L'Afrique et l'ordinateur

Étienne Brunet

Etienne Brunet

Institut National de la langue française (UPR 6861, CNRS)
98 bd Herriot 06204 Nice

L'*Afrique* et l'ordinateur

J'ai deux ou trois raisons de me réjouir de ces Journées consacrées aux particularités lexicales du français en Afrique. D'abord la présente manifestation prolonge avec bonheur un colloque que j'avais eu l'honneur d'organiser à Nice, avec Danièle Latin, il y a près de cinq ans, et dont certains participants se souviennent peut-être sans trop d'amertume. En second lieu le laboratoire dont j'ai eu longtemps la charge s'est investi dans le programme *Francil* de l'Aupelf-Uref pour une enquête sur le français écrit au Maghreb et ces Journées m'offrent l'occasion de confronter les méthodes qui conviennent le mieux à ce genre d'étude. Enfin je trouve là une opportunité pour entrer en contact, pour la première fois de ma vie, avec l'Afrique profonde, dont je ne connaissais jusqu'ici que la frange méditerranéenne. Aussi ai-je préparé mon voyage, avec l'ardeur des néopohytes et des catéchumènes, en consultant une agence de voyage d'un type particulier: *Internet*.

- I -

Chacun sait qu'*Internet* est un énorme magasin d'informations de toute sorte: touristiques, scientifiques, commerciales. Dans ce souk à l'échelle de la planète, où l'utile voisine avec le futile, on peut trouver matière à réflexion, et parfois à découverte, si l'on sait naviguer dans cet océan de mots, parmi les images, les logos et la publicité envahissante. Les pages du *Web* constituent un immense texte discontinu et c'est une 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 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, et avec des automates installés à la surface du Net, ne pourrait-on pas découvrir, dans ce flux, des remous, des courants, des marées?

---

[1] *Lycos*, qui est le plus puissant moteur de recherche branché sur *Internet*, avoue explorer 16,5 millions de documents (la progression est foudoyante et augmente actuellement d'un million par mois). 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.

Choisissons par exemple la représentation des villes et des pays, en nous intéressant particulièrement à l'Afrique. Comment les différentes régions du globe sont-elles représentées dans *Internet*? Quelle est leur importance relative? Leur poids dans la communication est-il en rapport avec celui de leur population, de leur économie ou de leur célébrité? Et l'Afrique a-t-elle quelque part dans le concert des nations électroniques? Pour ce faire consultons *Lycos* qui est un moteur de recherche spécialisé apte à renseigner l'utilisateur sur la grande majorité (91%) des documents disponibles sur *Internet*. À la question formulée dans la figure 1 et relative au pays qui nous reçoit présentement, une réponse est donnée dans la figure 2.

Figure 1. L'interrogation de *Lycos* (16,5 millions de documents)
à propos du *Cameroun*

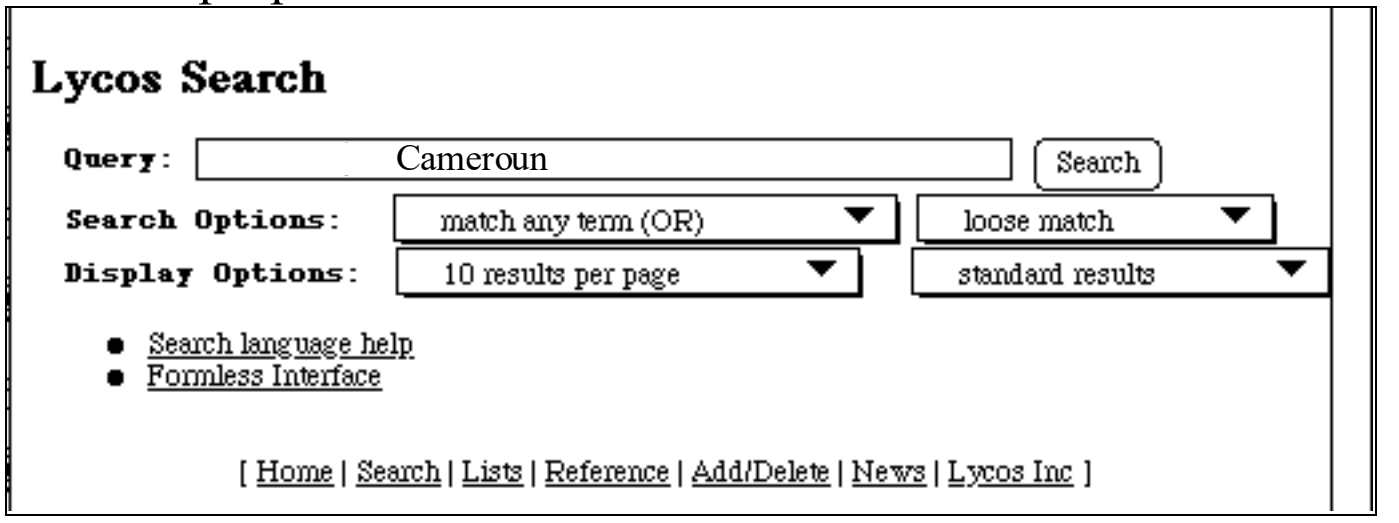


Figure 2. La présence du Cameroun sur Internet

Found 3 matching words (number of documents): cameroun (158), camerouns (1), camerounse (2)

3) B. ÉTUDES RÉGIONALES [0.9865]

Outline: B. ÉTUDES RÉGIONALES

Abstract: B1. AFRIQUE CHAHNAZARIAN Anouch Ed., Population et Santé à Niakhar. Niveaux et tendances des principaux indicateurs démographiques et épidémiologiques de la zone d'étude, 1984-1991. ORSTOM, Dakar, 1992. HERRY C., La démographie des pêcheurs du Delta Central du Niger , 1992, Bamako, 60 p. GENDREAU Francis, Quelques réflexions sur la démographie de l'Afrique au sud du Sahara , in IFRI, "Sociétés africaines et développement", Masson, Paris, 1992, pp. 27-43. GUBRY Patrick, **Cameroun**. In LOPEZ-ESCARTIN (Nuria), **Cameroun**. Paris : CEPED, 1991, 11 p., pp. 34. (Données de base sur la population, n.1). AGOUNKE
http://www.ostrom.fr/st6/demographes/F.html (13k)

---

4) gopher://rain.psg.com:70/1m/networks/connect/countries/cm [0.9830]

Abstract: Select one of: * Known nodes in Cameroon (CM) 93.4.21 * RIO- Information * Status of RIO Network 93.11.01 * OCLB CM - ORSTOM RIO node in Cameroon * Orstom Node in **Cameroun**
gopher://rain.psg.com:70/1m/networks/connect/countries/cm (1k)

---

5) ∞J∞™∞—™[∞„—¤˜aŸŒHome Page [0.9817]

Outline: ∞J∞™∞—™[∞„—¤˜aŸŒHome Page ∞J∞™∞—™[∞„—¤˜aŸŒ

Abstract: $B%+%a%k!
Embassy of The Republic of **Cameroun** in Tokyo
Enbassy of The Republic of **Cameroun** in Tokyo
Home Page of the Republic of Cameroon in Japan (Cameroon)
∞J∞™∞—™[∞„—¤˜aŸŒŒalg…U („'—1)

http://anakuria.himeji-du.ac.jp/Public/FLC/html/cameroun/home.html (6k)

On peut être déçu par la minceur des résultats (158 mentions seulement). Mais il y a un recours. Observons en effet que la question a été posée en utilisant la graphie française. Certes il y a peu de chance qu'*Internet* connaisse l'appellation locale et l'on voit dans la figure 2 combien le japonais pose problème au dialogue. Mais la clé anglaise ouvre davantage de portes que le nom français, et c'est 1246 documents qui font mention du *Cameroon* orthographié à l'anglaise, sans compter 98 emplois de *Yaoundé* (avec ou sans accent).

Le cas du Cameroun n'est pas unique et l'étude des doublets pourrait ainsi 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 26673 pour la graphie anglaise *London*). Ce simple rapport 218/26673 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 92/17455, soit 0,5% [2]. Et comme le présent colloque est consacré à l'étude du français en Afrique, il peut être intéressant de mener une enquête sur *Internet* pour apprécier la part du français dans les échanges africains. On s'est borné à une liste synoptique qui met en parallèle les scores obtenus en français et en anglais, la dernière colonne de la figure 3 établissant le quotient. Dès qu'il est question de l'Afrique, ce rapport est moins défavorable au français que lorsqu'il s'agit du reste du monde et la proportion du français double tout en restant très minoritaire (aux environs de 2%). Il ne faut guère faire confiance au témoignage ambigu de la Côte d'Ivoire, qui se classe en tête, à cause de la confusion entre le nom propre et le nom commun. De même avons-nous récusé le score trop flatteur que notre ville de *Nice* obtient dans l'enquête qui va suivre, par suite d'une homographie fâcheuse qui donne au toponyme la plus-value injustifiée de l'adjectif *nice*. Deux pays se détachent en ce qui concerne l'attachement à la langue française: le Maroc et le Cameroun, avec 13% des échanges. Ce n'est donc pas un hasard si les présentes journées ont lieu ici[3].

Mais en réduisant les doublets et en neutralisant les variations orthographiques (toutes les appellations d'un même site étant regroupées), on pourrait aussi dresser la carte *Internet* selon la place que chaque région du globe y occupe. La réponse est dans la figure 4, dont l'évidence décourage les commentaires. Les cités nord-américaines (on a choisi de mener l'enquête sur les villes) monopolisent les premières places et si *London* se hisse au troisième rang (derrière *Washington* et

---

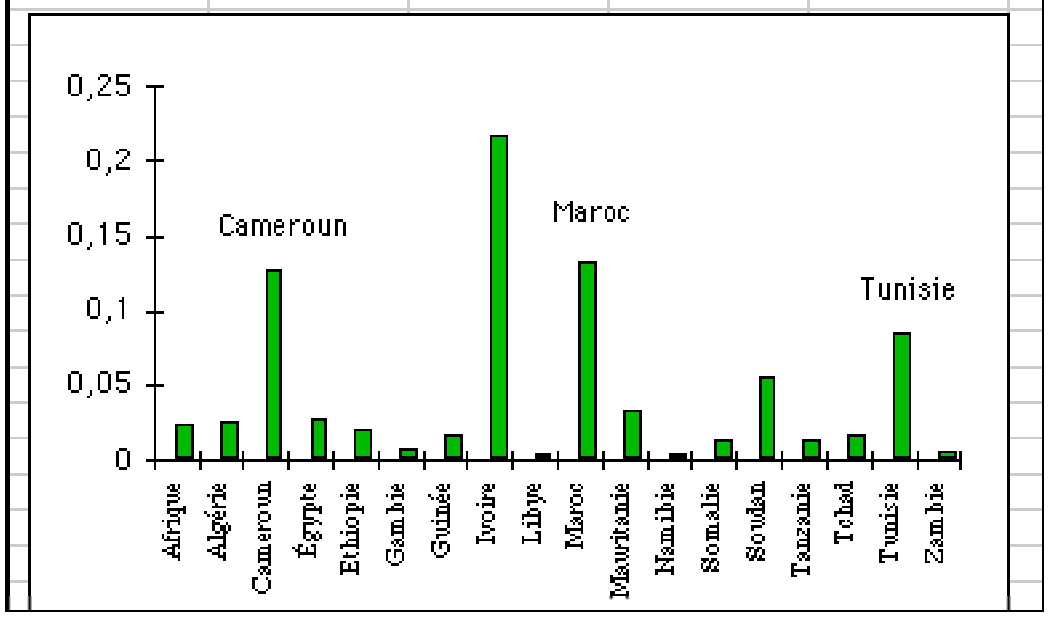
[2] Ce ne sont là que des indications fragiles, quoique vraisemblables. En pareille matière, il faut éviter le biais des villes nationales. Car les documents écrits en italien citent évidemment plus souvent que les autres langues les villes italiennes. Le rapport, d'ailleurs très variable, *Venezia/Venice* 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 *Genève* (876 contre *Geneva* 7004).

[3] Rappelons toutefois que ce classement ignore les pays, comme le Sénégal, le Congo ou le Gabon, dont la graphie est commune aux deux langues.

New York), *Paris* ne se situe qu'en dixième position et *Tokyo, Berlin* et *Melbourne* aux rangs 15, 16 et 20.

Figure 3. Part respective du français et de l'anglais dans les toponymes africains mentionnés sur *Internet*

Afrique	608	Africa	24304	0,0250
Algérie	45	Algeria	1688	0,0267
Cameroun	158	Cameroon	1246	0,1268
Égypte	165	Egypt	6055	0,0273
Ethiopie	39	Ethiopia	1859	0,0210
Gambie	6	Gambia	729	0,0082
Guinée	52	Guinea	3139	0,0166
Ivoire	693	Ivory	3187	0,2174
Libye	3	Libya	952	0,0032
Maroc	284	Morocco	2148	0,1322
Mauritanie	23	Mauritania	701	0,0328
Namibie	5	Namibia	1051	0,0048
Somalie	19	Somalia	1463	0,0130
Soudan	83	Sudan	1476	0,0562
Tanzanie	22	Tanzania	1636	0,0134
Tchad	58	Chad	3560	0,0163
Tunisie	118	Tunisia	1378	0,0856
Zambie	8	Zambia	1324	0,0060



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, Persépolis, Byzance*, ou de cités contemporaines établies sur des ruines antiques: *Jérusalem*, *Athènes* 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, les nations les plus développées accaparant les autres. Dans cette course à la modernité on ne trouve guère qu'un seul représentant africain: *Cairo*[4]. Là encore on peut estimer que les rares fois où les cités ou pays africains apparaissent dans *Internet* c'est comme objet du discours plutôt que comme émetteur des informations. En somme c'est le colonialisme qui se perpétue sous une forme nouvelle[5].

---

[4] Encore le Caire est-il à la jonction de l'Asie et de l'Afrique.

[5] En ce qui concerne les pays africains, voici dans son état brut le catalogue d'Internet qu'il ne nous appartient pas de commenter:

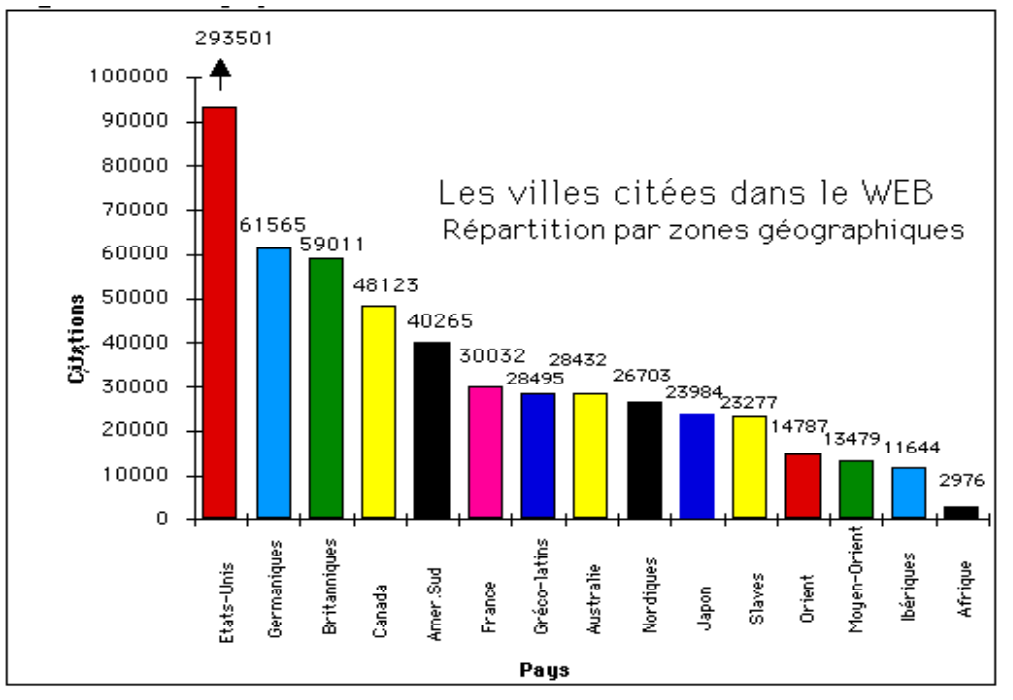
Figure 4. Les citations toponymiques d'*Internet*
classées par ordre de fréquence décroissante

Washington	64057	Sydney	8256	Glasgow	4402	Perth	2720	Granada	1424
New York	51593	Amsterdam	8178	Oslo	4337	Kobe	2712	Buenos Air.	1422
London	28891	Geneva	7817	Lyon	4120	Sao Paulo	2479	Lima	1395
Chicago	28323	Rome	7811	Petersburg	3989	Haiti	2375	Istambul	1295
San Francisco	24417	Edinburgh	7201	Athens	3905	**Cairo**	2292	Kiel	1263
Mexico	23244	Quebec	7133	Bern	3517	Basel	2148	Leipzig	1224
Boston	21948	Manchester	6885	Liverpool	3502	Lausanne	2145	Porto	1110
Los Angeles	18160	Nouvelle Orl.	6734	Venice	3461	Santiago	2061	Jakarta	1071
Seattle	17920	Dublin	6224	Prague	3378	Torino	1970	Marseille	1041
Paris	17455	Moscow	6056	Kyoto	3357	Belfast	1906	Firenze	1024
San Diego	16794	Brussels	5899	Winnipeg	3258	Nagoya	1895	Sofia	997
Toronto	16738	Wien	5638	Madrid	3160	Monaco	1859	Kawasaki	994
Buffalo	15902	Canberra	5389	Bonn	3158	Delhi	1856	Madras	942
Atlanta	15459	Frankfurt	5288	Zurich	3103	Bordeaux	1833	Seville	912
Tokyo	14345	Stockholm	5283	Florence	3006	Tel Aviv	1733	Gibraltar	879
Berlin	13957	Milan	4741	Barcelona	2996	Bombay	1700	Paz	853
Philadelphia	12194	Stuttgart	4733	Copenhagen	2920	Rio de Jan.	1683	Lisboa	820
Vancouver	11629	Jerusalem	4682	Panama	2740	Taipei	1567	Ankara	814
Montréal	9365	Beijing	4487	Luxembourg	2739	Toulouse	1529	Calcutta	785
Melbourne	9336	Munich	4456	Auckland	2731	Strasbourg	1451	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 5 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 5. Les pays et les zones sur *Internet*



Found 1315 matching words (number of documents): algeria (1688), angola (1415), benin (920), botswana (1220), burkina (872), faso (790), cameroon (1246), chad (3560), comoros (522), congo (1416), ivoire (693), djibouti (616), egypt (6055), equatorial (1775), guinea (3139), eritrea (727), ethiopia (1859), gabon (758), gambia (729), ghana (1707), bissau (519), ivory (3187), coast (23606), coastal (8703), kenya (2969), lesotho (729), liberia (836), libya (952), madagascar (1596), malawi (968), mali (1145), mauritania (701), morocco (2148), mozambique (1170)

Found 906 matching words (number of documents): namibia (1051), niger (961), nigeria (2338), rwanda (1869), burundi (1055), senegal (1231), seychelles (614), sierra (7400), leone (1258), somalia (1463), south (7723), southern (30838), southwest (10950), africa (24304), african (17195), sudan (1476), swaziland (713), tanzania (1636), togo (705), tunisia (1378), uganda (1396), sahara (1161), zaire (1646), zambia (1324), zimbabwe (2683), ...

- II-

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, en novembre, et à plus de 16 à la veille du présent colloque). Mais ce recul du temps est obtenu lorsque l'ordinateur s'adresse à la tradition littéraire. Nous prendrons pour exemple deux écrivains que séparent cinq siècles d'histoire et que réunit pourtant le même goût pour l'exploration du globe. Le premier, François Rabelais, est un voyageur de la Renaissance, observateur des moeurs et coutumes du temps et grand découvreur de pays imaginaires. Le second, Julien Gracq, est géographe et grand amateur de paysages. Aucun des deux n'a visité l'Afrique, sinon en rêve. Mais l'Afrique n'est pas absente de leur oeuvre comme en témoigne les deux concordances représentées ci-dessous.

Figure 6. Concordance de l'*Afrique* dans le CD-Rom Rabelais

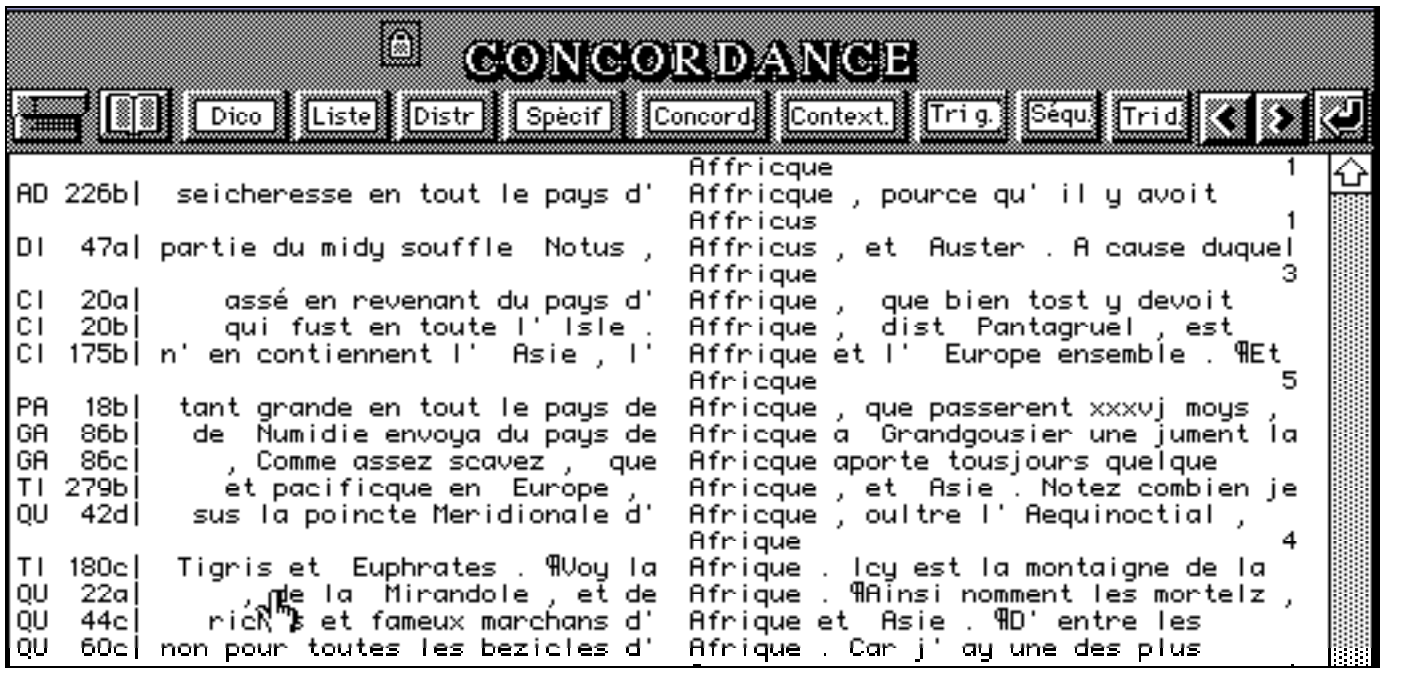
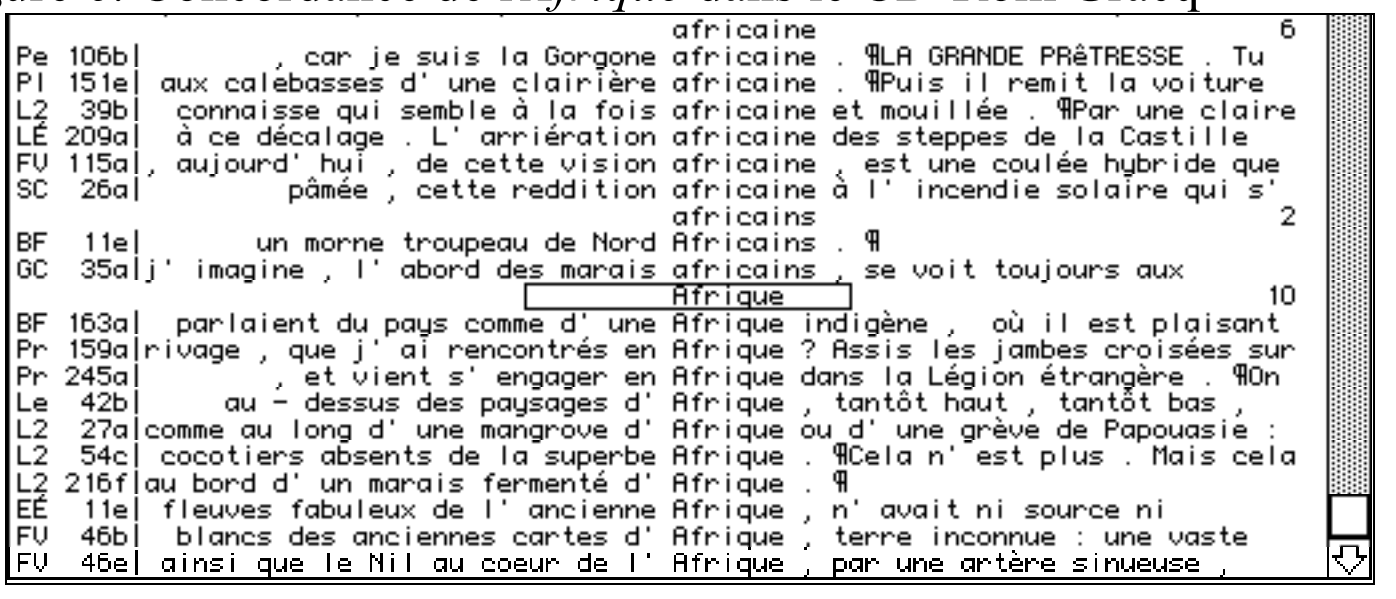


Figure 6. Concordance de l'*Afrique* dans le CD-Rom Gracq



On voit que l'orthographe n'est pas encore normalisée au temps de Rabelais, ni en ce qui concerne le redoublement de la consonne intérieure, ni en ce concerne la finale, qui hésite entre la forme archaïque *-icque* et la forme moderne *-ique*. Par

contre les maux climatiques dont souffre une grande partie de l'Afrique sont déjà présents au temps de Rabelais, comme en témoigne cette évocation d'une "seicheresse tant grande en tout le pays de l'Africque, que passerent XXXVI moys, troys sepmaines, quatre jours, treze heures, et quelque peu dadvantaige sans pluye, avec chaleur de soleil si vehemente que toute la terre en estoit aride". Voir figure 7.

Figure 7. Contexte du mot *Afrique* dans le corpus Rabelais (extrait).

Précisons que ces deux CD-Rom ont été réalisés grâce à notre logiciel Hyperbase. Il a fallu certes quelques aménagements pour adapter ce logiciel à des données spécifiques et à un support particulier dont la contenance est plus satisfaisante que la rapidité. Il serait oiseux d'entrer dans les détails techniques, d'autant qu'un manuel circonstancié de plus de cent pages accompagne ces deux produits. On renvoie le lecteur aux publications antérieures[6] qui décrivent les fonctionnalités d'Hyperbase et la marche à suivre pour les mettre en oeuvre. L'utilisateur a encore à sa disposition une aide en ligne plus concentrée, et, qui plus est, une page d'explication, immédiatement disponible, pour chaque fonction. Si le symbolisme d'un bouton n'est pas compris tout de suite et si son nom laisse perplexe, on peut par précaution faire apparaître sur l'écran les instructions précises avant de déclencher l'action correspondante. Les fonctions d'aide sont encore plus développées dans le CD-Rom Rabelais, puisqu'on y propose des séquences animées et sonores qui expliquent le mode d'emploi.

Qu'il s'agisse de la version courante qui s'adapte aux données de l'utilisateur, de la version CD-Rom qui livre tout ensemble les données et le programme d'interrogation, ou de la version Web dont on parlera plus loin, dans tous les cas le logiciel propose deux orientations, l'une documentaire, l'autre statistique.

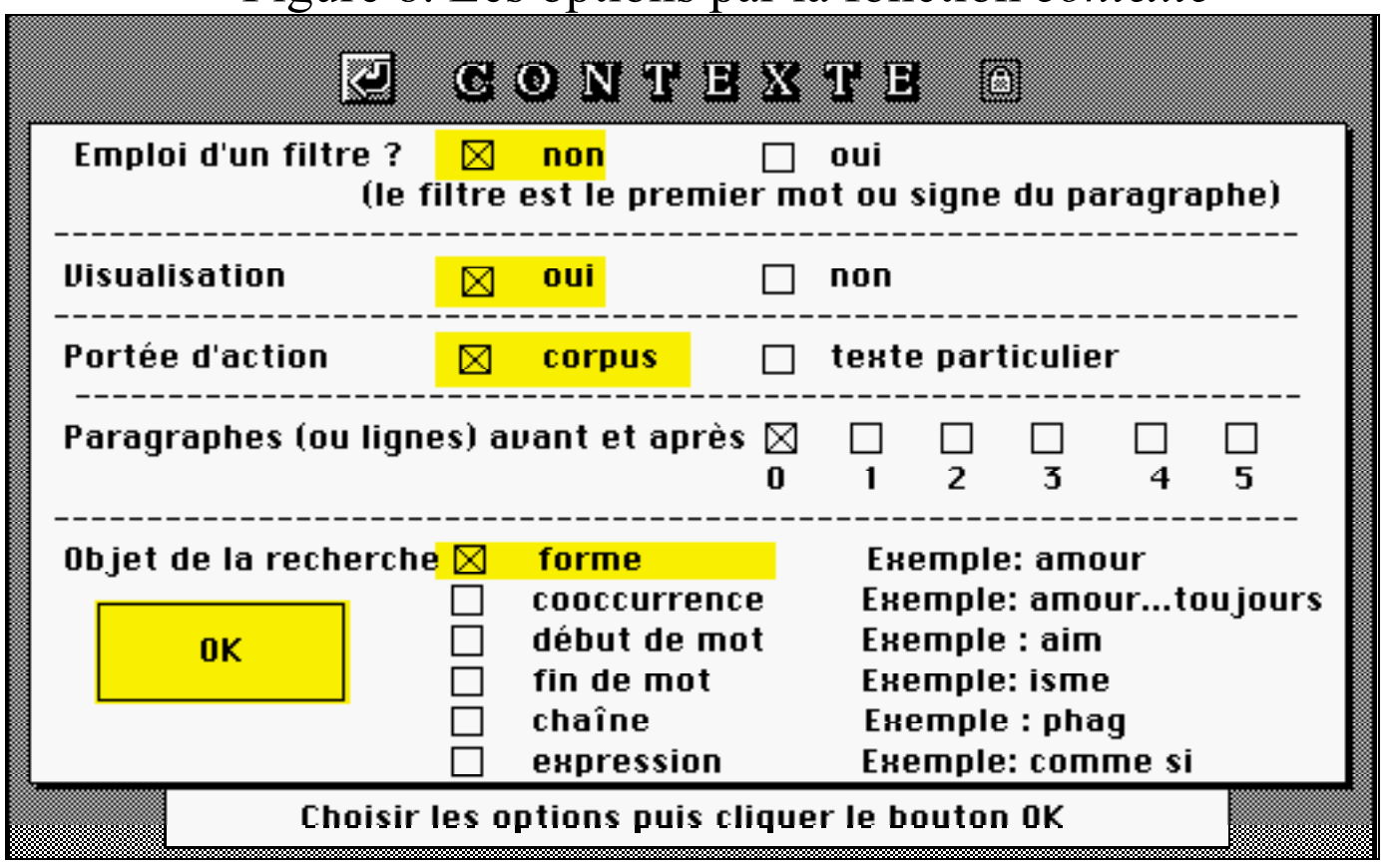
a - Le programme d'exploitation répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes: concordances de type *kwic* (avec tri des expansions droite ou gauche du contexte), index sélectifs ou systématiques, dictionnaires des fréquences, sélection de contextes larges, cooccurrences, filtrage et masquage des mots et constitution de listes, recherche des parties de mots (début, fin ou chaîne quelconque) ou des groupes de mots, limitation ou extension du corpus de travail, etc.. La fonction *Contexte* qui a produit le résultat de

---

[6] "Un hypertexte statistique: HYPERBASE", in *JADT 1993*, TELECOM, Paris, 1994, p. 1-16. "Hyperbase, synopsis.", in *Traitements informatisés de corpus textuels*, Didier-Erudition, 1994, p.169-184.

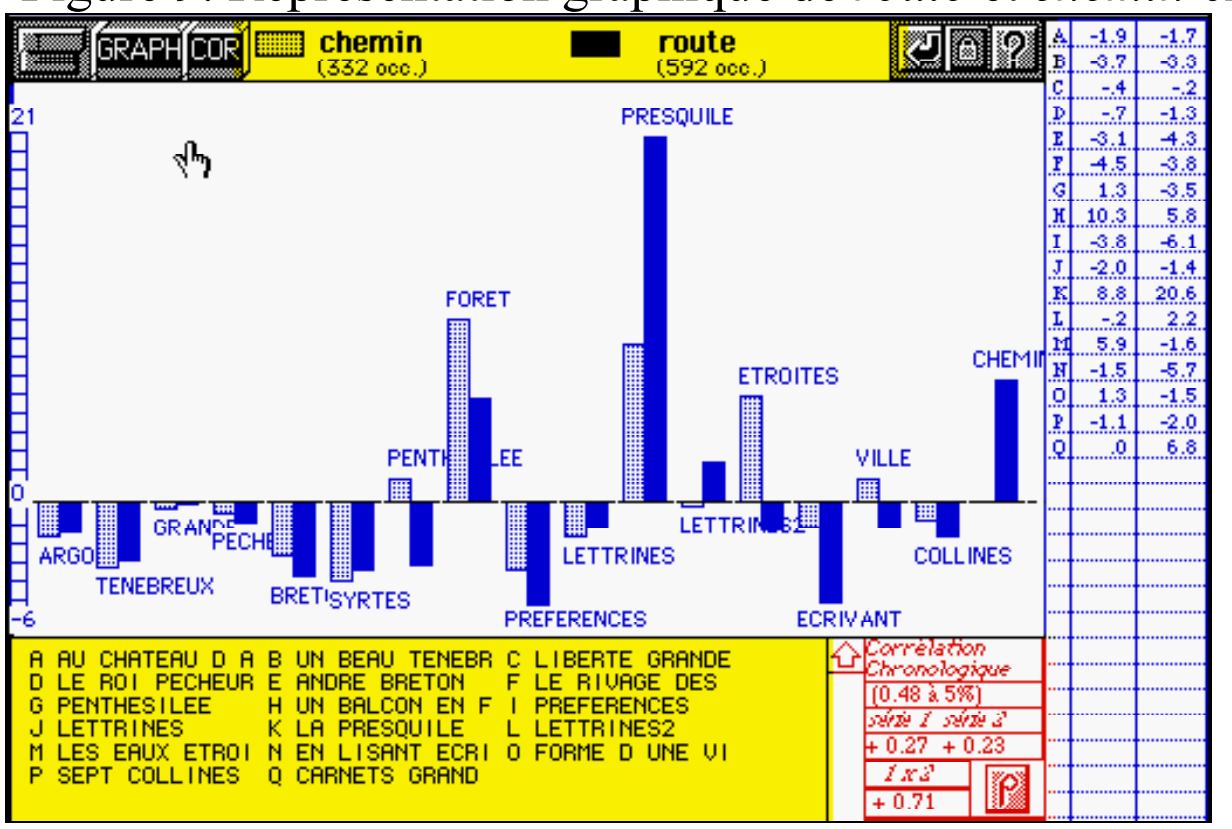
la figure 7 propose tout un choix de paramètres, dont rend compte le dialogue de la figure 8.

Figure 8. Les options par la fonction *contexte*



b - *Hyperbase* se distingue des hypertextes similaires par l'orientation statistique donnée au produit. D'une part, s'il s'agit d'un texte français, une comparaison est faite[7], sous forme d'écart réduit, avec le corpus du Trésor de la langue française (XIX-XXe, soit 70 millions de mots). D'autre part le corpus peut être partitionné pour permettre des comparaisons internes. *Hyperbase* restitue ainsi les mot-clés propres à chaque texte, de même qu'il dresse le profil caractéristique du corpus dans son ensemble, se détachant sur la toile de fond de l'usage littéraire de la langue depuis 1789. De même le profil d'un mot (ou de plusieurs que l'on superpose) est dessiné par *Hyperbase*, les sous-fréquences observées, judicieusement pondérées, se transformant à volonté en histogramme (figure 9).

Figure 9. Représentation graphique de *route* et *chemin* chez Gracq

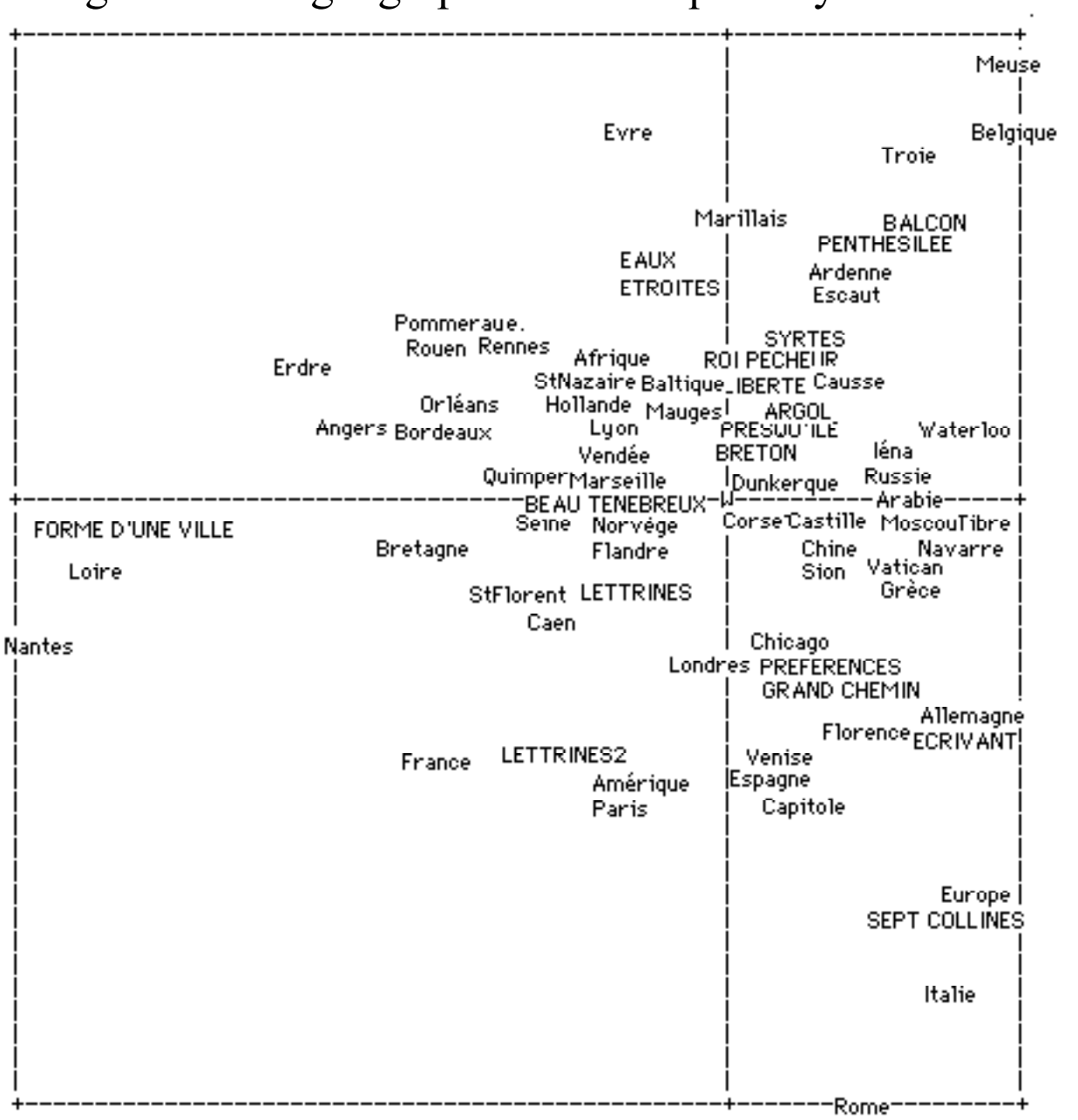


[7] Là où le calcul se justifie, c'est-à-dire quand la fréquence est suffisante dans le modèle (soit f=500 dans le TLF).

Des tableaux peuvent être constitués qui, à partir de critères, automatiques ou non, procèdent aux regroupements de mots ou de textes. Et ainsi peut-on pallier l'absence de lemmatisation. Par exemple *Hyperbase* permet de circonscrire une classe de mots, un champ thématique, voire même le système de la ponctuation.

Une fois constituées, ces listes - ce sont en réalité des tableaux à deux dimensions - peuvent être soumises aux méthodes multidimensionnelles (un programme d'analyse factorielle a été intégré à *Hyperbase*). Ainsi peut-on dresser une sorte de carte géographique de Gracq à travers les noms de lieu qui ont sa préférence. L'ouest est sa terre natale, évoquée dans les récits ou fragments autobiographiques comme *Le beau ténébreux*, *La forme d'une ville, Les eaux étroites, Lettrines* et *Lettrines 2.* Toute la moitié gauche est ainsi dévolue à la petite patrie, entre Nantes et Angers. L'est se divise en deux directions, l'une au nord, l'autre au sud. La première conduit au *Balcon en forêt*, sur les bords de la Meuse. La seconde s'oriente du côté de l'Italie, de Rome et des *Sept Collines.* L'Afrique n'est guère présente dans ce concert des nations, non plus que l'Asie et l'Amérique. Gracq est un marcheur à pied, qui emprunte parfois le train ou la voiture. Il ne va guère là où mène l'avion.

Figure 10. La géographie de Gracq. Analyse factorielle



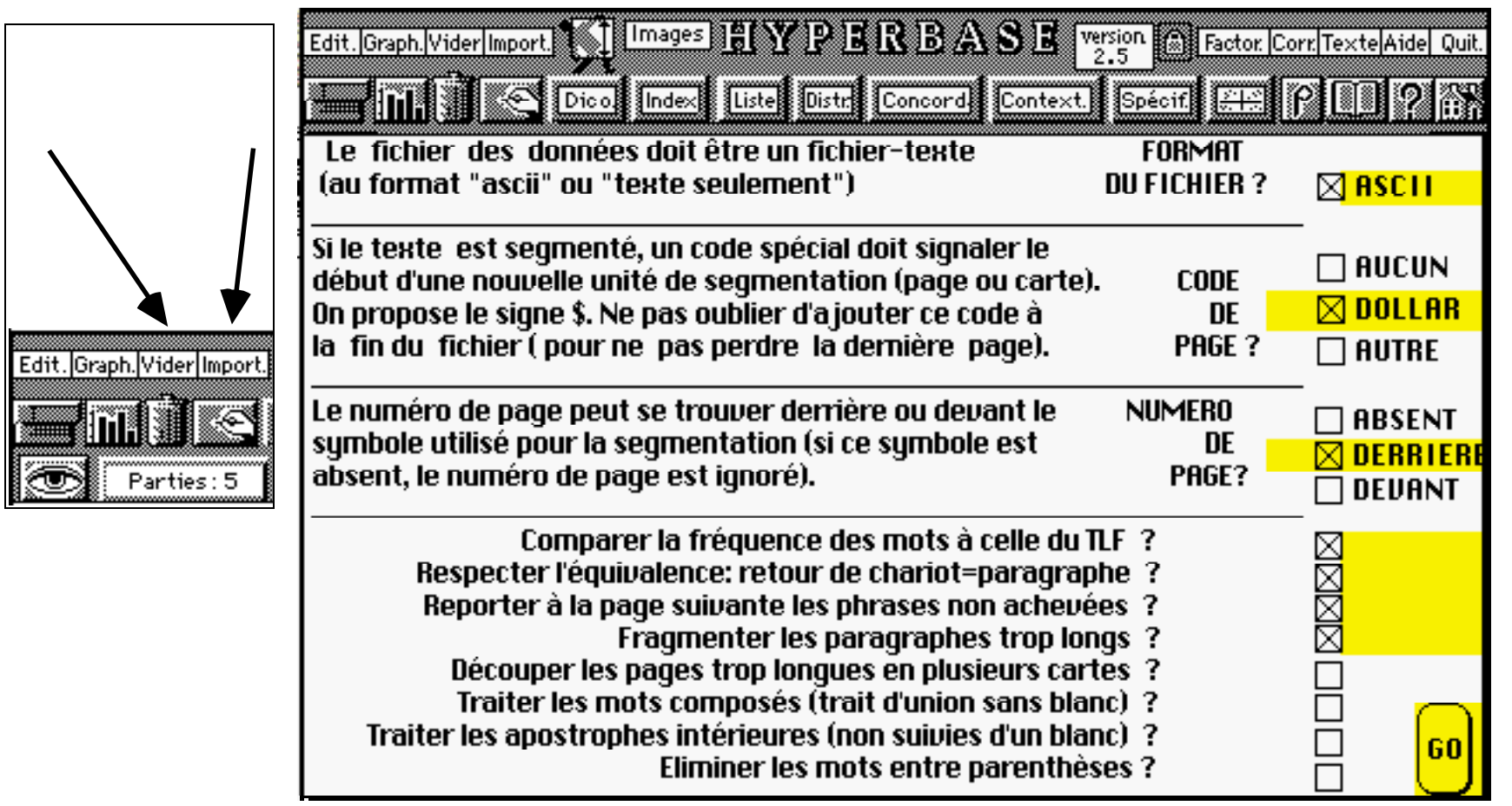
D'autres calculs lexicométriques sont assurés qui permettent d'apprécier la richesse relative du vocabulaire, la distribution des classes de fréquences, l'abondance, si l'on peut dire, des mots rares (ou hapax), l'accroissement et l'évolution du vocabulaire, etc.. En particulier une fonctionnalité nouvelle est

apparue dernièrement (version 2.5, Juillet 1995), qui mesure la distance que chaque texte établit avec tous les autres du même corpus, et qui est le rapport entre les vocables communs aux deux textes que l'on confronte et les vocables exclusifs que chacun des deux se réserve. Cela, qui peut s'appeler aussi la connexion lexicale, permet d'établir une typologie des textes à partir des similarités lexicales et principalement de leur composante sémantique[8].

c - Le traitement d'un nouveau texte

À la différence de beaucoup de logiciels où les fonctions sont absolument séparés des données, les unes et les autres sont mêlées dans une pile Hypercard, surtout lorsqu'il s'agit du type "standalone". La pile originale doit donc être recopiée dès que l'on veut traiter des données nouvelles. Sous son nouveau nom elle garde ses programmes - qu'il faut conserver - et ses anciennes données - qu'il faut évacuer. L'élimination de celles-ci se fait en sollicitant le bouton *Vider* (voir ci-dessous). Le résultat est une pile vierge, qui peut servir de modèle pour toutes les applications ultérieures. L'incorporation d'un texte est assurée par le bouton *Importer* (voir ci-dessous), qui montre la première page du fichier des données et s'enquiert des options souhaitables en affichant le dialogue de la figure 11.

Figure 11. Entrée des données



Les données textuelles doivent se trouver dans un fichier ASCII (ou "texte seulement"). On a pris en compte la plupart des alphabets européens. Aucun formatage particulier n'est obligatoire, le logiciel se chargeant de la pagination et de

[8] La fréquence ne joue ici aucun rôle. Les effectifs sont constitués sur le seul critère présence/absence.

la partition, si elles sont absentes du fichier. En ce cas les cartes (ou pages) ont environ 200 mots et l'ensemble du texte est découpé en dix parties de longueur voisine.

Mais il vaut mieux suivre le découpage naturel des données, s'il existe. Deux conventions doivent alors être respectées:

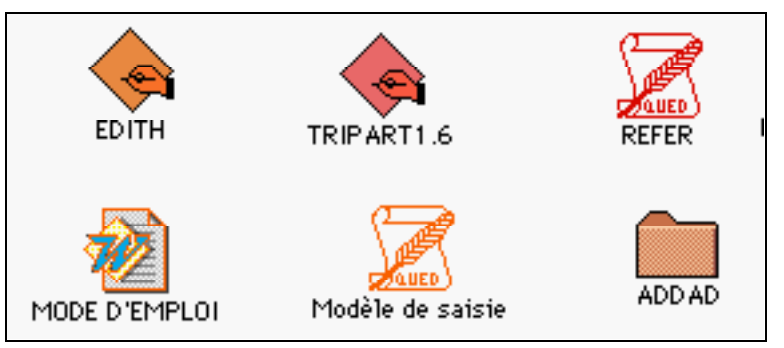
- les parties doivent être précédées d'une ligne où l'on indiquera le titre (en 20 caractères maximum, sans virgules ni apostrophes) en utilisant devant et derrière le symbole composite &&& (sans blanc). Veiller à bien choisir le dernier mot du titre qui sert d'abréviation lorsque la place manque, par exemple dans les graphiques, et qui doit être unique et distinctif.

- les pages sont indiquées en ajoutant une ligne (au début) et en y portant le numéro, immédiatement précédé ou suivi d'un code spécial (par exemple le symbole $; mais on peut choisir un autre code, si le symbole $ apparaît dans le texte même). Exemple:

```
&&&La vie en rose&&&
$1
texte de la page 1
$2
texte de la page 2, etc.
&&&Le travail au noir&&&
$62
texte de la page 62
$63
texte de la page 63, etc..
```

Le traitement d'un texte nouveau s'opère en trois phases: la première libère l'espace requis et transfère le texte dans la base, à raison d'une carte par page. C'est l'occasion de transcoder le texte afin d'uniformiser la présentation et en particulier de standardiser la ponctuation. En même temps est constitué un fichier formaté qui va servir d'entrée à la phase 2. Celle-ci suit la première étape de façon automatique ou manuelle. On choisira ce dernier mode, afin de libérer le maximum de mémoire, si le fichier à traiter est de grande taille et si la machine est de faible puissance. Dans ce cas la pile est abandonnée à l'issue de la phase 1 et un double clic sur l'icône *Tripart1.6* (ci-dessous) mettra en oeuvre le programme d'indexation, en lui réservant toutes les ressources de l'ordinateur.

Figure 12. Le programme d'indexation



Quand le tri est achevé et que les traitements subsidiaires ont pris fin, il suffit de revenir dans la pile par un double clic pour que les multiples résultats obtenus dans la phase 2 soient communiqués à la pile et définitivement enregistrés dans la phase 3. Les fichiers intermédiaires peuvent alors être détruits, la pile disposant de toutes les informations dont elle a besoin, et particulièrement du dictionnaire inverse.

Prévoir 15 minutes pour le dépouillement d'un texte de 500 pages (avec un Mac équipé d'un microprocesseur *8030*) et 20 minutes pour le tri. Après ce temps de préparation (qui comporte un transcodage, un découpage en cartes, un tri des formes, un dictionnaire des fréquences et divers tests statistiques), la pile est exploitable. Si l'on dispose d'un microprocesseur *8040* ou *PowerPC*, le temps de préparation est fortement raccourci.

d - Spécifications techniques

On ne s'appesantira pas sur les spécifications techniques. Il suffit de préciser qu'*Hyperbase* comprend un programme de préparation (écrit en *Pasca*l), un éditeur de texte (écrit en langage *C*), un programme d'analyse factorielle (écrit en *Fortran* et emprunté à *ADDAD*) et un programme d'exploitation (écrit en *Hypertalk* et complété par de nombreuses commandes externes). La configuration requise est peu exigeante et se contente d'une mémoire vive de 2000 Ko pour son usage propre (il faut ajouter la mémoire requise par le système et celle que réclament épisodiquement les applications externes auxquelles *Hyperbase* fait appel, traitement de texte ou analyse factorielle[9]). Le disque dur est indispensable et l'écran couleur recommandé. *Hyperbase* fonctionne indifféremment sur système 6 ou 7 et sur toute la gamme *Apple*, du *Mac Plus* au *PowerMac*. Précisons que la dernière version, totalement refondue (2.5), s'est affranchie de l'environnement Hypercard. Étant du type "standalone", l'application est devenue parfaitement autonome et ne dépend plus de l'installation de l'utilisateur. Mais elle échappe aussi à toute intervention intempestive dans le code même des programmes. Les scripts sont hors d'atteinte et ni la fenêtre de commande, ni la barre de menus ne livrent accès aux intrus.

Au bout de trois ans de commercialisation, les clients sont en majorité des chercheurs du secteur public, littéraires, linguistes, historiens et sociologues mais aussi des entreprises privées, spécialisées dans la veille technologique ou les systèmes experts. À l'étranger le logiciel est plus connu au Japon, aux États-Unis, au Canada et au Brésil. Le succès est venu là où on ne l'attendait pas: dans les instituts de sociologie ou de sondage, chaque fois que des questions ouvertes ou des développements textuels doivent être traités. *Infométrie*, l'agence *Harris* et la *Sofres* se servent d'*Hyperbase,* comme le prouve l'étude publiée par la *Sofres*, à la veille du scrutin présidentiel de 1995. Dans le domaine lexicologique qui est commun aux participants de ces journées, son application peut donner de bons résultats, pour autant qu'on dispose de données textuelles en continu. Mais pour traiter le lexique et

---

[9] Le seul moment où l'on a avantage à disposer d'une mémoire abondante et d'une machine rapide est celui de l'indexation, qui transforme un texte Ascii en base de données. Ce traitement exige précautions et patience mais, comme il n'a lieu qu'une fois pour chaque corpus, l'effort consenti se justifie sans peine.

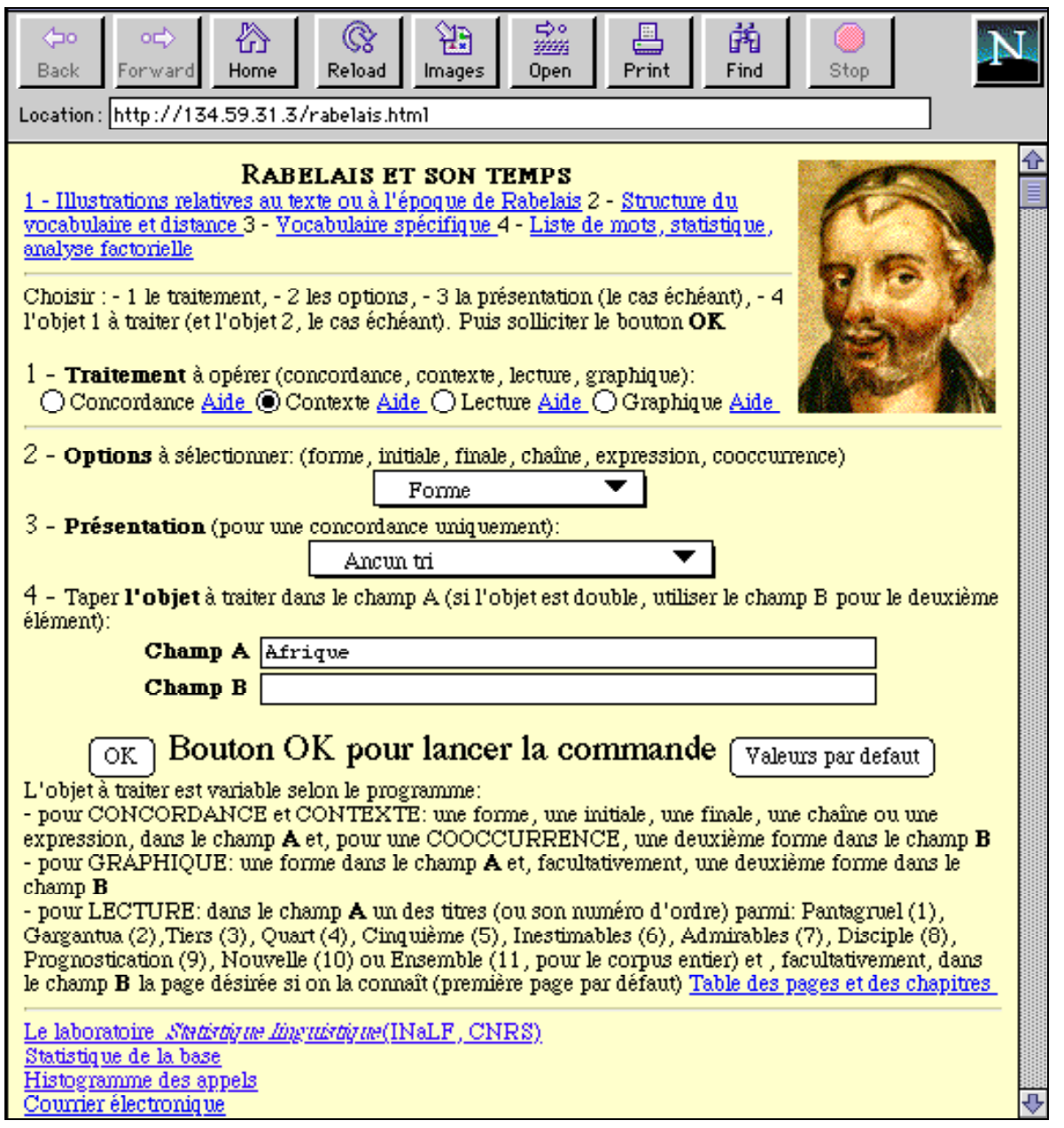
plus généralement les faits de langage, il est d'autres outils qui font appel aux méthodes des bases de données structurées ou à celles de l'intelligence artificielle et que nous n'aborderons pas ici.

- III -

Enfin le réseau *Internet* peut être aussi le lieu où s'accomplit - à distance et en temps réel - la recherche linguistique. On en montrera une illustration avec une base de données textuelles qu'on a proposée à la communauté scientifique sur le *Web* et qui reprend les données et les fonctions mises en oeuvre dans le CD-Rom Rabelais dont nous venons de parler. L'alternative qui oppose le CD-Rom à *Internet* est en effet moins exclusive que complémentaire - et les éditeurs ont vite compris l'intérêt des productions multimodales, dérivées les unes des autres. Chacun des deux supports à 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 13. Il contient peu d'éléments véritablement statistiques, étant orienté d'abord vers les requêtes documentaires, et en particulier vers la recherche des contextes.

Figure 13. L'écran d'accueil de la base Rabelais, disponible sur le *Web*



Le même exemple africain déjà mis en lumière à propos du CD-Rom peut nous servir à illustrer la fonction *Contexte*, qui livre le même résultat, avec la même rapidité. Il suffit de quelques secondes pour quérir les contextes où le mot est relevé dans le corpus, à quoi il faut ajouter le temps de la conversion en format *html* (format unique du réseau *Web*) et le temps de la transmission, lequel varie selon l'installation dont on dispose. On a prévu que l'interrogation pourrait se faire à partir d'un simple modem. Comme l'accès est alors moins rapide qu'un branchement direct[10], on a cru prudent de limiter le volume des résultats restitués à 100 000 caractères et à 1000 lignes dans le cas d'une concordance.

Figure 14. Le contexte de l'*Afrique* dans la base Rabelais (sur Internet)

Retour à la base Rabelais
FORME : " afrique" . CONTEXTES : 4

(1477 caractères)
Voy la Asie . Icy sont Tigris et Euphrates .
Voy la Afrique . Icy est la montaigne de la Lune . Voidz tu les Paluz du Nil ? Deça est Europe . Voydz tu Theleme ?
TIERS Page: 180c (Afrique) Chap. 28

Nous avons clos le passaige entre les Tartres et les Moscovites . Nous avons respondu a la requeste du Cheriph . Aussi avons nous a la devotion de Guolgotz Rays . L' estat de Parme est expedié : aussi est celluy de Maydenbourg , de la Mirandole , et de Afrique .
QUART Page: 22a (Afrique) Prologue

Adoncques descendit on havre , contemplant , ce pendent que les chormes des naufz faisoient aiguade , divers tableaulx , diverses tapisseries , divers animaulx , poissons , oizeaulx , et aultres marchandises exotiques et peregrines , qui estoient en l' allee du mole , et par les halles du port . Car c' estoit le tiers jour des grandes et solennes foires du lieu : es quelles annuellement convenoient tous les plus riches et fameux marchans d' Afrique et Asie .
QUART Page: 44c (Afrique) Chap. 2

Ouy vrayement , respondit le marchant , je le suys : et ne vouldrois ne l' estre pour toutes les lunettes d' Europe : non pour toutes les bezicles d' Afrique . Car j' ay une des plus belles , plus advenentes , plus honestes , plus prudes femmes en mariage , qui soit en tout le pays de Xantonge : et n' en desplaise aux aultres .
QUART Page: 60c (Afrique) Chap. 5

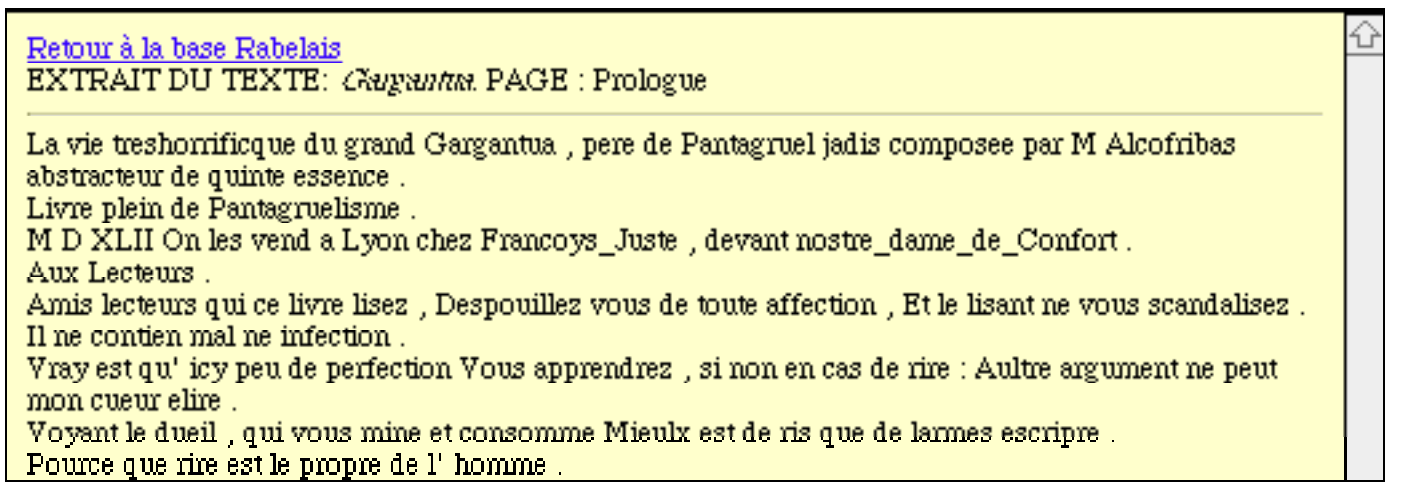
Paramètres:
Contexte Forme *Afrique*
Temps d'éxécution : 1 seconde(s)
Nombre d'appels de cette base: 1150

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 (figure 15).

---

[10] Les modems les plus rapides ne dépassent pas 28800 bauds, soit 3000 caractères/seconde, vitesse théorique qu'on atteint rarement, même avec protocole de compression et de correction. Le branchement direct sur une ligne spécialisée autorise des transferts beaucoup plus rapides. Mais ce luxe est réservé aux bâtiments "intelligents", qui sont relativement rares en France et plus encore en Afrique. Et cette vitesse potentielle - comme celle des voitures de course - ne peut rien contre les embouteillages des autoroutes de l'information.

Figure 15. Lecture et contrôle du texte (ici page 1 de *Gargantua)*.



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 16.

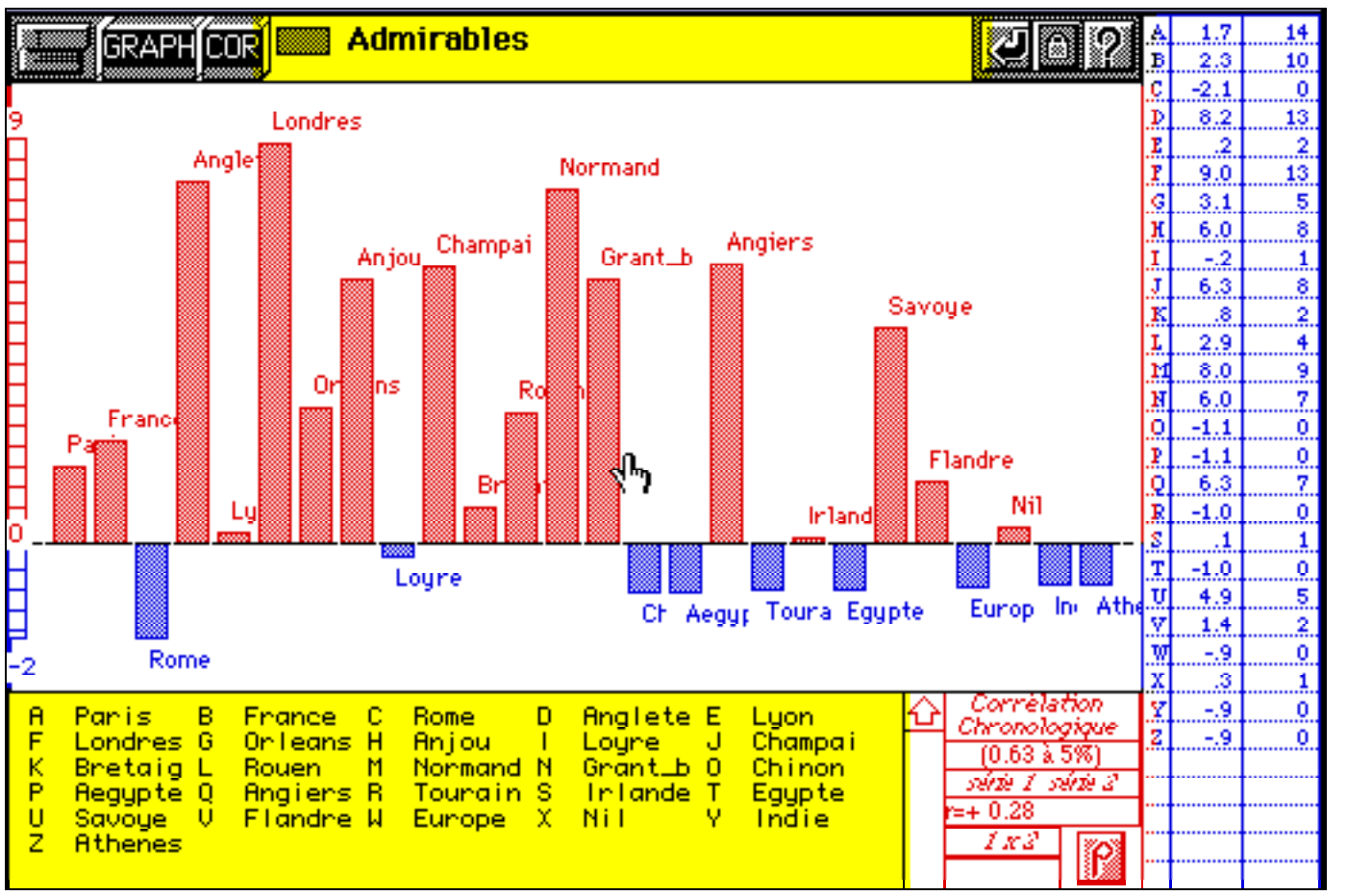
Figure 16. Le menu statistique de la base Rabelais

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, par exemple, l'histogramme de tous les mots bâtis, comme *Afrique,* sur le suffixe -*ique* ou -*icque*.

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, sa patrie d'adoption étant 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 Bretaigne ou la Flandre. Le centre de gravité s'y déplace vers le Nord et l'Ouest.

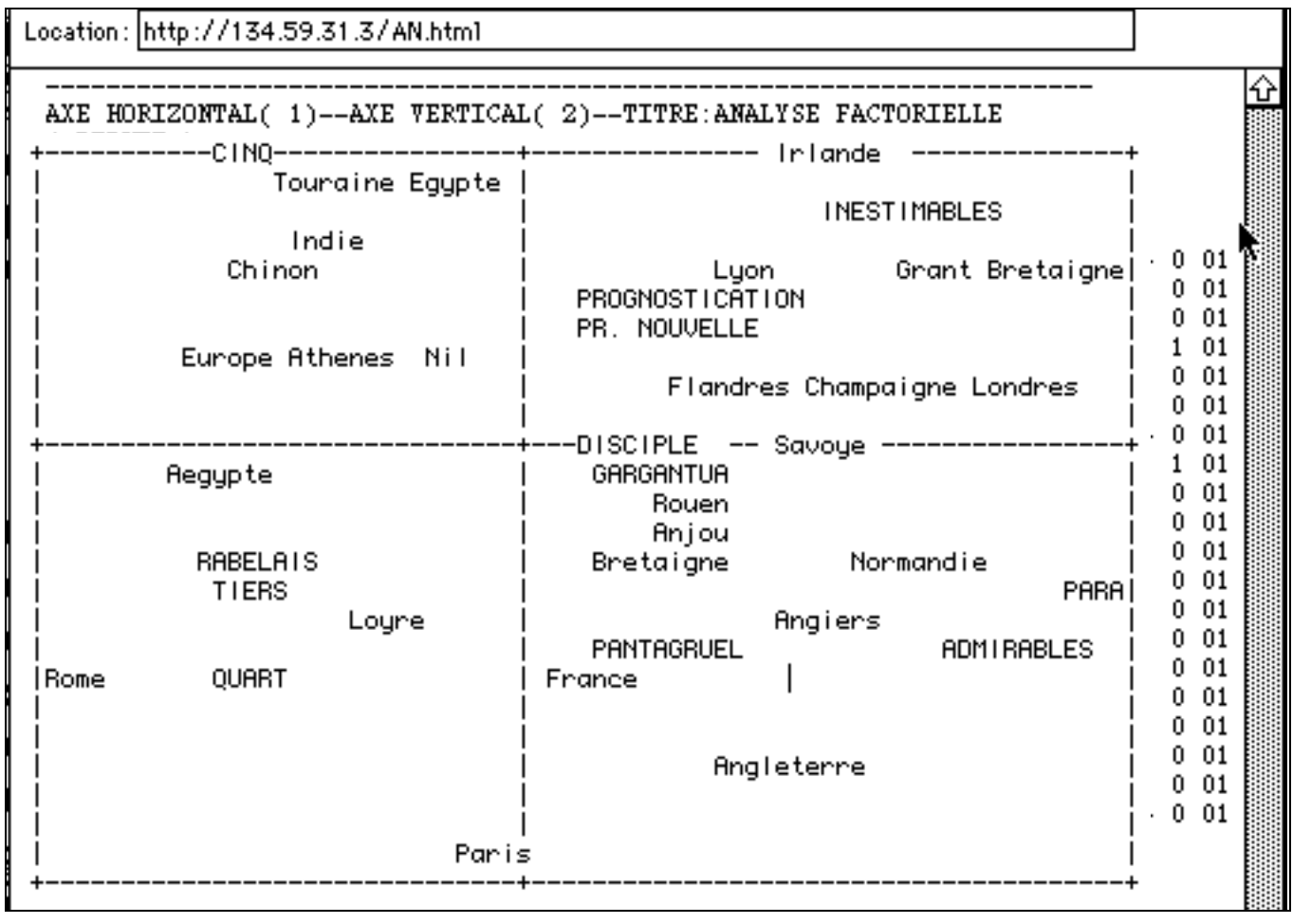
Figure 17. La représentation d'une colonne (ici les *Chronicques Admirables*)



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. On 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 et les rivages opposés d'Afrique et d'Asie, 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, Egypte, Nil, Indie*). En ce temps de découverte la force d'attraction de l'Orient et du Sud est irrésistible pour les esprits ouverts à l'aventure. 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 pararabelaisiens 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 pararabelaisiens 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.

- IV -

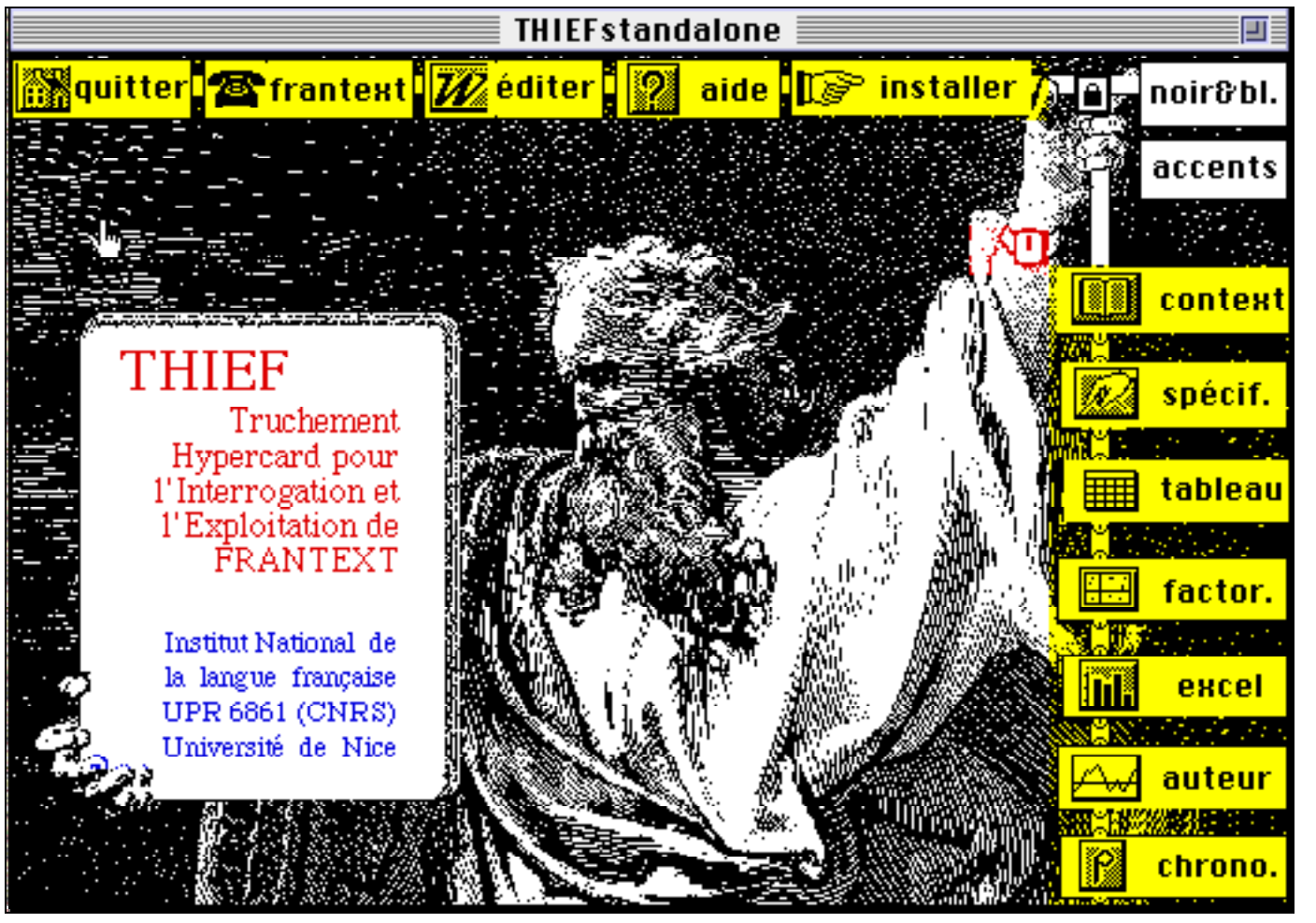
C'est pourquoi il importe de varier les données autant que les points de vue. On peut regretter la rareté des textes disponibles sur *Internet*. Mais il y a une notable exception que nous nous proposons d'aborder pour finir. C'est celle de *Frantext*. Cette base de données textuelles, constituée il y 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*.

Frantext est très riche en fonctions documentaires sans être démuni en fonctions statistiques. Pour un corpus choisi il peut fournir à volonté les passages du ou des mots qui intéressent l'utilisateur. Mais il peut livrer aussi 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, pour pour 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 *Tranpac*), le dialogue n'est pas encore celui du *Web*. Mais une version *Web* est déjà fonctionnelle et devrait être proposée aux chercheurs prochainement.

Comme l'unanimité ne règne par 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, seul 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

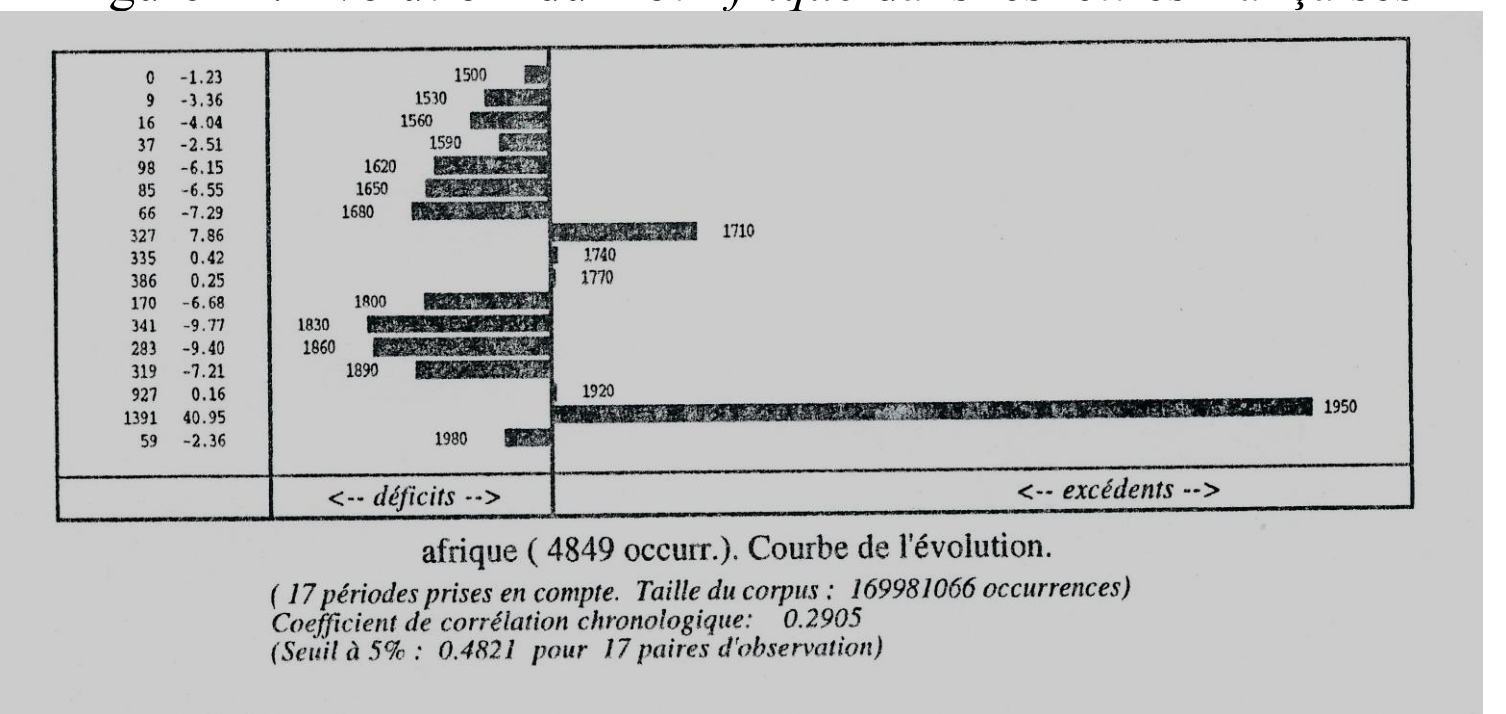
vont suivre et qui n'épuisent pas la variété des actions possibles, comme en témoigne le tableau de bord du logiciel:

Figure 20. *THIEF*
(Truchement Hypercard pour l'Interrogation et l'Exploitation de *Frantext*)



En reprenant le nom du mot *Afrique* 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 près de 5000 occurrences qu'on observe en cinq siècles de littérature française. Mais cette présence de l'Afrique dans les lettres françaises n'est pas constante. Elle est faible du 16e au 17e siècle et ne s'affirme qu'au siècle des Lumières qui est aussi le siècle de l'esclavage. Le terme générique Afrique n'est guère en faveur après la Révolution, l'impérialisme colonial préférant circonscrire des territoires particuliers. Mais tout change au 20e siècle dans la conscience française. L'Afrique y est au centre de l'actualité, surtout au moment de la seconde guerre mondiale et de la décolonisation qui a suivi. Voir graphique 21.

Figure 21. Évolution du mot *Afrique* dans les lettres françaises

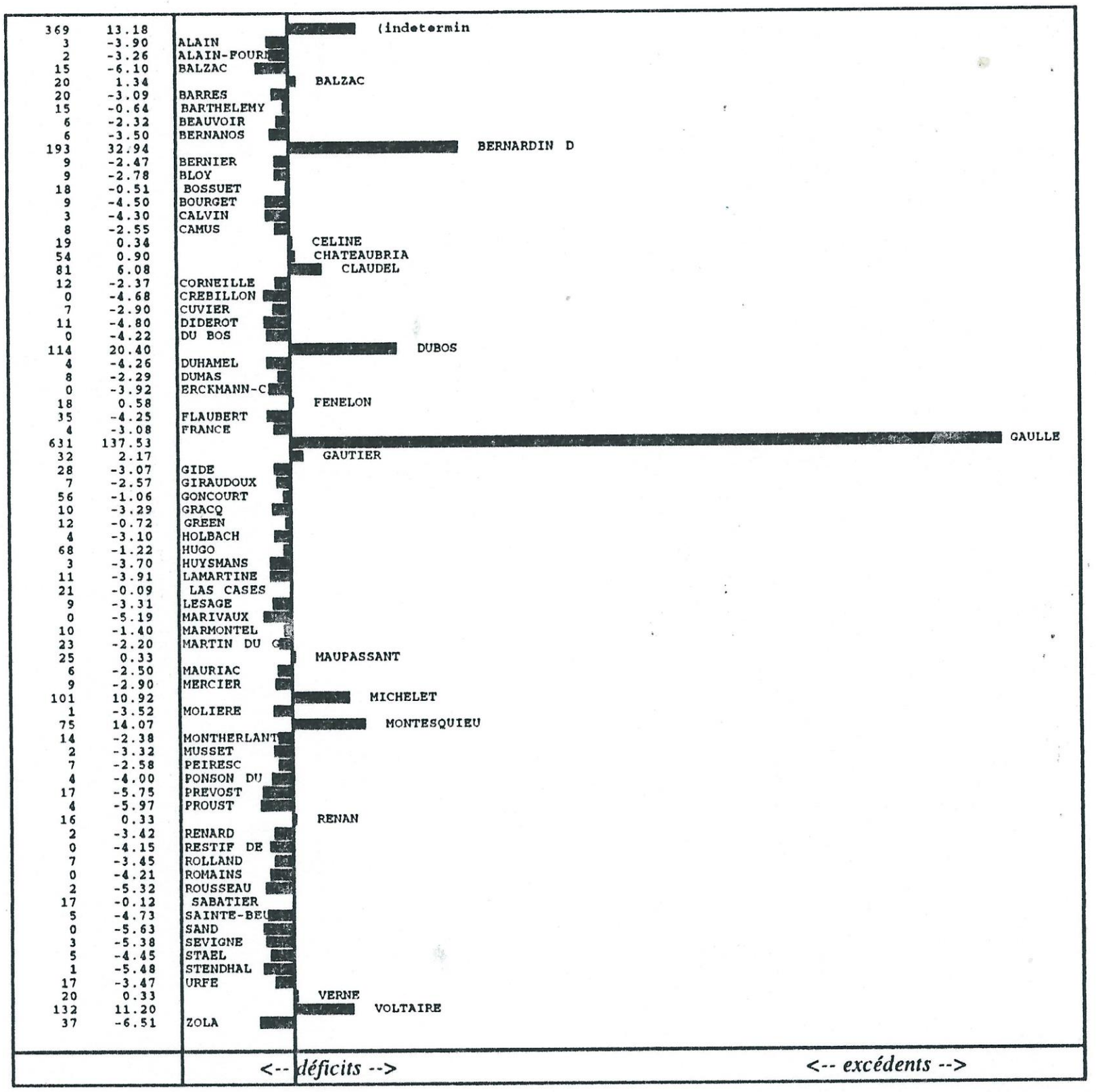


afrique (4849 occurr.). Courbe de l'évolution.

(17 périodes prises en compte. Taille du corpus : 169981066 occurrences)
Coefficient de corrélation chronologique: 0.2905
(Seuil à 5% : 0.4821 pour 17 paires d'observation)

Cette distribution se retrouve lorsqu'on isole les auteurs. Un millier d'écrivains ont été passés en revue dont 75 prennent place dans le graphique (ce sont ceux qui sont le mieux représentés dans le corpus, avec plus de 500000 mots). Ceux qui citent l'Afrique le plus souvent appartiennent au 18e et au 20e siècles et prennent place dans la moitié droite de l'histogramme 22.

Figure 22. Répartition du mot *Afrique* chez les écrivains française



afrique (4849 occurr.). Répartition selon les écrivains.

(75 écrivains représentés sur 989 pris en compte. Taille du corpus : 169981293 occurrences)
(limite d'étendue: >= 500000 occurrences)

Ils ne sont pas nombreux de ce côté. On y trouve des bourlingueurs comme Céline, des explorateurs comme Jules Verne, des diplomates, des marins ou des aviateurs tentés par l'aventure ou l'exotisme, comme Bernardin de Saint-Pierre, Loti, Chateaubriand, Claudel, Saint-Exupéry, des ethnologues, historiens ou essayistes comme Mably, Montesquieu, Voltaire, Michelet ou Gurvitch, ou tout simplement des voyageurs comme Maupassant. On y trouve surtout un général que le hasard de la guerre a conduit sur ce continent. Avec 631 mentions, les *Mémoires* de de Gaulle rendent hommage à l'Afrique plus qu'aucun autre texte de notre littérature.

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.

Parmi les mots que l'*Afrique* attire - et dont on trouvera la liste dans le tableau 23 - il y a les adjectifs qui s'accolent au pôle pour en définir une partie, de type géographique, comme les deux premiers de la liste: *équatoriale* et *occidental (*on trouve aussi *australe, orientale, orientales, occidentales)* ou de type administratif ou colonial (*française, françaises, français, portugais, arabe, romaine, anglais, britanniques, égyptiens, indiens, chinois, mondiale).* Quelques notations de couleur locale apparaissent ici et là, pour décrire, ou seulement nommer, les habitants (*nègres, nègre, indigènes, populations, habitants, population, sauvages, tribu*), le paysage (*plateaux, campagnes, plaines, sable*), la faune (*lions*), l'activité (*mines, ivoire*) ou le climat (*torride, climat, climats, brûlant, midi, glaces).* On notera surtout l'intérêt porté à la frange maritime, comme si le contact ne s'était établi que sur les côtes *(côtes, côte, cap, golfe, isles, océan, îles, navigation, ports, navires, île, pointe, isle).* On devine que la violence n'est pas absente et que le contact est militaire (*aviation, armée, flotte, armistice, amiral, offensive, militaire, armées, soldat, bloc, ennemi*). Mais les sujets tabous comme l'esclavage n'apparaissent guère. En somme c'est l'Afrique coloniale qui se révèle ici.

Mais élargissons le champ en considérant dans *Frantext* non plus la seule *Afrique* 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 (l'*Afrique* occupant l'une d'entre elles) et de 31 colonnes (on a choisi 31 écrivains du XVIIIe au XXe 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.

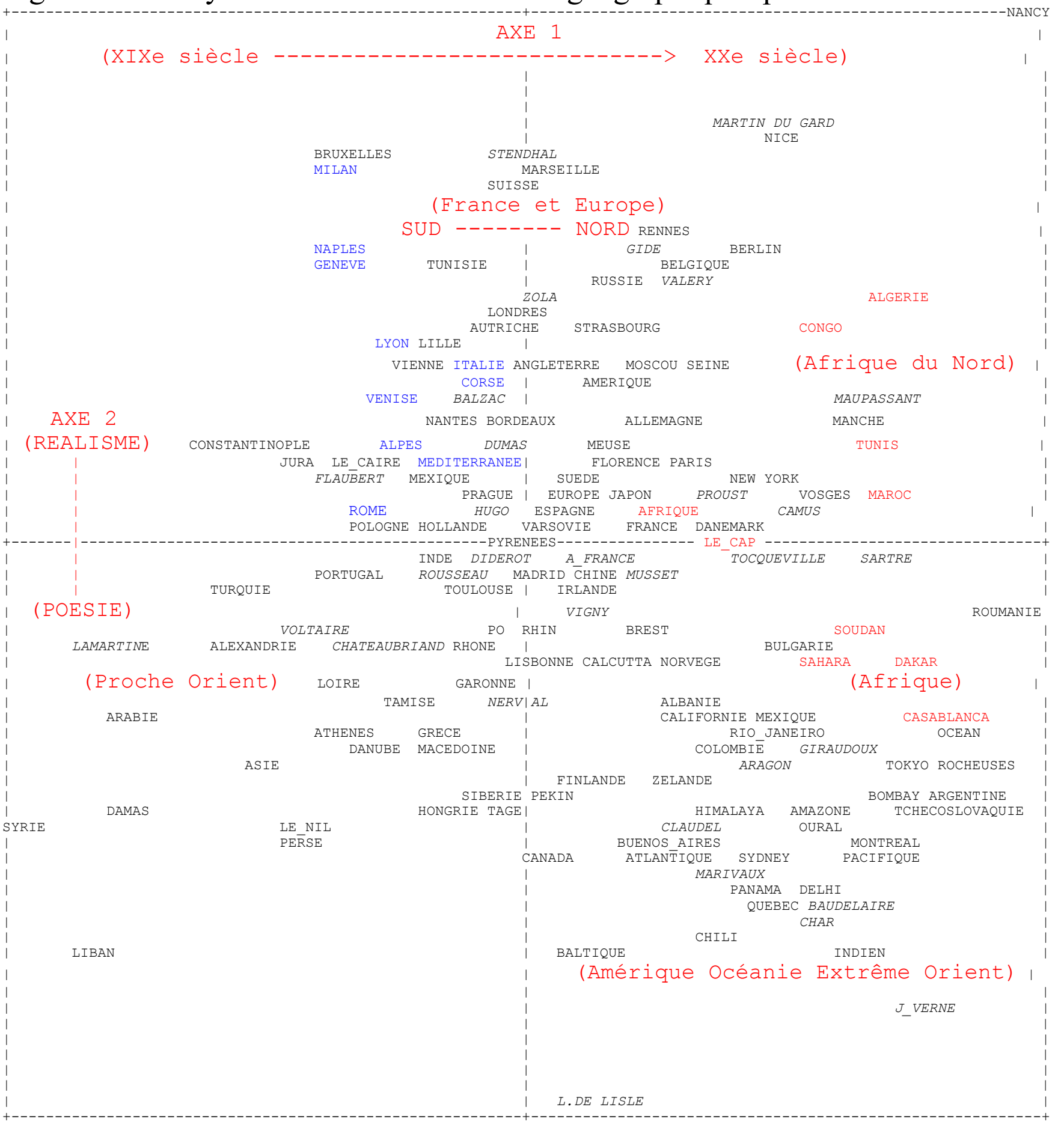
Tableau 23. L'environnement du mot *Afrique*

fréquence > 10 écart réduit > 3

Corpus de référence : Corpus total

64	142	équatoriale	156.2
208	1485	occidentale	156.1
578	12143	nord	149.2
93	1108	nègres	80.3
29	121	australe	76.5
152	3222	côtes	76.1
87	1239	orientale	70.9
221	11025	française	57.8
149	6817	côte	49.8
61	2253	cap	35.8
61	2838	françaises	31.6
33	1039	portugais	28.7
22	600	débarquement	25.4
27	974	équateur	24.2
28	1085	golfe	23.7
21	631	orientales	23.5
18	479	torride	23.2
59	4580	régions	23.1
27	1139	indigènes	22.2
20	654	isles	21.9
63	5879	orient	21.3
27	1299	lions	20.6
50	4125	océan	20.5
134	23201	français	20.4
36	2476	arabes	19.4
16	613	aviation	18.0
37	2968	zone	17.9
25	1457	populations	17.8
12	364	occidentales	17.7
38	3297	climat	17.3
28	1938	britannique	17.0
36	3387	îles	16.0
102	21352	armée	15.3
13	627	hémisphère	14.3
21	1545	tribus	14.2
17	1048	britanniques	14.2
23	1836	arabe	14.2
34	3927	militaires	13.7
29	2994	occident	13.6
19	1395	pôle	13.6
83	17820	libre	13.5
20	1660	navigation	12.9
37	5058	provinces	12.7
17	1295	royaumes	12.5
50	8627	romains	12.5
20	1794	indiens	12.3
36	5063	population	12.3
34	4702	sauvages	12.1
32	4350	habitants	11.9
12	769	anglo	11.7
17	1481	libération	11.6
24	2744	flotte	11.6
26	3165	voyages	11.5
11	679	armistice	11.4
11	726	missions	11.0
14	1206	mondiale	10.6
11	782	missionnaires	10.5
18	1910	ports	10.5
20	2312	navires	10.5
42	8285	île	10.3
18	2087	monstres	9.9
11	912	plateaux	9.6
31	5604	unis	9.5
22	3259	chinois	9.3
35	7261	province	9.1
53	13964	intérieur	9.0
13	1360	amiral	9.0
14	1567	autorités	9.0
12	1204	nègre	8.9
13	1417	ivoire	8.8
29	5573	région	8.8
32	6662	opérations	8.6
27	5124	libres	8.5
18	2717	lion	8.3
20	3332	romaine	8.1
13	1617	offensive	8.0
44	11971	anglais	8.0
11	1248	tribu	7.9
11	1274	répandue	7.8
31	7188	militaire	7.8
19	3265	voyageurs	7.7
13	1775	brûlant	7.6
16	2554	mines	7.5
11	1345	égyptiens	7.5
46	13855	midi	7.4
15	2401	plaines	7.2
22	4544	transport	7.2
26	6071	armées	7.1
13	1956	glaces	7.1
13	1976	climats	7.0
25	5775	bords	7.0
19	3780	habitans	6.9
24	5661	sable	6.7
37	11023	soldats	6.7
25	6122	pointe	6.6
19	3977	globe	6.6
12	1945	bloc	6.4
15	2828	isle	6.4
20	4524	national	6.4
20	4536	notamment	6.3
29	8238	race	6.2
17	3647	capitale	6.1
45	15905	voyage	6.1
13	2394	analogues	6.1
23	6005	déja	6.0
16	3419	campagnes	6.0
35	11536	ennemi	5.8
26	7479	nommé	5.8
16	3548	totale	5.8
12	2315	bassin	5.6
17	4040	oo	5.6
73	33123	soleil	5.4
12	2418	productions	5.4
17	4225	types	5.4
26	8080	millions	5.3
13	2870	économiques	5.2
49	20057	année	5.2
29	9745	nations	5.2
11	2259	tonnes	5.1
17	4520	forêts	5.1
19	5371	capitaine	5.0
16	4151	mission	5.0
11	2325	peste	5.0
38	14693	animaux	5.0
26	8640	jusques	5.0
44	18047	chef	4.9
12	2742	voyageur	4.9
12	2743	lointain	4.9
20	6056	soldat	4.8
19	5677	sociétés	4.8
22	7035	défense	4.8
27	9539	royaume	4.7
37	14857	armes	4.7
13	3274	prévoir	4.7
16	4504	romain	4.6
18	5385	découverte	4.6
30	11415	sol	4.5
27	9890	peau	4.5
28	10455	points	4.5
48	21826	avoient	4.4
12	3137	sépare	4.3
34	14100	anciens	4.3
20	6783	culture	4.3
18	5899	opération	4.2
28	10996	patrie	4.2
15	4599	principalement	4.1
16	5076	proche	4.1
33	14178	autorité	4.0
29	11955	importance	4.0
29	11957	ligne	4.0
49	23769	siècle	4.0
23	8809	lignes	3.9
29	12095	ancien	3.9
27	11076	éléments	3.9
26	10559	république	3.9
13	4009	sources	3.8
15	4973	fameux	3.8
16	5562	plantes	3.7
23	9265	accord	3.7
21	8448	production	3.5
31	14441	autrefois	3.4
12	3988	sécurité	3.4
13	4607	décidé	3.3
14	5132	importante	3.2
32	15667	situation	3.2
27	12650	nouveaux	3.1
58	32852	surtout	3.1
11	3812	plans	3.1
12	4375	histoires	3.0
20	8778	ancienne	3.0

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, Grèce, Macédoine*, sont ceux qui mènent aux lieux saints et que connurent les Croisés. Or les croisés des temps modernes se situent au début du XIXe 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. 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 XIXe siècle, et l'on y voit réunis Balzac, Stendhal, Flaubert et Zola. Le quadrant supérieur droit appartient plutôt aux prosateurs du XXe siècle: 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, Méditerranée, Corse, Alpes, Lyon, Genève, Vienne*), tandis que le nord tend à s'installer à droite: *Angleterre, Allemagne, Belgique, Suède, 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 trouve là l'Afrique, qui reste à cheval sur l'axe des x, avec, d'un côté, l'Afrique du Nord (*Algérie, Tunis, Casablanca, Maroc*), et, de l'autre, l'Afrique noire: *Dakar, Congo, Soudan, Sahara, Le Cap*. Là se trouve aussi l'Amérique: *Californie, Mexique, Canada, Québec, Colombie, Rocheuses, Pacifique, Argentine, Chili, Buenos Aires, Rio de Janeiro, Montréal, Panama, Atlantique, Pacifique, Amazone*. C'est ici enfin 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, Océan Indien, Chine, Pékin, 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 l'Afrique 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 continent mythique n'a d'autre but que d'illustrer, en profitant des circonstances, les possibilités qui s'ouvrent à l'explorateur lancé sur les autoroutes de l'information[11]. La statistique et l'informatique ont toujours aimé les grands espaces, les grandes masses de données et la loi des grands nombres. Internet leur ouvre ses richesses, sans contrôle, sans retard, sans dépense et sans limites.

---

[11] Nous avons négligé un autre service que le réseau *Web* peut rendre à la linguistique: 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 qui peut accueillir tout de suite certaines des communications du présent Colloque. En voici l'adresse:

http://www.msh-paris.fr/~salem/revue.html .