Projet de base de données textuelles pour l'Institut de la Langue Française

Étienne Brunet

▶ **To cite this version:**

Étienne Brunet. Projet de base de données textuelles pour l'Institut de la Langue Française. bases de données dans les humanités et les sciences sociales, Jun 1980, Madrid, Espagne. hal-01431077

HAL Id: hal-01431077
https://hal.science/hal-01431077

Submitted on 10 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Etienne Brunet

Projet de base de données textuelles pour l'Institut de la Langue Française

Résumé

L'exploitation systématique de l'immense corpus du Trésor de la langue française, gros de 70 millions d'occurrences, a fait l'objet d'une Table Ronde à l'Institut de la langue française (27-29 mai 1980, Nancy). On rend compte ici des critiques et des propositions qui furent faites alors, et particulièrement d'un projet de création d'une base de données textuelles, applicable au corpus littéraire du XIX^{e} et du XX^{e} siècle. En constituant un fichier des lignes et un fichier inverse des mots, l'un et l'autre d'un demi-milliard de caractères, et en recourant à quelques fichiers annexes (de lemmatisation et de statistique), on disposerait d'une base de données, qu'un langage simple d'interrogation conversationnelle mettrait à la disposition de la communauté scientifique. Des index, des concordances et des statistiques de types variés seraient livrés au consultant, dont le choix pourrait s'exercer sur les formes, les vocables, les constituants du mot, les catégories grammaticales, les homographes, les cooccurrences, le vocabulaire significatif, la recherche pouvant en chaque cas être limitée à un texte ou sous-texte, à un auteur, à un genre littéraire ou à une époque particulière.

SUMMARY

A symposium was held at the Institut de la Langue Française (I.L.F.) in Nancy, France (27-29 mai May, 1980). It was devoted to the systematic exploitation of the huge corpus of the "Tresor de la Langue Française" (T.L.F.) which contains 70 million word-tokens.

This is an account of the criticisms and proposals that were then made, with special emphasis on a plan for creating a textual data base concerning the literary corpus of the 19 th and 20 th centuries. This would imply the creation of a line file and a reverse word order file both of half a milliard characters, and the use of a few auxiliary files (for lemmatization and statistics),and would provide a data base that a simple conversational interrogation language would make available to the community of scholars. Indexes, concordances and statistics of various description could be supplied to the user who would be free to direct his attention to word-forms, word-types, morphemes, parts of speech, homographs, collocations or significant vocabulary at will. In each case the field of investigation could be restricted to a single text or sub-text, to a single author, a whole literary genre or a particular period..

Lors de la première conférence sur les bases de données dans les humanités qui s'est tenue en août dernier en Amérique, j'ai eu l'occasion d'évoquer et même de détailler les trésors qui gisent à Nancy au sein de l'Institut de la Langue Française. Faut-il rappeler ici qu'au bout de quinze ans de capitalisation, les données textuelles qui y ont été amassées représentent actuellement 140 millions de mots, soit une bibliothèque de 2000 volumes sur support magnétique, ou si l'on veut un seul gros volume qui aurait 500 000 pages. De ces données la moitié a été exploitée à des fins lexicographiques, l'objectif premier étant de réaliser un dictionnaire de la langue contemporaine, le Trésor de la Langue Française, dont on prévoit l'achèvement aux environs de 1987 (la rédaction actuellement a parcouru la moitié de l'alphabet).

Les données exploitées concernent uniquement le corpus du XIX-XX^{e} siècle. Quant aux textes antérieurs, principalement ceux du XVII^{e} et du XVlll^{e}, leur enregistrement (non achevé) se justifie par des objectifs plus lointains qui peuvent être la rédaction de nouveaux dictionnaires représentatifs de la langue classique, ou de la langue du XVI^{e} siècle, ou la constitution d'un dictionnaire électronique qu'on pourrait consulter à l'aide d'un terminal et qui éviterait les écueils où tombent les dictionnaires traditionnels : l'encombrement de la masse totale mais aussi l'exiguïté de chaque article, le coût et la lenteur de l'impression, de la diffusion et de la consultation, et surtout le vice originel de toutes ces entreprises lexicographiques qui meurent le jour même où elles naissent, car la mise à jour y est impossible, sinon sous la forme incommode de suppléments. L'idée qui semble prévaloir à l'Institut de la Langue Française est qu'il faut s'orienter vers cette deuxième réalisation qui s'apparente directement à l'objet de ce colloque puisqu'il s'agirait d'une base de données lexicographiques.

I - LA TABLE RONDE DE NANCY (27-29 Mai 1980) sur les bases de données textuelles -

1 - Ce n'est pourtant pas de ce type de base de données que je veux parler aujourd'hui mais d'un sujet plus actuel puisqu'il vient de faire l'objet d'une Table Ronde de trois jours à Nancy, où les participants avaient été conviés à réfléchir aux problèmes d'une base de données textuelles. Je retrouve ici plusieurs des 45 participants invités à cette Table Ronde, dont la plupart venaient de l'extérieur de l'ILF. Car en voulant créer cette base de données textuelles on avait le souci d'ouvrir le Trésor à l’ensemble de

la communauté scientifique et cette communauté devait d'abord être entendue.

Il faut bien avouer que cette communauté fit entendre quelques critiques, et, comme cela arrive, les critiques étaient souvent contradictoires.

a) Certains s'effrayaient de la boulimie d'une machine à saisir, qui absorbe chaque année 5 millions de mots supplémentaires et qui engloutit des sommes considérables en vue de recherches seulement virtuelles. D'autres au contraire constataient avec regret les lacunes du corpus, où certains auteurs sont peu représentés, où manquent certains textes essentiels, où font défaut les notes, les variantes, les versions d'un texte dont on fixe un état et non le devenir, où le choix même de cet état du texte et de l'édition de référence prêtait à discussion. Dans ce corpus énorme chacun estimait qu'il y avait trop et trop peu : trop pour les autres et trop peu pour soi, trop peu pour l'objet particulier, littéraire, linguistique ou historique, où un chercheur se trouve engagé.

b) Les critiques portaient non seulement sur la constitution du corpus et le choix des textes, mais sur la représentation des textes en mémoire et sur la perte d'information qui résulte d'un système insuffisant de codage, où les signes de ponctuation ont été respectés jusque dans leur ambiguïté (qu'on songe aux fonctions bien distinctes du point : marque de fin de phrase mais aussi de suspension, d'abréviation ou de siglaison),et où ont été négligés les éléments qui permettent de situer le locuteur et d'attribuer le discours à l'auteur ou à ses personnages. Une fidélité typographique trop courte ou trop myope favorise la distorsion qui accompagne la transcription en linéaire continu d'une information qui se présente au lecteur sous la forme structurée d'une surface (et même d'un volume). L'ensemble des textes du corpus est déformé quand il est réduit à une ligne très longue qui ferait le tour de la terre. De ce côté les critiques ont été constructives, plusieurs systèmes de codage ayant été proposés (notamment par Lafon et Tournier) qui peuvent atténuer la perte d'information tout en ménageant la standardisation et la compatibilité.

c) A côté des critiques il y eut des comparaisons et plusieurs orateurs parlèrent de réalisations similaires qui sont en cours au Canada, à Oxford, à Paris, à Grenoble, à Toulouse et ailleurs.

d) A l'heure du bilan, on peut porter à l'actif de la Table Ronde de Nancy d'abord un peu plus de clarté terminologique, les notions de base et de banque. de données ayant fait l'objet de débats contradictoires, et si chacun ne met pas le même sens dans ces termes, chacun sait au moins en quel sens d'autres les emploient. Des décisions, il n'y en eut point et ce n'était pas le rôle de ce conseil de réflexion. Mais il y eut des voeux quasi unanimes : un voeu d'ouverture et de fédéralisme, chacun souhaitant une circulation plus rapide des textes, des données, des moyens et des idées. Et sur le sujet principal de cette Table Ronde, on peut avancer que le principe même d'une base de données n'a pas été combattu, même si certains continuent à s'interroger sur cette mode ou cette fièvre qui s'est emparée des milieux scientifiques. Ce qui a été discuté ce sont les modalités, les échéances, les moyens et les objectifs de cette base. Et je ne cacherai pas que des difficultés techniques - ou plus précisément financières - sont apparues le dernier jour lorsque les implications concrètes du projet ont été examinées.

2 - Ce projet n'est donc encore qu' un projet et s'il est mis un jour à exécution ce sera probablement sous une forme modifiée, simplifiée et planifiée selon les urgences du calendrier et les besoins des utilisateurs.

Pour analyser ces besoins une étude de marché s'impose. Elle devrait s'appuyer sur une enquête menée aux Etats-Unis par M. Morissey auprès d'une centaine de chercheurs spécialisés dans l'histoire ou la littérature françaises. Elle mettrait à profit également la synthèse opérée par Mme Martin de toutes les demandes, qui ont été présentées à l'ILF et qui ne sont pas négligeables puisque ces dernières années le service des prestations extérieures en a reçu 1500. Mais on peut imaginer que bien des demandes n'ont pas été présentées, ou n'ont même pas été exprimées, faute de savoir précisément ce qui peut être extrait des données de Nancy.

Il s'agit donc de tenir compte de besoins virtuels autant que réels et pour les mesurer il faut d'abord les éveiller par une sensibilisation du public scientifique, ce que soulignait le directeur de l'ILF, Bernard Quemada, à l'ouverture du Colloque et ce que le fondateur du Trésor, le recteur lmbs, résumait dans cette formule : du savoir d'abord, du savoir-faire ensuite, et enfin du faire savoir. Ce souci de faire

savoir est d'autant plus nécessaire que les données amassées sont plus dignes d'être exploitées. Depuis plusieurs années que je les explore je puis dire qu'elles présentent plusieurs avantages considérables qui restreignent la portée des critiques qu'on a entendues à Nancy : le premier est le plus évident c'est la masse énorme de la documentation qui est certainement jusqu'à ce jour la plus étendue qui- ait été constituée dans le monde en matière de langage. Le second réside dans la constance des options de saisie et de traitement. On peut n'être pas entièrement satisfait de ces options, regretter les insuffisances du codage, de la lemmatisation, de la distinction des homographes ou de celle des catégories grammaticales. Mais on doit reconnaître qu'elles n'ont pas changé en quinze ans et qu'elles s'imposent à tous les textes, à tous les auteurs, à toutes les périodes et à tous les genres littéraires - ce qui permet les confrontations et les études comparatives. A cette cohérence dans l'enregistrement s'ajoute l'homogénéité dans le corpus dont on peut contester qu'il soit représentatif de la langue française (car la langue parlée n'y figure guère) mais qui est certainement représentatif de la littérature française : le corpus du Trésor est délibérément littéraire et les grands écrivains y tiennent une place prépondérante. Et cela facilite également les études comparatives et la méthode la plus opératoire qui est la distinction des semblables. Un dernier avantage enfin peut être trouvé dans la longue expérience d'une équipe ancienne et nombreuse qui ne s'est pas contentée d'amasser les données et d'empiler les textes et qui a élaboré une série très abondante de traitements et de longues chaînes de programmes. Au point qu'on peut dire que la base de données existe déjà de façon latente, puisqu'elle dispose du matériau et des outils et qu'elle permet et a permis des formes partielles d'exploitation. Reste à affiner les matériaux, à les rendre disponibles immédiatement à l'ensemble des utilisateurs potentiels, en proposant un système d'interrogation simple et proche du langage naturel, et en faisant sauter le verrou technique qui décourage encore aujourd'hui les consultants. Voici comment l'on pourrait faire.

Il - LA CONSTITUTION DE LA BASE DES DONNEES TEXTUELLES

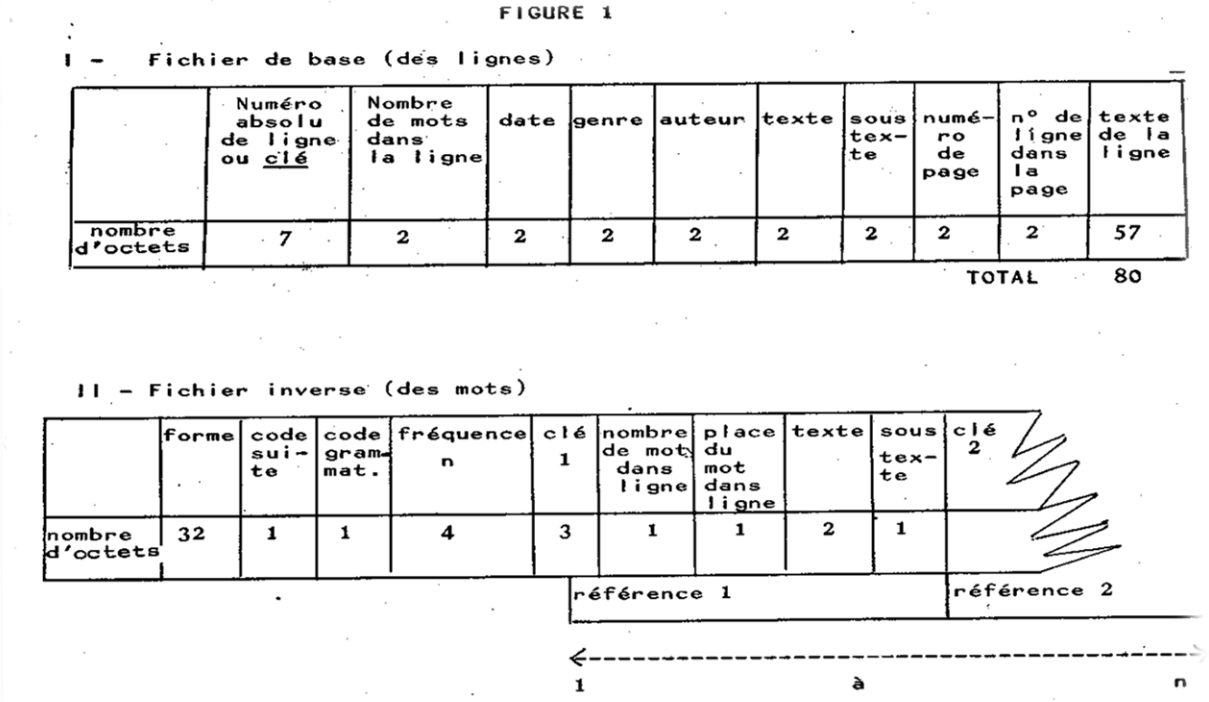
1 - Il faudrait d'abord délimiter l'étendue de la base. Il me semble que le corpus XIX-XX, gros de 70 millions d'occurrences, présente une priorité. Constitué depuis longtemps et support permanent de la rédaction du dictionnaire, il a été soumis à divers traitements de lemmatisation, de codage grammatical, de distinction des homographes qui résolvent une partie des problèmes. Et pour aller plus vite je proposerais qu'on ferme le corpus et qu'on en fasse une base autonome, déjà suffisamment

encombrante. Et cela n'exclut pas d'autres bases qu'on pourra établir plus tard à partir des textes du XVIII^{e} ou du XVII^{e} siècle.

2 - Une fois délimité, le corpus doit être unifié pour constituer le fichier de base. L'opération est déjà à moitié engagée puisqu'on dispose à l'heure actuelle de la totalité des textes en continu sur bande magnétique. Il suffit de transposer ce fichier séquentiel en fichier sélectif en y ajoutant une clé qui peut être un numéro de ligne de 7 chiffres puisqu'il y aura vraisemblablement quelques millions de lignes (aux environs de 7 millions si l'on compte en moyenne 10 mots par ligne sur un total de 70 millions de mots). Cela représente près de 600 millions de caractère, chaque ligne ayant un format fixe de 80 caractères répartis selon le schéma de la figure 1.

3 - La même figure 1 décrit également la structure du fichier inverse qui contiendra des mots (des formes) et des renvois aux enregistrements du fichier de base et dont la taille avoisine aussi le demi-milliard de caractères. Ce fichier est à constituer car on ne peut reprendre un produit qui existe déjà sous le nom de fichier-répertoire et qui répond mal au présent objectif puisqu'il ignore la localisation de la ligne. Le fichier inverse n'est pas obtenu d'un coup tel qu'on le montre dans la figure 1. Il faut d'abord dégrouper les mots dans chaque ligne, les compter et les entreposer dans un fichier qui sera soumis au tri, à la fusion et au tassement, toutes opérations de routine qui ne posent aucun problème. Le fichier inverse est au bout du compte un fichier de format variable, l'enregistrement ayant une longueur proportionnelle à la fréquence. Une limite s'impose toutefois pour les mots très fréquents qui donneront lieu à plusieurs enregistrements autant qu'ils contiendront de milliers d'occurrences. Un code spécial avertira que l'enregistrement courant comporte une suite. Un code grammatical sera également introduit en l'empruntant à certain fichier qui le possède déjà, le fichier de correspondance des graphies et des vedettes ou fichier de lemmatisation. Les homographes seront désignés par un code spécial qu'on peut extraire du même fichier.

FIGURE 1 - Fichier de base (des lignes)



III – L'INTERROGATION

1 - L'interrogation sollicite d'abord le fichier inverse en lui proposant une forme dont on veut connaître la fréquence et les références. Elle peut réclamer aussi le secours de fichiers annexes et notamment d'un fichier de lemmatisation où sont établies les correspondances entre les graphies et les vedettes. Par le truchement de ce fichier latéral on peut par exemple obtenir toutes les formes conjuguées d'un verbe dont on ne fournit que l'infinitif. Le fichier inverse contenant un code grammatical, certains choix et certaines limitations peuvent s'exercer à ce niveau, soit qu'on précise les catégories désirées, soit qu'on désigne les catégories exclues, grâce à l'opérateur SAUF. Le fichier inverse répond aussi, aux questions qui portent sur les constituants du mot, radicaux, racines, suffixes, préfixes, et enfin lettres ou combinaisons de lettres. L'opérateur de masque en position finale ou initiale permet de tronquer la chaîne de caractères qu'on veut isoler. Les renvois (ou références) contenant non seulement le numéro de ligne (ou clé), mais aussi le nombre de mots dans la ligne et la place du mot dans la ligne, le fichier inverse suffit à signaler les cooccurrences soit immédiates soit Grâce aux opérateurs AND, ADJ ou OR, on peut recenser les associations lexicales, les constructions syntaxiques, les clichés. Dans certaines recherches de contexte, le fichier inverse peut encore être consulté, par exemple lorsqu'on veut

distinguer les homographes par l'examen de l'environnement immédiat, ce qui peut se faire aussi en recourant au texte du fichier de base. Le même fichier inverse est consulté pour l'examen des critères limitatifs que sont les précisions de genre, d'époque, d'auteur, de texte ou de sous-texte. Ainsi on peut circonscrire le champ de la recherche à l'époque 1830-1840, aux textes en vers, à tel auteur ou à tel texte. Quand les contraintes sont fortes et que le champ est resserré - par exemple quand le champ ne dépasse pas l'étendue d'un texte - le critère primaire peut devenir indéfini: en utilisant le symbole * on signifie au système qu'on ne veut pas se limiter à une forme ou à un vocable, et qu'on les désire tous. On obtiendra alors l'index alphabétique du texte (ou du sous-texte) considéré.

Le tableau 2 ci-après regroupe les commandes et les opérateurs que le système offre au consultant. Remarquons qu'il n'est pas utile de détailler les précisions de genre ou d'époque à l'intérieur du fichier inverse, l'identification du texte suffisant à restituer la date et le genre, par la consultation préalable de tables appropriées. Le principe est évidemment de solliciter en priorité le fichier inverse, pour des raisons évidentes de rapidité, et de recourir le moins souvent possible au fichier de base.

2 - Le recours au fichier de base ne s'impose que lorsqu'on doit restituer le contexte qu'il s'agisse d'une phrase, de n lignes ou de n mots. La clé de chaque référence est alors utilisée pour repérer l'adresse du contexte souhaité.

3 - Notre système prévoit un fichier de fréquences qui peut jouer plusieurs rôles. Tout d'abord un rôle préventif pour éviter les dépenses inconsidérées. L'utilisateur peut ne pas mesurer précisément le volume des résultats demandés et le coût du traitement, et le danger est grand lorsqu'il s'agit d'index. Si le système connaît l'étendue de chaque auteur, de chaque texte ou sous-texte, comme aussi l'étendue des genres et des époques, et des genres dans chaque époque, il peut le signaler à l'utilisateur, lequel prendra sa décision en connaissance de cause. Le fichier inverse d'ailleurs fera les mêmes réponses de précaution s'il s'agit d'une forme fréquente. Dans notre schéma n° 5 la réponse initiale du système, préalable à la recherche effective, est indiquée en pointillé. Mais le fichier des fréquences peut jouer un second rôle et aider les chercheurs qui se soucient de linguistique quantitative et qui par exemple souhaiteraient circonscrire le vocabulaire significatif d'un texte ou d'un auteur, ou savoir si telle forme ou tel vocable est véritablement caractéristique

du texte ou de l'auteur. C'est la statistique qui répond à ces questions par l'emploi de la loi normale ou de la loi hypergéométrique.

TABLEAU 2

LA CONSULTATION DU FICHIER INVERSE (ET DU FICHIER DE LEMMATISATION)

COMMANDES			
	FORME	GENRE	TEXTE
	VOCABLE	DATE	SOUS TEXTE
	CATEGORIE	AUTEUR	

Exemples
/FORME/ nature
/FORME/ nature monde univers
/VOCABLE / aimer
/CATEGORIE / substantifs
/ GENRE / Vers
/ DATE / 1860
/ AUTEUR / Hugo
/ TEXTE / Madame Bovary

Opérateurs

1) totalité *
/ FORME / *

2) Exclusion SAUF
/ FORME / * SAUF de le la les
/ CATEGORIE / * SAUF grammaticaux

3) Masque +
/ FORME / amour +
/ FORME / anti +
/ FORME / + isme
/ FORME / + ch +
/ FORME / + w +

4) Limitation chronologique < >
/ DATE / > 1800 < 1811

5) Cooccurrences AND ADJ OR
/ FORME / feu ADJ rouge
/ FORME / homme AND femme
/ FORME / (système OR théorie) ADJ métaphysique

6) distinction des homographes H
/ CATEGORIE / Verbes (H)

Si le fichier statistique contient le dictionnaire des fréquences par genre, par période, par genre dans chaque période et par auteur (tout au moins lorsqu'il s'agit des auteurs principaux), les éléments sont réunis pour le calcul de la probabilité attachée à telle ou telle fréquence observée dans le texte. Reste à choisir la norme de comparaison qui peut être le corpus entier ou telle subdivision que l'on voudra, genre, époque, auteur, ou intersection du genre et de l'époque. Le tableau 3 fournit les commandes nécessaires à ce genre de recherche.

TABLEAU 3 :- RECHERCHE STATISTIQUE

consulte les fichiers statistiques

commande	STATISTIQUE
opérateurs	BASE ET +
mots-clés	TEXTE AUTEUR GENRE TRANCHE CORPUS
options	NORMAL HYPER
exemples	
/ STATISTIQUE /	TEXTE BASE GENRE
/ STATISTIQUE/	AUTEUR BASE TRANCHE
/ STATISTIQUE/	TEXTE BASE (TRANCHE ET GENRE)
/ STATISTIQUE/	TEXTE BASE AUTEUR
/ STATISTIQUE/	AUTEUR BASE CORPUS (HYPER)
/ STATISTIQUE/	TEXTE BASE TRANCHE (1 + 2)

L'option par défaut est celle de la loi normale.

4 - L'utilisateur enfin doit avoir le choix des sorties et dans l'interrogation un ordre spécial EDIT lui permet de préciser quels éléments il désire et dans quel ordre. Les résultats diffèrent en effet selon qu'il s'agit d'un mot dans tout le corpus ou d'un texte dans tout son vocabulaire. Ils diffèrent selon qu'on veut un index ou une concordance. S'agissant de concordance, le choix demeure entre un contexte d'une phrase ou de n lignes ou de n mots. Quant aux calculs de statistique, ils font appel à des mots-clés spécifiques comme FREQUENCE ou SEUIL. L'ensemble de ces mots-clés est détaillé dans le tableau 4 consacré à la commande EDIT.

On peut imaginer un second type de sortie qui consisterait en une simple copie d'une partie du fichier de base. Ce serait le moyen d'avoir communication directe d'un texte que l'on se proposerait de soumettre sur place à des traitements locaux. Ce transfert de la base textuelle devrait se limiter à de petits volumes, tels qu'ils puissent tenir dans les limites des disquettes actuelles. S'il s'agit de textes

longs la duplication de bandes sur le site central me paraît être un moyen plus approprié et moins coûteux. Dans la figure 5 qui résume l'ensemble du fonctionnement du système la demande de copie des textes est la seule qui ne transite pas par le fichier inverse et qui s'adresse directement (ou presque) au fichier de base.

TABLEAU 4 - LES COMMANDES DE SORTIE

COPIE
EDIT

commande COPIE
provoque la transmission d'un texte sur le support (disquette) de l'utilisateur
/ COPIE / Hernani

commande EDIT
mots-clés FORME DATE GENRE AUTEUR TEXTE SOUSTEXTE PAGE LIGNE PHRASE LIGNES (n) MOTS (n) FREQUENCE SEUIL

exemples

- édition de l'index d'un texte
/ EDIT / FORME PAGE LIGNE

- concordance avec contexte d'une phrase
/ EDIT / FORME PAGE LIGNE PHRASE

- interrogation portant sur un mot dans tout le corpus
/ EDIT / FORME DATE GENRE AUTEUR TEXTE PAGE LIGNE PHRASE

- impression du vocabulaire significatif d'un texte au seuil de probabilité de 0,05
/ EDIT / FORME FREQUENCE SEUIL (0,05).

En conclusion répétons que notre projet n'est guère autre chose qu'un cahier des charges établi par un utilisateur littéraire qui n'a pas eu peur d'entrer dans la salle-machine. La réalisation d'un tel système doit bien entendu être confiée à un spécialiste des bases des données qui maîtrise les problèmes de distribution sur le réseau et qui connaisse les contraintes techniques posés par l'implémentation sur un matériel particulier. On peut seulement souhaiter que la base soit établie à Nancy, non seulement parce que les données s'y trouvent réunies, mais aussi parce que les possibilités du Centre de Calcul y ont été grandement améliorées (puisqu'on y dispose actuellement d'une capacité de stockage de 1 milliard 400 millions de caractères - ce qui suffit à. notre base de données - et que les interfaces ou organes de liaison (ou frontal) ont été récemment acquis qui permettent le raccordement au réseau Transpac.

On peut certes admettre des étapes progressives et n'envisager au début qu'un logiciel réduit et un corpus partiel. Mais la simplicité

volontaire du projet devrait permettre de rapprocher les échéances et on doit souhaiter que la réalisation ne piétine pas trop longuement et que l'on ne tourne pas trop longtemps autour d'un arbre dont les fruits sont mûrs.

TABLEAU 5