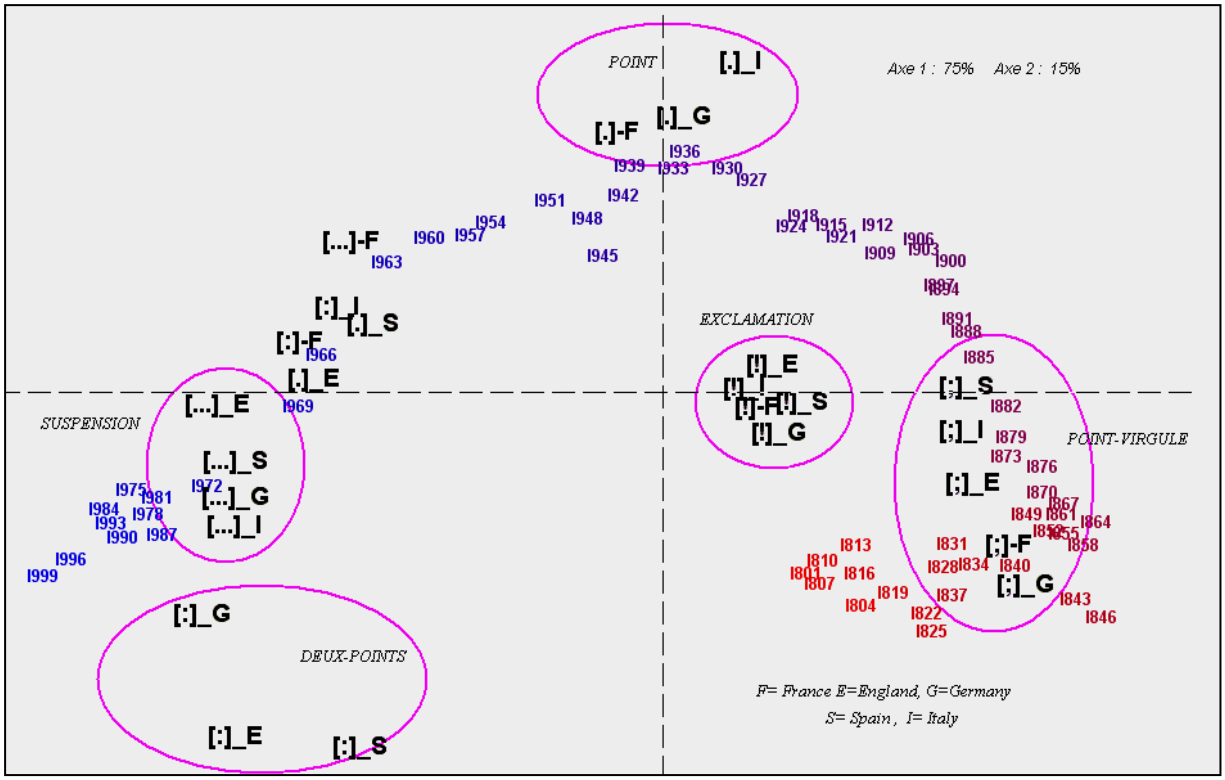
[12] Le même jeu de huit étiquettes grammaticales a servi pour tous les corpus : NOUN, VERB, PRON, ADJ, ADV, DET, CONJ et ADP (=préposition + postposition). Il n'est pas certain qu'il convienne de façon homologue à toutes les langues. S'y ajoutent deux codes de moindre intérêt : PRT (sigles) et X (mots étrangers).

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[13]). 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 perdition parce que, leur intérêt se maintenant, des rééditions ont eu lieu qui ont augmenté leur chance de survie. C'est là le privilège des œuvres littéraires, rarement le cas des publications techniques, que le progrès condamne très vite.

[13] 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.

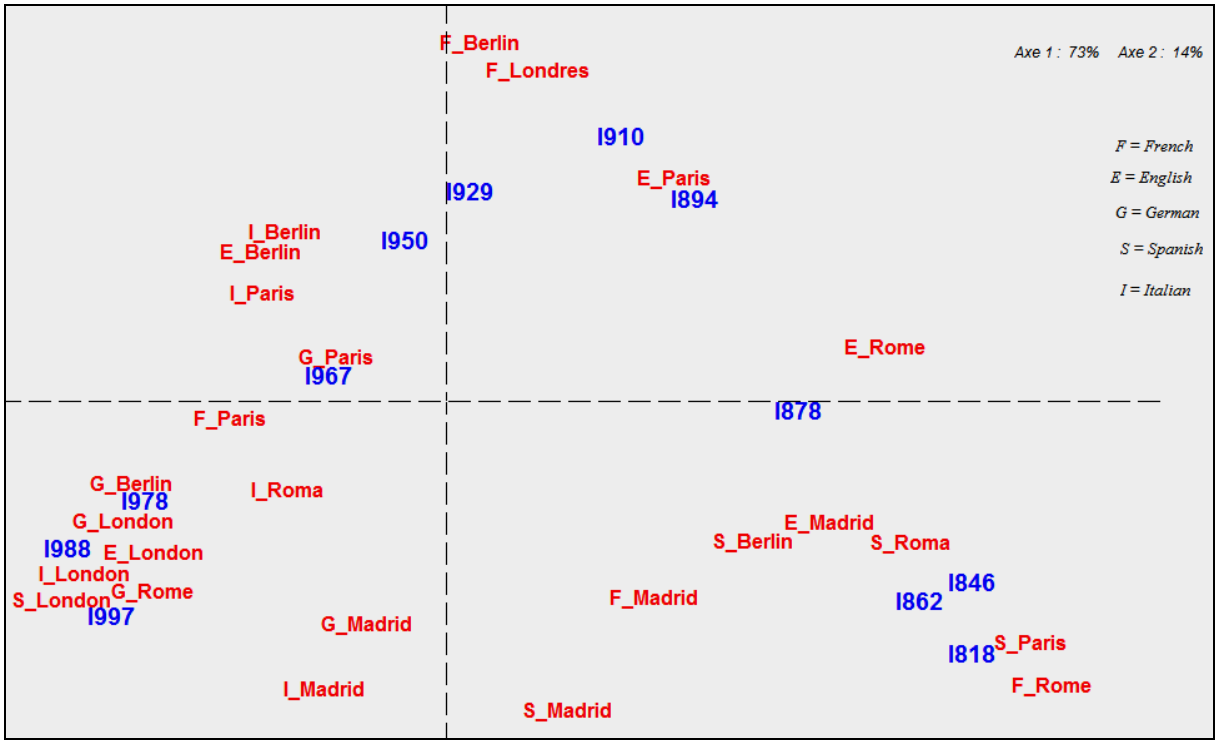
Fig. 8 - Analyse factorielle des signes de ponctuation



L'orthographe des toponymes comme celle des personnages historiques varie peu en passant les frontières. Avec quelques ajustements[14], 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.

[14] Ainsi *Rome* se dit *Roma* en italien et *London Londres* en français. Mais ni *Paris*, ni *Berlin*, ni *Madrid* n'admettent de variantes nationales.

Fig. 9 - Analyse factorielle de cinq capitales dans cinq langues européennes

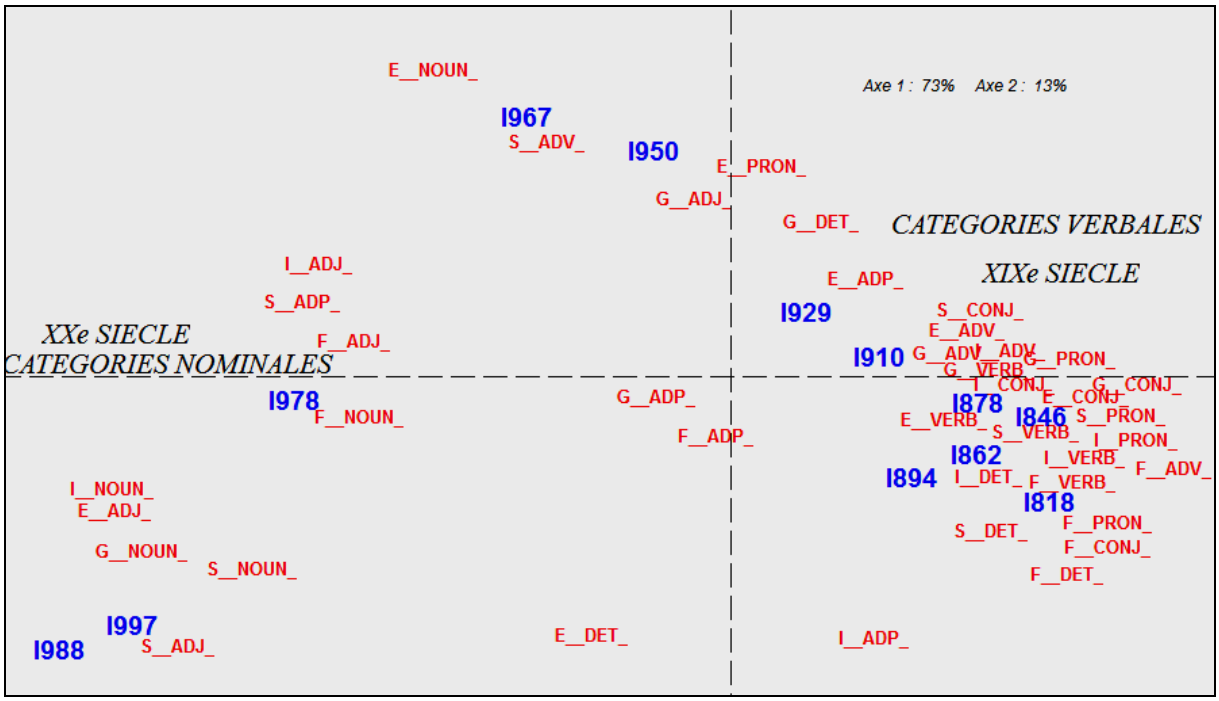


Prenons l'exemple des cinq capitales européennes correspondant aux cinq langues disponibles. Les mêmes noms, dans les mêmes divisions du temps, sont successivement recherchés dans les cinq corpus établis en 2012. L'analyse factorielle (figure 9) souligne un déplacement de l'intérêt historique qui tend 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 : même dans les tranches récentes *Paris* reste populaire 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).

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). 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.

Fig. 10 - Analyse factorielle des parties du discours (cinq corpus, 600 milliards de mots)



Il semble pourtant 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 et prépositions). Des mesures complémentaires semblent le prouver quand deux constructions sont en concurrence, par exemple *Prép+ Nom* préféré à *Prép+Dét+Nom*, ou *Verb+Nom* plutôt que *Verb+Dét+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, 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 suivre la progression.

Doutes et vertiges

Mais 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?

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 une frange infinitésimale des données, et si d'aventure la référence annoncée se révélait fausse 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[15], 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 43424969) 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.

Quant au volume des rebuts, nous ne pouvons en faire 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.

[15] 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.

Fig. 11 - Les avatars du mot *été*

249614 ete	20076 eté_adv	2103 éte_no un	39 êté_no un	51 èté_no un
130 ete_	48605 eté_no un	276 éte_pro n	212 êté_ver b	923 èté_ver b
5601 ete_adj	1493 eté_pro n	37184 éte_ver b	773 été	1000 ète
95 ete_ad p	54731 eté_ver b	58 éte_x	81 ëté_adv	67 ète_adv
41892 ete_adv	1485 eté_x	12925768 1 été	109 ëté_det	135 ète_det
62 ete_co nj	415 etè	444 été_	36 ëté_no un	751 ète_ver b
16211 ete_det	99 etè_adv	410 été__ve rb	515 ëté_ver b	6259 ête
99993 ete_no un	119 etè_no un	2171786 été_no un	2482 ète	90 ête_adj
2853 ete_pro n	91 etè_ver b	12708585 2 été_ver b	459 ète_adv	297 ête_adv
78 ete_prt	75 etê		131 ète_det	769 ête_det
79049 ete_ver b	26 etê_adv	962 étè	1310 ète_no un	1511 ête_no un
3271 ete_x	66 etë	53 étè_det	131 ète_pro n	148 ête_pro n
126767 eté	64022 éte	41 étè_no un	366 ète_ver b	41 ête_prt
123 eté_	3960 éte_adj	823 étè_ver b	40 ète_x	3224 ête_ver b
111 eté_ad p	12181 éte_adv	345 étê	981 ètè	66 ête_x
	8065 éte_det	48 étê_det		9955 êté

114	**355**	**120 êtê**	**46**	**39 ëté_det**
êté_adj	**êté_no**	**39**	**ëte_adv**	**531**
354	**un**	**êtê_adv**	**75 ëte_det**	**ëté_ver**
êté_adv	**8954**	**39**	**71**	**b**
102	**êté_ver**	**êtê_ver**	**ëte_no**	**109 ëtë**
êté_det	**b**	**b**	**un**	**44**
	42 êtè	**234 ëte**	**614 ëté**	**ëtë_adv**

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[16].Les deux tiers restants viennent le la lemmatisation. Il y a d'abord un doublage mécanique, qui rend compte du deuxième tiers. Chaque forme doit pouvoir être interrogée sous deux formes : 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 2171786 avec l'étiquette nominale et 127085852 comme verbe. Il ne manque qu'une broutille au total : à peine 43 occurrences. 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 lacis inextricable où le fil est perdu. Le second traitement est probablement dû à des décisions autoritaires qui obligent le choix dans l'alternative. Faut-il accorder crédit à la bonne foi du premier traitement qui se

[16] 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.

trompe souvent mais de façon aléatoire, ou à l'autorité péremptoire du second qui peut conduire à l'erreur systémique ?

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	9446 ver_det	88034 ver_verb
42188 ver_adj	528603 ver_noun	83771 ver_x
135 ver_adp	1503 ver_pron	
526 ver_adv	3506 ver_prt	

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 tous les mots fréquents. On a observé tous les mots dont la fréquence dépasse le million. Or il s'en est trouvé 6001 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 qu'un lemmatiseur loyal n'ait pas repéré – à tort ou à raison – des homographes dans ce lot énorme qui représente la grande majorité des occurrences, soit 60 milliards sur 70.

Espérances et persévérance

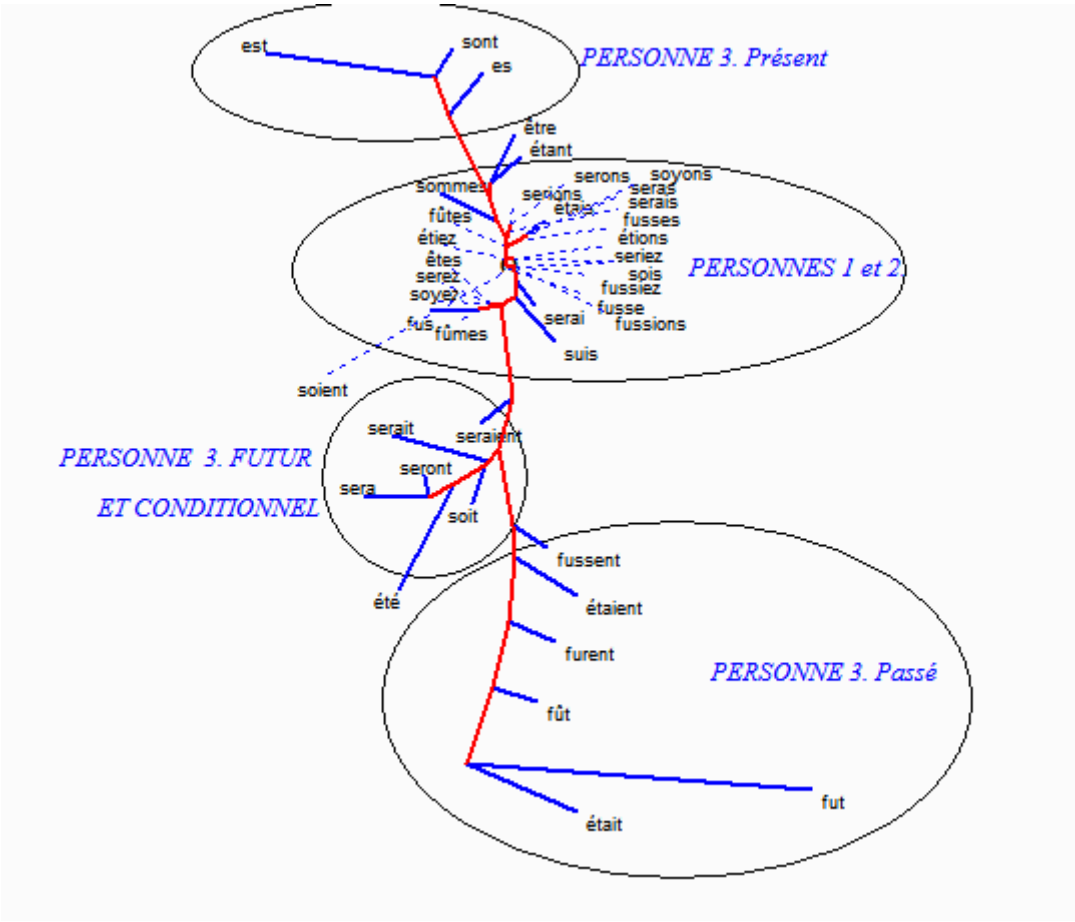
Pouvait-on faire autrement ? On peut en douter, vu la double contrainte d'une masse gigantesque à traiter en un temps limité. Les données de *Frantext* ont certes une fiabilité supérieure, due à une saisie manuelle et à des contrôles réitérés pendant plus de quarante ans. Avant que l'informatique ait 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* (figure 12). Maintenant que les formes verbales ont été désambiguïsées, même grossièrement, une seconde suffit pour réunir dans un même tableau toutes 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 XIXe siècle. A la simplification de la ponctuation et de la syntaxe, s'ajoute donc celle de la conjugaison[17]. Mais si l'analyse arborée de la figure 12 plante le verbe dans le temps avec la tête au présent et les racines dans le passé, futur et conditionnel occupant l'espace intermédiaire, elle est aussi sensible aux personnes, la troisième tenant les deux bouts de la chaîne , tandis qu'une pelote compacte concentre les deux autres.

[17] 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.

Fig. 12 - Analyse arborée des formes désambiguïsées du verbe *être*.



Des milliers de graphes semblables peuvent être extraits de la base GOOFRE2 que nous offrons au public dans sa virginité première[18]. Qu'on en profite sans attendre. L'espérance de vie de tels produits est courte : une version nouvelle de Culturomics est en préparation.

GOOFRE VERSÃO 2: VER E TRATAR 600 BILHÕES DE PALAVRAS

RESUMO: Os dados do Google Books dobraram em dois anos, ultrapassando a marca dos 500 bilhões de palavras. Um novo tratamento retomou as imagens escaneadas para propor uma leitura mais fiel delas. E, pela primeira vez, os textos registrados se beneficiaram com a desambiguação e a lematização. Enfim, o site Culturomics forneceu as ferramentas necessárias para assegurar sua difusão. Convinha então proceder a uma nova estimativa e criar uma nova base, provida de toda aparelhagem estatística que exige, em rede ou local, a exploração dos grandes corpora.

PALAVRAS-CHAVE: Google Books. Culturomics. Estatística textual. Vocabulário francês.

Bibliographie

[18] 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 (figure 12).

BOHANNON, John. Google Books, Wikipedia, and the future of Culturomics. **Science**, v. 331, 14 jan. 2011. Disponible à: <http://www.terceracultura.net/tc/wp-content/uploads/2011/01/culturomics.pdf>.

BOHANNON, John. The Science Hall of Fame. **Science***,* v. 331, 14 jan. 2011. Disponible à: <http://www.sciencemag.org/content/331/6014/143.3.full>.

BRUNET, Etienne. Au fond du GOOFRE, un gisement de 44 milliards de mots. **JADT**, p. 7-21, 2012. La base GOOFRE est téléchargeable à l'adresse: <http://logometrie.unice.fr/pages/bases> et sur le site <http://ancilla.unice.fr/GOOFRE.EXE>.

DELAHAYE, J.P.; GAUVRIT, Nicolas. **Culturomics**: Le numérique et la culture. Paris: Odile Jacob, 2013. 224 p.

LIEBERMAN, Erez. et al. Quantifying the evolutionary dynamics of language. **Nature**, Nature publishing Group, p. 713-716, 2007.

MICHEL, J.B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books. **Science**, déc. 2010. Dispobible à: <http://www.sciencemag.org/content/331/6014/176.full.html>.

Texto recebido em: 10/11/2014.
Texto aceito em: 05/12/2014.