Introduction aux études littéraires assistées par ordinateur

MICHEL BERNARD

puf écritures électroniques

INTRODUCTION
AUX ÉTUDES LITTÉRAIRES
ASSISTÉES PAR ORDINATEUR

ÉCRITURES ÉLECTRONIQUES

COLLECTION DIRIGÉE PAR

BÉATRICE DIDIER

ET NATHALIE FERRAND

INTRODUCTION AUX ÉTUDES LITTÉRAIRES ASSISTÉES PAR ORDINATEUR

Michel Bernard
Maître de conférences en littérature française
à l'Université de la Sorbonne-Nouvelle (Paris III)

Presses Universitaires de France

ISBN 2 13 049703 9

Dépôt légal — 1re édition : 1999, avril

108, boulevard Saint-Germain, 75006 Paris

Sommaire

De nouvelles bases

HISTORIQUE

Disons immédiatement que l'objet de ce livre n'est pas de se demander une fois de plus si les études littéraires doivent s'ouvrir à l'utilisation des nouvelles technologies. Outre le fait qu'il y a quelque ridicule à parler de « nouvelles technologies » quand les ordinateurs ont maintenant un demi-siècle d'existence, on se bornera à la constatation que les « littéraires » utilisent déjà, dans leur immense majorité, les outils informatiques[1]. La quasi-totalité des mémoires, des articles et des livres sont aujourd'hui composés avec un traitement de texte. Un grand nombre de chercheurs et d'étudiants consultent régulièrement les bibliographies et encyclopédies électroniques que mettent à leur disposition les bibliothèques universitaires. Le courrier électronique a également conquis beaucoup de chercheurs et les sites littéraires fleurissent à l'envi sur l'Internet.

On ne se demandera donc pas ici si l'on doit utiliser l'informatique dans le cadre des études littéraires mais on mon-

1. Une enquête menée en 1990 indiquait déjà que 60 % des universitaires spécialistes de Lettres utilisaient le traitement de texte pour leurs travaux (Meyer, 1991, p. 134). Je ne connais pas d'enquête plus récente sur ce sujet mais il est évident que la proportion a encore augmenté.

trera comment elle est mise en œuvre par la recherche et comment elle peut permettre une plus grande efficacité. Il y a en effet plus de trente ans que les ordinateurs sont utilisés pour l'étude de la littérature. La bibliographie proposée ici (p. 155) ne donnera qu'un pâle aperçu de l'énorme quantité de travaux publiés dans ce domaine. Il s'agit d'un champ de la recherche qui a une histoire assez longue pour que l'on puisse tracer un tableau assez sûr de ses méthodes, de ses acquis et de ses limites[1].

Plus encore, il est important de considérer que les travaux informatisés ne font que continuer, avec de nouveaux outils, des traditions séculaires de la recherche philologique et littéraire. On a par exemple fabriqué des concordances et des index depuis le Moyen Age[2], on a pratiqué une forme manuelle de la statistique textuelle depuis le XIX^e^ siècle[3] (à moins qu'on la fasse remonter aux kabbalistes...). Il ne s'agit donc pas de l'irruption dans le champ littéraire d'un intrus exogène et imposé de l'extérieur mais de la rencontre assez naturelle entre les techniques de la recherche littéraire et des outils qui la facilitent en la déchargeant de ses tâches les plus ingrates.

La question de savoir si les études littéraires ne vont pas se dévoyer dans cette utilisation de la technologie n'a pas plus de sens que de se demander si l'on n'aurait pas mieux fait jadis de s'en tenir au manuscrit ou naguère à la plume d'oie. Les appréhensions de cet ordre viennent souvent d'idées reçues, ou d'une mauvaise connaissance de ce qui est et demeure un outil. Le présent ouvrage voudrait lever certains de ces malentendus.

1. Voir une étude sur les premières références dans la littérature spécialisée (Pellen, 1983), les rétrospectives de Galli Pellegrini, 1978 ; Oakman, 1975 et 1980 ; Fortier, 1971 (sur la littérature française), de Chisholm, 1985 (sur la littérature allemande), Saez-Godoy, 1975 (pour l'Espagne). Vuillemin, 1990, représente le bilan français le plus récent.

2. Voir Sekhraoui, 1995.

3. Voir Salem, 1993, p. 13 et sq. L'Allemand F. W. Käding publia en 1897 le lexique d'un corpus de 11 millions de mots (Dugast, 1980, p. 8).

MÉTHODES

Mais si l'ordinateur n'est qu'un outil, c'est un outil très particulier. Il serait trompeur de vouloir l'assimiler aux autres instruments traditionnels de la recherche littéraire : il n'y a jamais eu de colloques, de revues, de stages de formation, de polémiques autour de l'utilisation des fiches cartonnées[1] ou du stylo à bille ! Si l'informatique suscite autant de réactions et de réflexions[2], c'est qu'elle introduit dans le champ littéraire des pratiques, des notions, des possibilités et des préoccupations qui lui étaient jusque-là étrangères. Examinons certains de ces concepts.

Coûts

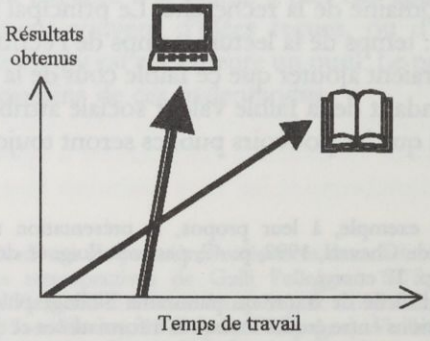
Le premier problème posé par l'informatique aux littéraires est celui du coût. Jusqu'alors, la recherche littéraire était la moins coûteuse de toutes en matériel. Une bibliothèque (dont le fonds vieillit beaucoup mieux que celui d'une bibliothèque scientifique), du papier, une machine à écrire, et voilà le littéraire équipé pour les études les plus poussées. Son « laboratoire » est si simple qu'il peut généralement travailler chez lui, ce qui favorise l'individualisme propre à ce domaine de la recherche. Le principal investissement, c'est le temps : temps de la lecture, temps de l'écriture. Des esprits chagrins pourraient ajouter que ce faible coût de la recherche littéraire est le pendant de la faible valeur sociale attribuée à ses résultats : il est vrai que les pouvoirs publics seront toujours plus dispo-

1. Voir par exemple, à leur propos, la présentation traditionnelle bien qu'assez récente de Chevrel, 1992, p. 47, pas très éloignée de celle qu'en faisait Rudler en 1923 (p. 12 et sq.).

2. Il serait difficile de tracer un panorama bibliographique complet de la question des relations entre études littéraires informatisées et théorie de la littérature, mais on trouvera des éléments de réflexion dans le numéro de *Computers in the Humanities* consacré à l'examen des critiques de Mark Olsen (Olsen, 1993-1994), ainsi que dans Landow, 1992 ; Delany, 1994 ; Vuillemin, 1990 ; Zwaan, 1987 ; Corns, 1986, et Smith, 1981 (intéressant à titre historique).

sés à financer la découverte d'un nouveau vaccin ou d'un alliage plus résistant qu'une nouvelle lecture des œuvres de Sedaine...

L'introduction de l'informatique change les données du problème. Les machines, les logiciels, les temps de calcul, les techniciens, les salles spécialisées, les heures de connexion et de formation ont un coût important, ce qui entraîne une réévaluation des procédures, des objectifs et de l'organisation de la recherche. Prenons l'exemple concret d'une étude sur les images de la mort dans *Voyage au bout de la nuit*[1]. Fait « à la main », ce travail va demander une (re)lecture du roman au cours de laquelle on relèvera sur fiches les contextes et leurs références. L'ordinateur, de son côté, ne peut travailler que sur un texte numérisé, qu'il faudra donc préalablement constituer. Au mieux, il existe déjà et l'on pourra se le procurer (mais il ne s'agit pas ici d'un texte du domaine public), au pire il faudra le numériser soi-même[2]. Les opérations seront ensuite beaucoup plus rapides : la recherche et l'affichage de tous les contextes d'une forme ou d'une liste de formes ne demanderont que quelques secondes à la machine. Tout le reste (tri, typologie, élaboration d'un modèle, rédaction) sera accéléré par l'utilisation d'un traitement de texte (copie et manipulation des contextes, par exemple). Au total, les caractéristiques de chacune de ces méthodes peuvent être figurées par ce schéma :



1. Voir Phalèse, 1993, p. 70.
2. Voir p. 26 pour cette opération.

Alors que le travail traditionnel donne immédiatement des résultats, l'usage de l'informatique impose une préparation mais permet, ensuite, de rattraper rapidement le temps perdu. Le point de jonction des deux courbes correspond à un « seuil de rentabilité » qu'il est important de fixer pour pouvoir choisir entre les deux méthodes, et ne pas se retrouver dans la situation de se dire au bout de plusieurs jours de saisie : « En travaillant à la main, j'aurais déjà fini. »[1] S'il ne s'agit, par exemple, que de faire la liste des occurrences d'un seul mot, le passage par la numérisation paraît superflu. Il faut cependant envisager la question dans tous ses développements : est-on sûr que l'on n'aura pas besoin, après un premier repérage, de chercher les contextes de mots auxquels on n'avait pas songé au départ ? N'y a-t-il pas d'autres chercheurs qui souhaiteraient travailler sur le même texte et seraient donc prêts à partager la tâche de numérisation ? A-t-on besoin d'un comptage absolument sûr ou d'une simple estimation ?

Au-delà de ces questions méthodologiques, la recherche littéraire assistée par ordinateur est amenée à se reposer la question de la valeur de ses résultats. Cette question nouvelle pour les littéraires : « Combien cela va-t-il coûter ? » entraîne forcément d'autres questions, tout aussi inouïes : « Qui va payer ? », « A qui, à quoi cette recherche va-t-elle servir ? », « Vaut-elle le prix des dépenses envisagées ? » Ces problématiques de la recherche sont fort communes en sciences expérimentales, où les laboratoires doivent se préoccuper de leur financement, mais peu familières aux chercheurs en littérature. Seules de grandes entreprises comme la constitution par l'Institut National de la Langue Française de la base de données Frantext[2] et du Trésor de la langue française peuvent nous donner une idée – dans un domaine voisin qui est celui de la lexicographie – de ce qu'est le fonctionnement d'un centre de recherche coûteux.

1. Voir par exemple la manière dont un médiéviste (Heinemann, 1994) pose la question : « Le grand problème de l'analyse informatisée est celui du rendement par rapport à l'investissement en efforts » (p. 30).
2. Voir p. 21. L'histoire de l'entreprise est racontée dans Gorcy, 1985.

Travail en groupe

Bien loin de favoriser son autonomie, l'ordinateur impose en effet au chercheur littéraire de travailler en groupe[1]. Les investissements importants – en temps, en matériel, en formation – exigent d'être partagés. L'acquisition d'une machine puissante, ou d'une banque de textes, peut être mise en commun par plusieurs chercheurs. De fait, ce domaine des études littéraires a vu se créer plusieurs centres de recherche (pour la France : Nancy, Saint-Cloud, Nice, Clermont-Ferrand, Paris III, Paris VIII, Montpellier...) fédérant une équipe autour d'un laboratoire offrant des moyens de calcul, des logiciels, une assistance technique et méthodologique. Les programmes de recherche impliquent souvent une démarche interdisciplinaire[2], ne serait-ce que parce que le littéraire doit souvent faire appel à un informaticien pour écrire ou adapter des programmes.

Par ailleurs, l'informatique, qui est une technologie en perpétuel renouvellement, exige une remise à jour permanente non seulement des matériels et des logiciels mais aussi des méthodologies. Un chercheur isolé ne peut mener à bien cette *veille technologique* qui lui permet d'être au fait de toutes les possibilités ouvertes dans sa discipline par les derniers perfectionnements de l'outil informatique. L'utilisation des méthodes de la statistique exige aussi une confrontation systématique des résultats et des méthodes pour les valider et les perfectionner. Ces « laboratoires de littérature » d'un nouveau genre risquent fort de mettre à mal, dans les années qui viennent, l'image du chercheur solitaire. Il est à noter par exemple que la plupart des travaux publiés dans ce domaine sont cosignés par plusieurs auteurs, à l'instar de ce que l'on constate dans le domaine des sciences exactes.

1. Sur ce thème, voir Laurette, 1993, p. 12 et le chapitre « Littérature, pluridisciplinarité et technologie » (p. 45). Voir également Denley, 1990, et Massonie, 1986.

2. Voir Denley, 1990.

Rapidité

La rapidité de fonctionnement des ordinateurs est un autre facteur de changement des mentalités. L'informatique permet en effet de réaliser en un temps très bref des tâches traditionnellement longues et fastidieuses. Un seul exemple fera comprendre le phénomène : la *collation* des différentes versions d'un même texte, c'est-à-dire la comparaison minutieuse, deux à deux, de toutes les éditions d'une œuvre, est un travail extrêmement pénible, mais indispensable à l'établissement d'une édition critique et des *variantes* d'un texte. L'opération peut être effectuée par l'ordinateur[1], sur des textes numérisés, avec une grande précision et dans un délai très court. Toutes sortes de tâches de ce type (recensions, comparaisons, comptages, copies, références, ...) sont ainsi accélérées par les moyens informatiques.

Ce saut quantitatif a d'ores et déjà généré un changement qualitatif dans les études littéraires[2]. Le chercheur, en effet, peut aujourd'hui envisager des entreprises devant lesquelles il aurait reculé naguère. De vastes synthèses comme celles d'Étienne Brunet sur le vocabulaire de Giraudoux, de Hugo, de Zola ou de Proust sont à peu près inimaginables sans moyens informatiques. Plus encore, cette facilité permet au chercheur de vérifier rapidement des hypothèses hasardeuses. Qui se lancerait dans une lecture intégrale des œuvres de Hugo simplement pour vérifier si elles mentionnent le nom de Confucius, au risque de conclure que ce n'est pas le cas ? L'investissement serait de toute façon beaucoup trop important par rapport au résultat. Une simple requête informatique fournira une réponse : aucune occurrence, par exemple, dans les œuvres enregistrées par Frantext.

1. Un simple traitement de texte peut comparer les versions d'un document.
2. Cf. Olsen, 1993-1994 [2], p. 309 : « [...] l'analyse textuelle assistée par ordinateur permet aux chercheurs de se poser des questions nouvelles, sans rapport avec les perspectives des lectures traditionnelles » (« [...] computer methods in textual analysis allow scholars to ask new questions which do not correspond to the traditional notions of reading texts »).

Il en est de même dans le domaine documentaire. Les encyclopédies et dictionnaires électroniques permettent de réaliser à une grande vitesse des opérations devant lesquelles tout chercheur traditionnel aurait rechigné. Il est envisageable par exemple, avec le *Robert électronique*, de vérifier la date de première attestation de *tous* les mots d'un poème d'Apollinaire pour se rendre compte de la part d'archaïsmes et de néologismes dans le texte : qui aurait fait une telle démarche en feuilletant les pages d'un dictionnaire étymologique ? Les perspectives peuvent être comparées à celles qui se sont ouvertes au monde savant de la Renaissance[1], quand la mise à disposition commode et rapide d'une grande quantité de textes de bonne qualité a généré des découvertes d'une immense portée dans tous les domaines de la connaissance, et spécialement en philologie.

Exactitude

Pour les raisons que je viens de présenter, la recherche littéraire s'est souvent contentée de procéder par sondages. On étudie un auteur, une œuvre, quelques œuvres représentatives d'une époque ou d'un genre. Les synthèses les plus vastes reposent sur la connaissance d'un corpus nécessairement réduit à ce qu'un être humain peut lire et assimiler, et sur des aperçus de seconde main. L'existence de bibliothèques numériques considérables change les données du problème. Il est possible dès aujourd'hui – et il sera de plus en plus aisé – d'étudier des corpus énormes dont on ne lira que ce que l'analyse informatique fera apparaître comme pertinent. Il ne s'agit pas de lire moins mais de lire mieux.

L'outil informatique permet ainsi l'émergence de nouvelles catégories en matière d'analyse des textes. On peut affirmer, par exemple, et avec une grande sécurité, qu'un terme est absent d'un corpus. Un repérage manuel ne peut arriver à une telle certitude.

1. Le parallèle, fréquent, est par exemple développé dans Denley, 1990 ; Delany, 1994 (p. 9-10).

Il est possible ainsi de caractériser une œuvre non seulement par ce qu'elle dit mais aussi par ce qu'elle évite, ce qu'elle cache[1], ce qu'elle travestit. La grande précision des comptages permet aussi d'utiliser dans le domaine littéraire les méthodes de la statistique, selon un vœu qu'exprimait dès 1894 le latiniste Paul Lejay : « Par la statistique seulement, grammaticale et lexicographique, on peut introduire dans la littérature un peu de rigueur et de certitude. »[2] Mais ce souci d'exactitude, qui était celui des Lanson et des Rudler, avait malheureusement paralysé toute la recherche littéraire du début de ce siècle parce qu'il imposait au chercheur des tâches écrasantes dans les conditions de travail de l'époque. Les nouveaux outils informatiques permettent enfin de concilier érudition et esprit de synthèse, précision du détail et vision d'ensemble.

CE LIVRE

Bien loin de vouloir traiter à fond ces questions théoriques, le présent volume ne veut être qu'un guide pratique et actuel des applications de l'informatique dans le domaine littéraire. Sa plus grande prétention est de se situer dans la lignée des manuels des études littéraires de Rudler[3], Bouvier et Jourda[4] ou, plus récemment, Yves Chevrel[5]. On insistera plus ici sur le « comment » que

1. Lire à ce propos l'étude de C. Lautier, qui détecte dans *Locus Solus* la présence d'un « mot-thème » que la simple lecture ne pourrait mettre en évidence (Lautier, 1984).

2. Cité par Charles Muller (Muller, 1979, p. 347).

3. Rudler, 1923.

4. Émile Bouvier et Pierre Jourda, *Guide de l'étudiant en littérature française,* PUF, 6e éd., 1968.

5. Chevrel, 1992. Voir p. 117 (« Informatique et littérature ») : « Nous sommes effectivement au début de métamorphoses importantes dans les conditions de travail dans ce domaine : accès aux documents, stockage des données, recherche d'informations, traitement de texte, etc., autant de domaines dans lesquels une révolution est en train de s'opérer. »

sur le « pourquoi », plus sur les outils que sur les débats qu'ils suscitent.

Qu'il soit cependant bien entendu que ce refus de s'interroger sur la légitimité d'une approche informatisée des problèmes littéraires constitue, en soi, une prise de position. Il me semble en effet que l'ordinateur ne génère rien qui ressemblerait à une « nouvelle critique », à une approche radicalement originale du phénomène littéraire mais qu'il peut, en revanche, se mettre au service de toutes les lectures, de toutes les pratiques de la recherche, des plus traditionnelles aux plus nouvelles. C'est pour cela qu'il sera considéré ici comme un outil, dont chacun fera un usage d'autant plus libre que l'on en connaîtra mieux les principes et les limites.

Pour rendre cet ouvrage lisible à plusieurs niveaux, j'ai choisi de rejeter en notes de bas de page[1] toutes les indications bibliographiques, pour suggérer des lectures complémentaires. Il est ainsi possible de lire une présentation rapide et accessible des différentes applications et d'en approfondir tel ou tel aspect en se reportant à la bibliographie.

Celle-ci, malgré sa taille, ne constitue qu'une sélection, nécessairement subjective. Je peux simplement indiquer que j'ai privilégié :

— les références les plus récentes par rapport aux études anciennes (sauf les textes importants qui ont marqué l'histoire de la discipline) ;
— les études rédigées en français (mais on constatera que les textes en anglais sont très nombreux dans ce domaine) ;
— les études portant sur la littérature française (j'ai exclu beaucoup de travaux méthodologiquement intéressants mais dont les textes d'application n'étaient pas littéraires) ;
— les travaux originaux, qui inventent une méthode en même temps qu'ils l'appliquent.

1. C'est aussi une manière de rappeler que ce dispositif éditorial est le précurseur de ce que l'on nomme *hypertexte* dans l'univers informatique (voir p. 149).

Certains domaines, voisins, ne seront pas explorés ici : la production de textes littéraires avec l'informatique[1], l'enseignement de la littérature assisté par ordinateur[2], les études spécifiquement linguistiques[3]. Je n'aborderai pas non plus les aspects strictement techniques de la question : types d'ordinateurs, configurations, performances, maintenance, mode d'emploi des logiciels, etc. La raison en est que ces paramètres se démodent si vite que le livre n'est pas un média très adapté pour en parler. On s'adressera plutôt aux revues spécialisées et à l'Internet pour s'en informer en temps réel.

Par ailleurs, ce manuel (et donc ses choix) doit beaucoup aux travaux du Centre de Recherches Hubert-de-Phalèse[4], dont je fais partie depuis sa création. *Hubert de Phalèse* (du nom d'un abbé du XVII[e] siècle, éditeur d'une des premières concordances de la Bible) est le pseudonyme d'un groupe d'enseignants-chercheurs qui souhaitent diffuser les méthodes informatiques dans les études littéraires. Nous publions chaque année, depuis 1991, des ouvrages[5] consacrés aux auteurs du programme d'agrégation et réalisés avec des outils informatiques, dans le but de montrer quel parti on peut en tirer. Je ne fais donc ici, en quelque sorte, que le bilan de ces travaux, qui nous ont permis d'utiliser et de tester la plupart des outils que j'aurai à présenter.

1. Sur ce sujet, voir les travaux de Balpe et Clément, ainsi que la partie IV de Vuillemin, 1990.

2. Lire des indications sur cet aspect dans Bertrand, 1986 ; Otman, 1988 (enseignement supérieur seulement).

3. Le partage est souvent très difficile, tant les approches critiques du dernier quart de siècle ont utilisé les techniques et les savoirs de la linguistique. De même, on constate que beaucoup d'études de lexicologie portent sur des corpus littéraires.

4. Le Centre Hubert de Phalèse (Université de la Sorbonne-Nouvelle - Paris III) est dirigé par Henri Béhar. Adresse électronique : <phalese@msh-paris.fr>. Site Internet : <http ://www.cavi.univ-paris3.fr/phalese/hubert1.htm>.

5. Collection « Cap'Agreg » (Nizet). On en trouvera une liste dans la bibliographie, sous le nom de Phalèse.

Traitements du texte

Un *computer* (calculateur) ne peut opérer que sur des séries de codes numériques. Il ne traite même, en réalité, que des valeurs binaires. Rechercher la lettre A dans un texte revient, pour la machine, à rechercher le code 00100001 dans une longue suite de 0 et de 1. C'est la répétition extrêmement rapide d'opérations aussi élémentaires qui peut donner l'impression que la machine est dotée d'une certaine forme d'intelligence symbolique. Même quand l'utilisateur travaille sur un texte, ou une image, l'unité arithmétique et logique de l'ordinateur ne manipule en dernière analyse que des nombres.

L'opération de *numérisation* consiste précisément à transformer une information signifiante pour l'homme en une série de codes lisibles par la machine. Le passage de l'un à l'autre ne va pas sans pertes et il faut être conscient des limites de cette traduction. Prenons l'exemple d'un calligramme d'Apollinaire[1]. Comment coder un tel document ? Si je tape simplement le texte (mais par où commencer ?), je perdrai certainement une partie de la signification. Si j'enregistre tous les points du dessin un par un, je ne pourrai pas reconnaître ce qui est écrit. Ce cas est peut-être extrême mais dans toute numérisation, il y a perte d'information (typographique par

1. Voir Wackernagel, 1991.

exemple). L'information dont nous sommes entourés est analogique (et donc infiniment détaillée), alors que l'ordinateur ne manipule que de l'information digitale (simplifiée). Si l'on examine par exemple un tableau au microscope, on peut découvrir, selon le grossissement, de plus en plus de détails, alors que l'image numérique de ce même tableau, même si elle satisfait celui qui la regarde d'assez loin sur son écran d'ordinateur, ne permettra par d'aller au-delà des points élémentaires *(pixels)* qui la composent. Cette caractéristique est celle de toutes les informations numérisées : texte, image, son. On applique sur le réel une grille plus ou moins fine[1] et on considère que le résultat est satisfaisant quand l'être humain, dans des conditions standard, ne peut plus distinguer l'original de sa représentation.

Mais cela ne doit pas nous troubler. L'imprimerie nous a habitués depuis le début de son histoire à une traduction de l'infinie richesse graphique du manuscrit en une suite de caractères interchangeables, fondus dans le même moule. L'informatique ne fait que prolonger ce mouvement. Il faudra simplement être averti des distorsions subies par les données pour pouvoir juger correctement des résultats. C'est dans cet esprit que nous allons envisager les différentes manières d'établir un texte numérisé. Mais gardons à l'esprit que ces précautions d'emploi ne sont pas d'une nature différente de celles que l'on prend en choisissant d'étudier *A la recherche du temps perdu* dans l'édition Clarac ou dans l'édition Tadié, de citer Rabelais dans l'orthographe d'origine ou dans une version modernisée, de tenir compte des variantes typographiques ou des manuscrits de Flaubert.

1. C'est ce paramètre que l'on appelle *résolution* pour une image ou *échantillonnage* pour un son.

COMMENT SE PROCURER UN TEXTE NUMÉRISÉ ?

Les banques de données textuelles

Fort heureusement, on trouve déjà une grande quantité de textes numérisés dans les bibliothèques électroniques. Il est donc possible de travailler sur ces textes sans avoir besoin de les saisir soi-même. Lorsqu'il ne s'agit que de vérifier la présence d'un mot dans un corpus considérable, ou d'opérer un comptage très simple, la disponibilité d'une version numérisée est bien entendu déterminante. Les banques de données textuelles sont de plus en plus nombreuses. Le panorama qui en est fait ici ne saurait être définitif ou exhaustif. Je m'attacherai surtout à une évaluation de ces vastes collections documentaires : quantité, qualité, disponibilité.

FRANTEXT. — La plus grande banque de données textuelles du monde est française. Il s'agit de FRANTEXT, une collection de quelque 3 600 textes intégraux[1] constituée par l'Institut National de la Langue Française depuis plus de trente ans. C'est pour rédiger le *Trésor de la langue française*, le fameux dictionnaire de la langue des XIX^e^ et XX^e^ siècles, que ce laboratoire du CNRS a numérisé une masse de textes représentatifs des usages linguistiques les plus divers[2]. On trouve en effet dans FRANTEXT des textes littéraires (c'est la majorité), mais aussi techniques, administratifs, documentaires, philosophiques, etc. La base de données est accessible seulement aux abonnés. Elle est interrogeable aujourd'hui par l'In-

1. Par « texte » il faut entendre ici une unité propre à FRANTEXT, qui ne correspond pas toujours à une œuvre (par ex., *Les Misérables* représentent dix références dans la banque). Cela devrait correspondre à environ 2 000 titres. Voir Brunet, 1981 (p. 57) sur les inconvénients de cette disposition.

2. Lire dans Dendien, 1988, un historique du projet. On lira aussi dans la préface du tome I du *TLF* les intentions de ses promoteurs.

ternet[1] mais on peut encore y accéder par une connexion télématique directe (TRANSPAC). Pour une consultation ponctuelle, on pourra toujours s'adresser à une bibliothèque abonnée (la plupart des bibliothèques universitaires le sont) ou à un centre de recherches.

Il est à noter que FRANTEXT ne permet pas de télécharger des textes entiers sur son ordinateur personnel mais propose un certain nombre de traitements prédessinés[2] : recherche d'un mot (ou de plusieurs) dans un corpus défini, établissement d'index, affichage de contextes d'environ deux pages (mais seulement pour les textes du domaine public), recherche de cooccurrences. Il est possible également de rechercher toutes les flexions d'un verbe, ou de constituer une liste des formes correspondant à un modèle. L'ensemble de ces outils fait de FRANTEXT une ressource majeure pour les études littéraires mais il faut être conscient de certaines limitations du corpus. Conçu dans une perspective lexicographique, il ne contient pas toujours les textes essentiels à une approche littéraire.

Un sondage fait sur les seuls auteurs dont le nom commence par A permet par exemple de faire les restrictions suivantes : Anouilh n'est représenté que par trois pièces *(La Sauvage, La Répétition, Antigone)* alors qu'un auteur aussi obscur que Baculard d'Arnaud s'en voit attribuer huit, il n'y a aucune pièce d'Augier, d'Adamov ou d'Arrabal, aucune œuvre d'Alexis, ni d'Alphonse Allais ; en revanche, on peut s'étonner de voir la place accordée à Jacques Abbadie *(Traité de la vérité de la religion chrétienne)*, ou la présence de deux pièces de D'Audiguier. Mais, au demeurant, les grands auteurs attendus sont bien représentés : Apollinaire avec la quasi-intégralité de son œuvre poétique, d'Aubigné avec *Les Tragiques*, le *Journal* d'Amiel, la correspondance d'Alain-Fournier avec Jacques Rivière, des textes significatifs de Marcel Aymé,

1. <http://www.ciril.fr/~mastina/frantext>. Pour l'Amérique du Nord, il est plus commode de s'abonner auprès d'ARTFL (<http://humanities.uchicago.edu/artfl>) mais on ne trouvera là qu'une collection de textes plus réduite (2 000 textes). En revanche, le site d'ARTFL propose en accès libre des documents intéressants : bibles, dictionnaires, etc.

2. Jacques Dendien est le maître d'œuvre des programmes de consultation de FRANTEXT (voir Dendien, 1986).