La base textuelle Batelier

Étienne Brunet

La base textuelle *Batelier*[1]

Etienne Brunet

Ceux qui s'attendent à voir le directeur de l'*Institut National de la Langue Française*, ceux qui se réjouissent d'entendre Bernard Cerquiglini seront déçus. Bernard Cerquiglini est ailleurs, sans doute sur un autre continent, peut-être une autre planète et des moyens de communication trop lents ne lui ont pas permis d'être ici et maintenant devant vous. Et l'on devra se contenter d'un remplaçant de fortune, qui ne sait trop quel sujet traiter puisque le premier programme prévoyait un exposé sur *Frantext* et le second une synthèse plus large sur les bases textuelles. Je devine toutefois la raison de ce changement : depuis quelques mois en effet un nouveau projet de base textuelle étendue a vu le jour à l'INaLF qui ne se confond pas avec *Frantext* et qui sous le nom de *Batelier* représente une des opérations d'envergure dont l'INaLF a besoin pour envisager l'avenir et justifier sa sérénité ou sa survie.

1. *Frantext*

1- Un mot tout de même sur *Frantext* qu'on ne devrait plus avoir à présenter vu son importance et son ancienneté et qui est encore méconnu des milieux scientifiques. On en connaît le contenu : quelque 180 millions de mots disponibles en permanence sur *Internet*, soit près de 3 000 textes de notre littérature nationale, de Rabelais à nos jours. Au moment où ces textes ont été dépouillés, il y a 30 ans (à l'époque on disait *perforés*, car cela se faisait sur bande ou sur carte perforée), on n'appelait pas cela base textuelle. C'était seulement un répertoire d'exemples à l'usage quasiment exclusif des rédacteurs du *Trésor de la Langue Française*. Je crois avoir été l'un des premiers bénéficiaires de ces données, au début des années 70, pour une utilisation non purement lexicographique, que le recteur Imbs avait cru devoir favoriser. Mais c'est Quemada qui a donné son plein essor à cette activité plus lexicologique que lexicographique et c'est le sens qu'il faut donner à la création de

[1] Communication présentée au colloque *Les trois révolutions de l'imprimerie*, 16-21 nov 1998, Bibliothèque municipale de Lyon.

l'*Institut National de la Langue Française* à côté du Trésor. Et l'un des colloques importants qui a marqué l'histoire de ce laboratoire, et un peu l'histoire de la linguistique française, est celui qui eut lieu à Nancy à la fin des années 70 avec pour thème – déjà – le sujet d'aujourd'hui : bases ou banques textuelles.

2- En quelques années, le projet est devenu réalité, grâce aux progrès techniques réalisés dans les supports de masse et dans les réseaux. Je passerai vite sur les premières étapes qui ont emprunté les canaux de la première heure, *Transpac* et le *minitel*, et les débuts d'*Internet*. Le logiciel d'interrogation créé par Jacques Dendien sous le nom de *Stella* distribuait sa lumière avec quelques éclipses, à cause des trous de la communication. Et pour traverser l'Atlantique, on avait cru bon d'avoir une tête de pont à Chicago qui distribuait les mêmes informations à partir d'un site miroir. À l'heure actuelle, la nouvelle version de *Stella* emprunte le canal du *Web*, dont l'usage est familier à tous les chercheurs, même littéraires. Pour ceux qui en ignoreraient encore les vertus, une démonstration est assurée par son créateur dans une salle attenante, à côté d'une autre base, qui n'est encore qu'un prototype et dont je vais parler maintenant.

3- *Frantext* se développe toujours : une version plus conviviale est en expérimentation et les textes s'amoncellent, surtout aux deux bouts de la chaîne : aux 16e et 20e siècles. Certes il n'est pas très difficile actuellement de réunir des monceaux ou des montagnes de texte : *Internet* en offre abondamment et des centaines de corpus ont été constitués au hasard des thèses. Le problème est qu'on ne peut pas accepter n'importe quoi. Quoique les fautes ne soient pas difficiles à trouver dans la meule énorme de *Frantext*, il faut reconnaître cependant d'autres qualités à cette base que sa taille : de ce côté, ce qui a été un record en linguistique risque de ne pas le demeurer. Il est très facile avec les journaux ou *Internet* d'élever des terrils de texte plus hauts que *Frantext* et les pyramides d'Égypte. Ce qui importe avant tout, c'est la qualité et l'homogénéité. Or *Frantext* a mis parmi ses objectifs la qualité littéraire des textes de sa base (le technique ne représentant que 20%). Quant à l'homogénéité, elle conditionne l'exploitation ultérieure : si les mêmes principes de dépouillement et de traitement ne sont pas assurés, que valent les comparaisons ? Il faut reconnaître qu'en trente ans les progrès techniques auraient pu justifier l'assouplissement et même l'amélioration des consignes de saisie. On a résisté à cette tentation. Avec raison, car on aurait perdu ainsi la compatibilité et ruiné d'un coup le stock ancien.

2. *Batelier*. Objectifs

1- Vient un moment pourtant où l'évolution oblige à infléchir la courbe. Face à la concurrence, face aux besoins nouveaux de la communauté scientifique, l'objectif est de proposer et de constituer une base nouvelle plus pure, plus stricte et, par certains aspects, plus riche. Il faut dire qu'il y a trente ans, le dépouillement des textes n'obéissait qu'à un impératif pratique : fournir des exemples aux rédacteurs. La qualité de l'édition importait assez peu pour les courts extraits retenus. Et les coquilles qui subsistaient dans les données pouvaient aisément être corrigées par le rédacteur. Mais une base textuelle créée dans ce seul but a nécessairement des critères plus stricts et plus sévères.

– D'une part, on veut exercer plus de rigueur dans le choix des éditions. Plutôt que d'examiner dans chaque cas pour chaque texte les éditions en présence, il a paru préférable de s'en tenir toujours au même principe et de choisir le dernier texte corrigé par l'auteur. Si pareille édition n'est pas disponible, on la réalise sur les documents originaux. Comme l'INaLF n'est pas un atelier de saisie, c'est une maison d'édition privée qui a été associée à cette tâche, en l'occurrence les éditions Champion.

– D'autre part, on a plus d'ambition dans l'étendue des corpus enregistrés. L'utilisateur apprécie peu d'être soumis au choix, toujours arbitraire, du créateur de la base. On tâchera donc de lui offrir, autant que faire se peut, des intégrales, des œuvres complètes. Bien sûr tous les écrivains ne seront pas présents d'un coup dans la base mais ceux qui le seront seront dignement représentés.

– En outre, le codage et la préparation des textes doivent permettre une exploitation sophistiquée. Des normes de plus en plus répandues ont acquis un statut de standard : en particulier le codage SGML et les spécifications plus précises encore du groupe TEI (*Text Encoding Initiative*). Les textes enregistrés doivent contenir un code de ce type, même simplifié.

– Enfin la base doit se prêter aux interrogations qui portent sur les structures grammaticales et qui mettent en œuvre non plus les formes et les mots individuels mais des critères syntaxiques et sémantiques. Ici s'impose la nécessité de la lemmatisation. Or à l'INaLF deux programmes de lemmatisation ont été expérimentés ou créés de toutes pièces : le premier qui vient du MIT, a été adapté au français par Josette Lecomte et Gilles Souvay, sous le nom de *Winbrill*, l'autre qu'on

commence à appeler *Maupin* (car réalisé par Maucourt et Papin) est plus prometteur et plus efficace pour le français, mais n'est pas implanté encore sur les systèmes courants, notamment sous *Windows*.

2- Cette base qu'on veut nouvelle et qui ne se substitue pas à *Frantext* non plus qu'aux bases mises en place par la *BNF* a un nom et un commencement d'existence. Pour passer d'un siècle à l'autre et d'une technologie à l'autre, on a choisi un mot de passe, un mot de passeur. C'est donc *Batelier* (base des textes littéraires pour l'enseignement et la recherche) qui nous transportera d'une rive à l'autre. On remarquera que la pédagogie est présente dans le titre même de la base et que le ministère de l'Éducation Nationale est partie prenante dans cette opération. Car *Batelier* doit être une base disponible non seulement dans les universités mais aussi dans des lycées et collèges (comme aussi *Frantext*). La distribution de masse qu'on envisage impose ainsi au produit les règles habituelles de la pédagogie et de l'ergonomie, dont la première est de plaire en étant simple. Or pour plaire, le batelier doit se faire un peu bateleur, sans craindre de donner au produit un caractère attractif et presque ludique, où les chiffres sont associés aux lettres comme dans un jeu célèbre de la télévision. Car la population scolaire à laquelle le produit est destiné reçoit quotidiennement des cours de mathématique et des leçons de français. Elle est plus réceptive que celle des aînés aux exercices où ces deux disciplines sont conjuguées. Au delà des années de lycée, pareille gymnastique est difficilement praticable. Les littéraires ont oublié les formules, et les scientifiques n'ont plus guère le loisir de s'intéresser à la littérature et au langage. Au collège, le divorce des deux cultures n'est pas encore consommé. Il faut en profiter.

3. *Batelier*. Choix du corpus

Ayant reçu de son responsable (François Rastier) la charge du logiciel à mettre au point pour de telles données, je prends le risque de montrer, pour la première fois, le prototype qui doit être remis dans un court délai au Ministère. Il s'agit d'un cédérom au standard *Windows* qui met en œuvre mon logiciel HYPERBASE appliqué à un ensemble de textes qui va de Rabelais à Proust et qui va s'enrichir dans les prochains mois. En voici le menu initial :

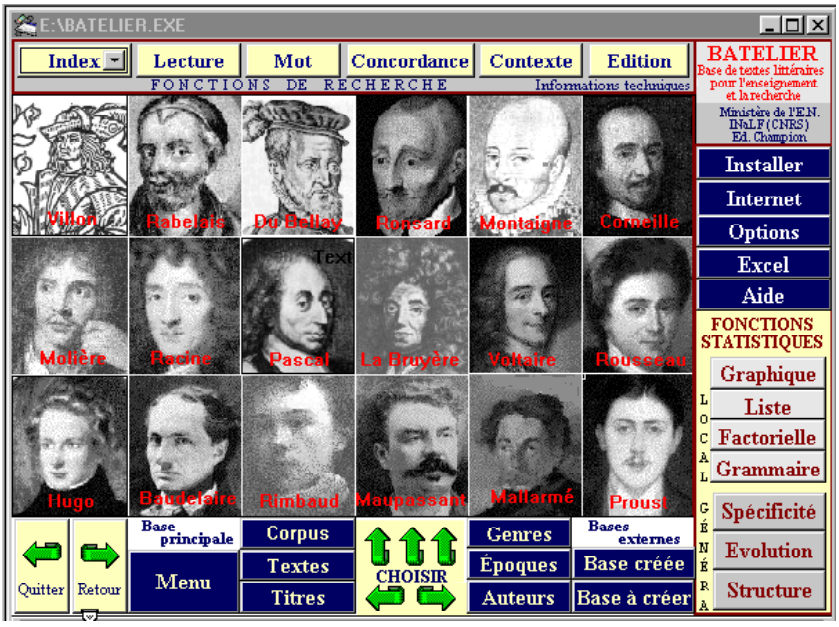


Figure 1. La base *Batelier*. Menu principal

Ceux qui connaissent le logiciel HYPERBASE reconnaitront 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.

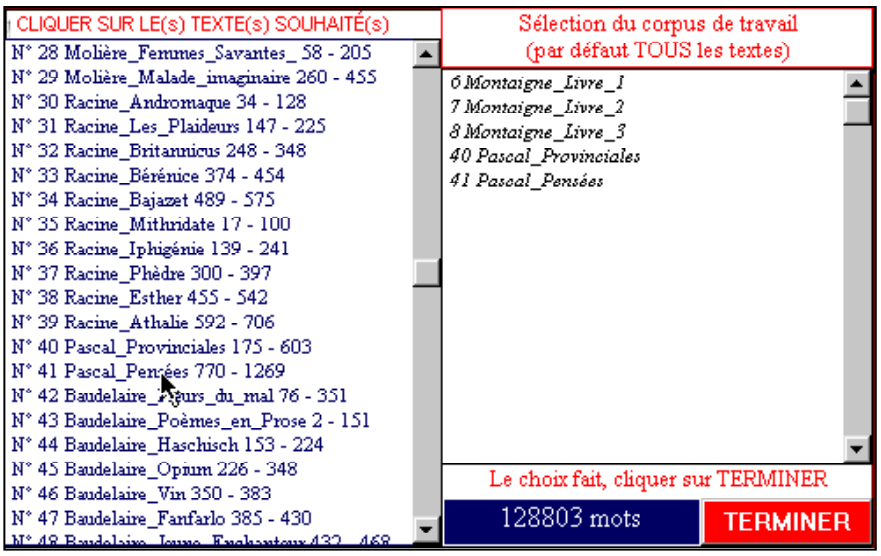


Figure 2. Choix du corpus de travail

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 *fonctions 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.

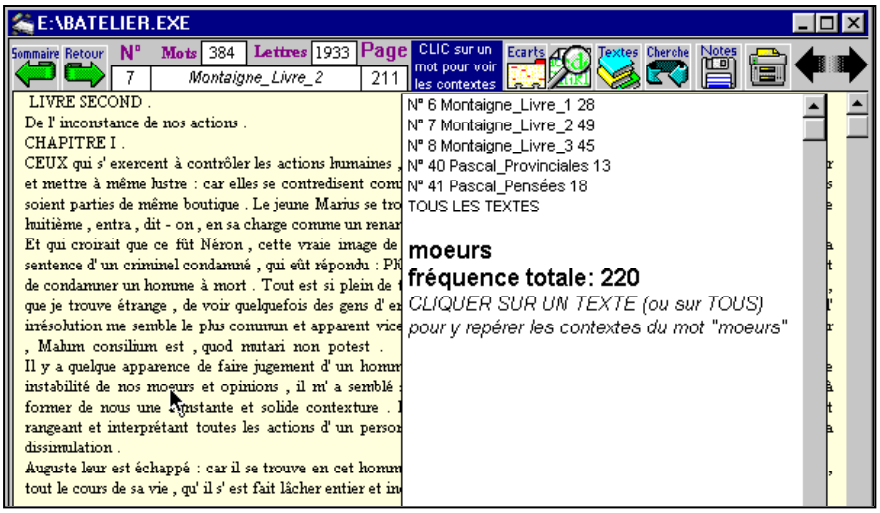


Figure 3. Une page de texte

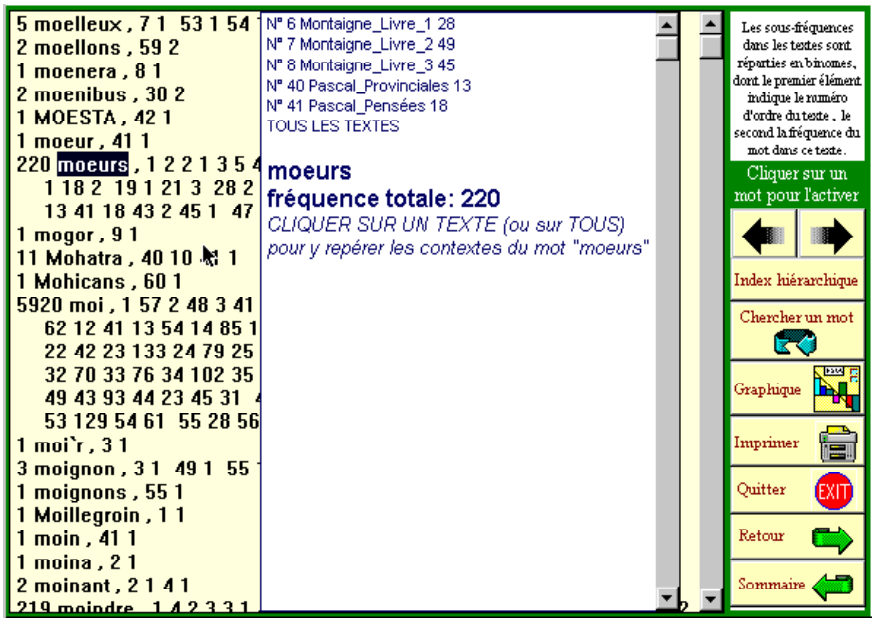


Figure 4. Une page de l'index

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, PASSE, TEMPS, ANNEES, ŒUVRE, qui sont récurrents dans le dernier roman de Proust (figure 5).

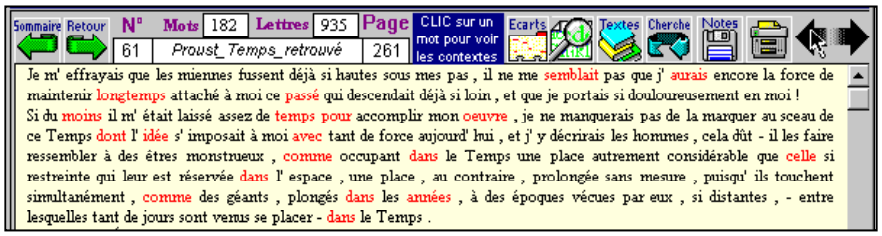


Figure 5. Les spécificités de la dernière page de Proust

4. Nouvelles fonctions de recherche

Les 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)

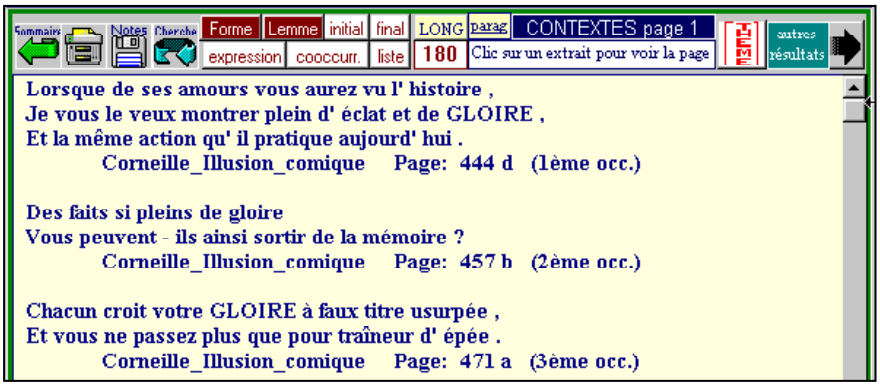


Figure 6. Résultats du programme *Contexte*

– 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 MEMOIRE 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, GENEREUX, ILLUSTRE, HONNEUR, VAINQUEUR n'excluent pas les antonymes SOUILLER, IGNOMINIE). Voir figure 7.

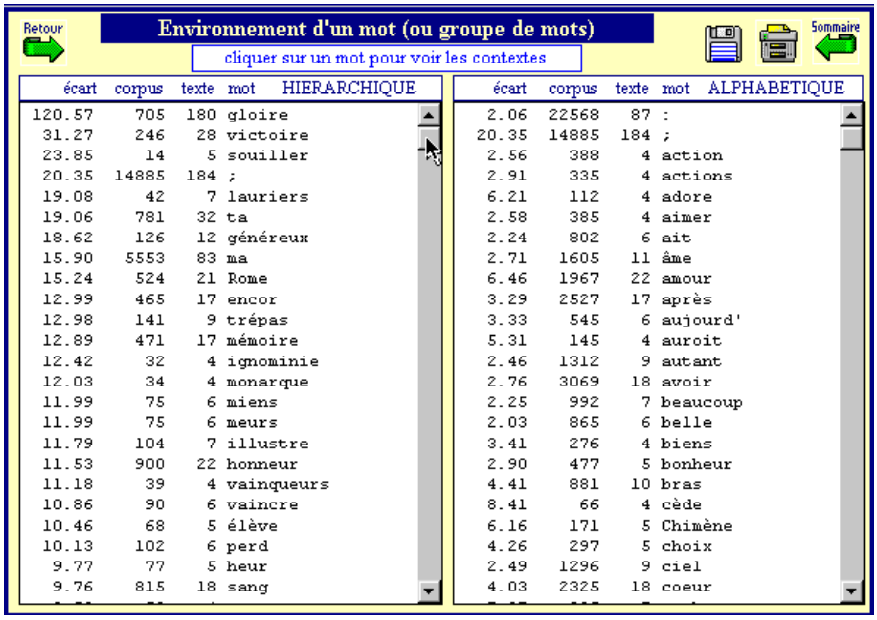


Figure 7. La fonction thématique. L'entourage du mot GLOIRE chez Corneille

5. Nouvelles fonctions statistiques

Les 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 superposés et mettre en parallèle deux distributions. Enfin, ils peuvent être imprimés ou *transférés* dans un fichier, via le presse-papier. Voir la figure 8 qui compare l'emploi des mots CIEL et AZUR chez Baudelaire, Rimbaud et Mallarmé.

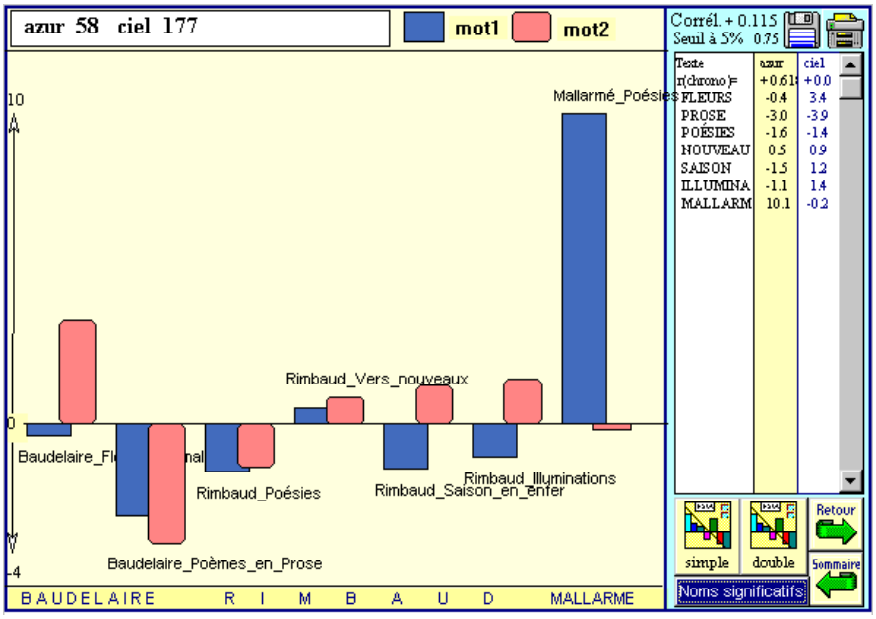


Figure 8. Histogramme de CIEL et AZUR

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 9). Enfin le programme se prête plus souplement aux diverses manipulations qu'on peut faire en jouant sur le *regroupement* ou la *séparation* des mots.

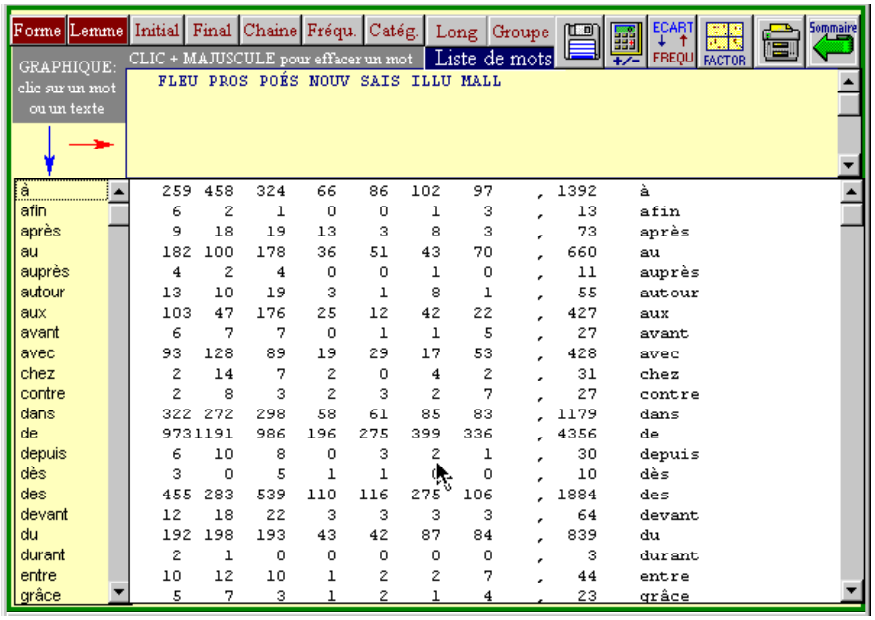
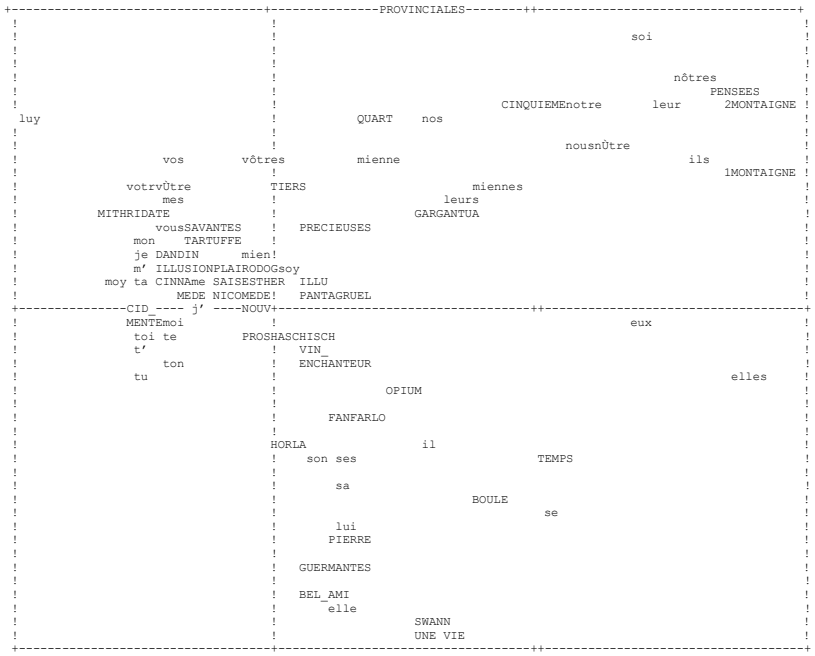


Figure 9. La page *Liste*

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.



NOMBRE DE POINTS SUPERPOSÉS : 24

3MONTAIGNE (leur) MISANTHROPE (vous) tes (mon) JUAN (mon) BOURGEOIS (mon) mien (mien) HORACE (mien) AMPHITRYON (je) AVARE (je) MALADE (DANDIN) IPHIGÉNIE (DANDIN) ÉCOLE (ILLUMINATIONS) BÉRÉNICE (ILLUMINATIONS) FLEURS (m') POLYEUCTE (me) SCAPIN (ta) ANDROMAQUE (CINNA) PHÈDRE (ta) POÉSIES (ESTHER) BRITANNICUS (MÉDECIN) BAJAZET (MÉDECIN) ATHALIE (NICOMÈDE) MALLARMÉ (NICOMÈDE) ma (moi)

Figure 10. Analyse factorielle des pronoms personnels (corpus entier)

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.

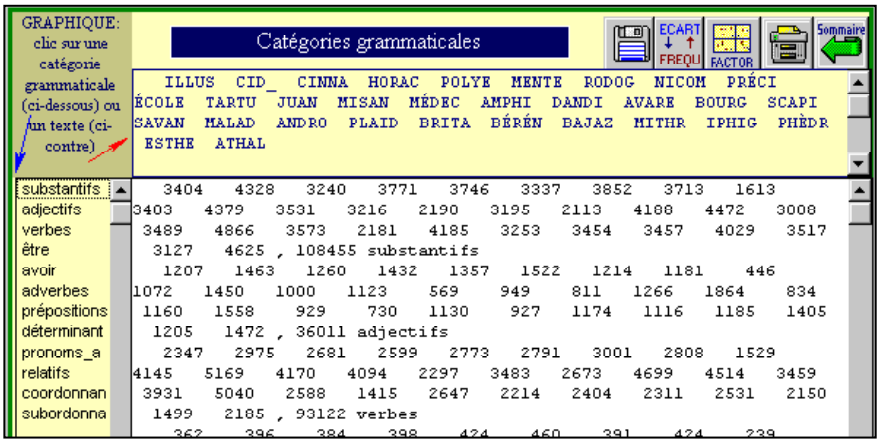


Figure 11. La page Grammaire.

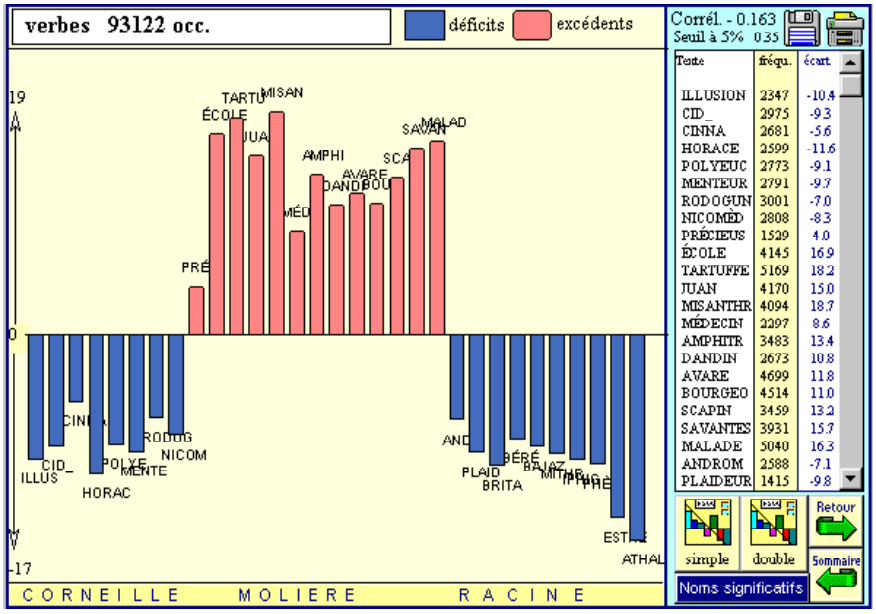


Figure 12. Les verbes dans le théâtre classique (ETRE et AVOIR exclus)

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. L'exemple de la figure 12 est de ce dernier type. On y voit Molière s'opposer à Racine et à Corneille relativement à l'emploi des substantifs. La différence des genres ne suffit pas à expliquer cet écart, puisque le *Menteur* et les *Plaideurs* ne rejoignent pas les comédies dans la zone des excédents.

6. 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. Même si l'œuvre de Mallarmé est mince (elle constitue un texte unique dans la base principale), elle a donné lieu à une segmentation en quatre sous-ensembles afin de donner du grain à moudre à la statistique.

Ces bases sont conçues à l'image de la base principale et partagent les mêmes fonctions. Comme elles sont orientées vers l'étude d'un écrivain, le corpus de travail ne peut être redéfini. Comme elles sont plus compactes, elles sont plus rapides. En outre elles bénéficient d'un traitement particulier de l'italique, qui n'a pu être maintenu dans la base principale, à cause de l'hétérogénéité de ce code typographique, quand on passe d'un auteur à l'autre. Chez le même écrivain, l'italique a une fonction plus aisée à cerner. Ainsi chez Montaigne (voir présentation ci-dessous), elle sert aux citations latines, de sorte qu'on a pu dans ce cas précis détacher le lexique français du lexique latin et ne pas confondre les homographes qui ont un sens dans une langue et un autre dans l'autre.

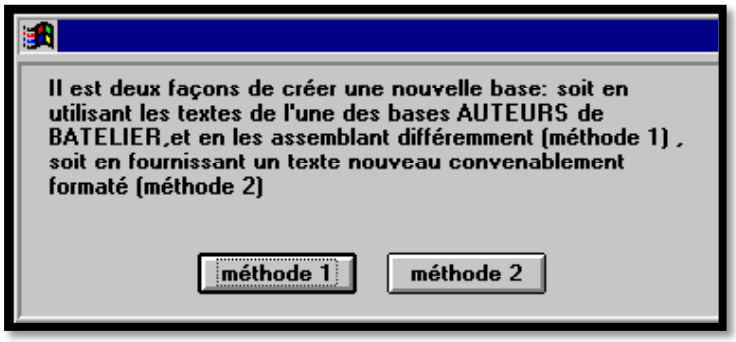
Figure 13. La base Montaigne

Lorsqu'une de ces bases est lancée à partir de la base principale, elle se maintient au premier plan tant que le bouton *Quitter* n'a pas été sollicité, après quoi la base principale reprend le contrôle. Mais on peut fort bien ignorer la base principale et s'introduire sans intermédiaire dans l'une de ces monographies, chacune étant autonome.

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).

Ce souci du formatage lui est épargné s'il choisit ses textes parmi ceux que contient la base *Batelier*. Il suffit de solliciter le bouton « Base à créer » et de choisir la première option du dialogue ci-dessous :



La liste des textes est alors offerte au clic de la souris. Quand le choix est clos, la base se construit sur le disque dur, dans le répertoire *C:\Hyperbas* , sous le nom qu'on est invité à donner. Noter que toute base nouvelle se fait par copie d'un modèle, *Hyperbas.exe* ou *Bateleur.exe* selon le cas, et que ces modèles ne doivent être ni changés, ni déplacés. Ils sont transférés sur le disque dur au moment de la première installation, en même temps que tous les fichiers nécessaires (programme *Setup.exe*).

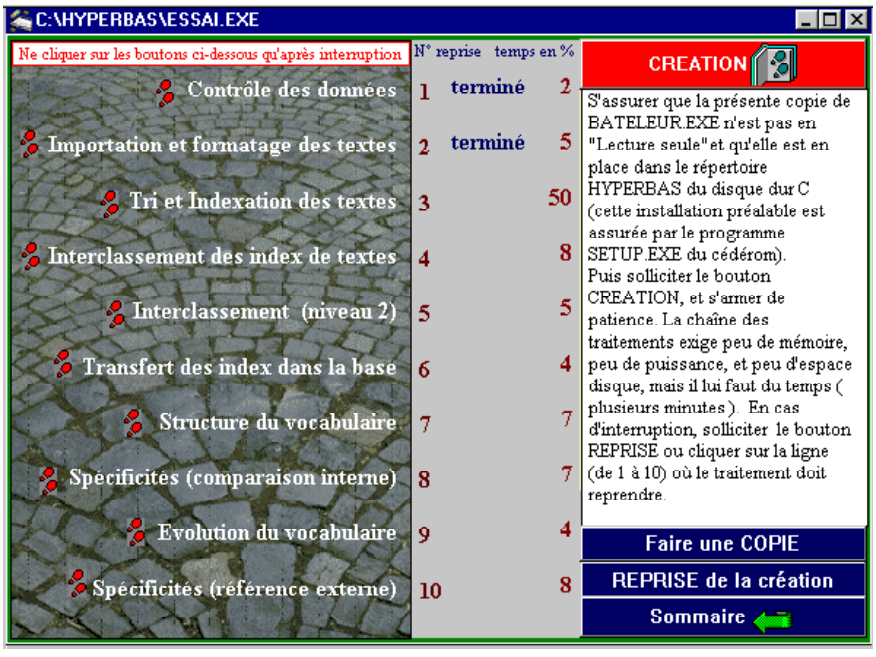


Figure 14. Création d'une base nouvelle

On n'abusera pas de cette procédure exceptionnelle de création, parce qu'elle coûte du temps (quelques minutes) et de l'espace (plusieurs *Mo* sur le disque dur). Dans la plupart des cas, la souplesse de la base *Batelier* permet de choisir toutes les combinaisons possibles. 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-dessus (graphique 14), la page qui est affichée au moment de la création et où l'on peut contrôler et diriger le progrès des opérations.

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 (fig. 16). 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. Elles sont d'autant plus précieuses que le genre littéraire est un champ de recherche encore peu exploité où l'on peut s'attendre à des surprises et des découvertes. L'analyse factorielle représentée ci-dessous (fig.15) en est une illustration. Elle analyse la distance qui sépare chaque texte de tous les autres quand on mesure pour chaque paire le rapport entre mots exclusifs et mots communs. La spécificité des trois écrivains y est excellemment soulignée puisque chacun accapare un coin du graphique. Mais la loi suprême du genre est respectée : le *Menteur* et les *Plaideurs* passent dans le camp de la

comédie, où les pièces en vers (en bas) se distinguent des pièces en prose (en haut).

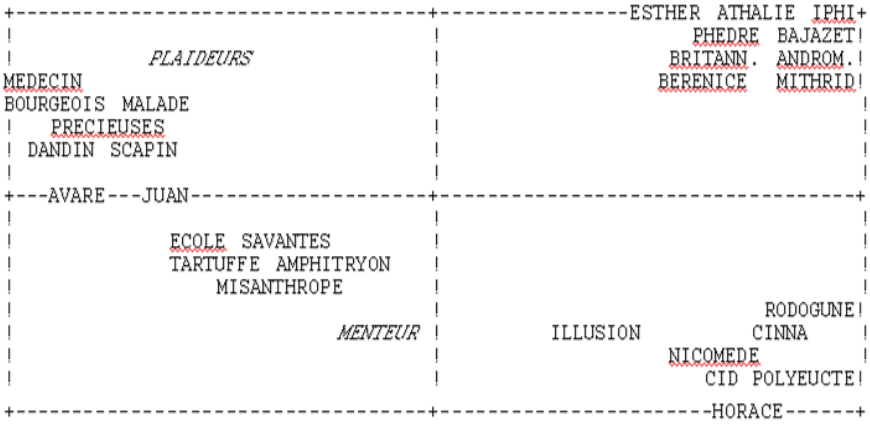


Figure 15. Carte des distances lexicales dans le théâtre classique

Figure 16. La base du théâtre classique

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. Le corpus constitué est celui qui était disponible sur le réseau en 1996, au moment où cette base chronologique a été réalisée, sous le nom de THIEF (*Tools for Helping Interrogation and Exploitation of Frantext*). La présente base est extraite de THIEF, dont elle a repris les fonctions *off line* (mais la consultation *on line* de *Frantext* est ménagée par un bouton de *Batelier*). La figure 17 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.

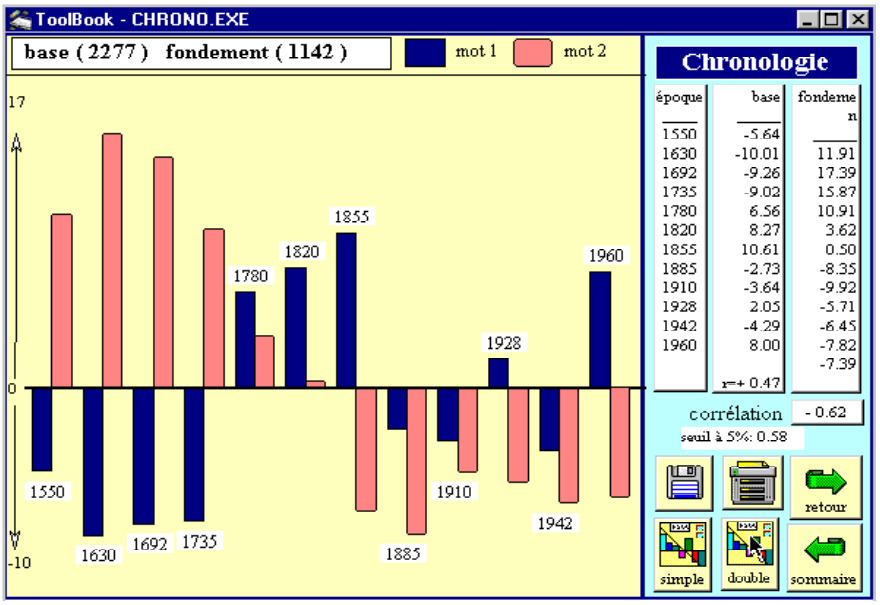


Figure 17. La base chronologique issue de Frantext

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). Car la base *Chrono* nous a averti que l'instabilité de l'orthographe au 16e siècle rendait les comparaisons délicates, d'autant que dans *Frantext* les textes anciens ne sont pas souvent modernisés (Rabelais et Montaigne l'ont été dans *Batelier*). La liste des auteurs retenus est affichée dans le menu principal de la base (voir figure 18). Un clic sur l'un d'entre eux fait apparaître la fiche bibliographique qui le concerne, à l'image de celle de Giono, livrée dans la figure 19 (en regrettant que le choix soit trop mesquin pour un auteur qui ne l'est pas).

Figure 18. La base *Auteurs* (56 millions d'occurrences)

La présente base est purement quantitative. Elle rend compte de l'usage comparé des mots chez les grands écrivains de notre littérature.Pour avoir accès aux textes qui ont servi à constituer cette base, consulter FRANTEXT.
LISTE des textes de Giono :

K629 GIONO.J/COLLINE/1929
PROSE,ROMAN
PARIS : GRASSET, 1929.

K628 GIONO.J/UN DE BAUMUGNES/1929
PROSE,ROMAN
PARIS : GRASSET, 1929.

K633 GIONO.J/REGAIN/1930
PROSE,ROMAN
PARIS : GRASSET, 1937.

K632 GIONO.J/LE GRAND TROUPEAU/1931
PROSE,ROMAN
PARIS : GALLIMARD, 1931.

Figure 19. Les textes de Giono présents dans la base

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 calcul de « connexion lexicale » est le même que celui de la figure 15. Son interprétation, alors que tous les mots sont entrés dans le calcul, offre la même lisibilité : c'est le temps qui parcourt l'espace de droite à gauche (du XVII^{e} au XX^{e} siècle) et c'est le genre qui oppose le haut (la prose) et le bas (la poésie).

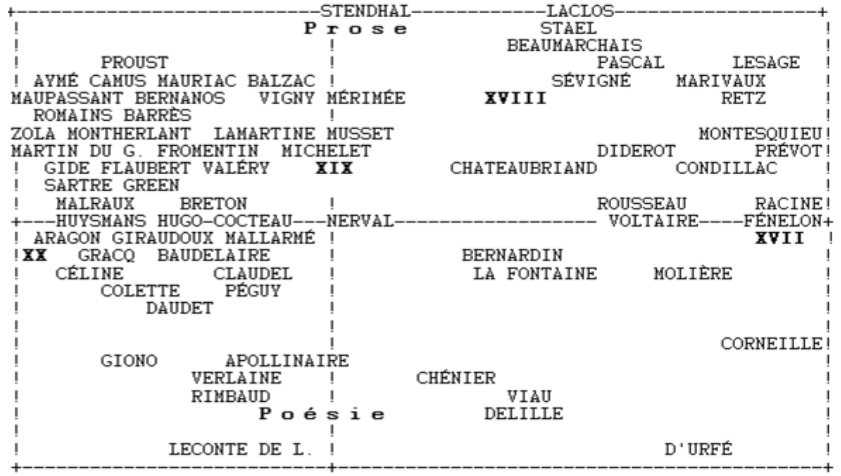


Figure 20. Analyse factorielle de la distance lexicale dans la base Auteurs

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. Verra-t-on l'une se substituer à l'autre au fil des ans, comme se succèdent les générations ? Rien n'est moins sûr. Car loin de décroître, *Frantext* est une entreprise qui monte et s'élargit et son audience 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. *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 au pilote et à l'équipage.