70 MILLIONS DE MOTS ON LINE

Étienne Brunet

70 MILLIONS DE MOTS ON LINE

Projet de base de données textuelles

applicable au corpus littéraire de l'Institut de la Langue française

(mai 1980)

Introduction

1) Le public visé est un public de spécialistes ou de chercheurs, et j'imagine qu'il en est de même de la base de données lexicographiques dont je ne dirai rien. Contrairement à la banque de données d'orthographe et de grammaire à laquelle songe le Comité International de la Langue Française, on ne vise pas ici le grand public et les options qui se présentent dans ce dernier cas ne sont pas applicables. En particulier, il ne serait pas réaliste d'offrir une banque de données textuelles aux abonnés futurs du système Télétel ou du videotex Antiope. Le système Titan conviendrait sans doute mieux, étant voué à l'exploitation des banques de données, mais comme les utilisateurs de notre base seront principalement des universitaires, c'est dans le réseau universitaire des six centres nationaux ou régionaux prévus par le schéma directeur de l'informatique qu'il faut imaginer le fonctionnement d'une base de données textuelles. Et puisque le centre de calcul de Nancy risque d'être choisi pour l'un des serveurs de l'hexagone, au moins comme serveur spécialisé dans le domaine linguistique, c'est à Nancy que devrait être créée cette banque, d'autant que les données s'y trouvent disponibles.

2) Public spécialisé, public large tout de même, car il ne s'agirait pas seulement des linguistes, grammairiens, lexicologues mais aussi des chercheurs "littéraires" qui depuis des siècles constituent des fichiers manuels, notant l'apparition, dans le texte qu'ils étudient, de tel thème, de tel mot, de telle figure, et qui seraient bien soulagés si leurs hypothèses pouvaient être vérifiées sur le champ. Une base de données textuelles serait plus précieuse encore aux téméraires qui s'intéressent, non plus à un texte, mais à un auteur, à une époque, à une école, à un genre littéraire, voire à l'ensemble de la littérature des deux derniers siècles. Quand on atteint de telles hauteurs, qu'on ne distingue des ouvrages que leur titre et qu'on ne voit des auteurs que leur crâne, on ne saurait faire que des sondages fragiles où l'intuition linguistique et littéraire tient lieu de radar ou de sonar. Offrons-leur une base, non un bazar où le hasard conduit la recherche.

Ces chercheurs ont d'ailleurs défini clairement ou confusément ce besoin. Une étude de marché a été menée récemment aux Etats-Unis par Monsieur MORISSEY sur un échantillon d'une centaine d'universitaires américains qui étudient l'histoire et la littérature françaises. Tous tombent d'accord sur l'utilité d'une base de données textuelles que l'Institut de la Langue française, par quelque moyen que ce soit et notamment par le truchement du Centre de CHICAGO, pourrait leur procurer.

Le service des prestations extérieures de l'ILF est mieux encore le lieu où s'expriment ces besoins. Son témoignage est essentiel sur la nature et l'importance des demandes enregistrées. Mais on peut imaginer que bien des demandes n'ont pas été formulées ou n'ont pas été présentées, parce que les utilisateurs éventuels ignoraient que l'ILF pouvait les satisfaire, ou peut-être parce qu'ils savaient trop bien que, en l'état actuel des services, l'ILF ne pouvait pas les satisfaire.

3) L'auteur des considérations qui suivent est lui-même un utilisateur littéraire dont les demandes ont été satisfaites. Mais, parce que les produits qu'il recevait de NANCY étaient bruts ou demi-finis, il a dû devenir analyste et programmeur, sans cesser d'être littéraire. Cela lui donne l'avantage de considérer le problème dans ses deux aspects, celui de l'objectif et celui des moyens et de prendre la chaîne par les deux bouts. Au reste, les deux bouts se touchent et à la limite n'en font qu'un, n'en déplaise à Raymond DEVOS.

I - LA CONSTITUTION DE LA BASE DE DONNEES TEXTUELLES

La solution telle que je la conçois serait la suivante : on aurait une base constituée de plusieurs fichiers liés entre eux par des clés ou pointeurs pour parler le langage des informaticiens, par des renvois pour parler le langage de tout le monde.

I) Le fichier des textes

Le premier serait un fichier à accès direct (ou sélectif) comportant, mis bout à bout, tous les textes enregistrés, et, pour réduire les coûts, tels qu'ils ont été enregistrés et tels qu'on les retrouve sur les fiches-textes de 18 lignes que tout le monde connaît. Deux modifications toutefois doivent être apportées aux fiches-textes, une soustraction et une addition :

- d'abord une soustraction : il faut enlever dans les fiches-textes toutes les lignes répétées, c'est-à-dire les 5 premières et les 5 dernières qui se retrouvent nécessairement sur la fiche précédente ou suivante. On aurait ainsi un continuum bien serré, comme un gros volume où seraient reliés les 1000 ouvrages du XIX^{e} et du XX^{e} du corpus ;

- mais il faudrait aussi procéder à une addition et compléter chaque ligne par une indication, une clé ainsi faite qu'on puisse d'un coup retrouver la ligne en question, si elle contient un mot qui intéresse l'utilisateur et si toutes les contraintes imposées par celui-ci sont satisfaites qu'elles concernent l'époque, le genre, l'auteur, le texte, ou tel sous- ensemble du texte.

Voici la disposition de la clé telle que je l'imagine :

(la première ligne désignant le nombre de chiffres, la seconde les octets nécessaires)

clé									
numéro absolu de ligne	nombre mots dans la ligne	date	genre	auteur	texte	sous-ensemble du texte	page	ligne dans la page	texte de la ligne
7	2	3	2	3	4	3	4	2	
4	2	2	2	2	3	2	3	2	58

soit 22 octets à ajouter à chaque ligne ce qui porterait à 80 caractères la longueur de chaque enregistrement. Précisons qu'il n'est nul besoin de reprendre une à une chacune des lignes. Pour un sous-ensemble donné presque tous les octets de définition sont répétés de la même façon, sur toutes les lignes ; seul change le numéro de ligne qui s'incrémente d'une unité à chaque ligne par un procédé tout à fait automatique. Le coût de constitution de ce fichier ne serait pas très élevé, à peine plus qu'une simple recopie de bande. Et le travail manuel serait aussi très réduit et se bornerait à préparer les quelques 1000 étiquettes correspondant aux 1000 textes différents, ce nombre étant multiplié par le nombre moyen de sous-ensembles par texte, mettons 5, soit 5000 étiquettes d'une quinzaine de chiffres, soit un travail d'une semaine pour une équipe de 2 ou 3 personnes.

Du côté de la taille requise sur les disques, le calcul qui la définit est simple : on a affaire à une masse de 70 millions de mots qui, groupés par 10 en moyenne, occupent 7 millions de lignes, soit 7 millions d'enregistrements de 80 caractères, soit 560 millions de caractères. Un système peut-il supporter une telle masse en ligne, c'est-à-dire immédiatement accessible en une fraction de seconde ? La réponse est oui, si l'on a affaire à un engin de la puissance de l'IRIS 80 dont le centre de calcul de NANCY est équipé. Il faut savoir que les disques qu'on trouve sur le marché contiennent aisément 200 millions de caractères. Il suffirait d'en aligner 3. A titre de comparaison la banque de données technologiques implantée près de Nice sous le nom de Télésystèmes dispose d'une batterie de 64 unités de disques, soit 13 milliards d'octets on-line.

Si l'on objecte que le centre de calcul de Nancy ne possède pas de disque à si haute capacité, il suffit de multiplier par 2 le nombre de disques en réduisant de moitié la capacité de chacun. Ainsi 6 disques de 100 millions d'octets rempliraient la même fonction. J'ajoute que l'on ne doit peut-être pas raisonner à partir des équipements existants, mais à partir de ceux dont on prévoit l'installation dans l'Est de la France au cours des prochaines années à Nancy ou à Strasbourg, et qui seront probablement du type de ceux dont on vient d'équiper le centre de Grenoble, c'est-à-dire une machine Honeywell C.I.I. dotée d'une grande capacité de stockage et apte à travailler en mode conversationnel , ce qui est très exactement le profil adapté à une base de données.

Si l'on craint enfin que le gel permanent de trois unités de disques, pour le profit d'utilisateurs dispersés dans l'espace et dans le temps, puisse n'être pas rentable au début, il faut répondre que c'est le sort de toute banque de données au stade initial. La clientèle est un continuum qui part de zéro et n'atteint que progressivement le volume optimal et rentable. Tandis que la banque de données est un tout quasi indivis, un produit qu'il faut offrir opérationnel dès le premier jour de fonctionnement. L'adéquation

entre l'investissement et le bénéfice ne se réalise qu'à moyen terme. Il en est ainsi par exemple de Télésystèmes qui au bout de deux ans est encore largement déficitaire. On peut toutefois envisager un accommodement dans la période de lancement, qui consisterait à utiliser pour la banque des disques amovibles qu'on mettrait en place à certaines heures de la journée réservées à l'interrogation de la banque, certains jours de la semaine.

Comme on peut voir, la constitution de ce fichier a été conçue pour utiliser au mieux les fichiers existants, et sans remonter à leur source la chaîne des traitements, et notamment sans qu'il soit nécessaire de reprendre des bandes perforées vieilles de 10 ou 15 ans.

2) Le fichier des mots

Par contre, dans l'étape suivante que nous allons décrire, les fichiers existant à Nancy paraissent inadaptés à la fonction souhaitée. On pense immédiatement aux fichiers-répertoires, sorte de fichiers inverses ou d'index qui, pour chaque forme rencontrée dans un texte-machine donné, précisent le numéro d'ordre du mot dans le texte-machine et le numéro de la fiche-texte où on le retrouve. Précisons qu'un texte-machine est un découpage artificiel d'au plus 100000 mots qui recoupe rarement les divisions naturelles des textes, et qu'il faut abandonner définitivement ce lit de Procuste. Je reprocherai d'abord aux fichiers-répertoires de n'être pas standard. Conçue en des temps révolus, leur structure pose des problèmes d'accès. Plus gravement, ces fichiers-répertoires sont étroitement associés aux fiches-textes, auxquelles nous avons renoncé dans notre fichier de base au profit des lignes. Or ces index ne précisent pas l'emplacement exact de la forme, et notamment le numéro de la ligne, mais seulement une zone de 8 lignes ce qui n'est pas une localisation suffisante. Certes on pourrait imaginer un sous-programme qui, à partir de la fiche-texte fournie par la clé associée au mot étudié, rechercherait le mot en question en explorant les 8 lignes de la fiche. L'interrogation serait plus longue et plus coûteuse si 1 'on considère que pour un mot donné le travail devrait être recommencé à chaque interrogation. Et comment résoudre l'ambiguïté lorsque la même forme est employée plusieurs fois dans le même contexte de 8 lignes ? Enfin ces fichiers-répertoires, qui ignorent les lignes, ignorent aussi chacun des paramètres que nous avons incorporés à notre fichier de base et qui concernent la date, le genre, l'auteur, le texte et le sous- ensemble du texte.

Il nous faut donc créer notre fichier inverse à partir du fichier de base tel que nous l'avons constitué dans la première phase. A vrai dire, l'informaticien peut faire tout à la fois, dans le même traitement et le fichier de base et le fichier inverse, comme il peut aussi séparer les deux étapes.

Pour la clarté pédagogique, supposons constitué notre fichier de base. Nous allons faire défiler l'une après l'autre, les informaticiens disent en accès séquentiel, les 7 millions de lignes (ou d'enregistrements) que nous aurons préalablement contrôlées (par programme et par sondage). En une fraction de seconde la ligne est explorée, les mots détachés les uns des autres et rangés au fur et à mesure de leur apparition dans un fichier de mots (et non plus de lignes) où la forme rencontrée est transcrite, accompagnée du numéro de la ligne en cours, ce qu'on appelle la clé. Ici des problèmes linguistiques se posent pour lesquels il faudra trancher : qu'appelle-t-on séparateurs de mots ? Comment traiter le trait d'union et l'apostrophe ? Faut-il reprendre les options choisies il y a 15 ans ? Pour ma part, je serais d'avis que oui, afin de rendre compatible cette base de données avec

les résultats déjà acquis et publiés. De toute façon, l'arbitraire est le terminus où débouchent toutes les voies que l'on pourrait choisir en ce domaine.

Afin de ne rien perdre de l'information, je proposerais d'ajouter à la clé de référence, c'est-à-dire au numéro de ligne, le numéro d'ordre du mot dans la ligne. Et quand chaque ligne est épuisée, j'ajouterais au fichier de base une information nouvelle : le nombre de mots contenus dans la ligne. Deux octets suffisent à transcrire dans l'un et l'autre fichiers cette information qui ne fait pas partie de la clé mais qui peut être utile lorsqu'on aura affaire à des questions portant sur un contexte de n mots ou sur la place des mots. Dans tous les traitements réalisés jusqu'ici à Nancy, on a jugé utile de porter à 32 le nombre maximum de caractères contenus dans un mot - ce qui paraît sage si l'on tient compte des mots composés. Les enregistrements du fichier à ce stade intermédiaire auront donc : 32 + 8 = 40 caractères au maximum. S'ils sont de longueur fixe, on aura besoin de stocker 3 milliards de caractères et s'ils ont un format variable, 1 milliard. On utilisera quelques bandes magnétiques que l'on soumettra ensuite au tri et à la fusion. On procédera ensuite à l'opération de tassement du fichier trié, afin de gagner de la place et pour une forme donnée on se contentera d'une transcription unique accompagnée de toutes les références de la forme dans le corpus. Toutes ces opérations de tri, de fusion et de tassement appartiennent à la routine. La seule difficulté dans l'application présente relève non de la taille des fichiers - car l'ordinateur en a rencontré de bien plus gros encore sans s'effrayer - mais de la nécessité de faire un sort spécial aux mots fréquents. S'il s'agit d'un mot de fréquence basse ou moyenne, on aura des enregistrements variables conçus sur le modèle suivant :

	Forme	Code gramm.	Fréquence	clé 1	clé 2	clé 3	etc.
Nombre chiffres			7	7 2 2	7 2 2	7 2 2	
Nombre octets	32	1	4	4 2 2	4 2 2	4 2 2	

L'information donnée par la fréquence n permet de lire les n référence qui suivent. Dans le cas d'un hapax on aura donc un enregistrement de

32 + 1 + 1 + 4 + 8 = 46 caractères.

Si la fréquence est de 1000, il faut réserver un espace de plus de 6000 octets. Mais on ne peut imaginer que pour une fréquence de 3 millions d'occurrences (la préposition *de*) on constitue un enregistrement de 18 millions de caractères. Dans le cas des mots fréquents, on devrait prévoir une possibilité de dépassement, ou de remplissage successif du buffer d'entrée par paquets de 1000 renvois. On aurait donc pour ces mots autant d'enregistrements que de milliers de renvois. Un code spécial en position 33 indiquerait l'existence d'une suite. Quelle serait la taille d'un tel fichier inverse ? Chaque renvoi exigeant 8 caractères, on aurait besoin

de 70 millions x 8 = 560 millions d'octets,

à quoi il faut ajouter la transcription des formes et des fréquences

soit 200000 x 40 = 8 millions de caractères.

II - L'INTERROGATION DE LA BASE DE DONNEES

Une fois constitués le fichier de base (fichier des lignes) et le fichier inverse (fichier des mots), le système d'interrogation peut déjà fonctionner. Si, par exemple, je m'intéresse au substantif nature, je peux interroger le fichier inverse qui m'indiquera la fréquence totale de ce mot et si je ne suis pas découragé par la multitude de ses occurrences, je peux demander le détail de ses localisations. Plutôt que d'éditer tout uniment le numéro de la ligne du fichier de base, le système pour être plus explicite utilisera ce numéro du fichier inverse comme clé du fichier de base et trouvera à la ligne ainsi repérée les codes de date, de genre, d'auteur et de texte qui seront convertis en langage naturel par la consultation de tables appropriées. L'édition en clair de toutes les références apparaîtra sur l'écran ou le téléscripteur de l'utilisateur. Si ce dernier est plus exigeant encore, le système lui fournira le contexte d'une ou plusieurs lignes, ou un contexte plus naturel comme celui de la phrase. Mais l'utilisateur peut aussi restreindre et préciser le champ de recherche et n'être intéressé que par les emplois d'une forme

– qu'on rencontre avant ou après telle date ou entre deux dates,
– qui appartiennent à tel genre littéraire,
– qu'on relève chez tel auteur,
– que l'on trouve dans tel texte ou telle partie de texte.

Ces critères, dont la liste n'est pas limitative, devront être définis au moment de l'interrogation grâce à une conversation assez semblable à celle du programme MISTRAL. Par exemple si l'on veut connaître les occurrences du mot *nature* chez HUGO, entre 1840 et 1860, et dans les seuls ouvrages en vers, on peut imaginer un dialogue du type suivant :

1, IDEN, DUPONT
2, CRPR,/FORME/nature
3, CRSE,/DATE/ > 1839 ETL 1861
4, CRSE,/GENRE/VERS
5, CRSE,/AUTEUR/HUGO
6, RCSE, AUTEUR ET GENRE ET DATE 7, EDIT, DATE, GENRE, TEXTE, LIGNE (3)

L'ordre EDIT permet à l'utilisateur de choisir sa mise en page, c'est-à-dire les éléments de la fiche qui l'intéressent et l'ordre dans lesquels il souhaite les voir apparaître ; les différents enregistrements intéressés par la question seront délivrés dans l'ordre chronologique qui a présidé à l'élaboration du fichier de base. L'ordre EDIT en particulier permet de préciser si la concordance qu'on souhaite est de 1, 3, 5, n lignes, ou si elle est d'une phrase, ou si elle se réduit à 3, 5, 7, n mots. Le dialogue d'interrogation peut être considérablement simplifié, si l'on se contente des options par défaut. On peut négliger de préciser la date, l'auteur ou le genre et dans ce cas c'est la totalité du corpus qui est explorée pour le mot intéressé. L'édition sera standard si l'on omet l'ordre EDIT et dans ce cas elle ne donnera pas le contexte.

1) Le critère principal

Le critère principal ne peut être absent, sans quoi l'interrogation n'a pas de sens. Mais il peut être double, multiple, variable ou soumis à diverses contraintes.

a) par exemple on peut demander la liste de plusieurs formes non seulement nature mais aussi monde, homme, Dieu, etc.

b) mieux même, la liste peut être indéterminée et dans ce cas les mots du fichier seront restitués s'ils satisfont aux critères secondaires. Par exemple, en choisissant le symbole * et en précisant un texte particulier, on pourra obtenir l'index alphabétique de ce texte.

c) si on veut limiter cet index et exclure tel ou tel mot, il faudra utiliser l'opérateur SAUF. Ainsi les deux ordres

CRPR,/FORME/ * SAUF DE
CRSE,/TEXTE/LES MISERABLES

provoqueront l'édition de l'index des Misérables sans la préposition de. L'opérateur SAUF peut être suivi de plusieurs mots qui seront pareillement exclus.

2) La catégorie grammaticale

La notion de mot grammatical devra être connue du système si l'on veut que l'exclusion concerne cette catégorie encombrante. Cela suppose que le fichier inverse soit pourvu d'un code grammatical, tel qu'on le trouve sur les dictionnaires de fréquences qui existent déjà. Bien sûr, le cas des homographes pose problème : on le résoudra sans élégance en leur attribuant un code spécial, H par exemple. Au niveau de l'interrogation on utilisera l'ordre CATEGORIE qui précisera les catégories désirées et celles qui sont exclues. Par exemple l'ordre

CRSE,/CATEGORIE/ * SAUF G

signifie qu'on désire toutes les catégories sauf les mots grammaticaux.

3) La lemmatisation

Le fichier inverse est un fichier de formes non un fichier de vocables ou entrées de dictionnaire. On doit pouvoir cependant interroger la base sur un vocable plutôt que sur une forme, ce qui pose un nouveau problème, celui de la lemmatisation. Il existe déjà un fichier de correspondance graphie- vedette qu'on pourrait reprendre sous la forme vedette-graphie et qui constituerait le préalable nécessaire par où passerait le système avant de renvoyer aux formes constitutives du vocable demandé. On aurait alors un ordre VOCABLE sous la forme suivante :

CRPR,/VOCABLE/AMOUR, AIMER

ce qui permettrait d'atteindre toutes les occurrences du substantif amour, au singulier ou au pluriel, ou du verbe aimer, à toutes les formes de la conjugaison.

4) Les composants du mot : racines, suffixes, préfixes, caractères et chaînes de caractères

Dans certains cas, un résultat assez proche de la lemmatisation peut être obtenu directement si l'on dispose d'un opérateur qui masque la fin ou désinence d'un mot. On pourrait alors choisir le symbole + collé à la racine, comme dans le système MISTRAL. Ainsi l'ordre

CRPR,/FORME/AMOUR+

permettrait d'atteindre non seulement amour et amours mais aussi amoureux, amoureuse, amoureuses, etc. ...

Ainsi pourrait-on distinguer non seulement le radical d'un nom ou la racine d'une famille mais aussi les préfixes ou l'élément initial des mots composés. De plus le symbole de masque pourrait être utilisé non seulement en position finale, mais aussi en position initiale, ou intérieure, ce qui permettrait de délimiter une chaîne de caractères, voire un caractère unique. Ainsi l'ordre

CRPR,/FORME/+TION

donnerait accès au suffixe TION, tandis que l'ordre

CRPR,/FORME/+CH+

restituerait toutes les formes qui contiennent la graphie CH en position initiale, intérieure ou finale.

5) Les cooccurrences

Les recherches qui portent non sur les combinaisons de lettres mais sur les cooccurrences de mots devraient être rendues possibles par des opérateurs spécifiques, ceux qu'on trouve dans les systèmes de documentation automatique, soit AND OR ADJ. Par exemple la commande

CRPR,/FORME/PROBITE ADJ CANDIDE

nous dira si d'autres que Booz sont ainsi vêtus. La contrainte exercée par l'opérateur ADJ est plus forte que celle de AND qui n'exige pas la contiguïté immédiate, mais une proximité moins étroite qu'on pourrait limiter à un contexte de n lignes, ou de n mots de part et d'autre. Je pencherais assez pour la solution d'un contexte de 3 lignes. Le système n'aurait d'ailleurs pas à rechercher ces 3 lignes, car le fichier inverse peut suffire à donner la réponse par la simple comparaison des renvois de probité et de candide, pour reprendre notre précédent exemple. Si deux clés coïncident ou ne diffèrent que d'une unité, la cooccurrence immédiate est établie. Naturellement on peut associer les opérateurs AND OR ADJ et par un jeu de parenthèses affiner la question. Par exemple :

CRPR,/FORME/(SYSTEME OR THEORIE)ADJ(METAPHYSIQUE OR PHILOSOPHIQUE)

6) Les homographes

Le sous-programme précédent qui relève les cooccurrences peut aussi servir à dissoudre certains cas d'homographie, et notamment l'ambiguïté très fréquente qui mêle les emplois du verbe et du substantif, sur le modèle la marche/ je marche. Charles MULLER a montré que certains algorithmes simples et peu coûteux étaient efficaces dans plus de la moitié des cas, même en ne remontant pas loin dans la chaîne syntagmatique et en se contentant d'un trousseau restreint de critères (par exemple la présence immédiate d'un démonstratif comme ce devant le substantif ou celle d'un pronom comme je devant le verbe).

L'ordre CATEGORIE que nous avons déjà vu pourrait alors être complété par une option HOMOGRAPHE. Ainsi en écrivant

CRSE,/CATEGORIE/VERBE (HOMOGRAPHE)

on obtiendrait parmi les formes recensées celles qui appartiennent à la catégorie du verbe, par nature ou par position, ainsi que les cas que l'analyse syntaxique n'aura pu régler, et qui seront soumis, pourvus d'un code interrogatif, à l'examen du chercheur.

7) La statistique

Peut-on en interrogeant la base aborder le domaine de la statistique ? Je ne parle pas ici des statistiques de routine que le système prend en compte en même temps que la facturation et qui permettent de savoir quels auteurs, quels mots, quels textes ont été soumis à l'interrogation, ce qui peut intéresser non seulement le créateur de la base de données, mais aussi le critique, l'historien ou le sociologue de la littérature ou de la langue. Mais je pense ici plutôt aux études linguistiques et principalement lexicologiques qui se fondent sur la statistique et qui mettent en rapport les fréquences des mots. Par exemple un chercheur qui s'intéresse à un mot dans un texte aimerait savoir si ce mot est spécifique de ce texte, c'est-à-dire savoir si la fréquence de ce mot est suffisamment forte pour qu'il puisse être rangé dans le vocabulaire significatif du texte en question, ou de l'auteur en question, ou de l'époque ou du genre en question. Pour que le système puisse répondre à des voeux de ce genre, il faut qu'une table ait été préalablement constituée donnant la taille de tous les textes et de leurs parties, de tous les auteurs, de toutes les tranches chronologiques, de tous les genres, mais aussi de chaque genre à l'intérieur de chaque tranche, ce qui est très simple à réaliser à partir du fichier des lignes. Mais il faut aussi disposer d'un dictionnaire des fréquences où toutes les formes du corpus (il y en a environ 200 000) apparaîtront avec leurs fréquences dans chacune des tranches, chacun des genres et au croisement des tranches et des genres. Ce fichier existe déjà sous une forme un peu lâche. Concentré, le fichier ne dépasserait pas 60 millions de caractères. Pour le consulter, il faudrait prévoir un ordre particulier par exemple STATISTIQUE et un opérateur, par exemple BASE, ce dernier introduisant l'ensemble (la tranche ou le genre ou une tranche dans un genre ou le corpus entier) qu'on choisit pour norme. Ainsi l'ordre

CRSE,/STAT1STIQUE/TEXTE BASE TRANCHE

demande qu'on choisisse pour norme l'époque à laquelle appartient le texte choisi et que, pour le ou les mots précisés, soient calculés les fréquences théoriques, les écarts réduits et les probabilités attachées à ces écarts. Bien entendu la loi normale peut céder la place, dans ces calculs, à la loi hypergéométrique qui est plus exacte mais plus coûteuse. Peut-être peut-on laisser le libre choix à l'utilisateur qui préciserait l'option de calcul (NORMAL ou HYPER) dans une parenthèse.

Doit-on souhaiter d'aller plus loin dans cette voie et d'offrir au calcul les fréquences par auteur ? Cela obligerait à constituer un dictionnaire des fréquences des 350 auteurs du corpus, ce qui exigerait beaucoup de place (300 millions de caractères environ). En se contentant des 50 auteurs les mieux représentés le fichier des fréquences passerait de 60 à 120 millions de caractères, ce qui est raisonnable. Avec ce complément les questions posées pourraient prendre des formes plus complexes, par exemple :

TEXTE BASE AUTEUR

OU AUTEUR BASE CORPUS

De plus les tranches pourraient s'ajouter les unes aux autres, ou former une intersection avec le genre, comme dans les exemples suivants :

TEXTE BASE TRANCHE (1 + 2 + 3 + 4)

ou

TEXTE BASE (TRANCHE AND GENRE)

Si l'utilisateur désire faire lui-même les calculs et souhaite seulement les éléments du calcul, un ordre plus simple pourrait être lancé qui demanderait seulement les informations du dictionnaire des fréquences. Par exemple pour un mot précisé préalablement l'ordre :

/FREQUENCE / TRANCHE

ou

/FREQUENCE / GENRE

donnera le nombre d'occurrences du mot considéré dans la période ou le genre qui enveloppent le texte choisi.

Bien entendu on pourra demander toutes les périodes ou tous les genres ou le croisement d'un genre ou d'une période ou certaines périodes cumulées dont les numéros d'ordre seront reliés par le signe +. Et si on dispose d'un dictionnaire des fréquences des auteurs l'ordre fréquence pourra demander la fréquence du mot chez un ou plusieurs auteurs :

/FREQUENCE / HUGO AND BALZAC

J'ajoute que de toute façon des éléments statistiques doivent être incorporés à la base, au moins pour prévenir le consultant du volume approximatif des réponses qui lui seront fournies, ce qui peut le prémunir contre les dépenses inconsidérées et l'aider à préciser (ou à élargir) le champ de l'interrogation. Par exemple le système, consulté sur une forme, éditera immédiatement la fréquence de cette forme dans le corpus et, si des contraintes de genre, d'époque ou d'auteur ont été imposées par l'utilisateur, le système, avant toute recherche effective, calculera et imprimera la fréquence théorique de la forme en question dans le sous-ensemble ainsi limité. De même, avant que soit réalisé l'index, complet ou limitatif, d'un texte, l'utilisateur sera averti de l'importance attendue des sorties, et pourra renoncer à donner suite à sa demande, ou la maintenir en différé ou la détourner sur les imprimantes rapides du site central.

La transmission des textes n'a pas été prévue dans ce système. C'est pourtant ce qui serait le plus facile à faire, s'il n'y avait des problèmes juridiques à poser et à trancher d'abord. Imaginons un instant qu'ils soient réglés. Alors notre système d'interrogation pourrait ajouter une commande supplémentaire, par exemple :

/COPY / Texte (ou sous texte ou auteur)

Bien entendu le système avertira l'usager que la transmission du ou des textes envisagés occupe tel ou tel volume, afin que les possibilités de stockage de l'usager soient mises en rapport avec la masse de documents attendus. D'ailleurs il faudrait réserver ce mode de transmission aux petits volumes et s'en tenir à la capacité des disquettes actuelles, au standard IBM, soit 250 000 caractères. La duplication des textes longs devrait emprunter d'autres voies, la poste par exemple. J'imagine aussi qu'on pourrait commercialiser facilement quelques disquettes et les offrir au marché du secondaire. Si l'on place dans

l'Education Nationale les 10 000 micro¯ordinateurs envisagés, il faudra bien donner une matière aux exercices des élèves. Et cette question déjà abordée hier est à traiter avec l'organisme qui a succédé à l'IPN et avec le responsable de la prospective et de l'informatique au Ministère de l'Education, Monsieur TREYSSEL.

Conclusion

1) La modularité

Si la base de textes constitue un tout auquel on pourra difficilement retrancher ou ajouter, par contre les différentes options que nous venons d'énumérer et qui précisent le mode d'interrogation peuvent n'apparaître que progressivement, au fur et à mesure que se préciseront les besoins et que seront élaborés les programmes et les outils correspondants. L'essentiel est que la base soit conçue de telle façon qu'elle permette ces extensions de la consultation.

2) La faisabilité

A Nancy (mais aussi à Paris) existent depuis longtemps de multiples logiciels qui donnent réponse à la plupart des points ici soulevés. L'auteur de ce projet a lui-même une longue pratique des données dépouillées à Nancy et il a été confronté aux divers problèmes qu'il évoque ici. A partir des fiches-textes il a eu à constituer des index, des concordances-phrases, à résoudre les problèmes de lemmatisation et d'homographies, à distinguer les constituants du mot, et surtout à assurer l'exploitation statistique de ces données. Pour toutes ces recherches des programmes expérimentaux, en Cobol, PLI et Fortran, ont été réalisés, ce qui l'autorise à penser qu'ils sont réalisables.

3) La fiabilité

Il est souhaitable toutefois que des professionnels et des praticiens éprouvés de l'informatique soient chargés de la réalisation du projet, afin d'assurer la fiabilité, l'optimisation, la généralité et la transparence des traitements et, en fin de compte, la rentabilité du projet. Cela suppose qu'on attache à cette réalisation un informaticien de haut niveau, rompu aux techniques les plus récentes de la télématique et des bases de données.

4) Les contraintes

Le fait que les données soient déjà constituées représente tout à la fois un avantage et un handicap. L'avantage est de disposer d'un coup de la plus grosse masse de données textuelles jamais constituée dans le monde. Et d'une masse cohérente : car d'un bout à l'autre du dépouillement, les mêmes principes ont été respectés. Et sans doute certains de ces principes méritent révision, eu égard au progrès des techniques et des matériels. Je regrette en particulier la recomposition malencontreuse des lignes du texte, et un codage insuffisamment précis des signes de ponctuation ambigus. Mais dans l'ensemble la perte d'information a été évitée et la constance des choix maintenue.

Cet avantage décisif crée des obligations et des contraintes. L'obligation est de réaliser une base digne des données, dont la souplesse et la sûreté répondent à l'ampleur et à l'homogénéité des données. La contrainte vient de la présence et de la nature

de ces données, qui ont été constituées avant la base, et avant même toute conception d'une base de données. Il y a là une inversion du processus à laquelle il faut s'adapter.

5) Les échéances et les coûts

Comme notre projet est un compromis systématique qui s'accommode au mieux des données de départ, les échéances et les coûts devraient être réduits d'autant. Point de toilettage trop onéreux des textes, pour rétablir les lignes originelles du texte ou pour imposer de meilleures éditions ou pour souligner les articulations de l'énoncé. Cela me semble devoir être trop long et trop cher. A ce compte il vaudrait peut-être mieux abandonner les dépouillements réalisés et s'entendre avec les éditeurs pour la cession directe de la bande d'origine qui sert à l'impression ou à la photocomposition, ou, ce qui me semble moins incertain, adapter un lecteur optique à la typographie d'une grande collection littéraire comme celle de la Pléiade.

Le traitement ne représente qu'une valeur ajoutée à un produit qui n'est livré que demi-fini entre les mains du chercheur. Et ce traitement est lui-même simple de conception et, je l'espère, de réalisation. Ce caractère devrait permettre de rapprocher les échéances. Car il ne convient pas de trop creuser l'écart entre le moment où l'on accumule les richesses et celui où on les exploite. Si on tarde trop les biens entassés se dégradent, se démodent et se dérobent.

Le projet a été conçu dans une optique conversationnelle. Si on estime qu'il s'agit là d'un luxe peu nécessaire, on peut réaliser une version batch de ce même projet avec des coûts moindres dans la réalisation et l'exploitation. Les critères de l'utilisateur seraient fournis sous forme de paramètres. On peut concevoir aussi un système mixte où l'interrogation serait conversationnelle et la recherche différée en batch. Il faut se rendre compte toutefois que renoncer au conversationnel c'est aller à contre-courant de l'évolution de l'informatique et amoindrir (en rapidité et en souplesse) la qualité du service.

En résumé trouvons un homme qui puisse consacrer 1 ou 2 années à cette réalisation. Trouvons aussi une machine qui puisse disposer d'l ou 2 milliards de caractères on line. Trouvons enfin un commanditaire qui puisse débourser une ou deux centaines de milliers de francs. Ce serait suffisant au moins pour réaliser l'étude préalable et un logiciel expérimental et pour jeter les bases de notre base.

Etienne BRUNET

Pièces jointes : exemples d'index et de concordances-phrase de l'Émile, réalisés à partir des données de l'ILF et publiés chez Slatkine-Champion dans la collection des Index de J.J. Rousseau. (Les codes alphabétiques, de A à G, indiquent la zone de la page dans l'édition de la Pléiade.)

	LIV.1	LIV.2	LIV.3	LIV.4	LIV.5
	296 d	.	.	.	825 e
	296 e	.	.	.	.
	7	4	0	2	6
accentuée		2			
	285 e	404 e	.	.	.
accentüent		1			
	.	404 g	.	.	.
accepta		1			
	.	.	.	.	789 g
accepté		1			
	264 b	.	.	.	.
accepte		4			
	.	.	.	.	789 f
	.	.	.	.	805 f
	.	.	.	.	806 f
	.	.	.	.	863 g
accepter		5			
	264 b	306 b	440 a	.	759 f
	.	.	.	.	823 g
acceptera		1			
	.	.	.	.	753 f
accepterez		1			
	.	334 e	.	.	.
acceptés		1			
	.	.	.	521 d	.
acception		1			
	.	313 b	.	.	.
acceptions		1			
	.	.	.	548 a	.
acceptoit		1			
	.	.	.	.	760 g
acceptons		1			
	.	.	.	.	758 a
acçés		1			
	.	328 b	.	.	.
accessoire		1			
	.	.	.	.	832 a
accessoires		1			
	.	.	.	499 f	.
accidens		10			
	252 e	376 a	.	524 a	712 g
	259 d	378 b	.	537 g	.
	.	388 e	.	565 e	.
	.	390 e	.	.	.
	2	4	0	3	1
accident		3			
	259 a	375 e	.	.	812 f
accidentelles		1			
	.	.	.	.	851 b
accidentels		1			
	243 d	.	.	.	.
acclamations		2			
	.	394 c	438 b	.	.
accomodant		1			
	.	.	.	528 a	.
accomode		4			
	.	396 a	.	690 e	790 f
	.	.	.	.	805 d
accomoder		3			
	.	.	.	.	749 e
	.	.	.	.	761 g
	.	.	.	.	778 f

	LIV.1	LIV.2	LIV.3	LIV.4	LIV.5
accompagné			2		
	285 f	.	.	.	810 d
accompagne			4		
	.	405 g	.	519 b	809 c
	.	.	.	592 e	.
accompagnée			2		
	.	.	.	567 a	742 f
accompagnées			1		
	.	.	.	651 a	.
accompagnement			1		
	.	.	.	.	747 e
accompagnent			2		
	.	392 e	.	515 c	.
accompagner			2		
	.	.	.	.	717 c
	.	.	.	.	754 d
accomplir			1		
	.	.	.	611 e	.
accompliasement			2		
	.	.	.	617 d	.
	.	.	.	617 d	.
accomplit			2		
	.	.	.	.	740 d
	.	.	.	.	810 b
accord			17		
	247 c	334 a	.	548 c	757 f
	256 b	334 b	.	580 b	843 b
	274 e	366 c	.	598 e	843 b
	.	.	.	598 g	.
	.	.	.	602 d	.
	.	.	.	605 g	.
	.	.	.	627 a	.
	.	.	.	660 g	.
	3	3	0	8	3
accordant			1		
	.	.	.	.	798 d
accordé			1		
	.	.	.	.	712 a
accorde			11		
	251 f	332 d	.	494 c	730 d
	.	332 d	.	514 g	766 b
	.	.	.	556 a	790 f
	.	.	.	.	817 e
	.	.	.	.	866 g
	1	2	0	3	5
accordée			1		
	.	.	.	.	780 b
accordent			7		
	261 g	371 d	.	530 a	720 c
	.	.	.	568 d	766 d
	.	.	.	.	863 d
	1	1	0	2	3
accorder			10		
	251 b	312 e	.	515 a	734 g
	264 d	422 a	.	585 a	782 d
	290 c	.	.	.	.
	290 d	.	.	.	.
	4	2	0	2	2
accordera			1		
	.	401 a	.	.	.
accorderai			1		
	.	.	.	574 c	.

abus LIVRE I PAGE 258 D F. 21 Z.+01,7
Ainsi de ce
—) seul abus corrigé resulteroit bientôt une réforme générale ;

abus LIVRE I PAGE 294 D
—) Mais un abus d'une toute autre importance et qu'il n'est pas moins aisé de prévenir est qu'on se presse trop de les faire parler, comme si l'on avoit peur

abus LIVRE II PAGE 310 A
Mais cet attachement peut avoir son excès, son
—) défaut, ses abus. Des parens qui vivent dans l'état civil y transportent leur enfant avant l'age.

abus LIVRE II PAGE 409 F
nature, tenant immédiatement au sens, et que la seconde est un ouvrage de l'opinion sujet au caprice
—) des hommes et à toutes sortes d'abus. La gourmandise est la passion de l'enfance ;

abus LIVRE III PAGE 462 E
En toute chose il importe de bien
—) exposer les usages avant de montrer les abus.

abus LIVRE IV PAGE 495 B
Mais on peut se tromper sur les causes, et souvent attribuer au physique ce qu'il faut imputer
—) au moral : c'est un des abus les plus fréquens de la philosophie de nôtre siècle.

abus LIVRE IV PAGE 582 E
Ton génie
dépose contre tes principes, ton coeur bienfaisant
—) dément ta doctrine, et l'abus même de tes facultés prouve leur excellence en depit de toi.

abus LIVRE IV PAGE 587 C
Mais
—) elle a tellement borné ses forces que l'abus de la liberté qu'elle lui laisse ne peut troubler l'ordre général.

abus LIVRE IV PAGE 587 G
—) C'est l'abus de nos facultés qui nous rend malheureux et méchans.

abus LIVRE IV PAGE 633 A
état de parler aux hommes, ne leur parlez jamais que selon vôtre conscience, sans vous embarrasser s'ils
—) vous applaudiront. L'abus du savoir produit l'incrédulité.

abus LIVRE IV PAGE 664 B
Qui croit devoir fermer
les yeux sur quelque chose se voit bientot forcé de
—) les fermer sur tout, le prémier abus toléré en amène un autre, et cette chaine ne finit plus qu'au renversement de tout ordre et au mépris de toute loi.

abus LIVRE V PAGE 700 B
les deux sèxes dans les mêmes emplois, dans les mêmes travaux, et ne peut manquer d'engendrer les
—) plus intolérables abus ; je parle de cette subversion des plus doux sentimens de la nature, immolés à un sentiment artificiel qui ne peut

abus LIVRE V PAGE 705 F
l'usage de ces corps de baleine par lesquels les nôtres contrefont leur taille plustôt qu'elles ne
—) la marquent. Je ne puis concevoir que cet abus poussé en Angleterre à un point inconcevable n'y fasse pas à la fin dégénérer l'espèce, et je

abus LIVRE V PAGE 708 D
Il y en
—) a bien peu qui ne fassent plus d'abus que d'usage de cette fatale science, et toutes sont un peu trop curieuses pour ne pas l'apprendre sans qu'on les y

abus LIVRE V PAGE 709 E
Pour prévenir cet
—) abus aprenez-leur surtout à se vaincre. Dans nos insensés établissemens la vie de l'honnête femme est un combat perpétuel contre elle même ;

abus LIVRE V PAGE 711 E
Il ne s'agit que
—) d'en prévenir l'abus.

abus LIVRE V PAGE 714 E
Cependant en général elles sont mises,
au rouge près, avec autant de soin que les dames,
—) et souvent de meilleur gout. L'abus de la toilette n'est pas ce qu'on pense, il vient bien plus d'ennui que de vanité.

abus LIVRE V PAGE 729 G
Voila la véritable
religion, voila la seule qui n'est susceptible ni
—) d'abus ni d'impieté ni de fanatisme. Qu'on en prêche tant qu'on voudra de plus sublimes, pour moi, je n'en reconnois point d'autre que celle-là.

abus LIVRE V PAGE 764 E
—) Voulez-vous prévenir les abus et faire d'heureux mariages ?

abus LIVRE V PAGE 826 C
—) L'abus des livres tûe la science. Croyant savoir ce qu'on a lû, on se croit dispensé de l'apprendre.

abus LIVRE V PAGE 841 D
raisonables et sans danger des engagemens qui sans cela seroient absurdes, tirranniques et sujets aux
—) plus énormes abus.

abusant LIVRE IV PAGE 587 B F. 2
Elle ne veut point le mal
—) que fait l'homme en abusant de la liberté qu'elle lui donne, mais elle ne l'empêche pas de le faire ;