Proceedings of VEXTAL'99 – Venezia, San Servolo V.I.U. – 22-24 November 1999

La base textuelle *Batelier*

Étienne Brunet
Institut national de la langue française

La base *Batelier* (Base de Textes Littéraires pour l'Enseignement et la Recherche) est un prototype actuellement à l'essai dans une vingtaine de lycées français. Ce projet met en jeu trois opérateurs: le Ministère de l'Éducation Nationale, l'Institut National de la langue française et les Éditions Champion. Il ne fait pas double emploi avec la base FRANTEXT, car ni les objectifs, ni les données, ni la taille, ni les moyens, ni le support ne sont les mêmes. Les éditions qui entrent dans cette base obéissent à un principe constant qui est de proposer la dernière version corrigée par l'auteur. Et le texte est préparé et balisé de façon à assurer à l'exploitation la sûreté, la puissance et la cohérence. La qualité, pour un projet qui démarre, importe plus en effet que la quantité, et dans une première étape on n'envisage qu'une centaine de textes et une vingtaine d'écrivains dont le portrait est aisément reconnaissable dans la page d'accueil représentée ci-dessous. Le support choisi, le cédérom, est le moins coûteux et le plus disponible sur les machines actuelles. Quant au logiciel retenu pour réalisation de la base et son exploitation, il s'agit d'une version spéciale d'Hyperbase, dont nous nous proposons de détailler quelques aspects originaux.

Batelier. Choix du corpus

Le cédérom a est au standard *Windows* mais une version particulière tourne sur Macintosh avec des données identiques et des fonctionnalités semblables. Le catalogue des textes disponibles va de Rabelais à Proust et devrait s'enrichir dans les prochains mois. Voici le menu proposé dans la page d'accueil:

Figure 1. La base* Batelier. *Menu principal

Ceux qui connaissent le logiciel *Hyperbase* reconnaîtront les deux axes qui commandent l'exploitation :
En haut, à l'horizontale, une série de boutons à fonction documentaire. Il s'agit d'abord d'ouvrir le texte à la page désirée, ou l'index à la lettre souhaitée, ou bien d'éditer les documents divers qui sont en relation avec la base :

fichiers des données, des résultats et des notes. C'est là que sont proposés surtout les deux boutons prévus pour la recherche hypertextuelle : *Concordance* et *Contexte*.
Dans la marge droite, outre les programmes d'installation ou de liaison avec l'environnement informatique, sont disposés les outils propres à assurer l'exploitation statistique de la base. On distinguera ceux dont la portée est partielle ou locale, parce qu'ils prennent en compte une sélection de la base, et ceux qui traitent la base dans son intégralité.

La présente version offre en effet la possibilité de **choisir un corpus de travail** parmi les textes disponibles. Cette fonction, peu utile lorsqu'un chercheur construit sa base avec ses propres données (car le choix des textes se fait avantageusement au moment de la saisie et du traitement initial), s'avère indispensable lorsqu'un corpus déjà constitué est proposé à des tiers. Un peu de liberté et de souplesse dans l'exploitation doit compenser la rigidité des choix imposés dans la phase de réalisation. L'étude globale reste néanmoins possible et c'est même l'option par défaut, pour les fonctions quantitatives comme pour les programmes documentaires. À l'opposé la sélection d'un texte unique parmi la centaine disponible n'est nullement interdite ou déconseillée. Mais dans ce cas les comparaisons, qui sont à la base de toute statistique, faisant défaut, seules les fonctions de recherche seront justifiées.

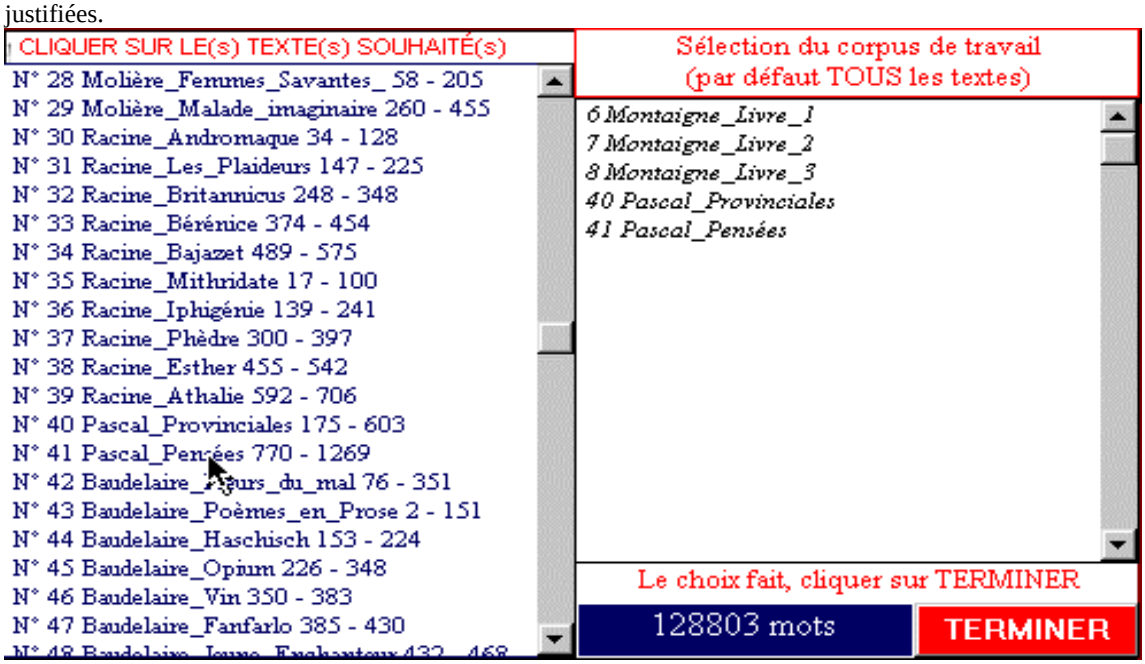

Figure 2. Choix du corpus de travail

La sélection d'un corpus de travail s'opère par le bouton corpus, situé en bas de l'écran. La liste des textes s'offre alors au clic de la souris (les textes retenus passent dans le champ de droite), jusqu'à ce que l'utilisateur estime son panier suffisamment rempli (en sollicitant le bouton Terminer). Pour chaque texte sélectionné on est averti du nombre de pages et du nombre de mots.
Le bouton Lecture donne accès au texte choisi parmi ceux du corpus de travail, après que de brèves informations bibliographiques ont été affichées. La lecture suivie, page après page, est possible grâce aux flèches de navigation, mais pour ce rôle traditionnel le papier offre un confort supérieur. L'écran prend l'avantage dans ses f**onctions hypertextuelles**: il suffit de cliquer sur un mot pour connaitre sa répartition dans le corpus et être conduit dans les passages où le mot se trouve employé (figure 3). Ces excursions verticales peuvent se faire aussi à partir de l'index, c'est-à-dire de la liste alphabétique à laquelle un menu déroulant donne accès dans la page d'entrée (bouton Index). Voir figure 4.

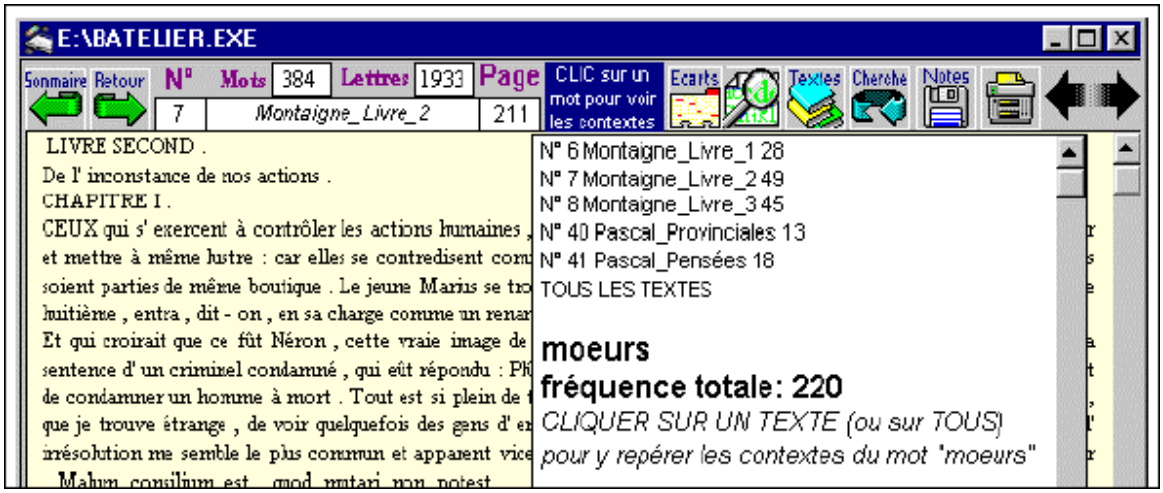

Figure 3. Une page de texte

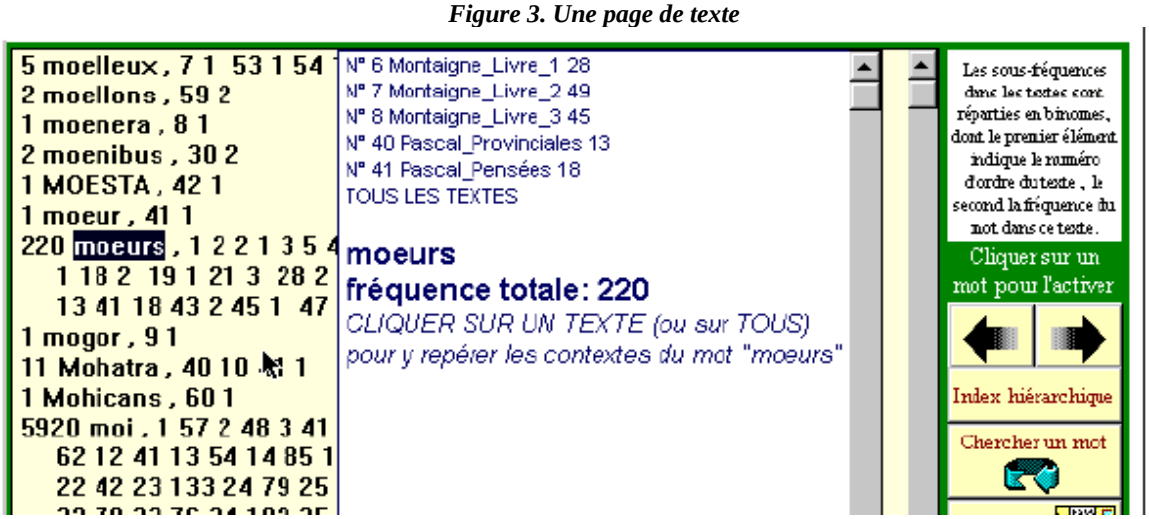

Figure 4. Une page de l'index

Sans reprendre ici les explications développées dans le manuel, nous ne signalerons que les nouveautés introduites dans le logiciel à l'occasion de Batelier, par exemple dans les pages-index le bouton de renvoi à l'index hiérarchique ou encore dans les pages-texte la **mise en relief** (en rouge) des mots qui, à l'intérieur de la page, reflètent les caractéristiques du texte considéré (par rapport au corpus d'ensemble). Le bouton Écarts remplit cet office. Ainsi sont soulignés dans la dernière page du Temps retrouvé les mots longtemps, passé, temps, années, oeuvre, qui sont récurrents dans le dernier roman de Proust (figure 5).

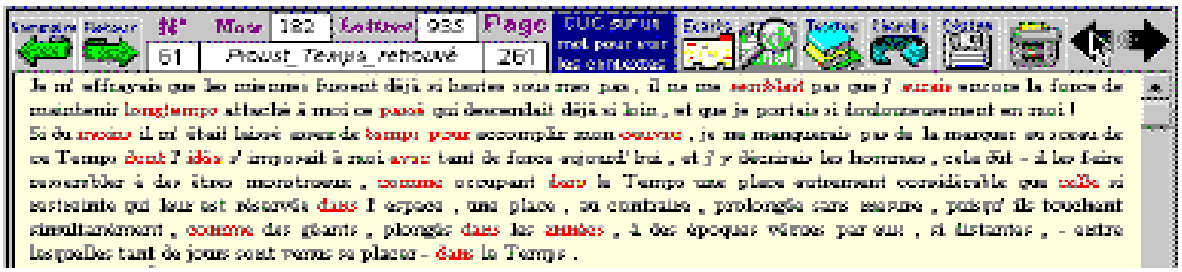

Figure 5. Les spécificités de la dernière page de Proust

Nouvelles fonctions de recherche.

*L*es fonctions **documentaires** Concordance et Contexte restent ce qu'elles étaient dans le passé. Elles restituent soit une ligne de texte, soit le contexte plus large du paragraphe, dès lors que l'usager a précisé l'objet de sa recherche: un mot, une expression (plusieurs mots contigus), une initiale, une finale ou une chaine quelconque. S'y ajoute la

cooccurrence de plusieurs formes dans le cas de la fonction Contexte. Ces programmes classiques ont été complétés par de nouvelles fonctionnalités:
- le paragraphe n'est plus la **taille** imposée au contexte. On peut exiger plus (jusqu'à 1000 caractères) ou moins (jusqu'à 50 caractères)
- l'objet de la recherche peut être un **vocable**, dont les formes fléchies seront regroupées sous l'entrée habituelle aux dictionnaires. Comme le corpus s'étend sur plusieurs siècles, le conjugueur tient compte des graphies anciennes.
- le programme Concordance peut s'appliquer au lexique entier avec divers critères de sélection liés à la fréquence
- la fonction "**zoom**" est commune aux deux programmes: un clic sur une ligne de la concordance ou un paragraphe du contexte déclenche l'affichage de la page entière dans le texte d'origine, avec mise en relief du mot étudié
- dans le cas des **vers**, une distinction est faite entre les fins de vers et les fins de paragraphe, même si le même code ambigu - le retour de chariot - sert à marquer les unes et les autres dans les données d'origine (voir exemple ci-dessous)

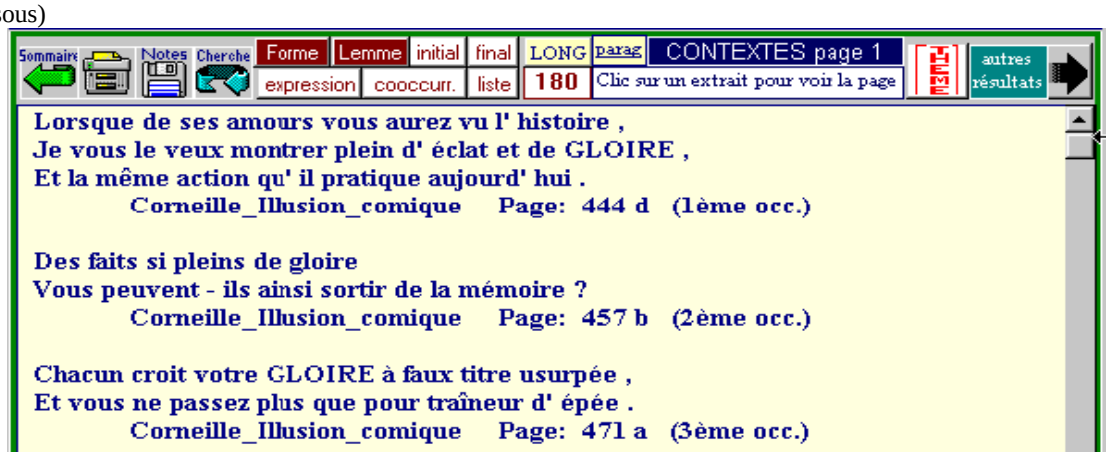

Figure 6. Résultats du programme Contexte

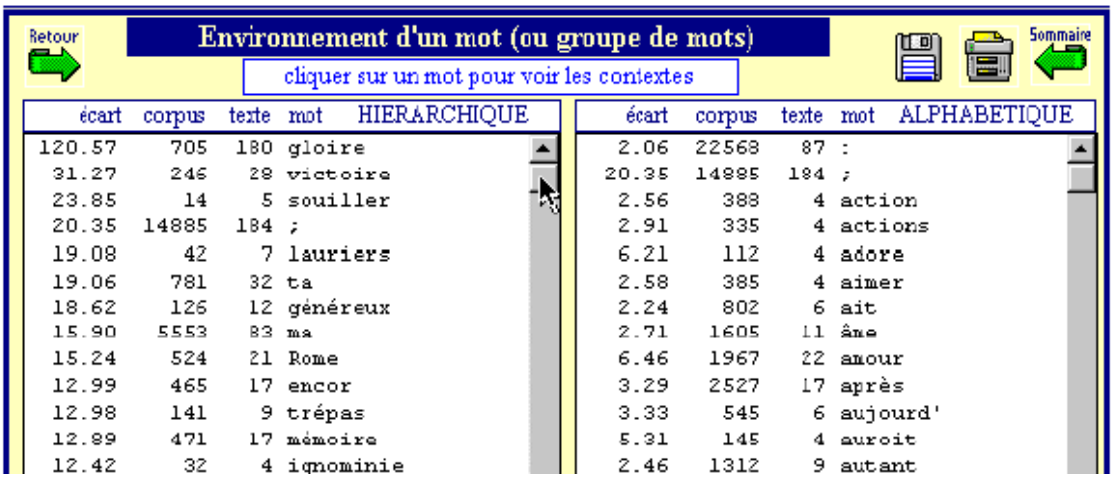

Figure 7. La fonction thématique.
L'entourage du mot gloire chez Corneille

- enfin une fonction "**thématique**" est maintenant disponible (bouton Thème de la page Contexte): quand l'environnement d'un mot (ou d'un groupe de mots) atteint une certaine amplitude qui autorise l'emploi des tests statistiques, le programme procède au relevé et au tri des mots rencontrés dans l'entourage du thème (c'est-à-dire du mot) proposé. Après filtrage statistique le contenu du thème (ce qu'on appelle les corrélats) est affiché dans une liste, triée selon la plus ou moins grande accointance des corrélats avec le mot-pôle. Reste alors à expliquer cette proximité récurrente qui peut être d'ordre syntaxique (le mot gloire impose souvent les possessifs féminin ma ou ta dans la phraséologie du temps et les tragédies de Corneille), de caractère métrique (victoire et mémoire accourent à la rime dès que la gloire les y appelle), ou dont la raison plus généralement appartient au sens ou au thème (dans le même exemple, les liaisons synonymiques lauriers, généreux, illustre, honneur, vainqueur n'excluent pas les antonymes souiller, ignominie). Voir figure 7.

Nouvelles fonctions statistiques

*L*es fonctions **statistiques** ont par défaut une portée maximum et s'exercent sur la totalité du corpus. Certaines n'ont de sens que si cette condition est remplie. On imagine mal par exemple quel pourrait être le résultat du programme Évolution s'il s'appliquait à un ensemble de textes trop étroit ou dépareillé. De la même façon les calculs portant sur la structure lexicale ou sur le profil spécifique d'un texte se justifient mieux si le corpus de référence est plus large, et s'étend à la totalité des textes. Les trois boutons Évolution, Spécificités et Structure donnent donc des résultats constants qui s'attachent à l'**ensemble de la base** et ne dépendent pas du choix d'un corpus de travail particulier. S'ils donnent des informations sur chacun des textes, c'est sans en privilégier aucun.

En revanche les autres fonctions statistiques (Graphique, Liste, Factorielle et Grammaire) s'appliquent au **corpus de travail** (qui peut aussi être le corpus intégral). Prendre garde toutefois qu'aucune statistique n'est possible si l'on a moins de deux textes à comparer et qu'aucune analyse factorielle n'est concevable si l'on en compte moins de trois. En dehors de cette modification sur la portée de telles fonctions, les programmes statistiques ont bénéficié de quelques retouches:

1 - Le grapheur permet d'ajouter les **légendes** de son choix, au bas du graphique. Il donne le choix du format et du contenu des titres, collectivement ou individuellement. Les histogrammes peuvent être surperposés et mettre en parallèle deux distributions. Enfin ils peuvent être imprimés ou **transférés** dans un fichier, via le presse-papier.

2 - La page Liste dispose de nouvelles fonctions relatives à la **longueur des mots** ou aux **classes de fréquences**. La sélection s'étend maintenant aux formes d'un même **vocable**, regroupées automatiquement. Des présélections de **mots-outils** sont disponibles en permanence (voir la catégorie des prépositions dans la figure 8). Enfin le programme se prête plus souplement aux diverses manipulations qu'on peut faire en jouant sur le **regroupement** ou la **séparation** des mots.

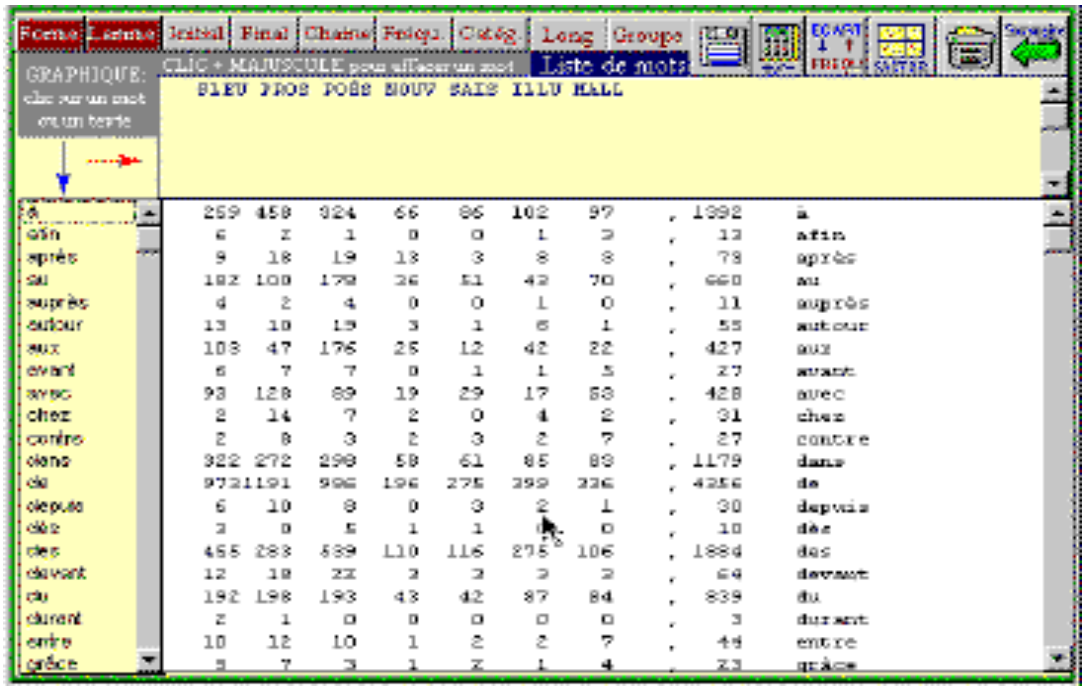

Figure 8. La page Liste (avec un corpus de travail réduit à 7 textes)

3 - Peu de changements sont à noter dans l'**analyse factorielle**, car il s'agit d'un programme extérieur dont le code ne nous appartient pas. Ce programme vénérable, écrit il y a vingt ans en Fortran, est très rapide mais l'interface laisse à désirer. On a eu recours à quelques aménagements pour améliorer la **lisibilité** des résultats: l'explicitation des points ne se limite plus à quatre lettres et les variables peuvent s'écrire en entier si elles n'en recouvrent pas une autre. De plus quand trop de points encombrent le graphique, une police plus petite est substituée dont l'effet est plus intéressant sur l'imprimante.

4 - L'innovation principale est à découvrir derrière le bouton *Grammaire*. On est renvoyé à une page nouvelle qui rend compte de la distribution des **catégories grammaticales**, y compris les classes ouvertes comme les adjectifs, les substantifs et les verbes, dont le recensement implique une véritable **lemmatisation**. Tout le corpus a effectivement été lemmatisé dans le texte d'origine, même si la trace du code grammatical n'est pas visible dans le texte restitué à l'écran. Car de tels codes sont assez opaques à la lecture et ils alourdissent considérablement le corpus en en doublant la taille. Mais l'essentiel de ce traitement (emprunté à *Winbrill*) est conservé dans la page *Grammaire*, qui autorise courbes et analyses factorielles appliquées aux parties du discours. Ainsi peut-on approcher les phénomènes linguistiques qui relèvent de la syntaxe.

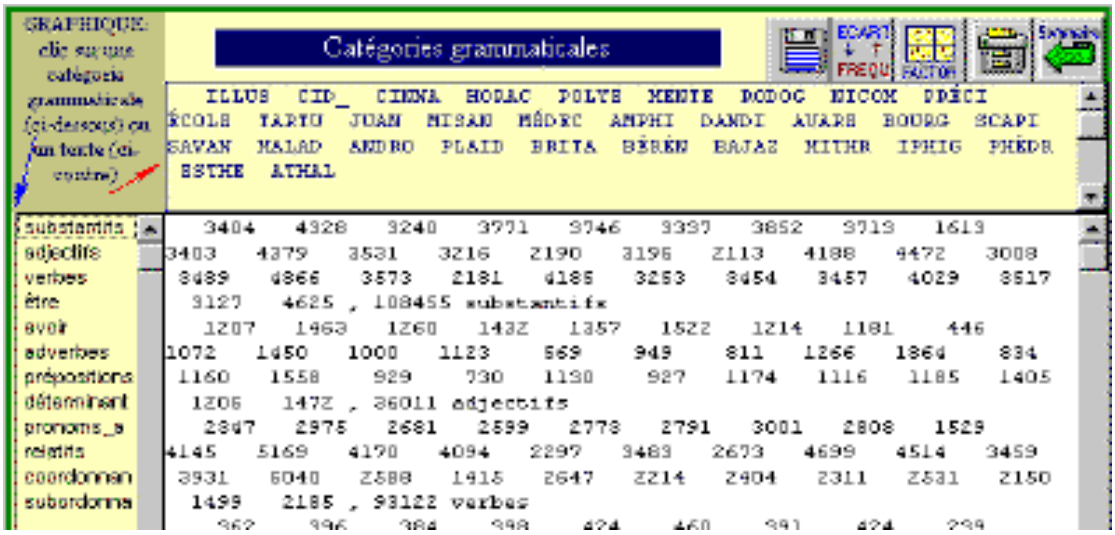

Figure 9. La page Grammaire

Comme dans la carte Liste, les histogrammes s'obtiennent sur les lignes ou sur les colonnes. Dans le premier cas - quand l'on clique sur un élément de la marge gauche - le profil est celui d'une catégorie, dont on suit la distribution à travers les différents textes du corpus. Dans le second - cliquer sur un texte de la marge supérieure - le programme dessine le profil d'un texte à travers le dosage des parties du discours qu'on observe dans ce texte.

Les bases externes

1 - Les monographies d'écrivains

La base Batelier contient tous les textes et peut se suffire à elle-même. Cependant on a jugé utile de la livrer aussi en monographies détachées, construites autour d'un écrivain. En cliquant sur l'une des vignettes portant l'effigie d'un écrivain, on ouvre la monographie correspondante. Le texte est le même que celui de la base principale, mais la segmentation peut être différente. Ainsi la division des Essais de Montaigne en trois livres n'est pas assez fine pour permettre une exploitation statistique. On lui a substitué une division moins grossière en 19 sous-ensembles où l'Apologie de Segond trouve sa vraie place. Il en est ainsi de Pascal, et de Proust.

2 - Création de bases nouvelles

Le logiciel Hyperbase dans sa version standard n'a pas de données propres. Il les reçoit de l'usager sous la forme d'un fichier texte (ou ASCII), et en assure l'indexation et le traitement statistique pour constituer une base hypertextuelle. On n'a pas cru devoir refuser à l'utilisateur de Batelier cette possibilité de créer des bases parallèles, avec les données de son cru ou de son choix, à charge pour lui de constituer le fichier des données dans le format attendu (consulter le manuel). Les créations nouvelles se justifient cependant lorsqu'on veut soumettre à la statistique un texte unique (il est alors divisé en 9 parties de longueur voisine), lorsqu'on souhaite éviter l'emploi du cédérom sur lequel Batelier est installé, ou, bien entendu, quand le texte à traiter n'est pas dans Batelier. Ci-dessous la page qui est affichée au moment de la création et où l'on peut contrôler et diriger le progrès des opérations (figure 10).

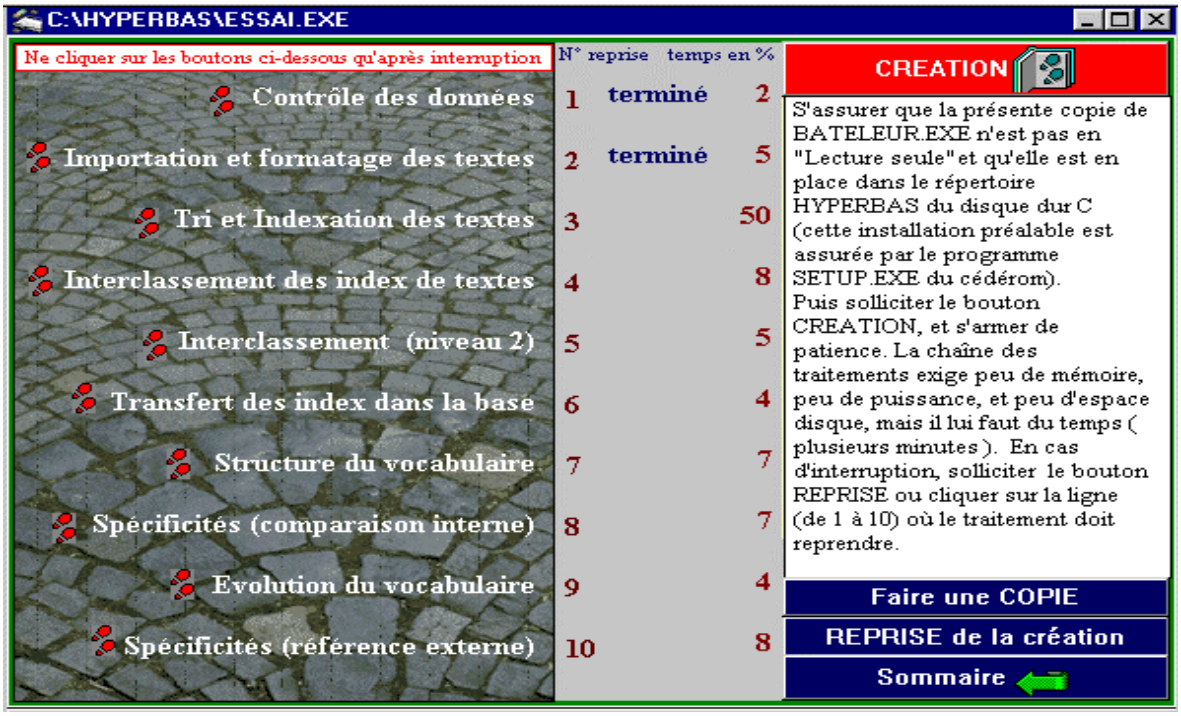

Figure 10. Création d'une base nouvelle

3 - Les bases transversales

Dans son état actuel de prototype, la base Batelier n'est ni assez étendue, ni assez représentative, pour permettre l'étude généralisée de la littérature française. Avec une douzaine d'auteurs seulement, elle ne peut donner qu'une idée très partielle des genres littéraires, des époques et des écoles, et bien évidemment des écrivains qui n'ont pas encore été retenus dans la base. À terme sont prévus des regroupements selon les genres ou les siècles.

a - Les genres littéraires

On disposera ainsi d'une base romanesque, d'une base de poésie, d'une base de théâtre, etc... Afin de donner un aperçu de cette exploitation des genres, on a constitué en corpus les textes du théâtre classique. On peut y accéder en activant le bouton Genres du menu principal, qui à son tour conduit à la base désirée. Précisons que les bases transversales, pour éviter des doublons inutiles, sont dénuées de texte et que les recherches documentaires n'y ont pas cours. Les informations nouvelles qu'elles donnent sont comparatives et quantitatives.

b - Les époques. La base Chrono

La base Batelier est ordonnée chronologiquement, de Rabelais à Proust. On peut certes y suivre l'évolution d'un mot, mais la courbe obtenue a trop de sauts et de lacunes et elle est trop dépendante de la minceur de l'échantillon. Frantext offre une meilleure assise pour ce type de recherche historique. Mais on peut y craindre le défaut inverse : trop de textes, trop de genres s'y trouvent mêlés et l'évolution, si on la constate, doit pouvoir être isolée des influences parasites. On a rendu le corpus plus homogène en écartant les textes techniques pour ne conserver que la littérature. Cela représente encore une masse énorme, de 117 millions d'occurrences. La figure 11 illustre ce qu'on peut attendre de cette base quand on croise deux mots, ici les mots base et fondement, dont la fortune historique s'oriente inversement.

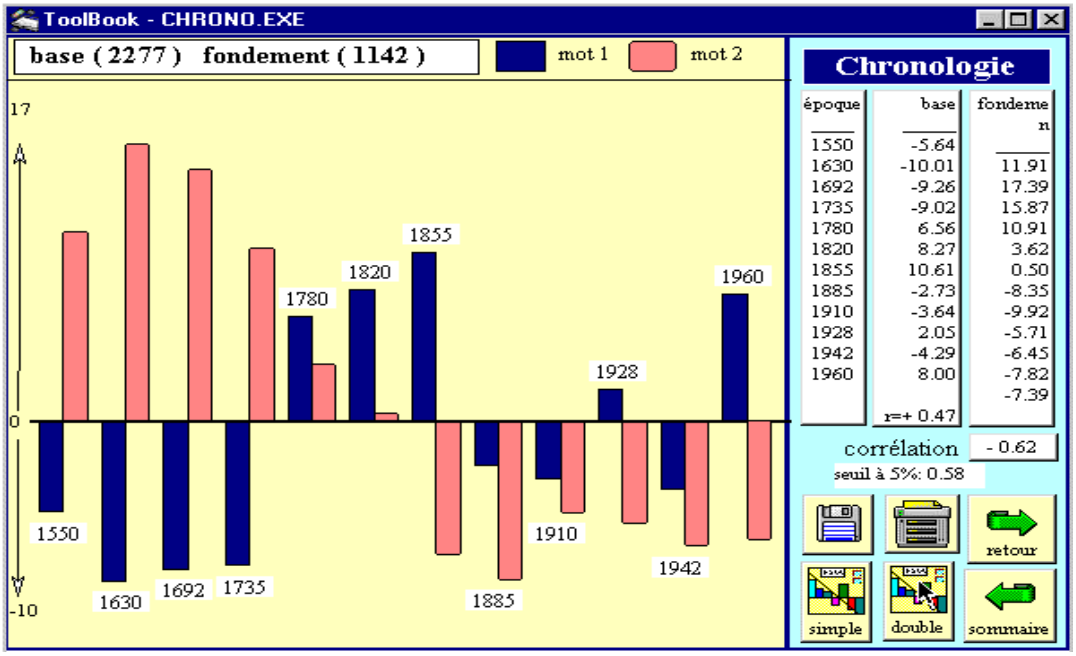

Figure 11. La base chronologique issue de Frantext

c - Les écrivains. La base Auteurs

Batelier passe en revue une douzaine d'auteurs, au stade préparatoire qui est le sien, alors qu'on en compte des centaines dans Frantext. Là encore c'est à Frantext qu'il faut s'adresser si l'on veut avoir une vue d'ensemble du paysage littéraire, région par région, auteur par auteur. Comme on ne traite ici que des données quantitatives, le copyright n'est pas embarrassant et les écrivains modernes ont été pris en compte, jusqu'à Gracq. On a ainsi réuni 70 écrivains sans dépasser le 17e siècle (le premier de la liste est Honoré d'Urfé). Au total cette base Auteurs enveloppe un corpus considérable de 56 millions de mots (et 236 000 formes différentes). Au risque de la déflorer, on en illustrera la richesse en dressant la carte des écrivains selon la distance lexicale où chacun s'établit au regard des autres. Ce programme de "connexion lexicale" fait entrer tous les mots dans le calcul, et son interprétation jouit d'une grande lisibilité: c'est le temps qui parcourt l'espace de droite à gauche (du XVIIe au XXe siècle) et c'est le genre qui oppose le haut (la prose) et le bas (la poésie).

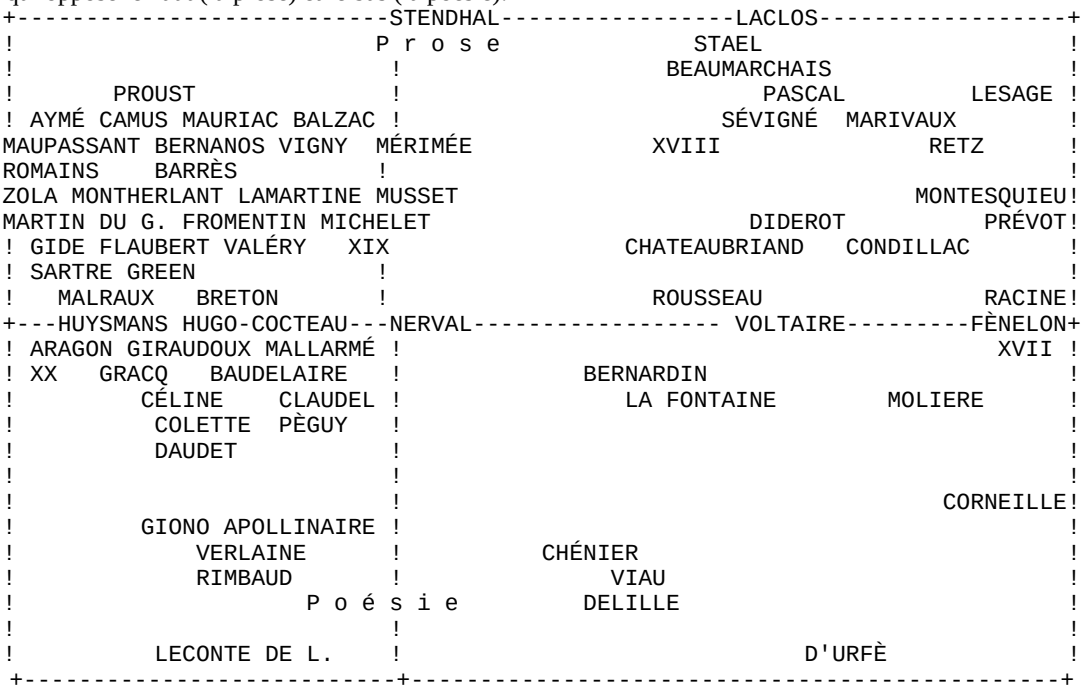

Figure 12. Analyse factorielle de la distance lexicale dans la base **Auteurs**

d - Lieux et milieux. La base *Francil*

Une dernière base enfin est proposée pour donner réponse aux questions que l'on peut se poser à propos des diverses variétés du français et des variables mises en jeu dans son exercice. Ces variables divergent autant que les populations qui partagent l'usage du français et que divers facteurs peuvent opposer: l'espace géographique, le temps historique, les conditions sociologiques, l'environnement économique, politique et culturel, sans compter le tempérament, les goûts et les choix personnels des écrivains. S'y superposent les variables proprement linguistiques qui opposent l'oral à l'écrit, l'information à la fiction, l'utilitaire au littéraire, . Pour tenter de maîtriser toutes ces variables et en dénouer les fils invisibles et entremêlés, on a étendu le champ de l'enquête à l'ensemble de la francophonie. Les données ont été empruntées aux observatoires du français établis au Québec (*Québétext* et *Catiq*), en Belgique (*Valibel* et *Beltext*), en Suisse (*Suistext*), en France et en Afrique (*Gars* et *Frantext*). Certaines données relatives à la presse ou à l'oral ont été recueillies expressément en vue de ce programme *Uref/Aupelf*. Au total le corpus recouvre 4,5 millions d'occurrences, où le français parlé est confronté à l'écrit, le littéraire à l'utilitaire, et le nord au sud. Les variables en jeu peuvent s'ajouter, se neutraliser, ou rester indépendantes. Comme la mer où s'exerce la force conjuguée ou contrariée des vents, des courants, des marées et des obstacles, le langage obéit à la mécanique des fluides, et la statistique a fort à faire pour en rendre compte.

On le voit, un air de famille et des relations de parenté lient les deux bases généralistes de l'Institut National de la langue française. La base naissante a besoin, pour longtemps encore, de l'appui de l'aînée. Nul besoin d'imaginer que l'une puisse se substituer à l'autre au fil des ans, comme se succèdent les générations. Car loin de décroître *Frantext* est une entreprise qui monte et s'élargit, en volume et en qualité. Les nouvelles fonctionnalités qu'on peut expérimenter présentement sur des données étiquetées sont d'une puissance inégalée, et l'audience de Frantext est appelée à se développer quand le réseau Internet sera plus familier aux français et que les limitations dues à la nécessité de s'abonner auront disparu - ce qui est déjà acquis pour l'expérimentation dans les lycées. *Batelier* ne vise pas à s'installer à la place de *Frantext*, mais à côté de *Frantext*. Il s'agit d'occuper les niches où *Frantext*, gêné par son poids, peut difficilement s'introduire. Ainsi voit-on certaines PME prospérer dans les créneaux vides que les multinationales laissent entre elles. Pour l'instant la voie libre, sur laquelle les grandes maisons d'édition ont hésité à se lancer, au moins dans le domaine littéraire, est celle du cédérom. Ce support, dont on connaît les limites (mais le DVD permet déjà de les dépasser) est, par son prix, sa fiabilité et sa facilité d'emploi, particulièrement bien adapté aux populations scolaires et universitaires auxquelles *Batelier* prête son vaisseau pour un voyage hypertextuel dans la littérature française. Au moment où s'amorce la première traversée, il reste à souhaiter bon voyage à l'équipage et aux passagers.