Il y a cinquante ans

Etienne Brunet

Il y a cinquante ans.

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

Dans un siècle ou deux, quand l'espérance de vie aura considérablement augmenté, les gens fêteront peut-être leurs retrouvailles au bout de cent ans. Aujourd'hui les anniversaires les plus courus se contentent de cinquante ans. C'est l'intervalle normal qui sépare l'entrée en classe de sixième de l'entrée dans la classe des retraités. La mémoire s'active alors soudainement lorsque la vie active fléchit et les photos de classe qui dormaient dans la poussière se multiplient dans les réseaux sociaux d'internet. L'histoire individuelle trouve sa garantie et son amplification dans le partage des souvenirs, comme l'histoire familiale dans la recherche généalogique.

Il en est ainsi des institutions, des collectivités, des sociétés savantes, des entreprises collectives qui aiment à se retourner, cinquante ans après (quand la chance les a maintenues en vie jusque là), pour contempler le chemin parcouru et teinter de nostalgie la satisfaction d'avoir survécu et donc plus ou moins réussi.

Or cinquante ans après la naissance de la statistique linguistique, laquelle peut être située , en France tout au moins, au début des années soixante, on a demandé au signataire de ces lignes d'écrire l'histoire de cette discipline. Un véritable historien jouit normalement de deux privilèges: n'être pas mêlé personnellement aux faits rapportés, et circonscrire un temps et un lieu qui échappent à l'actualité. Ces avantages m'étant refusés, je ne puis guère que proposer mon témoignage, particulièrement à l'époque des débuts, en limitant mon propos au domaine français et aux chercheurs personnellement rencontrés.

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.[1] 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.[2] Dès 1959 paraissait *La Machine à traduire* de E. Delavenay qui rendait compte des travaux des uns et des autres.[3]

---

[1] 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 Grammaticals Praxis* (ed. P.Desmet, L . Jooken, P. Schmitter, P. Swiggers) Louvain/Paris, Peeters, p. 369-394.

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

[3] 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.

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[4], 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.

La statistique linguistique

1 – C'est Pierre Guiraud qui imagine le premier le profit statistique qu'on pourrait tirer des données engrangées. Dès 1954 il publie une *Bibliographie de la statistique linguistique*. Et en 1959 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[5] ». Les idées de Guiraud ont pourtant germé alentour , à Paris autour de Wagner ou Georges Gougenheim, à Liège autour de 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[6].

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

---

[4] 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.

[5] 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.

[6] 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.

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.[7]

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 »[8] sur des générations de lexicologues vient de son manuel de 1968[9]. 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 mutidimensionnelles et la disponibilité des ressources informatiques, un autre catéchisme devenait nécessaire et ce sont deux mathématiciens, Lebart et Salem, qui prirent le relais de Muller[10]. 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. Le présent ouvrage est peut-être ce troisième manuel qui complèterait ou corrigerait les deux précédents. Mais étant à l'intérieur et à l'entrée, nous n'avons pas les moyens de le survoler et de l'analyser. Pourtant l'idée nous est venue d'examiner la liste des mots retenus pour ce dictionnaire. 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). Comme pareil glossaire n'existe pas et qu'il serait arbitraire et artificiel de le reconstituer, on se contentera de la nomenclature du présent ouvrage en la projetant sur un immense corpus issu de Google Books et gros de 44 milliards de mots du domaine

---

[7] 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.

[8] 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.

[9] 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.

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

français. 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[11].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[12]. 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[13].

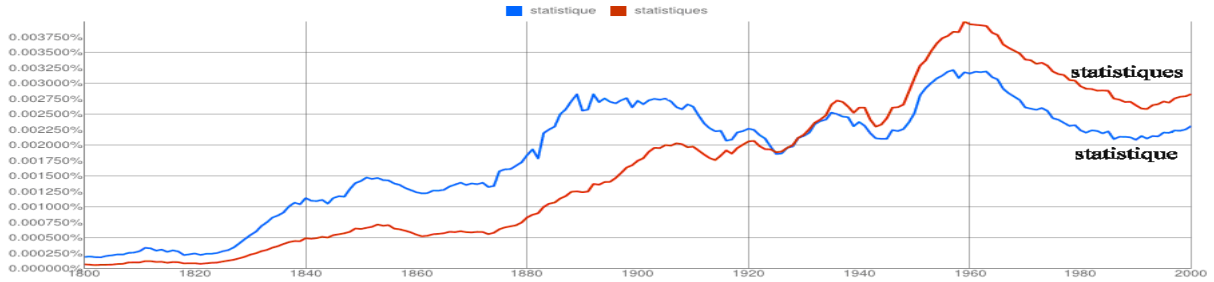


Figure 1. L'évolution de la *statistique* de 1800 à 2000

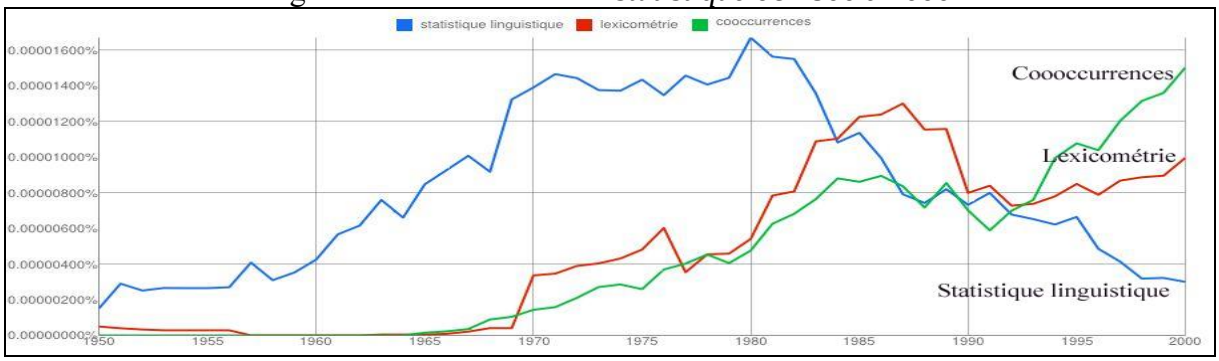
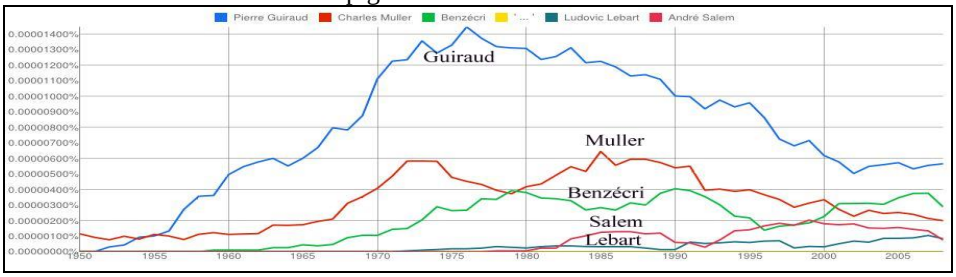


Figure 2. *Statistique linguistique* et *lexicométrie* de 1950 à 2000

On ne saurait 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 240 unités représentatives de la discipline. On a donc extrait de Google ses fichiers de données pour en faire une base exploitable, apte à fournir des synthèses. Cette base porte le nom de *Goofre* pour indiquer tout à la fois sa filiation et son immensité. La figure 3 restitue le

[11] 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.

[12] 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.



[13] On a proposé d'élargir le champ de la *lexicométrie* en abandonnant son radical, trop limité au lexique. Mais si la *stylométrie* se maintient, à un faible étiage, ni la *textométrie* ni la *logométrie* ne se sont imposées. En revanche on assiste au démarrage prometteur de la *proxémie* qui a l'avantage de mettre l'accent sur le voisinage des mots dans le discours.

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. Car la plupart des 240 mots de la liste se sont agglutinés dans les tranches récentes, de 1988 à 2008. Cela est de bon augure pour la modernité et l'actualité des choix, mais cela rendait le graphique illisible.

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.

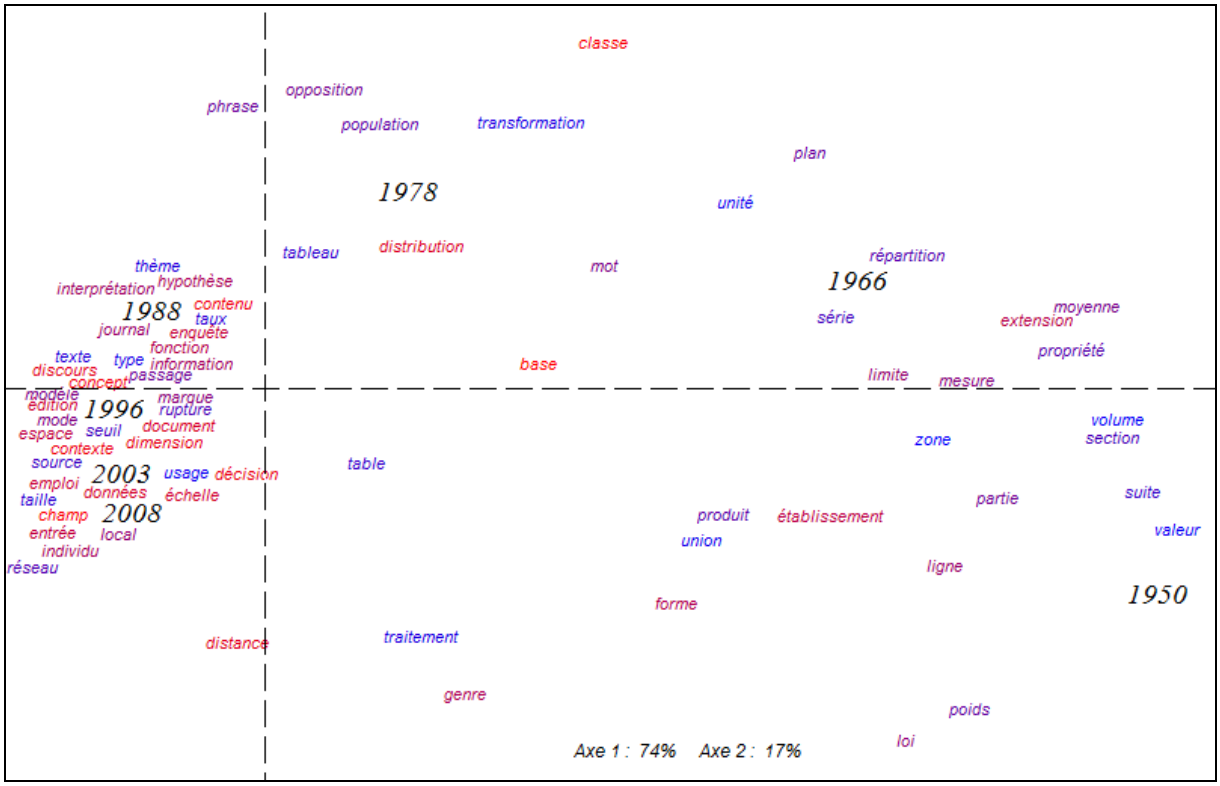


Figure 3. Analyse factorielle des 71 mots ayant plus d'un million d'occurrences dans *Goofre*

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*)[14].

III – Bilan

Puisque le présent ouvrage est une rencontre où le témoin passe du passé à l'avenir et du bilan au projet, il convient de s'interroger sur les succès et les échecs de la discipline. Celle-ci s'est imposée dans la plupart des sciences sociales : sociologie, psychologie, géographie humaine, économie, sciences politiques. Mais elle semble piétiner dans sa discipline initiale. Les compteux de mot, pour reprendre une expression qui a cours au Québec, n'ont pas bonne presse, ni chez les linguistes, ni chez les littéraires. Les premiers s'appuient sur des exempliers et trouvent la garantie dans l'attestation d'un fait de langue et non dans sa fréquence. Les seconds se fondent sur les sources, les lectures, la culture et cherchent la garantie dans l'accord avec les jugements d'autrui. En réalité les réticences s'adressent moins à l'informatique qu'à la statistique. Les vertus domestiques de l'ordinateur ont fini par être reconnues : rares sont les critiques ou écrivains qui écrivent à la main ou qui utilisent encore une machine à écrire. Et depuis qu'Internet a gagné en puissance, en extension et en rapidité, les mêmes littéraires sont sensibles aux facilités documentaires que permet le réseau mondial. Ce n'est plus seulement la référence d'un livre qu'on trouve dans le catalogue monstrueux de Google qui tend à se donner l'image de la bibliothèque universelle imaginée par Borges. C'est aussi à l'information primaire, au contenu d'un article ou d'un document, qu'Internet donne un accès immédiat. Plus besoin d'attendre, plus besoin de commander, ni même de payer. Certes les moteurs de recherche ne sont pas des critiques avertis : il mêlent l'ivraie au bon grain. Mais l'indice de notoriété qu'ils utilisent est assez souvent suffisant pour faire apparaître le document pertinent parmi d'autres qui le sont moins. Les détracteurs littéraires de l'informatique et d'Internet sont ainsi devenus moins virulents. Ils acceptent les services auxilliaires et ancillaires de la machine, qui réunit les matériaux sur la table de travail. Ils se réservent la part noble de la sélection, de l'exploitation et de l'interprétation. C'est là ce qui compte.

Et pour cela, croient-ils, nul besoin de ceux qui comptent. On ne compte, disent-ils, que ce qui est quantifiable : les prix et les produits, les carottes et les avions. On ne compte pas les idées… Pourtant la démocratie s'exerce en comptant les votes, les opinions, les hommes. Et le jugement littéraire lui-même n'est pas tout à fait dépourvu de compteurs inconscients : beaucoup des jugements qu'on croit qualitatifs sont inspirés par une statistique implicite qui n'avoue pas son nom et qui autorise l'emploi des mots *typique, spécifique, caractéristique*, si fréquents sous la plume de la critique lorsqu'elle analyse un auteur, un genre ou une époque. Le mot *fréquent* lui-même relève de cette approche, comme aussi *rare, original, banal, courant, cliché, surprise, rupture*. Les littéraires parlent d'horizon d'attente quand les statisticiens parlent d'espérance mathématique. L'espérance des uns et l'attente des autres, ce n'est qu'une prévision fondée sur les observations répétées que la conscience enregistre.

Cette réticence des milieux littéraires est d'autant plus regrettable que les données textuelles ont des propriétés très avantageuses. Ces données sont d'abord bon marché ; il suffit d'un traitement de texte pour les enregistrer, ou d'un scanner, ou tout simplement d'une liaison à Internet où les textes foisonnent.. Elles sont abondantes ; or la statistique se plaît

[14] 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 la mot *ligne* à la pêche ou à quelque autre domaine. Il y a trop de bruit dans le mot *bruit*.

dans les grands nombres et ses conclusions sont moins sûres quand les observations sont en nombre limité parce qu'elles sont chères, ce qui est le cas des enquêtes et des sondages, ou que la nature des choses le veut ainsi, comme il arrive en médecine. Les données textuelles sont faciles à contrôler, à reproduire, à transmettre, à corriger, et se prêtent volontiers à la répétition et aux variations des expériences. Elles sont exemptes de filtrage sélectif et subjectif, et la conscience du chercheur intervient peu à l'entrée. Enfin et surtout les textes sont immédiatement interprétables, sans le truchement des machines. La lecture et la connaissance du texte sont des garants contre les calculs aberrants et les manœuvres avortées. La conscience a au moins une idée de l'ordre de grandeur des résultats attendus et peut les rejeter s'ils sont exagérément erronés ou délibérément triviaux, alors que d'autres disciplines travaillent sans vision directe et se trouvent liées aux instruments, sans possibilité de récuser leur témoignage.

Or ce qui est un avantage est aussi un inconvénient. Puisqu'en matière littéraire ou linguistique on peut se débrouiller sans appareillage, pourquoi s'encombrer de méthodes indirectes et grossières qui obscurcissent l'évidence ou brutalisent les nuances. Pour beaucoup d'esprits, le refus d'accueillir les méthodes quantitatives est une manière de réserver un espace de liberté pour la conscience individuelle. Qu'il y ait au moins un domaine préservé, une réserve naturelle comme il y des parcs du même nom pour la sauvegarde des paysages, un refuge contre l'invasion technologique, où puissent vivre et subsister les disciplines menacées : l'art, le langage, la philosophie, la religion, la musique, la cuisine…