La lexicométrie française: naissance, évolution et perspectives

Etienne Brunet (Bases, Corpus et Langage, CNRS, Université de Nice)

Résumé

Au terme d'une carrière de cinquante ans entièrement consacrée à la statistique linguistique, l'auteur tente d'établir un bilan de la discipline, au moins pour le domaine français. Il s'attache d'abord à évoquer les premières initiatives auxquelles sont associés entre autres les noms de Guiraud, Quemada, Gougenheim,Tournier et Muller. Puis il suit l'évolution des méthodes qui tendent à s'éloigner du modèle inférentiel prôné par Muller pour adopter une démarche descriptive où l'analyse s'appuie sur des calculs multidimensionnels. En passant de la calculette à l'ordinateur l'outil informatique développe sa puissance sur des corpus de taille croissante, dont certains font l'objet d'un examen particulier : la *BNF, Frantext, SketchEngine* et enfin *Google Books*. La taille de ce dernier projet – qui atteint presque 100 milliards de mots pour la production française de ces deux derniers siècles – peut donner le vertige au jugement, sans effacer le doute, la composition du corpus, inégale et incertaine, faussant la chronologie. On en conclut que l'évidence aveuglante d'un résultat graphique ne doit pas en imposer à la raison. L'effet peut être massif, et la cause obscure. La lexicométrie s'est beaucoup étendue en surface ; il lui faut aussi gagner en profondeur.

La naissance de la *statistique linguistique* peut être située en France au début des années soixante. C'est sous cette appellation que Guiraud définit une nouvelle discipline qui ne l'est pas à l'étranger puisque Guiraud en propose une bibliographie dès 1954[1]. Le modèle anglais se nomme *quantitative linguistics* et son représentant le plus connu Gustav Herdan[2]. Au même moment les thèses de Muller[3] faisant appel à l'adjectif *lexical* semblent restreindre le champ d'application et annoncer le terme de *lexicométrie* sous lequel la discipline va se développer vingt ans plus tard[4].

À l'origine, deux événements expliquent l'intérêt porté à cette voie de recherche : le premier est un ouvrage de G.K. Zipf[5] où l'on apprend que la fréquence d'utilisation d'un mot est inversement proportionnelle à son rang, les mots d'un texte étant classés par fréquence décroissante. En réalité les travaux de Zipf sont connus depuis 1935 sous la fameuse loi de Zipf dont les applications se multiplient même en dehors du langage. Le second facteur favorable tient à la disponibilité de l'outil informatique, capable de réaliser traitements et calculs.

---

[1] P. Guiraud et J. Whatmough , *Bibliographie de la statistique linguistique*, Utrecht-Anvers, Mouton, 1954. Il faut attendre quelques années pour que Guiraud digère cette bibliographie et expose les principes de la discipline dans *Problèmes et méthodes de la statistique linguistique*, D. Reidel Publishing Company, Dordrecht-Holland 1959, réédité aux P.U.F. en 1960.

[2] Gustav Herdan, *Quantitative Linguistics*, Londres, Butterwoths, 1964, 284 p. et The advanced theory of Language as Choice and Chance,

[3] Ch. Muller, *Essai de statistique lexicale, l'Illusion comique de P. Corneille*, Paris Klincksieck, 1964, 204 p. et *Essai de statistique lexicale, Le vocabulaire du théâtre de P. Corneille*, Paris, Larousse, 1967, 380 p.

[4] P. Lafon, *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, Genève-Paris, 1984.

[5] Zipf GK (1949). *Human Behavior and the Principle of Least Effort*, Cambridge, Massachusetts: Addison-Wesley.

I - La chiquenaude initiale

La traduction automatique

Les perspectives et les illusions créées par l'avènement de l'ordinateur ont engagé d'emblée la recherche dans le problème le plus ardu : celui de la traduction automatique. Le premier vol dans l'espace accompli par Gagarine en 1961 avait montré le retard de l'Amérique non seulement dans la technologie des fusées mais aussi dans l'information et la maîtrise des langues étrangères. Des crédits inhabituels ont alors alimenté les travaux de linguistique, la plupart orientés vers la traduction.[6] La France s'est aussi engagée dans cette voie, un peu plus tard mais plus longtemps, notamment à Grenoble dans un laboratoire (CETA, puis GETA) fondé et animé par Bernard Vauquois.[7] Dès 1959 paraissait *La Machine à traduire* de E. Delavenay qui rendait compte des travaux des uns et des autres.[8]

Index, Concordances et Dictionnaires

Là aussi les débuts ne sont pas proprement français. Le premier qui imagine un traitement automatique (ou mécanographique) des données textuelles est un jésuite italien, Roberto Busa (qui vient de s'éteindre en 2011, presque centenaire). Le père Busa aimait à raconter la visite qu'il fit en 1949 au siège d*'IBM*. Dans l'antichambre qui menait au bureau de Thomas J. Watson, le fondateur, il s'était saisi d'un écriteau vantant la puissance et la célérité de l'entreprise : « Pour les urgences, c'est déjà fait. Pour les miracles c'est en cours. » Le Père Busa brandit l'écriteau sous le nez du directeur, et, comme il croyait au miracle, il l'obtint sous la forme d'un mécénat de trente ans qui aboutit à l'*Index Thomisticus* en 56 volumes, grand format, reliés cuir. Un ingénieur d'*IBM*, Antonio Zampolli, avait été attaché à ce projet, avant de fonder le *CNUCE* de Pise, de présider l'*ALLC (Association for Linguistic et Litterary Computing*) et de devenir l'un des acteurs majeurs de l'informatique européenne.

En France, quelques années plus tard, grâce à l'appui de René Moreau[9], directeur du développement scientifique d'*IBM-France*, Bernard Quemada et les chercheurs bisontins mettaient en route le *Centre d'étude du vocabulaire français*, en prolongeant une entreprise antérieure initiée dès 1953 par Wagner et Guiraud et vouée à l'établissement d'un *Index du Vocabulaire du théâtre classique*. À la fin des années 50 le Recteur Imbs démarrait le chantier lexicographique qui allait devenir le TLF et où aucun exemple ne devait se trouver qui ne fût daté et signé. Tous ces projets sont d'ordre documentaire. La technique n'y est guère sollicitée que pour fournir des exemples, des références et des relevés. On utilise certes la notion de fréquence et en 1971 R. Martin publie le *Dictionnaire des fréquences* qui rend compte des données amassées au *Trésor de la langue française*. Mais brutes ou relatives, ces fréquences sont données telles quelles, sans donner lieu à une véritable exploitation.

---

[6] On trouvera dans les publications de Jacqueline Léon le récit documenté des essais tentés aux Etats-Unis et en Angleterre pour résoudre les difficultés de l'entreprise, avant qu'un rapport négatif n'aboutisse à la suspension du projet. Voir en particulier l'article « Traduction automatique et formalisation du langage, les tentatives du Cambridge Language Research Unot (1955-1960 » in *The History of Linguistics and Grammatical Praxis* (ed. P.Desmet, L . Jooken, P. Schmitter, P. Swiggers) Louvain/Paris, Peeters, p. 369-394.

[7] Christian Boitet (dir.), *Bernard Vauquois et la tao : vingt-cinq ans de traduction automatique : analectes*, Centre National de la Recherche Scientifique, 1989, 718 p.

[8] Delavenay (Emile) - *La machine à traduire*, Presses Universitaires de France, *Que sais-je ?, 1959*, 1ère édition, 128 pages. La même année, l'auteur fonde l'*ATALA* (Association pour la traduction automatique et la linguistique appliquée), devenue plus tard l'Association pour le traitement automatique des langues.

[9] René Moreau est le coauteur de la première étude statistique, publiée en France, dans le domaine socio-politique : *Le vocabulaire du Général de Gaulle*, en collaboration avec le Professeur Jean-Marie Cotteret ed. Fondation Nouvelles des Sciences Politiques, 1969. Je ne puis éviter d'ajouter que je lui dois mon entrée dans la salle-machine : à un moment où une heure de calcul coûtait le salaire mensuel d'un ouvrier, le mécénat d'IBM m'a permis, trois années durant, d'utiliser librement et gratuitement les ordinateurs du Centre IBM de La Gaude.

La statistique linguistique

1 – C'est Pierre Guiraud qui imagine le premier le profit statistique qu'on pourrait tirer des données engrangées. Mais isolé en France, il note non sans quelque dépit: « La linguistique est la science statistique type; les statisticiens le savent bien; la plupart des linguistes l'ignorent encore[10] ». Les idées de Guiraud ont pourtant germé alentour, à Paris autour de Wagner ou Georges Gougenheim, à Liège autour d'Etienne Evrard, à Strasbourg autour de Ch. Muller.

2 - Le CREDIF, créé en 1959 sous la direction des linguistes Georges Gougenheim et Paul Rivenc se proposait une mission pédagogique : constituer un corpus d'énoncés oraux afin d'établir la liste des mots les plus utiles à la communication en français. Aux simples fréquences s'ajoutaient des calculs plus complexes de disponibilité[11].

3 - Sur le même site de l'ENS de Saint Cloud, un autre laboratoire s'installait sous l'égide de Robert-Léon Wagner et Maurice Tournier : le *Centre de lexicologie politique*. L'équipe qui était épaulée par des mathématiciens ou informaticiens comme G.Th.Guibaud et, plus tard, Pierre Lafon et André Salem, allait pousser plus loin la méthodologie statistique et l'instrumentation informatique, en produisant un manuel de saisie (le *machinal*), divers programmes de traitement (dont le logiciel *Pistes* de P. Muller), des colloques (dont le premier en 1968), et une publication périodique, la revue *Mots.*

4- Fondé à l'Université de Liège en novembre 1961, le *Laboratoire d'Analyse Statistique des Langues Anciennes* (**L.A.S.L.A.**) se donnait pour objectif d'analyser les textes classiques - latins ou grecs - en recourant aux technologies nouvelles. Il ne s'agit plus seulement de produire des index, comme *l'Index Thomisticus* du Père Busa, mais d'analyser chaque forme avant de la soumettre aux tris et aux comptages. Pour la première fois la statistique apparaît pleinement dans le titre et la pratique. Pour la première fois aussi la lemmatisation reçoit l'aide des machines. Il est vrai qu'un latiniste doué de tous les dons, Etienne Evrard, savait déjà maîtriser les ordinateurs, les programmes et les calculs, et en exploiter savamment les résultats.[12]

5 – Au même moment, à Strasbourg, Charles Muller s'employait à établir la méthodologie de la discipline, en l'appliquant au français, et principalement à Corneille. Ses deux thèses sont publiées respectivement en 1964 et 1967. Mais l'influence prépondérante du « lexicomaître »[13] sur des générations de lexicologues vient de son manuel de 1968[14]. Au lieu d'être enseignées par un mathématicien sévère, les leçons de probabilité et de raisonnement statistique ont trouvé sous la plume de Muller une clarté rigoureuse mais souriante. Il s'agissait alors d'une statistique inférentielle fondée sur le schéma d'urne et accessible aux calculettes de l'époque. Dix ans plus tard avec l'abondance des données, la taille des tableaux, la puissance des méthodes multidimensionnelles et la disponibilité des ressources

---

[10] P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, D. Reidel Publishing Company, Dordrecht-Holland, 1959, p. 15. Je possède ce livre précieux, épuisé depuis longtemps. L'auteur m'avait donné son dernier exemplaire, à un moment où la statistique ne l'intéressait plus guère. Guiraud dans sa thèse sur Valéry et dans ses premiers travaux s'était beaucoup investi dans la saisie et l'exploitation statistique des données textuelles. Mais venu trop tôt, sans personnel et sans moyens informatiques, il a renoncé à poursuivre une tâche trop ingrate où les relevés et les calculs devaient se faire à la main.

[11] Georges Gougenheim, René Michea, Paul Rivenc, Aurélien Sauvageot, *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base,* Didier, Paris, 1956. Nouv. éd. refondue et augmentée sous le titre *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Didier, Paris, 1964.

[12] La première publication du *Lasla* date de 1962 : *Sénèque, Consolation à Polybe. Index verborum, relevés statistiques*, La Haye, Mouton, 219 p., éd. Delatte Louis, Évrard Étienne, Govaerts Suzanne, Hazette Pierre.

[13] C'est sous ce titre que Muller apparaît dans *Mélanges offerts à Charles Muller pour son centième anniversaire*, textes réunis par Christian Delcourt, CILF, Paris, 2009, 426 p.

[14] Ch. Muller, *Initiation à la statistique linguistique*, Collection Langue et Langage, Larousse, 1968, 266 p. Une nouvelle édition en deux volumes paraît chez Hachette en 1977, puis chez Champion en 1993.

informatiques, un autre catéchisme devenait nécessaire et ce sont deux mathématiciens, Lebart et Salem, qui prirent le relais de Muller[15]. L'ordinateur remplaçait la calculette et l'analyse des données complétait la statistique inférentielle.

II - L'évolution

La comparaison de ces deux manuels montre assez l'évolution de la discipline. Dans les années 70 on se préoccupait de richesse lexicale, de spécificités, de corrélation et on appliquait aux fréquences les lois statistiques (normale, binomiale et hypergéométrique). Vingt ans plus tard les textes littéraires ont cédé la place aux données commerciales, sociologiques, psychologiques ou politiques. Sous la pression des instituts de sondage, des études de marché et de la veille technologique, de nouveaux logiciels sont nés dont les résultats acquièrent un impact économique. Et les méthodes ont gagné en puissance ce qu'elles perdaient en prétention. Devenue plus modeste et seulement descriptive, l'analyse multidimensionnelle des données a offert des vues synthétiques, mettant de la lumière dans l'opacité des tableaux. Reste à savoir si vingt ans plus tard une nouvelle étape n'a pas été franchie.

Au moment où un dictionnaire de la discipline est en gestation, l'idée nous est venue d'examiner la liste des mots retenus pour cet ouvrage. Si on avait consulté Guiraud ou Muller il y a quarante ou cinquante ans, la liste eût été différente, avec des lacunes (là où ni le mot ni la chose n'existaient) et des points de rencontre (car le dictionnaire d'une discipline doit recouvrir toute son histoire, y compris ses débuts). On se contentera de cette nomenclature en la projetant sur un immense corpus issu de Google Books.

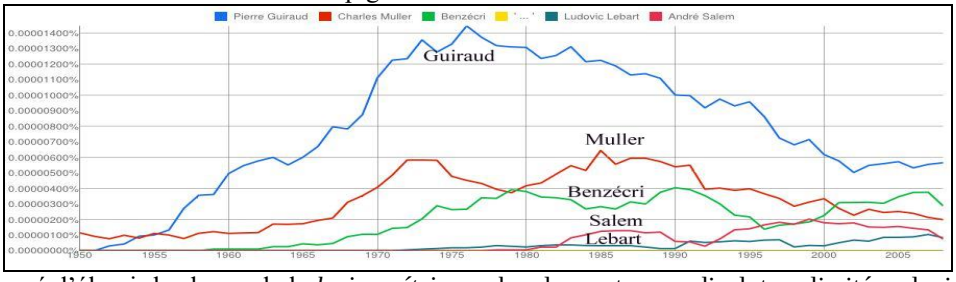
Cette base qui répond au nom de *Culturomics* peut être interrogée pour n'importe quel mot ou expression, comme par exemple la notion de *statistique*. Il suffit de s'adresser au site http://books.google.com/ngrams, en choisissant le français parmi sept langues et en précisant les dates de départ et d'arrivée. La figure 1 montre le progrès du mot de 1800 à 2000, avec une pointe à la fin du XIXe et au milieu du XXe et un fléchissement dans la phase finale[16].Cet essoufflement est moins celui de la discipline que du mot qui la désigne et que d'autres expressions peuvent remplacer. Ainsi en s'en tenant à la période qui nous intéresse on voit dans la figure 2 que l'appellation initiale *statistique linguistique* domine jusqu'en 1980, puis s'efface devant la *lexicométrie*, laquelle à son tour donne des signes de faiblesse, tandis que des termes concurrents semblent vouloir assurer la relève[17]. Le mot *cooccurrences*, que nous avons ajouté au graphique et qui est en croissance rapide, n'est pas un candidat crédible, n'ayant pas le profil idoine, mais il indique la tendance où se porte la recherche actuelle[18].

---

[15] Ludovic Lebart, André Salem, *Statistique textuelle*, Dunod, 1994,| 336 pages.

[16] Observons le destin croisé du singulier (en rouge) rejoint et dépassé par le pluriel en bleu. La statistique tend ainsi à apparaître moins comme une méthode que comme un ensemble de données.

[17] Le profil des promoteurs français accompagne ce décalage qui voit Guiraud céder la place à Muller qui s'incline à son tour devant Benzécri et ses épigones.



[18] On a proposé d'élargir le champ de la *lexicométrie* en abandonnant son radical, trop limité au lexique. Mais si la *logométrie* commence à se répandre, sous l'impulsion de Damon Mayaffre, ni la *textométrie* ni la *stylométrie* ne se sont imposées.

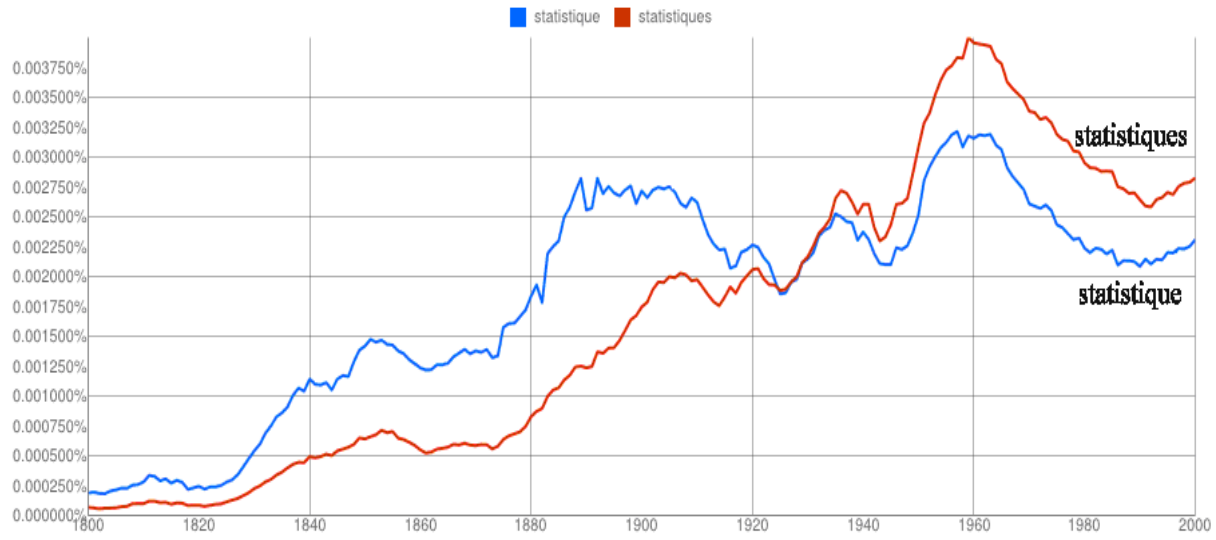


Figure 1. L'évolution de la *statistique* de 1800 à 2000

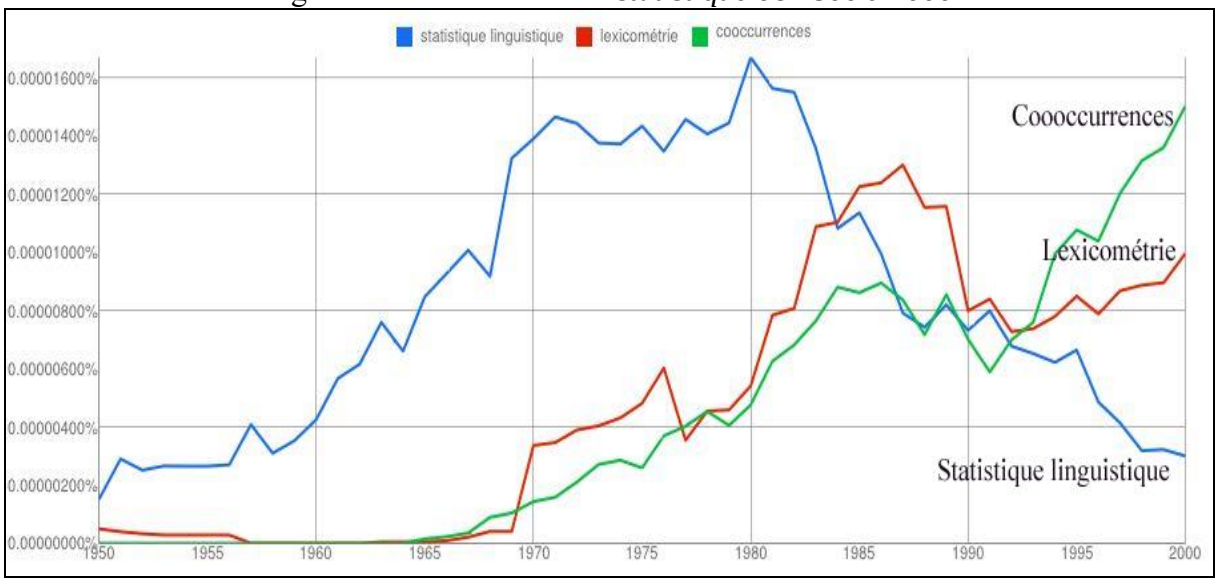


Figure 2. *Statistique linguistique* et *lexicométrie* de 1950 à 2000

On est amené à multiplier l'interrogation de *Culturomics* autant de fois que le présent dictionnaire compte d'entrées. On attend une vue d'ensemble qui situe dans l'histoire les unités représentatives de la discipline. La figure 3 restitue le résultat que la statistique obtient quand elle s'applique à elle-même. On n'y trouvera cependant que les 71 mots qui ont plus d'un million d'occurrences dans le corpus.

Or la chronologie est parfaitement reconnaissable dans le croissant qui parcourt l'espace de droite à gauche. Les premières tranches mettent en oeuvre les ingrédients habituels de la statistique inférentielle. On isole des unités (*unité, forme, mot, ligne, paragraphe*). On établit des partitions (*partie, population, section, volume, série, classe*). On fait des calculs probabilistes (*produit, mesure, moyenne, union, écart réduit, probabilité*) d'après un modèle théorique ou expérimental (*mesure, coefficient, indice, loi, pondération, limite*). On étudie la *fréquence* des mots, leur *distribution*, leur *répartition*, leur *richesse*, leur *accroissement*. On fabrique des *concordances*. On dresse des *courbes*. On délimite le vocabulaire *caractéristique*. Bref on suit l'enseignement de Muller.

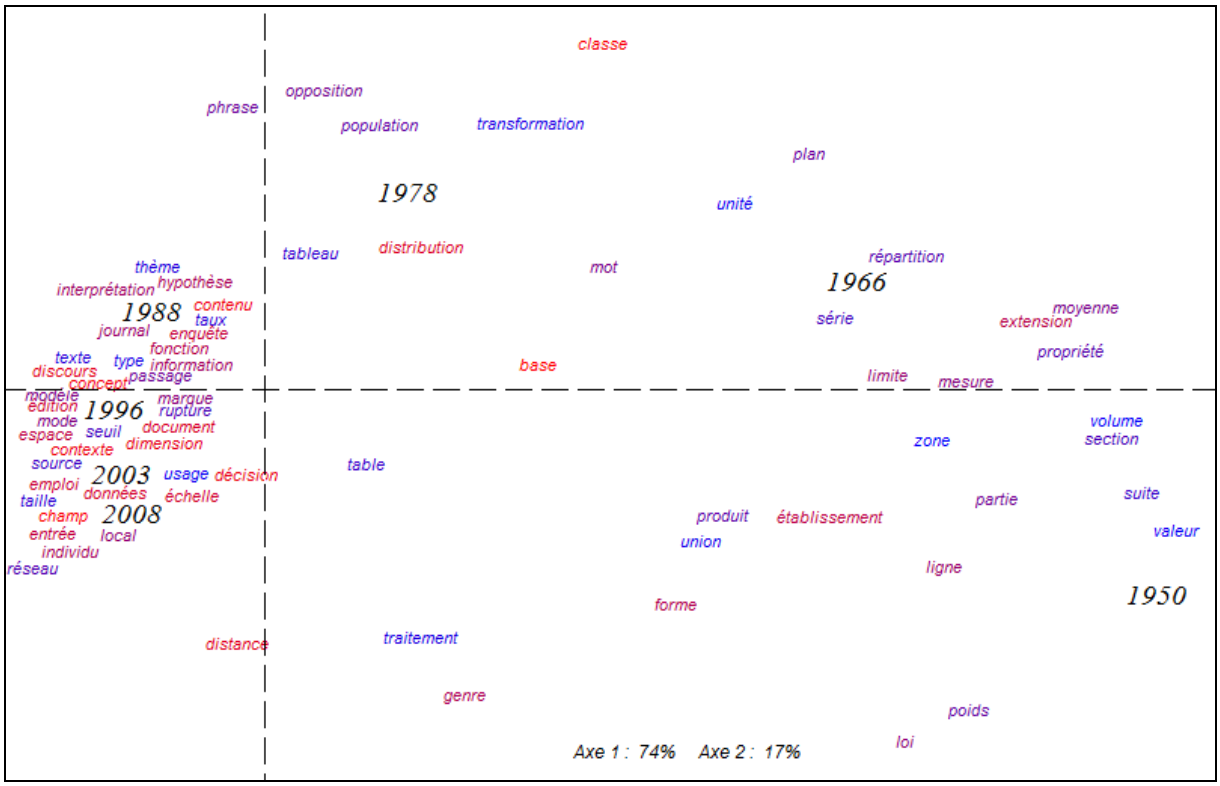


Figure 3. Analyse factorielle des 71 mots ayant plus d'un million d'occurrences dans *Goofre*

Les mots ou valeurs qui se concentrent dans la partie gauche réservée aux tranches récentes correspondent plutôt à ce qu'on attend de l'analyse des données textuelles. L'accent y est mis sur la variété des données qui peuvent être politiques, scientifiques, commerciales aussi bien que littéraires (*données, corpus, texte, chaîne, document, discours, enquête, entretiens, questionnaire, journal, oral, corpus, contexte*). Un soin particulier est porté à la préparation de ces données et à leur étiquetage (*édition, source, OCR, norme, normalisation, standard, désambiguïsation, catégorisation, transcription, annotation, étiquetage, filtrage, balise, TEI*). Si l'analyse factorielle est connue dans les périodes précédentes, son vocabulaire tend à se préciser et à se formaliser (*cluster, AFC, ACP, dimension, contribution, paramètre, vecteur*). On semble porter intérêt au contenu sémantique plutôt qu'aux questions morphologiques ou syntaxiques (*thésaurus, dictionnaire, concept, contenu, champ, motif, pôle, focus*). Enfin la recherche paraît s'intéresser moins aux mots individuels qu'à leur association (*cooccurrences, collocation, réseau, proximité, distance*)[19].

III – Perspectives

1 - S'agissant du français, on aurait pu s'adresser à la Bibliothèque Nationale Française qui est riche de 14 millions de documents dont 11 millions de livres sur le site de Tolbiac. Ce serait comparable à l'offre de *Google Books*, si l'accès était pareillement

[19] Il est toutefois difficile de donner foi à beaucoup de mots de la liste qui ont un sens commun à côté d'un sens spécialisé. Comme les textes saisis par *Google* ne sont pas limités au domaine de la lexicométrie, il y a chance que la place d'un mot polysémique soit déterminée par des emplois qui n'ont rien à faire avec la statistique. Ainsi la domination de Guiraud dans le graphique qui précède ne tient pas à ses seules interventions, vite abandonnées, dans le domaine statistique, mais à l'abondance de ses autres publications, purement linguistiques, et à la notoriété qu'il s'est ainsi acquise, hors de la statistique. Dans les données indifférenciées de Google, on peut craindre que le mot *champ* soit lié à l'agriculture, le mot *classe* à l'enseignement et le mot *ligne* à la pêche ou à quelque autre domaine. Il y a trop de bruit dans le mot *bruit*.

électronique. Malheureusement le nombre de documents accessibles par Internet, principalement dans la base *Gallica*, est loin d'atteindre ce chiffre. On a certes accès au catalogue et un choix sophistiqué des paramètres de métadonnées permet une sélection aussi fine que l'on veut. Mais le texte même de l'ouvrage échappe le plus souvent à l'internaute. Et quand le texte est transmis, il n'est lisible le plus souvent qu'en mode image. Certes la transcription en mode texte est parfois proposée. Mais c'est souvent le résultat brut du lecteur optique, avec la mention du taux d'erreur probable. Quand ce taux s'abaisse au dessous de 99%, cela signifie qu'un caractère sur cent est douteux ou fautif et donc qu'un mot sur vingt l'est aussi (la longueur moyenne du mot tournant autour de cinq lettres). Naturellement le pourcentage de réussite diminue à mesure qu'on s'éloigne dans le passé, les documents anciens souffrant des outrages du temps et proposant souvent des polices inhabituelles. Ces défauts sont communs à tous les corpus ou bases fondés sur la lecture automatique des documents, mais ils sont plus sensibles dans *Gallica* parce que les siècles passés y sont moins sous-représentés. Encore ne voit-on dans les résultats que les lectures correctes du mot proposé, parce qu'il ne vient pas à l'esprit de chercher des lectures erronées. Quoique *Gallica* soit une base déjà ancienne et largement antérieure à *Google Books*, son extension n'a pas la même ampleur et c'est ce qui limite son intérêt statistique. Le nombre de documents exploitables en mode texte est limité à 200000 alors que le rival américain en propose des millions. Et les informations proprement statistiques sont réduites au minimum, à la seule mention de la fréquence du mot cherché. Difficile avec si peu d'éléments d'établir une courbe, encore moins un tableau[20].

2 - En réalité les textes les plus fiables de *Gallica*, en dehors des plus récents transmis par les éditeurs sous forme numérique, sont ceux qui relèvent de l'héritage de *Frantext.* Ceux-là ne doivent rien à la lecture optique, dont l'invention par Ray Kurzweil en 1974 est postérieure à la saisie initiale, réalisée par des clavistes sur ruban perforé. Cette saisie manuelle, dûment revue et corrigée durant cinquante ans, a résisté à tous les changements de systèmes ou de supports, passant sans encombre du ruban perforé à la bande magnétique, puis au disque et enfin à toutes les variétés de mémoires disponibles aujourd'hui.

À cette fiabilité du texte, même lorsqu'il s'agit d'éditions anciennes, *Frantext* ajoute bien d'autres vertus : un certain équilibre entre les époques, ce qui autorise les comparaisons et donne une assise solide aux calculs d'évolution ; un large empan chronologique recouvrant cinq siècles de publication ; une homogénéité voulue des textes dont le choix obéit à des critères précis, quant au genre et au niveau de langue ; une constance dans les services offerts à la communauté scientifique, le même logiciel Stella étant maintenu sans changement sur Internet depuis vingt ans ; un accroissement modéré et un enrichissement contrôlé des données qui ménage la compatibilité avec l'exploitation antérieure. Bref dans le projet initial du *Trésor de la Langue Française* comme dans les produits dérivés que sont le *TLFI* et *Frantext* on observe une conception claire des objectifs et une définition précise des moyens qui ont fait de la réalisation française un modèle. Or une particularité de ce modèle nous intéresse : c'est la part qu'y prend la statistique. Dès l'origine le *TLF* réserve pour chaque article une rubrique finale où l'on relève la fréquence du mot dans l'ensemble du corpus mais aussi dans les sous-ensembles constitués autour de l'époque et du genre. Pendant toute la fabrication du dictionnaire les rédacteurs ont eu à leur disposition, outre les concordances, des informations chiffrées portant sur les fréquences et les cooccurrences[21] et s'appliquant aux

[20] Google Books n'offre pas davantage de précisions quantitatives et se contente pareillement d'indiquer le nombre de documents intéressés par l'objet de la requête. Mais ce nombre est d'un autre ordre de grandeur et l'exploitation proprement statistique y est assurée par un site dérivé Culturomics, qui n'a pas de correspondant à la BNF.

[21] Le nom de «groupes binaires » avait été donné à ces relevés cooccurrentiels, triés par catégories grammaticales. Bien avant la fin de la rédaction, et l'avènement d'Internet, les données quantitatives du *TLF*

graphies, aux lemmes, aux parties du discours, aux expressions, aux constructions. La plupart des fonctions documentaires et statistiques qui ont fait le succès de Frantext étaient déjà opérationnelles en mode local sur le site de Nancy ou même, en mode distribué, sur les réseaux nationaux (Transpac ou le minitel) qui ont précédé Internet. Elles trouvaient aussi leur application dans le cédérom *Discotext* produit et distribué en 1984. Mais c'est en 1998, avec Frantext sur Internet, que les recherches proprement statistiques se trouvent grandement facilitées. L'utilisation reste avant tout documentaire et les résultats chiffrés sont assez discrets. Car, pour ne pas effrayer les populations littéraires, *Frantext* se contente souvent de fournir des pourcentages ou fréquences relatives. Mais il est facile d'en déduire les fréquences réelles, de calculer des écarts, et de bâtir des courbes, des tableaux de distribution et des analyses multidimensionnelles, en opposant les textes les uns aux autres, ou les auteurs, ou les genres ou les époques. Le traitement statistique n'étant pas totalement pris en charge par *Frantext*, l'utilisateur a besoin de programmes complémentaires et spécialisés[22].

Pourtant quels que soient son renom et son mérite, Frantext a des perspectives limitées. Cinquante ans de passé font de l'ombre à son avenir. Cela tient pour une part à la timidité de son appareillage statistique : se contenter de distribuer du texte sous forme de listes de mots ou de séries de nombres, à l'heure où les images ont envahi Internet, c'est se priver le la lisibilité immédiate propre à la représentation graphique. Mais le handicap le plus lourd vient des données : on a loué leur fiabilité et leur homogénéité, mais elles ne représentent guère qu'un seul usage du français, celui de la langue relevée, littéraire et classique. C'est le français qu'on apprend dans les manuels de l'école ou qu'on lit dans les livres des bibliothèques. Ce n'est pas le français qu'on parle ou qu'on utilise dans la vie courante, dans les journaux et les medias. Il porte le témoignage de la culture, non le reflet de l'actualité. Certes le catalogue s'agrandit en s'ouvrant à la production récente: on en est présentement à 4000 références et 270 millions de mots. Mais la *BNF* pèse dix fois plus, *Google Books* mille fois plus et le rythme de croissance y est bien plus rapide. Enfin *Frantext* reste bridé par une convention, faite avec les éditeurs, qui limite la taille des extraits et contraint l'utilisateur à un abonnement préalable. Cette souscription peut se justifier s'il s'agit de communiquer un texte ou un extrait sous copyright. Mais on n'en voit nullement la légitimité juridique s'il s'agit d'informations quantitatives réalisées à partir du texte, qu'il soit ou non du domaine public. Enfin *Frantext* n'a pas retenu la solution intermédiaire qui consisterait à déporter le texte, au moins le texte libre de droits, et à l'offrir au téléchargement afin que l'utilisateur lui applique le traitement informatique et statistique de son choix. Cette fonction de distribution a été sous-traitée à un organisme annexe, le *CNRTL*, dont le catalogue actuel est très restreint.

3 – *Sketchengine* n'a qu'un point commun avec *Frantext* : la nécessité de souscrire un abonnement, mais ici pour des raisons de rentabilité et non de copyright[23]. Pour le reste tout les oppose. *Frantext* a des données propres. *Sketchengine* écume le *Web* et en tire sa subsistance et ses corpus. Le premier s'attache aux livres et aux textes complets, le second à des contextes courts, au mieux à des documents de peu d'ampleur. Le premier est diachronique, le second synchronique.

étaient accessibles aux chercheurs. Nous avons ainsi reçu dès les années 1970 les index et les dictionnaires de fréquences dont nous avions besoin pour nos travaux.

[22] Notre base *THIEF (Tools for Helping Interrogation and Exploitation of Frantext*) répond à ce besoin, en proposant la panoplie usuelle des outils statistiques et en les appliquant aux données de *Frantext*, qu'elles soient chargées préalablement ou téléchargées à la demande.

[23] Les droits sont modérés dans les deux cas, soit 41 euros pour *Frantext* et 52 £ pour *Sketchengine* ***(souscription académique pour un an).***

Comme beaucoup d'applications branchées sur Internet, *Sketchengine* explore les données du *WEB*, non pour en tirer des informations sur les nouveautés ou les tendances comme fait le *data mining*, mais pour en extraire de larges corpus représentatifs d'une langue donnée. Le point de départ est une liste de quelques centaines de mots de moyenne fréquence, que l'on considère comme la semence[24]. Pour engranger la récolte, un processus utilisant les moteurs Google, Bing ou Yahoo est réitéré des milliers de fois, à la recherche des sites et des pages où apparaissent simultanément trois mots tirés de la liste au hasard. À chaque fois on recueille les extraits en les empilant dans un corpus cumulatif avec les étiquettes correspondant aux métadonnées disponibles (à tout le moins l'adresse et le titre du site et la date de l'enregistrement). Vient ensuite un nettoyage des données qui expurge les doublons (grâce au logiciel « onion ») et élimine (grâce au logiciel « justext ») la gangue des balises et des hors-textes divers dont s'entoure abondamment le standard HTML (surtout au début et à la fin du document[25]). Divers filtres sont alors appliqués pour extraire le contenu intéressant : pour être retenu le document doit remplir plusieurs conditions : être de type TXT ou HTML, avoir une longueur suffisante (au moins 500 mots) et contenir un dosage minimum de mots grammaticaux sans lesquels un texte ne saurait être estimé normal. Ce contrôle automatique exercé sur des critères simples, sinon toujours pertinents, rejette beaucoup de sites et beaucoup de données dans les sites acceptés : le rapport entre ce qui est retenu et ce qui est examiné se situe entre 1/10 et 1/1000. De toute façon l'exploitation d'un site est limitée en quantité pour assurer au corpus plus de diversité. L'automate peut ainsi empiler jusqu'à 1 milliard de mots par jour. Les données textuelles recueillies reçoivent alors des traitements linguistiques pour en assurer la lemmatisation (TreeTagger est utilisé pour les langues occidentales) et toute une série d'opérations statistiques pour en permettre une consultation sophistiquée.

La demande la plus simple est documentaire, c'est la recherche de concordances. Le concordancier de *Sketchengine* restitue le contexte (ligne ou phrase) de tout objet qu'on lui propose : graphie, lemme ou expression avec filtres divers. Il peut aussi examiner l'environnement du mot-pôle et classer les mots cooccurrents selon le type de rapport qu'ils entretiennent avec celui-ci (objet, sujet, qualification, coordination, attaches prépositionnelles, etc) et selon la force de l'aimantation exercée par le pôle. Ainsi peut-on faire une monographie sur le mot *« samedi »* et observer que les mots qui l'accompagnent le plus souvent sont *dimanche, dernier , prochain , pluvieux* et *ensoleillé,* exprimant ainsi le rôle majeur que joue le week end dans les préoccupations des gens. En ajoutant à l'étude le profil des autres jours de la semaine, on obtient une typologie sociologique du rythme hebdomadaire, ou du rythme saisonnier si l'on s'attache aux mois de l'année. Des sondages de toutes sortes s'offrent à la recherche, quand on met en balance des valeurs comme la liberté, la justice, l'égalité, la communauté, ou quand on répertorie les péchés capitaux.

À titre d'exemple, les hommes politiques de la scène nationale pourraient observer sans grands frais l'image qui est la leur dans Internet et dont rend compte la figure 4. Il s'agit d'une analyse factorielle qui réunit les contextes où l'on rencontre le nom des acteurs majeurs de la cinquième république. Le sous-corpus extrait du corpus *FrTenTen12*[26] de Sketchengine

---

[24] Pour la plupart des langues, la semence est extraite du corpus de Wikipedia. Quand tous les articles que Wikipedia consacre à une langue donnée ont été réunis, on en tire un dictionnaire des fréquences, dont on élimine les 1000 mots les plus fréquents. Les 5000 unités qui viennent ensuite sont considérées comme appartenant aux moyennes fréquences et sont considérées comme la semence. Généralement il suffit d'un triplet (une combinaison de 3 mots sur 5000) pour explorer le Web. Mais suivant la taille de la semence, le « triplet » peut se réduire à 2 ou s'étendre à 4.

[25] On coupe le début et la fin pour ne retenir que le corps, présenté comme un poulet sans la tête et les pattes.

[26] Ce corpus, comme quatorze autres, consacrés aux principales langues du monde, doit son nom générique *TenTen* à la dimension de dix milliards de mots adoptée pour chacun (soit 10 puissance 10). Les deux premières lettres désignent la langue et les deux derniers chiffres l'année où le corpus a été établi. Ajoutons que d'autres

est le fruit de 37 consultations, dont chacune est consacrée à un homme politique (président, ministre ou chef de parti) et relève de façon aléatoire 5000 concordances-citations[27]. Le calcul porte sur l'environnement des 279 substantifs les plus fréquents du corpus (où figurent par définition les noms des hommes politiques concernés, chacun ayant au moins 5000 occurrences). Le tableau, carré et symétrique, soumis à l'analyse croise ces mêmes mots en ligne et en colonne (à l'intersection de la ligne i et de la colonne j on a le nombre de fois où mot i cooccurre avec le mot j). S'agissant d'une enquête sur des hommes liés au temps historique, on se sera pas surpris de voir la chronologie régner sur le premier facteur et séparer les anciens des modernes. La dérive du temps, évidente dans les noms propres, s'observe aussi parmi les noms communs. Ceux qu'on trouve à gauche appartiennent à l'actualité politique[28] des années 2000 (c'est-à-dire des batailles électorales et particulièrement des présidentielles de 2007 et 2012). On y voit s'affronter les candidats au milieu des *sondages*, des *campagnes* et des *votes (débat, déclaration, programme, émission, media, opinion, parti, soutien, militant, candidature, primaire, présidentielle, tour, résultat, victoire).* Avec le recul du temps, la confrontation est plus apaisée dans la moitié opposée. Mort ou retraite, certains des acteurs ont quitté la scène politique. On y voit leur œuvre plutôt que leur ambition et l'histoire plutôt que l'actualité.

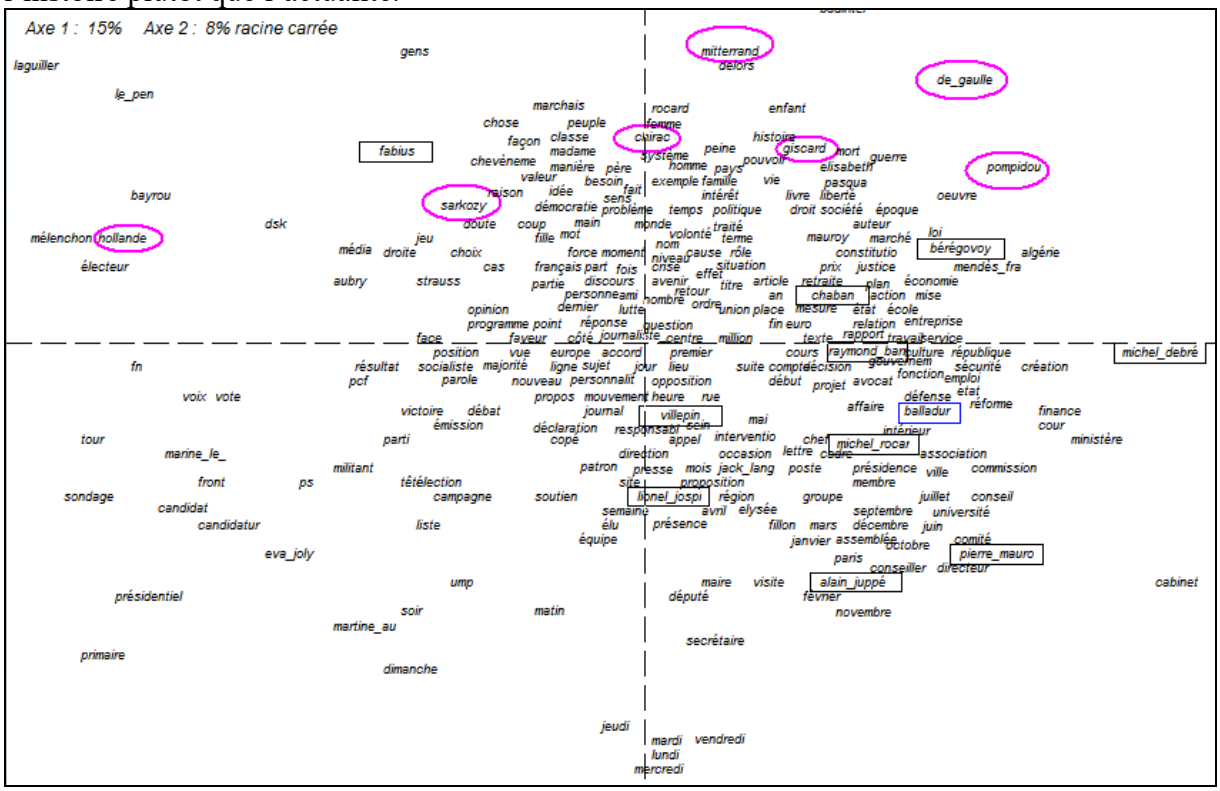


Figure 4. Analyse factorielle des corrélats dans le corpus des hommes politiques

Quant au second facteur qui oppose le haut et le bas de la figure, on pourrait s'attendre à y trouver le clivage gauche-droite. Et ce serait peut-être le cas si les discours des hommes

---

corpus de moindre dimension sont aussi disponibles pour d'autres langues et qu'on peut aussi constituer directement son propre corpus à partir d'Internet en utilisant les programmes et les *API* que *Sketchengine* met à la disposition de l'utilisateur.

[27] Le corpus ainsi constitué est gros de 7 millions d'occurrences. Une fois téléchargées les données ont été exploitées avec notre logiciel *Hyperbase.*

[28] Rappelons que les corpus de *Stetchengine* sont synchroniques, étant établis à partir des ressources qu'offre Internet à un instant donné. On ne préjuge pas de la date où les extraits ont été écrits et mis sur le réseau. Mais étant donné l'explosion exponentielle d'Internet, la plupart des textes recueillis sont de peu antérieurs à la collecte, et représentent une actualité circonscrite à la première décennie du siècle.

politiques étaient directement en cause. Or ce qu'ils disent n'importe pas, mais seulement ce qu'on dit d'eux. Et dans les propos des tiers, où la comparaison est fréquente, la droite et la gauche peuvent cohabiter, et Mitterrand voisiner avec de Gaulle – ce qu'on observe en effet dans le graphique. L'opinion tend à classer – et à confronter - les gens selon leur rang. Les présidents de la république occupent le haut du pavé, dans la moitié supérieure, où ils se donnent la main dans l'ordre à peu près chronologique[29]. Les premiers ministres sont relégués dans la moitié inférieure, où Balladur voisine avec Rocard, Jospin avec Villepin, Mauroy avec Juppé. Tandis que les présidents ont le privilège des valeurs et des objectifs nobles de la politique (*peuple, famille, homme, femme, pays, société, valeur, liberté, justice, loi, démocratie, politique, guerre, mort*), les premiers ministres n'ont droit qu'à la gestion administrative des affaires (*ministère, cabinet, comité, conseil, commission, directeur, secrétaire, conseiller, assemblée, groupe, chef, membre, député, maire, poste, finance, université, presse, fonction, réforme, emploi*, etc.).

4 – De telles analyses cooccurrentielles ne sont pas possibles sur le site *Culturomics* que nous avons rencontré précédemment à l'occasion de l'évolution de la lexicométrie (figures 2 et 3). Car si le texte d'une citation est nominalement accessible dans *Google Books*, il disparaît dans le produit dérivé *Culturomics* en ne laissant que des traces de la lecture : des *ngrammes* ou tronçons de texte dont le plus long ne dépasse pas cinq mots. Mais l'entreprise de *Google* offre des avantages appréciables en qualité comme en quantité. En extension elle l'emporte largement dans le domaine français, avec un volume dix fois supérieur : près de 100 milliards de mots dans la version de 2012[30]. Et la couverture chronologique, si rétrécie et incontrôlable dans *Sketchengine*, s'étend ici sur des siècles, en ouvrant des perspectives sur l'histoire des mots et des réalités dont les mots portent témoignage. La qualité des sources est aussi à l'avantage de *Google*. Internet est un fourre-tout où la parole est libre et les contrôles impraticables. Le pompage auquel se livre le « spider » de *Sketchengine* a beau multiplier les filtres, il reste insensible aux barbarismes qui se multiplient sur les réseaux sociaux. Comme *Google Books* ne s'intéresse qu'aux livres, de même que la *B.N.* et *Frantext*, on accède par là même à un certain niveau de langue et de culture, que *Facebook* ne peut garantir. Un simple sondage en donne la mesure : le rapport de *fesait* à *faisait* est de 388213/1474862, soit 2,6% dans *Sketchengine* alors qu'il s'abaisse à 55584/9494928, soit 0,6 % dans *Culturomics*[31]. Sans viser ni atteindre la langue relevée et littéraire de *Frantext*, *Google Books* accepte le tout-venant de l'édition, surtout dans les publications de l'actualité technique, sociale ou médiatique. Mais la barrière du livre imprimé le protège contre la logorrhée envahissante et insignifiante qui se répand sur les blogs et les réseaux sociaux.

Jean Véronis, qui vient de nous quitter, ne cachait pas son enthousiasme à la naissance de *Culturomics*, à Noël 2010. Il a dû saluer pareillement la version 2012 qui corrige certains défauts de la version 2009 et en multiplie la puissance et la souplesse. La sélection ne se

---

[29] C'est là aussi qu'on trouve les chefs de parti, surtout les irréductibles : Marchais, Le Pen, Bayrou, Mélenchon, Laguiller.

[30] De 2009 à 2012, la hauteur a doublé, pour le français comme pour les autres langues On en est à 89 milliards de mots 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 corpus sont disponibles, dont nous ne dirons rien faute de connaissances et de clavier : le russe, le chinois et l'hébreu.

[31] Il est vrai qu'en retour *Google Books* est livré aux fautes de lecture, inévitables quand on utilise le scanner. Mais ces erreurs, étant aléatoires, introduisent du bruit plutôt que des déviations. Quant aux erreurs de codage elles sont communes à tous les lemmatiseurs, ceux de *Google Books* comme ceux de *Sketchengine*. Sur la mesure et l'étendue de ces erreurs dans les *big data*, voir notre article « Data hygiénisme : nettoyer les données de Google », in *Documentaliste. Sciences de l'information*, n° 1, vol 51, mars 2014, p. 12-15

limite plus aux graphies ou groupes de mots. Elle s'exerce maintenant sur les lemmes (par exemple *faire_INF*, pour solliciter le détail des formes conjuguées de ce verbe), sur les catégories seules *(_DET_)* ou suffixées (*est_VERB*), sur les jokers (symbole *), et sur le choix des corpus (symbole :). 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. Les auteurs de *Culturomics* ont donc proposé une API téléchargeable[32] qui pour un mot donné distribue les 201 pourcentages observés le long de la chronologie, de 1800 à 2000. Mieux même : les données brutes qui servent à la fabrication des tableaux et des courbes ont été livrés au libre téléchargement, dont nous avons profité pour constituer une base offrant en local l'exploitation des unigrams (ou mots individuels) du domaine français. À titre d'illustration la figure 5 livre une synthèse sur l'évolution syntaxique de la phrase française, où le verbe et ses acolytes, pronoms, adverbes et conjonctions, perdent du terrain, au bénéfice des classes liées au nom : substantifs, adjectifs et prépositions. Or cette évolution n'est pas propre au français : on la retrouve pour la même période dans les autres langues occidentales. Cette tendance quoique unanime ne laisse pas d'être un peu suspecte.

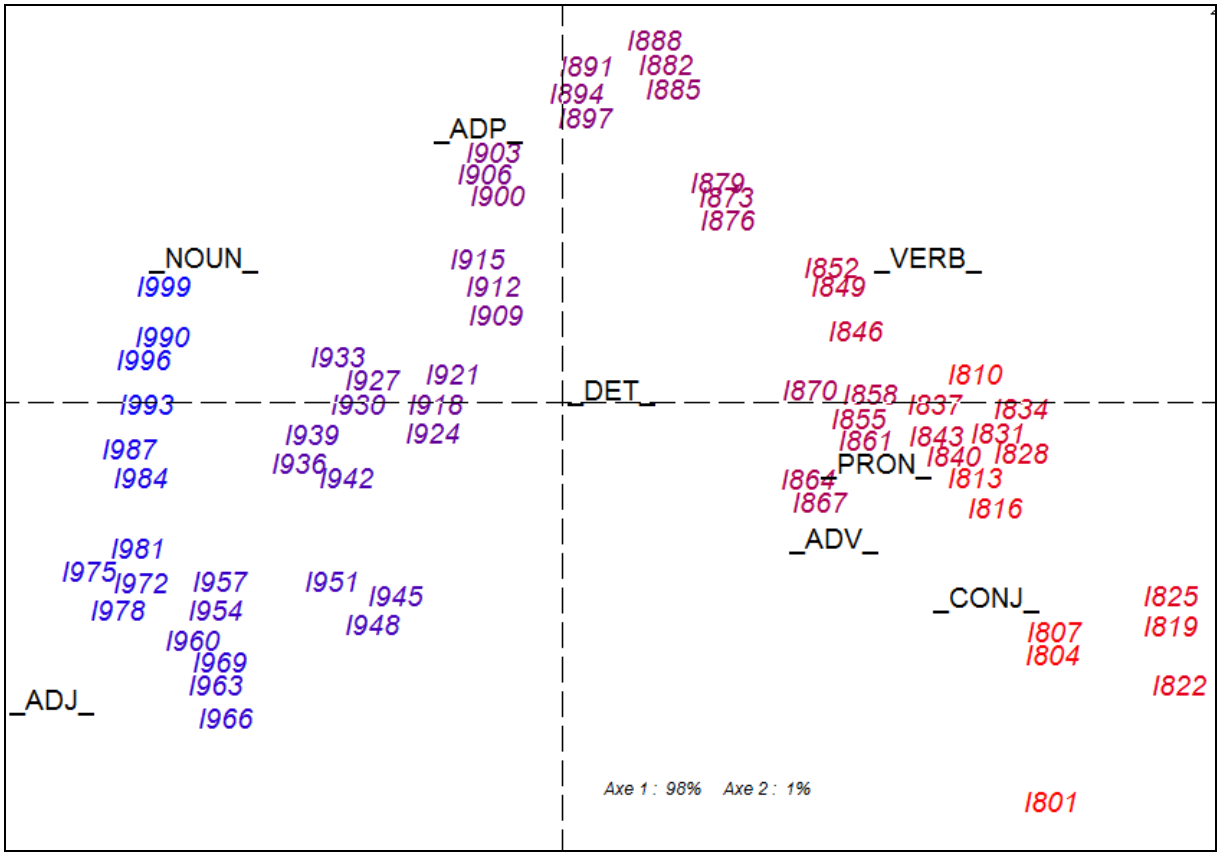


Figure 5. Le dosage variable des parties du discours (24 milliards de substantifs contre 10 milliards de verbes)

[32] 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 (CNRS, Nice), s'est chargé des rectifications nécessaires et pourra assurer le maintien de ce service.

On a le soupçon qu'il s'agit moins d'un changement dans l'usage du français, que d'un changement dans la composition du corpus. Les textes récents étant les plus nombreux et les plus techniques ont donné aux tranches modernes une coloration abstraite et impersonnelle où l'information passe par les catégories nominales, tandis que les époques plus anciennes puisaient sur les rayons des bibliothèques les livres qui avaient subsisté et qui appartenaient plus souvent à la littérature qu'à la science et à la technique, lesquelles se démodent plus vite. Or le verbe est plus présent dans le discours littéraire, les pronoms et le dialogue y sont plus vivants et plus personnels. Le genre aurait donc créé une hétérogénéité du corpus, donnant l'illusion d'une évolution[33].

On voit que la taille gigantesque des données et des résultats statistiques peut donner le vertige au jugement, mais le doute subsiste dans les failles de la muraille. Il s'accroît d'autant plus que l'obscurité règne sur la composition du corpus. On voit aussi que l'évidence aveuglante d'un résultat graphique ne doit pas en imposer à la raison. L'effet peut être massif, et la cause obscure, laissant à l'interprétation des chances incertaines et à la lexicométrie une perspective fuyante.

[33] L'exploitation des données de Culturomics a été facilitée, pour le domaine français, dans une base que nous avons nommée GOOFRE2 et qui est téléchargeable sur le site http://logometrie.unice.fr (et aussi sur le site ftp://ancilla.unice.fr). Ce n'est pas le lieu 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 5), 3 – exploitation entièrement locale pour les unigrammes du corpus français.