Le français sur Internet

Etienne Brunet

Étienne Brunet

Le français sur Internet

Résumé.

Dans l'immense magasin multimédia qu'est devenu Internet, les informations textuelles sont les mieux traitées parce qu'elles se prêtent mieux que les autres à l'indexation. Cette bibliothèque virtuelle, dont aurait rêvé Borges, a ses rayons éparpillés aux quatre coins du globe sans qu'aucun document pourtant échappe au contrôle des moteurs de recherche, dont la puissance permet d'apprécier les parts de marché que les pays et les langues se disputent sur les réseaux de communication et dont nous montrerons l'évolution depuis dix ans. La consultation des compteurs et mouchards qui enregistrent les flux d'échanges n'est guère favorable à la langue française, dont l'emploi sur Internet stagne ou régresse au fil du temps. Cependant dans le domaine culurel l'exception française n'est pas sans effet sur Internet où les populations littéraires disposent de bases de données spécifiques, ouvertes à l'appétit des chercheurs, et notamment Frantext, que l'on explorera, la lampe statistique à la main.

- I –

Rappel historique

La croissance exponentielle d'Internet fait songer à un Big Bang dont l'explosion rend aveugle aux débuts et à l'étincelle initiale. Aussi comme pour le Big Bang cosmique la naissance du phénomène revêt un intérêt particulier. Si les mots « site » et « Toile » ou « Internet » sont clairs maintenant pour tout le monde, ils sont moins anciens que la chose. Auparavant on parlait de station de travail et de réseau. Et dès les années 70, le Centre IBM de La Gaude (près de Nice) était en liaison avec des ordinateurs de Californie. Le premier réseau européen a été un réseau IBM, gratuitement mis à la disposition des universitaires européens sous le nom de EARN.

En France le réseau Transpac a pris le relais au début des années 80. Des ressources, dont certaines linguistiques et littéraires comme celles du Trésor de la langue française à Nancy, pouvaient transiter sur le réseau, parfois même avec un simple minitel. En ce sens on peut dire que le Trésor de la langue française a peut-être été le premier serveur de données littéraires et linguistiques, en même temps que le site miroir installé à Chicago sous le nom d'ARTFL par Morissey. Même si la version WEB de FRANTEXT n'existe dans sa forme actuelle que depuis 1995, des services similaires étaient rendus depuis des années par le logiciel d'interrogation Stella, réalisé par Jacques Dendien à Nancy. Et parallèlement sur le réseau Minitel à la fin des années 80 Charles Muller proposait sa base interactive Orthotel (qui existe toujours, mais maintenant sur Internet et sous le nom d'Orthonet).

C'est dans la dernière décennie que le WEB voit le jour. Progressivement il va se substituer, grâce au codage HTML, aux logiciels de travail à distance comme Telnet, de transmission de fichiers comme Fetch ou FTP, d'information comme Gopher ou News, ou de messagerie comme Eudora.

Les navigateurs Mosaic, puis Netscape, puis Microsoft Explorer s'imposent définitivement, en même temps que les premiers métaserveurs, en particulier Lycos qui a eu son heure de gloire bien avant Altavista et Google. Mais parallèlement les procédures anciennes se maintenaient encore et les grandes bibliothèques étaient ouvertes au dialogue, comme celles de l'université de Californie (tous sites réunis), la bibliothèque du Congrès ou, en France, la Bibliothèque Nationale. Bien sûr on ne téléchargeait pas les textes, mais on avait accès au catalogue à distance. C'est au moment où se répand le code HTML (qui, comme XML, n'est qu'une instanciation du code SGML antérieur) qu'on peut vraiment parler de site et d'Internet. Et cela ne permet guère de remonter au delà de 1993.

Qu'on nous permette un instant d'évoquer notre modeste contribution aux débuts de l'Internet littéraire francophone (je n'ose trop utiliser le sigle ILF, qui est déjà requis par l'Institut de Linguistique Française et que je ne puis ignorer, puisque mon laboratoire en fait partie). On se contentera de noter l'apparition de son site (ancilla.unice.fr) dès 1994 et un peu avant celui de l'INaLF. Ce site est encore accessible actuellement, sous une forme à peine changée, sur un vieux Macintosh qui a plus de dix ans. Le laboratoire proposait alors sur un serveur Sun un autre site (lolita.unice.fr) qui a disparu, parce que le nom pouvait prêter à confusion au moment où le sexe a envahi la toile. Et dès juillet 1995 le laboratoire mettait sur son site la base interactive Rabelais et son temps (base CGI, à l'adresse http://**ancilla**.unice.fr/**rabelais**.html) et quelques mois plus tard la base Balzac (adresse:http://**ancilla**.unice.fr/~brunet/**BALZAC/balzac**.html) Ces bases sont toujours opérationnelles, sans changement, au bout de dix ans, la dernière traitant plus de 10 000 requêtes par jour (en entendant par requête tout fichier transmis)[1].

On se gardera de toute nostalgie en évoquant ces temps héroïques, dont les vertus sont surtout négatives. Le spam était inconnu et les échanges se faisaient de gré à gré, entre gens de bonne compagnie. Les virus ne circulaient guère sur le réseau et restaient confinés dans les disquettes et disques locaux. Mais quelle lenteur dans les communications et quelle insécurité. Et surtout quelle pénurie dans les entrepôts de l'information. On n'était jamais sûr que le document cherché fût disponible quelque part ni que la question posée pût avoir une réponse. On circulait dans le silence et le désert, en rêvant d'autoroutes.

Les autoroutes de l'information sont venues très vite, amenant le bruit et la cohue. Quel pépiement inextricable dans le gigantesque marché qu'est devenu Internet, quel vacarme dans ce souk à l'échelle du monde, où chacun propose et vante sa marchandise. Universités, laboratoires, grandes sociétés et petites entreprises, et jusqu'aux individus même, chacun arrange sa vitrine, avec cette impudeur que permet la distance et l'anonymat. Un plaisantin a pu appeler cela BabelWeb.

II
La part du français sur Internet

Si le Web est une tour de Babil, ce n'est peut-être pas une tour de Babel. Car la chute de la cité antique est venue soudain d'une perversion de la communication et de la multiplication des langues, au lieu qu'Internet paraît tendre à l'unification des langues et à la généralisation de l'anglais. La réduction de l'alphabet à un code ascii réduit et incomplètement défini a gêné dans les premiers temps la diffusion des langues pourvues d'accents, au point que certains usagers français ont parfois renoncé à utiliser sur leur clavier les lettres accentuées et ont fini par renoncer au français même. La technique a depuis lors proposé des solutions, dont la plus radicale est celle de l'unicode, et les langues non occidentales peuvent désormais s'exprimer sur Internet en dehors de l'alphabet latin. Le français a-t-il bénéficié de ce progrès technique, comme aussi de la

---

[1] Pour de plus amples détails, on peut consulter un article publié sur internet dès 1995 sous le titre "Le triple double V:" , à l'adresse http://ancilla.unice.fr/~brunet/PUB/lessive.html.

multiplication des ordinateurs dans les bureaux et les foyers nationaux? On avait l'espoir il y a dix ans que le retard du français serait ainsi comblé. Or il semble que ce vœu ait été vain et que la part du français dans les échanges sur Internet soit restée faible. Et les ressources d'Internet peuvent précisément aider à en prendre la mesure. On y trouve des moteurs de recherche comme *Lycos, Altavista, Hotbot* ou *Google* qui couvrent une énorme masse de documents et autorisent des sondages statistiques dans cet océan de mots. Pour mesurer et comparer la part respective de l'anglais et du français (ou de toute autre langue), il suffit de choisir deux mots ou deux séries de mots qui se correspondent d'une langue à l'autre et d'observer le nombre de leurs occurrences dans les données explorées par le serveur. Prenons l'exemple du couple *homme/man* où le moteur *Lyco*s propose le rapport 2 081 084 / 51 446 825, soit 4%, chiffre confirmé par *Hotbot* et *Google*. Au pluriel le rapport *hommes/men* est plus défavorable au français : 1 613 000 / 131 710 000, soit 1,2% (données de Google fin Juillet 2005). La *femme* française semble mieux résister, le rapport femme/woman étant de 3 951 000 / 74 520 000, soit 5,3%, mais c'est oublier qu'en anglais la femme en tant qu'épouse se dit *wife*, dont l'effectif (57 430 000) doit être ajouté à celui de *woman* et le pourcentage baisse alors à 3%. Simple, la méthode de calcul reste fruste et peu robuste, car elle se heurte aux idiotismes nationaux qui donnent aux choses et aux mots une extension inégale d'une langue à l'autre et empêchent l'équivalence biunivoque. On ne se fiera que médiocrement au couple *sexe/sex*, l'un des plus présents dans le commerce électronique. Le rapport, qui est de 4,2 % (3 600 000 / 83 800 000 , données de Google en juillet 2005) dépend des connotations et des inhibitions variables qui accompagnent le mot dans la conscience des peuples. Et ce qui est vrai pour les mots dits pleins, à forte charge sémantique, l'est plus encore pour les mots grammaticaux : le rapport *rien/nothing* qui atteint une valeur exceptionnelle de 6% n'a pas de sens si l'on ignore les différences dans l'expression de la négation en français et en anglais.

Il n'en reste pas moins que ces rapports, peu fiables dans l'absolu, aident efficacement à déceler une évolution relative au cours du temps. Il se trouve que nous les avons calculés il y a neuf ans exactement, le 16 août 1996, à un moment où le serveur Lycos revendiquait avantageusement 60 millions de documents, soit quatre fois plus que l'année précédente à la même date, mais mille fois moins qu'aujourd'hui. Voici les rapports observés alors :

Homme /man : 3673/65111 = 5,6 %
Femme/woman : 2891/38118 = 7,6 %

Si l'on ajoute foi à ce mince échantillon, le français, loin de rattrapper son retard sur l'anglais, s'est laissé distancer plus encore. Les pessimistes disaient alors que la part du français était de 5 % . Aujourd'hui les optimistes n'osent plus l'affirmer.

On peut récuser le témoignage de l'*homme* et de la *femme* et exiger des chiffres moins suspects. Plutôt que d'explorer le dictionnaire à la recherche de témoins neutres et fiables, il y a avantage à se fonder sur des noms propres car ils ont souvent la propriété d'être exclusifs, c'est-à-dire de ne souffrir ni synonyme, ni hyperonyme, ni appellation concurrente. Les noms géographiques peuvent jouer ce rôle, du moins lorsque leur forme graphique est différenciée selon la langue, et permet le partage. Car l'univocité de certaines graphies comme *Berlin* ou *Madrid* laisse le jugement en suspens. On prendra garde que les pays voisins comme l'Italie ou l'Espagne sont plus fortement représentés dans la conscience française, et donc dans les pages françaises, et bénéficient d'une plus-value. Mais s'il s'agit d'un pays lointain qu'aucun lien particulier, ni historique, ni géographique, n'attache à la France, la part du français se réduit, jusqu'à devenir négligeable.

Italie / Italy : 1 849 000 / 40 790 000 = 4,5 %
Espagne / Spain : 6 490 000 / 139 000 000 = 4,6 %
Chine / China : 4 760 000 / 285 000 000 = 1,6 %
Japon / Japan : 4 890 000 / 319 000 000 = 1,5 %

(données de Google, juillet 2005).

Il est toutefois possible de varier les procédés d'investigation et d'être plus précis dans l'estimation. Ainsi Google propose de choisir la base des documents exploités, de l'étendre à l'ensemble de ce qui est accessible sur le réseau ou bien de limiter l'exploration aux documents de la francophonie ou de l'hexagone. Cette fois, pour que la comparaison ait un sens, il faut opter pour des mots dont la forme ne varie pas, d'une langue à l'autre et, à tout le moins, du français à l'anglais. Mais il faudrait aussi que le sens, la valeur et l'intérêt porté au signifié soit constant. Cela exclut les mots nationaux ou régionaux, comme France (10%), Paris (20 %), Europe (6%), ou euro (7%), qui sont naturellement plus présents dans la conscience et le discours des français, ou inversement les mots comme Chicago (0,5 %), Washington (0,4 %), ou même dollar (1,8 %) dont l'intérêt est moins vivement ressenti dans l'hexagone. Mais si le mot choisi est de portée universelle, sans préférence nationale trop visible, le rapport entre les bases consultées donne la mesure approximative de la part du français dans les échanges. Proposons par exemple à Google le mot *Google* précisément. L'image que Google donne de soi est flatteuse puisque 225 millions de pages répondent à l'appel. Comme dans le même temps on relève 4 millions d'occurrences dans la base francophone, le pourcentage ne dépasse pas 1,8% pour le public français. C'est à peu près le même résultat qu'on trouve avec les mots ou sigles *IBM* ou *Microsoft*.

IBM : *1 510 000 / 86 100 000 = 1,7 %*
Microsoft *3 650 000 / 231 000 000 = 1,6 %*

Si l'on objecte que Google, IBM et Microsoft sont des sociétés américaines et qu'à ce titre la préférence nationale peut être tentée de les écarter, on pourrait envisager de faire porter l'enquête sur des éléments neutres, comme les lettres de l'alphabet. Le rapport pour la lettre *a* est de 2%, de 3 % pour le *b*, de 4,7% pour le *c*. Mais l'exemple du *k* et du *w*, dont le français use fort peu, montre que la composition alphabétique des langues n'est pas constante et que ce critère manque de stabilité. On fera plutôt confiance aux chiffres dont la liste, élément par élément, se présente comme suit dans les données de Google, le 23 juin 2005 :

Chiffre	*Web français*	*Web total*	*%*
0	13 700 000	412 000 000	3,3
1	35 000 000	1 470 000 000	2,4
2	42 500 000	1 410 000 000	3,1
3	28 300 000	1 010 000 000	2,8
4	25 300 000	894 000 000	2,8
5	24 900 000	910 000 000	2,7
6	28 200 000	762 000 000	3,7
7	20 000 000	695 000 000	2,8
8	19 100 000	689 000 000	2,8
9	23 000 000	632 000 000	3,6

Cette fois, les chiffres s'accordent et ne s'éloignent guère de 3%, pourcentage que nous retiendrons en fin de compte pour l'estimation de la part du français sur Internet.

Quoique apparemment assez précis et très simple, le calcul n'est pourtant pas exempt d'ambiguïté. Il semble se fonder sur le suffixe attribué à chaque pays pour l'adressage des serveurs

et la provenace des documents. Ceux où figure le suffixe *.fr* sont estimés être situés en France[2]. Ce n'est pourtant pas le cas de tous puisque certains sites nationaux peuvent obtenir un autre suffixe. Et il ne suffit pas non plus d'inclure les suffixes *.ca* (Canada) et *.be* (Belgique) pour faire le tour de la francophonie. Inversement le français n'est nullement obligatoire dans la rédaction des documents catalogués sous la rubrique *.fr*. Enfin si certains objets neutres, comme les chiffres, peuvent échapper aux préférences nationales, il n'en va pas de même pour les mots dont la mesure n'est pas seulement liée à la langue, mais aussi à l'intérêt variable qui s'attache au référent dans la conscience nationale. Le *blé*, le *riz* et le *maïs* n'ont pas la même extension dans les sites d'Europe, d'Asie ou d'Amérique. Ce n'est pas une question de langue, mais de culture, au double sens, agricole et culturel, du mot.

III
L'exception culturelle

La faible présence du français sur Internet et, ce qui est plus grave, son recul au cours des ans, même quand la France a comblé son retard en équipement informatique, ne permettent guère de grandes espérances dans le marché linguistique, qui s'est mondialisé comme le marché économique. La part du français s'est encore amincie, si l'on considère les autres langues que l'anglais, surtout celles qui n'utilisent pas l'alphabet latin et qui, pour des raisons techniques, n'avaient pas encore accès à Internet il y a dix ans. Sans doute souffrent-t-elles aussi de la confrontation à l'anglais, mais disposant d'une clientèle locale très étendue, comme le chinois, le russe, l'arabe, l'espagnol ou le japonais, elles ont un large marché captif, privilège refusé à une nation comme la France, dont la population reste moyenne. En outre même dans le marché intérieur, l'anglais s'offre en France comme une alternative facile et séduisante, puisque l'anglais et le français sont très proches, au moins dans la forme écrite, et que la plupart des français, au sortir de l'école, peuvent s'exprimer vaille que vaille en anglais. Il suffit pour s'en convaincre de consulter la publicité des entreprises ou des laboratoires français, le mode d'emploi des produits fabriqués en France, ou le programme des colloques annoncés en France (excepté le présent colloque).

Mais, si paradoxal que cela puisse paraître, la marginalisation du français comme langue de communication ordinaire, à l'échelle mondiale, n'empêche pas la France et sa langue de bénéficier d'un intérêt soutenu à l'étranger. Certes Internet a dépassé depuis longtemps le cercle étroit des happy few et le commerce y occupe désormais une part prépondérante et souvent vulgaire. Mais de ses origines universitaires, Internet a gardé un parfum libertaire, un esprit jeune, et le goût des biens culturels. Et les produits français, notamment son art et sa langue, peuvent répondre à ce goût. Paris, qui évoque au plus haut point la France, n'a rien perdu de son attrait et les mentions qu'on en fait sur Internet en font une des villes les plus citées au monde. Dans le tableau ci-dessous qui établissait le palmarès mondial des villes en 1995, Paris n'occupait que la dixième place. Avec des effectifs mille fois supérieurs, le classement n'est pas bouleversé et London est toujours troisième derrière New York et Washington, mais Paris gagne deux places.

---

2 On pourrait d'ailleurs tenter de mesurer directement le volume des pages cataloguées comme françaises. Google évalue à 121 millions les documents portant la marque .fr, à peine plus que ceux qui viennent d'Espagne (suffixe .es : 108 millions) et nettement moins que les pages rattachées à l'Angleterre (suffixe :uk : 420 millions). Les sites américains ne sont que partiellement repérables (suffixe .gov : 488 millions, .edu 658 millions), les sites commerciaux (suffixe.com : 1800 millions) n'étant pas exclusifs.

Tableau 1. Le palmarès des villes sur Internet en 1995 (données de Lycos)

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

Nous laissons au lecteur le soin d'actualiser ce tableau, dix ans plus tard, en choisissant tel ou tel moteur de recherche. Voici ce que proposait Google dans les derniers jours de juillet 2005 (en millions d'occurrences) :

1	New York	129 millions	10	Atlanta	31 millions
2	Washington	107	11	Philadelphia	31
3	London	103	12	Seattle	28
4	Mexico	63	13	Madrid	28
5	Chicago	56	14	Toronto	25
6	Boston	48	15	San Diego	25
7	Los Angeles	44	16	Berlin	22
8	Paris	44	17	Rome	21
9	San Francisco	40			

Dans bien des cas le nom de *Paris* est évoqué comme étant la matière ou la cible du discours et non pas sa 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 qui profite non seulement à Paris, mais aussi à Mexico, Rome et Madrid, et dont ne bénéficient pas les sites modernes et purement économiques comme *Seattle* ou *Buffalo*.

Mais le nom de Paris n'est qu'un symbole ambigu et un témoin confus de la présence française sur la Toile. Les moteurs de recherche permettent d'affiner la recherche et de recenser les sites où s'affirme cette présence. Mais tant de sites naissent chaque jour que les tâches de dénombrement et de classement sont périmées dès qu'elles sont achevées. Nous

laisserons aux observateurs plus assidus le soin d'établir la carte des ressources linguistiques et culturelles qu'on peut trouver sur Internet et qui contribuent au rayonnement de la France. Afin d'échapper au chauvinisme, nous citerons un site universitaire, établi à New York (Lehman, Cuny) et consacré à la culture française. L'extrait très partiel représenté ci-dessous témoigne de l'intérêt que la langue française suscite encore dans le monde et des efforts qui ont été faits pour répondre à cette attente.

Tableau 2. Extrait du panorama des ressources linguistiques française
(http://www.lehman.edu/deanhum/langlit/french/lit.html)

Certes cette attente ne vise guère à la satisfaction de besoins utilitaires. Elle se mêle à d'autres aspirations que la France polarise pareillement : au goût du luxe, des parfums, des vins fins, de la haute couture et de la haute cuisine. Les amateurs de haut langage se tournent ainsi vers la France, comme jadis les romains cultivés vers la Grèce. Ils lui savent gré de garder une langue pure, peu accueillante aux emprunts et aux réformes et jalouse de ses subtilités, de ses conjugaisons et de ses difficultés, estimant que l'usage généralisé de l'anglais dans les affaires ne peut qu'abaisser cette langue à un rôle de truchement commercial et l'entraîner à la trivialité et à la déchéance.

Il se trouve aussi que la France et plus généralement l'Europe sont riches en biens culturels, accumulés tout au long de l'histoire. Dès qu'il s'agit de musique, de peinture, d'architecture ou de littérature, la France est une mine pour les chercheurs d'art, particulièrement dans les siècles passés où l'Amérique est démunie. De là ces tentatives renouvelées pour acquérir, sinon les trésors de nos musées ou de nos bibliothèques, du moins leurs doubles numériques et le droit de les reproduire et de les diffuser. Après la razzia des Mormons sur les registres paroissiaux de nos provinces, après le raid infructueux de Microsoft sur le Louvre, on assiste présentement à un essai d'*opa* de Google sur le patrimoine textuel et littéraire de la France et de l'Europe. Google propose de numériser, en six ans, quinze millions de livres représentant une bonne part de la mémoire du monde, soit près de 5 milliards de pages qui deviendraient accessibles gratuitement sur Internet. Google répondrait ainsi à une grave lacune d'Internet où abondent les informations sur l'actualité, même les plus insignifiantes, et où manquent les documents essentiels touchant au passé. Il est facile d'imaginer les réticences des responsables de la *BNF*, comme ceux du Louvre naguère[3], et l'on peut s'attendre à ce que les éditeurs soient peu disposés à céder leur copyright.

Pour mesurer où en était ce projet, nous avons interrogé *Google-Print* en lui proposant un personnage célèbre de notre littérature. Le nom de *Homais* aurait dû attirer au premier rang celui de son créateur. Or il n'en est rien. Flaubert est, bien sûr, cité par les critiques, la plupart anglophones, qui parlent de Homais, mais le texte de *Madame Bovary* n'apparaît pas.

On aurait mauvaise grâce à se réjouir de cet échec si l'Europe se contentait de dire non sans rien proposer. Or le ministre français de la Culture vient d'annoncer la semaine dernière qu'il était en charge d'un grand projet européen de bibliothèque numérique. S'il semble reprendre à son compte l'idée de *Google-Print*, en réalité le projet se situe dans la perspective de la *BNF*, dressée de longue date et maintenant agrandie à la dimension de l'Europe. Le dernier état des lieux publié par la *BNF* déclare 70 000 volumes engrangés en mode image, soit 21 millions de pages, et 1500 volumes en mode texte[4]. On est loin du projet pharaonesque de Google. Mais l'expérience est acquise, la clientèle fidélisée[5] et les procédures rôdées. Il suffirait d'accélérer la cadence de saisie et d'étendre le protocole à l'Europe entière. On a dit du *TLF* que c'était un *Concorde* littéraire, en entendant par là une entreprise de prestige, onéreuse et peu rentable. On peut espérer que la « Bibliothèque Numérique Européenne » sera plutôt un *Airbus* littéraire.

IV
THIEF ou l'exploitation statistique de *Frantext*

En réalité, du *TLF* à la *BNF* il y a une relation d'héritage, comme entre *Concorde* et *Airbus*. Les 1500 volumes en mode texte que *Gallica* compte à son catalogue viennent pour la plupart des données de *Frantext*, dont la saisie manuelle remonte aux années 70. La fidélité électronique de la lecture optique et des scanners modernes dont se prévaut Google ne semble pas devoir l'emporter sur celle des dactylos de cette époque reculée. Il est vrai que les textes disponibles sur les sites de *Frantext* et de *Gallica* ont bénéficié d'un toilettage ultérieur, dont la nécessité a été démontrée dans l'expérience de la revue *Europe*, entreprise et commentée, ici même, par Henri Béhar. Souhaitons que le défi de *Google-Print* serve à orienter la *BNF* vers le mode texte plutôt que le mode image. Les textes en mode image sont certes plus

---

[3] La réponse de Jean-Noël Jeanneney, au nom de la BNF, a été publiée dans un livre en avril 2005 sous le titre « Quand Google défie l'Europe » (éditions « Mille et une nuits, coll. « Essai »).

[4] « Les bibliothèques numériques », in Culture et recherche, n° 100, janvier-mars 2004, p.6.

[5] Le même rapport revendique 140 000 pages visitées chaque jour en 2002 sur le site Gallica.

faciles à préparer et plus fidèles à l'orignal, mais ils sont aussi plus encombrants, leur stockage plus onéreux et leur transmission plus lourde. Et surtout ils sont rebelles à l'indexation et leur contenu échappe aux moteurs de recherche. Sans doute peut-on proposer le déchiffrage immédiat du document transmis, mais la portée de cette lecture et donc de la recherche restera limitée à ce seul document, sauf à imaginer un système double où chaque document est visible sous les deux formes, comme dans le DVD *Europe*.

En attendant la réalisation de cette Bibliothèque Européenne Numérique (qu'on appellera peut-être la *Big BEN*), la France dispose depuis plus de trente ans d'un gisement de textes exceptionnel, accessible depuis près de dix ans sur le réseau. On pourrait dire de *Frantext* ce qu'on a pu dire du *TLF* : un trésor est caché dedans. Trésor *Frantext* l'est par la quantité, la qualité et l'homogénéité des 3000 textes traités à ce jour. Mais pour beaucoup d'internautes, le trésor reste enfoui sous terre, à cause de la nécessité de l'abonnement, imposé par le copyright féroce des éditeurs. Cet obstacle juridique - qui n'a pas de justification économique, l'abonnement étant d'un rapport faible – décourage la consultation. Un autre obstable, technique celui-là, en empêche l'exploitation statistique. Certes la plupart des consultations sont d'ordre purement documentaire et les fonctions de Frantext répondent parfaitement à ceux qui cherchent des exemples. Mais quand une base a cette taille, répartie sur cinq siècles et recouvrant des milliers de textes, des centaines d'écrivains et plus de 200 millions de mots, la tentation est grande d'en extraire des enseignements sur l'évolution de la langue, la spécificité des écrivains ou la typologie des genres littéraires. Et de telles questions suscitent l'outil statistique.

Or *Frantext* est riche en fonctions statistiques :

1 - Pour un corpus choisi il peut fournir la liste intégrale du vocabulaire avec l'indication de fréquence pour chacune des formes.

2 - Il permet de constituer à volonté des listes de mots et d'en extraire les fréquences dans les textes que l'on veut.

3 - Pour un mot donné, ou pour une liste de mots 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 ou les genres).

4 – Pour un mot choisi (ou une liste de mots), il peut explorer l'environnement lexical, et constituer le catalogue des mots qui entourent le mot-pôle.

Et pour ces fonctions puissantes, il laisse à l'utilisateur le choix des mots et le choix du corpus (selon quatre critères qu'on peut croiser : textes, auteurs, genres, et dates).

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) fonctionne en direct ou non, selon qu'on se contente des données de Frantext déjà enregistrées (selon une optique chronologique) ou qu'on s'adresse à Frantext pour une sélection particulière. Comme la place nous manque pour détailler les fonctions disponibles, nous nous

limiterons à un quelques exemples d'exploitation, en laissant au lecteur la liberté d'imaginer ce qu'on obtient avec le menu principal représenté ci-dessous [6]:

Figure 3. Le menu du logiciel THIEF
(Tools for Helping Interrogation and Exploitation of Frantext)

Pour garder à notre propos un peu d'unité, au moins l'unité de lieu, et retrouver la thématique du début de notre exposé, choisissons pour exemple des noms de lieu, en l'occurrence les toponymes 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 et de 31 colonnes (on a choisi 31 écrivains du XVIIIe au XXe siècle et proposé la liste à Frantext pour chacun de ces écrivains). L'analyse factorielle que reproduit la figure 4 et que THIEF a réalisée donne une image de cette géographie mentale que projette l'usage des écrivains.

6

Introduction

Les fonctions documentaires de Frantext
THIEF. Version Windows.
THIEF. Version Mac.
La statistique lexicale

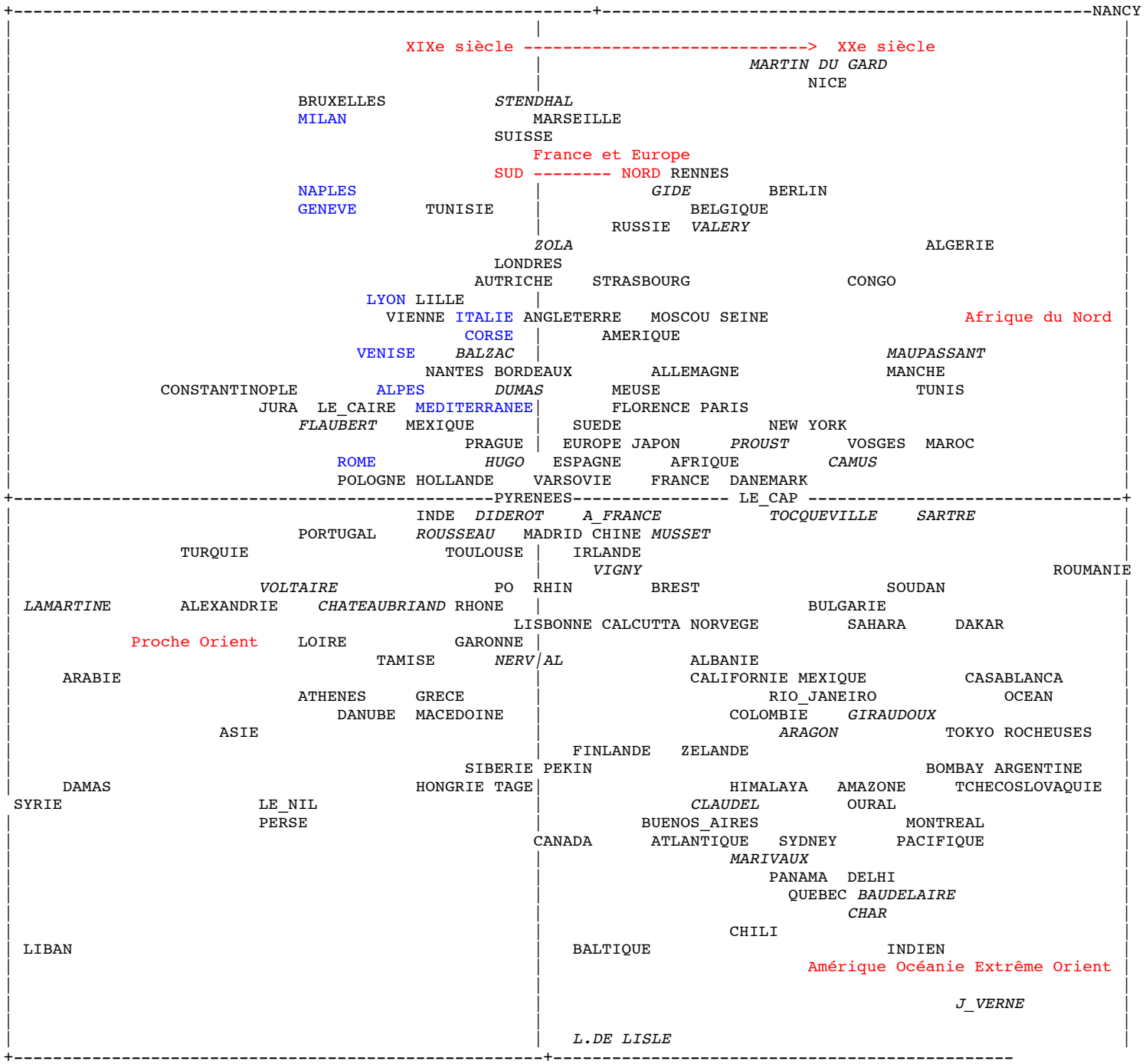
Première partie. THIEF off line.
La consultation de la base locale.

L'exploitation statistique de la base locale.
Le dictionnaire des fréquences.
Courbes et histogrammes
Les listes de mots
L'analyse factorielle
Le vocabulaire spécifique
L'évolution du vocabulaire
La structure lexicale
La distance lexicale

Deuxième partie. THIEF on line.
La consultation directe de FRANTEXT

Installation
L'évolution d'un mot ou d'une liste de mots
La répartition selon les auteurs ou les textes
Tableaux à deux dimensions
Le vocabulaire spécifique d'un texte
Les corrélats spécifiques d'un mot ou d'une liste de mots
Analyse factorielle

Figure 4 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.

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: 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 a là l'Amérique: *Californie, Mexique, Canada, Québec, Colombie, Rocheuses, Pacifique, Argentine, Chili, Buenos Aires, Rio de Janeiro, Montréal, Panama, Atlantique, Pacifique, Amazone*. C'est ici qu'on rencontre les pays les plus reculés de l'Asie, de l'Inde, de l'Extrême-Orient et de l'Océanie (*Delhi, Bombay, Himalaya, 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.

Étienne BRUNET
Université de Nice-Sophia Antipolis