L'exploitation des données de l'Institut National de la langue française

Étienne Brunet

L'exploitation des données de l'Institut National de la langue française

Étienne Brunet
(CNRS — Institut National de la langue française - Nice)

Les grands travaux publics se passent de publicité, étant eux-mêmes les meilleurs supports publicitaires. Nul besoin d'apposer des banderoles au sommet de la pyramide du Louvre, ou entre les deux pans de l'Arche de la Défense ou entre les quatre tours de la future Bibliothèque mondiale. Nul besoin non plus, vingt ans après le démarrage de l'entreprise, de parler du *Trésor de la langue française*. Les dernières lettres de l'alphabet se pressent à la sortie des presses et ce grand pro-gramme ne mérite plus son préfixe, même si, entre le *projet et le rejet,* il reste encore un large espace, même si les lourds bâtiments de l'espèce du *TLF* courent longtemps sur leur erre, comme les grands châtaigners qui mettent cent ans à grandir, cent ans à produire et cent ans à mourir. Ce n'est donc ni le lieu, ni le temps de parler d'une entreprise si célèbre et si peu nouvelle. Au reste, n'ayant en rien participé à la rédaction du *TLF,* je serais moins que quiconque qualifié pour le faire.

Mais l'Institut National de la langue française ne se confond pas avec le *TLF,* et Bernard Quemada, son directeur, ayant été cet été invité à prendre en charge les destinées du Conseil Supérieur de la langue française, on voit que de nouvelles perspectives se font jour et qu'au bout d'un quart de siècle, le *TLF,* plus rapide que l'Académie et plus loyal que Pénélope, a mis le dernier point à sa tapisserie. Il appartient à Bernard Quemada de faire le bilan du passé et de dresser les perspectives de l'avenir. Mais entre le passé, pour moi trop antérieur, et le futur, pas assez simple, il y a un présent, qui m'est accessible et qui n'a peut-être pas été suffisamment présenté à la communauté scientifique.

Il porte le nom de FRANTEXT, et derrière cette appellation de consonance anglaise se cache une réalité bien française: l'ensemble des textes qui ont été enregistrés à Nancy depuis vingt cinq ans, pour fournir des exemples aux rédacteurs du dictionnaire. Chacun sait en effet que le *Trésor* a été fondé sur une table rase, préalablement débarrassée de tous les dictionnaires précédents. C'était se couper ainsi de la source traditionnelle et cumulative où des générations de rédacteurs avaient puisé jusqu'ici, chacun pillant le voisin sachant qu'il serait pillé à son tour. Même si les dictionnaires expulsés s'étaient réfugiés sous la table, la

règle fut établie de n'emprunter les exemples qu'aux bons auteurs, avec des citations de première main. Et c'est ainsi que les grands écrivains de notre littérature furent cités à comparaître devant l'ordinateur, qui, en dépouillant des milliers de textes, devint le grand magasinier du stock lexical.

I. Les données.

1. Le *TLF* portant sur la langue du XIXe et du XXe siècles, les textes d'abord retenus n'étaient pas antérieurs à 1789 ni postérieurs à 1964, date à laquelle la liste fut fixée. Cela représentait toutefois une masse considérable de plus de 70 millions d'occurrences. À l'heure actuelle cette mine est toujours en exploitation, alors même que le principal client, le *TLF,* a cessé d'exprimer des besoins et de passer des commandes. C'est que d'autres clients sont apparus, qui ont souvent des exigences spécifiques. Ainsi le Dictionnaire du Moyen français a été mis en chantier avec Robert Martin pour maître d'oeuvre et une entreprise semblable est envisagée pour la langue classique. Le choix des textes disponibles s'est donc étendu au XVIIIe, au XVIIe et maintenant au XVIe siècle. Mais en même temps à l'autre bout de la chronologie, le chantier se prolonge pareillement. Depuis 1964, la littérature a coulé sous le pont Mirabeau et, sous peine d'apparaître comme un monument antique, la base de textes est tenue à une mise à jour permanente qui laisse leur place aux célébrités que le présent consacre. Il faut dire que la technologie offre des opportunités nouvelles à l'enregistrement des textes: d'une part la lecture optique permet une saisie automatique, qui n'engendre pas plus de fautes que la frappe manuelle[1]. D'autre part les éditeurs peuvent consentir à communiquer leurs bandes de photocomposition, ce qui est en principe une formule plus sûre et plus légale. Le monde de l'édition, longtemps jaloux de ses procédures, a fini par s'ouvrir à l'ordinateur et le déchiffrage des codes typographiques est devenu moins épineux[2]. Ajoutons que l'échange de données entre centres de recherche pourrait être d'un rendement avantageux, si l'INaLF avait la même aptitude à recevoir qu'à donner. Mais on s'engage plus à recevoir qu'à donner et l'INaLF n'accepte qu'avec circonspection les apports extérieurs, qu'ils viennent des chercheurs, des éditeurs ou des machines. Car il y a moins d'avantages à ajouter une pierre à un édifice déjà si imposant que d'inconvénients à accepter des matériaux impurs qui risqueraient de déparer ou dépareiller l'ensemble.

2. À l'heure actuelle la base de données FRANTEXT contient plus de 3000 textes complets, soit 180 millions de mots. Il n'est pas d'exemple équivalent au monde dans le domaine linguistique et littéraire. Mais cet avantage de l'étendue serait de peu de valeur si la qualité des données ne répondait pas à leur quantité. Il n'est pas très difficile d'amonceler un

immense terril lexical en entassant les matériaux qui ont servi à telle ou telle recherche et que l'utilisateur a abandonnés après usage. Les normes de saisie étant disparates et inconstantes, on ne peut guère construire ainsi qu'une tour de Babel, à laquelle on pourra donner au mieux le nom d'archives. Une véritable base de données requiert une sélection et une standardisation plus rigoureuses.

Ces données sont issues d'une saisie ancienne, exécutée il y a plus de vingt ans avec des machines Friden qu'on ne trouve plus que dans les musées. Et l'on pourrait songer à invoquer les faiblesses de la technologie de l'époque. Mais on aurait tort. Le code Friden qui fut adopté alors respectait l'essentiel du contenu textuel, et en particulier la distinction des majuscules et des minuscules et l'intégralité de l'accentuation française. Cela semble aller de soi à notre époque où les micro-ordinateurs se sont substitués aux machines à écrire. Mais en 1960 l'informatique n'offrait guère sur les claviers que les 26 lettres majuscules de l'alphabet anglais et il a fallu l'intransigeance des fondateurs du *Trésor* pour obliger les constructeurs à se plier aux contraintes de l'alphabet national[3]. Cette saisie initiale a été soigneusement contrôlée, mais comme il arrive aux meilleures éditions quelques fautes subsistent, en faible nombre.

S'y ajoutent quelques erreurs engendrées par la **transmission** de ces données initiales. C'est qu'en près de trente ans l'informatique a constamment évolué et les matériels ont été plusieurs fois renouvelés. On est passé d'un constructeur à un autre, d'un système à un autre et les codages ont changé en même temps que les supports de l'information. Les bandes originales de papier perforé ont été transposées en fichiers magnétiques, sur bandes ou sur disques, et plus récemment sur disque optique. Tous ces transferts forment une longue chaîne et il est arrivé qu'un des maillons de la chaîne ait eu une défaillance et qu'un caractère ait été oublié. Ceux qui ont interrogé longuement FRANTEXT ont peut-être surpris, ici ou là, une erreur, due à l'homme ou à la machine. Quoique les fautes soient peu nombreuses et qu'elles passent généralement inaperçues, une campagne de correction systématique est actuellement menée qui dans la prochaine version de FRANTEXT fournira un texte plus pur.

Certains esprits rigoureux se demanderont s'il était opportun d'inviter le public à l'ouverture de FRANTEXT, alors que le ménage n'était pas tout à fait achevé. Rappelons ici que le *TLF* n'avait pas vocation à l'origine à devenir serveur de bases de données (le mot n'existait pas plus que la chose) et il aurait été fondé à s'en tenir à son objectif premier, la rédaction du dictionnaire, pour lequel les imperfections qu'on vient de dire étaient sans importance, les rédacteurs ayant toute facilité de redresser les erreurs. Mais c'eût été prolonger trop longtemps la rétention de données que la communauté scientifique réclamait à cor et à cri, alors que la technique permettait leur **diffusion large,** souple, et facile. Les responsables de

l'Institut National de la langue française n'ont pas cru qu'il fallait par purisme exagéré refuser l'entrée du temple aux fidèles qui se pressaient derrière les grilles de Nancy et dont on pouvait espérer l'indulgence durant cette phase de transition. Ils avaient aussi le souci de ne point trop se laisser distancer par les avancées technologiques. Car si la réalisation d'une base de données suppose des données exactes et copieuses, elle exige aussi une grande rigueur dans la fabrication des logiciels à mettre en oeuvre et il convenait dans ce domaine d'acquérir l'expérience et de parvenir à la puissance, à la fiabilité, à la rapidité et à la convivialité nécessaires. Et ces qualités ne peuvent être trouvées et prouvées que par des essais en grandeur réelle. Or le talent de l'ingénieur à qui on doit le logiciel STELLA a fait de ce coup d'essai un coup de maître. Les spécialistes de la programmation s'accordent sur ce point avec les utilisateurs.

3. Mais avant d'en vanter les mérites et d'en détailler le mode d'emploi, il faut souligner que les données de Nancy ont remarquablement résisté, sinon à l'évolution de la technique, du moins aux changements de la volonté humaine. Il était tentant de profiter de l'expérience acquise, et des nouvelles possibilités offertes par le progrès de la technologie, pour corriger les objectifs et les méthodes de saisie. Or les consignes ont été maintenues à travers le temps sans modification majeure. Le fruit de cette constance héroïque est la cohérence et l'homogénéité des données. Et cette qualité est sans doute aussi essentielle que leur étendue, et certainement plus rare. Et ce qui est unique au monde, c'est la conjonction de ces deux objectifs habituellement inconciliables. Les normes dans la définition du mot et particulièrement des mots composés, dans le repérage des noms propres, dans la lemmatisation et la catégorisation grammaticale, ou dans le traitement des signes de ponctuation, n'ont pas varié en vingt ans quoiqu'elles aient fait l'objet de critiques, parfois justifiées. Et même le choix délibérément littéraire des textes retenus n'a guère été remis en question. Cela donne à l'ensemble une base solide pour établir les comparaisons d'un texte à l'autre ou d'un corpus à l'autre et pour définir l'usage littéraire de la langue.

II. FRANTEXT.

Si le créateur de STELLA, Jacques Dendien, n'est pas en ces lieux, du moins s'y trouve Madame Martin à qui la tâche incombe de promouvoir FRANTEXT et d'organiser des stages (l'initiation ou de perfectionnement dans les bibliothèques de France, de Navarre et de Galicie. Elle est mieux placée que moi pour parler de l'audience actuelle de cette base de données textuelles, évaluer le nombre et le type des questions posées, apprécier l'intérêt, la satisfaction ou les regrets exprimés par les utilisateurs,

mesurer les tendances d'un marché nouveau et montrer sur quels points des améliorations ou des compléments pourraient être cherchés. Elle peut apporter toutes les précisions utiles sur les modalités d'accès, les lieux d'interrogation, les coûts de la consultation. Je me bornerai à indiquer à grands traits les principes de fonctionnement de cette base, en soulignant d'emblée qu'il serait difficile d'en trouver une semblable en matière linguistique, qui allie pareillement la disponibilité, la puissance, la rapidité et la souplesse.

1. L'utilisateur est invité d'abord à choisir un corpus. Ce peut être d'ailleurs le corpus tout entier, si la question posée est suffisamment précise. Assez souvent on s'intéresse à un auteur, ou à un genre littéraire ou à une période particulière. Rien n'est plus aisé que de circonscrire un territoire de recherche en utilisant la commande BIBLIO et en introduisant (en les conjuguant au besoin) les critères de sélection qui peuvent être relatifs:

— à l'auteur (par exemple *a = Hugo).*
— au genre ou au domaine (par exemple *g = roman).*
— à la date (par exemple *d = 1815-1848).*
— au titre ou à une partie du titre (par exemple *t = guerre)*[4].

Ainsi la commande *a = Hugo g = poésie d = 1800-1850* permettra de choisir les premiers recueils poétiques de Hugo jusqu'aux *Contemplations* et de constituer une liste de textes qu'on pourra cataloguer sous un nom générique, qu'on pourra trier selon la chronologie (normale ou inverse) ou selon le classement alphabétique, et qu'on pourra rapprocher d'une autre liste pour cumuler les deux séries ou procéder à leur intersection. (La combinaison des corpus s'opère par la commande LISTEREFS). On peut ainsi se constituer plusieurs corpus virtuels, parmi lesquels on choisit au dernier moment celui qu'on veut explorer et qui devient le corpus courant. On doit alors préciser ce choix par la commande DEFCORPUS. Une variante permet de sélectionner les textes non point en vertu de critères bibliographiques mais à partir du contenu même. Pour être sélectionné un texte doit alors contenir au moins une fois tel ou tel mot que l'on indique ou l'un ou l'autre des constituants d'une liste de mots préétablie. Cette commande MOTCORPUS est souvent la meilleure approche quand on s'intéresse à un thème, à un mot, ou à une tournure et que l'on veut recueillir aux moindres frais le maximum de documentation.

2. Dans la seconde phase de la consultation, l'utilisateur choisit le traitement, qui peut le conduire à un index, à une sélection de contextes ou à des indications de fréquences. Dans le cas le plus simple, on souhaite relever le contexte d'une forme. En déclenchant la commande CHERCHE et en répondant au dialogue qui s'instaure, on obtient le détail, avec un contexte de la longueur choisie, de tous les emplois de la forme considérée dans le corpus

sélectionné. Par exemple une demande portant sur le lieu privilégié qui accueille le présent congrès a restitué les 69 passages où le mot ***Compostelle*** a été évoqué. On en donnera une illustration dans le tableau 1 où Madame d'Aulnoy (exemple 2) accorde à ce toponyme l'explication étymologique qui eut longtemps cours: ***Campo Stella,*** tandis que l'ironie de Voltaire perce dans les derniers exemples. Si on élargit l'enquête à la province toute entière, on aura intérêt à réunir dans une même liste (grâce à la commande LISTE-MOTS) les mots ***Galice, Galicie, Compostela,*** et tous ceux qu'on voudra, et à solliciter la commande INDEX, qui est moins gourmande en temps et en espace, puisqu'elle se contente de donner les références, et qui convient mieux lorsque les résultats sont trop abondants ou lorsqu'on n'a pas besoin du contexte explicite. Le tableau 2 illustre cette façon d'interroger la base, en montrant les deux voies qui s'ouvrent en de tels cas: ou bien on sollicite la base entière en sollicitant la commande BIBLIO et en choisissant un critère extensif (par exemple une datation comprise entre 1000 et 2000), ce qui alourdit considérablement la recherche; ou bien une sélection préalable opérée par la commande MOTCORPUS permet d'obtenir les mêmes résultats de façon plus rapide et moins onéreuse.

3. Le logiciel STELLA offre des possibilités sophistiquées pour la sélection des mots et la recherche des contextes.

— Les listes de mots peuvent être soumises, comme les listes de textes, aux opérateurs booléens et être combinées pour produire union, intersection ou différence. Elles peuvent aussi être générées automatiquement à partir d'un critère de filtrage (dans la commande MOTS), grâce auquel on peut repérer toutes les formes qui commencent ou finissent par telle ou telle chaîne de caractères, ou toutes celles qui contiennent telle suite de lettres.

— La commande CHERCHE possède toute la puissance et la généralité qu'on peut souhaiter. Elle est apte à sélectionner les cooccurrences, orientées ou non, et les expressions, figées ou non. Elle tient compte des critères de filtrage, des listes de mots préétablies et par exemple, si l'on souhaite retrouver toutes les variantes de l'expression ***gagn(er) son*** (ou ***mon, ton...) pain,*** peut être formulée comme suit:

```
$ & pain
```

le signe $ suscitant un dialogue auquel on pourra répondre par: ***gagn%,*** si l'on veut toutes les formes verbales du mot ***gagner,*** et le signe & demandant qu'on précise la liste des mots qui peuvent se trouver en seconde position dans l'expression et qu'on aura préalablement remplie des formes du possessif.

Tableau 1: Consultation de FRANTEXT
Commandes MOTCORPUS, DEFCORPUS et CHERCHE

Commande: --> **motcorpus**

Bibliographie d' un mot ou d' une liste de mots DANS LA TOTALITÉ DES TEXTES.
Vous pouvez faire une étude bibliographique
1: pour un seul mot
2: pour une liste de mots
Indiquez votre choix: —> **1**

Donnez votre mot: —> ***COMPOSTELLE**

Traitement du mot —> *COMPOSTELLE
Nom du fichier qui contiendra la liste bibliographique: —> COMPOSTELLE

***Debut de la recherche
***État de la recherche: un quart du corpus
***État de la recherche: trois quarts du corpus
***Fin de la recherche
Nombre de références bibliographiques

selectionnées: **38**

Commande: —> **defcorpus**

Définition du corpus courant
donnez le nom de la -liste des références: —> **COMPOSTELLE**

Le corpus de travail comprend:
38 références bibliographiques
2440600 mots

Commande: —> **cherche**
Le corpus de travail comprend:
38 références bibliographiques
2440600 mots
composant 1: --> ***COMPOSTELLE**
composant 2: —> *
Entrée des positions des composants
? --> *
Veuillez faire un des choix indiqués
Vous pouvez travailler:
1: à l' intérieur d' une phrase
2: à l' intérieur d' un contexte de largeur donnée
Indiquez quel est votre choix: —> **1**
Nombre de phrases avant: —> 0
Nombre de phrases après: —> 0

Ce n' est pas que la ville de *Saint-*Jacques *De *COMPOSTELLE ne soit considérable;
page 111/Q437/AULNOY. M.C.D'/RELATION DU VOYAGE D'ESPAGNE/1691 LETTRE 4

Un homme de ma connaissance, grand chercheur d'étymologies, assurait que la ville de *COMPOSTELLE se nommait ainsi, parce que *Saint *Jacques devait souffrir le martyre dans le lieu où il verrait paraître une étoile à *Campo *Stella.
page 113/Q437/AULNOY. M.C.D'/RELATION DU VOYAGE D'ESPAGNE/1691 LETTRE 4

Il est vrai, reprit-il, que quelques gens le prétendent ainsi, mais le zèle et la crédulité du peuple vont bien plus loin, et l'on montre à *Padron, proche de *COMPOSTELLE, une pierre creuse, et l'on prétend que c'était le petit bateau dans lequel *Saint *Jacques arriva après avoir passé dedans tant de mers, où, sans un continuel miracle, la pierre aurait bien dû aller au fond.
page 113/Q437/AULNOY. M.C.D'/RELATION D11 VOYAGE D'ESPAGNE/1691 LETTRE 4

Les terres, pour la plupart, y demeurent incultes, et, du côté de *Saint-*Jacques-*De-*COMPOSTELLE, il semble que ce soit un désert;
page 115/Q437/AULNOY. M.C. D'/RELATION DU VOYAGE D'ESPAGNE/1691 LETTRE 4.

Il se seroit étendu davantage sur des louanges qui la réjouissoient fort, si on ne l'eût avertie que l'archevêque de *COMPOSTELLE venoit d'arriver et qu' il étoit déjà dans son appartement.
page 517/Q3984ULNOY. M.C. D'/FINETTE CENDRON/1698.

Elle venoit les querir pour chanter devant l'archevêque de *COMPOSTELLE, mais ils le connoissoient trop pour hasarder de paroître devant lui;
page 520/Q398/AULNOY. M.C. D'/FINETTE CENDRON/1698.

Ils restèrent dans ce lieu jusqu' à ce qu' ils eussent entendu passer l'archevêque, qui retournoit à *COMPOSTELLE.
page 520-521/Q398/AULNOY. M.C. D'/FINETTE CENDRON/1698.

"Mangez, buvez, dormez, lui dit-elle, et que notre dame d'*Atocha, monseigneur st *Antoine *De *Padoue, et monseigneur st *Jaques *De *COMPOSTELLE prennent soin de vous.
page 43/N577/VOLTAIRE/CANDIDE OU L'OPTIMISME/1759 CHAPITRE 7.

— Par st *Jaques *De *COMPOSTELLE, dit *Cacambo vous alliez faire la guerre aux jésuites;
page 79/N577/VOLTAIRE/CANDIDE OU L'OPTIMISME/1759 CHAPITRE 14.
ETC... Nombre d'exemples sélectionnés: **69**

Tableau 2: Commandes LISTEMOTS, BIBLIO et INDEX

Commande: —> **listemots**

PROGRAMME DE TRAITEMENT DE LISTES DE MOTS
Si vous ne savez que repondre a la question ACTION, tapez ? pour obtenir
des informations. Le nom de la liste courante n'est pas defini
ACTION: —> créé
CRÉATION D'UNE LISTE. Nom de la liste a créer: GALICE

Chaque fois que le programme le demande, entrez un mot ou entrez le caractère * pour terminer. Il est indifferent de taper des majuscules ou des minuscules
mot 1: --> *COMPOSTELLE
mot 2: --> *COMPOSTELA
mot 3: --> *GALICIE
mot 4: —> *GALICE
mot 8: --> *
Le nom de la liste courante est: GALICE.mots

Commande: --> **biblio**

Recherche bibliographique
Entrez votre expression de sélection
—> d = 1000-2000
nombre d'oeuvres sélectionnées: 2330
Vous pouvez:
1: visualiser la liste des oeuvres sélectionnées.
2: reprendre une recherche avec d'autres critères.
3: créer une liste de références.
4: arrêter la recherche
tapez le numéro de votre choix: —> 3
Donnez le nom de la liste: —> INTÉGRAL

Commande: —> **motcorpus**

Bibliographie d'un mot ou d'une liste de mots DANS LA TOTALITÉ DES TEXTES
Vous pouvez faire une étude bibliographique
1: pour un seul mot
2: pour une liste de mots
Indiquez votre choix: —> 2
Donnez le nom de la liste de mots: —> GALICE
Voulez-vous sélectionner les références bibliographiques:
1: contenant au moins un des mots de la liste
2: contenant simultanément tous les mots
Indiquez votre choix: --> 1
Nom du fichier qui contiendra la liste bibliographique: —> GALICE

Nombre de références bibliographiques sélectionnées: 99

Commande:-***index**
PROGRAMME D'INDEX
nom de la liste de mots: —> GALICE
Étude du mot: *COMPOSTELA
399/CARON-HUTIN/LES ALCHIMISTES
1 occurrence(s). Page: 10
278/TSERSTEVENS.A/L'ITINÉRAIRE ESPAGNOL
3 occurrence(s). Pages: 257 269 274
Nombre total d'occurrences dans le corpus: 4
Étude du mot *COMPOSTELLE
399/CARON-HUTIN/LES ALCHIMISTES
2 occurrence(s). Pages: 9 17
328/BRUNHES.J/LA GÉOGRAPHIE HUMAINE
1 occurrence(s). Page: 284
463/LAMBERTIE.R-M/INDUSTRIE PIERRE ET MARBRE
ETC... Nombre total d'occurrences dans le corpus de travail: 69

Étude du mot: *GALICE
284/VIDAI, DE LA BLACHE.P/TABLEAU GÉOGRAPHIE FRANCE 1
2 occurrence(s). Pages: 20 20
589/HADDON.A-C/LES RACES HUMAINES...
1 occurrence(s). Pate: 116
642/FAURÉ.E/HISTOIRE ART. L'ART MÉDIÉVAL
1 occurrence(s). Page: 310
274/CENDRARS.B/BOURLINGUER
2 occurrence(s). Pages: 33 34
563/CLAUDELP/SOULIER DE SATIN, VERSION SCÈNE
1 occurrence(s). Page: 966
273/CAMUS.A/REVOLTE DANS LES ASTURIES
1 occurrence(s). Page: 425
ETC... Nombre total d' occurrences dans le corpus de travail: 75

Étude du mot *GALICE
354 // ARTS ET LITT. SOCIÉTÉ CONTEMP.
1 occurrence(s) . Page: 5401
342/LEFEBVRE.G/LA RÉVOLUTION FRANÇAISE
3 occurrence(s). Pages: 93 222 224
343/LEFEBVRE.G/LA RÉVOLUTION FRANÇAISE
ETC... Nombre total d'occurrences dans le corpus de travail: 53

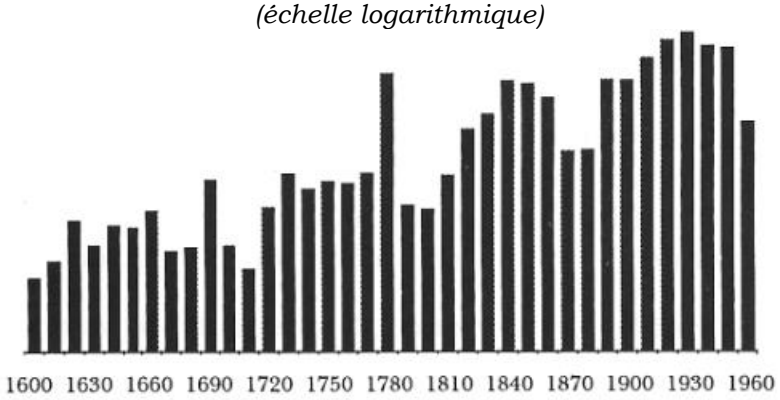
Commande: —,*
FIN D'EXÉCUTION DE STELLA
r 22:46 35.801 4555

logout
Brunet.URL9 déconnecté le 09/01/89 2246.4 fst Fri
cpu: 41 sec, mémoire: 3841.7 unités, coût: 16.36F.

— Bien d'autres commandes sont accessibles dont l'explication serait trop longue et qui servent à repérer des suites répétitives qui comportent un mot considéré comme pôle (commande SUITE) ou à extraire les contextes qui contiennent un mot pôle dans un environnement lexical variable qu'on peut affiner en précisant la position et la nature de chacun des éléments de cet environnement (commande ..ÉTUDE). On dispose ainsi d'un substitut d'analyse syntaxique qui dans beaucoup de cas permet de dissoudre les ambigüités et de repérer les tournures complexes.

— Enfin si l'on prétend se diriger vers la statistique lexicale, des outils sont proposés qui mesurent la fréquence des mots d'une liste dans un corpus défini (commande FRÉQUENCES). Comme la statistique est essentiellement comparative, nous avons ajouté au logiciel la commande SOUSFREQ qui permet de construire un tableau à deux dimensions où pour une même liste de mots (ce sont les lignes du tableau) on dispose des fréquences dans les différents textes ou corpus confrontés (ce sont les colonnes). Nous choisirons un exemple inattendu, qui concerne un nom propre (la statistique s'exerce habituellement sur le vocabulaire commun): le mot *Jacques* qui peut désigner le saint patronyme, ou plus souvent un lieu comme celui où ce colloque se déroule, ou bien le nom ou le prénom d'un personnage de l'histoire ou d'un héros de roman. Tant d'ambiguïtés empêchent de parvenir à des conclusions fines. La courbe obtenue à partir des 13716 occurrences relevées semble indiquer pourtant un succès grandissant de ce patronyme.

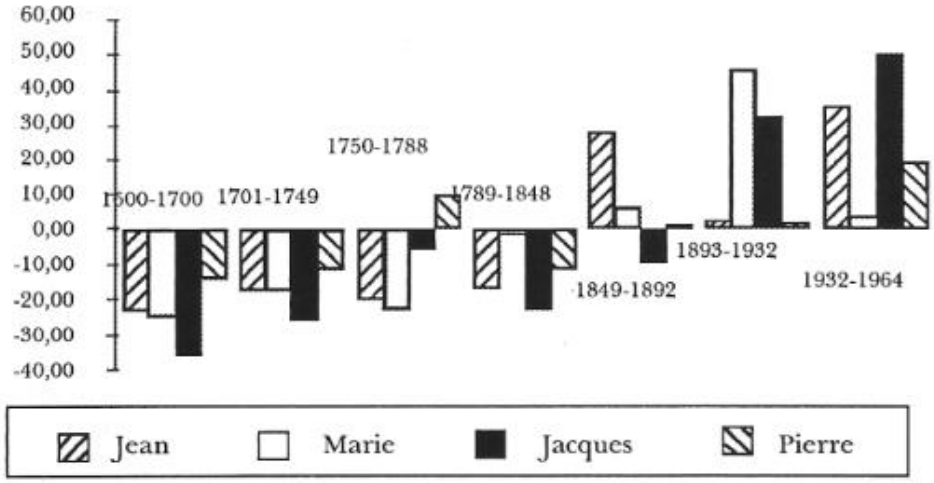
Graphique 3: "JACQUES" de 1600 à 1970 par tranches de 10 ans (échelle logarithmique)



Mais il faut prendre garde que la progression concerne l'ensemble des prénoms du calendrier et que les tranches de 10 ans qu'on a découpées dans la chronologie n.'ont pas la même étendue (en nombre de textes). Il faut donc pondérer, ce qui est fait dans la courbe 4 qui propose une comparaison avec d'autres prénoms populaires. *Jacques,* qui occupe le troisième rang derrière *Jean* (18370 occurrences) et *Louis* (15069), montre

une progression régulière qui lui permet les plus grands espoirs, à l'inverse de certains prénoms trop aristocratiques comme *Charles* ou *Philippe* dont le déclin accompagne, au moins en France, la chute des rois.

Graphique 4: Quelques prénoms très courants (en progression)

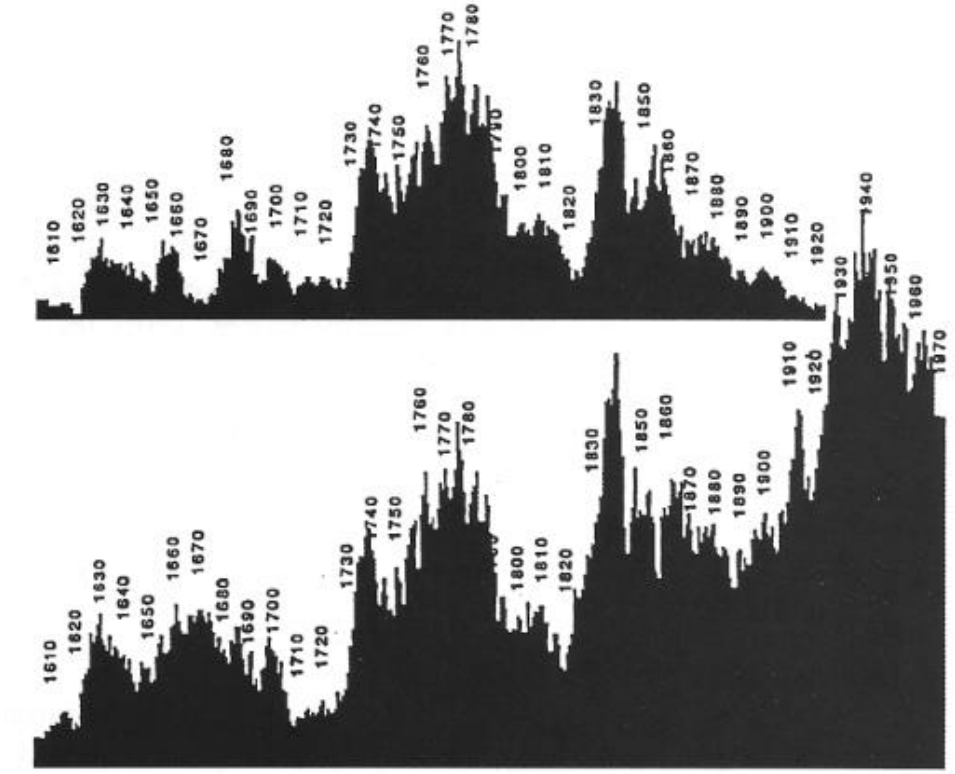


III. Les perspectives.

FRANTEXT offre à l'utilisateur d'autres opportunités qui lui permettent de choisir la présentation des résultats, leur mode de restitution (immédiate ou différée) et la possibilité à tout moment d'interrompre un traitement, de modifier les options de sortie, ou de contrôler les paramètres, les fichiers et les résultats enregistrés. Mais FRANTEXT ne peut s'affranchir des contraintes de la télématique que l'utilisateur non-spécialiste peut estimer trop lourdes. Ces contraintes matérielles, techniques, financières et juridiques sont si fortes que la consultation ne peut guère se faire que dans les bibliothèques universitaires, par le truchement d'un terminal spécialisé (un simple minitel ne suffit pas) et d'une personne qualifiée.

1. On s'emploie présentement à faire sauter le **verrou juridique** qui interdit de reproduire les textes qui ne sont pas du domaine public. Des accords ont été passés avec les éditeurs qui autorisent la consultation de FRANTEXT et même la restitution de contextes point trop larges, dont le statut peut être assimilé à celui de la citation. Naturellement aucune limite ne pèse sur les produits discrets que sont les index et les listes de fréquences. Mais s'il s'agit de contextes et à plus forte raison de textes entiers la distinction juridique doit être maintenue entre ce qui appartient au domaine public et ce qui lui échappe. Le graphique ci-dessous indique la part respective des deux lots, dont le plus important, hélas, n'est pas celui des textes libres.

Graphique 5. Les textes libres (en haut) et ceux qui ne le sont pas (en bas)



On pourrait penser en effet que tout ce qui est en dehors du XXe siècle est libre de tout droit. C'est oublier que les droits des éditeurs s'ajoutent à ceux des auteurs et comme les grands textes sont ceux qu'on réédite le plus souvent, les oeuvres de Voltaire, Molière ou Hugo, n'appartiennent pas au domaine public parce que l'édition de référence (c'est souvent la Pléiade) est trop récente pour être exempte de toute redevance. Il arrive ainsi que les textes libres sont souvent ceux qui ont cessé d'être disponibles et d'être lus et la valeur de ce stock s'en trouve diminuée. On ne pourra exploiter véritablement FRANTEXT que lorsqu'un accord global aura été négocié avec les auteurs et les éditeurs, quitte à acquitter un droit forfaitaire pour l'ensemble des consultations. Bien entendu cette redevance serait répercutée dans la contribution demandée à l'usager de la base. Des discussions sont en cours qui vont dans ce sens et des accords limités ont déjà été acquis.

2. Supposons réglé ce problème juridique et commercial. Restent les **obstacles techniques.** La télématique a des vertus, notamment pour la mise à jour des informations. Mais cet avantage a peu de poids dans le cas d'une base de données littéraires. Une fois qu'on a dépouillé (correctement) le texte le Rabelais ou de Zola, on voit mal quel séisme de l'édi-

tion imposerait une modification radicale du texte enregistré. Par contre le poids de la télématique est d'autant plus insupportable aux esprits littéraires que les connaissances technologiques leur sont assez souvent étrangères. Passe encore d'apprendre à se servir du clavier d'un terminal. Mais il faut encore s'initier à l'emploi du logiciel de communication, mettre en oeuvre un modem, maîtriser les protocoles de TRANSPAC, s'adresser au centre serveur (et donc connaître peu ou prou son système d'exploitation) et décliner son identité (et son mot de passe), et bien entendu enfin être rompu aux finesses du logiciel d'interrogation STELLA, qui rend FRANTEXT accessibles. Ce parcours d'obstacles peut effrayer l'utilisateur et même le décourager de s'adresser aux intermédiaires. Dans les faits la base FRANTEXT, qui est disponible depuis plus de deux ans, n'a pas franchi le seuil de rentabilité —comme il arrive souvent aux bases de données qui démarrent— et on peut douter qu'elle y parvienne un jour.

Car les utilisateurs —qui sont virtuellement légion— aimeraient conserver les mêmes possibilités d'interrogation des textes sans avoir à faire aucun déplacement physique ni aucune reconversion intellectuelle. La plupart consentiraient tout au plus à se servir d'un micro-ordinateur, d'autant que cet outil tend à remplacer la machine à écrire. Habitués aux ressources de l'audiovisuel, ils accepteraient encore de mettre en route un lecteur de disque d'autant que la technologie du laser leur est familière depuis que la musique utilise la lecture optique et le son laser. Or la même technologie a été adaptée à l'enregistrement des données binaires, forme sous laquelle se présentent les textes. Et le codage des textes est si économique que l'étroite surface d'un seul disque —que les mains d'un enfant suffisent à recouvrir— permet l'enregistrement de plus d'un demi-milliard de caractères, soit l'équivalent de plus de 1000 textes complets. L'évolution des techniques documentaires éloigne de plus en plus l'utilisateur des gros systèmes et la technologie du CD-ROM —c'est le nom de ce disque optique— offre une alternative intéressante à la télématique. Le marché de ce disque compact (c'est l'appellation française), longtemps hésitant, se développe actuellement aux Etats-Unis et au Japon, et, le mouvement gagnant l'Europe, plus de cinquante projets sont en chantier qui utilisent en France cette technologie, à l'exemple de celui de la Bibliothèque Nationale[6].

Or l'Institut National de la langue française s'intéresse depuis longtemps aux possibilités offertes par les mémoires optiques, en matière de capacité de stockage et de souplesse de diffusion et il prépare une version de sa base, qui autorisera un transfert sur de nouveaux systèmes et de nouveaux supports, y compris le disque optique.

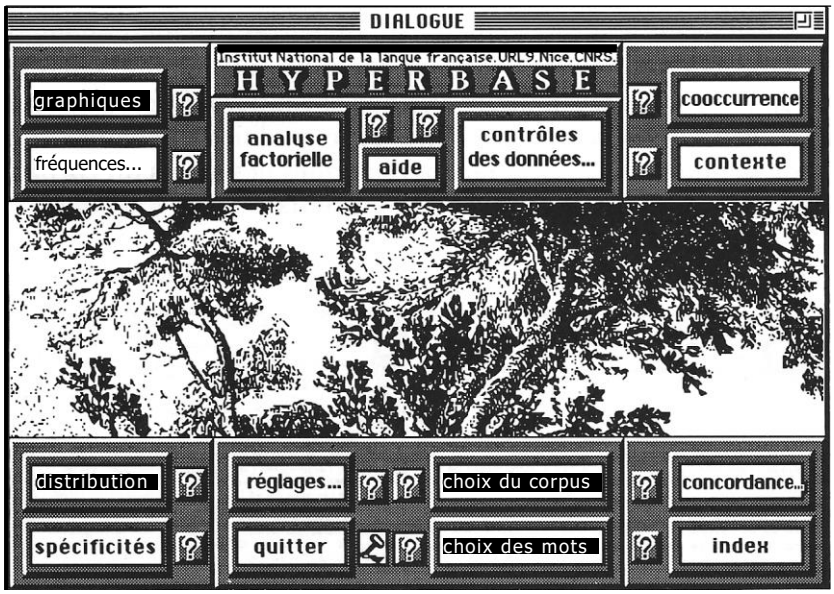
3. Pour frayer la voie à cette vaste opération, plusieurs prototypes ont été expérimentés dont l'un porte le nom d'HYPERBASE et a été mis à

la disposition du public de Beaubourg pour la consultation des textes de l'époque révolutionnaire.

Pour le mettre en oeuvre il suffit d'un Macintosh et d'un disque dur. Nulle liaison n'est nécessaire avec l'extérieur. L'accès aux données est libre, immédiat et facile. Mieux encore le texte même est disponible et les résultats, au lieu d'être figés dans les index et les concordances qu'on a publiées jusqu'ici sur papier ou microfiche, naissent sous les yeux du chercheur qui peut à tout instant les modifier, les orienter, les interrompre, les élargir, les préciser. Avec cette maîtrise retrouvée, l'utilisateur, tout en acceptant d'être aidé par un logiciel d'interrogation, a le sentiment de partir seul et le premier à la découverte, sans devoir suivre les pas d'un guide.

Le graphique 6 reproduit le menu d'accueil auquel l'utilisateur est conduit d'emblée et auquel il est ramené systématiquement dès qu'un traitement est accompli. Deux branches y sont proposées: à droite on s'engage dans la recherche documentaire; à gauche on s'oriente vers les traitements statistiques.

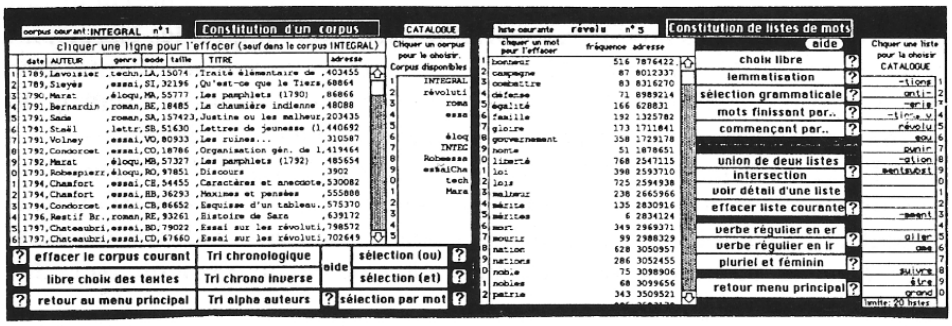
Graphique 6. Le menu d'accueil du logiciel HYPERBASE



a — Mais avant de choisir une de ces deux directions (d'ailleurs non exclusives), le chercheur est invité à fixer ses choix. Il délimite lui-même le sous-ensemble de textes qu'il veut consulter (par un programme *Choix du Corpus,* qui multiplie les critères de sélection: genres, auteur, date, titre, contenu, et leur applique les opérateurs booléens. Voir figure 7). Il établit

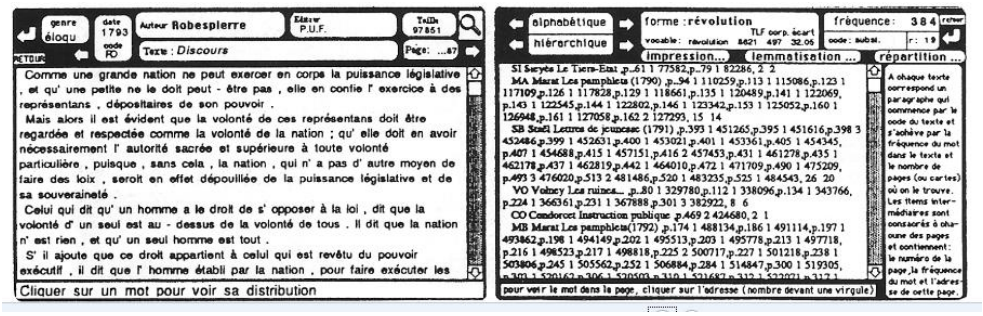
aussi librement la liste des mots qui l'intéressent (programme CHOIX DES MOTS, qui propose des listes automatiques fondées sur le suffixe, le préfixe ou la lemmatisation, et qui autorise les ajouts, les suppressions, les croisements. Voir figure 8).

Figure 7. Choix du corpus *Figure 8. Choix des mots*



b — L'utilisateur peut vouloir d'abord procéder à des opérations de contrôle et examiner les textes, page après page (graphique 9), ou consulter le fichier des mots (graphique 10) ou passer instantanément, par un simple "clic", des textes aux mots et des mots aux textes, selon les méthodes de l'hypertexte. S'il désire vérifier le contenu du corpus courant et de la liste courante —qui sont variables et se modifient à volonté— un aperçu rapide fait défiler les fiches bibliographiques des textes du corpus ou les fiches signalétiques des mots de la liste en cours.

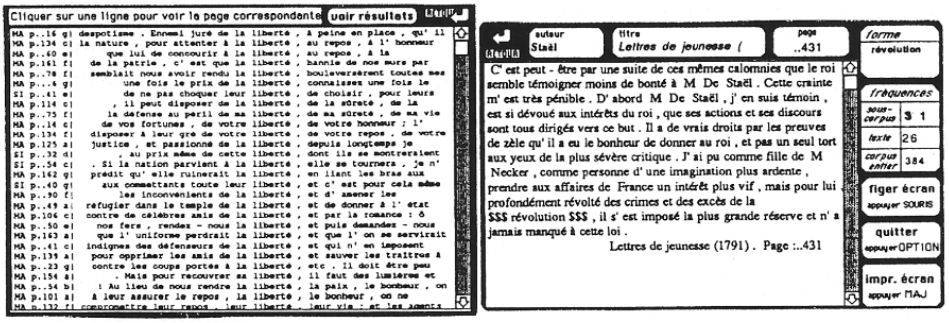
Figure 9. Une fiche-page *Figure 10. Une fiche-mot*



c — Quand le chercheur veut s'engager dans la recherche documentaire, il choisit dans les boutons de droite les actions à mener, qui peuvent conduire à un index (programme INDEX), à une concordance (programme CONCORDANCE, figure 11), à une liste de contextes-phrases (voir le programme CONTEXTE, graphique 12), ou à une

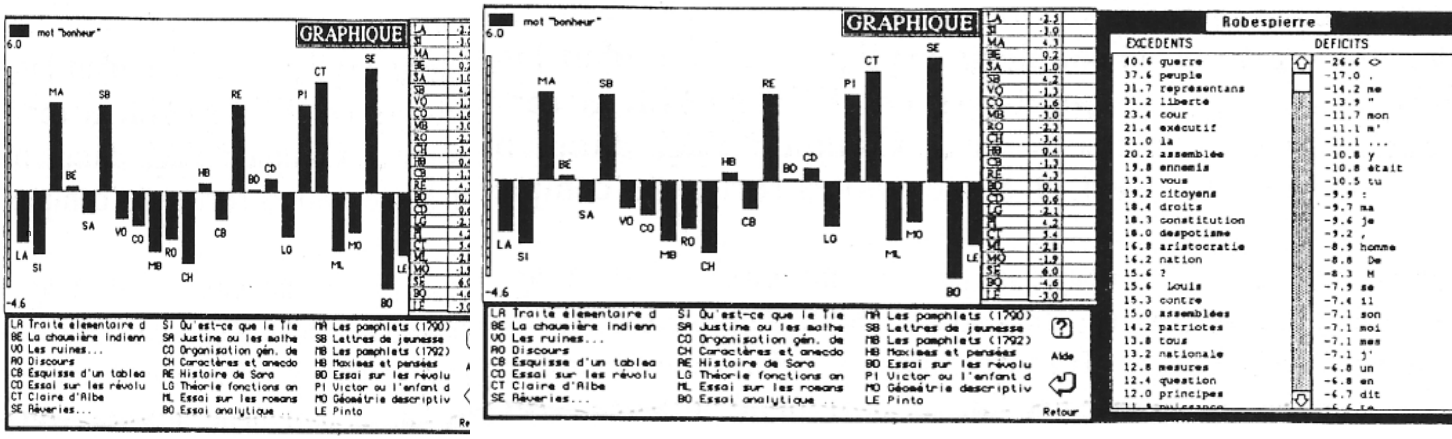
recherche de cooccurrence (programme COOCCURRENCE). Liberté lui est donnée pour chacun de ces traitements de fournir soit un mot, soit un vocable (dont les différentes formes seront automatiquement fournies), soit une chaîne de mots (une expression), soit une liste de mots.

Figure 11. Concordance *Figure 12. Contexte*



d - Le logiciel HYPERBASE mène aussi du côté des méthodes quantitatives. Il établit des tableaux de contingence (programme FRÉQUENCES), et il dresse des courbes qui comparent la distribution de deux mots dans les différents textes, ou le profil de deux textes à travers une série de mots (programme GRAPHIQUE, figure 13). On dispose d'une interface pour l'exploitation de ces tableaux par les méthodes d'analyse multi-dimensionnelle (programme ANALYSE FACTORIELLE). Enfin divers modules sont proposés qui font apparaître la distribution des classes de fréquences, dans chaque texte et dans l'ensemble (boutons DISTRIBUTION et LOI DE ZIPF) ou qui détaillent le vocabulaire spécifique d'un texte ou les propriétés spécifiques d'un mot (bouton SPÉCIFICITÉS. Voir graphique 14).

Figure 13. Graphique *Figure 14. Vocabulaire spécifique*



e — Enfin les résultats apparaissent sous deux formes qui ne sont pas toujours identiques: à l'écran, au moment même où ils sont obtenus, et dans un fichier où ils sont enregistrés, avant d'être contrôlés et imprimés. Les résultats sont soumis à des opérations de tri, et notamment les concordances qu'on offre sous diverses présentations: ordre chronologique, ordre chronologique inverse, ordre alphabétique des auteurs, tri du contexte gauche ou tri du contexte droit. Un éditeur (avec module d'impression) est fourni qui permet les retouches et la mise en page. Notons que les résultats eux-mêmes ne constituent pas un terminus, et que les méthodes de l'hypertexte leur sont applicables: en particulier il suffit de désigner une ligne de la concordance obtenue pour faire apparaître aussitôt la page correspondante du texte en question.

Précisons que le moteur de recherche mis en oeuvre tient compte des contraintes liées au CD-ROM, et en particulier de la relative lenteur des accès-disque de ce support. Pour toute exploration, il suffit de deux déplacements de la tête de lecture. Les accès eux-mêmes ont été optimisés (à partir de la fréquence présumée des appels). Mais la rapidité nécessaire n'a pas conduit aux sacrifices habituellement consentis par les produits documentaires, qui s'intéressent rarement aux mots-outils. Ces mots ont au contraire beaucoup de prix dans une exploitation de type linguistique, qui vise à l'exhaustivité. Le fichier inverse contient tous les mots, et incorpore aussi les marqueurs du récit (en particulier les signes de ponctuation).

Tout le logiciel a été écrit en gardant le souci de la convivialité. Même s'il s'adresse à un public averti, on n'a pas cru devoir lui donner un visage sévère. Au contraire on a tenté d'exploiter les ressources graphiques de l'écran (on montre par exemple le portrait des écrivains ou le fac-similé de l'édition originale), et l'on a utilisé pleinement la facilité opératoire des menus déroulants ou des boutons de commandes et la spontanéité de l'utilisation que permet la "souris". Le son même n'a pas été banni. Et une aide a été prévue pour chacune des fonctionnalités, au moment même où celle-ci est proposée: il suffit de solliciter le point d'interrogation qui accompagne chaque bouton.

Ce souci de convivialité et de souplesse explique le choix d'un langage objet pour l'ensemble de ce logiciel écrit en HYPERTALK et complété par des commandes et fonctions externes. Comme il s'agit d'un langage interprété le logiciel est ouvert et admet toutes sortes de compléments.

Dans son état actuel, le prototype exploite une base de 24 textes complets de l'époque révolutionnaire, de 1789 à 1800. Robespierre y voisine avec Sade, Sié ès avec Marat, Mme de Staël avec Chateaubriand, et

Chamfort avec Condorcet. Si ce corpus fait la part belle aux essayistes (parmi lesquels Volney, Marmontel, Bonald), il n'exclut pas les romanciers (par exemple Restif de la Bretonne, Madame Cottin ou Bernardin de Saint Pierre), ni les auteurs dramatiques (Pixérécourt, Lemercier), ni les savants (Lavoisier, Monge, Lagrange). Au total c'est un ensemble de 1.300.000 mots qui se trouve disponible.

Bien entendu on a pensé à traiter d'autres données, par exemple les textes du Moyen Âge, ou la poésie au XIXe. Et on a réalisé une version dépourvue de données mais munie de programmes de préparation. L'utilisateur n'aurait plus alors qu'à introduire ses propres données. Enfin pour les textes soumis aux droits de copyright, une version du logiciel a été étudiée qui escamoterait le texte intégral pour ne délivrer que des contextes. En somme les versions envisagées s'orientent tour à tour sur les deux versants où se sont engagés jusqu'ici les producteurs de bases de données: le premier est plus touffu, plus passionnant peut-être, parce que la découverte y est plus sensible, c'est celui du texte intégral. Le second est balisé, canalisé. Les voies y ont été tracées à l'avance et l'on y circule plus vite. C'est le versant des bases structurées. Il est à parier que les esprits littéraires, plus sensibles à la liberté, choisiront la première voie[7].

NOTES

1. L'INaLF s'est doté de tels appareils de lecture, qui associent un scanner à un logiciel de reconnaissance des formes.
2. Un accord de ce type lie la maison Gallimard et l'INaLF.
3. Ces exigences portaient non seulement sur la saisie mais aussi sur la restitution et l'on a vu à Nancy à cette époque des chaînes d'imprimantes à alphabet riche qu'on ne voyait nulle part ailleurs.
4. Il existe trois autres critères, moins souvent sollicités: le nom de l'éditeur, la référence interne (mais pour la préciser il faut disposer du répertoire), et le code d'accès qui permet de différencier les textes libres et ceux du domaine public.
5. Ajoutons qu'une base de données interrogeable par la voie télématique est sujette aux aléas des communications. Le réseau n'est pas toujours disponible, les lignes sont parfois coupées ou ralenties, TRANSPAC est parfois embouteillé, les transferts sont parfois fautifs et l'ordinateur auquel on s'adresse peut être en panne. De plus la base est dépendante de la machine sur laquelle on l'a installée. Qu'on vienne à changer ce serveur —ce qui doit se produire prochainement à Nancy— et tout le logiciel est à revoir.
6. Cf. l'exposé de ce projet en cours par Emmanuel Le Roy Ladurie, dans *PC Informatique,* septembre 1988, n° 46, p. 18.

7 Entre la rédaction du présent(?) article et sa publication tardive, les choses ont beaucoup évolué. Et les projets qu'on annonce ici ont été réalisés depuis longtemps. En particulier la base FRANTEXT est accessible de tous les points du globe sous une nouvelle forme, plus puissante et plus conviviale et un CD-ROM en a été extrait pour les textes du domaine public (DISCOTEXT 1). Quant au logiciel HYPERBASE, il a été commercialisé sous différentes versions, dont la dernière (2.5) est "standalone" et a servi à la réalisation de plusieurs CD-ROM, proposés au public (sur Rabelais et Gracq).