Balzac sur ordinateur

Étienne Brunet

Étienne Brunet
Institut National de la langue française

Balzac sur ordinateur

Il y a belle lurette que Shakespeare est disponible sur support informatique, bien avant l'avènement d'Internet et du CD-Rom. En proposant, il y a dix ans, sa machine *Next*, Steve Jobs n'avait pas craint de sacrifier un peu d'espace sur le disque dur pour y engranger l'oeuvre de l'écrivain anglais et faciliter la recherche des citations. En France, aucun auteur ne jouit d'une priorité reconnue sur tous les autres, comme Shakespeare en Angleterre, Dante en Italie ou Cervantès en Espagne. Et la littérature française, quoique fort bien représentée dans les données de *Frantext*, n'a accordé à aucun écrivain le privilège de l'exhaustivité, si ce n'est aux moins prolixes comme Rimbaud ou Mallarmé. Certes on y trouve en entier les *Mémoires d'outre-tombe*, les *Rougon-Macquart* et la *Recherche du temps perdu* mais l'intégrale n'est jamais proposée ni pour Molière (35 textes), ni pour Voltaire (56 textes), non plus que pour Hugo (44), Zola (45), Flaubert (55), ou Claudel (56). Le mieux représenté est actuellement Balzac avec 69 titres disponibles. Encore ne l'était-il pas à la date où s'est tenu le colloque sur les *Hypertextes littéraires.* Le catalogue de Nancy n'offrait alors que 22 textes de Balzac, dont 17 de la *Comédie humaine.* Et c'est pourquoi nous avions proposé en cette occasion de combler cette lacune et d'installer Balzac à la place qui semble devoir être la sienne : la première.

Les circonstances s'y prêtaient: d'une part nous disposions du texte entier de la *Comédie humaine.* Un chercheur de Tokyo, le Professeur Kazuo Kiriu, avec qui nous collaborions depuis longtemps, avait patiemment soumis au scanner les quelque cent textes de cet immense corpus. Une fois assurés tous les contrôles, ce chercheur généreux nous avait transmis le précieux dépôt, ainsi qu'à l'Institut National de la langue française. D'autre part au même moment un groupe de

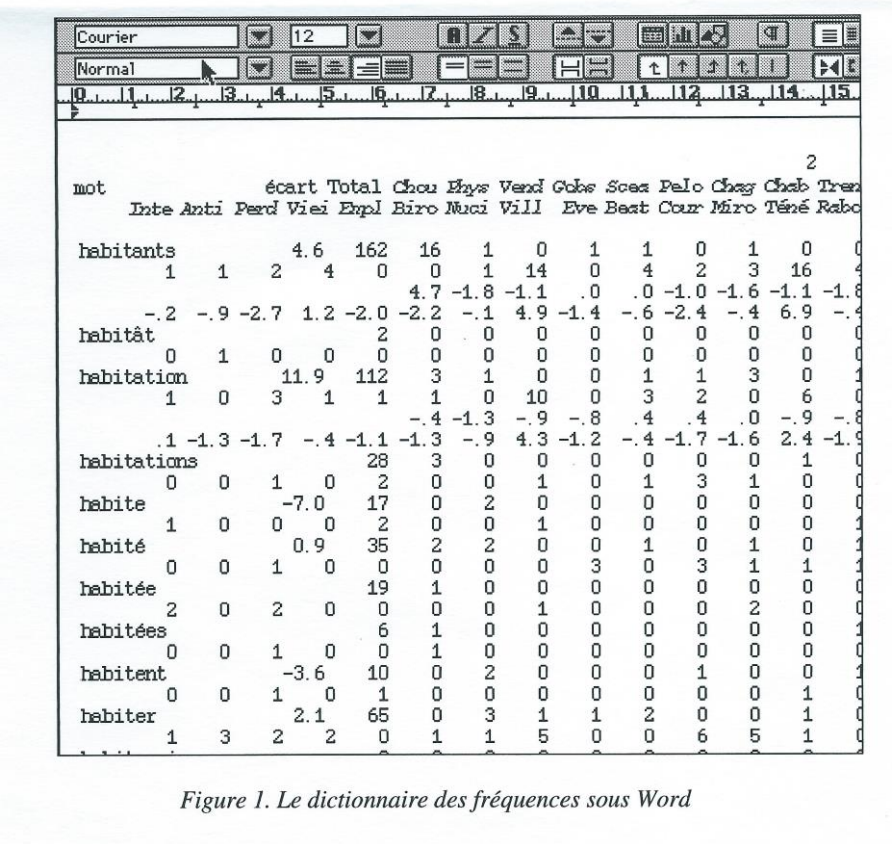
balzaciens (le GIRB) s'apprêtait à préparer dignement le bicentenaire de la naissance de Balzac, à l'initiative de Nicole Mozet, et songeait à produire un cédérom. Enfin précisément un cédérom littéraire venait d'être publié sur Rabelais et son temps, sous la conduite de Marie-Luce Demonet, et la partie logicielle de ce cédérom, qui nous avait été confiée, semblait pouvoir être adaptée à des projets similaires.

De là est né le prototype dont nous allons rendre compte. Disons tout de suite, à l'heure où nous écrivons ces lignes, c'est-à-dire presque trois ans après le colloque, que le projet a abouti à un cédérom commercialisé (éditions *Acamédia*, 1999) mais que sa forme définitive n'est pas celle qu'on avait initialement imaginée. Une raison essentielle à cela : l'édition livrée sur support informatique était celle de la *Pléiade*. Et des obstacles liés au copyright se sont immédiatement dressés, qu'on a contournés en faisant appel à une édition libre de droits, l'édition Fume, qu'au reste beaucoup de spécialistes préféraient, comme étant la plus proche du texte original corrigé par l'auteur. Mais c'était repartir de zéro et se condamner à une longue marche, dont je n'exposerai pas les étapes. Car je n'ai participé qu'aux premières, à un moment où plusieurs orientations se présentaient et où l'expérimentation d'un prototype pouvait être utile. Trois ans plus tard, est-il utile d'exhumer ce prototype ? Par définition, un prototype est un produit raté, un mort- né, un monstre dont les difformités font sourire, quand on a le recul de l'évolution. Mais aussi bien un prototype, sans être nécessairement un mutant, a pour objectif de faciliter les mutations, et de s'ouvrir aux idées qui germent plutôt que d'aboutir au produit fini. Il laisse des pierres d'attente dans la construction, des rameaux que la taille ménage. Et cela permet d'autres départs, d'autres bourgeonnements, auxquels l'avenir réserve une chance.

-1 -
Index et concordances sous Word

Il s'agit là de bois mort, que la sève a définitivement abandonné. Ceux qui veulent se contenter obstinément de leur traitement de texte peuvent souhaiter disposer d'un dictionnaire, d'un index, ou d'une concordance au format Word. Nous n'avons pas voulu les décevoir. Le dictionnaire est complet et restitue les sous-fréquences de chaque forme dans chacun des textes (avec comparaison externe et interne, et

calcul de l'écart réduit quand la fréquence le permet). La concordance est pareillement exhaustive, avec cette réserve que seules les 360 premières occurrences des mots très fréquents sont accompagnées de leur contexte. Même si les forêts ne sont pas menacées, quand on se sert d'un support autre que le papier, encore convient-il d'être économe de l'espace du disque et du temps du chercheur. Ceux qui ont l'habitude du papier ou des microfiches ne seront pas bousculés. L'ordre alphabétique les protège du vertige. Ils auront tout de même gagné en confort, en vitesse et en souplesse. Nul effort pour atteindre le tome utile ou décrypter la fiche nécessaire. Word les conduira gentiment à la lettre ou au mot souhaités, en proposant ses services d'adressage direct, de duplication ou d'impression.


Courier 12
Normal

mot				écart	Total	Chou	Rhys	Vend	Gobe	Sceu	Pelo	Chag	Chab	2 Trem
	Inte	Anti	Perd	Viei	Empl	Biro	Nuci	Vill	Eve	Beat	Cour	Miro	Téné	Rabo
habitants				4.6	162	16	1	0	1	1	0	1	0	0
	1	1	2	4	0	0	1	14	0	4	2	3	16	4
						4.7	-1.8	-1.1	.0	.0	-1.0	-1.6	-1.1	-1.8
	-.2	-.9	-2.7	1.2	-2.0	-2.2	-.1	4.9	-1.4	-.6	-2.4	-.4	6.9	-.4
habitât					2	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0	0	0	0	0	0
habitation				11.9	112	3	1	0	0	1	1	3	0	1
	1	0	3	1	1	1	0	10	0	3	2	0	6	0
						-.4	-1.3	-.9	-.8	.4	.4	.0	-.9	-.8
	.1	-1.3	-1.7	-.4	-1.1	-1.3	-.9	4.3	-1.2	-.4	-1.7	-1.6	2.4	-1.9
habitations					28	3	0	0	0	0	0	0	1	0
	0	0	1	0	2	0	0	1	0	1	3	1	0	0
habite				-7.0	17	0	2	0	0	0	0	0	0	0
	1	0	0	0	2	0	0	1	0	0	0	0	0	1
habité				0.9	35	2	2	0	0	1	0	1	0	1
	0	0	1	0	0	0	0	0	3	0	3	1	1	1
habitée					19	1	0	0	0	0	0	0	0	0
	2	0	2	0	0	0	0	1	0	0	0	2	0	0
habitées					6	1	0	0	0	0	0	0	0	1
	0	0	1	0	0	1	0	0	0	0	0	0	0	0
habitent				-3.6	10	0	2	0	0	0	1	0	0	1
	0	0	1	0	1	0	0	0	0	0	0	0	1	0
habiter				2.1	65	0	3	1	1	2	0	0	1	0
	1	3	2	2	0	1	1	5	0	0	6	5	1	0

Figure 1. Le dictionnaire des fréquences sous Word

De tels services auraient été fort appréciés il y a vingt ans. Ils correspondent à une démarche traditionnelle, qui est antérieure à l'informatique et qui s'avance timidement dans le champ hypertextuel. Mais il lui manque une chose essentielle : le texte lui-même. Ni les fréquences, ni les références, ni même un contexte réduit à une ligne, ne donnent l'accès immédiat au texte. Dans certains cas, où le respect du copyright interdit de montrer le texte, ce peut être paradoxalement un avantage, ou une garantie.

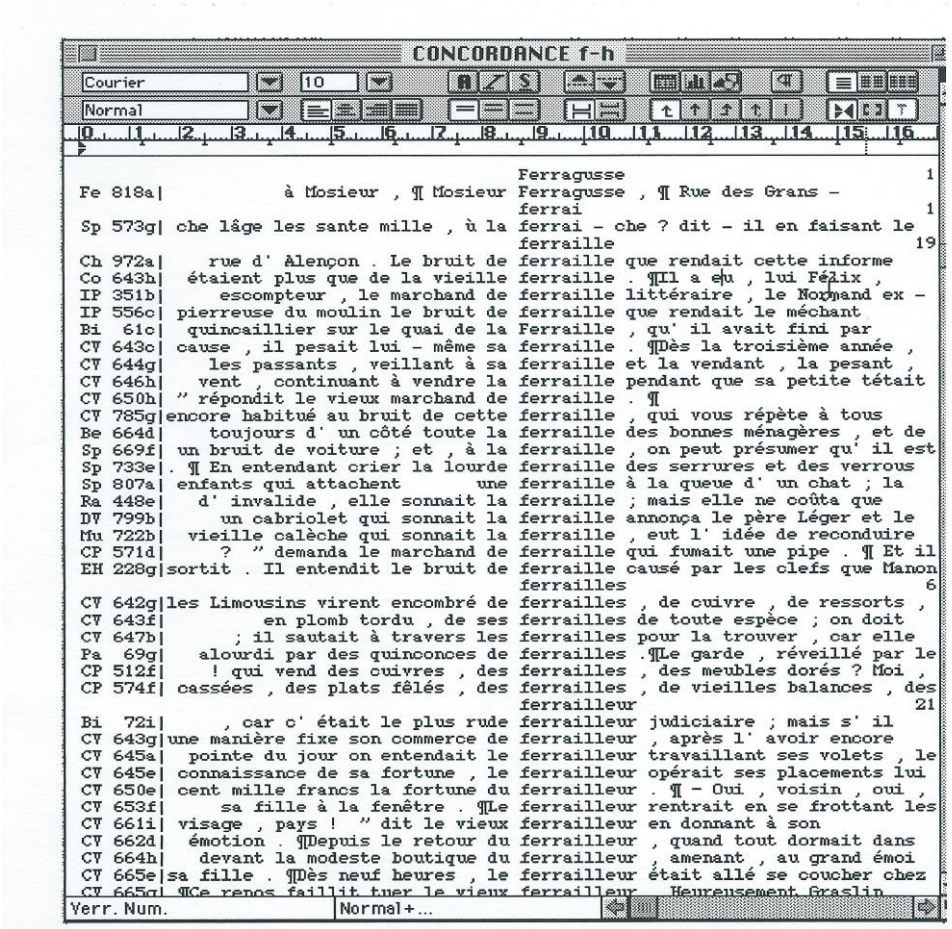


Figure 2. Une page de la concordance de Balzac sous Word

- II -

Balzac sur Internet

La même prudence à l'endroit du copyright doit s'exercer sur Internet. Certes il n'est pas difficile de glaner des textes sur les serveurs du Web, où la prudence prend souvent le visage de la dissimulation. Même si parfois on y trouve des intégrales intègres, comme celles de La Fontaine (http://www.lafontaine.net) ou de Maupassant (http://lib.univ-fcomte.fr/PEOPLE/selva/Maupassant.html), il s'agit rarement d'éditions critiques ou originales. La plupart sont issues d'éditions du commerce qu'on a dépouillées de tout signe distinctif et qu'on recycle en changeant la pagination, la présentation et les annotations (mais les fautes sont pieusement respectées par le scanner). Quant à nous, si nous n'avons pas hésité à corriger les coquilles, rencontrées très rarement il est vrai, nous n'avons pas cru devoir cacher nos sources et le texte de Gallimard est reconnu comme tel, avec sa pagination d'origine.

Car la collection de la Pléiade passe communément pour être la plus répandue des éditions de qualité. Et les références ne sont utiles que si elles renvoient à un ouvrage disponible.

Comment donc échapper aux poursuites ? L'ingénuité n'est pas l'innocence et il fallait offrir une garantie sûre. Il n'y en a pas sur un CD-Rom, parce que le produit une fois répandu échappe à tout contrôle et offre sa vulnérabilité aux violations des malins. Le créateur d'une base sur Internet est mieux armé contre les intrusions, car il est le maître du serveur, dont il peut changer les clés à sa guise, tandis qu'un espion le renseigne à tout moment sur les tentatives de viol et l'identité des sauvageons. Il suffit donc de ne livrer à la lecture qu'une page à la fois, en empêchant qu'on puisse accéder à la suivante ou à la précédente. La page qu'on montre a dès lors le statut d'une citation, dont la limite traditionnelle est acceptée autour de 300 mots.

Pour y parvenir, l'entrée est le mot, qu'on précise à l'aide d'index et de listes déroulantes (d'abord l'initiale, puis l'entrée la plus proche dans le dictionnaire). On est plongé alors au coeur de la concordance, dans la zone du mot choisi. On peut circuler en avant et en arrière et observer les contextes des mots que l'ordre alphabétique situe dans le voisinage et qui appartiennent à la même famille. Enfin un zoom donne accès à la page entière dès qu'un clic s'exerce sur une référence, dans la zone réservée à cet effet au début de chaque ligne de la concordance. Ainsi trois clics de la souris suffisent pour accéder à n'importe quel mot et n'importe quelle page de la *Comédie humaine*. Nul besoin de dérouler des menus, de cocher des cases ou de solliciter le clavier. La simplicité de l'interrogation tient au fait que la base ne recourt qu'au code HTML. Nulle trace de Java, ou de CGI. Pas le moindre calcul. Le moteur de recherche ne manipule que des liens (il y en a deux millions) et des pages HTML (on en compte 20 000). Toutes les informations sont préparées à l'avance et immédiatement disponibles, même les courbes et les résultats statistiques. L'inconvénient est l'impossibilité de répondre à des questions non prévues, comme les cooccurrences, dont la combinatoire est infinie. L'avantage réside dans la simplicité, la solidité, la rapidité et l'universalité. La base est indifférente au type de machine utilisé. Implantée sur un *Sun* sous *Unix*, elle tourne aussi bien sur un *Mac* ou sur un *PC*. Le support, disque dur ou cédérom, n'importe pas, non

plus que le mode d'interrogation, en local ou en réseau. La comparaison avec une autre base installée sur le même site avec des principes différents (base active *CGI* sur *Rabelais et son temps*) montre que le choix des utilisateurs est plus sensible à la robustesse qu'à la sophistication : le nombre de transferts de fichiers exécutés sur le Web y est dix fois supérieur (3000 par jour dans le cas de Balzac). Il est vrai que l'audience de Balzac sans le monde est plus forte que celle de Rabelais.

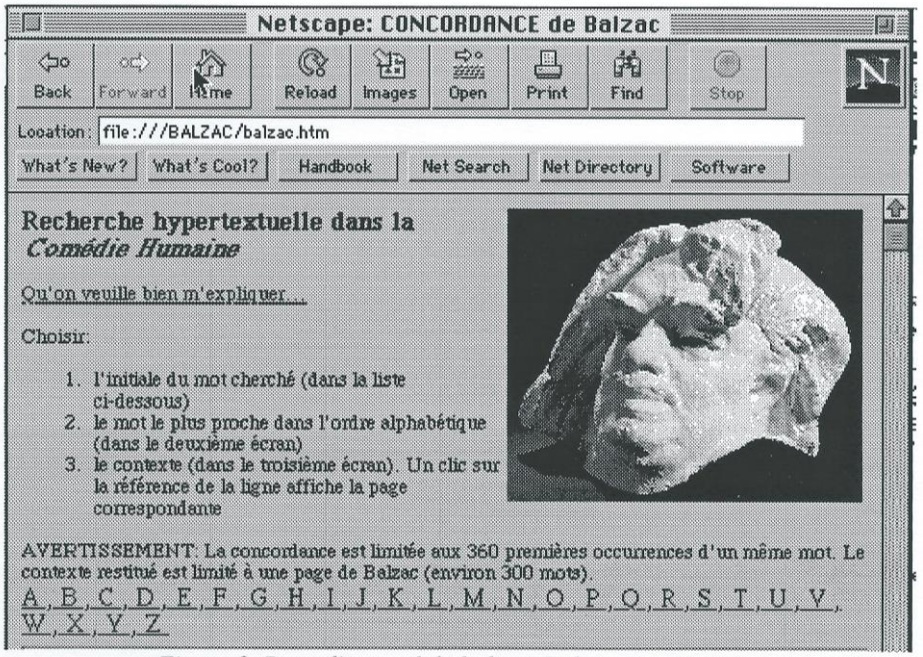


Figure 3. Page d'accueil de la base Balzac sur Internet (adresse http://lolita.unice.fr*)*

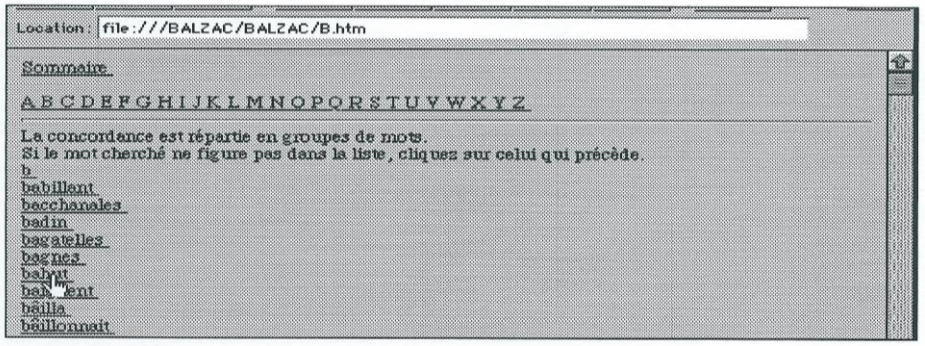


Figure 4. L'index de la base Balzac

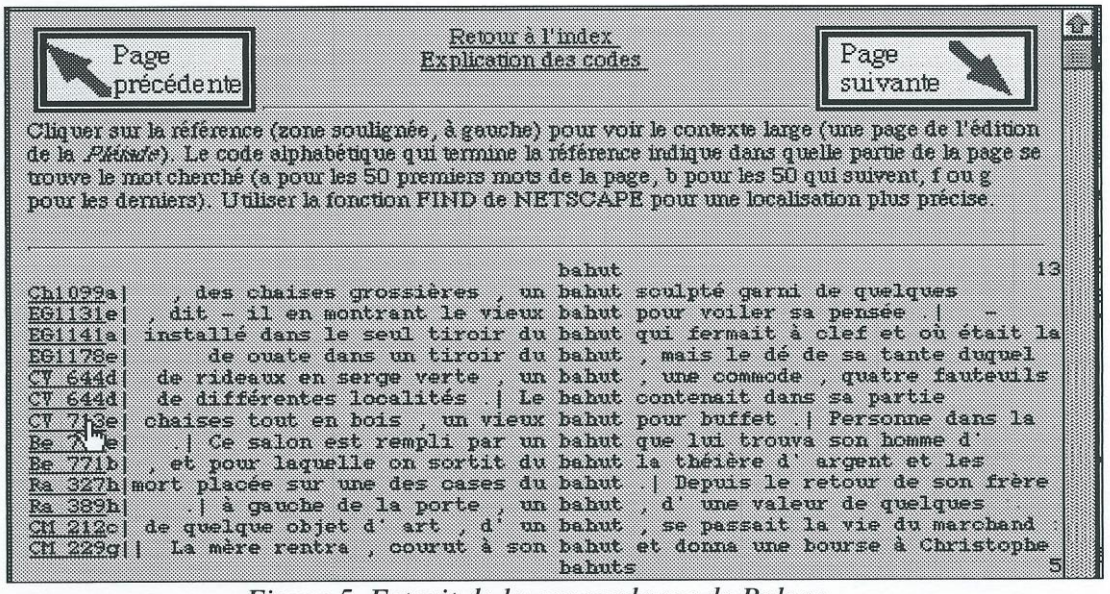


Figure 5. Extrait de la concordance de Balzac

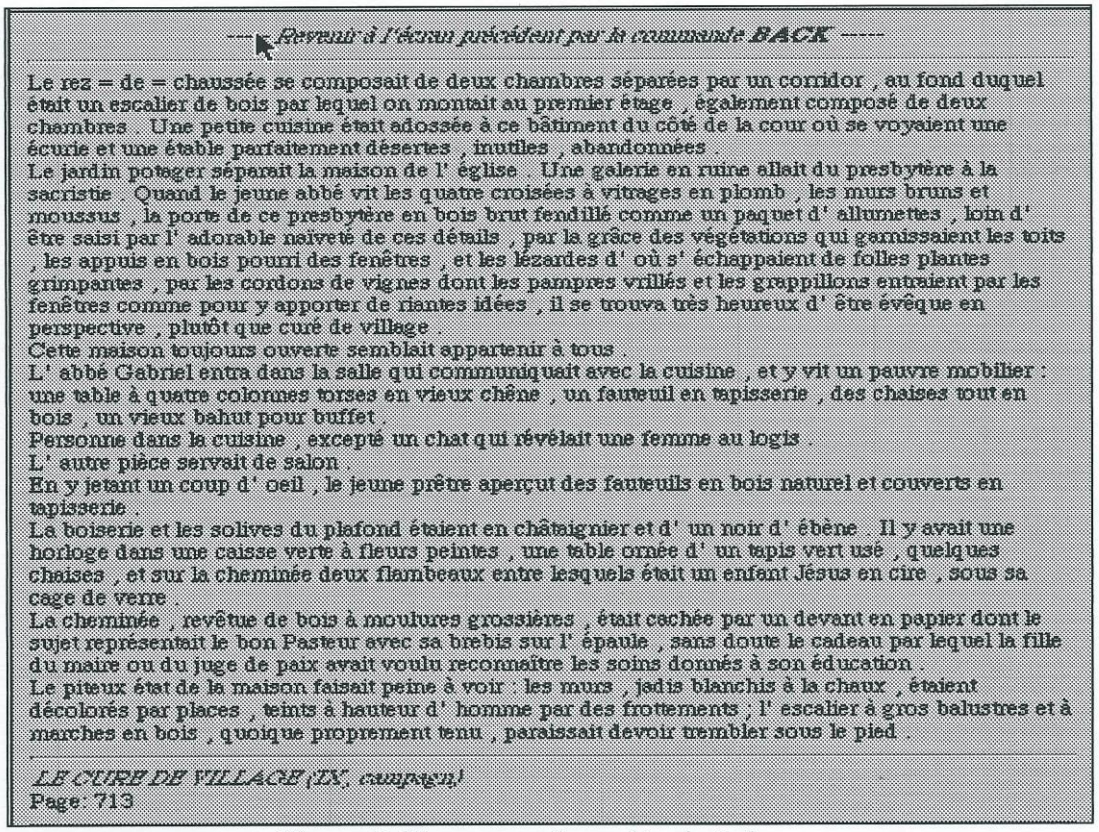


Figure 6. Une page adressable de Balzac

- III –
Balzac sous Acrobat

La technologie Acrobat offre une alternative intéressante au langage HTML. Elle offre une plus grande fidélité au texte, puisque le document d'origine est reproduit sans modifier la mise en page. Il ne s'agit pourtant pas d'une reproduction en mode image. Car le texte est considéré comme tel et la recherche hypertextuelle peut s'y exercer sans entrave. Comme les pages HTML, les documents au format Acrobat (ils ont le suffixe PDF) peuvent circuler sur le Web. Ainsi les lecteurs américains du journal le *Monde* peuvent prendre connaissance de l'édition du jour en consultant Internet, avant même que les premiers exemplaires papier ne circulent dans Paris. Cependant le format d'Acrobat est plus répandu quand la consultation est locale et que le support fait appel au cédérom. C'est le format qu'on trouve habituellement lorsqu'une information doit être transmise avec exactitude, la seule modification permise étant le zoom avant et arrière par quoi s'opère l'adaptation du texte à la taille de l'écran. Aussi avons-nous choisi cette dernière orientation dans le prototype ci-dessous :

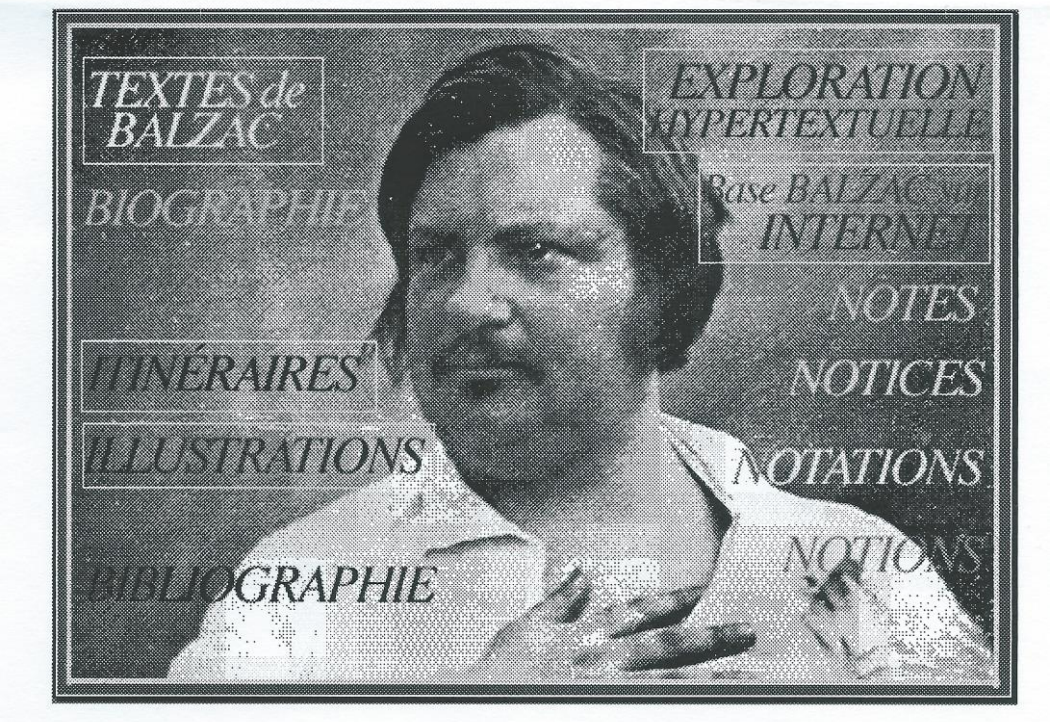


Figure 7. Menu principal de la base Balzac sous Acrobat

Pour aider l'utilisateur dans l'exploration, un plan est proposé (figure 8) qui peut conduire à la lecture du texte, mais aussi aux commentaires et aux illustrations.

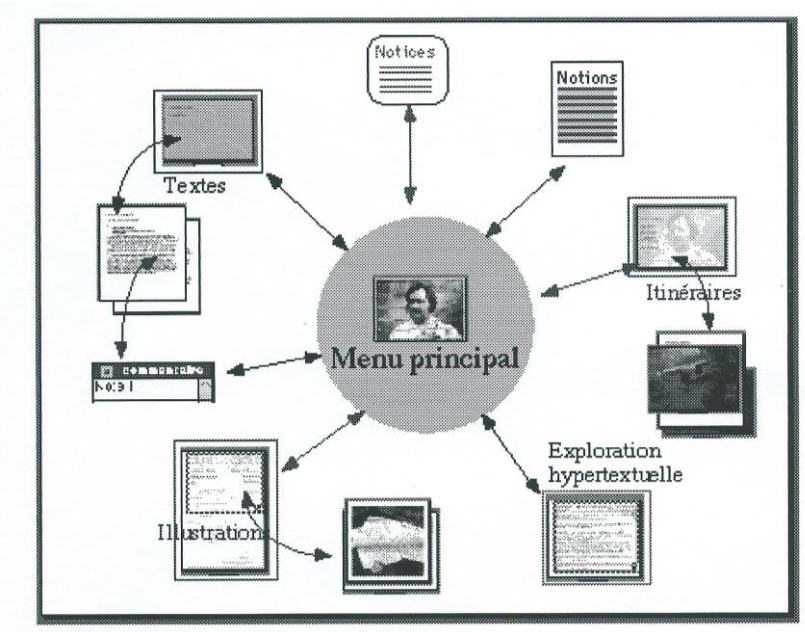
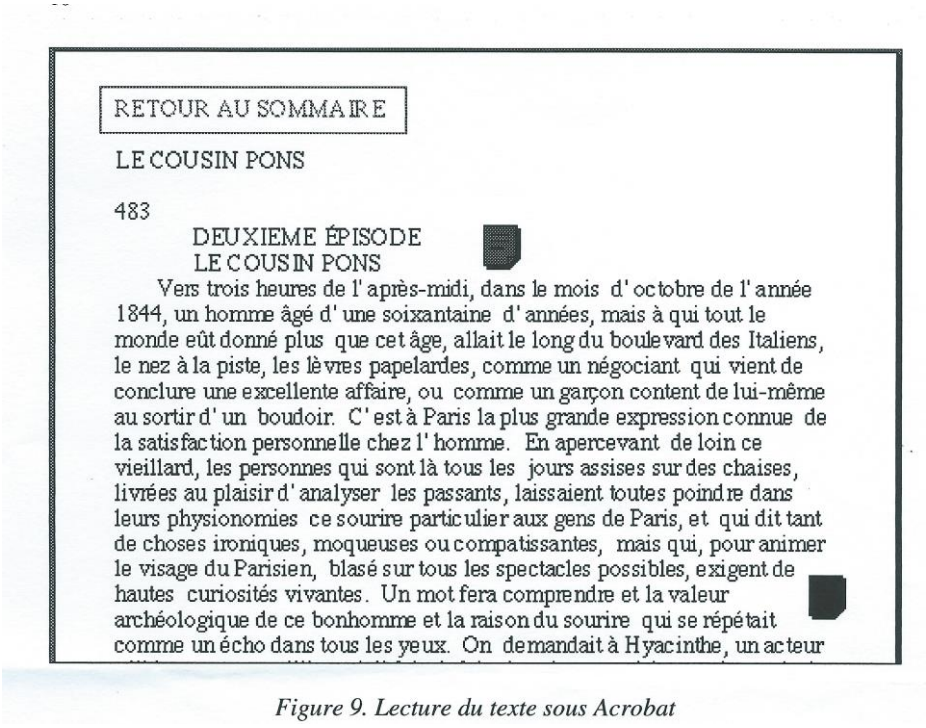


Figure 8. Le plan d'exploration

On n'insistera pas sur la lecture car pour cette fonction traditionnelle le papier offre un confort supérieur. Le seul avantage apparent est qu'on n'a pas à ouvrir l'un après l'autre les douze tomes de la Pléiade. On n'a pas non plus à tourner les pages, même pour consulter les notes. Car les appels de note apparaissent en marge du texte, avec une couleur adaptée à leur contenu. Le clic sur la vignette dévoile ce contenu en superposition, selon la technique bien connue des bulles d'aide. Un second clic fait disparaître la fenêtre et rétablit le texte sur l'écran. La souplesse de l'écran permet ainsi une signalisation discrète des notes, leur visualisation immédiate, un symbolisme qui en indique la nature, et aussi un affichage récapitulatif. Les vignettes sensibles au clic de la souris peuvent aussi désigner la page en raccourci, ce qui est précieux lorsque le texte comporte des illustrations reconnaissables. Comme les pages HTML, les pages Acrobat admettent des liens qui renvoient à d'autres pages ou d'autres textes. Dans la figure 9, c'est le cas du texte encadré qui conduit au sommaire.


RETOUR AU SOMMAIRE

LE COUSIN PONS

483

DEUXIEME ÉPISODE
LE COUSIN PONS

Vers trois heures de l'après-midi, dans le mois d'octobre de l'année 1844, un homme âgé d'une soixantaine d'années, mais à qui tout le monde eût donné plus que cet âge, allait le long du boulevard des Italiens, le nez à la piste, les lèvres papelardes, comme un négociant qui vient de conclure une excellente affaire, ou comme un garçon content de lui-même au sortir d'un boudoir. C'est à Paris la plus grande expression connue de la satisfaction personnelle chez l'homme. En apercevant de loin ce vieillard, les personnes qui sont là tous les jours assises sur des chaises, livrées au plaisir d'analyser les passants, laissaient toutes poindre dans leurs physionomies ce sourire particulier aux gens de Paris, et qui dit tant de choses ironiques, moqueuses ou compatissantes, mais qui, pour animer le visage du Parisien, blasé sur tous les spectacles possibles, exigent de hautes curiosités vivantes. Un mot fera comprendre et la valeur archéologique de ce bonhomme et la raison du sourire qui se répétait comme un écho dans tous les yeux. On demandait à Hyacinthe, un acteur

Figure 9. Lecture du texte sous Acrobat

Mais Acrobat permet en outre l'indexation du texte, ce que le langage HTML n'a pas prévu. Maintenant que le complément SEARCH est intégré gratuitement à Acrobat Reader, on dispose d'un accès direct et indexé à chaque mot (ou expression ou cooccurrence ou famille de mots) du corpus, si du moins le créateur de la base a procédé à cette indexation avant de livrer le produit au public. Si tel est le cas, une icône particulière apparaîtra à côté de celle qui désigne l'outil de recherche simple d'Acrobat et que symbolise une paire de jumelles. Un dialogue (voir figure 10) permet d'indiquer de quel index on se sert et de choisir l'objet de la recherche (le mot *art* dans la figure 11) et les options désirées (mots de même racine, ou de sens équivalent, ou d'emplacement voisin). Lorsqu'un texte n'a pas été soumis à l'indexation préalable, la recherche hypertextuelle est cependant possible dans Acrobat. Mais dépourvue d'adressage direct, elle a recours à un balayage linéaire qui est nécessairement plus lent.

Naturellement la base Balzac sous Acrobat est apte à montrer des illustrations aussi bien qu'un navigateur. Le contraire eût été étonnant de la part d'un constructeur qui a créé le langage Postscript et reste un des maîtres de l'image et des applications multimédia. On en trouvera un exemple dans la figure 12 qui représente une rue familière à Balzac.

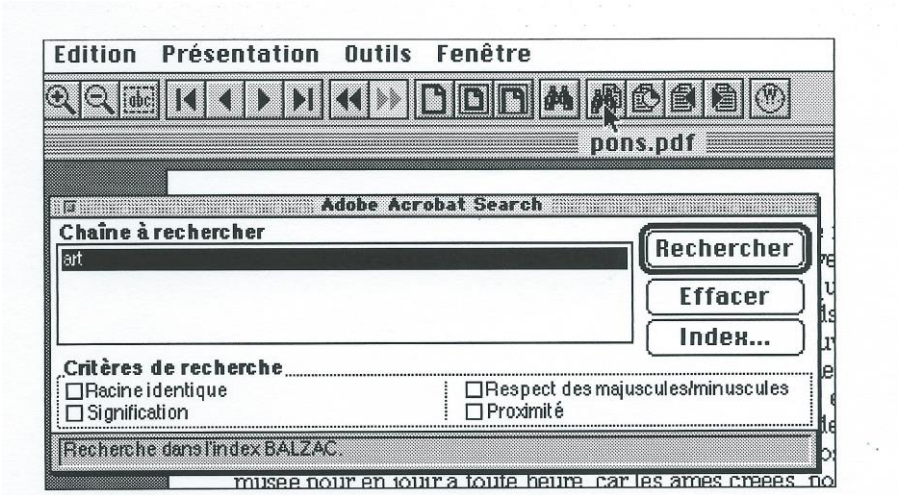


Figure 10. Dialogue pour une base indexée sous Acrobat

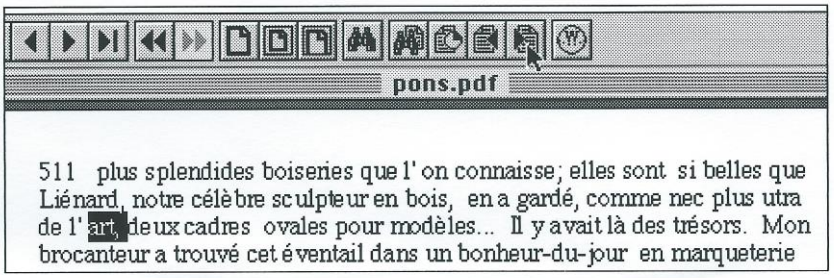


Figure 11. Résultat de la recherche dans une base indexée

Figure 12. Exemple d'illustration : La rue Berton

- IV –
Balzac sous HYPERBASE

Les solutions qu'on vient d'explorer relèvent de la démarche hypertextuelle. Toutes, à des degrés divers, permettent, à la faveur d'un mot, de relier un texte à un autre et de se déplacer par sauts et gambades dans un espace textuel à deux dimensions. Mais il y a plus de sauts que de gambades et les

mouvements gardent la raideur de l'automate. Car on y suit des pistes tracées à l'avance, qui laissent peu de place à l'imprévu, à la rencontre, et à l'aventure. Les langages à balises — HTML et Acrobat sont de ce type — ne peuvent conduire qu'à des chemins balisés. Il faut des procédures plus souples pour accéder à la pleine liberté de mouvements et à la véritable recherche hypertextuelle. Et cela ne peut guère s'obtenir que dans une application spécifique utilisant un langage adéquat.

Parmi les langages de programmation, certains semblent mieux adaptés à l'objet textuel. Ce sont les langages-objets, particulièrement ceux qui reposent sur la métaphore du texte. C'est le cas de *Toolbook* (langage *Openscript)* dans le monde *Windows* et de *Hypercard* (langage *Hypertalk)* dans le monde *Apple*. Dans ces deux langages on manipule des objets hiérarchisés qui vont du caractère *(char)* au livre *(book* ou *stack),* en passant par l'item, le paragraphe *(textline* ou *line),* les champs *(field)* et la page *(card* ou *page).* Il est bien sûr d'autres objets qui s'appliquent aux données (nombres, fichiers, images, sons) ou aux outils. Mais ce n'est pas le lieu d'entrer dans le détail technique. Il nous suffit de montrer ce qu'on peut obtenir quand une application s'engage dans cette voie.

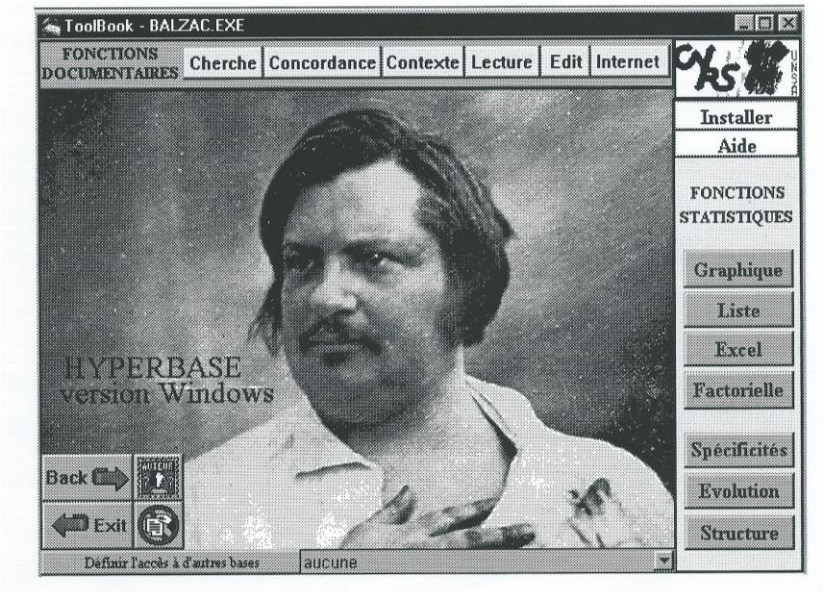


Figure 13. Le menu principal de la base Balzac sous Windows

Le menu principal représenté ci-dessus distingue deux sortes de fonctions, auxquelles répondent des boutons spécifiques, répartis à l'horizontale, au haut de l'écran, lorsqu'il s'agit de recherche textuelle ou documentaire, à la verticale, à droite de l'écran, quand la statistique linguistique est en jeu.

Les fonctions documentaires (qui aboutissent soit à une concordance d'une ligne, soit à un contexte plus large) ont ici la variété et la puissance nécessaires. Dans le dialogue de la figure 14, on remarquera le bouton *Lemme* qui réunit autour de la vedette canonique toutes les formes variables d'un verbe, d'un adjectif ou d'un substantif. On regrettera toutefois que la lemmatisation ne s'exerce pas comme il faudrait dans le texte, après examen de l'environnement syntagmatique, mais dans le dictionnaire, à un moment où il est trop tard pour dissoudre les ambiguïtés et choisir le bon code. Mais une version récente d'Hyperbase est adaptée aux données étiquetées, telles que les produit un lemmatiseur comme Winbrill.

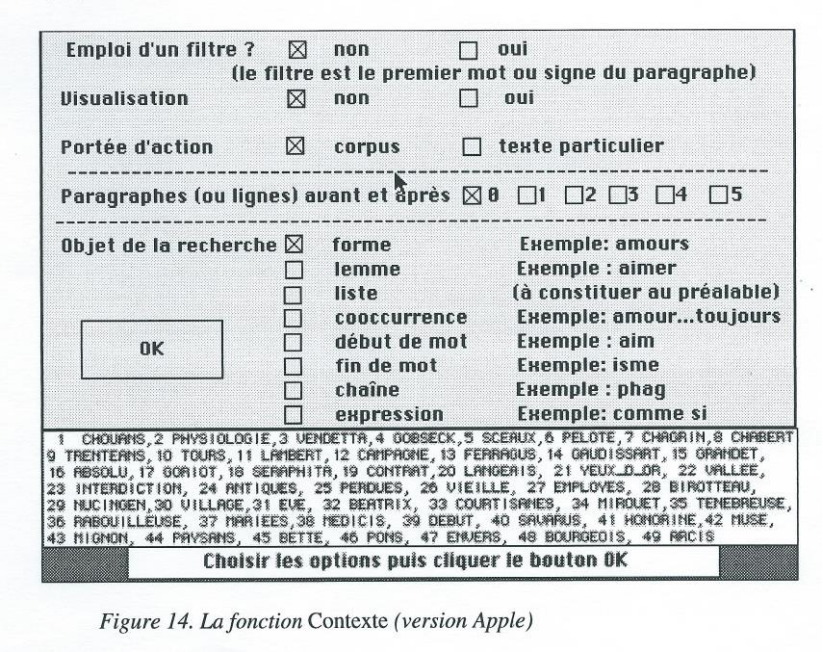


Figure 14. La fonction Contexte *(version Apple)*

Le programme d'exploitation répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes : index sélectifs ou systématiques, dictionnaires des fréquences, concordances, sélection de contextes élargis, cooccurrences, recherche des parties ou groupes de mots. *Hyperbase* se distingue toutefois des produits traditionnels : - par l'exhaustivité de l'indexation, qui prend en compte tous les mots (ponctuations incluses), - par 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, qui sont reproduits en clair, - par la souplesse et la convivialité de l'exploitation -et surtout par l'orientation statistique donnée au produit. Une comparaison est faite avec le corpus du Trésor de la langue française. Une autre, interne, met

en relation les textes de la base, ce qui engendre des courbes, des listes de spécificités, des analyses factorielles, et des mesures diverses appréciant la richesse lexicale, l'évolution du vocabulaire, la distance ou connexion des textes, etc. On trouvera ci-dessous, parmi beaucoup d'autres, le profil lexical d'un texte du corpus: *La Peau de chagrin.* Il s'agit d'un portrait dont les reliefs sont à gauche (ce sont les mots en excédent) et les ombres à droite (ce sont les déficits). Quoique l'ordre adopté (tri selon la valeur de l'écart réduit) favorise la désarticulation des traits, le thème du roman s'y lit sans peine, et même quelques touches liées à la syntaxe et à la pragmatique s'exercent parmi les mots grammaticaux et jettent quelque lueur sur le style particulier de cette oeuvre.

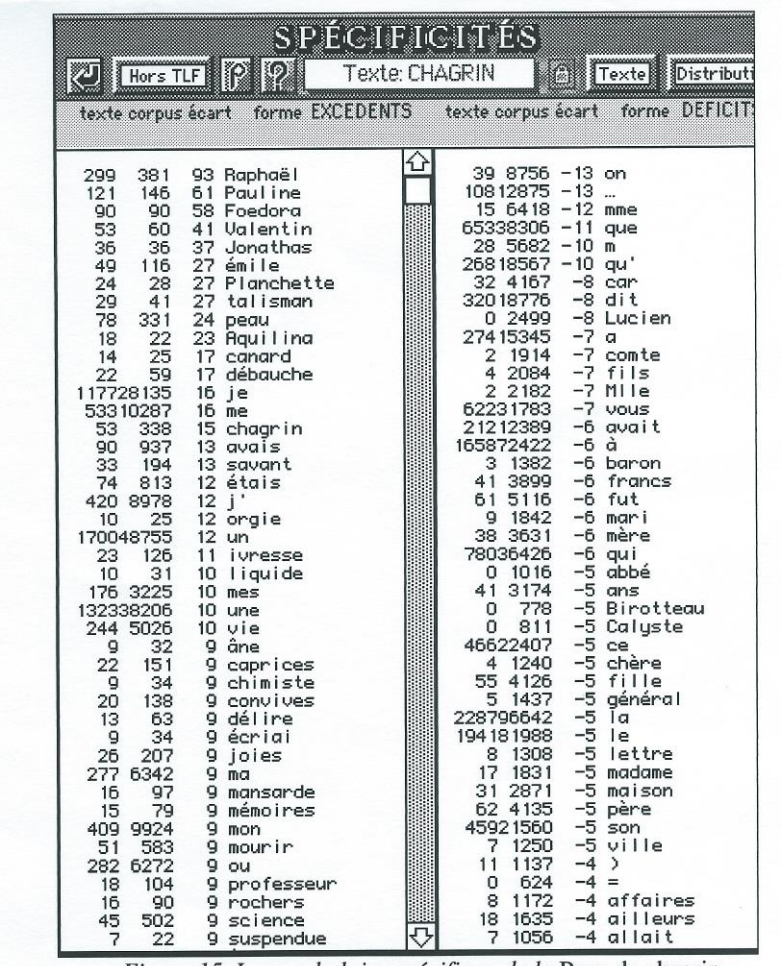


Figure 15. Le vocabulaire spécifique de la Peau de chagrin

En inversant le point de vue, au lieu d'embrasser tous les mots d'un texte pour repérer ceux qui émergent, la statistique peut envisager tous les textes à propos d'un mot ou d'un groupe de mots et mettre en relief les textes où l'on relève un emploi excédentaire ou déficitaire du ou des mots en question.

L'exemple ci- dessous (figure 16) est relatif à l'argent. Aucune tendance chronologique ne se dessine des premiers textes (à gauche) aux derniers (à droite). Mais la répartition du mot (et donc du thème) n'est nullement homogène : il y a alternance des romans voués à l'argent et ceux où interviennent d'autres ressorts, politiques, philosophiques ou amoureux. Dans les premiers — au haut du graphique - s'agitent ou se débattent les usuriers et leurs victimes : Gobseck, Grandet, Nucingen, Goriot, Birotteau. Dans la zone basse s'établissent les textes plus sensibles à l'amour qu'à l'argent : *Physiologie du mariage, Femme de trente ans, Séraphita, Duchesse de Langeais, Lys dans la vallée, Béatrix*, etc...

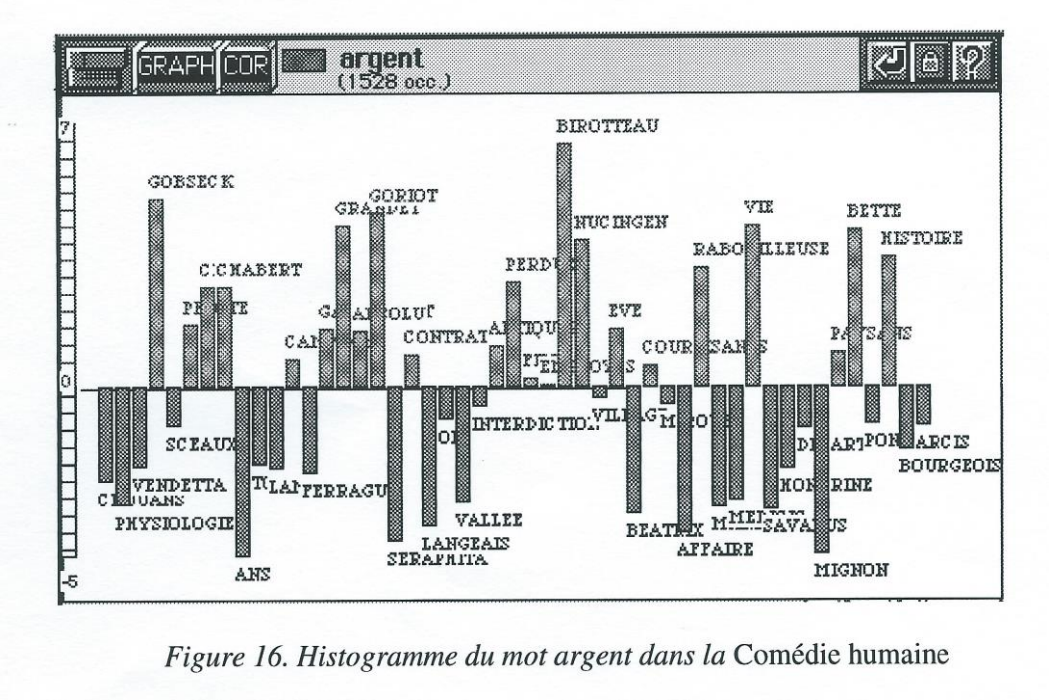


Figure 16. Histogramme du mot argent dans la Comédie humaine

Peut-on réunir les deux perspectives et s'élever assez haut au-dessus du corpus pour considérer à la fois tous les mots et tous les textes ? La machine ne répugne pas à construire un gigantesque tableau de 60000 lignes (autant de lignes que de mots) et de 49 colonnes (le corpus comprend 49 textes). Au croisement d'une ligne et d'une colonne, on enregistre ainsi la fréquence du mot i dans le texte j. Il faut alors une méthode synthétique pour traiter d'un coup tant de données réunies. C'est le rôle de l'analyse factorielle, qui peut s'appliquer pareillement à toute sélection partielle des mots et des textes.

L'exemple ci-dessous envisage la totalité du lexique. Mais au lieu de s'intéresser aux fréquences, il ne retient que la présence ou l'absence d'un mot dans un texte. Ou plus exactement on se préoccupe ici de mesurer la distance lexicale -

et sans doute thématique - qui s'établit entre les textes. Pour apprécier la distance entre deux textes, on établit le rapport des mots qui sont exclusifs à ceux qui sont communs aux deux textes considérés. On comprend aisément que la distance est d'autant plus grande que la part privative l'emporte sur les parties communes du vocabulaire. Les dénombrements sont certes longs mais les calculs sont rapides puisque l'analyse porte cette fois sur un tableau carré et symétrique (49 x 49) dont on peut attendre une sorte de carte thématique, analogue à la carte géographique que le même programme peut reconstruire quand on lui communique les distances kilométriques des villes deux à deux. La carte ci-dessous n'est pas sans analogie avec la courbe précédente : les textes qui sont répartis sur la gauche sont ceux que l'amour irrigue, ou quelque inspiration idéaliste, tandis qu'à droite le réalisme prévaut, avec le règne de l'argent. Les mots *amour* et *argent* n'ont pourtant aucune influence dans l'analyse, étant noyés comme des gouttes dans l'océan lexical. Il s'agit ici de tendances profondes qui parcourent l'oeuvre balzacienne.

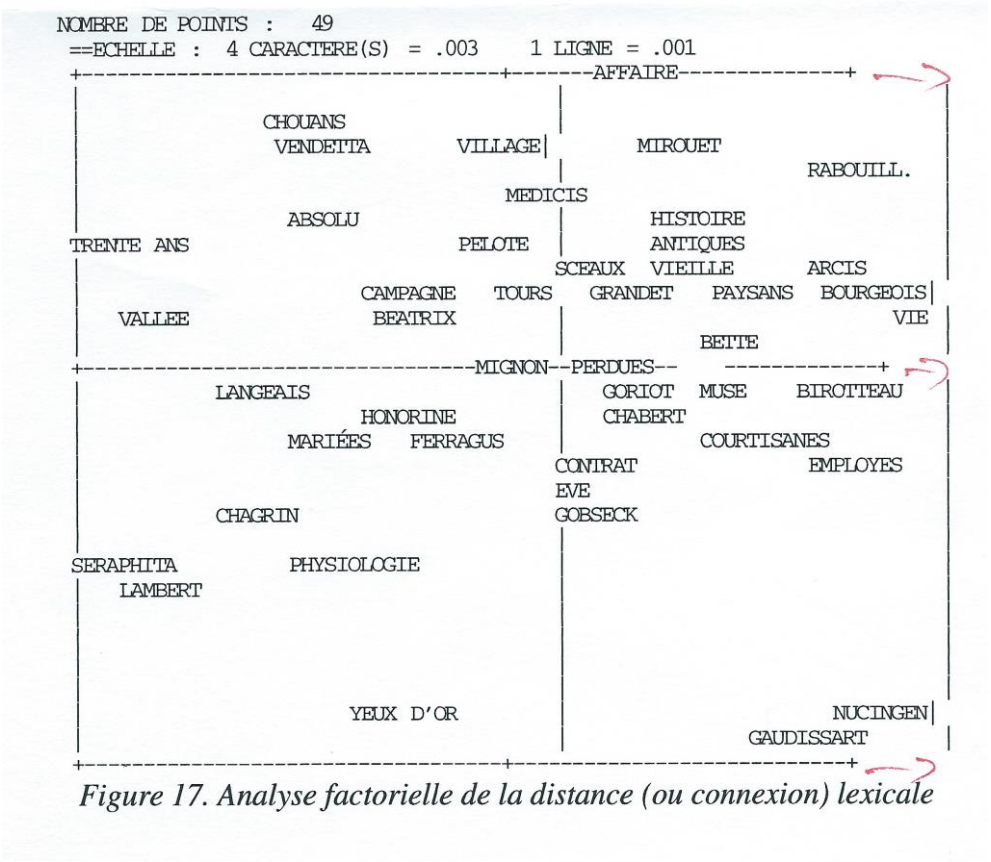


Figure 17. Analyse factorielle de la distance (ou connexion) lexicale

Bien d'autres analyses sont possibles qui portent sur les différents éléments comptables qu'on rencontre dans un texte. À travers les mots et leur assemblage, à travers les parties du

discours et les structures grammaticales, à travers les signes de ponctuation et la segmentation du texte, différentes perspectives s'ouvrent dans la forêt balzacienne. Nous n'avons pas le temps de les explorer ici.

Les grands corpus comme celui de Balzac offrent plus de chances de découvertes que les petits. Ou du moins l'informatique et la statistique ont plus d'intérêt et d'efficacité quand s'ouvrent de grands espaces, que la mémoire humaine n'arrive plus à contenir. Mais, si grand soit-il, un territoire ne peut être parcouru que s'il est d'abord situé, en faisant référence aux pays voisins. C'est en rapprochant Balzac de Stendhal, de Flaubert ou de Georges Sand que l'originalité balzacienne prend du relief. Le logiciel *Hyperbase* a donc été utilisé pour des monographies qu'on peut comparer à celle de Balzac, et qui sont disponibles pour Sand, Verne, Nerval, Baudelaire, Maupassant. Plusieurs, qui intéressent moins directement Balzac, sont au catalogue des éditions Champion (Rabelais, Pascal, Rimbaud, Proust). Vient naturellement à l'esprit l'idée d'une base récapitulative où tous les écrivains seraient représentés. Cette idée n'est pas nouvelle, et ce n'est pas seulement une idée, mais une réalité qui porte un nom connu : *FRANTEXT*, et qui est bien la tentative la plus ambitieuse, et, à notre sentiment, la mieux réussie, dans le domaine des hypertextes littéraires, d'autant que des prolongements récents donnent maintenant accès au texte lemmatisé.

Comme l'exploitation statistique n'est pas poursuivie à son terme dans *Frantext*, *Hyperbase* a servi de relais pour créer deux bases statistiques dérivées de *Frantext*. L'une est déjà ancienne et s'attache à explorer la perspective chronologique dans la littérature française (base *THIEF*). L'autre, grosse de 55 millions de mots, s'emploie à comparer les écrivains, du moins 70 des plus représentatifs de notre littérature. Balzac y figure au premier plan.

Un prototype, même enterré, peut par des voies souterraines réapparaître dans le paysage. Un nouveau cédérom Balzac est à l'étude auquel s'intéressent les éditions Champion, qui distribuent notre logiciel. On peut donc espérer quelque résurgence à l'occasion du bicentenaire de Balzac. Quoi qu'il en soit, toute étude exploratoire est plus profitable, du point de vue méthodologique, que la réalisation définitive. L'occasion nous a

été ainsi donnée d'approfondir quatre approches hypertextuelles et de les ordonner. La première ne peut convenir qu'aux esprits timides à qui la navigation hypertextuelle fait peur et qui, sujets au mal de mer, resteront au port en compagnie de leur WORD familier. La seconde est adaptée à une consultation sur Internet, mais elle reste figée et fermée, comme une borne d'information qui ne sait que restituer. La troisième méthode a des possibilités plus étendues et une présentation meilleure, mais elle partage aussi le même défaut. Reste la dernière à laquelle, faute de mieux, on donnera la préférence.