Web et la statistique L'exemple du mot Rome

Étienne Brunet

Web et la statistique
L'exemple du mot *Rome*

Etienne Brunet

-I-

Internet est devenu un énorme magasin d'informations et les chercheurs en sciences humaines peuvent y trouver les statistiques dont ils ont besoin, même si celles qui ont un intérêt stratégique et commercial ne sont pas nécessairement gratuites. Il peut exister des producteurs dispendieux qui sèment à tout vent et offrent leurs données sur le *Web*. Mais plus nombreux sont les avaricieux qui protègent jalousement leur secret. Entre ces deux démarches, il y a place pour un marché raisonnable des données statistiques et leur mise à disposition sur le réseau. Bien des organismes nationaux ou internationaux ont vocation pour être des observatoires de la vie économique, sociale et culturelle et jouer le rôle de fournisseur de relevés chiffrés. Jusqu'ici, ils l'ont fait surtout sous la forme de rapports épais, dont la publication sur papier arrivait trop tard et trop tôt, trop tard pour les faits dépassés dont ils rendaient compte et trop tôt pour l'actualité présente dont ils ne disaient rien. Le livre, ralenti par les délais de la publication et les lourdeurs de la commercialisation, impose toujours une distance entre les faits recensés et leur divulgation. Cela peut contenter l'historien, qui ne travaille pas dans l'urgence, mais le sociologue et l'économiste ne peuvent s'en satisfaire puisque leur objet d'étude est dans l'actualité. Le réseau télématique leur convient mieux parce qu'il assure aux données de base une diffusion très large, immédiate et peu coûteuse.

N'étant ni économiste, ni sociologue, je ne puis guère aborder ce sujet, sinon de manière marginale, en me contentant d'évoquer les données littéraires. Leur importance commerciale présumée n'est pas si négligeable, si l'on en juge par le soin jaloux dont les maisons d'édition protègent leur copyright. Certains textes tombés dans le domaine public sont disponibles sur le réseau, sans constituer, hélas, un véritable catalogue, même pour les langues les mieux représentées. S'il est à la portée de quiconque de se procurer, par les canaux électroniques les plus divers, l'œuvre de Shakespeare, ou certaines versions de la Bible, certains

titres sont moins faciles à obtenir, même parmi les plus connus. Et dans la plupart des cas, on a accès à des échantillons, voire à des traductions, rarement à l'intégrale. Il semble que le catalogue offert par quelques serveurs ait été constitué sans fil directeur, au gré des contributeurs et au hasard des cessions. Même le projet Gutenberg, qui est le plus ambitieux, n'offre qu'une liste dépareillée de quelques centaines de titres. C'est loin de représenter la littérature anglaise. Les autres langues sont moins bien traitées encore, et le catalogue du projet ABU ne propose guère qu'un échantillon symbolique de textes français. Le texte brut ne représente pas d'ailleurs un marché plausible, s'il s'agit seulement de lecture. Il est tellement moins confortable de lire un roman sur écran que dans un livre. L'intérêt du support électronique n'existe que pour les chercheurs désireux de soumettre le texte aux recherches documentaires et à l'analyse quantitative. Cet intérêt trouve souvent des occasions de se manifester et rarement de se contenter.

-II-

Mais, à défaut de données textuelles proprement dites, on peut considérer que les pages du *Web* constituent un immense texte discontinu, qu'on pourrait soumettre à une étude de type sociologique, en isolant les variables géographiques ou culturelles qui caractérisent le serveur et l'émetteur, et les variables internes qui définissent le contenu ou la forme des pages d'information. *Internet* est en soi un fait de société nouveau qui suscite l'interrogation des observateurs et envahit la presse à grand tirage. Sujet à la mode, il fait l'objet de multiples articles de vulgarisation et plusieurs revues spécialisées lui sont consacrées exclusivement, sans parler des ouvrages qui se multiplient sur le même sujet. Que ce succès soit éphémère ou durable, cela mériterait dans tous les cas l'analyse statistique, laquelle aurait au moins l'avantage initial d'évoluer dans les grands nombres. Car c'est une immense forêt vierge qui s'offre à l'exploration. Certes les milliards de mots[1] qu'on y trouve constituent une masse informe, mouvante, éparpillée aux quatre coins du monde et difficile à appréhender, même si l'usage dominant de l'anglais lui donne

1 *Lycos*, qui est le plus puissant moteur de recherche branché sur Internet, avoue explorer près de 14 millions de documents. Il suffit que chacun d'entre eux compte une centaine de mots pour que le milliard de mots soit dépassé pour l'ensemble, ce qui constitue un record dans le domaine du texte intégral. Mais si l'on évalue en octets la masse d'information gérée par le réseau Internet, la taille devient démesurée, vu le nombre des images.

une certaine homogénéité. Les échanges incessants qui s'y perpétuent font penser à la mécanique des fluides. Mais avec les méthodes convenables, raisonnées ou aléatoires, avec des automates installés à la surface du Net et procédant par sondage, ne pourrait-on pas découvrir, dans ce flux, des remous, des courants, des marées ?

Choisissons par exemple la représentation des villes. Comment sont-elles représentées dans *Internet* ? Quelle est leur importance relative ? Leur poids dans la communication est-il en rapport avec le poids de leur population, de leur économie ou de leur célébrité ? Et puisque nous sommes dans la Ville Éternelle, voyons si l'actualité s'accorde avec l'éternité, en consultant *Lycos* (graphique 1). Le « loup » électronique a tôt fait de retrouver la louve romaine et en signale 7 811 traces dans le monde *Internet*, dont 5 445 citations sous la forme internationale ROME et 2 366 sous la forme locale ROMA. Les célébrités italiennes changent leur nom lorsqu'elles passent la frontière et c'est le cas de FLORENCE (3 006 citations contre 1024 à FIRENZE), de VENISE (VENICE 2 715, VENEZIA 677, VENISE 69), de MILAN (2 609 contre 2 132 à MILANO).

Accessoirement, on pourrait mesurer par ce moyen la notoriété extérieure des cités, en établissant le rapport entre l'appellation étrangère et le nom d'origine et l'on observerait par ce moyen que le rayonnement international de Rome l'emporte sur celui de Milan, ce qui se vérifie aussi pour Venise et Florence. L'étude des doublets, italiens ou non, pourrait aussi servir à apprécier le poids d'une langue dans les communications mondiales. Si l'univocité de certaines graphies comme BERLIN ou MADRID laisse le jugement en suspens, il n'en va pas de même avec LONDRES (la forme francisée n'apparaît que 218 fois contre 26 673 pour la graphie anglaise LONDON). Ce simple rapport 218/26 673 mesure approximativement la part de marché de la langue française dans les transactions d'*Internet*, qui n'est guère que de 0,8%. Celui de la langue italienne est encore moins favorable, du moins si l'on se fonde sur le quotient PARIGI/PARIS, qui est de 92/17 455, soit 0,5% [2].

2 Ce ne sont là que des indications fragiles, quoique vraisemblables. En pareille matière, il faut éviter le biais des villes nationales. Car les documents écrits en italien citent évidemment plus souvent que les autres langues les villes italiennes. Le rapport, d'ailleurs très variable, VENEZIA/VENICE ou MILANO/MILAN ou ROMA/ROME peut mesurer la notoriété d'un site, comme on l'a vu, mais non l'importance relative de la langue utilisée dans les documents Internet. Ce même rapport serait pareillement faussé si l'on partait d'exemples francophones, comme BRUXELLES (1075 contre BRUSSELS 3672) ou GENEVE (876 contre GENEVA 7004).

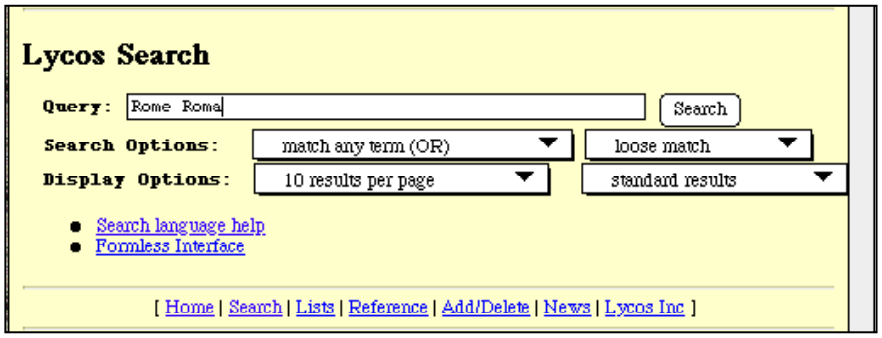


Figure 1. L'interrogation de *Lycos* (14 millions de documents)

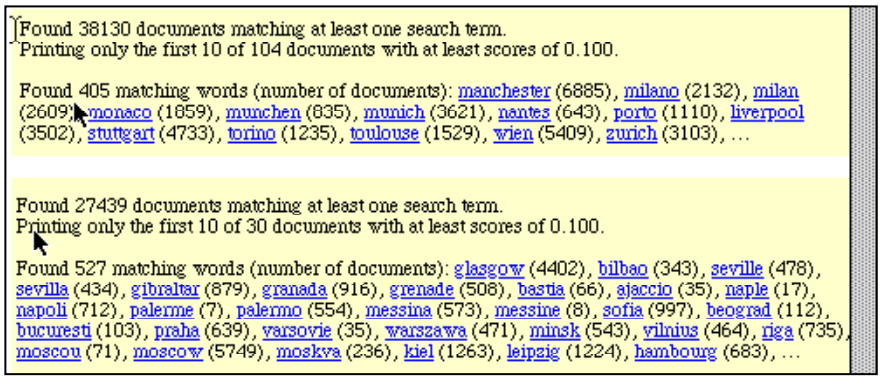


Figure 2. Quelques réponses de *Lycos*

Mais en réduisant les doublets et en neutralisant les variations orthographiques (toutes les appellations d'un même site étant regroupées), on voudrait dresser la carte *Internet* selon la place que chaque ville du globe y occupe. La réponse est dans la figure 3, dont l'évidence décourage les commentaires. Les cités nord-américaines monopolisent les premières places et si LONDON se hisse au troisième rang (derrière WASHINGTON et NEW-YORK), PARIS ne se situe qu'en dixième position et TOKYO, BERLIN et MELBOURNE aux rangs 15, 16 et 20. Encore peut-on penser que dans bien des cas les noms de PARIS ou BERLIN sont évoqués comme étant la matière ou la cible du discours et non pas leur source. Et cela peut être plus vrai encore de cités mythiques ou historiques comme BABYLONE, PERSEPOLIS, BYZANCE, ou de cités contemporaines établies sur des ruines antiques : JERUSALEM, ATHENES ou ROME par exemple. Cette charge culturelle donne ainsi au vieux monde une plus-value dont ne bénéficient pas les sites modernes comme SEATTLE ou BUFFALO. Malgré ce handicap, l'Amérique du Nord truste

les premières places et ROME, tous emplois et toutes graphies confondus, n'arrive qu'en 24e position, derrière GENEVE. Mais le présent Colloque qui a lieu dans la cité de César va sûrement améliorer ce classement.

Tableau 3. Les citations toponymiques d'*Internet* classées par ordre de fréquence décroissante

Washington	64 057	Quebec	7 133	Winnipeg	3 258	Bombay	1 700
New York	51 593	Manchester	6 885	Madrid	3 160	Rio de Jan.	1 683
London	28 891	Nouvelle Orl.	6 734	Bonn	3 158	Taipei	1 567
Chicago	28 323	Dublin	6 224	Zurich	3 103	Toulouse	1 529
San Francisco	24 417	Moscow	6 056	Florence	3 006	Strasbourg	1 451
Mexico	23 244	Brussels	5 899	Barcelona	2 996	Granada	1 424
Boston	21 948	Wien	5 638	Copenhagen	2 920	Buenos Air.	1 422
Los Angeles	18 160	Canberra	5 389	Panama	2 740	Lima	1 395
Seattle	17 920	Frankfurt	5 288	Luxembourg	2 739	Istambul	1 295
Paris	17 455	Stockholm	5 283	Auckland	2 731	Kiel	1 263
San Diego	16 794	Milan	4 741	Perth	2 720	Leipzig	1 224
Toronto	16 738	Stuttgart	4 733	Kobe	2 712	Porto	1 110
Buffalo	15 902	Jerusalem	4 682	Sao Paulo	2 479	Jakarta	1 071
Atlanta	15 459	Beijing	4 487	Haiti	2 375	Marseille	1 041
Tokyo	14 345	Munich	4 456	Cairo	2 292	Firenze	1 024
Berlin	13 957	Glasgow	4 402	Basel	2 148	Sofia	997
Philadelphia	12 194	Oslo	4 337	Lausanne	2 145	Kawasaki	994
Vancouver	11 629	Lyon	4 120	Santiago	2 061	Madras	942
Montréal	9 365	Petersburg	3 989	Torino	1 970	Seville	912
Melbourne	9 336	Athens	3 905	Belfast	1 906	Gibraltar	879
Sydney	8 256	Bern	3 517	Nagoya	1 895	Paz	853
Amsterdam	8 178	Liverpool	3 502	Monaco	1 859	Lisboa	820
Geneva	7 817	Venice	3 461	Delhi	1 856	Ankara	814
Rome	7 811	Prague	3 378	Bordeaux	1 833	Calcutta	785
Edinburgh	7 201	Kyoto	3 357	Tel Aviv	1 733	Pretoria	770

Bien que cela ne soit ni très nécessaire, ni très précis, on a tâché de regrouper les villes en pays ou en zones assez larges pour donner une idée synthétique de la carte géographique d'*Internet*. La figure 4 qui en rend compte montre une domination des États-Unis si écrasante qu'on a dû tronquer l'histogramme. Le poids des îles britanniques ou des populations germaniques atteint à peine le sixième du géant américain, et au moins deux fois l'importance quantitative de la France, de l'Italie, du Japon ou

des pays slaves, Russie comprise. Quant à l'Afrique, tous pays réunis, le total des mentions qu'elle réunit ne représente que le centième des États-Unis. Le fossé entre les nations développées et les pays en espoir de développement est donc énorme, même si rien dans nos données ne permet de dire s'il se creuse ou se réduit.

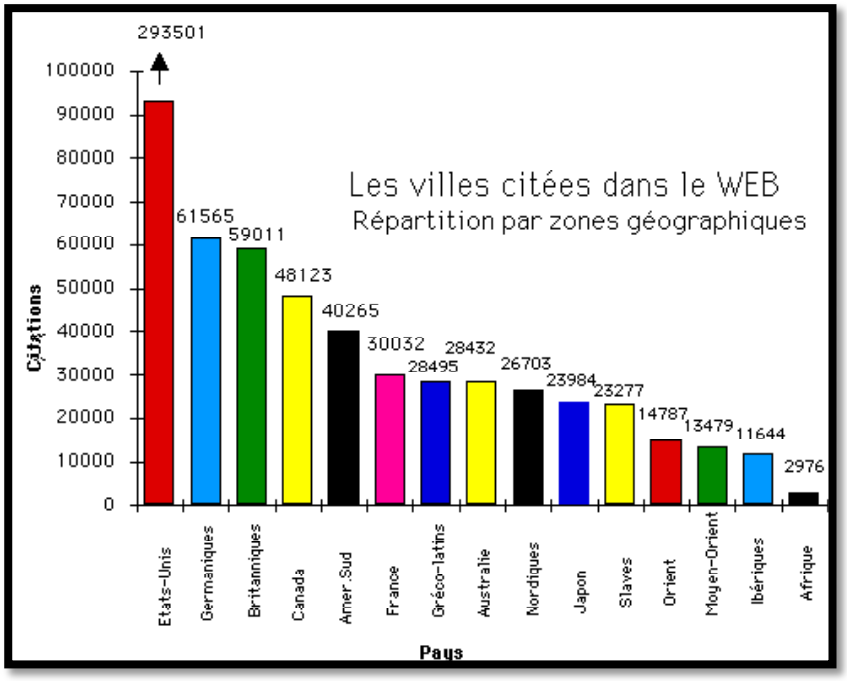


Figure 4. Les pays et les zones sur *Internet*

-III-

Pour qu'on puisse suivre les mouvements et les courants, il faudrait poser à *Lycos* les mêmes questions à intervalles réguliers, afin de disposer du paramètre temps. Tout nous dit que cela en vaudrait la peine, vu l'évolution rapide du réseau et la croissance exponentielle de *Lycos* (en trois mois le nombre de documents que ce serveur déclare prendre en compte est passé de 9 millions, en septembre 1995, à près de 14 millions, à la veille du présent Colloque). Mais les serveurs ont une mémoire et un appareil enregistreur toujours en éveil qui remplit sans discontinuer un fichier historique où les statistiques, temporelles ou non, sont piégées. Loin de donner sans compter, on note pointilleusement le jour, l'heure, l'origine, l'objet de la demande et le nombre de caractères transmis au

demandeur. Voilà un protocole d'enquête idéal : questionnaires exhaustifs et systématiques, neutralité brutale et aveugle du sondeur, transparence innocente du sondé qui ignore la caméra invisible. De telles pratiques d'espionnage ont plus d'intérêt sur les grands serveurs, quand le contact est sollicité à chaque seconde. Les petits serveurs ne peuvent guère accéder qu'à la satisfaction – souvent amère – de compter le nombre trop restreint de leurs lecteurs. L'exemple dont nous nous servirons est de cette dernière espèce. Il correspond à l'un des deux modestes serveurs que nous avons installés dans notre laboratoire. Précisons toutefois qu'il ne propose pas ces indigestes pages de publicité qu'on voit trop souvent sur *Internet*, mais une véritable base de données, apte à répondre à des questions complexes qui entraînent des recherches croisées, des tris, et des opérations particulières, impossibles à préparer à l'avance, comme on le verra plus loin. L'enregistrement automatique des transactions est assuré dans un fichier (généralement pourvu du suffixe *log*) dont on fournit un extrait dans la figure 5 et qui comporte autant de lignes que d'appels reçus et traités. Reste à soumettre au tri un tel fichier et à regrouper les lignes selon l'heure, la date, l'origine de l'appel ou l'objet de la requête. C'est l'objet du programme *Web*stat qu'on voit souvent associé aux serveurs et qui livre les statistiques par jour du mois, jour de la semaine, heure du jour, région géographique et requête traitée, en indiquant pour chaque rubrique le nombre de fichiers explorés et le nombre de caractères transmis. Un extrait de la liste fournie est présenté dans la figure 6.

date	heure	résultat	adresse du client	requête	volume (nb caractères)
08/29/95	00:01:37	OK	arena.unice.fr.	:Rabelais.cgi	1 178
08/29/95	00:10:14	OK	arena.unice.fr.	:rabelais.html	3 700
08/29/95	00:10:16	OK	arena.unice.fr.	:Rabelais75.GIF	8 081
08/29/95	00:14:44	OK	arena.unice.fr.	:listex.html	2 809
08/29/95	00:15:59	OK	arena.unice.fr.	:Rabelais.cgi	1 796
08/29/95	13:54:01	OK	kangourou.unice.fr.	:rabelais.html	3 700
08/29/95	13:54:05	OK	kangourou.unice.fr.	:Rabelais75.GIF	8 081

Figure 5. Le fichier historique des transactions

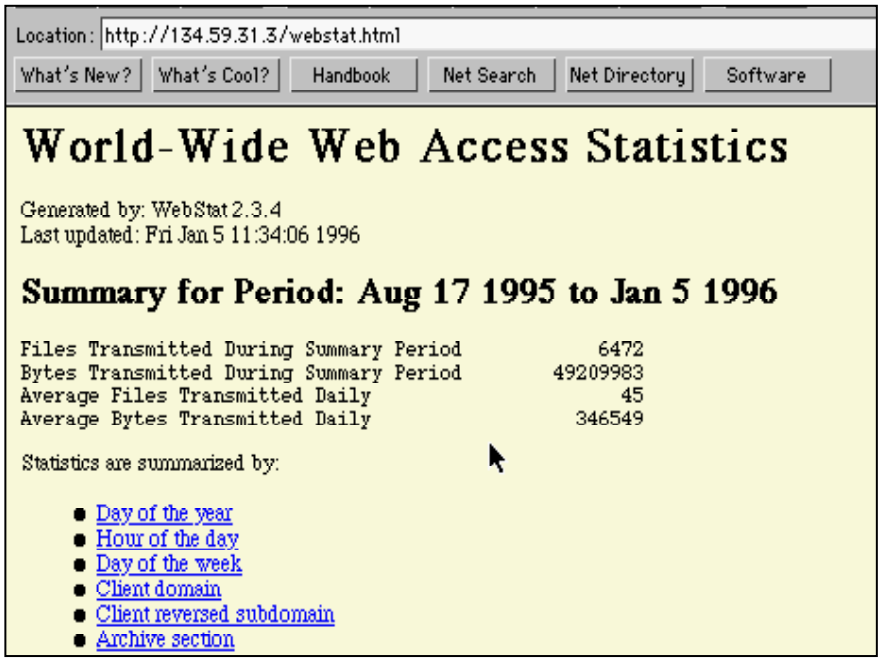


Figure 6. L'espion professionnel : *Web*stat

On a converti de telles listes en histogrammes[3], eux-mêmes accessibles par la voie télématique, en distinguant pour chaque série le nombre de fichiers transmis et le volume de la transaction. On trouvera ci-dessous (figure 7) la représentation graphique de la charge supportée à chaque heure du jour. Visiblement le serveur en question se repose la nuit entre 2 heures et 7 heures, n'étant guère connu au-delà de l'Atlantique. Il s'agit en effet d'un serveur expérimental, qui vient de naître et dont l'existence n'a pas été divulguée. Quand un serveur s'est constitué une clientèle internationale, la charge est répartie plus équitablement, l'est et l'ouest faisant contrepoids.

3 Ce programme (original) de représentation graphique prend appui sur le fichier restitué par *Web*stat et reproduit dans la figure 6.

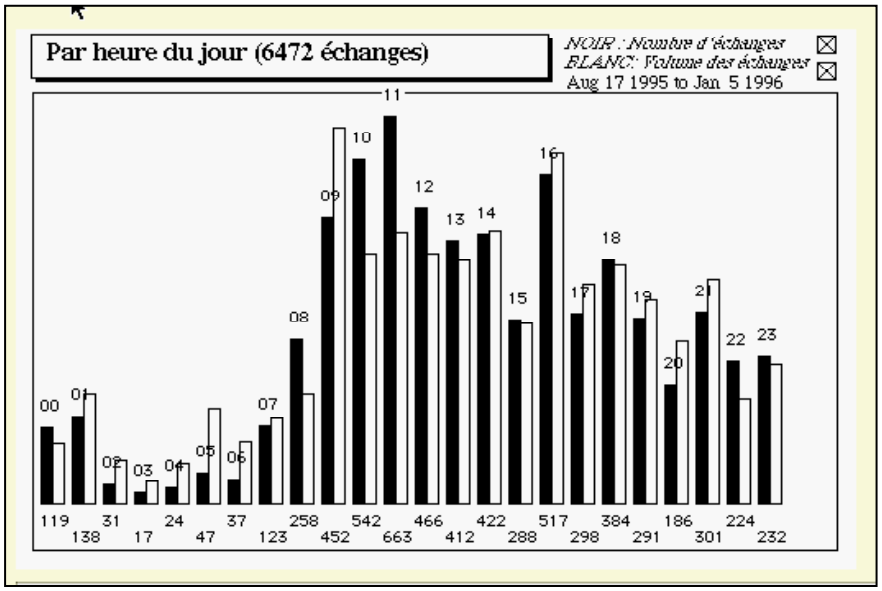


Figure 7. La charge horaire du serveur

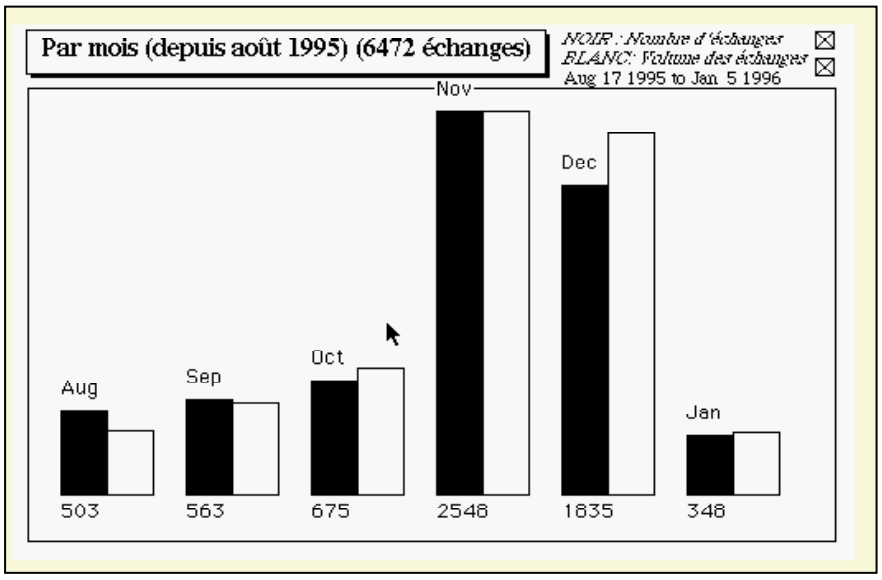


Figure 8. Évolution des appels sur quelques mois

Pour les mêmes raisons qui tiennent à la trop grande jeunesse du serveur, la figure 8 n'est guère significative. Elle le serait si le service

était assuré depuis longtemps et plus connu. Observons qu'on n'a pas affaire à un audimat partiel et contestable, mais à une mesure exhaustive qui rend compte avec une fidélité absolue de l'évolution de la demande.

Quant à la répartition géographique, elle peut être établie à partir du même fichier si l'on sait décrypter les adresses des clients. Celles qui apparaissent en lettres sont à demi transparentes, le suffixe indiquant clairement le pays d'origine. Les autres – comme celles représentées ci-dessous – peuvent être déchiffrées en ayant recours au serveur DSN qu'on trouve sur les nœuds d'*Internet*.

Tableau 9. L'origine géographique des clients (extrait)
(l'adresse est ici numérique)

adresse	nb fichiers	nb caract.	%fichiers	% caract.
192.5.146.173	2	55047	0.05	0.21
192.70.115.190	43	817710	1.12	3.19
192.134.44.9	9	206726	0.24	0.81
192.134.45.100	13	118687	0.34	0.46
193.48.184.41	2	11553	0.05	0.05
193.51.74.16	2	11553	0.05	0.05
193.51.78.14	1	3472	0.03	0.01
193.54.152.10	5	26239	0.13	0.10
193.104.34.34	4	14697	0.10	0.06
193.145.222.4	5	15907	0.13	0.06
193.145.222.10	2	11553	0.05	0.05
194.5.53.35	11	401874	0.29	1.57

Enfin le même fichier produit par *Webstat* donne le détail complet des articles qui attirent la clientèle ou la laissent indifférente. On a ainsi le moyen de connaître ses goûts et les tendances du marché.

Tableau 10. Les fichiers transmis (extrait)

nom du fichier	nb appels	nb caract.	%appels	% caract.
/Rabelais.cgi	992	3908261	25.93	15.26
/rabelais.html	481	955370	12.58	3.73
/rabelais3.html	2	7400	0.05	0.03
/Rabelais75.GIF	393	2040508	10.27	7.97
/redpoint.gif	27	2400	0.71	0.01
/resu1.html	6	533791	0.16	2.08
/resu2.html	16	947241	0.42	3.70
/rireest.html	1	1355	0.03	0.01
/Sophiste.html	1	2533	0.03	0.01
/specifrabe.html	1	13636	0.03	0.05
/specif.html	32	929342	0.84	3.63
/structure.html	18	336197	0.47	1.31
/test.html	1	539	0.03	0.00
/tiers.gif	1	10297	0.03	0.04
/tiers.html	1	10533	0.03	0.04
/WEBGRAPHE/graphestat.html	10	829	0.26	0.00
/WEBGRAPHE/hist1.gif	68	352840	1.78	1.38
/WEBGRAPHE/hist2.gif	65	292362	1.70	1.14
/WEBGRAPHE/hist3.gif	44	186452	1.15	0.73
/WEBGRAPHE/hist4.gif	32	149308	0.84	0.58
/webstat.html	42	533241	1.10	2.08
/~brunet/pub/axes.html	3	19302	0.08	0.08

On pourrait aller beaucoup plus loin dans l'analyse, si l'on procédait à des tris croisés ou à des regroupements par périodes, par zones géographiques ou par catégories de produits. Rien de bien nouveau dans

cette exploitation statistique du résultat des ventes. Les entreprises commerciales savent faire cela depuis longtemps. Ce qui est nouveau, c'est le champ d'application des transactions – qui ne sont pas à proprement parler commerciales, même si l'enjeu économique y est considérable. *Internet* n'est peut-être pas le lieu où se divulguent à la cantonade les secrets stratégiques mais des milliers d'informations y circulent avant même d'avoir fait l'objet d'une publication ou d'un brevet. La demande même des clients peut être une indication précieuse pour l'orientation de la recherche. De tout cela la veille technologique peut tirer le plus grand profit, quand elle a accès à un serveur puissant et rayonnant – ce qui n'est pas le cas de l'exemple que nous avons choisi, faute de mieux.

-IV-

Enfin, le réseau *Internet* peut être aussi le lieu où s'accomplit – à distance et en temps réel – la recherche statistique. On en montrera une illustration avec une base de données textuelles qu'on a proposée à la communauté scientifique sur le *Web* et qui concerne Rabelais et son époque. Certes un CD-Rom a d'abord été produit sur ce thème et l'on aurait pu s'arrêter là. Mais l'alternative qui oppose le CD-Rom à *Internet* est moins exclusive que complémentaire – et les éditeurs ont vite compris l'intérêt des productions multimodales, dérivées les unes des autres. Chacun des deux supports a ses avantages et ses contraintes spécifiques. Si le *Web* n'a pas la vivacité, la souplesse et la puissance dont dispose le CD-Rom, il l'emporte manifestement pour la simplicité et la généralité du dialogue. Et il n'interdit pas certaines interrogations complexes et des traitements sophistiqués comme l'analyse factorielle – ce qu'on va tenter de montrer. L'écran d'accueil est reproduit dans la figure 11.

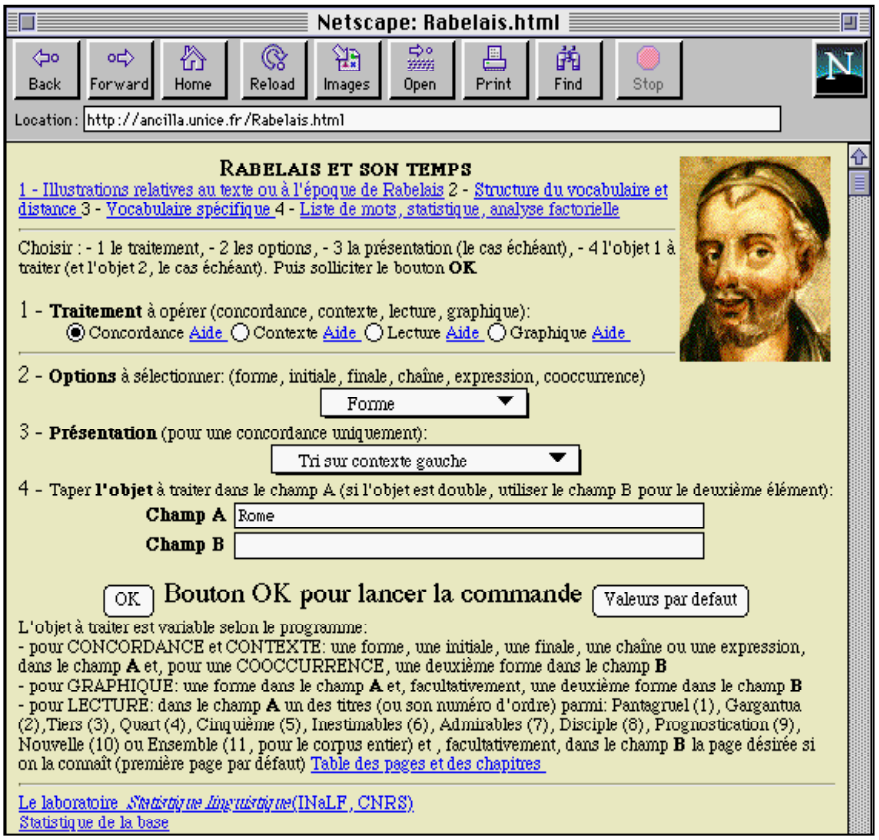


Figure 11. L'écran d'accueil de la base *Rabelais*, disponible sur le *Web*

Il contient peu d'éléments véritablement statistiques, étant orienté d'abord vers les requêtes documentaires, et en particulier vers la recherche des contextes. Une fois de plus, nous nous intéresserons à ROME, ce qui d'ailleurs correspond à un intérêt particulier de Rabelais. Rabelais, attiré par Rome comme tous les humanistes de la Renaissance, y a fait son premier séjour en 1534 comme médecin et conseiller du cardinal Jean du Bellay, lui-même conseiller du Roy et familier du Pape. Mais tous les chemins ne conduisent pas à Rome et le sujet populaire de Gargantua n'aurait pas dû mener de ce côté. Et de fait Rome n'est citée ni par les devanciers, ni par les imitateurs de Rabelais. Mais la mention de ROME revient 45 fois dans le texte de Rabelais, surtout dans le *Tiers Livre*. Voir figure 12.

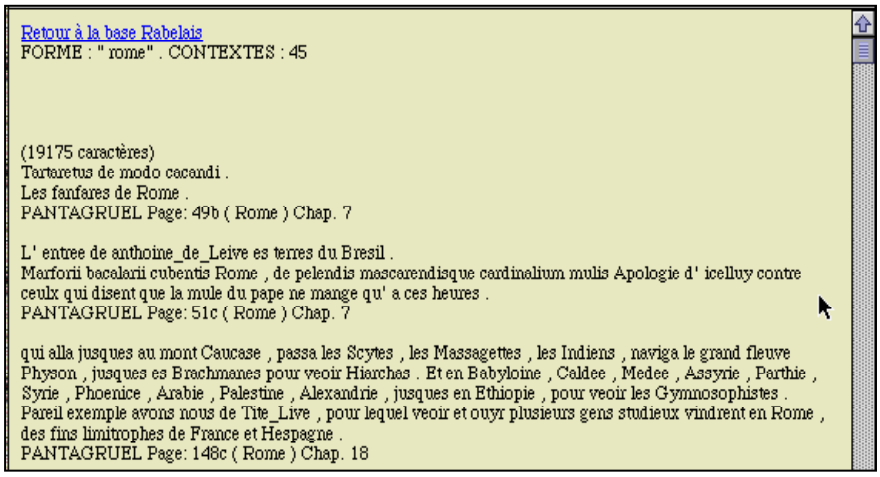
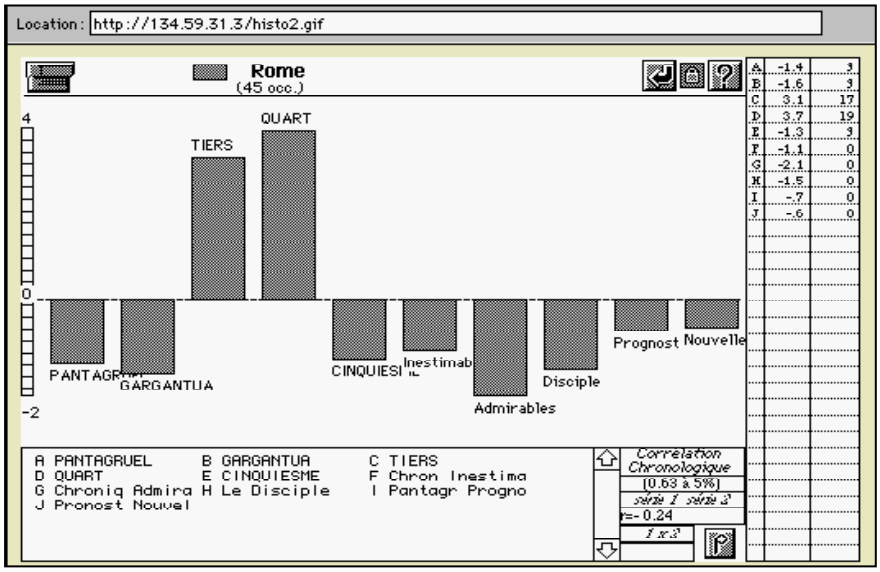


Figure 12. Le contexte de *ROME* chez Rabelais (extrait)



Graphique 13. Histogramme de *ROME* dans le corpus Rabelais

D'autres routes documentaires s'ouvrent plus largement sur des sources iconographiques et font découvrir 400 illustrations empruntées aux ouvrages de l'époque que Rabelais a connus ou inspirés. Et bien entendu, le texte même dont le corpus est constitué est accessible sur l'écran dans l'édition originale. La statistique n'est pourtant pas absente de cette page d'accueil, puisqu'on y voit une proposition de graphique,

simple ou double, pour tout mot choisi par l'utilisateur, comme par exemple le mot ROME (graphique 13).

La statistique reste discrète dans cette page d'accueil pour ne pas effrayer les populations littéraires. Mais elle montre le bout du nez derrière les invitations innocentes regroupées au haut de l'écran d'accueil. Les « ancres » déposées là établissent un lien avec des tableaux de chiffres, des listes de mots, des courbes et même des analyses factorielles, tous résultats relatifs à la structure du vocabulaire, à la distance lexicale ou aux spécificités thématiques des textes. L'invitation la plus explicite conduit les téméraires à un menu copieux où la sauce statistique est commune à tous les plats et qu'on a reproduit dans la figure 14 :

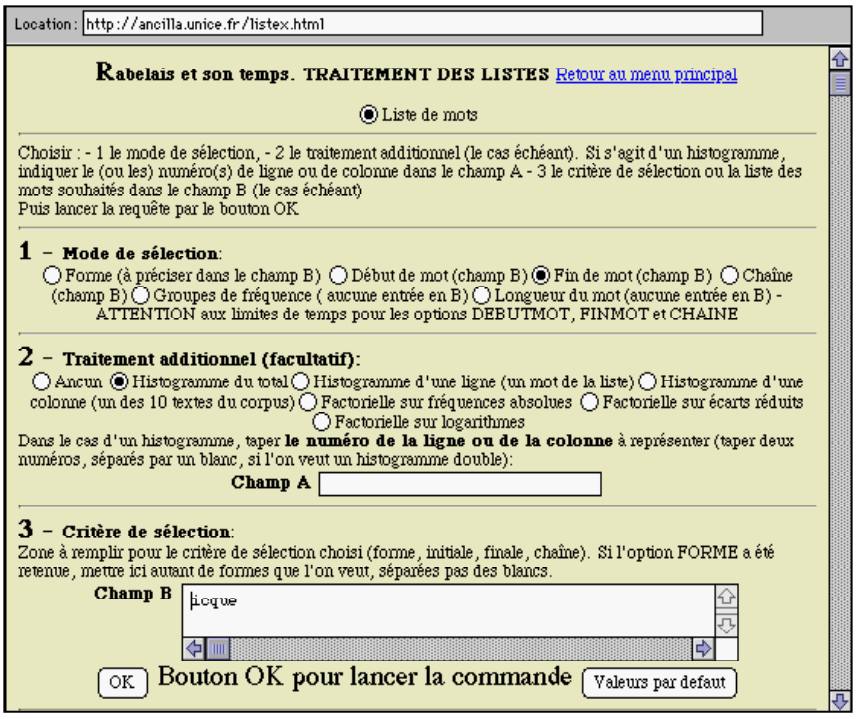


Figure 14. Le menu statistique de la base Rabelais

L'objectif est ici d'abord de constituer des tableaux à deux dimensions dont les lignes sont dévolues aux mots et les colonnes aux textes, et qui sont la matière même de la plupart des traitements quantitatifs. Au croisement de la ligne *i* avec la colonne *j*, on observe donc la fréquence du mot *i* dans le texte *j*. Le choix des mots peut se faire

librement, l'utilisateur les proposant dans un champ réservé à cet effet. Mais des facilités sont aussi offertes pour établir des sélections à partir de l'initiale ou de la finale, ou de la présence d'une chaîne de caractères. D'autres critères sont proposés qui portent sur la fréquence ou la longueur des mots. Enfin, l'utilisateur, avant de lancer sa commande, est invité à préciser le traitement statistique souhaité. Ce peut être la distribution d'ensemble de la série à travers la totalisation des lignes du tableau. Ainsi obtient-on l'histogramme de tous les mots bâtis sur le suffixe -IQUE. Voir figures 15 et 16. Avant de conclure imprudemment à une explosion soudaine dans le *Cinquième Livre* d'un suffixe appelé à un grand avenir, il convient de prendre la mesure d'un frère aîné, le suffixe -ICQUE, dont la distribution est complémentaire, en sorte qu'on a affaire en réalité à une modification de l'orthographe, la graphie -IQUE s'étant substituée à l'archaïsme -ICQUE, par un mouvement semblable à celui qu'on observe pour les finales ULX (AULX, EULX, OULX) et UX (AUX, EUX, OUX).

pratique	2	-1.3	0	0	0	0	2	0	0	0	0	0	2	0
prophetique	2	1.7	0	0	0	0	2	0	0	0	0	0	2	0
pythagorique	4		0	0	0	1	3	0	0	0	0	0	4	0
republique	2		0	1	0	0	1	0	0	0	0	0	2	0
rethorique	2		0	1	0	0	1	0	0	0	0	0	2	0
rubrique	1		1	0	0	0	0	0	0	0	0	0	1	0
rustique	2	-0.6	0	0	0	0	2	0	0	0	0	0	2	0
satyrique	1		0	0	0	1	0	0	0	0	0	0	1	0
sphaerique	1		0	0	0	1	0	0	0	0	0	0	1	0
spherique	2	6.5	0	0	0	0	2	0	0	0	0	0	2	0
symbolique	1	-0.8	0	0	0	0	1	0	0	0	0	0	1	0
synecdochique	1		0	0	1	0	0	0	0	0	0	0	1	0
tetrique	1		0	0	0	1	0	0	0	0	0	0	1	0
theologique	1		0	0	0	1	0	0	0	0	0	0	1	0
titanique	1		0	0	0	1	0	0	0	0	0	0	1	0
tritonique	1		0	0	0	0	1	0	0	0	0	0	1	0
tyrannique	2	-0.4	0	0	0	1	1	0	0	0	0	0	2	0
unique	5	-3.0	0	0	2	1	2	0	0	0	0	0	5	0
venereique	1		1	0	0	0	0	0	0	0	0	0	1	0
veronnique	1		0	0	0	0	1	0	0	0	0	0	1	0
vivifique	1		0	0	0	0	1	0	0	0	0	0	1	0

L'histogramme est au format GIF .ATTENDRE QUELQUES SECONDES avant de solliciter l'ancre HISTOGRAMME.
histogramme

Paramètres:
graphique /
Temps d'éxécution : 23 seconde(s)
Nombre d'appels de cette base: 1081

Figure 15. Recensement des formes en -IQUE

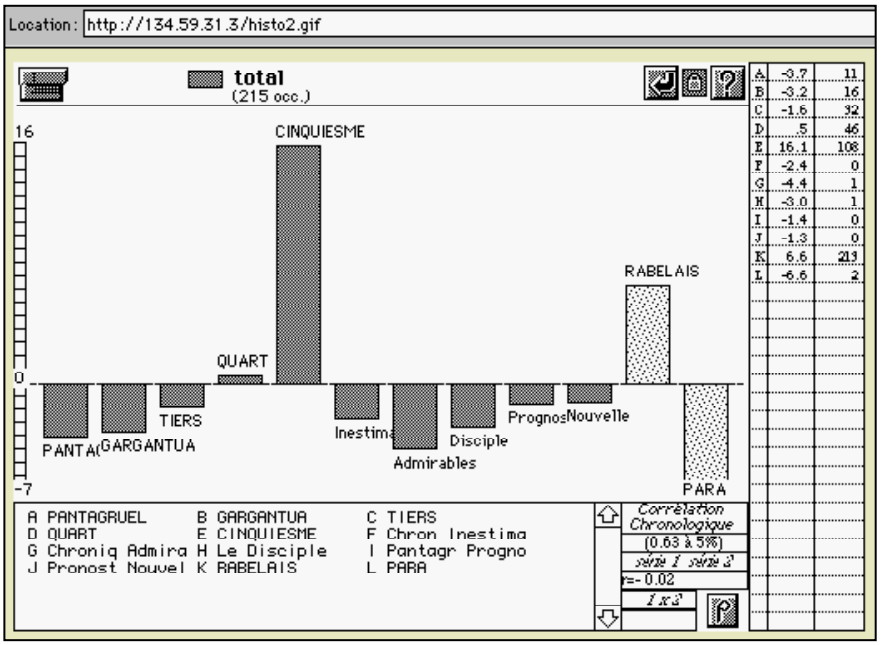


Figure 16. Histogramme des formes en -IQUE

On peut aussi focaliser l'attention sur une ou deux lignes du tableau, si l'on précise leur numéro d'ordre dans la liste traitée. À cette possibilité de dessiner la courbe des mots, s'ajoute celle de représenter le profil des textes, à travers la distribution des mots ou catégories choisis. Si par exemple on dresse la liste des sites géographiques les mieux représentés dans le corpus, ROME y occupe le troisième rang dans l'ordre de la fréquence, derrière PARIS (100 occurrences) et FRANCE (56 occ.). Si l'on s'arrête à la fréquence 10, le tableau contient une trentaine de lignes correspondant aux toponymes en faveur à l'époque. Mais la faveur n'est pas constante pour les mêmes lieux selon qu'on a affaire à Rabelais ou à un autre auteur. Le terroir de Rabelais est celui de ses racines, les pays de Loire, et sa patrie d'adoption Rome et l'Antiquité. Tout autre est le paysage de l'auteur des *Chronicques Admirables*, comme on le voit dans le profil du graphique 17. Nul tropisme méditerranéen : l'auteur regarde à l'opposé, vers l'Angleterre, Londres, la Champaigne, la Bretagne ou la Flandre. Le centre de gravité se déplace vers le Nord et l'Ouest.

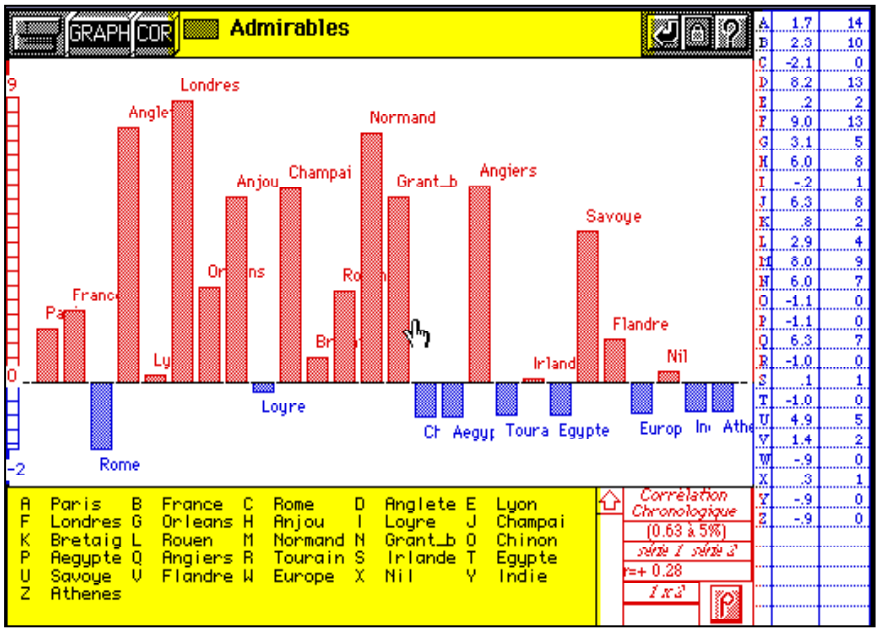


Figure 17. La représentation d'une colonne (ici les *Chronicques Admirables*)

Quelle que soit l'évidence graphique de la représentation d'une ligne ou d'une colonne, ce n'est là qu'un coup de projecteur porté sur une zone limitée du tableau. Mais avec des méthodes spécifiques, on peut ouvrir l'angle du faisceau pour éliminer toute zone d'ombre. Cet éclairage collectif est fourni par les analyses factorielles. Le dialogue de la figure 14 en propose trois variétés, selon qu'on souhaite ou non pondérer les données. L'éclairage qui souligne le mieux les reliefs est souvent celui qui utilise le filtre de l'écart réduit. Les logarithmes constituent un filtre plus neutre, qui corrige plus faiblement l'effet de taille. Si les données brutes sont traitées sans filtre (cette possibilité reste offerte), on peut craindre en effet que l'étendue variable des textes et le poids inégal des mots retenus ne précipitent au centre du graphique les éléments les plus lourds et les plus aptes à faire la loi.

Il suffit de 10 secondes pour obtenir le résultat de la figure 18, obtenue avec le filtre logarithmique. On voit qu'à partir du *Tiers Livre*, l'intérêt de Rabelais s'écarte de la petite patrie et, suivant le vent de la Renaissance, s'oriente vers la Méditerranée, en atteignant d'abord ROME, puis en poussant vers les contrées lointaines de la Grèce, de l'Égypte et de l'Inde (ATHENES, EUROPE, AEGYPTE, ÉGYPTE, NIL, INDIE). Au centre, le *Gargantua* et le *Pantagruel* ne sortent guère du cercle étroit de

l'Ile de France et des pays de Loire. Quant aux textes para-rabelaisiens ils se tournent traditionnellement à l'Ouest et au Nord et regardent vers l'Angleterre, qui avait tant fait parler d'elle au Moyen-âge.

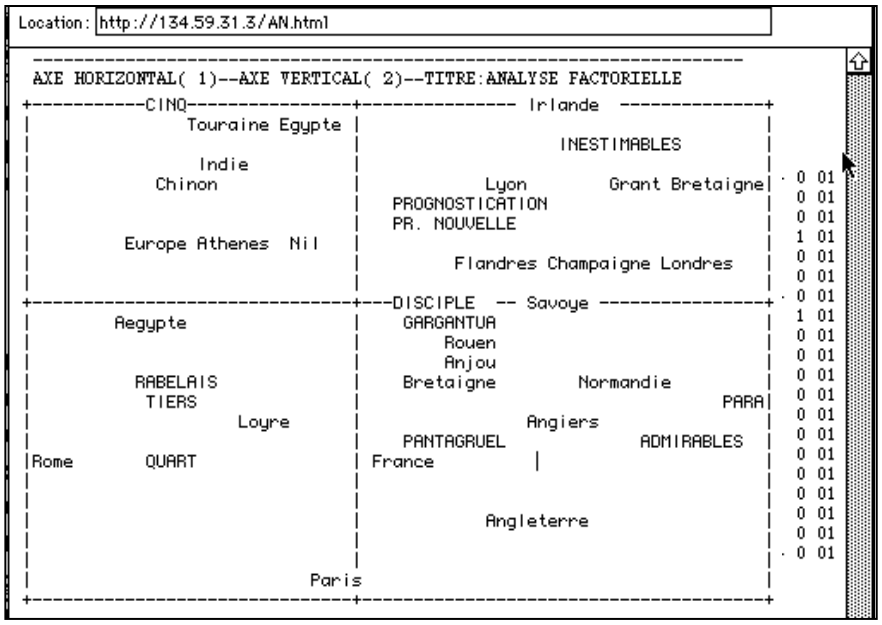


Figure 18. La géographie de Rabelais. Analyse factorielle

Bien d'autres analyses attendent le chercheur, avec les données assemblées à son goût. Il est probable que la typologie des textes y sera analogue à celle que suggère la figure 18 : d'un côté les livres 3 à 5, de l'autre les textes para-rabelaisiens, et au milieu *Pantagruel* et *Gargantua* qui prolongent et renouvellent une tradition. Il est possible aussi que l'exploitation du même gisement produise un effet de saturation : quand on tourne autour d'une boule, on a beau varier les angles de vue, c'est toujours la même image qu'on obtient.

-V-

C'est pourquoi il importe de varier les données autant que les points de vue. On a regretté au début de cet exposé la rareté des textes disponibles sur *Internet*. Il y a une notable exception que nous nous proposons d'aborder pour finir. C'est celle de *Frantext*. Cette base de données textuelles, constituée il y a plus de trente ans pour approvisionner en exemples les rédacteurs du *Trésor de la langue*

française est à notre connaissance la plus importante et en tous cas la plus homogène base qu'on ait constituée au monde en matière littéraire et linguistique. La base française contient quelques milliers de textes, représentant 160 millions de mots. Quant à son double américain, établi à Chicago sous l'appellation *ARTFL*, si son étendue est un peu moins considérable, son approche est plus aisée puisque le canal de diffusion choisi est celui du *Web*.

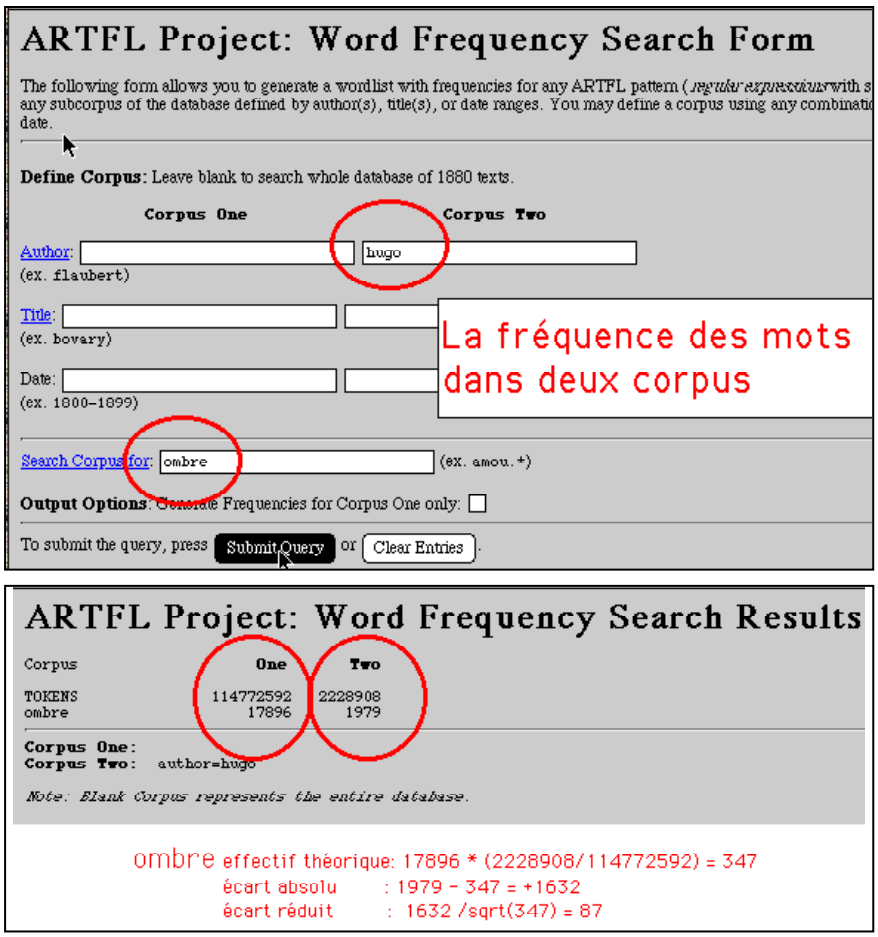


Figure 19. Éléments de statistique dans *ARTFL* (sur le *Web*)

Comme les fonctions documentaires d'*ARTFL* sont réservées aux abonnés américains, il n'est guère utile d'en parler de ce côté-ci de l'Atlantique, où les interrogations d'ordre statistique sont cependant

autorisées. Mais elles sont d'une extrême rusticité et se bornent à des indications de fréquence pour un mot donné dans deux corpus choisis. Aucun calcul n'est réalisé et l'utilisateur a la charge de la comparaison et du test statistique. Mais cette lacune est moins gênante que la nécessité de recommencer l'opération pour remplir les cellules du tableau, deux cases seulement étant livrées à chaque consultation. Pour mémoire, on trouvera ci-dessous l'exemple du mot OMBRE chez Hugo et les quatre résultats obtenus qui permettent le calcul de l'écart réduit (ou, comme on voudra, le calcul hypergéométrique). La version simplifiée du test (z= 87) est suffisante pour mettre en lumière, si l'on peut dire, l'OMBRE hugolienne. Voir figure 19.

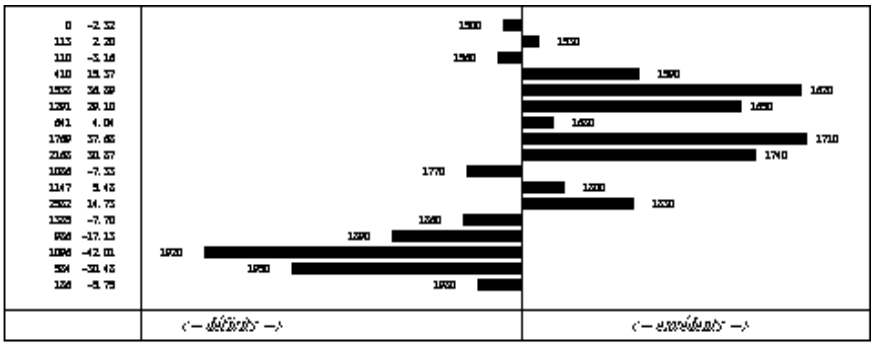
Frantext est beaucoup plus riche en fonctions statistiques. Pour un corpus choisi, il peut fournir la liste intégrale du vocabulaire avec l'indication de fréquence pour chacune des formes. Il permet de constituer à volonté des listes de mots et d'en extraire les fréquences dans les textes que l'on veut. Pour un mot donné, une liste préétablie, il autorise les recherches portant sur l'évolution (en opposant les tranches chronologiques les unes aux autres) ou sur la répartition (en comparant les auteurs). Et pour ces fonctions puissantes, il laisse à l'utilisateur le choix du corpus, le choix des mots et le choix du pas de progression dans le temps. Le dialogue avec l'utilisateur est évidemment plus complexe puisque beaucoup d'options sont proposées. Et si l'interrogation emprunte le canal d'*Internet* (on accède aussi par *Transpac*), le dialogue n'est pas encore celui du *Web*. Mais une version *Web*, mise au point par Jacques Dendien, est déjà fonctionnelle et devrait être proposée aux chercheurs prochainement.

Comme l'unanimité ne règne pas parmi les spécialistes sur les meilleurs tests statistiques à appliquer aux données, les auteurs de *Frantext* n'en ont choisi aucun et les données numériques sont livrées dans un état presque brut, seule étant réalisée la transformation en fréquences relatives. Pour chaque élément d'une distribution, on dispose donc des fréquences réelle et théorique, grâce à quoi il est facile de restituer l'étendue de chaque sous-corpus et d'utiliser les tests statistiques que l'on préfère.

Néanmoins, comme ces manipulations sont longues et délicates, nous avons réalisé un automate qui dirige et enregistre le dialogue avec *Frantext*. Le produit du pompage est entreposé dans des fichiers avant d'être canalisé dans des stations de traitement spécialisées qui livrent des courbes, des listes triées ou des analyses factorielles. Cette industrie de

transformation (on l'a appelée THIEF pour souligner la filiation naturelle à la source-mère) est responsable des quelques exemples qui vont suivre et qui n'épuisent pas la variété des actions possibles.

En reprenant le nom de ROME qui nous a servi de prétexte tout au long de ce parcours, on parvient à des résultats dignes d'intérêt. En explorant la base entière, de 1500 à nos jours, c'est plus de 17 000 occurrences qu'on observe en cinq siècles de littérature française. Mais cette présence massive de Rome dans les lettres françaises n'est pas constante. Elle s'amplifie du 16e au 17e siècle et se maintient à un palier élevé jusqu'à la Révolution. Un regain de faveur se marque encore à l'époque romantique, les charmes artistiques et touristiques de cette ville-musée se substituant à l'autorité de la cité pontificale. Puis, à partir de 1850, un nuage s'interpose sur la frontière et le rayonnement de ROME en France perd de son intensité. Ce déclin – confirmé par un coefficient de corrélation chronologique significatif (r = -50) – se lit aisément dans l'histogramme 20 :



rome (17092 occurr.). Courbe de l'évolution.

(17 périodes prises en compte. Taille du corpus : 1699981066 occurrences)
Coefficient de corrélation chronologique: -0.5061
(Seuil à 5% : 0.4821 pour 17 paires d'observation)

Figure 20. Évolution déclinante du mot *ROME* dans les lettres françaises

Cette faveur de Rome aux siècles classiques se retrouve lorsqu'on isole les auteurs. Ceux qui citent Rome le plus souvent appartiennent au 17e et au 18e siècles et prennent place dans la moitié droite de l'histogramme 21. Un millier d'écrivains ont été passés en revue dont une centaine ont des écarts significatifs, en plus ou en moins (ce sont les seuls représentés dans le graphique).

Les écrivains contemporains se situent dans la zone des déficits, mis à part le critique Charles Du Bos et Julien Gracq, dont un essai *Les Sept*

Collines est consacré à Rome. Au XIX[e], l'attirance que Rome exerce sur Mme de Staël, Chateaubriand, Lamartine, Balzac et surtout Stendhal se vérifie, mais la présence romaine semble moins directement rattachée aux préférences personnelles qu'au sujet traité et au genre littéraire mis en œuvre. Rome a partie liée avec les mémoires (Argens, Dangeau, Retz) et les écrits historiques (Michelet, Las Cases, Renan), avec les essais politiques ou philosophiques (Marmontel, Montesquieu, Voltaire) ou avec les considérations artistiques (Faure), ou religieuses (Calvin, Bossuet, Fénelon).

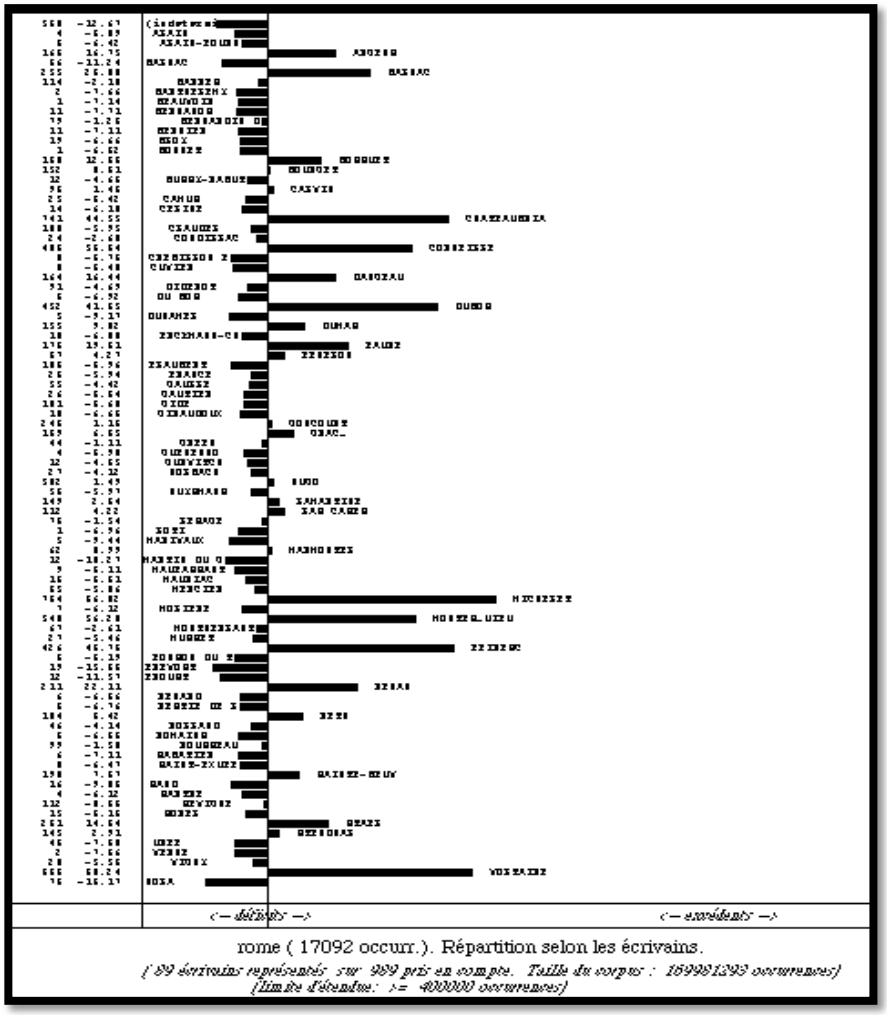


Figure 21. Répartition du mot *ROME* chez les écrivains français

Tableau 22. L'environnement de *ROME*. Spécificités positives
Fréquence > 50 écart réduit > 3 Corpus de référence : Corpus total

59	22	évesque	197.9	60	3493	st	12.2
664	6791	pape	121.7	70	4497	vn	12.2
733	16388	rome	82.1	88	6573	triomphe	11.9
416	8627	romains	64.7	109	9745	nations	11.1
275	4504	romain	60.3	180	20057	année	11.0
175	2080	sénat	57.6	55	3459	débris	11.0
180	2745	sénat	50.8	140	14178	autorité	10.9
744	37108	ville	48.6	102	9045	princes	10.8
103	1175	fondation	45.2	187	21482	gouvernement	10.8
77	1120	pontife	34.1	373	54357	saint	10.4
180	5434	évêque	33.8	132	13685	univers	10.3
63	870	tribuns	31.8	134	14100	anciens	10.2
54	679	patriciens	31.0	84	7230	vouloit	10.2
217	8778	ancienne	30.5	72	6005	déjà	9.7
125	3332	romaine	30.5	94	8924	siècles	9.7
228	10559	république	28.4	97	9406	murs	9.6
342	21826	avoient	27.1	70	5879	orient	9.5
92	2363	évêques	26.7	68	5732	trône	9.3
313	19846	église	26.0	82	7886	porté	8.9
172	9557	après	21.5	161	20313	lors	8.7
86	3397	église	19.5	66	5868	dignité	8.7
252	20721	ay	18.4	103	11294	donna	8.5
61	2103	dict	18.0	123	14674	ordinaire	8.3
57	2039	siège	17.0	173	23356	prix	8.1
135	8640	jusques	17.0	51	4414	superbe	7.9
68	2948	églises	16.3	173	23769	siècle	7.9
171	13196	palais	16.1	118	14441	autrefois	7.8
150	11023	soldats	15.8	102	12095	ancien	7.6
132	9080	durant	15.8	98	11470	prise	7.6
169	13288	campagne	15.8	90	10382	ny	7.4
62	2659	italien	15.7	247	38294	histoire	7.4
58	2384	italiens	15.6	203	30204	nouvelle	7.3
145	10730	rêve	15.5	55	5307	tableaux	7.3
144	11399	an	14.5	85	9787	quoy	7.2
279	31429	prince	13.5	96	11754	françois	7.0
57	2903	jésuites	13.2	206	32247	sang	6.6
58	2999	tyran	13.2	115	15861	beaux	6.4
56	2872	murailles	13.1	93	12077	droits	6.3
86	5745	séjour	13.1	232	38705	droit	6.1
100	7305	icy	13.0	52	5892	camp	5.8
223	23873	rue	12.9	81	10662	ordres	5.8
129	10996	patrie	12.7	54	6432	sçavoir	5.5
55	2896	tyrannie	12.7	126	19437	vit	5.4
107	8405	portes	12.6	86	12180	école	5.3
113	9129	victoire	12.6	56	6993	auparavant	5.2
199	20830	puissance	12.5	140	22505	parti	5.2
216	23421	gloire	12.5	103	15739	vint	4.9
61	3510	païs	12.5	58	7605	passa	4.9
55	2994	occident	12.4	52	6680	tousjours	4.9
57	3211	parmy	12.2	60	3493	st	12.2

Il est un moyen de mesurer plus directement la coloration thématique d'un mot. Il est fourni par une fonction puissante de *Frantext*, appliquée à l'ensemble des contextes où apparaît le mot choisi pour pôle (ce peut être aussi une liste). Cette fonction relève tous les termes qui environnent le mot-pôle et en livre la liste alphabétique, fréquences à l'appui. En soumettant cette liste à des calculs de pondération, on obtient une constellation lexicale qui circonscrit l'environnement privilégié du mot étudié et met en relief ses corrélats préférés. On trouvera dans la figure 22 les mots que Rome attire et dans la figure 23 ceux qu'elle repousse.

Figure 23. Déficits dans l'entourage de *ROME* (ou spécificités négatives)
Fréquence > 50 écart réduit < -3 Corpus de référence : Corpus total

253	128753	vie	-11.7	125	49275	bonne	-5.3
69	57564	sens	-10.8	56	27161	instant	-5.1
100	62188	nature	-9.6	56	26919	sentiment	-5.1
70	51298	quoi	-9.6	58	27241	petits	-5.0
169	82705	amour	-9.1	88	35115	ami	-4.5
170	80788	femme	-8.7	86	34070	vue	-4.4
58	40821	eau	-8.3	102	38809	femmes	-4.4
109	58362	raison	-8.3	55	24221	oeil	-4.3
78	46896	sais	-8.1	157	54097	va	-4.2
149	64102	moment	-6.9	101	37635	forme	-4.2
80	41022	fond	-6.7	108	39720	donne	-4.2
155	64937	tête	-6.6	131	46430	parler	-4.1
162	66932	père	-6.6	107	39190	besoin	-4.1
55	32110	vérité	-6.6	89	33426	pauvre	-4.0
56	31423	idées	-6.3	52	22064	francs	-3.9
257	93572	hommes	-6.2	76	28821	autour	-3.8
88	41057	bras	-6.1	51	21043	sœur	-3.7
184	70330	bon	-6.0	74	27608	longtemps	-3.6
81	38059	bas	-5.9	607	172906	homme	-3.5
171	65603	doit	-5.8	62	23585	grâce	-3.4
98	43087	VRAI	-5.8	58	21828	passe	-3.2
109	45430	idée	-5.5	74	26398	pieds	-3.2
151	58268	porte	-5.5	51	19567	ombre	-3.2
75	34641	pensée	-5.5	58	21483	justice	-3.1
74	34069	matin	-5.4	58	21245	travers	-3.0
61	29787	manière	-5.4				

Pas besoin d'être grand clerc pour relever la présence massive du haut clergé dans la ville aux trois cents églises, qui abrite le Saint-Siège. La tradition littéraire française, comme on le voit déjà dans les *Regrets* de Joachim du Bellay, voit d'abord dans Rome la ville du Pape (PAPE, PONTIFE, SAINT, SIEGE, EVESQUE, EVEQUE, EVEQUES, JESUITES, EGLISE, EGLISE, EGLISES). Et, comme dans les *Antiquités* du même Du Bellay, on y sent aussi l'héritage de l'empire romain, à la faveur de la polysémie qui

s'attache à l'adjectif dérivé aussi bien qu'au nom propre (ROMAIN, ROMAINS, ROMAINE, ANCIEN, ANCIENNE, ANCIENS, AUTREFOIS, SENAT, SENAT, TRIBUNS, PATRICIENS). Deux millénaires ont été nécessaires pour que ces deux visages se substituent ou se superposent l'un à l'autre et l'histoire est là, plus présente qu'ailleurs (HISTOIRE, FONDATION, ANNEE, SIECLE, SIECLES), avec ses guerres (SOLDATS, CAMP, FUREUR, SANG), ses expériences politiques et juridiques (GOUVERNEMENT, REPUBLIQUE, TYRANNIE, PATRIE, PARTI, POLITIQUE, DROIT, DROITS) et une splendeur jamais démentie, malgré les misères (PALAIS, PRINCE, PRINCES, TRONE, SUPERBE, TRIOMPHE, VICTOIRE, GLOIRE, PUISSANCE, AUTORITE, DIGNITE).

Cet aspect princier et théocratique est renforcé par le contraste que fournissent les ombres du portrait. On entend par là les termes avec lesquels Rome n'est pas en bons termes. Ils se trouvent dans le tableau 23 et définissent un univers laïc et humble dont l'homme est devenu le centre (HOMME, HOMMES, FEMMES). Cet humanisme modeste (PAUVRE, PETITS, BESOIN) n'exclut pas les liens du cœur (AMOUR, AMI, SENS, SENTIMENT) ou de la famille (PERE, SŒUR), non plus que les réalités du corps (TETE, BRAS, ŒIL, PIEDS), de la nature (NATURE, EAU, VIE) et de l'existence humaine dont les courts instants sont précieux (MATIN, INSTANT, MOMENT). À l'absolutisme religieux que Rome symbolise s'opposent un rationalisme à visage humain qui prône la RAISON et la VERITE (IDEE, IDEES, PENSEE, VRAI), et une morale sociale en quête de JUSTICE *(*BON, BONNE*)*. Ce monde sans Dieu et sans majesté, si peu compatible avec la grandeur romaine, caractérise sans doute aussi le cadre bourgeois et réaliste dans lequel le roman français s'est installé depuis deux siècles.

Mais élargissons le champ en considérant dans *Frantext* non plus la seule ville de Rome mais les lieux géographiques les plus connus, soit 130 au total. Distinguons les écrivains, afin de savoir à qui attribuer les préférences ou les rejets. Et constituons un vaste tableau de 130 lignes (ROME occupant l'une d'entre elles) et de 31 colonnes (on a choisi 31 écrivains du XVIII^{e} au XX^{e} siècle). Une dernière fois, nous allons recourir à l'analyse factorielle. Celle que reproduit la figure 24 donne une image de cette géographie mentale que projette l'usage des écrivains.

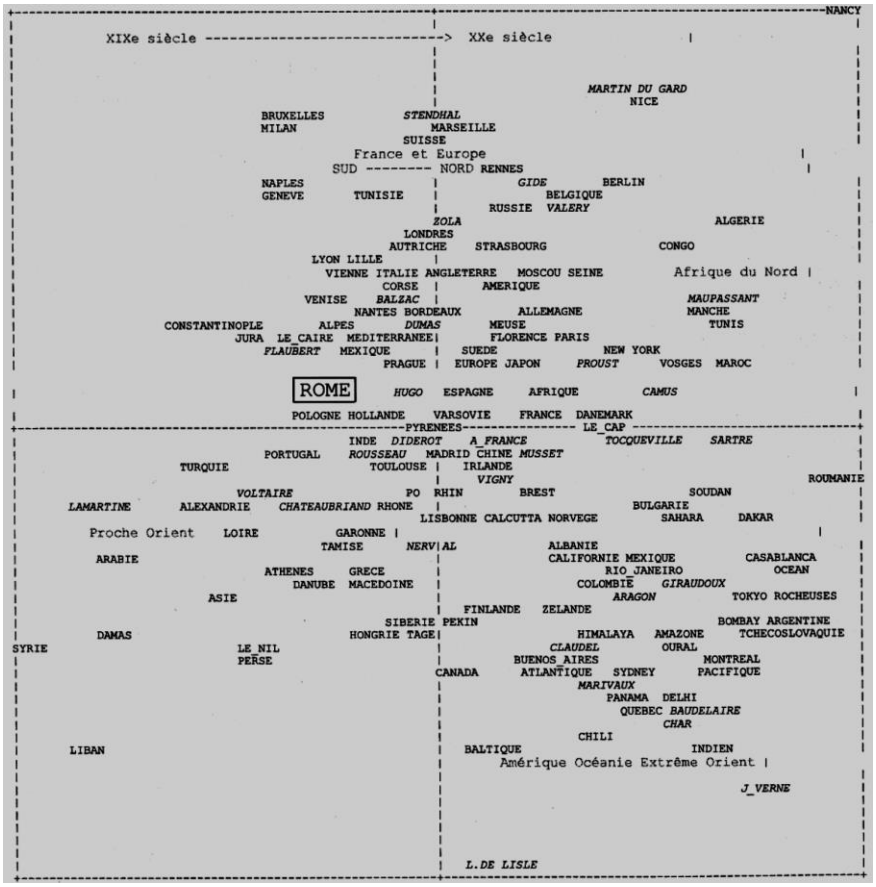


Figure 24. Analyse factorielle de 130 noms géographiques parmi 31 écrivains

Les lieux qu'on relève dans le quadrant inférieur gauche dessinent un contour bien précis qui est celui de l'Orient, au sens restreint que l'on donnait à ce mot dans les siècles passés et qui correspond à la méditerranée orientale. Les pays évoqués à cet endroit : LIBAN, SYRIE, DAMAS, ASIE, PERSE, NIL, ALEXANDRIE, LE CAIRE, CONSTANTINOPLE, TURQUIE, GRECE, MACEDOINE, sont ceux qui mènent aux lieux saints et que connurent les Croisés. Or, les Croisés des temps modernes se situent au début du XIX^e^ siècle quand le voyage à Jérusalem devient le rêve d'une génération. Chateaubriand entreprend jusqu'au bout cet « itinéraire » et, à sa suite, Lamartine et Flaubert. Précisément, le graphique situe à cet endroit les noms de Chateaubriand et Lamartine. Voltaire aussi lorgne de ce côté aussi bien que Nerval. Or Rome – c'est le

lieu que Voltaire cite le plus souvent – se situe non loin de là, sur l'axe des *x*, à l'endroit où le graphique passe en Europe.

Si l'on examine de plus près le quadrant supérieur gauche, on voit qu'il est l'apanage du roman français du XIX^e^ siècle, et l'on y voit réunis Balzac, Stendhal, Flaubert et Zola. Le quadrant supérieur droit appartient plutôt aux prosateurs du XX^e^ : Gide, Proust, Valéry, Martin du Gard et Camus. Or parallèlement à cette différenciation chronologique, on croit déceler aussi un mouvement géographique : les villes et pays du midi sont plutôt à gauche près de ROME (MILAN, NAPLES, VENISE, ITALIE, MEDITERRANEE, CORSE, ALPES, LYON, GENEVE, VIENNE), tandis que le nord tend à s'installer à droite : ANGLETERRE, ALLEMAGNE, BELGIQUE, SUEDE, DANEMARK, RUSSIE, BERLIN, MOSCOU, STRASBOURG, NANCY, MEUSE, SEINE, MANCHE. Le mouvement de l'histoire semble donc favorable au nord, la Méditerranée perdant sa force d'attraction.

Enfin le dernier quadrant, en bas et à droite, est le plus excentrique. Les distances y sont plus grandes, comme elles le sont dans la réalité physique. On a là l'Amérique : CALIFORNIE, MEXIQUE, CANADA, QUEBEC, COLOMBIE, ROCHEUSES, PACIFIQUE, ARGENTINE, CHILI, BUENOS AIRES, RIO DE JANEIRO, MONTREAL, PANAMA, ATLANTIQUE, PACIFIQUE, AMAZONE. C'est ici qu'on rencontre les pays les plus reculés de l'Asie, de l'Inde, de l'Extrême-Orient et de l'Océanie (DELHI, BOMBAY, HIMALAYA, OCEAN INDIEN, CHINE, PEKIN, TOKYO, SYDNEY). Dira-t-on que ces contrées lointaines sont devenues accessibles au tourisme littéraire ? Les écrivains que le graphique situe dans ces parages sont en effet parfois des diplomates qui ont voyagé loin, comme Claudel et Giraudoux. Mais ce sont surtout des poètes, comme Aragon, Char, Baudelaire ou Leconte de Lisle, auxquels s'ajoute un représentant de la science-fiction : Jules Verne. La part du rêve semble donc ici l'emporter sur celle de la réalité, comme c'était le cas du mirage oriental un siècle plus tôt. Comme les frontières du monde se sont rétrécies, il a fallu aller chercher le rêve plus loin.

****

Nous arrêterons là notre enquête, sans décider si les traits de Rome qu'on vient d'observer appartiennent au modèle ou au reflet déformant et archaïque qu'en livre la littérature française. Au reste cette monographie, partielle et provisoire, d'un site célèbre n'a d'autre but que d'illustrer, en

profitant des circonstances[4], les possibilités qui s'ouvrent à l'explorateur lancé sur les autoroutes de l'information[5]. La statistique a toujours aimé les grands espaces, les grandes masses de données et la loi des grands nombres. Internet lui ouvre ses richesses, sans contrôle, sans retard, sans dépense et sans limites. À elle d'en profiter.

[4] Le présent exposé a fait l'objet d'une communication à Rome, en décembre 1995, à l'occasion des *Journées d'Analyse Statistique des Données Textuelles*.

[5] Nous avons négligé un autre service que le réseau Web peut rendre à la statistique : celui de la publication des études. Internet est devenu la première maison d'édition mondiale. Plus de retard entre la rédaction et la diffusion, plus de commandes, ni de livraisons, ni de stocks. Plus de gestion ni d'argent. La communauté scientifique pourrait être comme un grand monastère où tout se partage. Dans notre discipline même, des forums se créent, se procréent et se multiplient. Je n'en citerai qu'un exemple : la revue télématique Lexicometrica qui vient de naître, à l'initiative de André Salem, et dont voici l'adresse : http://www.msh-paris.fr/~salem/revue.html. Le présent article y a trouvé un gîte provisoire avant d'être couché sur le papier.