Tous les chemins mènent-ils au CD-ROM?

Étienne Brunet

Etienne BRUNET

Tous les chemins mènent-ils au *CD-ROM*?

On se propose ici de comparer le CD-ROM aux autres procédures (télématique, disque dur et autres supports) qu'offre la technologie actuelle (celle de 1990) pour réaliser des bases textuelles et les proposer au public.

I - FRANTEXT (version 2)

Il n'est guère utile de présenter FRANTEXT. 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.

1 - Le TLF portant sur la langue du XIX^{e} et du XX^{e} 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. A 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 XVIII^{e}, au XVII^{e} et maintenant au XVI^{e} 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 - A l'heure actuelle la base de données FRANTEXT contient près de 3000 textes complets, soit plus de 150 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.

Or 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

---

[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.

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.

3 - Mais la qualité des données ne serait rien sans l'**efficacité** du logiciel d'interrogation. On a eu l'occasion d'exposer les principes et les vertus de STELLA, que Jacques Dendien a réalisé pour permettre l'accès télématique aux données de FRANTEXT. Or depuis quelques mois une nouvelle version a été implantée qui multiplie les avantages de ce logiciel exceptionnel et dont les fonctions sont restées semblables aux précédentes - l'utilisateur déjà familier des anciennes procédures ne sera pas dépaysé.

a - Il est invité comme précédemment à choisir son **corpus** de travail. Ce peut être d'ailleurs le corpus tout entier, si la question posée n'engendre pas des résultats trop volumineux. Mais le plus 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 introduisant (et en conjuguant au besoin) ces différents critères de sélection. On peut aussi se constituer plusieurs corpus virtuels, parmi lesquels on choisit au dernier moment celui qu'on veut explorer et qui devient le corpus courant.

b - De même qu'il circonscrit son corpus à sa guise, le chercheur peut délimiter le **champ lexical** qui l'intéresse, en dressant la liste des formes qui constituent ce champ. Diverses aides lui sont proposées pour établir ces listes, par exemple pour conjuguer un verbe, ou pour filtrer par un masque les formes qui appartiennent au même radical ou partagent le même suffixe. Ces séries de mots peuvent être soumises, comme les listes de textes, aux opérateurs booléens et être combinées pour produire union, intersection ou différence.

c - Après ces phases de préparation, 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 et le plus naturel, on souhaite relever le contexte d'une forme. La commande qui répond à cette fonction 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 peut traiter des formes ou des listes de mots préétablies, en leur appliquant des critères de filtrage très

sophistiqués qui s'étendent même aux délimiteurs textuels (fin de lignes, de phrases, de paragraphes).

4 - Mais la version 2 de Stella offre de grandes **nouveautés**.

a - Ce qui est nouveau dans la version 2 de Stella, c'est d'abord l'extrême **rapidité** des traitements. Rien n'interdit désormais d'interroger la base entière, même si la demande est complexe. Même alors le temps de réponse se compte en secondes, jamais en minutes. Au besoin, plusieurs requêtes peuvent être traitées en parallèle, en tirant profit des possibilités de multitraitement qu'offre le système UNIX.

b - Ce qui frappe aussi l'utilisateur, c'est la **convivialité** grandement améliorée du dialogue. Le système du menu a été généralisé, rendant inutile la mémorisation des commandes et de leur syntaxe. Quoique la hiérarchie des opérations reste complexe, l'utilisateur n'est jamais perdu et il lui est toujours possible de choisir ou de retrouver son chemin.

c - La **souplesse** du logiciel est remarquable, compensant les contraintes liées à la télématique. Frantext s'adapte maintenant sans problèmes au goût de l'usager et à la multiplicité des claviers.

d - Enfin la **richesse** des fonctions s'est amplifiée, notamment du côté de la **statistique** lexicale. Des outils sont proposés qui mesurent la fréquence des mots d'une liste dans un corpus défini et les sous-fréquences des mêmes mots parmi les textes, les auteurs ou les tranches chronologiques de ce corpus. Mieux même, l'étendue des sous-ensembles est prise en compte pour calculer les fréquences relatives et délivrer des histogrammes. On en donnera ci-dessous deux exemples, légèrement retouchés, à partir des dérivés du mot *raison,* le premier rendant compte de la distribution dans les genres (les essais font une plus large place à la *raison*), le second manifestant l'évolution du champ lexical à travers le temps, de 1600 à nos jours (le culte de la *raison* est évidemment plus fort durant les siècles classiques).

Figure 1. La *raison* et les genres littéraires

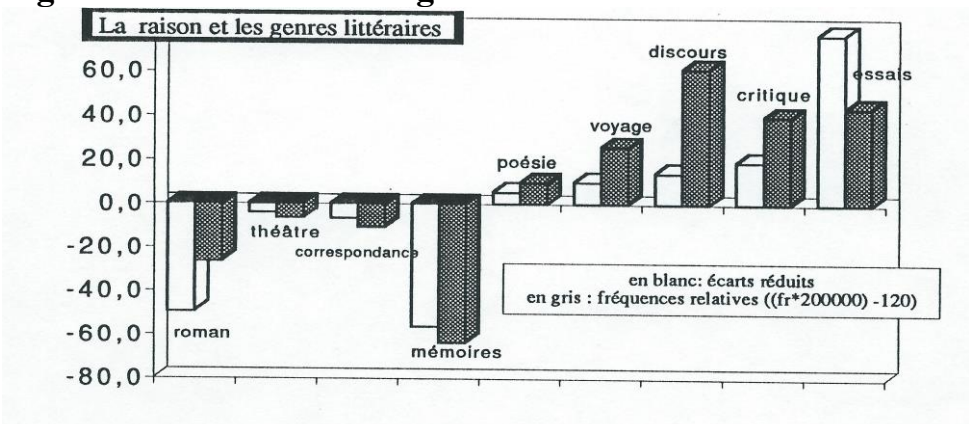
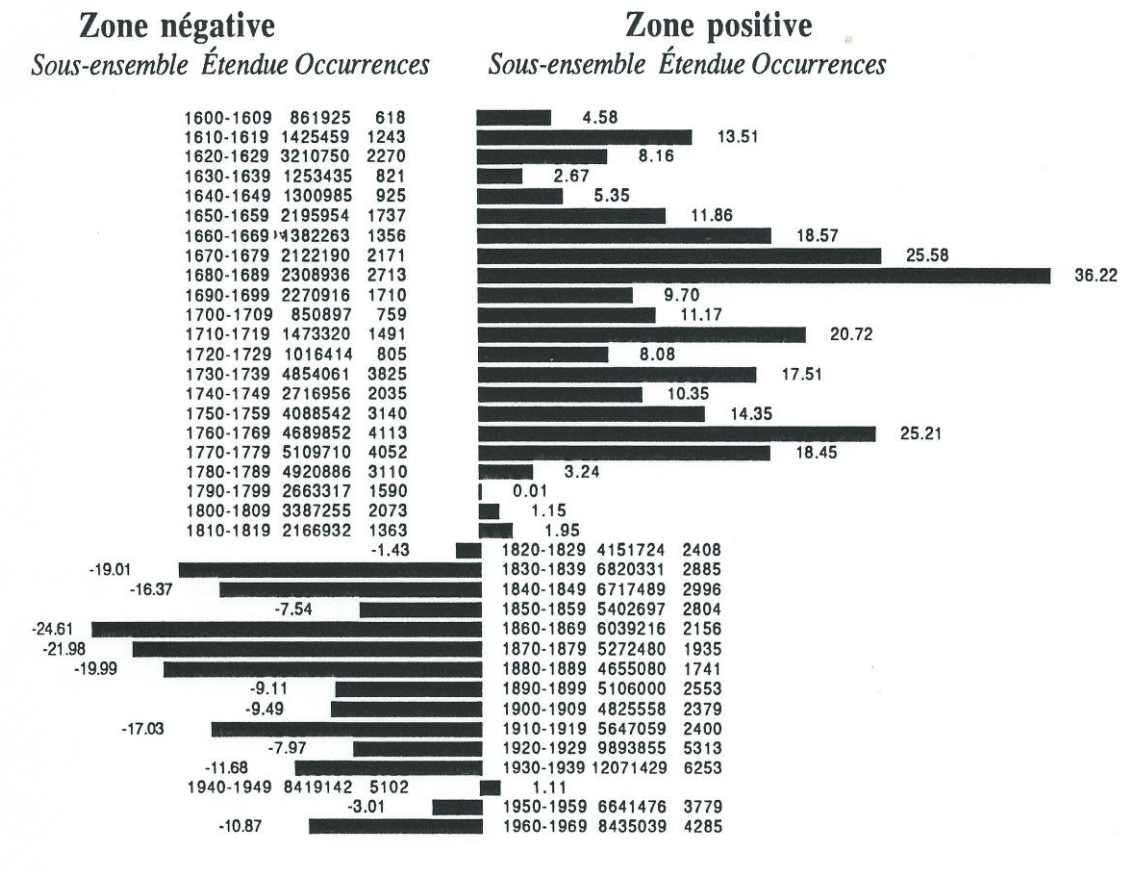


Figure 2. Les dérivés de *raison* de 1600 à 1970



II – DISCOTEXT

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 ou d'un microordinateur pourvu d'un modem (un simple minitel peut suffire) et d'une documentaliste 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 avec les éditeurs ont été passés 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 (la limite est de 300 mots). 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.

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'édition 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), décliner son identité (et son mot de passe), et bien entendu enfin être rompu aux finesses du logiciel d'interrogation STELLA, qui rend FRANTEXT accessible[1]. Ce parcours d'obstacles peut effrayer l'utilisateur et même le décourager de s'adresser aux intermédiaires.

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 disque compact. 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

[1] 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 s'est produit récemment à Nancy - et tout le logiciel est à revoir.

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 anglais de ce disque optique, dont la désignation française est tardive et flottante: DOC ou DON - offre une alternative intéressante à la télématique. Le marché de ce disque compact, longtemps hésitant, se développe actuellement aux Etats-Unis et au Japon, et, le mouvement gagnant l'Europe, de nombreux projets sont en chantier qui utilisent en France cette technologie[1], à l'exemple de celui de la Bibliothèque Nationale[2].

Or l'Institut National de la langue française s'est intéressé dès l'origine aux possibilités que les mémoires optiques offraient en matière de capacité de stockage et de souplesse de diffusion. Un premier CD–ROM (qui porte un nom branché: DISCOTEXT) a été réalisé et sa commercialisation est imminente (éditeur: Hachette). Si le logiciel d'interrogation a été réalisé par le Bureau van Dijk, la collaboration étroite de Jacques Dendien a grandement favorisé la ressemblance et la compatibilité de *Discotext* avec *Frantext* et le maintien des mêmes procédures. On s'en assurera en consultant les différents menus qui se présentent à l'utilisateur et que nous avons reproduits dans les pages suivantes.

Naturellement les vertus de Frantext se retrouvent ici. L'étendue des données offertes - quoique inférieure à celle de **Frantext** - reste considérable. Il y a là 575 textes entassés sur une étroite surface, avec leurs index exhaustifs et tous les fichiers nécessaires à l'exploitation. Une telle concentration de l'information représente un exploit technique dont le secret est révélé par Jacques Dendien[3].

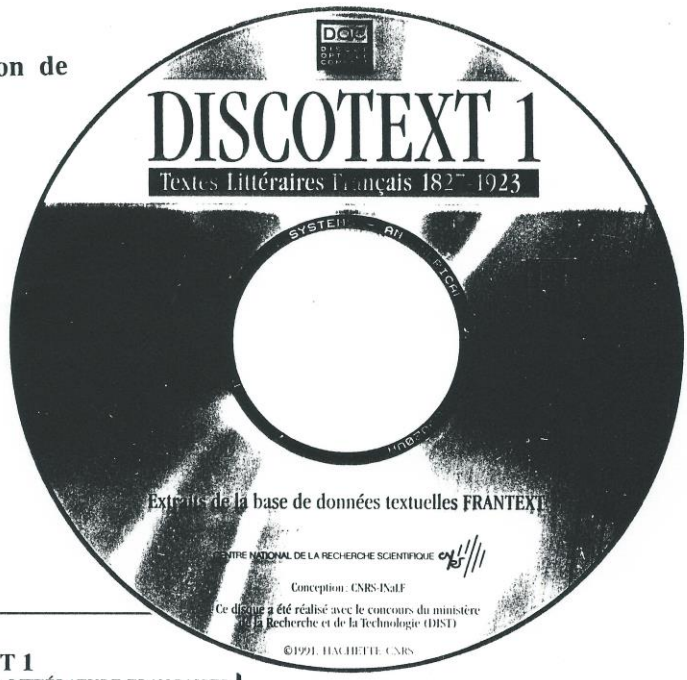
---

[1] On a compté 308 CD-ROM répertoriés et analysés dans le dernier numéro d'une revue spécialisée *Kim Stacks & CD-ROM*.

[2] Cf l'exposé de ce projet en cours par Emmanuel Le Roy Ladurie, dans *PC Informatique*, septembre 1988, n° 46, p. 18).

[3] Dans *Research in Humanities Computing,* Clarendon Press, Oxford, 1991, pp. 308-324. L'économie des index réalisés par Jacques Dendien est telle qu'elle s'approche de la limite idéale que dessine la théorie de l'information. "The index of all the forms of a text, including textual markers (ends of lines, breaks between sentences, chapters and pages, etc.) only represents 25 per cent of the naturel volume of the text (coded in ASCII, for example).", p.324.

Figure 3. La présentation de *Discotext 1*



PAGE D'ACCUEIL

DISCOTEXT 1

POUR L'ÉTUDE DE LA LANGUE ET DE LA LITTÉRATURE FRANÇAISES

Textes Littéraires français 1827 - 1923

Extraits de la base de données textuelles élaborée par l'Institut National de la Langue Française (INaLF) du CNRS.

Logiciel d'analyse réalisé par le Bureau van Dijk, avec la collaboration de J. Dendien (INaLF).

Conception : **CNRS-INaLF** Edition : **Hachette** Supérieur

Réalisation : **van Dijk/CNRS-INaLF**

avec le concours du MRT (DIST) (c) Hachette/CNRS

Appuyez sur une touche pour démarrer Esc *Retour au Dos*

<---------- ***Page d'accueil***

Menu principal ----->

MENU PRINCIPAL

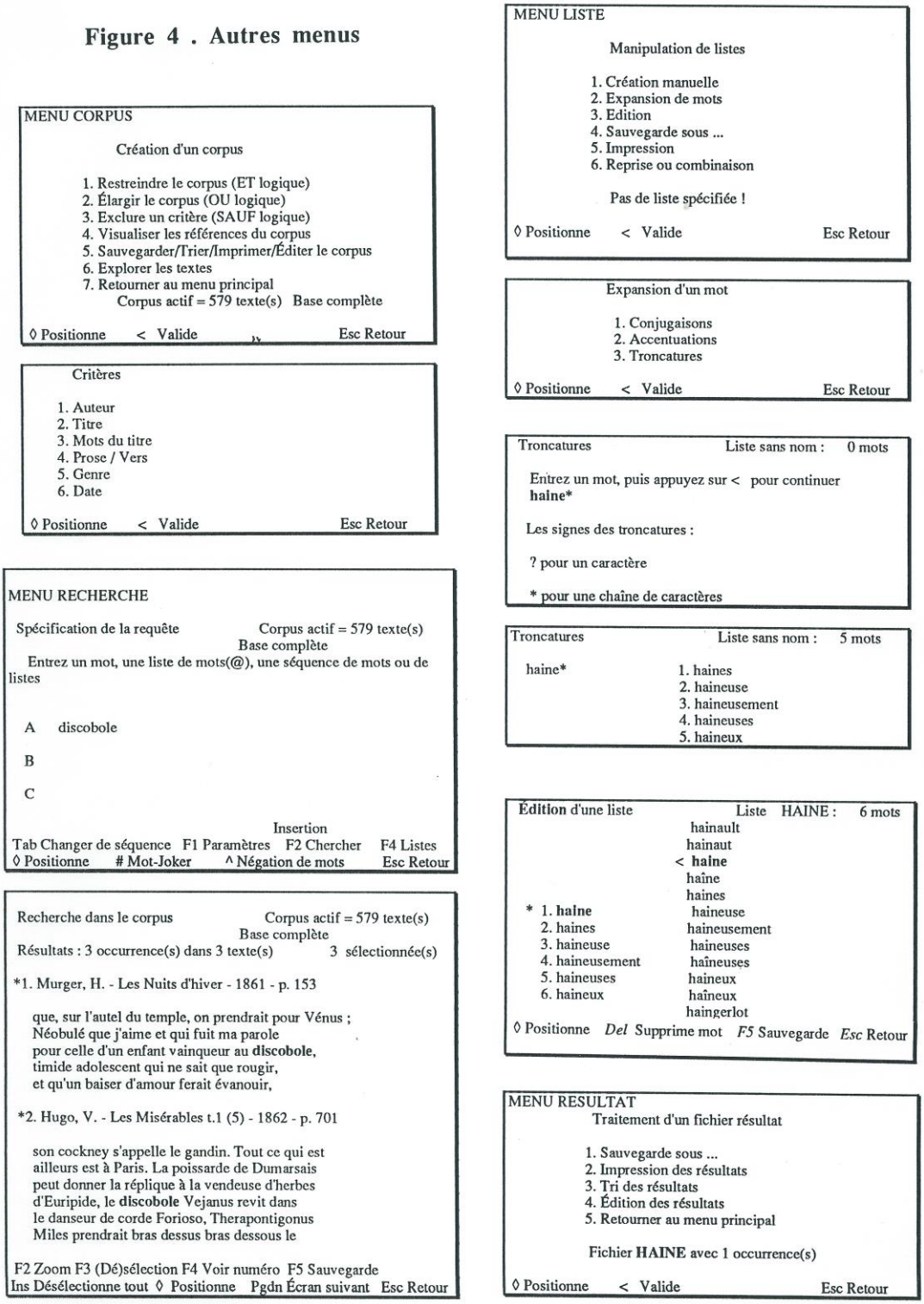
Voulez-vous

1. Créer un nouveau corpus
2. Reprendre ou combiner un corpus
3. Explorer les textes
4. Reprendre un fichier résultat
5. Consulter les services annexes
6. Retourner au DOS

Corpus actif = 579 texte(s)
Base complète

◊ Positionne ¬ Valide

Figure 4 . Autres menus



MENU CORPUS

Création d'un corpus

1. Restreindre le corpus (ET logique)
2. Élargir le corpus (OU logique)
3. Exclure un critère (SAUF logique)
4. Visualiser les références du corpus
5. Sauvegarder/Trier/Imprimer/Éditer le corpus
6. Explorer les textes
7. Retourner au menu principal

Corpus actif = 579 texte(s) Base complète

◊ Positionne < Valide Esc Retour

Critères

1. Auteur
2. Titre
3. Mots du titre
4. Prose / Vers
5. Genre
6. Date

◊ Positionne < Valide Esc Retour

MENU RECHERCHE

Spécification de la requête Corpus actif = 579 texte(s)
Base complète

Entrez un mot, une liste de mots(@), une séquence de mots ou de listes

A discobole

B

C

Insertion

Tab Changer de séquence F1 Paramètres F2 Chercher F4 Listes
◊ Positionne # Mot-Joker ^ Négation de mots Esc Retour

Recherche dans le corpus Corpus actif = 579 texte(s)
Base complète
Résultats : 3 occurrence(s) dans 3 texte(s) 3 sélectionnée(s)

*1. Murger, H. - Les Nuits d'hiver - 1861 - p. 153

que, sur l'autel du temple, on prendrait pour Vénus ;
Néobulé que j'aime et qui fuit ma parole
pour celle d'un enfant vainqueur au **discobole**,
timide adolescent qui ne sait que rougir,
et qu'un baiser d'amour ferait évanouir,

*2. Hugo, V. - Les Misérables t.1 (5) - 1862 - p. 701

son cockney s'appelle le gandin. Tout ce qui est
ailleurs est à Paris. La poissarde de Dumarsais
peut donner la réplique à la vendeuse d'herbes
d'Euripide, le **discobole** Vejanus revit dans
le danseur de corde Forioso, Therapontigonus
Miles prendrait bras dessus bras dessous le

F2 Zoom F3 (Dé)sélection F4 Voir numéro F5 Sauvegarde
Ins Désélectionne tout ◊ Positionne Pgdn Écran suivant Esc Retour

MENU LISTE

Manipulation de listes

1. Création manuelle
2. Expansion de mots
3. Edition
4. Sauvegarde sous ...
5. Impression
6. Reprise ou combinaison

Pas de liste spécifiée !

◊ Positionne < Valide Esc Retour

Expansion d'un mot

1. Conjugaisons
2. Accentuations
3. Troncatures

◊ Positionne < Valide Esc Retour

Troncatures Liste sans nom : 0 mots

Entrez un mot, puis appuyez sur < pour continuer
haine*

Les signes des troncatures :

? pour un caractère

* pour une chaîne de caractères

Troncatures Liste sans nom : 5 mots

haine*

1. haines
2. haineuse
3. haineusement
4. haineuses
5. haineux

Édition d'une liste Liste HAINE : 6 mots

hainault
hainaut
< **haine**
haîne
haines
haineuse
haineusement
haineuses
haîneuses
haineux
haîneux
haingerlot

* 1. **haine**
2. haines
3. haineuse
4. haineusement
5. haineuses
6. haineux

◊ Positionne *Del* Supprime mot *F5* Sauvegarde *Esc* Retour

MENU RESULTAT

Traitement d'un fichier résultat

1. Sauvegarde sous ...
2. Impression des résultats
3. Tri des résultats
4. Édition des résultats
5. Retourner au menu principal

Fichier **HAINE** avec 1 occurrence(s)

◊ Positionne < Valide Esc Retour

Quelques avantages sont offerts par *Discotext* qui étaient hors de portée de *Frantext*. Comme ce *CD-ROM* est dédié à un standard particulier - celui des *PC-IBM* - il est possible d'utiliser les propriétés de ce standard et d'en exploiter les spécificités sans être astreint au respect des autres standards, et à la multiplicité de leurs écrans, de leurs claviers et de leurs systèmes. Sans aller jusqu'à faire appel à la souris - ce qui aurait limité l'emploi aux possesseurs de *Windows* -, *Discotext* utilise la couleur de l'écran *VGA* et les touches de fonctions du clavier *IBM*. L'affichage, qui outre la couleur utilise les ressources de la surbrillance et du multi-fenêtrage, a des vertus ergonomiques supérieures à celles de *Frantext.*

Pourtant l'utilisateur familier de *Frantext* ressent parfois des frustrations. La plus fréquente est liée à la lenteur des recherches. À quoi sert de disposer d'un affichage plein écran, rapide et coloré, si l'on doit attendre trop longtemps le transfert des données? Le programme n'est pas en cause, non plus que l'ordinateur utilisé (un simple *IBM-AT* a une puissance suffisante[1]). Ce qu'on doit incriminer, c'est le support optique qui n'a pas été conçu pour l'adressage direct et immédiat des informations enregistrées et qui réagit avec un retard de près d'une demi-seconde à toute sollicitation[2]. Les autres limitations sont plus acceptables: l'impossibilité d'écrire sur la surface optique est aisément contournée par le recours au disque dur, qui sert de réceptacle aux résultats et aux fichiers intermédiaires. L'orientation obligatoire des résultats vers l'imprimante et le refus d'enregistrer les contextes en clair (seuls sont consignées les adresses) ne sont pas une gêne pour les demi-habiles, qui auront vite fait de détourner le périphérique d'impression vers un fichier d'enregistrement. Quant au cryptage des données textuelles, imposé par des raisons de confidentialité, on ne le regrette qu'assez peu, la lecture suivie sur écran étant par trop inconfortable.

---

[1] L'émulation d'un *IBM-AT* est tout-à-fait réalisable, à partir de matériels étrangers à ce standard. Disposant d'un *MAC II*, d'un lecteur de *CD-ROM Apple* et du logiciel *SoftPC*, nous avons pu mettre en oeuvre *Discotext*, sans dégradation appréciable des performances.

[2] Le disque compact à lecture laser vient d'une technologie étrangère à l'informatique, celle du son. Or si la lecture linéaire convient à l'écoute musicale et se satisfait d'une inscription continue, en spirale, les données informatiques au contraire s'accordent mieux avec la discontinuité des pistes concentriques, divisées en secteurs et adressables avec précision. D'autre part les hauts débits sont inutiles dans le flux régulier des informations musicales, nécessairement synchronisées avec l'écoute, alors que la vitesse du transfert acquiert une importance capitale en informatique, dans le domaine des réseaux comme dans celui du stockage. Or ici encore les limites du *CD-ROM* sont basses, avec une capacité de l'ordre de 150 Ko par seconde.

III - Tout GRACQ sur ordinateur

A - Les considérations qui précèdent nous ont fait surseoir à un projet qui nous tenait à coeur et pour lequel nous avions réalisé un prototype adapté, comme *Discotext*, aux spécificités du *CD-ROM***.** Conçue initialement en 1989 pour une manifestation du Bicentenaire à Beaubourg (il s'agissait de mettre à la disposition du public un ensemble de textes de la Révolution), la version première d'*HYPERBASE* ne nécessitait que deux déplacements de la tête de lecture pour tout accès sélectif à l'information. Nous éviterons d'en décrire le détail, puisque deux publications sont explicites là-dessus[1]. Nous nous contenterons de reproduire, dans la figure 5, 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.

La figure 6 montre le détail du corpus de Julien Gracq, auquel manque le dernier texte, publié en 1992: *Carnets du grand chemin.* Ce texte a pourtant été soumis, dès sa parution, à la saisie optique. Mais la pile primitive n'a pas été modifiée pour recevoir ce complément. Car entre temps, une nouvelle pile avait été créée, qui se fonde sur des principes différents et n'est plus orientée vers le *CD-ROM*. C'est en effet la grande capacité du *CD-ROM* qui justifie son emploi et fait pardonner sa lenteur. C'est le cas de Discotext. C'est le cas des grands dictionnaires, comme le Grand Robert, l'Oxford English Dictionnary, l'encyclopédie Grolier. Et la situation est semblable lorsqu'on doit emmagasiner des sons, ou des images, les uns et les autres étant grands dévoreurs d'espace. Mais l'oeuvre de Julien Gracq n'a pas une taille démesurée. Les 17 titres réunis qui constituent son oeuvre complète tiennent dans cinq millions de caractères. Est-il raisonnable de réserver inutilement un espace cent fois plus grand, en acceptant les contraintes techniques et commerciales qui pèsent sur les fabrications industrielles. Aux États-Unis se développe une fabrication artisanale, décentralisée, et peu coûteuse de *CD-ROM* amateurs qui sont

---

[1] "Computer processing and quantitative text analysis: HYPERBASE, an interactive software for large corpora", in *Data analysis, learning symbolic and numeric knowledge*, INRIA, Nova Science Publishers, New York - Budapest, 1989, p. 207-214.

"What do statistics tell us?", in *Research in Humanities Computing*, Clarendon Press, Oxford, 1991, p. 70-92

Figure 5 .La version CD-ROM d'HYPERBASE. Menu principal

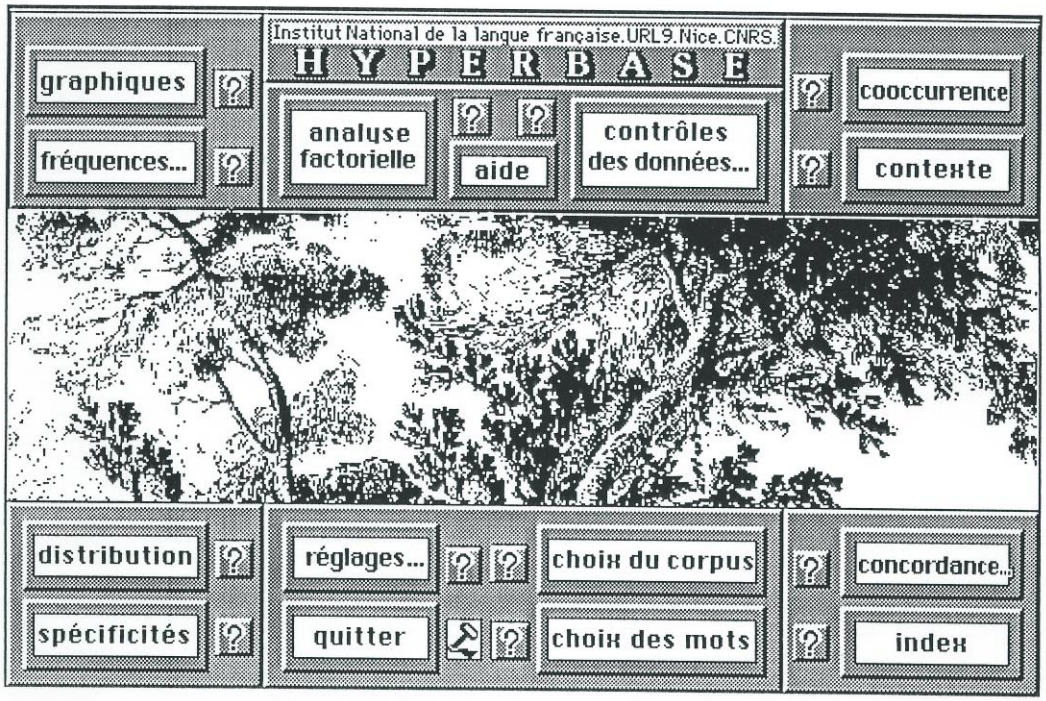
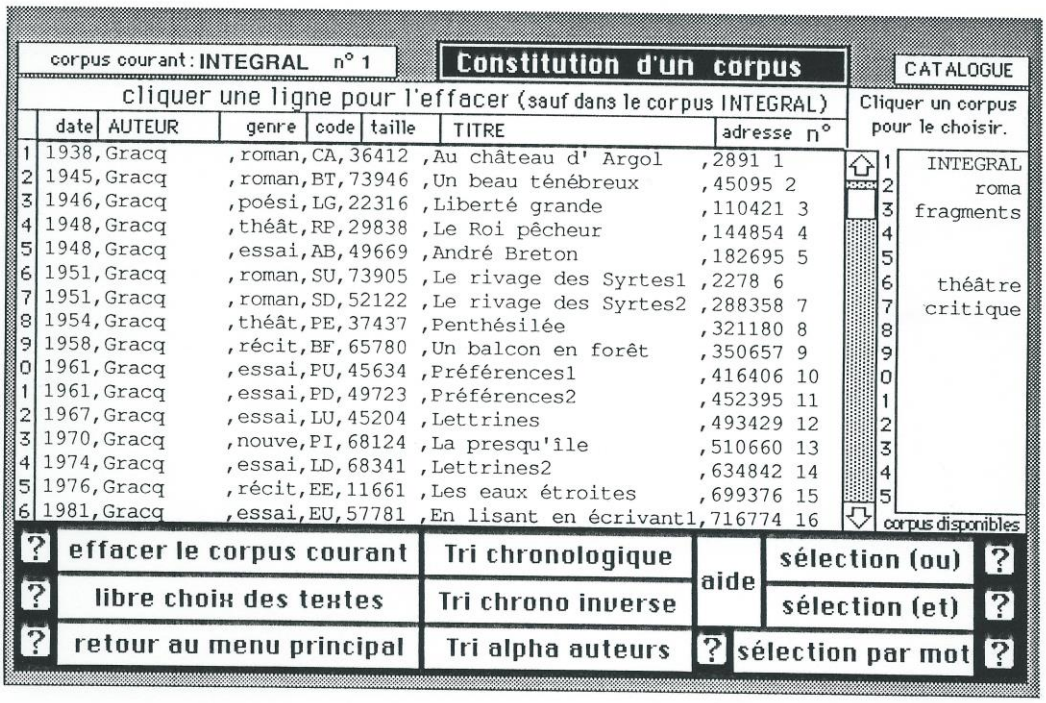


Figure 6 - Le corpus GRACQ



fabriqués à la demande et à l'unité. Tant que de telles techniques ne seront pas disponibles en Europe, il faudra passer par la chaîne pesante du mastering et du pressage, qui se réalise en usine, au prix fort.

B - Or la technologie offre des **substituts** au *CD-ROM*, qui ne souffrent ni de sa lenteur de déplacement, ni de sa lourdeur de fabrication. Tous ont en commun un support inscriptible et effaçable qui peut être rempli par simple copie, ce qui convient aux petites séries et donc aux bases de données littéraires ou linguistiques qui ne s'adressent guère au grand public et ne visent que la clientèle limitée des bibliothèques universitaires. Le support le plus proche du *CD-ROM* est le WORM ou disque laser inscriptible une fois (et non effaçable). Notre base a été transférée avec succès sur ce support, mais la diffusion n'en est pas envisageable, à cause du trop faible nombre de lecteurs WORM dans le monde. La technologie mixte du support opto-magnétique a plus de souplesse (les enregistrements sont effaçables), plus de rapidité, une plus grande modularité (la contenance varie de 100 Mo à 600 Mo) et un avenir mieux assuré. Mais la guerre des standards y fait rage et le marché est hésitant entre le 3 pouces et le 5 pouces, entre les marques Ricoh et Sony et beaucoup d'autres dont les lecteurs ne sont pas compatibles. Quoique notre base fonctionne fort bien sur un tel support, il est risqué de s'aventurer dans un choix prématuré, et prudent d'attendre que le marché désigne un vainqueur. Le disque dur est pour l'utilisateur le périphérique le plus commode, le plus commun, le plus rapide, le plus sûr et le mieux standardisé (à la norme SCSI). Mais personne n'a jamais envisagé de distribuer une base de données sur un tel support dont le coût est nécessairement élevé puisqu'il incorpore non seulement la surface traitée et les données enregistrées mais aussi la tête de lecture et la mécanique d'enregistrement. La seule voie possible ici est celle de la cartouche amovible qui reste indépendante du lecteur et dont le prix s'abaisse en conséquence. Nous étudions cette formule, qui reste encore coûteuse (les cartouches de 44 Mo valent environ 400 frs) et ne peut atteindre que les possesseurs de lecteurs *Syquest*. Reste l'option la moins glorieuse: celle de la disquette. Ce support est le plus ancien, le plus disponible, le moins cher et le plus léger. C'est par cette voie que se distribuent les logiciels et beaucoup de données. On connaît ses défauts: lenteur, faible fiabilité, faible contenance. Aucun n'est rédhibitoire. Lenteur et fragilité sont tolérables pour un support qui est appelé à ne servir qu'une fois, au moment où le contenu est transféré sur le disque dur de l'utilisateur. Quant à la faible capacité, divers procédés permettent d'y porter remède. Les plus connus sont le compactage, qui double en général le volume utile, et le chaînage qui permet de répartir le contenu des gros fichiers sur plusieurs disquettes associées (selon une technique analogue à celle du fichier multivolume qu'on rencontre sur les gros ordinateurs de gestion). Comme

le corpus Gracq a une taille moyenne (10 millions de caractères au total, texte et index compris), il semble que l'option d'une diffusion par disquette soit la plus modeste, la plus raisonnable et la moins coûteuse[1].

C - Elle a cependant nécessité la **refonte** entière du programme d'exploitation. Il n'était plus nécessaire de se lier aux contraintes du *CD-ROM* et de prévoir des accès optimisés, ce qui rendait un peu de liberté et autorisait des objectifs nouveaux. Certes la pile *HYPERBASE* a gardé le même nom, le même environnement *Hypercard*, les mêmes fonctionnalités. Mais sa structure est radicalement différente et l'orientation fort différente. L'ancienne version suivait le modèle des bases de données commerciales, qui sont proposées sur les serveurs ou sur *CD-ROM*. On y livre les données et le logiciel d'exploitation, mais non les programmes de préparation. En somme les clients sont invités à la salle à manger, mais non à la cuisine. L'autre logique est celle des logiciels vides de données (comme les traitements de texte, les tableurs, etc.), mais riches d'outils variés, conçus pour aider l'utilisateur à traiter son bien propre. Nous avons tenté de concilier les deux points de vue. La pile consacrée à Julien Gracq contient effectivement l'oeuvre entière de l'écrivain et toutes les informations (dictionnaire, index, spécificités...) propres à cette application. Mais la pile peut être vidée de son contenu et recevoir d'autres données. Programmes de préparation et d'exploitation sont fournis conjointement dans le même produit. Ainsi espère-t-on satisfaire une communauté scientifique plus large que l'église où se réunissent les fidèles de Julien Gracq.

HYPERBASE dans sa version nouvelle vise à la généralité et à la simplicité, même en perdant un peu de puissance. Le produit est conçu pour s'adapter immédiatement aux données de l'utilisateur et réaliser l'indexation dans un temps acceptable et sans manipulation excessive. Les deux objectifs antérieurs ont été maintenus qui orientent l'exploitation vers la recherche documentaire et la statistique.

1 - Le programme d'exploitation répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes:

---

[1] Une nouvelle espèce de disquette est maintenant proposée sur le marché, dont la contenance n'est pas si faible (de 5 à 10 Mo). Hélas, les lecteurs requis sont d'un type spécial, ce qui limite la diffusion du produit et élève un peu plus haut la tour de Babel. Verra-t-on un jour un lecteur universel, capable de lire tous les formats? On songe avec un peu de nostalgie et d'envie à l'universalité du papier, qui depuis cinq siècles s'est accommodé de tous les formats, de toutes les écritures, de toutes les machines à imprimer ou à reproduire.

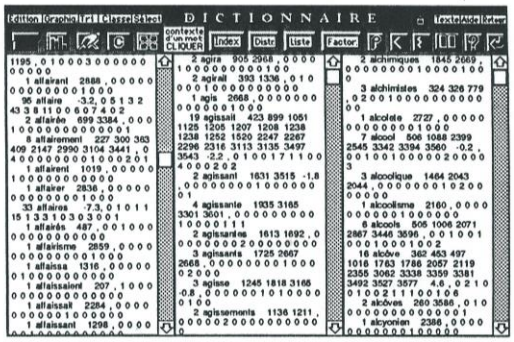
concordances de type kwic (avec tri des expansions droite ou gauche du contexte), index sélectifs ou systématiques, dictionnaires des fréquences, sélection de contextes larges, cooccurrences, filtrage et masquage des mots et constitution de listes, recherche des parties de mots (début, fin ou chaîne quelconque) ou des groupes de mots, limitation ou extension du corpus de travail, etc.. On accède par la souris à toutes ces fonctions documentaires et à leurs multiples options. Les résultats sont mémorisés en même temps qu'ils apparaissent à l'écran et peuvent être à tout moment dirigés sur l'imprimante ou sur la mémoire de masse.

HYPERBASE, qui entre dans la famille des hypertextes, se distingue toutefois des produits traditionnels

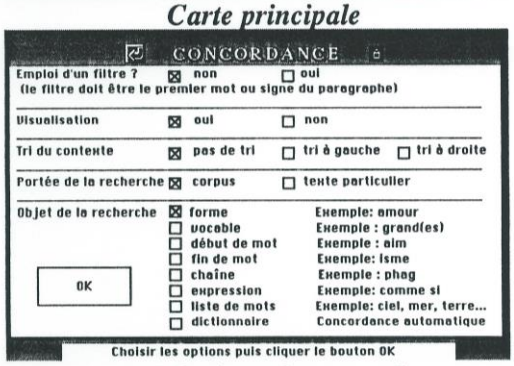
- par l'exhaustivité de l'indexation, qui prend en compte tous les mots et même les signes de ponctuation
- par le respect des signes diacritiques et l'adaptation des routines de tri et de recherche aux alphabets européens
- par la variété des filtres d'interrogation, des options de traitement et des résultats obtenus
- par l'accessibilité du dictionnaire et du texte, l'un et l'autre reproduits en clair. L'exploration du corpus peut se faire librement par le va-et-vient contrôlable de l'un à l'autre.
- par la souplesse et la convivialité de l'exploitation (facilitée par *HYPERCARD* et l'interface *Apple*)

2 - Il s'en distingue surtout par l'orientation statistique donnée au produit. D'une part, s'il s'agit d'un texte français, une comparaison est faite, 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* délivre alors pour chaque mot les sous-fréquences, qui judicieusement pondérées se transforment à volonté en courbes ou en listes spécifiques. *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. D'autres calculs lexicométriques sont assurés qui permettent d'apprécier la richesse relative du vocabulaire, la distribution des classes de fréquences, l'abondance, si l'on peut dire, des mots rares (ou hapax), l'accroissement et l'évolution du vocabulaire, etc…

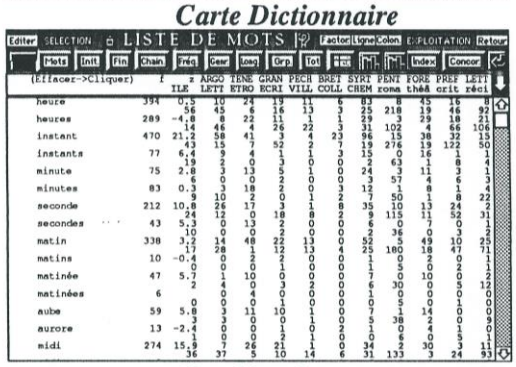
Carte principale



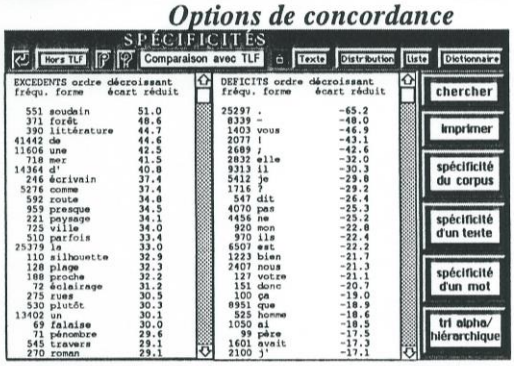
Carte Dictionnaire



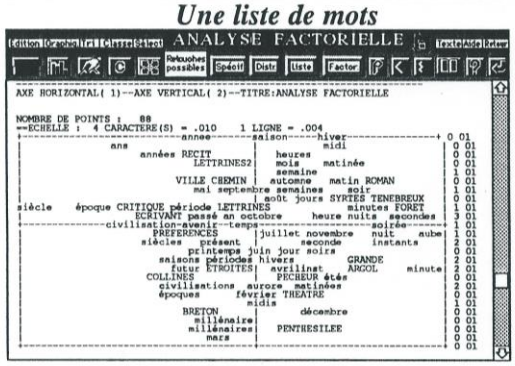
Options de concordance



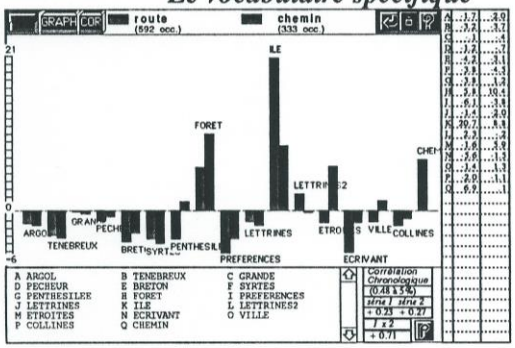
Une liste de mots



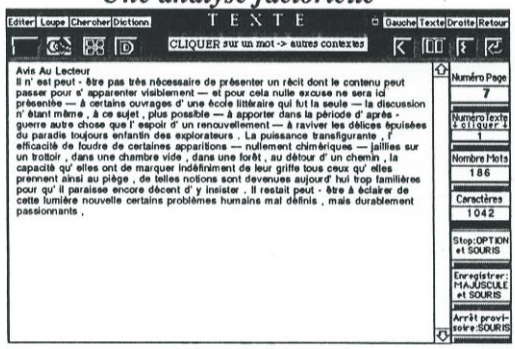
Le vocabulaire spécifique



Une analyse factorielle



Graphique



Une fiche-page

Figure 7. Quelques fonctionnalités d'HYPERBASE

Surtout Hyperbase permet de constituer à son gré des listes manuelles ou automatiques, et de circonscrire ainsi une catégorie grammaticale, un champ thématique, voire même le système de la ponctuation. Une fois constituées, ces listes - ce sont en réalité des tableaux à deux dimensions - peuvent être soumises aux méthodes multidimensionnelles (un programme d'analyse factorielle a été intégré à *HYPERBASE*).

3 - On ne s'appesantira pas sur les spécifications techniques. Il suffit de préciser que *HYPERBASE* comprend un programme de préparation (écrit en Pascal), un éditeur de texte (écrit en 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). HPERBASE requiert la configuration suivante: mémoire vive minimum: 2000K; disque dur indispensable; Mac II couleur ou SE30 recommandés. *HYPERBASE* fonctionne indifféremment sur système 6 ou 7 et avec les versions 1.2 ou 2.0 d'*HYPERCARD*.

4 - Pour illustrer les ressources de ce logiciel, revenons à Julien Gracq[1]. La figure 7 illustre la manière dont *HYPERBASE* rend compte d'un corpus littéraire[2]. Mais il n'y a pas lieu ici d'analyser les divers aspects de l'oeuvre de l'écrivain qu'on découvre à l'aide d'un tel programme: l'étendue et la variété de son lexique, la structure grammaticale et rythmique de sa phrase, la spécificité de ses thèmes, et, sur tous ces points, l'originalité de Julien Gracq dans son siècle, l'évolution interne qui gouverne sa production littéraire, enfin la distinction et le mélange des genres dans une oeuvre qu'il a voulue composite et rebelle aux étiquettes.

---

[1] On a choisi Julien Gracq, d'une part pour la qualité rare de son écriture qui le place au premier rang en France et peut-être même à la première place, mais aussi pour des raisons d'opportunité qui nous ont permis d'obtenir de Julien Gracq le droit d'enregistrer son oeuvre (grâce au lecteur optique), d'en exploiter le contenu dans une base de données en texte intégral (créée par *Hyperbase*) et de diffuser l'ensemble sous forme de pile *Hypercard* (dans une série limitée). En somme il s'agissait d'entreprendre pour un auteur vivant - en dépit du copyright - ce qu'on a récemment réalisé pour Shakespeare, la Bible, les classiques grecs ou les Pères de l'Église. Les premiers résultats ont été communiqués lors d'un colloque consacré à Julien Gracq qui s'est tenu à Cerisy-la-Salle en Août 1991 et présentés à Julien Gracq lui-même, qui ne s'en est trouvé aucunement paralysé (comme le malheureux romancier de *Small World*) puisque quelques mois plus tard paraissaient les *Carnets du grand chemin.*

[2] Nul besoin que le corpus ait cette qualité littéraire. Ce sont les sociologues qui sont les plus nombreux à utiliser *Hyperbase*, pour le traitement des questions ouvertes de leurs enquêtes.