



**HAL**  
open science

## L'Afrique et l'ordinateur

Étienne Brunet

► **To cite this version:**

Étienne Brunet. L'Afrique et l'ordinateur. Le corpus lexicographique, 1997, Louvain la Neuve, Belgique. pp.315-34. hal-01569007

**HAL Id: hal-01569007**

**<https://hal.science/hal-01569007>**

Submitted on 26 Jul 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 néophytes 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<sup>1</sup> 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?

---

<sup>1</sup> *Lycos*, qui est le plus puissant moteur de recherche branché sur *Internet*, avoue explorer 16,5 millions de documents (la progression est foudroyante 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*

The screenshot shows the Lycos Search interface. At the top, the title "Lycos Search" is displayed. Below it, the search query "Cameroun" is entered in a text box, with a "Search" button to its right. Underneath, there are search options: "match any term (OR)" and "loose match", both in dropdown menus. Below that, display options are shown: "10 results per page" and "standard results", also in dropdown menus. At the bottom left, there are two links: "Search language help" and "Formless Interface". At the bottom center, there is a navigation bar with links: "[ Home | Search | Lists | Reference | Add/Delete | News | Lycos Inc ]".

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.orstom.fr/ct6/demographies/E.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∞M∞—M\[∞,—\\*`aYCEHome Page](#) [0.9817]  
**Outline:** ∞J∞M∞—M[∞,—\*`aYCEHome Page ∞J∞M∞—M[∞,—\*`aYCE  
**Abstract:** \$B%+%a%k!  
 Embassy of The Republic of **Cameroun** in Tokyo  
 Embassy of The Republic of **Cameroun** in Tokyo  
 Home Page of the Republic of Cameroon in Japan (Cameroun)  
 ∞J∞M∞—M[∞,—\*`aYCEalg...U (∞'—1)  
<http://www.karia.kameji-dt.ac.jp/JPPublic/HLF.html/cameroun/home.html> (6k)

### *L'Afrique et l'ordinateur*

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%<sup>2</sup>. 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<sup>3</sup>.

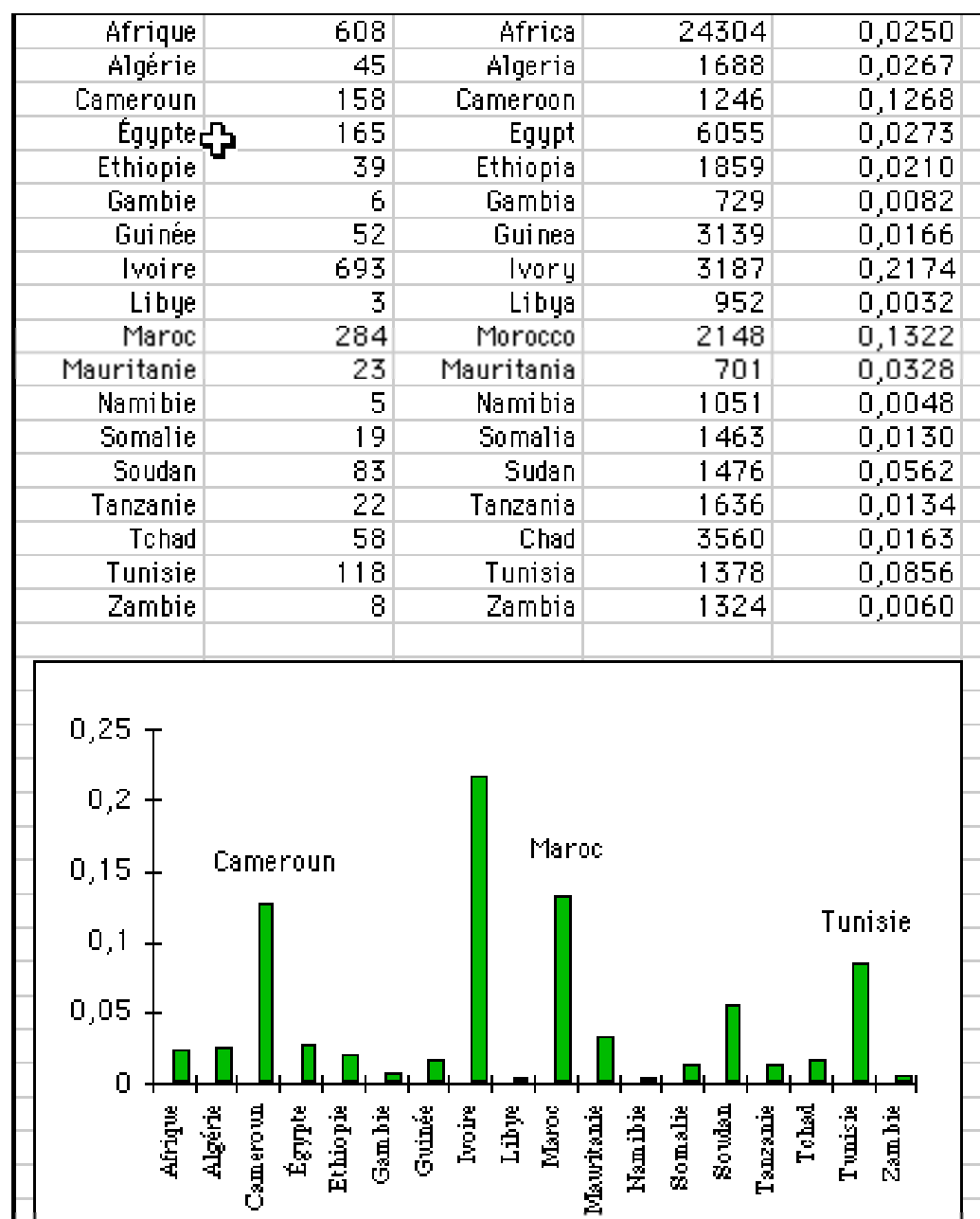
Mais en réduisant les doublets et en neutralisant les variations orthographiques (toutes les appellations d'un même site étant regroupées), on pourrait aussi dresser la carte *Internet* selon la place que chaque région du globe y occupe. La réponse est dans la figure 4, dont l'évidence décourage les 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

<sup>2</sup> 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).

<sup>3</sup> 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*



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 trône 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*<sup>4</sup>. 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<sup>5</sup>.

<sup>4</sup> Encore le Caire est-il à la jonction de l'Asie et de l'Afrique.

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



*L'Afrique et l'ordinateur*

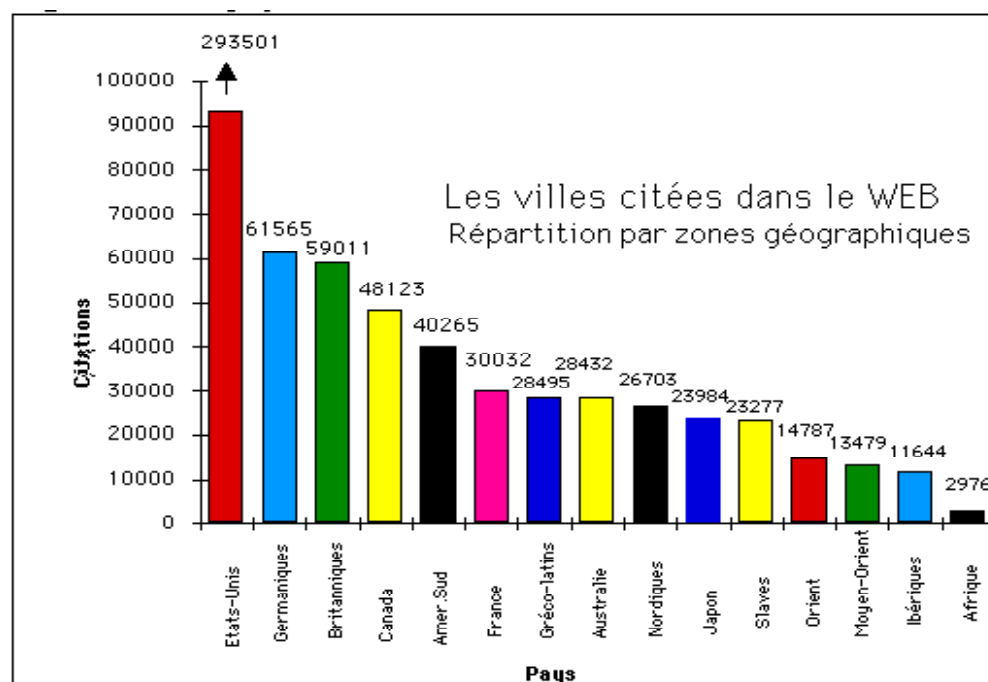
Figure 4. Les citations toponymiques d'*Internet*  
classées par ordre de fréquence décroissante

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	<b>Cairo</b>	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), [djbouti](#) (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 mœurs 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

CONCORDANCE			
Dico Liste Distr Spécif Concord Context Trig Séqu Trid			
AD 226b	seicheresse en tout le pays d'	Affricque Affricque , pource qu' il y avoit Affricus	1 1
DI 47a	partie du midy souffle Notus ,	Affricus , et Auster . A cause duquel Affricque	3
CI 20a	assé en revenant du pays d'	Affricque , que bien tost y devoit	
CI 20b	qui fust en toute l' isle .	Affricque , dist Pantagruel , est	
CI 175b	n' en contiennent l' Asie , l'	Affricque et l' Europe ensemble . ¶Et Affricque	5
PA 18b	tant grande en tout le pays de	Affricque , que passerent xxxvj moys ,	
GA 86b	de Numidie envoya du pays de	Affricque a Grandgousier une jument la	
GA 86c	, Comme assez scavez , que	Affricque aporte tousjours quelque	
TI 279b	et pacifique en Europe ,	Affricque , et Asie . Notez combien je	
QU 42d	sus la poincte Meridionale d'	Affricque , oultre l' Aequinoctial , Affricque	4
TI 180c	Tigris et Euphrates . ¶Uoy la	Affricque . icy est la montaigne de la	
QU 22a	de la Mirandole , et de	Affricque . ¶Ainsi nomment les mortelz ,	
QU 44c	riches et fameux marchans d'	Affricque et Asie . ¶D' entre les	
QU 60c	non pour toutes les bezicles d'	Affricque . Car j' ay une des plus	

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

Pe 106b	, car je suis la Gorgone	africaine	6
PI 151e	aux calé basses d' une claière	africaine . ¶LA GRANDE PRÊTRESSE . Tu	
L2 39b	connaisse qui semble à la fois	africaine et mouillée . ¶Par une claire	
LE 209a	à ce décalage . L' arriération	africaine des steppes de la Castille	
FU 115a	, aujourd' hui , de cette vision	africaine , est une coulée hybride que	
SC 26a	pâmée , cette reddition	africaine à l' incendie solaire qui s' africains	2
BF 11e	un morne troupeau de Nord	Africains . ¶	
GC 35a	j' imagine , l' abord des marais	africains , se voit toujours aux Afrique	10
BF 163a	parlaient du pays comme d' une	Affricque indigène , où il est plaisant	
Pr 159a	rivage , que j' ai rencontrés en	Affricque ? Assis les jambes croisées sur	
Pr 245a	, et vient s' engager en	Affricque dans la Légion étrangère . ¶On	
Le 42b	au - dessus des paysages d'	Affricque , tantôt haut , tantôt bas ,	
L2 27a	comme au long d' une mangrove d'	Affricque ou d' une grève de Papouasié :	
L2 54c	cocotiers absents de la superbe	Affricque . ¶Cela n' est plus . Mais cela	
L2 216f	au bord d' un marais fermenté d'	Affricque . ¶	
EE 11e	fleuves fabuleux de l' ancienne	Affricque , n' avait ni source ni	
FU 46b	blancs des anciennes cartes d'	Affricque , terre inconnue : une vaste	
FU 46e	ainsi que le Nil au coeur de l'	Affricque , par une artère sinueuse ,	

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 qui concerne la finale, qui hésite entre la forme archaïque *-icque* et la forme moderne *-ique*. Par



## L'Afrique et l'ordinateur

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<sup>6</sup> 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

<sup>6</sup> "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*

**CONTEXTE**

Emploi d'un filtre ?  non  oui  
(le filtre est le premier mot ou signe du paragraphe)

---

Disualisation  oui  non

---

Portée d'action  corpus  texte particulier

---

Paragraphes (ou lignes) avant et après  0  1  2  3  4  5

---

Objet de la recherche  forme  cooccurrence  début de mot  fin de mot  chaîne  expression

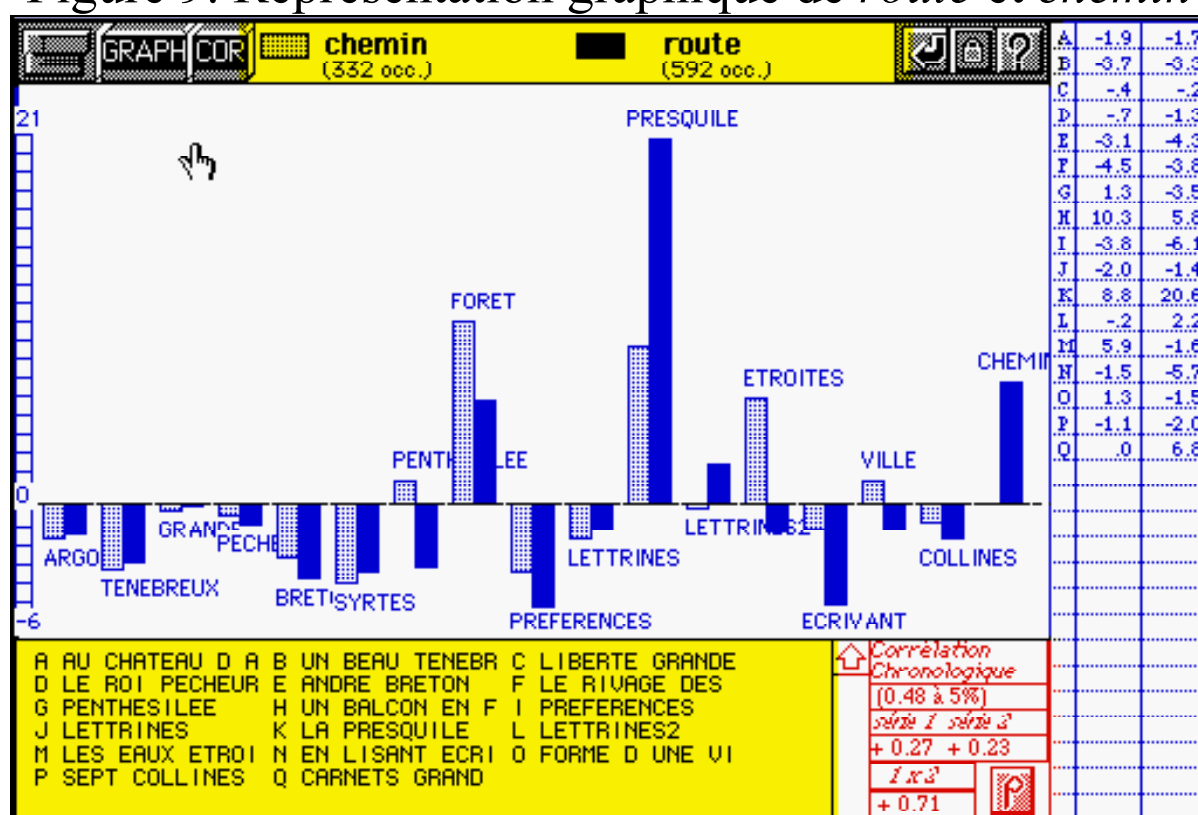
Exemple: amour  
Exemple: amour...toujours  
Exemple : aim  
Exemple: isme  
Exemple : phag  
Exemple: comme si

**OK**

Choisir les options puis cliquer le bouton OK

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<sup>7</sup>, 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



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

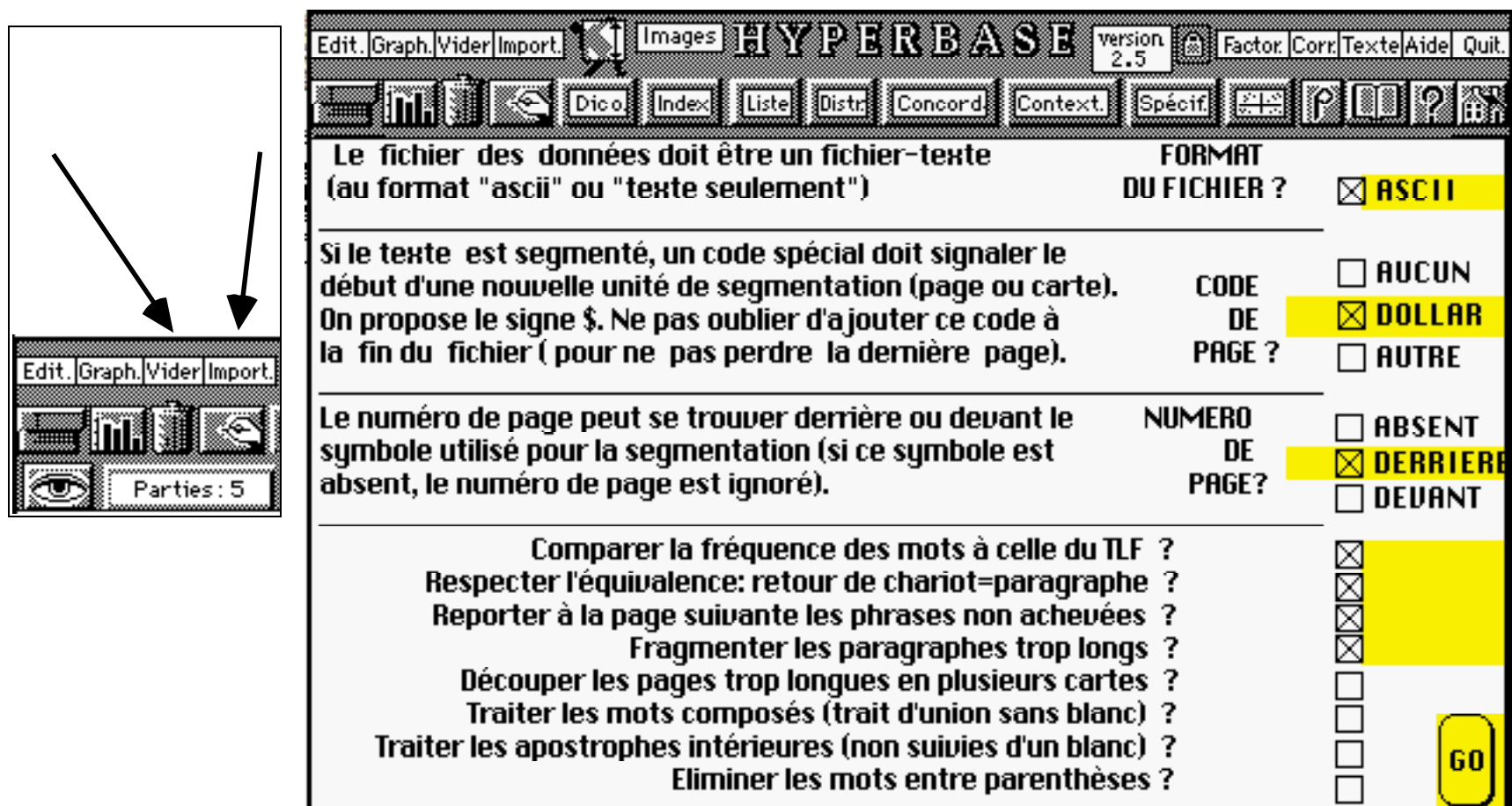


apparue dernièrement (version 2.5, Juillet 1995), qui mesure la distance que chaque texte établit avec tous les autres du même corpus, et qui est le rapport entre les vocables communs aux deux textes que l'on confronte et les vocables exclusifs que chacun des deux se réserve. Cela, qui peut s'appeler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

### 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

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

## *L'Afrique et l'ordinateur*

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 *Pascal*), 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<sup>9</sup>). 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

---

<sup>9</sup> 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.

## L'Afrique et l'ordinateur

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.

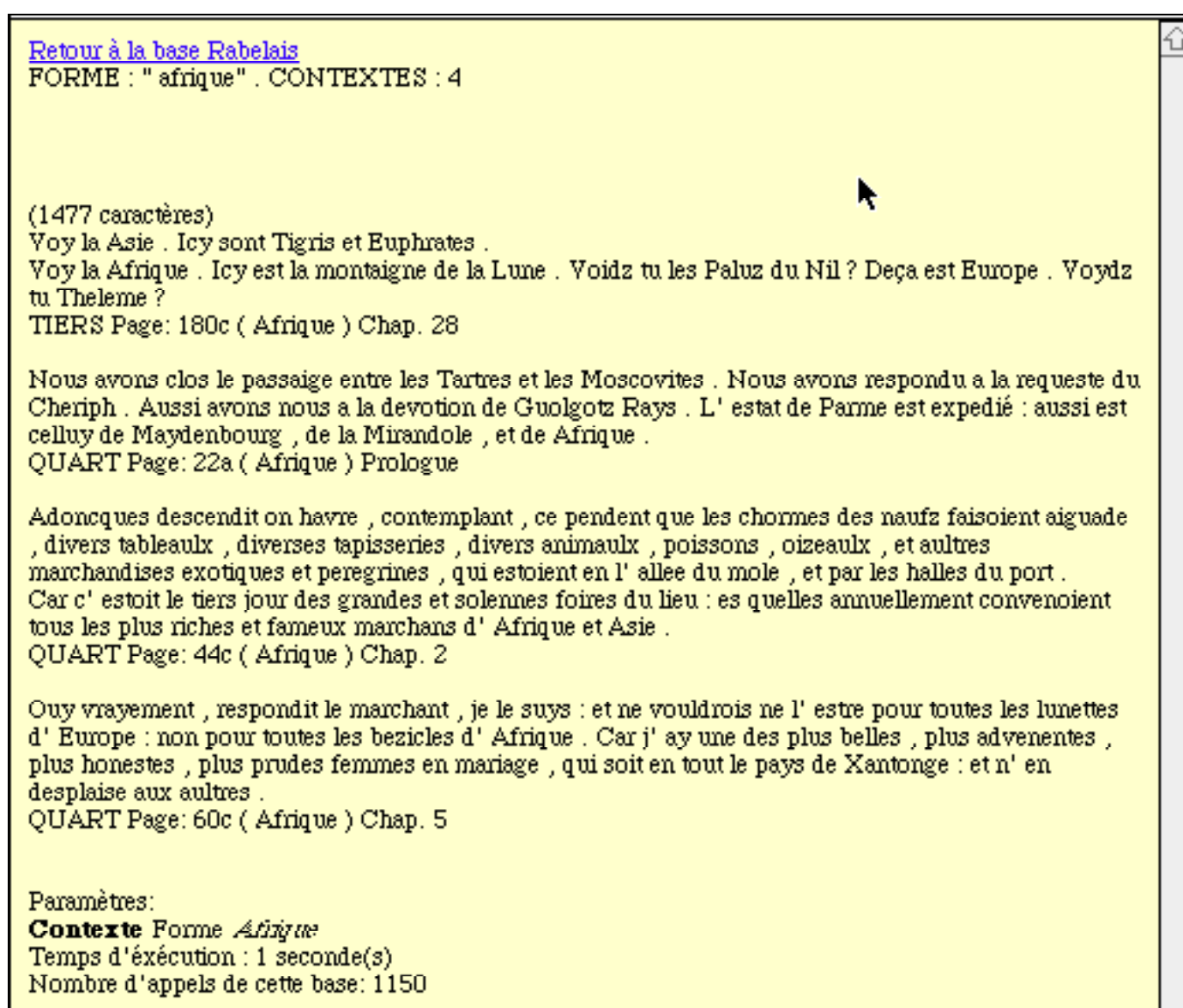
L'écran d'accueil est reproduit dans la figure 13. Il contient peu d'éléments véritablement statistiques, étant orienté d'abord vers les requêtes documentaires, et en particulier vers la recherche des contextes.

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

The screenshot shows a web browser window with a navigation toolbar at the top (Back, Forward, Home, Reload, Images, Open, Print, Find, Stop) and a location bar containing 'http://134.59.31.3/rabelais.html'. The main content area has a yellow background and is titled 'RABELAIS ET SON TEMPS'. It contains a list of links: '1 - Illustrations relatives au texte ou à l'époque de Rabelais', '2 - Structure du vocabulaire et distance', '3 - Vocabulaire spécifique', and '4 - Liste de mots, statistique, analyse factorielle'. Below the links is a section for choosing options: '1 - Traitement à opérer (concordance, contexte, lecture, graphique)' with radio buttons for 'Concordance Aide', 'Contexte Aide' (selected), 'Lecture Aide', and 'Graphique Aide'. '2 - Options à sélectionner: (forme, initiale, finale, chaîne, expression, cooccurrence)' with a 'Forme' dropdown menu. '3 - Présentation (pour une concordance uniquement):' with an 'Aucun tri' dropdown menu. '4 - Taper l'objet à traiter dans le champ A (si l'objet est double, utiliser le champ B pour le deuxième élément):' with input fields for 'Champ A' (containing 'Afrique') and 'Champ B'. There are 'OK' and 'Valeurs par défaut' buttons. At the bottom, there is explanatory text about the object to be treated and a list of titles: 'Pantagruel (1), Gargantua (2), Tiers (3), Quart (4), Cinquième (5), Inestimables (6), Admirables (7), Disciple (8), Prognostication (9), Nouvelle (10) ou Ensemble (11, pour le corpus entier) et, facultativement, dans le champ B la page désirée si on la connaît (première page par défaut)'. There are also links for 'Le laboratoire Statistique linguistique (INaLF, CNRS)', 'Statistique de la base', 'Histogramme des appels', and 'Courrier électronique'.

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<sup>10</sup>, 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.

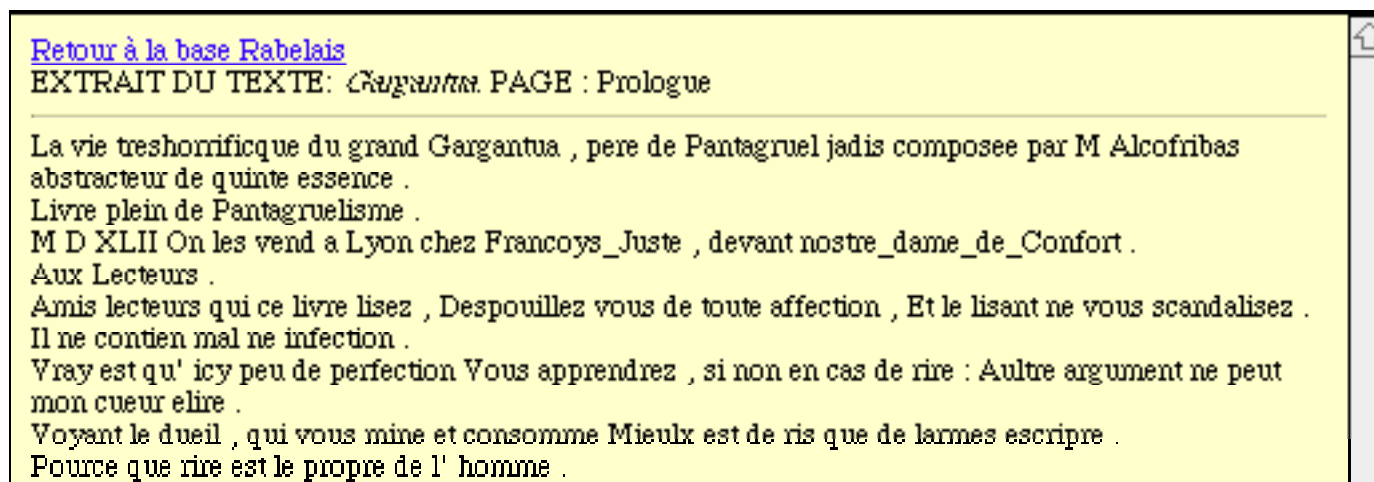
Figure 14. Le contexte de l'*Afrique* dans la base Rabelais (sur Internet)



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).

<sup>10</sup> 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

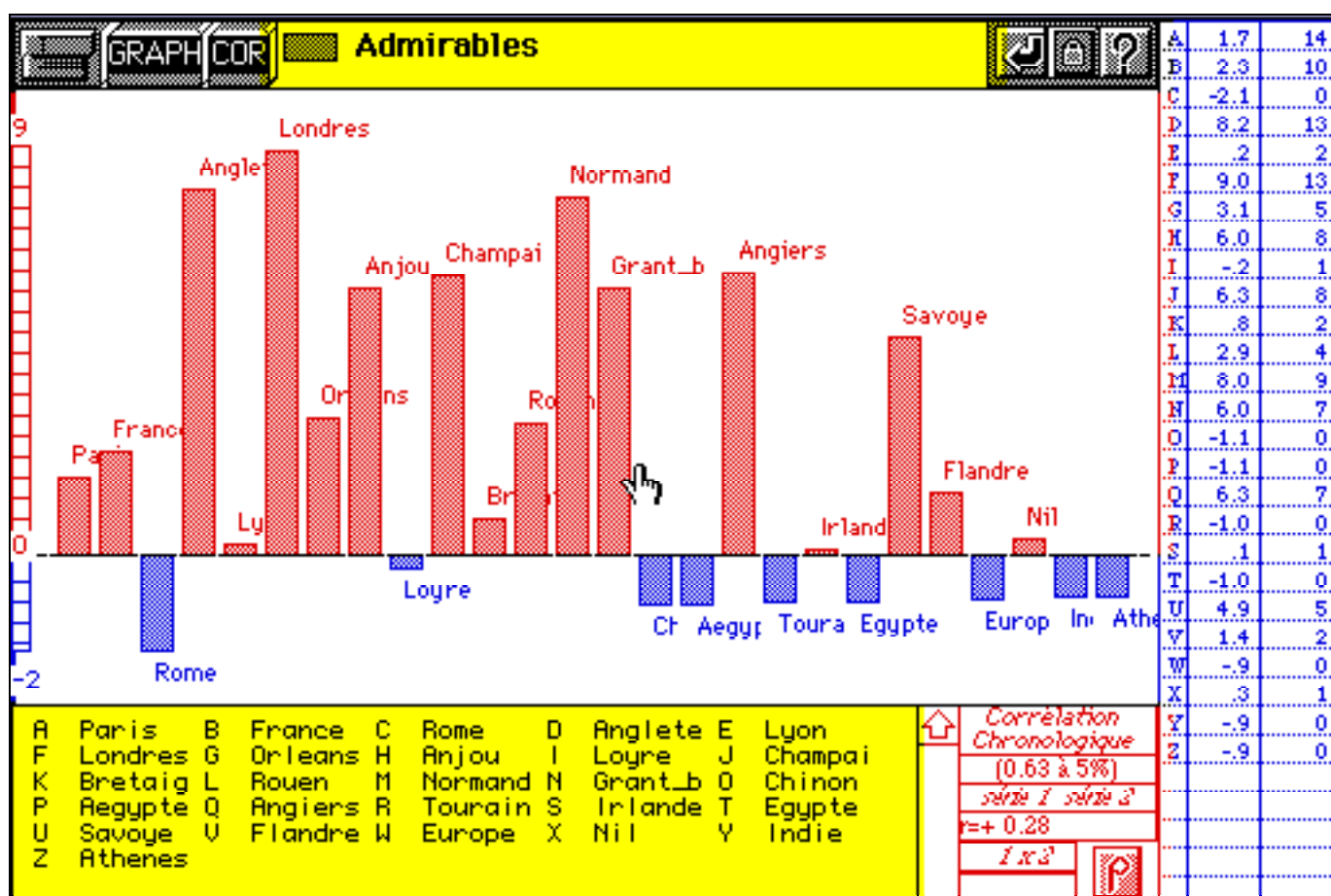
L'objectif est ici d'abord de constituer des tableaux à deux dimensions dont les lignes sont dévolues aux mots et les colonnes aux textes et qui sont la matière même de la plupart des traitements quantitatifs. Au croisement de la ligne  $i$  avec la



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 Champagne, la Bretagne 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*)

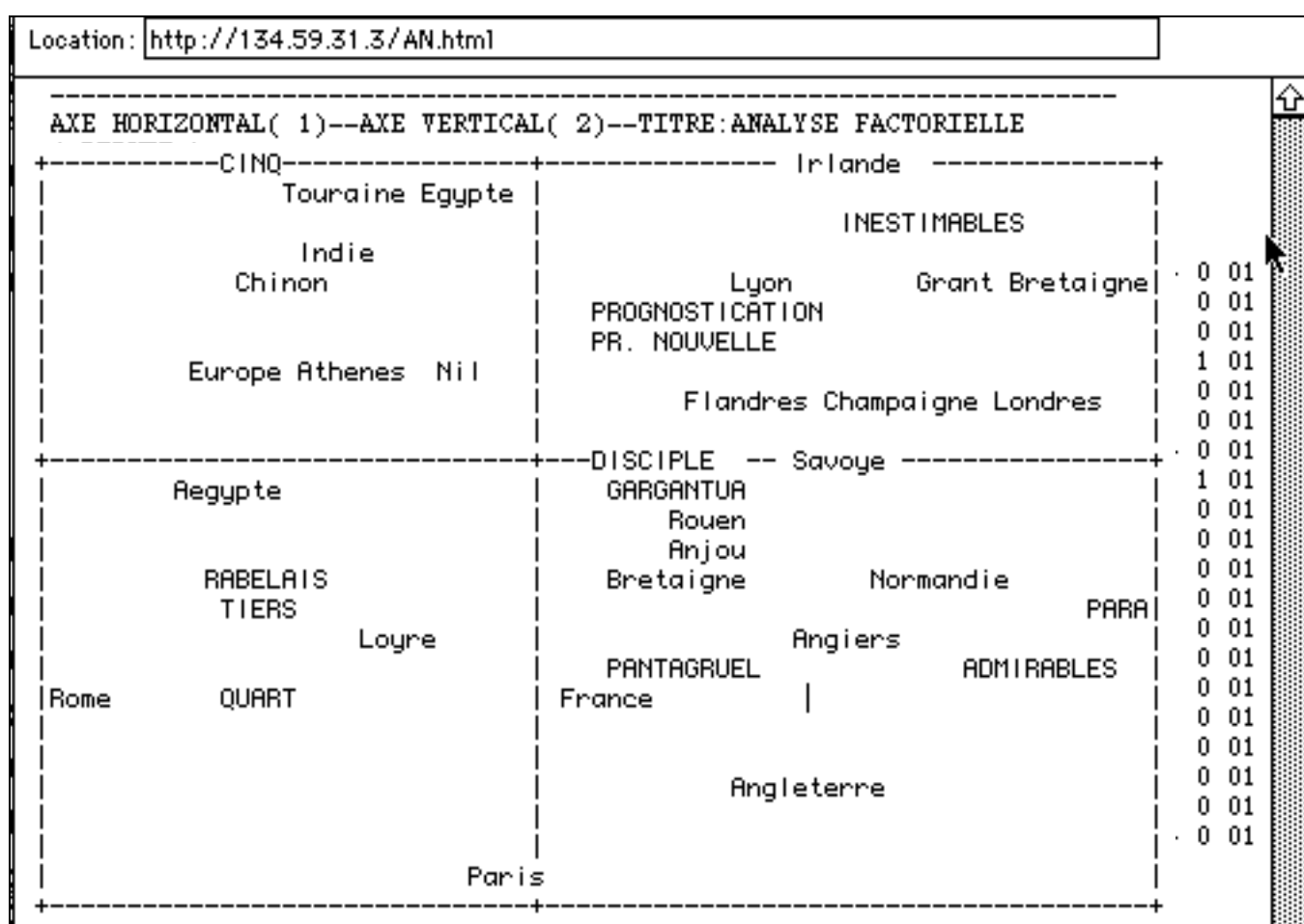


*L'Afrique et l'ordinateur*

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 démunie 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 pas parmi les spécialistes sur les meilleurs tests statistiques à appliquer aux données, les auteurs de *Frantext* n'en ont choisi aucun et les données numériques sont livrées dans un état presque brut, 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

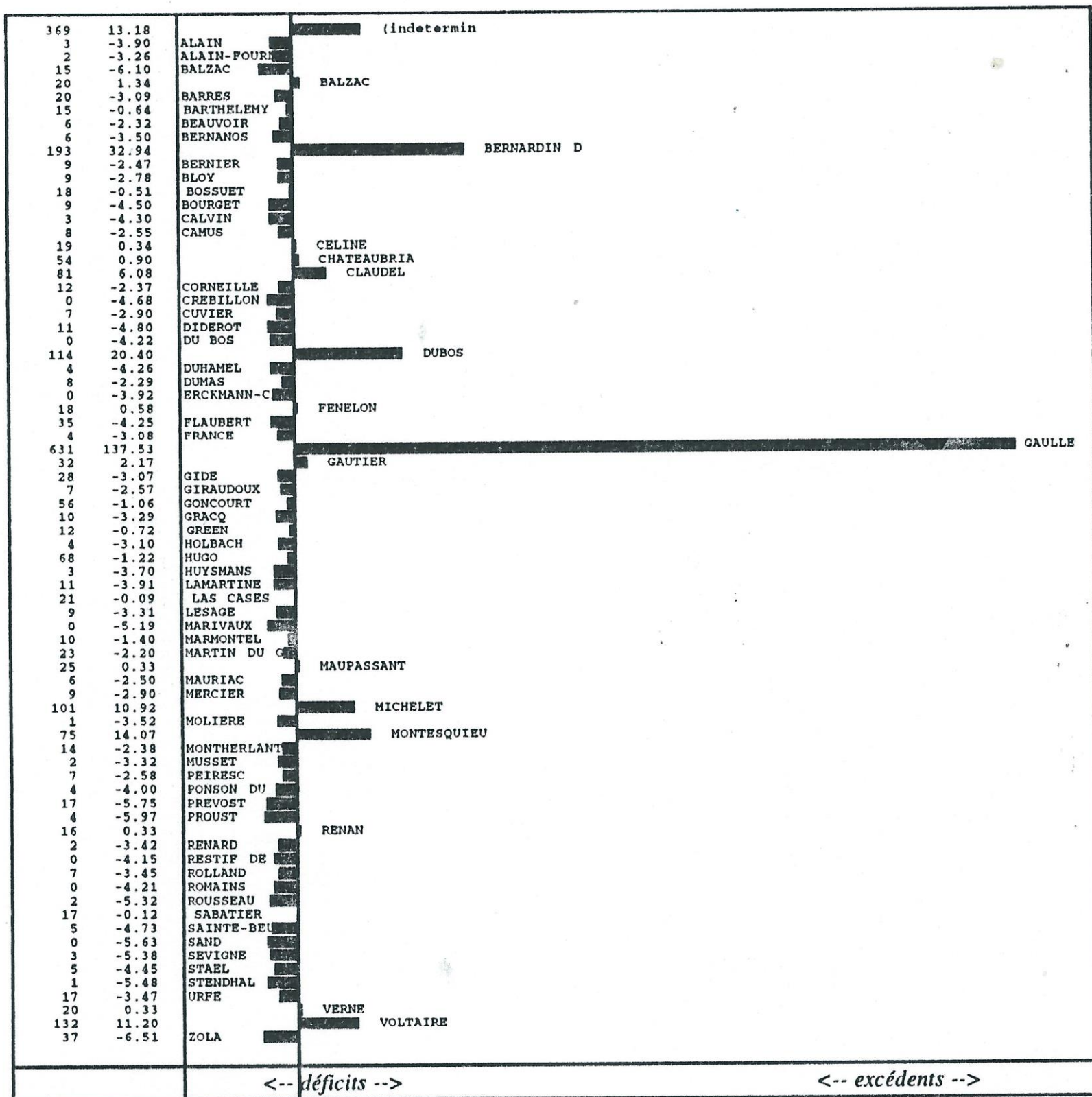






Cette distribution se retrouve lorsqu'on isole les auteurs. Un millier d'écrivains ont été passés en revue dont 75 prennent place dans le graphique (ce sont ceux qui sont le mieux représentés dans le corpus, avec plus de 500000 mots). Ceux qui citent l'Afrique le plus souvent appartiennent au 18e et 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 Gervitch, 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'accrochent au pôle pour en définir une partie, de type géographique, comme les deux premiers de la liste: *équatoriale* et *occidentale* (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*

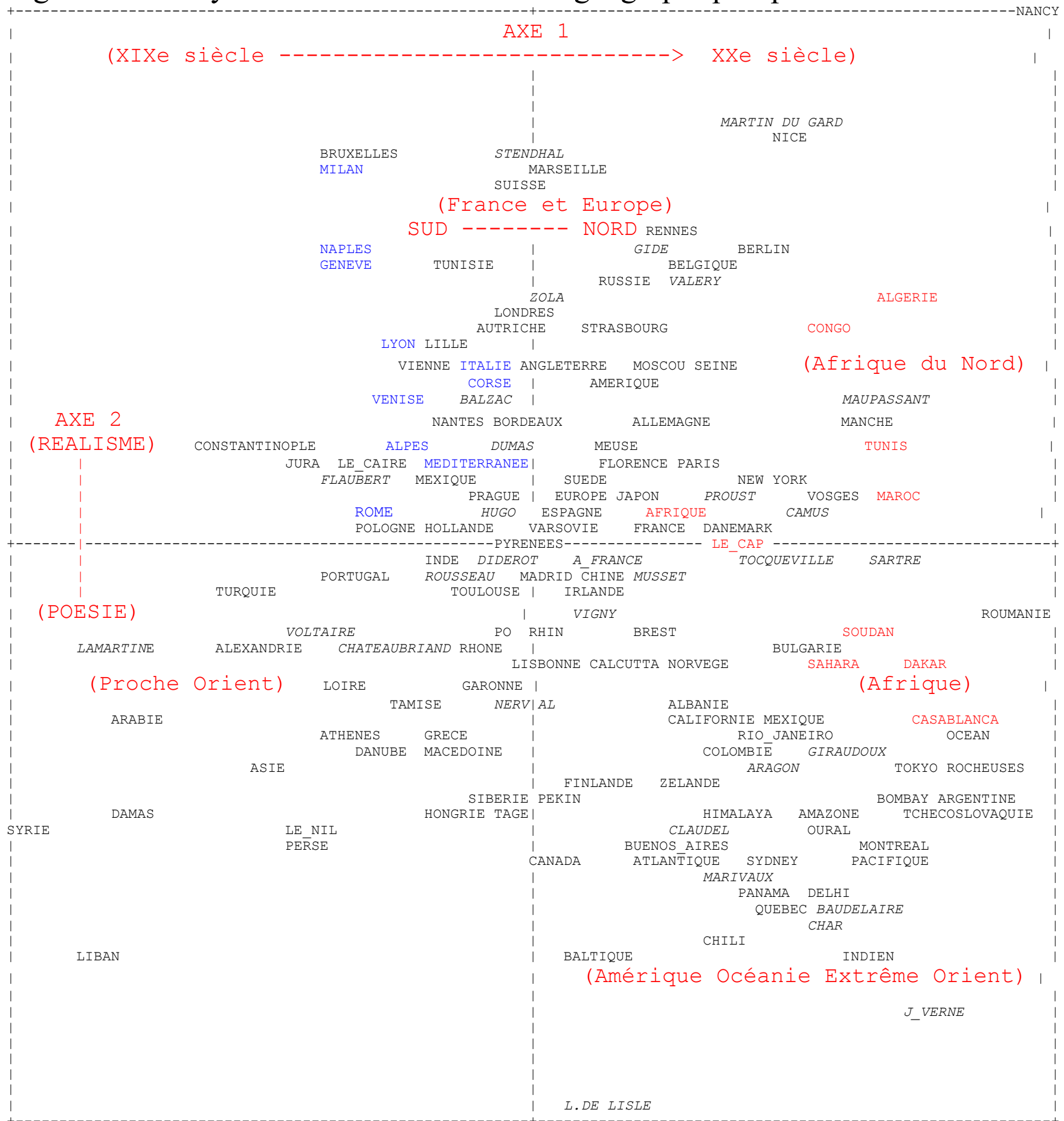
fréquence &gt; 10 écart réduit &gt; 3

Corpus de référence : Corpus total

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

## L'Afrique et l'ordinateur

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<sup>11</sup>. 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.

<sup>11</sup> 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 *Lexicométrica* 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> .