LES BASES DE DONNÉES EN TEXTE INTÉGRAL. RÉALISATION ET DIFFUSION.

Étienne BRUNET

BCL, Université de Nice
brunet@unice.fr

Les organisateurs, qui ont le souci de la composition, m'ont donné le double rôle de parenthèse ouvrante et fermante. Ils savent que les parenthèses ne sont là que pour encadrer le contenu et qu'elles ne font pas partie du message. Aussi bien ne suis-je guère versé dans la discipline qui fait la matière de ce Colloque, et, de la dialectologie, je ne connais guère que mon patois natal (celui des Mauges, en Anjou) qui n'appartient pas à la langue d'oc et que Le Du peut commenter mieux que moi. Car si je puis encore le comprendre, je ne saurais rien en dire, étant dans ce domaine aussi ignorant que Monsieur Jourdain faisant de la prose sans le savoir. Mais l'objet du Colloque est large et ne fait mention ni de la dialectologie ni des régionalismes et ce qu'il met en question ce sont les bases de données linguistiques dans leur ensemble, en mettant l'accent moins sur le contenu que sur l'aspect méthodologique et opératoire de leur fabrication. Dès lors mon propos peut n'être pas tout-à-fait extérieur, et, comme la croûte fait partie du fromage, la parenthèse n'est pas sans parenté avec ce qu'elle enveloppe1.

1. Un modèle de données linguistiques sur abonnement: Frantext

Depuis quelques années la France détient le cordon bleu en matière de base de données linguistiques. *Frantext* n'a d'équivalent dans aucune langue au monde, pour la taille, l'homogénéité et la disponibilité. Avec 160 millions de mots et près de 3000 textes complets de la littérature nationale, le corpus livré à la communauté scientifique dépasse ceux qui ont été réunis autour de l'anglais, de l'italien ou du suédois. Certes il n'est plus très difficile de dépouiller des textes et de les transférer sur un support informatique. Avec le scanner et la lecture optique, on a un moyen d'engranger de grandes masses de texte. Encore faut-il que les caractères typographiques appartiennent à une édition moderne, les polices anciennes faisant encore difficulté. Les bandes de composition constituent un autre moyen d'approvisionnement, à la double condition que la provision existe (les éditions fondues au plomb n'ont laissé d'autre trace que le papier), et que des raisons commerciales n'empêchent pas leur emploi détourné. Et s'il est relativement aisé d'avoir accès aux données périssables des quotidiens ou des hebdomadaires, dont la réédition est hautement improbable, le marché du livre, qui a mal résisté aux infiltrations du photocopieur et de l'audiovisuel, se barricade devant l'ordinateur et le *CD-Rom*. Aux deux bouts de la chaîne chronologique les données se raréfient, et c'est précisément à ces deux extrêmes que le catalogue de *Frantext* s'enrichit. Chaque année de nouveaux titres, puisés à l'actualité littéraire ou scientifique, s'ajoutent à la collection, tandis que recule, à reculons, la limite chronologique des textes admis dans la base. Le XVIIe et le XVIIIe siècles sont désormais aussi bien représentés que l'époque contemporaine. Mais l'échantillon du XVIe siècle reste encore mince et lacunaire et le Moyen Âge est absent de la base.

On ne regrettera que médiocrement cette lacune, car l'homogénéité des données, qui a été miraculeusement préservée en trente ans de saisie, pourrait souffrir des incertitudes de l'orthographe et de

1 Pardon pour ce premier et dernier calembour, qui a cependant ses lettres de noblesse, puisqu'on le trouve chez Proust, dans la bouche de Françoise confondant parentèle et "parenthèse".

l'instabilité de la langue dans les siècles reculés. Cette homogénéité est très remarquable dans les données actuelles de *Frantext*, par suite d'un double mouvement d'accélération et de freinage. Dans les années 60, il a fallu aux responsables du *Trésor* une faculté d'anticipation pour imposer les accents français, là où les constructeurs ne proposaient que les 26 lettres majuscules de l'alphabet anglais. Inversement quand l'explosion des machines, des standards et des systèmes a bouleversé le monde informatique, il a fallu une certaine force de résistance pour maintenir les mêmes principes de dépouillement, en refusant même les mesures qui pouvaient enrichir l'exploitation mais portaient en même temps le germe de la rupture et de l'incohérence. Car si la pérennité du monument repose sur son poids et sur le volume et la masse du socle, la garantie majeure est fournie par la solidité du matériau et la cohésion interne des éléments de sa structure. Et c'est précisément cette volonté organisationnelle de l'architecte qui a manqué dans les projets étrangers où la tradition centralisée est moins forte qu'en France. Les Archives de la langue anglaise, amassées à Oxford, constituent sans doute un terril lexical aussi haut que la pyramide française. Mais il s'agit de résidus de la recherche, de textes dépareillés, reçus en héritage et impropres à constituer une véritable base de données réellement opérationnelle.

1 - D'autres tours de Babel, plus hautes encore, se dressent à l'horizon d'*Internet*, et des moteurs de recherche d'une rare puissance parviennent à extraire quelques mots de ces millions de pages amoncelées sans ordre. Mais quand aucun tri n'est opéré à l'entrée des données et que le tout-venant s'installe sans contrôle, il est impossible d'endiguer l'entropie à la sortie et dans le fatras inextricable d'*Internet* aucun filtre ne saurait procéder à l'épuration de l'information restituée. Pour donner une idée de l'impureté de l'eau, voici ce que l'on obtient quand on ouvre le robinet du *WEB,* en proposant le mot *Corse.*

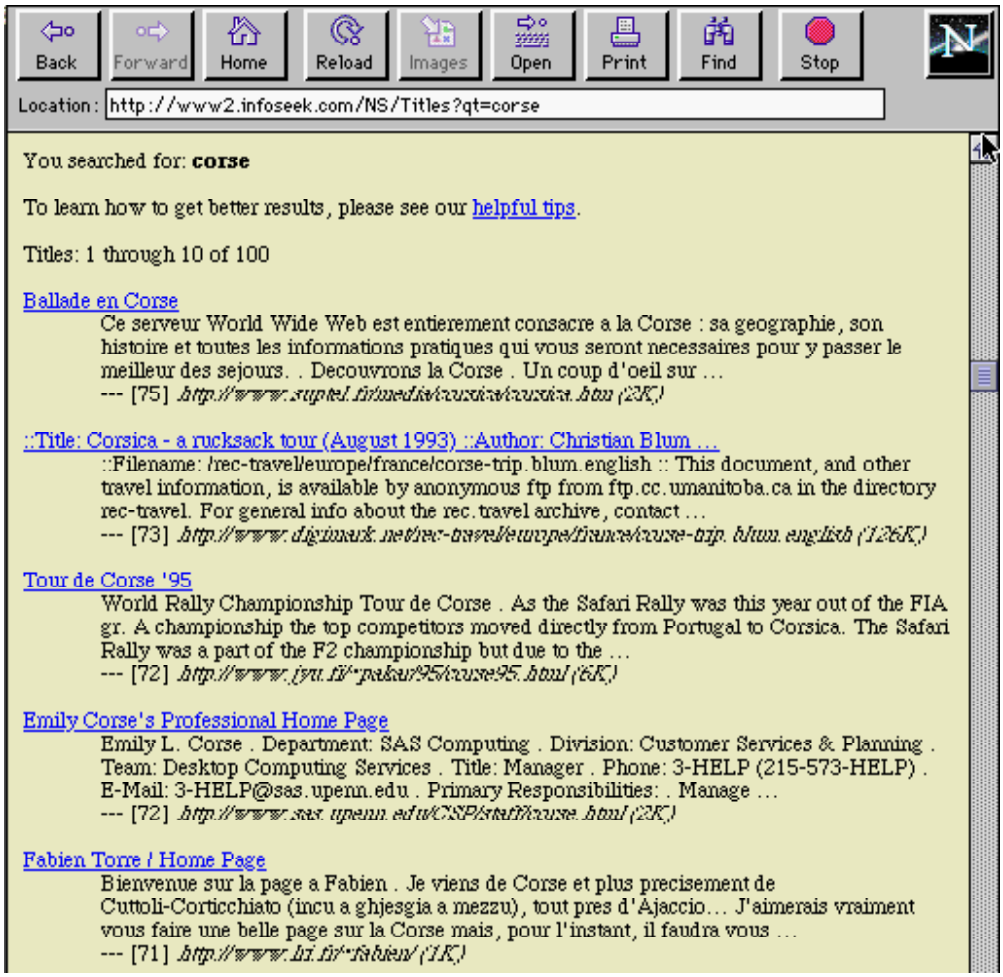


Figure 1. La Corse sur Internet (extrait)

Ce n'est là que le début d'une liste fort longue, dont les 100 premiers éléments sont immédiatement accessibles. En réglant différemment les paramètres, on en obtiendrait beaucoup plus, pour peu qu'on ait le goût des curiosités et qu'on ne soit pas difficile sur la qualité de l'information. Ces cinq propositions qui arrivent en tête de la liste mêlent les langues, le français voisinant avec l'anglais (on y trouve même quelques mots en corse), mélangent les informations commerciales, touristiques et personnelles et confondent toponymes et anthroponymes (ainsi a-t-on droit à la page promotionnelle d'une certaine Emily Corse qui communique ses titres et adresses - et peut-être même, dans la suite de l'article, ses mensurations). *Internet* est devenu un souk criard et vulgaire[2], livré à la réclame, et les marchands ont envahi ce qui fut naguère un temple universitaire.

La station de recherche est ici celle d'*Infoseek* (adresse indiquée dans la figure 1). Celle de *Lycos* donne un résultat encore plus édifiant, puisqu'elle exploite un gisement plus important - et plus désordonné - de 9 millions de références. Pour le mot *Corse,* elle fournit 349 ancres ou références, sans compter une multitude d'adresses qui concernent 44 "composés", par exemple *corset*, où les syllabes du mot *corse* peuvent se retrouver par le plus grand des hasards. Ci-dessous l'une des plus acceptables (figure 2).

Figure 2. Détail d'une référence donnée par Lycos

---

2 Non seulement la futilité règne dans le contenu, mais encore le débraillé et l'inculture s'exhibent sans pudeur dans l'écriture. Alors que le langage html a facilité la transmission des accents français (les auteurs qui travaillaient en Suisse y ont heureusement veillé), beaucoup de serveurs français ont gardé les habitudes négligentes acquises avec le courrier électronique. Et dans notre exemple les deux extraits écrits en français ont délibérément ignoré les diacritiques... et les critiques.

2 - On n'en mesurera que mieux la qualité des réponses fournies par *Frantext* à la même question. La quantité y est aussi puisque 1203 contextes sont fournis dans l'ordre que l'on veut (par auteurs, par textes, par périodes), chacun étant immédiatement accessible. On donnera ci-dessous le premier et le dernier exemples (figure 3). Une distance de quatre siècles les sépare, qui est fort visible dans le ton comme dans le contenu.

Une fonction particulièrement utile de *Frantext* permet pour un mot donné non plus seulement de restituer les contextes où le mot apparaît, comme nous venons de le faire, mais aussi d'explorer ces mêmes contextes et d'en faire une analyse statistique, en comptabilisant tous les corrélats qui s'associent de manière récurrente au mot choisi. Il est aisé de voir que la Corse se situe dans un environnement géopolitique qui intéresse les *génois*, les *italiens*, les *anglais* et surtout les *français* et leur administration (*départements, système, systèmes, préfet, département, député, république*).

```
*********POSSIBILITE DE PROCESSUS PARALLELE*******
  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%
----I----I----I----I----I----I----I----I----I----I
**************************************************
Recherche terminée. Nombre de solutions trouvées : 1203
Vous pouvez :
1 : Visualiser les résultats
2 : Arrêter le programme
Indiquez votre choix (1 ou 2) :1
Ex. 1 (Selec.) --Ex disponibles 1203 --Ex. selec. 1203
Q789/FAUCHET.CL / FLEUR MAISON DE CHARLEMAGNE / 1601
page 51 / LIVRE 1 CHAPITRE 10
Les lettres aportées et leuës en la presence du roy et de ses juges, ils les aprouverent : et lors il commanda à *Ithier son secretaire, d' ajouster aux donations ja faictes, <<<<*Corse>>>>, *Sardaigne, *Sicile (ceste-cy n' estoit lors, et ne fut oncques en la possession de *Charles, ne des lombards ou exarques) le territoire *Sabin, duché de *Spolette et de *Toscane, avec tous les cens que les ducs de ces terres devoyent tous les ans aux rois de *Lombardie : sauf la puissance royalle sur lesdits duchez.

?-aide s-suiv. p-prec. r-revoir i-increment n-voir_num. o-oter O-oter_tout
g-garder G-tout_garder e-echant. Z-zoom_AV z-zoom_AR a-archiv. f-fin t-tri
Donnez le numéro de l'exemple que vous voulez voir : 1203
Ex. 1203 (Selec.) --Ex disponibles 1203 --Ex. selec. 1203
R971/FORLANI.R / GOUTTIERE / 1989
pages 114-115 / La guerre du lapin
ça prit le temps que ça prit et un des gardiens, le noiraud qui avait du sang de braconnier <<<<corse>>>> dans les veines, finit par réussir le prodige d'empoigner Qu'une-Oreille par la queue, au pied de la statue grandeur nature d'un auteur de vaudevilles emporté par la grippe espagnole de mil neuf cent dix-huit.
```

Figure 3 . Les contextes du mot corse (ou Corse) dans Frantext. Extraits 1 et 1203.

Sans doute fait-elle l'objet de quelque récit militaire (principalement des *Mémoires de guerre* de De Gaulle), ce qui explique la présence des mots *militaire, libération, axe, colonel, régiment, commandant, chef, bataille, troupes, marins, marine, flotte*. Les *fusils* y abondent dont certains appartiennent non pas aux soldats mais aux *bandits*. Enfin nul ne peut ignorer à la lecture de cette liste que la Corse est une île (les quatre formes *île, îles, isle* et *isles* occupent les premières places) et plus précisément une montagne (*mont, rochers, village*) dans la mer (*rivages, rivage, ports, bleu*). On eût aimé que des notations poétiques ou morales qualifient plus intimement l'Île de Beauté: mais on ne trouve que de rares notation sur son aspect *primitif* et *sauvage*, son *odeur*, sa *fierté*. Sans doute la tonalité aurait été différente si la littérature corse avait été prise en compte. La Corse qu'on voit ici est celle des administrateurs, des soldats et des voyageurs, c'est-à-dire, en fin de compte, des étrangers.

12	320	génois	31.9	5	1586	rivages	7.3	15	18047	chef	4.9
62	8285	île	31.4	5	1605	matelots	7.2	6	4218	accident	4.8
38	3387	îles	30.2	6	2377	département	7.0	5	3259	fierté	4.6
10	536	méridien	26.8	11	7086	cercle	6.9	5	3359	marine	4.5
20	2828	isle	22.8	5	1774	député	6.8	7	6381	horizon	4.2
8	654	isles	19.3	6	2558	marin	6.7	7	6442	sud	4.2
12	1481	libération	19.0	6	2659	italien	6.5	5	3888	systèmes	4.1
9	860	dominique	18.8	5	1918	marins	6.5	6	5247	placé	4.1
7	600	bandit	17.6	8	4566	rochers	6.4	6	5268	comparaison	4.0
14	2682	axe	16.2	6	2895	régiment	6.2	5	3978	mont	4.0
10	1910	ports	13.7	6	2907	fusil	6.1	10	12095	ancien	4.0
5	564	barbier	12.9	5	2116	accidents	6.1	5	4016	comité	4.0
14	4578	colonel	11.9	7	3991	commandant	6.0	5	4031	abri	3.9
9	2062	primitif	11.8	7	4039	accent	5.9	7	7188	militaire	3.8
16	6046	longueur	11.7	6	3128	rivage	5.9	8	8985	bataille	3.8
5	743	détroit	11.1	36	55626	jeune	5.9	9	11025	française	3.7
7	1431	départements	11.0	5	2303	réseau	5.8	7	7502	odeur	3.7
15	7479	nommé	9.5	5	2315	bassin	5.8	5	4382	promenade	3.7
5	1108	nègres	8.9	8	5322	sauvage	5.7	8	9257	village	3.7
23	17447	système	8.9	5	2384	italiens	5.7	8	9539	royaume	3.6
27	23201	français	8.7	6	3325	anglaise	5.6	9	11971	anglais	3.4
10	4268	miss	8.6	27	38294	histoire	5.6	5	4989	lève	3.3
9	3962	préfet	8.0	5	2665	kilomètres	5.3	8	10430	mise	3.3
17	12143	nord	8.0	6	3687	branche	5.3	8	10522	hôtel	3.3
7	2699	poing	7.7	5	2744	flotte	5.2	8	10559	république	3.3
6	2099	écorce	7.5	5	2953	écrivit	4.9	7	9074	troupes	3.1
7	2838	françaises	7.4	9	8020	bleu	4.9	10	15539	second	3.1

Figure 4. L'entourage du mot corse- (fréquence > 4; écart réduit > 3; référence : corpus entier)

C'est pourquoi il importe de localiser chacun des contextes obtenus pour un mot donné, en le rattachant à l'auteur et à l'époque. D'autres fonctions de *Frantext* aident à faire de tels relevés. Comme leur maniement est assez délicat et contraignant, nous avons créé des procédures automatiques qui simplifient la collecte des données, en fournissant au dialogue de *Frantext* les réponses appropriées et en ne laissant au chercheur que le choix des mots et des corpus. Mais notre logiciel *Thief*, dont la page d'accueil est représentée dans la figure 5, aide aussi le chercheur dans l'exploitation documentaire et statistique des résultats. La conversation, enregistrée dans un fichier, est reprise après coup par des programmes d'analyse qui en tirent les enseignements. Et la figure 4 qui précède a été obtenue ainsi, les fréquences brutes de la cooccurrence avec le mot-pôle (colonne 1) étant rapprochées de la fréquence du mot isolé (colonne 2) pour obtenir un écart réduit (dernière colonne).

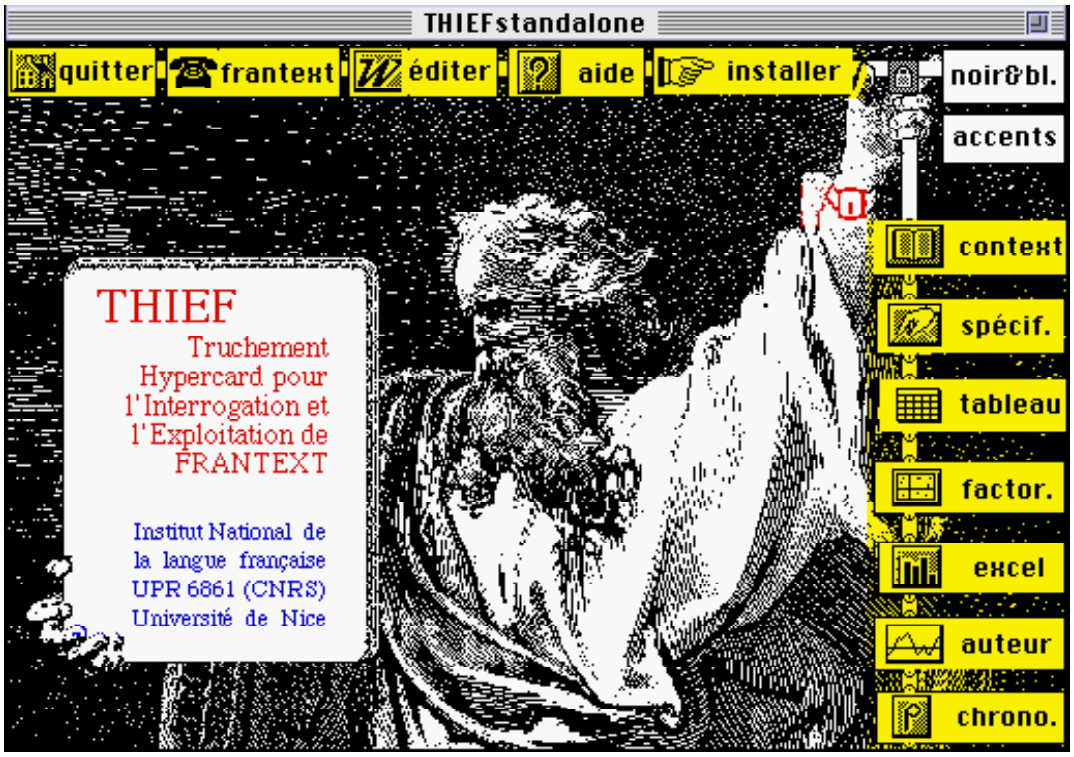
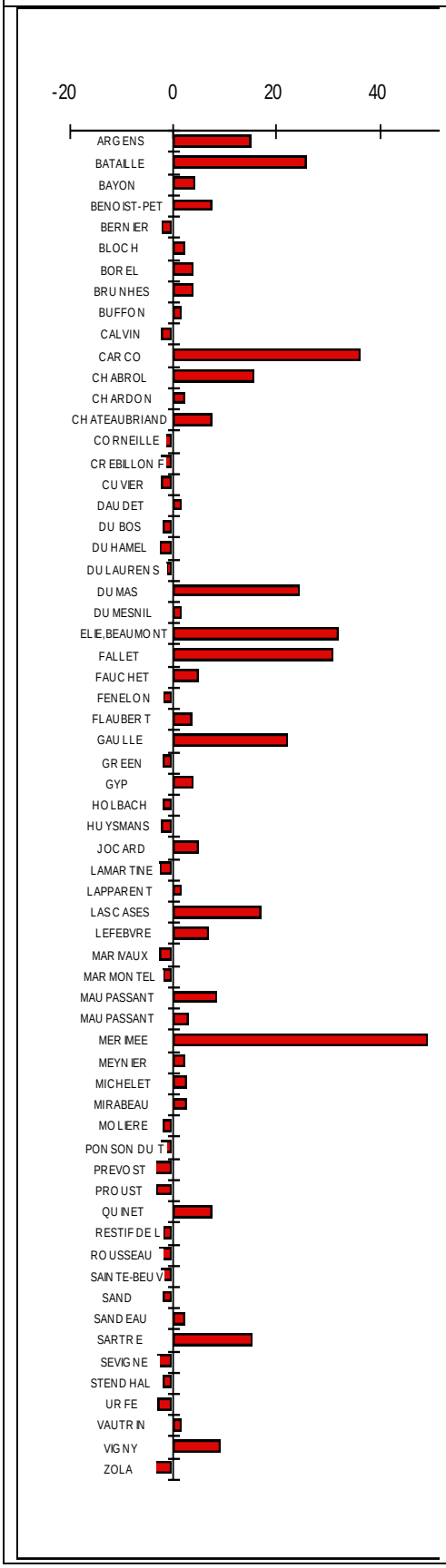


Figure 5. Le logiciel Thief (interrogation et l'exploitation statistique de Frantext)

La fonction *Auteur* (bas de l'écran) compare également la fréquence d'un mot chez les écrivains à l'étendue de leur œuvre (du moins de l'œuvre présente dans le corpus de *Frantext*). L'écart calculé – et pondéré – entre la fréquence observée et la fréquence théorique sert d'abscisse à la représentation graphique, où les excédents sont à droite et les déficits à gauche.

Nous n'avons gardé dans la figure 6 que les écarts les plus significatifs, en plus ou en moins. On notera l'absence du mot chez les auteurs classiques et ceux qui les ont précédés ou suivis immédiatement: ni Corneille, ni Racine, ni H. d'Urfé, ni Molière, ni Mme de Sévigné, ni Marivaux, ni Fénelon ne font la moindre allusion à la terre corse, dont la découverte littéraire semble dater de Napoléon. Les écrivains qui parlent de Bonaparte sont en effet amenés à évoquer sa terre natale, comme le font Las Cases et Chateaubriand. Mais indépendamment de l'Empereur, la Corse devient un décor romanesque pour les récits de Dumas, ou les nouvelles de Maupassant ou Mérimée. Chez les écrivains contemporains, la faveur semble avoir baissé, sauf circonstances exceptionnelles liées à l'histoire locale d'un conflit mondial (de Gaulle), ou à l'exposé technique d'un paysage géologique (Élie de Beaumont).



mot	Écart réduit	Fréquence réelle	Fréquence théorique	Étendue
ARGENS	15.22	32	4	496578
BATAILLE	25.93	20	1	79430
BAYON	4.44	6	1	166803
BENOIST-PET	7.92	6	0	68558
BERNIER	-2.24	0	5	703036
BLOCH	2.47	2	0	58502
BOREL	4.32	4	1	86843
BRUNHES	4.19	5	1	132799
BUFFON	1.99	3	1	142292
CALVIN	-2.45	0	6	842499
CARCO	86.60	106	1	206140
CHABROL	16.05	12	1	72384
CHARDON	2.49	2	0	57726
CHATEAUBRIAND	7.92	39	12	1675580
CORNEILLE	-2.42	0	6	822044
CREBILLONF	-2.33	0	5	763667
CUVIER	-2.23	0	5	698919
DAUDET	2.10	2	0	71454
DUBOS	-2.10	0	4	623329
DUHAMEL	-2.52	0	6	891385
DULAURENS	-1.28	0	2	230822
DUMAS	24.75	56	4	617222
DUMESNIL	2.01	2	1	75236
ELIE.BEAUMONT	65.24	51	1	84385
FALLET	31.27	54	3	381202
FAUCHET	5.22	10	2	314361
FENELON	-1.98	0	4	550221
FLAUBERT	3.74	33	17	2468193
GAULLE	22.62	55	5	695147
GREEN	-1.92	0	4	517170
GYP	4.13	5	1	135527
HOLBACH	-2.03	0	4	581347
HUYSMANS	-2.18	0	5	671506
JOCARD	5.36	5	1	92801
LAMARTINE	-2.55	1	8	1175819
LAPPARENT	2.03	2	1	74258
LASCASES	17.26	45	5	750204
LEFEBVRE	7.18	12	2	276511
MARIVAUX	-2.59	0	7	939218
MARMONTEL	-1.96	0	4	543690
MAUPASSANT	8.81	27	6	820787
MAUPASSANT	3.24	5	1	184310
MERIMEE	109.33	85	1	84257
MEYNIER	2.45	2	0	58982
MICHELET	3.06	18	9	1257380
MIRABEAU	3.08	8	3	399376
MOLIERE	-1.89	0	4	500907
PONSONDUT	-2.41	0	6	815330
PREVOST	-3.69	1	15	2172846
PROUST	-3.28	0	11	1503946
QUINET	7.73	11	2	213071
RESTIFDEL	-2.07	0	4	601063
ROUSSEAU	-2.47	1	8	1121791
SAINTE-BEUV	-2.44	1	8	1100394
SAND	-2.09	2	8	1104846
SANDEAU	2.47	4	1	176241
SARTRE	15.66	31	3	447389
SEVIGNE	-2.93	0	9	1206941
STENDHAL	-2.11	2	8	1116236
URFE	-3.10	0	10	1346693
VAUTRIN	2.02	3	1	140089
VIGNY	9.40	19	3	413836
ZOLA	-4.89	1	25	3578108

FIGURE 6. Le mot Corse parmi les écrivains français de 1500 à nos jours

Cette évolution observée à travers les auteurs peut enfin être étudiée en elle-même, abstraction faite des écrivains. C'est l'objet de la fonction *Chrono*. *Frantext* propose d'en dresser la courbe, pourvu que le chercheur précise le pas de la progression, c'est-à-dire la taille (en nombre d'années) des tranches chronologiques. On a choisi un pas de 30 ans dans la figure 7 qui livre le résultat sous forme d'histogramme. Celui de gauche, fondé sur les fréquences théoriques, appartient à *Frantext*, celui de droite, plus élaboré, aux transformations probabilistes de *Thief*. Tous les deux confirment la même tendance: la Corse est quasiment ignorée du 16e au 17e siècle (20 mentions à peine) et son apparition reste timide au XVIIIe. L'explosion littéraire se produit non pas durant le règne de Napoléon, mais après sa chute, avec le retard propre à tout écho. La mode corse se développe tout au long du XIXe siècle et se maintient, avec des fléchissements, au XXe.

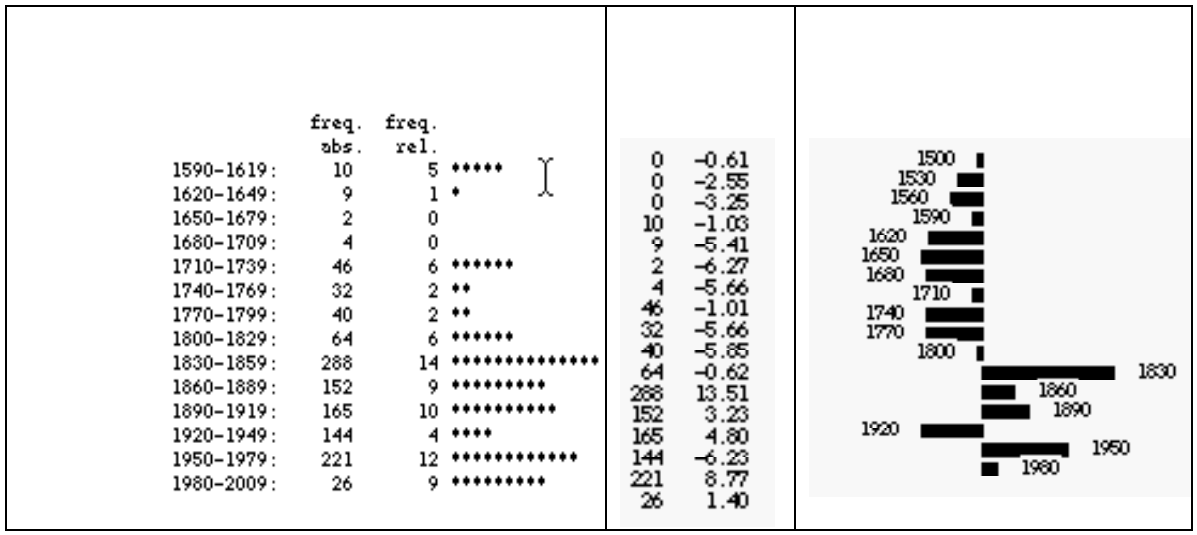


FIGURE 7. Évolution d'emploi du mot Corse de 1500 à nos jours
Fréquence absolue totale : 1203
DIAGRAMME DES FRÉQUENCES RELATIVES, CONVERSION EN ÉCARTS RÉDUITS
Échelle : un astérisque représente une fréquence relative de 1 millionième(s)

2. Un exemple de base de données linguistiques sur CD-ROM

La base de données *Frantext* existe aussi dans une version indépendante et partielle, sous la forme d'un CD-Rom. Ce produit (il se nomme *Discotext 1*) contient plus de 500 textes (de 1789 à 1925). Comme les fonctionnalités sont à peu près les mêmes que celles de la version télématique, il n'y a pas lieu d'insister là-dessus.

Mais puisque l'objet du colloque est relatif à la réalisation et à la problématique des bases de données linguistiques, il peut paraître opportun d'évoquer une expérience personnelle et de montrer, à titre d'exemple plutôt que de modèle, un CD-Rom qui vient d'être commercialisé et dont nous avons assuré la réalisation technique. L'initiative en revient à Marie Luce Demonet et au laboratoire EQUIL XVI qu'elle dirige à l'Université de Clermont II[3]. Il s'agit d'un hypertexte consacré à Rabelais et exploitant un corpus de dix textes dont la moitié appartient à l'auteur de *Gargantua*, les autres textes étant attribués aux

3 Il faut souligner le caractère collectif de la réalisation, qui a bénéficié de collaborations multiples mettant en cause des organismes publics et des entreprises privées. Outre le laboratoire EQUIL XVI, à qui reviennent l'initiative et la conduite du projet, et le centre niçois (UPR 63861, INaLF, CNRS) qui en a assuré la réalisation technique, l'opération a reçu le concours de la Bibliothèque Municipale de Lyon, de la Bibliothèque Nationale de France, du Centre National du Livre et de la société APPLE-France, à quoi s'est ajoutée en dernier ressort l'industrie de l'éditeur.

prédécesseurs ou aux imitateurs de Rabelais. On trouvera dans la figure 8 la composition du corpus traité, avec l'étendue relative de chaque élément.

	Nb.mots	Formes	prob.p	prob.q		
1	46204	8293	0.13805	0.86195	PANTAGRUEL	(PA)
2	51674	9368	0.15439	0.84561	GARGANTUA	(GA)
3	64511	11362	0.19274	0.80726	TIERS	(TI)
4	67018	11271	0.20023	0.79977	QUART	(QU)
5	44122	8002	0.13183	0.86817	CINQUIESME	(CI)
6	8802	1614	0.02630	0.97370	Chron Inestimables	(IN)
7	30252	4020	0.09039	0.90961	Chroniq Admirables	(AD)
8	16381	3132	0.04894	0.95106	Le Disciple	(DI)
9	3167	1135	0.00946	0.99054	Pantagr Prognost	(PR)
10	2569	560	0.00768	0.99232	Pronost Nouvelle	(NO)
total	334700	33328				

Figure 8. La composition du corpus Rabelais

a - La fonction première de tout hypertexte est bien entendu la recherche des contextes. Et notre logiciel *Hyperbase* y pourvoit de plusieurs façons. Il propose d'abord une fonction *Contexte* qui est analogue à celle de *Frantext*. Le mot sur lequel porte l'interrogation (ce peut être aussi une finale, une initiale, une chaîne quelconque, une expression, une cooccurrence ou une liste de formes) est recherché dans le texte et tous les paragraphes où il est repéré sont rassemblés sur l'écran. Nous ne chercherons pas la Corse, Rabelais n'ayant fait que deux allusions à l'Île de Beauté. Avec le mot *Paris* la récolte est évidemment plus abondante et les 100 contextes où la capitale est évoquée sont restitués en 6 secondes (voir figure 9).

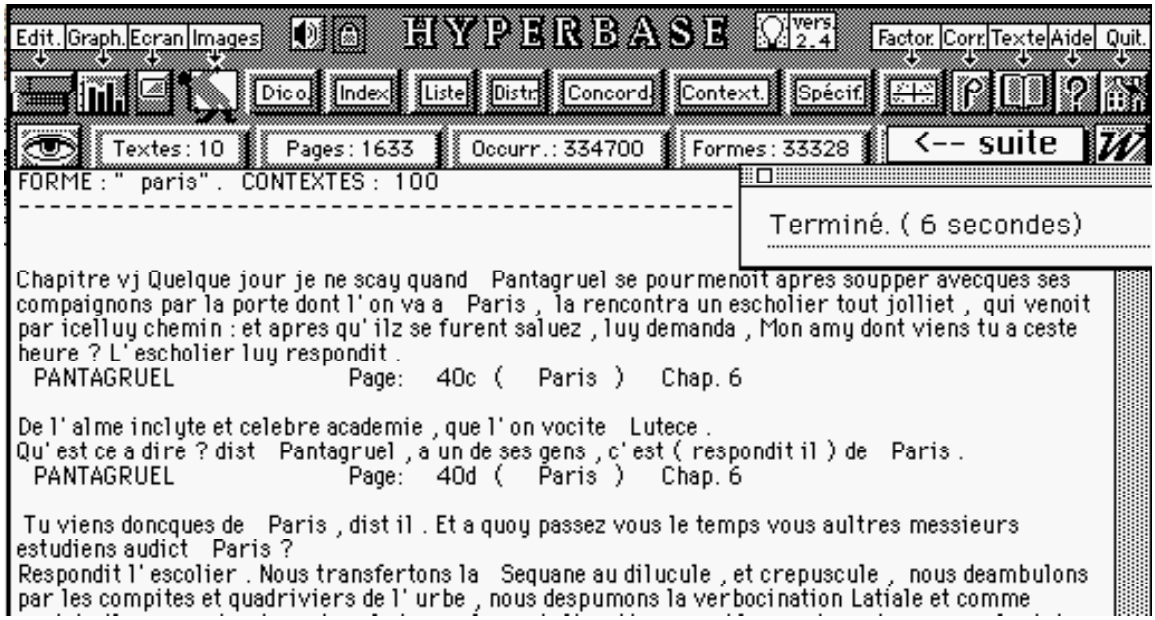


Figure 9. Les 100 contextes de Paris dans le CD-Rom Rabelais (extrait)

b - Une autre façon, plus traditionnelle, de livrer les contextes d'un mot est de les présenter sous forme de concordance, une ligne étant dévolue à chaque passage. Cette fois la restitution est immédiate et n'exige qu'une seconde, quelle que soit la fréquence du mot. Cette présentation concentrée et synoptique permet les rapprochements, surtout si l'on utilise les possibilités de tri du contexte droit ou de l'expansion gauche. Si la ligne et ses références ne sont pas jugées suffisamment explicites, un clic sur une ligne ouvre instantanément une fenêtre avec le paragraphe entier, et, si l'on insiste, toute la page. On peut par exemple hésiter sur le sens d'un extrait de la concordance du mot *Lyon* (figure 10)et rester perplexe devant

l'expression *tarabin tarebas*. Le clic libérateur élargit le discours qui étant celui d'un clerc n'est pas plus clair pour autant. Mais au moins sait-on qu'il s'agit d'un coq-à-l'âne.

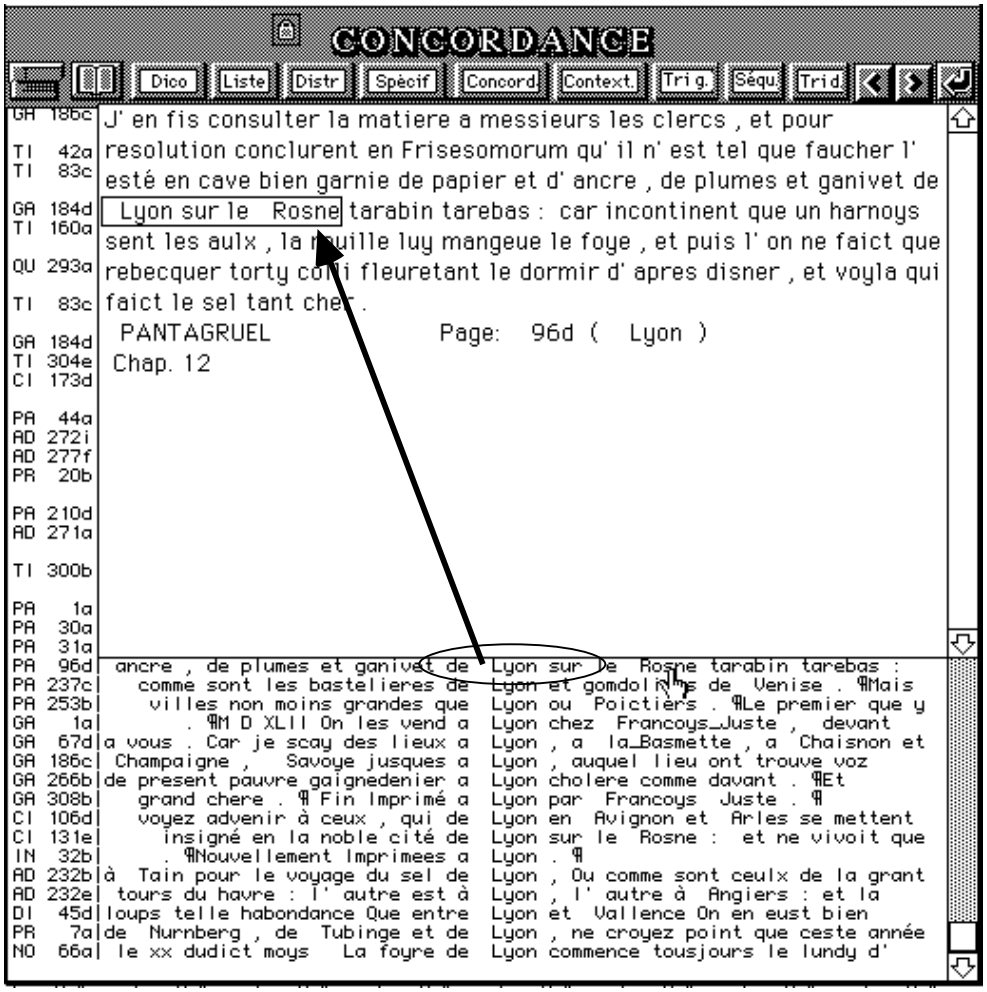


Figure 10. La fonction Concordance. Exemple de la ville de Lyon.

c - La circulation hypertextuelle peut être plus libre encore et se réaliser dans l'exercice même de la lecture. Toute page du texte est accessible immédiatement, par voie directe ou séquentielle. Si au fil du texte le regard est accroché par un mot remarquable ou mystérieux, un clic sur ce mot renvoie aux autres passages où son emploi se vérifie. Toute la surface de l'écran est ainsi sensible à la caresse de la souris, et réagit au moindre signe.

d - Une démarche complémentaire a été prévue pour ceux qui aiment voyager dans les dictionnaires. Quoique le dictionnaire ne comporte ici que des fréquences et des références et que la consultation en soit sévère, là aussi des liens sont établis avec le texte, et tout mot que l'on désigne dans la liste reporte le lecteur aux passages référencés. La Corsicque (c'est l'orthographe de l'époque pour désigner la Corse) se dessine soudain aux yeux du lecteur dans un passage où Picrochole rêve d'étendre son empire à la Sardaigne, aux Baléares et "aultres isles de la Mer Ligusticque". Voir figure 11.

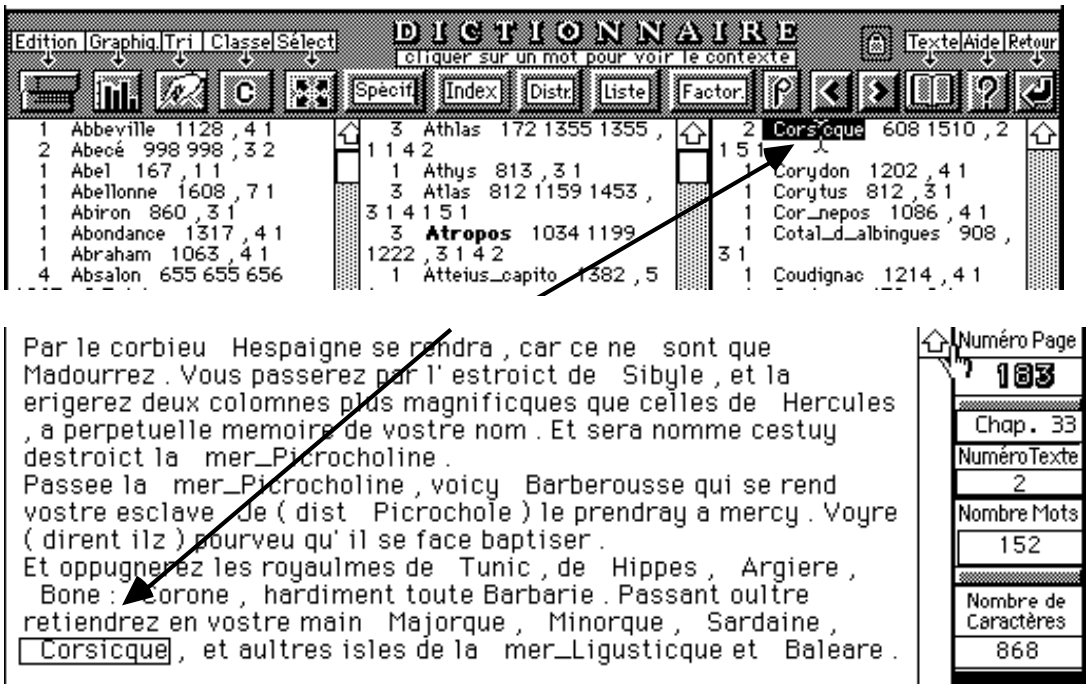


Figure 11. Le va-et-vient du dictionnaire au texte.

Outre ces outils documentaires, le CD-Rom Rabelais offre une comparaison sous forme synoptique de plusieurs éditions du même texte, du moins là où les variantes sont les plus intéressantes, c'est-à-dire dans le cas du *Pantagruel* et du *Quart Livre*. Un clic sur n'importe quel mot de l'une des versions renvoie au mot correspondant de l'autre version. Voir figure 12.

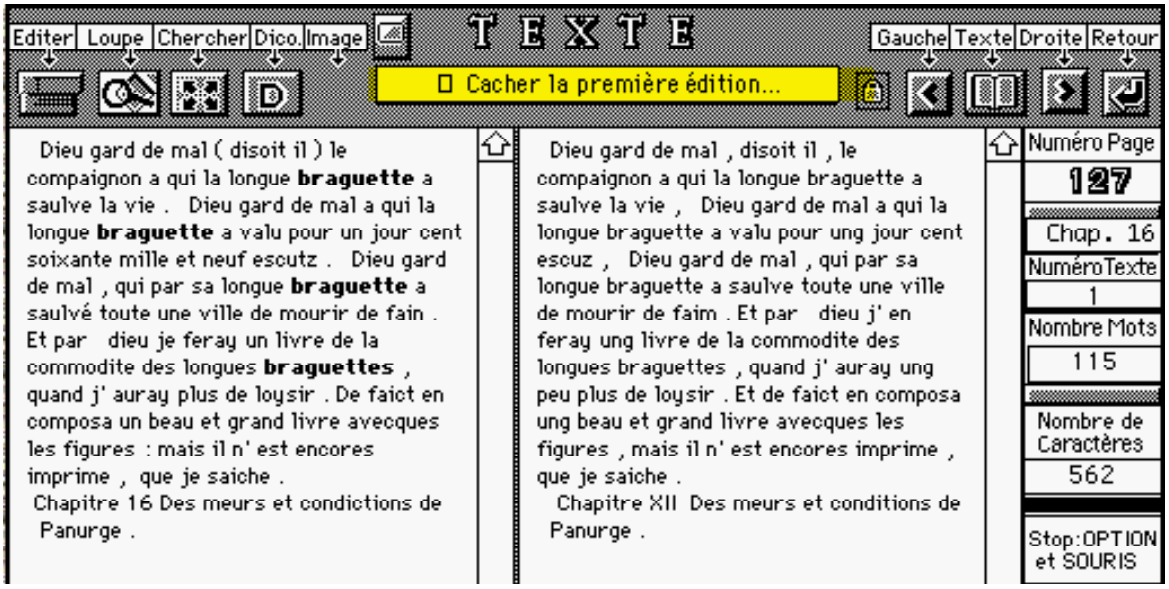


Figure 12. Comparaison de deux éditions du Pantagruel

De plus certains mots du texte de Rabelais ont été mis en relation avec les dictionnaires ou glossaires de l'époque. Ils apparaissent en caractères gras et réagissent au clic de la souris, montrant la définition du *Thresor* de Jean Nicot ou le commentaire de la *Brève déclaration*. L'italique désigne par ailleurs d'autres mots ou expressions qui bénéficient d'une explication et d'une illustration. Et de la même façon le clic sur le passage en italique fait apparaître successivement l'une et l'autre. Quatre cents documents iconographiques qui ont un rapport avec le texte du Gargantua ont été fournis par la Bibliothèque municipale de Lyon. Ils

datent tous de l'époque de Rabelais et éclairent certains aspects méconnus du texte. Ces illustrations accompagnent les pages du *Gargantua* dans leur défilement mais elles constituent aussi une base de données autonome qu'on peut consulter librement. L'exemple de la figure 13 est emprunté à un très bel ouvrage de cryptographie qui avait cours au temps de Rabelais et dont l'évocation est suscitée par l'expression *lettres non apparentes* de la page 13. Avant que l'image apparaisse, un commentaire bibliographique est proposé sur l'écran (figure 14).

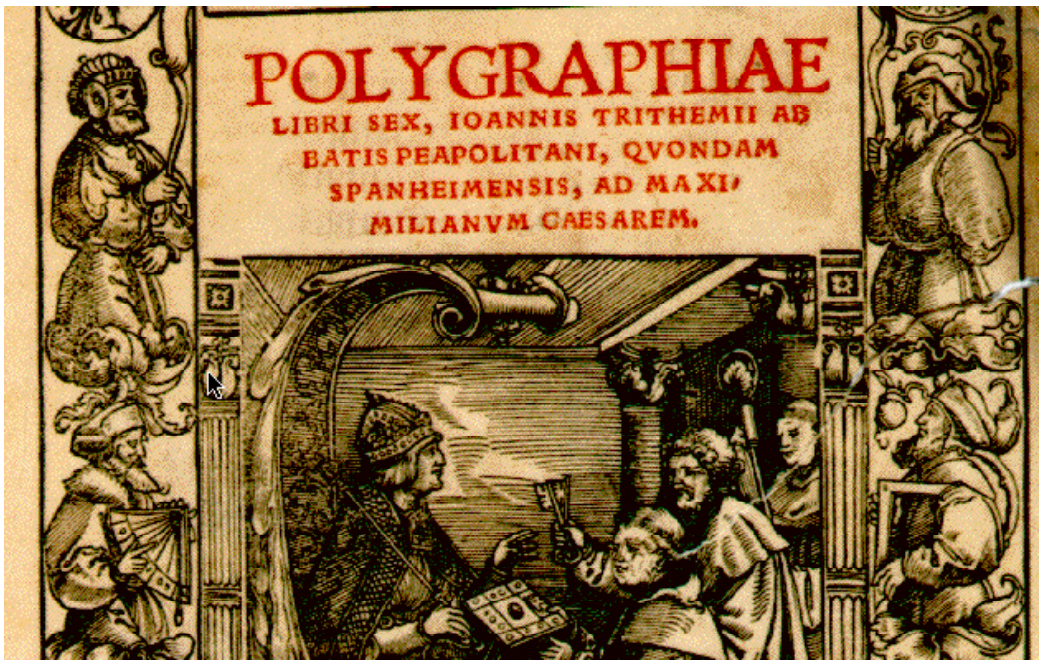


Figure 13. Un document iconographique (extrait)

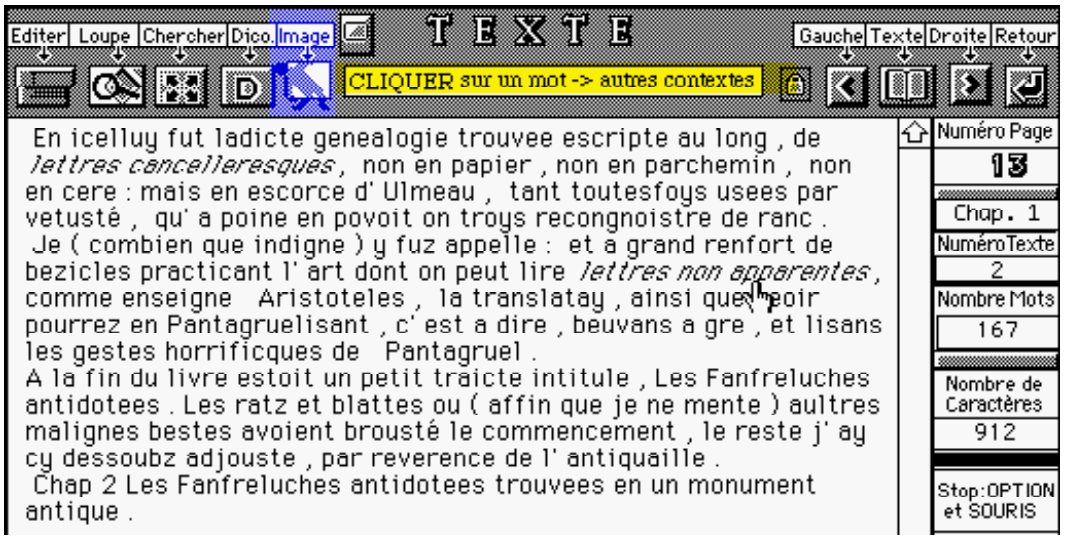


Figure 14. Le commentaire de l'image

Les facilités multimédia du CD-Rom ont été mises à profit, non seulement pour le traitement de la couleur (pourquoi s'en priver sur écran puisque le noir et blanc n'est pas plus économique), mais aussi pour l'incrustation de séquences animées et sonores qui expliquent et illustrent le fonctionnement de la base. Quand un bouton fait mystère, l'utilisateur a le moyen d'en exiger l'explication, brève ou détaillée. Un écran lui est d'abord montré avec un commentaire approprié et si cela ne suffit pas une séquence *Quicktime* lui est proposée, qui décompose les phases de l'opération, selon l'invitation reproduite dans la figure 15.

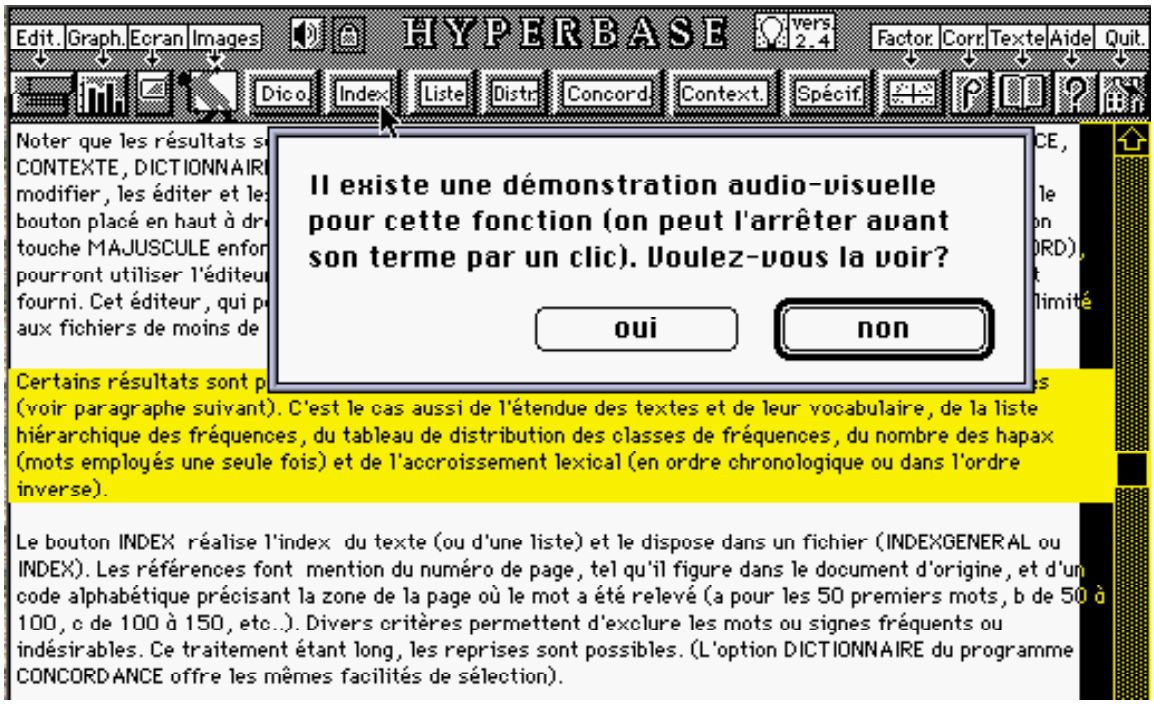


Figure 15. Les aides multimédia (la fonction Index expliquée).

Quant aux fonctions quantitatives, nous les exposerons plus loin, car elles sont communes à la présente version CD-Rom[4] et à une version nouvelle de la même base qu'on a rendue accessible par le réseau *Internet* et qu'il convient d'aborder maintenant. Ce faisant, nous fermons la parenthèse ouverte au début du colloque, en rejoignant les sujets abordés lors de la Table Ronde qui a clos les débats. Il s'agissait alors d'étudier les divers modes de diffusion des bases de données linguistiques, le CD-Rom étant l'un d'eux, parmi beaucoup d'autres, comme le papier, la microfiche, le microfilm et les multiples supports magnétiques qui foisonnent autour de l'ordinateur.

3. Un essai de bases de données linguistiques sur *Internet*

Entre la production d'une base et sa diffusion un paramètre intervient qui est d'ordre commercial. Qui dit diffusion dit distribution, circuit de vente, appareil de production, de gestion, de maintenance, maison d'éditions et de publicité, gains, profits et pertes. Les produits linguistiques dont nous avons débattu ont rarement une valeur marchande et s'il est arrivé durant ces trois jours qu'on parle d'argent, c'est pour s'en plaindre, pour regretter que les Atlas coûtent si cher à préparer, ou pour condamner l'attitude avaricieuse des éditeurs qui protègent leur copyright comme Harpagon sa cassette. Certes les industries de la langue ont une grande importance économique. Le marché des logiciels de traitement de texte ou des correcteurs d'orthographe, le marché des dictionnaires et des encyclopédies, celui de la traduction, celui des études de marché, des sondages d'opinion, des agences de presse, de la veille technologique, tout cela a à voir - peu ou prou - avec le texte et l'exploitation de l'information linguistique. Mais les bases de données linguistiques dont on a parlé dans cette enceinte représentent un marché marginal : celui de la

---

[4] Au moment d'abandonner la présentation - incomplète - de ce CD-Rom, nous renvoyons le lecteur à un article plus circonstancié "Le CD-Rom Rabelais", paru dans *Travaux du Cercle de Linguistique de Nice*, n° 11, 1995, p. 43-79, 1994 (cet article se trouve aussi in extenso sur *Internet* à l'adresse: *http://ancilla.unice.fr*). Bien entendu un manuel plus explicite encore fournit 130 pages d'explication aux acquéreurs du CD-Rom. Parmi ces explications figure un mode d'emploi de la base sur *Internet*, qui permet ... de se dispenser d'acquérir le CD-Rom. Quant au logiciel *Hyperbase* qui a servi à constituer la base Rabelais, sous ces différentes formes, son destin est indépendant. Il est commercialisé en dehors de toute donnée, prêt à recevoir celles du chercheur (s'adresser au laboratoire).

dialectologie, qui n'est pas la voie royale pour faire fortune[5]. Les considérations économiques prêchent donc pour une diffusion bon marché de nos bases et cela peut orienter le choix du côté d'*Internet* où règne encore une sorte de gratuité universitaire[6].

Le mot diffusion peut d'ailleurs prêter à confusion à cause de son préfixe *dis* qui implique dispersion, éclatement, mouvement centrifuge et multiplication. Or toute diffusion de ce type est nécessairement coûteuse en fabrication, en maintenance, en transports, alors qu'on a la possibilité d'inverser le mouvement. Au lieu de dupliquer l'original pour le répandre en autant d'exemplaires qu'il en faut pour contenter les clients, il suffit d'inviter la multitude, grâce aux télécommunications, à assister à un événement unique ou à utiliser un produit non reproduit, tiré à un seul exemplaire. C'est le principe de la radio et de la télévision. On assiste alors à un mouvement centripète. Au lieu de diffusion il faudrait parler de con-fusion si le mot n'avait pas un autre sens. L'information n'est d'ailleurs pas la denrée unique qui se prête aux deux systèmes de distribution: l'eau, le gaz, l'électricité sont proposés sous les deux modes: en réseau (EDF, GDF, Cie des Eaux) ou en paquets (bonbonnes de gaz, piles électriques ou bouteilles d'Evian). Pour ces trois fluides, l'économie, en temps, en énergie et en dépense, est du côté du réseau. Ainsi en est-il de l'information, qui revient moins cher en tuyaux qu'en bouteilles.

C'est pourquoi nous avons tenté l'expérience du *Web*, à partir de notre base de données *Rabelais*. C'était sans grand risque, le *CD-Rom Rabelais* étant parvenu en deux mois à la rentabilité (il est vrai facile à assurer, vu que les réalisateurs n'exigeaient ni salaire, ni bénéfice)[7]. Cela pouvait avoir même des avantages, la clientèle d'*Internet* assurant une plus large ouverture et pouvant faire office de publicité en faveur du CD-Rom. Car si la base est bien la même, le confort est plus grand quand on dispose du CD-Rom pour soi seul, puisqu'on n'a pas à partager avec d'autres le temps, les lignes et les ressources du serveur. Mais si les fonctionnalités sont moins riches, moins rapides et moins puissantes, elles ne diffèrent guère dans leur principe. Les actions documentaires sont menées grâce à un unique écran, qui propose des boutons à cocher, des menus à dérouler, des zones à remplir (par exemple pour y inscrire le mot ou le suffixe demandé) et aussi, ce qui est propre au *Web*, des *ancres* (ou *links*) à solliciter. Ces renvois apparaissent généralement en bleu et toujours avec le soulignement. Ils réagissent à l'action de la souris, en allant quérir le fichier demandé pour le mettre sur l'écran. Cette procédure convient très bien pour les documents iconographiques, les commentaires associés et la lecture du texte même du corpus, en sorte que les figures 13 et 14 issues du CD-Rom sont obtenues aussi facilement par le réseau. Les interrogations qui portent sur les spécificités lexicales ou la structure du vocabulaire parviennent immédiatement au résultat, puisque les fichiers de réponse, dans cette situation particulière, ont été préparés à l'avance.

Ce n'est plus le cas pour les questions ouvertes: graphique, concordance, recherche de contexte ou de cooccurrence, puisque l'objet à traiter est imprévisible et peut être n'importe quel mot ou chaîne du corpus. Dans une telle situation, il ne suffit plus de préparer des pages d'information, comme on écrit un livre, avec pour seule contrainte le respect du balisage imposé par le code *Html*. Encore faut-il prévoir des programmes interactifs pour gérer les appels, qui viennent par le réseau sous forme de grappes d'arguments, et fournir les réponses adaptées. Sans entrer dans le détail technique, on devine que l'interface homme-machine a nécessité de profonds changements et que beaucoup des 50000 lignes de

---

5 Ce n'est pas non plus la voie normale - même si on l'emprunte parfois - pour mener une politique linguistique. La diffusion des bases de données dialectologiques ne poursuit par un but d'évangélisation. Le prosélytisme en faveur des langues régionales peut être un noble but, ce n'est pas celui de la recherche.

6 Entendons par là qu'une fois le forfait payé à l'année pour le rattachement au réseau, la dépense n'est pas proportionnelle à l'usage. L'abonnement donne droit à un débit constant - comme l'eau distribuée "à la jauge" dans certaines campagnes méridionales. Le prix sera le même, qu'on ouvre ou non le robinet. Ajoutons que le forfait ne se négocie pas à titre individuel, mais à l'échelle de l'Université, de l'entreprise ou de l'administration.

7 La commercialisation est assurée par *Les Temps qui Courent*, 118-130 av. J. Jaurès, 75019 Paris.

code du CD-Rom Rabelais ont dû être réécrites. En fin de compte les résultats documentaires, concordances ou contextes, parviennent au demandeur dans la même présentation que précédemment (voir figures 9 et 10) et presque aussi rapidement. Il faut cependant ajouter au temps de la recherche proprement dite le délai supplémentaire du transcodage en document *html*[8] et le temps du transfert sur la ligne - lequel est tributaire des conditions du trafic.

Nous ne produirons ici que la partie statistique, elle aussi semblable à la version CD-Rom. Certaines questions traditionnelles auxquelles la lexicométrie sait répondre peuvent se traduire ainsi:

- qu'est-ce qui différencie le corpus ou tel texte du corpus?
- y a-t-il moyen de reconnaître Rabelais parmi ses contemporains?
- quelle est la distribution d'un mot ou d'un groupe de mots, par exemple d'un suffixe ou d'un thème?

Les écrans ci-dessous proposent des réponses circonstanciées à ces questions. La place nous manque pour en livrer le détail. On trouvera dans la figure 16 le sommaire des résultats relatifs à la richesse du vocabulaire et à la connexion lexicale (on appelle ainsi avec Ch. Muller, la distance qui sépare chaque texte de tous les autres quand pour chaque couple de textes on mesure la part commune du lexique et la part exclusive à chacun).

Figure 16. La structure du vocabulaire

---

[8] Si la réponse est encombrante (au delà de 30 000 caractères) le trancodage est assuré par un programme que nous avons écrit spécialement pour la circonstance et dont le débit est rapide (une seconde suffit pour 100 000 caractères). Ceux qui viendraient à s'étonner de la nécessité de ce transcodage doivent comprendre que le réseau *Web* est indépendant des standards, et qu'on interroge à partir d'une station *Unix*, d'un compatible *PC* ou d'un *Macintosh*, le serveur livre les mêmes informations où les signes diacritiques reçoivent un traitement universel (par exemple *é* est écrit *é*). Les logiciels *Mosaic* ou *Netscape* dont on se sert généralement sur le réseau *Web* savent traduire ce codage et l'adapter aux particularités de la machine du client, si bien que le résultat est constant - ce qui constitue un grand avantage.

La figure 17 montre sous ce rapport le profil du *Cinquième Livre* dont l'attribution à Rabelais fait problème. L'analyse du lexique montre que ce livre contesté est très proche du *Quart Livre* et la présomption statistique incline à penser que c'est la même plume qui a écrit l'un et l'autre livres.

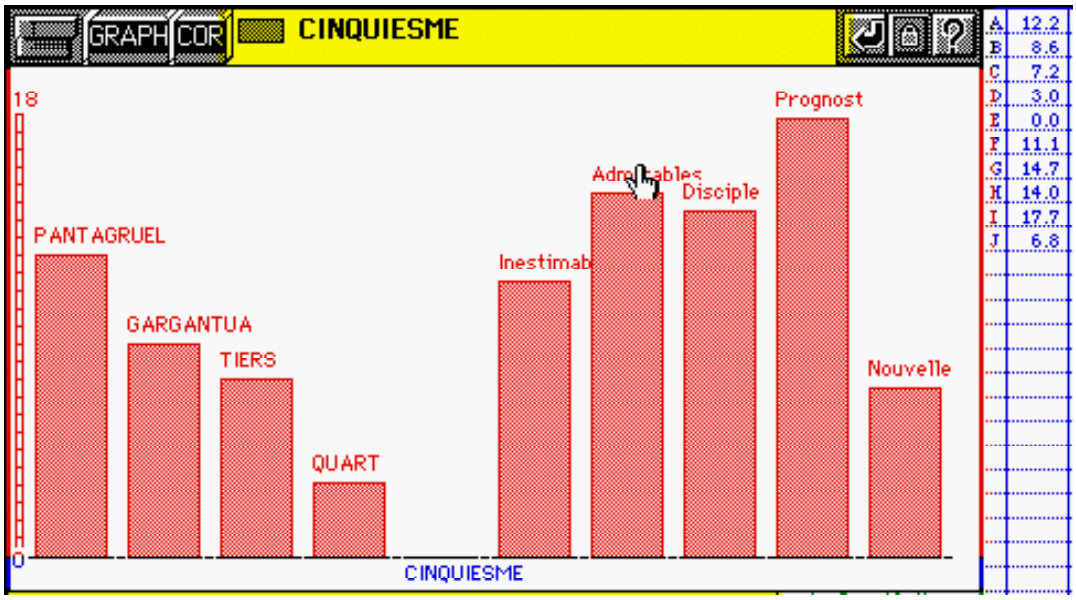


Figure 17. La distance du Cinquième Livre à tous les autres textes

Les spécificités de la base Rabelais (par rapport à un ensemble plus vaste constitué par les textes de *Frantext* appartenant au 16e siècle) ou celles qu'on peut relever à l'intérieur du corpus en opposant les textes les uns aux autres sont accessibles d'un simple clic sur une des lignes de la figure 18. On obtient alors une liste triée des formes significativement excédentaires ou déficitaires dans le texte considéré. De telles listes dessinent comme un portrait, fait de reliefs et d'ombres, du texte en question. Mais elles sont longues et encombrantes et nous ne donnerons ici que les têtes de liste: pour *Pantagruel* c'est le couple Pantagrel-Panurge; pour *Gargantua* c'est Grangousier et Picrochole mais aussi les *fouaces*, les *pèlerins*, l'*abbaye*, les *moynes* et les *ennemys* - grâce à quoi il est facile de reconnaître certains épisodes célèbres de la guerre picrocholine; pour le *Tiers Livre* la trilogie *cocu-femme-mariage* indique assez la problématique mise en oeuvre, tandis que les *chicanous* et les *andouilles* nous plongent dans l'univers mythique du *Quart Livre*, en compagnie de *Frère Jan*.

Figure 18. Le vocabulaire spécifique. Sommaire.

Le traitement des listes de mots est ouvert à toutes les combinaisons. On constitue ainsi des tableaux à deux dimensions où les lignes renvoient aux mots et les colonnes aux textes. À l'intersection d'une ligne et d'une colonne, il faut donc lire la fréquence du mot *x* dans le texte *y*. La sélection peut être manuelle, le chercheur énumérant dans le champ B (figure 19) les formes qui l'intéressent. Elle peut être aussi automatique et s'exercer sur les critères retenus par le chercheur: initiale, finale, chaîne quelconque ou seuil de fréquence. La liste obtenue peut être copiée et soumise à toutes les manipulations auxquelles se prêtent les tableaux de contingence. Mais les options proposées par le programme réalisent les plus courantes et principalement l'histogramme d'une ou deux lignes, d'une ou deux colonnes, ou du total en ligne ou en colonne.

Figure 19. Les listes de mots

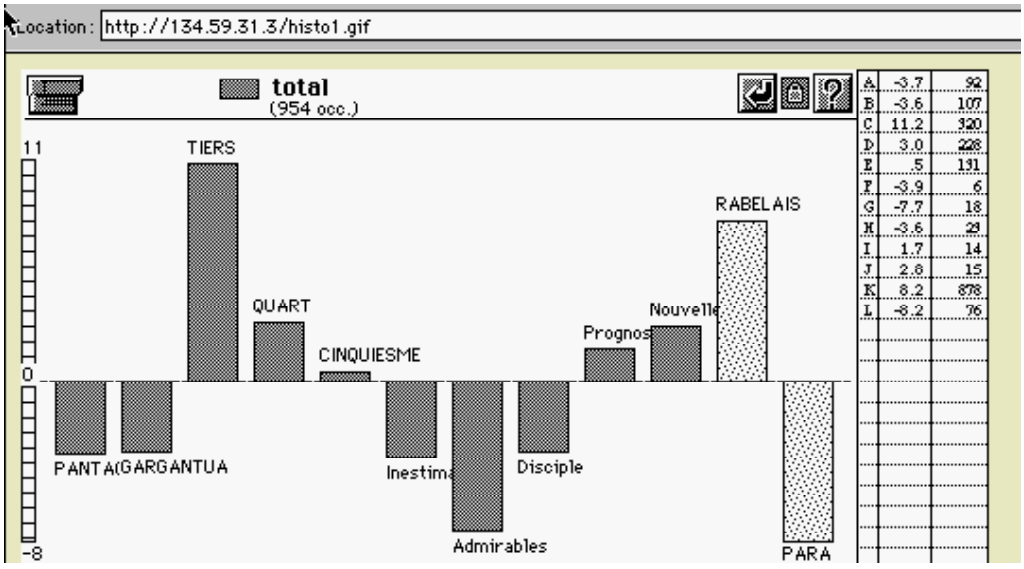


Figure 20. Histogramme des mots en -tion (322 unités, 954 occurrences)

Ainsi trouvera-t-on dans la figure 20 la courbe cumulative de tous les mots en *-tion*, dont l'emploi se répand à partir du *Tiers Livre,* quand la visée satirique et intellectuelle - qui fait usage des suffixes abstraits - supplante les propos pour rire des premiers livres.

Deux tableaux particuliers - dont la constitution a été réalisée à l'avance - sont immédiatement disponibles: l'un correspond aux groupes de fréquences et pour chaque texte établit l'effectif des mots rares et moins rares, soit 11 groupes échelonnés à partir de la fréquence observée dans le TLF (1: 512 ou moins, 2: de

512 à 1024, 3 de 1024 à 2048, etc..). L'autre reproduit le tableau de la longueur des mots: lg1=mots de 1 lettre, lg2=mots de 2 lettres, etc..[9]. Le tableau 21 reproduit cette distribution, dont l'éclairage collectif est fourni par les analyses factorielles. Le dialogue de la figure 19 en propose trois variétés, selon qu'on souhaite ou non pondérer les données. L'éclairage qui souligne le mieux les reliefs est 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 ou groupes de mots retenus ne précipitent au centre du graphique les éléments les plus lourds et les plus aptes à faire la loi.

À partir des données du tableau 21, il suffit de 10 secondes pour obtenir le résultat de la figure 22, obtenue avec le filtre de l'écart réduit. On voit dans la position du *Quart Livre (*dans lequel on rencontre deux "mots" de plus de 50 lettres à la page 106), à gauche de l'écran, au voisinage des mots longs[10], la confirmation de l'histogramme 20, qui envisageait la suffixation en *-tion* et donc englobait des mots plutôt longs. Mais cette caractéristique est commune aux textes de la dernière période. À l'opposé, sur la marge droite, se situent les textes qui ont précédé ou suivi Rabelais et qui privilégient les mots courts. Au centre le *Gargantua* et le *Pantagruel* représentent un compromis où Rabelais reste encore proche de ses devanciers et où la farce populaire n'a pas cédé la place au conte philosophique.

Retour au sommaire

	TOTAL	ecart	Pant	Garg	Tier	Quar	Cinq	Inse	Admi	Disc	Prog	Nouv	RABE	PARA
lg1	7599	-6.2	1111	1275	1810	1214	809	176	645	399	73	87	6219	1380
lg2	74263	-26.9	10964	11504	13755	14184	9162	2202	7247	3947	678	620	59569	14694
lg3	40105	-16.4	5890	6319	7193	7118	4805	1370	4449	2493	335	133	31325	8780
lg4	34707	-10.1	4994	5159	5897	6250	4606	1007	3878	2150	286	480	26906	7801
lg5	31016	13.6	4156	4720	5562	6001	3993	934	3381	1735	286	248	24432	6584
lg6	28600	33.0	4010	4472	5130	5577	3718	761	2701	1709	262	260	22907	5693
lg7	21540	16.0	2848	3458	4179	4509	2865	528	1754	1034	188	177	17859	3681
lg8	16699	16.9	2211	2744	3230	3730	2238	339	1260	663	155	129	14153	2546
lg9	18211	12.4	2212	2671	3923	4123	2661	375	1239	721	193	93	15590	2621
lg10	7048	7.4	937	1059	1621	1602	986	115	400	205	88	35	6205	843
lg11	740	-3.5	109	89	183	191	100	4	26	13	13	12	672	68

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

Tableau 21. La distribution des mots selon leur longueur

Ce critère de la longueur du mot, dont on a tout lieu de penser qu'il échappe à la conscience linguistique de l'auteur aussi bien que du lecteur, n'en distingue pas moins les textes avec netteté.

---

9 Il y a regroupement des mots longs, vu leur faible nombre: la classe *lg9* mêle les mots de 9 et 10 lettres, la classe *lg10* ceux de 11 à 13 lettres et la classe *lg11* ceux qui sont plus longs encore.

10 La position des points ne va pas parfois sans recouvrement: ainsi les points lg10 et lg11 sont superposés comme les points lg5 et lg6. Quand on a rapatrié le résultat, on a tout loisir pour séparer les points doubles et compléter les symboles.

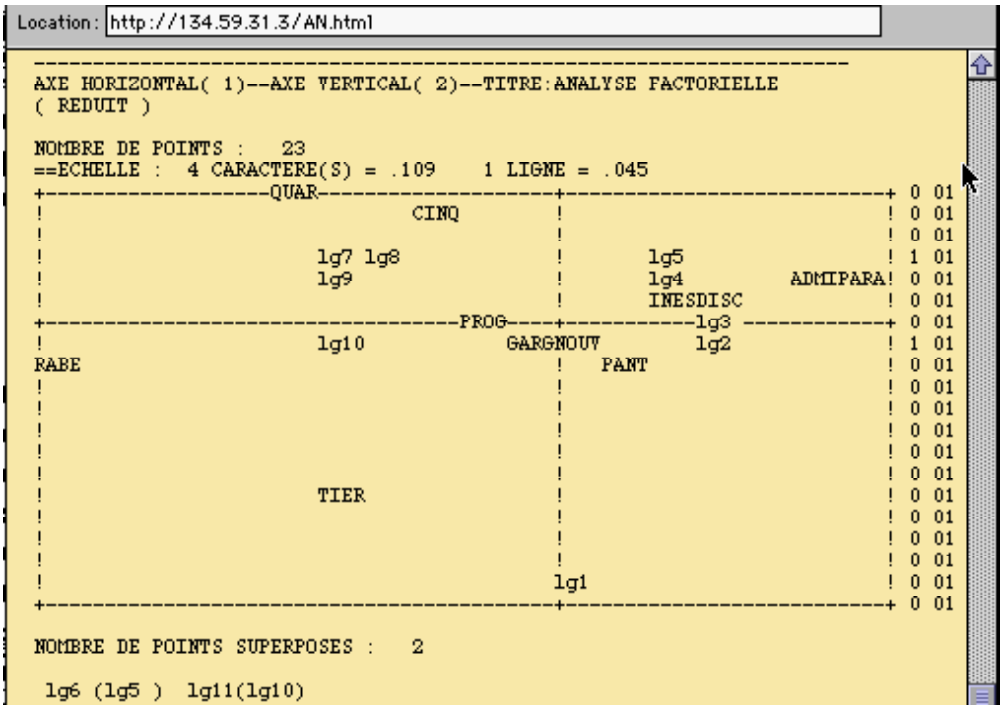


Figure 22. Analyse factorielle de la longueur du mot

****

On arrêtera là le catalogue. Bien d'autres exemples se trouvent sur *Internet*, dont la plupart cependant sont du type dictionnairique ou bibliographique. Car les bases élaborées sur le texte intégral sont encore rares. Elles correspondent d'ailleurs plus souvent à des visées historiques ou littéraires que proprement linguistiques. Mais la linguistique s'est trop longtemps appuyée sur des exempliers fabriqués de toutes pièces, à partir de la conscience linguistique du chercheur, pour qu'on puisse se passer des témoignages fournis par les textes, même s'il arrive que les textes soient indociles, par action ou par omission, quand ils proposent des exceptions non désirées ou qu'ils restent obstinément muets sur des faits évidents. Dans le domaine de la dialectologie et des langues régionales, les textes font parfois défaut et les témoignages sont souvent oraux. Mais qu'une langue vienne à disparaître, après quelques générations, seuls survivront les vestiges de l'écrit.

Reste à savoir si *Internet* est un recours ou un danger. Quand les spécialistes d'une discipline sont rares et dispersés ou quand manquent les moyens de diffusion traditionnels, le réseau offre une alternative intéressante et immédiatement disponible - du moins quand les bureaux des chercheurs seront convenablement reliés et que les bâtiments seront aussi "intelligents" que leurs occupants[11]. Mais on peut éprouver quelques craintes si l'on considère que notre discipline linguistique, et particulièrement la branche dialectologique, est essentiellement pluraliste et repose sur le postulat que toutes les langues sont égales en dignité. Observons que le titre de ce colloque comporte six mots, tous au pluriel, et qu'il se tient éloigné de l'universalisme réducteur. À cet égard le réseau mondial peut désenclaver les particularismes et leur permettre de communiquer, même à l'autre bout de la planète - ce qui est un bien -,

11 Cependant un modem rapide branché sur le téléphone peut suffire.

mais en même temps *Internet* tend à unifier le monde et à imposer un langage unique (l'anglais bien sûr), un dialogue réglementé, un standard passe-partout. Voilà que la malédiction de Babel est en passe d'être conjurée. Est-ce aussi un bien ? Les dialectologues, si soucieux de la variété, peuvent en douter.

On peut craindre aussi que les machines communiquent de plus en plus, et les hommes de moins en moins. Le temps est venu des colloques virtuels, où les réseaux véhiculeront en direct la voix et le visage du congressiste, mais non sa présence et son odeur, et où le chercheur ne voyagera plus que dans sa chambre, tandis que les données feront la farandole sur les ondes en tourbillons légers[12]. Quel dommage ce serait de perdre l'odeur, surtout en Corse, au milieu du maquis. Aussi bien en fermant la parenthèse de ce Colloque, ne fermé-je point la porte, dans l'espérance du retour.

Nous reviendrons en Corse,

À l'hôtel Colonna[13].

La science nous y force

Et l'accueil que l'on a.

---

12 Ah! quelle belle ronde
On dansera demain,
Quand les données du monde
Se donneront la main!

13 L'hôtel Colonna - qui recevait royalement et démocratiquement les congressistes - doit son nom à un célèbre gardien de but, reconverti en gardien des traditions corses.