Frantext, la base de données textuelles du français de l'INaLF

Évelyne Martin

FRANTEXT LA BASE DE DONNÉES TEXTUELLES DU FRANÇAIS

Évelyne MARTIN

Ces quelques lignes destinées à faire connaître la Base de données textuelles FRANTEXT s'adressent à ceux qui, familiers de ce mode de documentation de création relativement récente, savent fort bien que les données textuelles sont des données comme les autres et que tout système qui peut permettre de les reconnaître, de les sélectionner, de les regrouper et de les dénombrer est utile en lui-même. Elles s'adressent aussi à ceux qui pensent, et on peut les comprendre, que le seul support acceptable du texte est le livre et que la seule façon de tirer le meilleur parti du livre est de le lire, voire de le relire et de le relire jusqu'à plus faim. Ceux-là, liront d'autant plus et d'autant mieux que FRANTEXT les aura dispensés de la première lecture, la plus rébarbative, celle qui n'a pour but que la constitution d'un fichier de travail, ou la recherche hasardeuse et laborieuse de l'aiguille dans le tas de foin.

Nous verrons successivement ce qu'est au juste la Base de données textuelles FRANTEXT, comment on l'utilise, l'exploitation qu'on peut en faire pour l'étude des signifiants comme pour celle des signifiés, de même que, marginalement, comme source de création (morceaux choisis, génération de texte, par exemple). Nous évoquerons enfin les programmes de développement dont plusieurs sont en cours de réalisation.

La création de la Base FRANTEXT correspond à l'un des principaux objectifs que B. QUEMADA a fixés à l'Institut National de la Langue Française lors de sa fondation. Conçue pour valoriser les ressources textuelles informatisées déjà existantes, elle porte la marque des grands programmes lexicographiques développés par le CNRS depuis 50 ans et en particulier celle de l'élaboration du dictionnaire *Trésor de la Langue Françaisel*. Pour la rédaction de celui-ci il a fallu rassembler, analyser et traiter avec l'aide de l'informatique d'immenses fonds documentaires bibliographiques, lexicologiques et surtout textuels, qui constituent par eux-mêmes un véritable autre Trésor, celui-là de

données, c'est-à-dire à la fois beaucoup moins et beaucoup plus qu'on n'en trouve dans le Trésor-dictionnaire. En ce qui concerne, par exemple, les données qui constituent la Base FRANTEXT, elles ne bénéficient pas pour l'instant des classements de sens lexicographiques, mais présentent une masse d'exemples souvent beaucoup plus de cent fois supérieure à ce que présentent pour la même époque les dictionnaires les plus riches[2].

FRANTEXT rend accessibles en permanence, en interactif, 160 millions de citations résultant du traitement informatique de quatre siècles de littérature et d'une collection représentative de textes appartenant aux domaines des sciences, des arts et des techniques. En tout quelque 2 600 textes[3], de 1600 à 1969, d'Honoré d'URFE à René CHAR.

La Base est actuellement gérée par le logiciel STELLA (Système de Textes en Ligne et Libre Accès) conçu par J. DENDIEN, et fonctionne sous MULTICS au Serveur CIRIL (Centre Interrégional d'Informatique de Lorraine) de Nancy.

Le Corpus de FRANTEXT

Quelque 900 écrivains (450 éditeurs environ)[4] sont représentés ici, et donc, directement ou indirectement ce qu'ils ont vu, vécu et pensé.

Les corpus d'origine ont été choisis par une équipe de spécialistes parmi les titres retenus pour leur fréquence d'apparition dans les grandes bibliographies[5]. En ce qui concerne les textes littéraires à proprement parler, qui représentaient environ 80% de l'ensemble, un juste équilibre a été observé en fonction de la répartition dans le temps (6 millions de mots environ par décennies) et de la représentation des genres dans chaque époque, approximativement le plus souvent dans cet ordre (décroissant) roman, théâtre, poésie, mémoires, correspondances, récits de voyage, pamphlets, art oratoire. On notera que s'il s'agit là, par décision initiale, de langue écrite, les journaux intimes, la correspondance, les mémoires, le théâtre en prose, les dialogues de roman sont susceptibles de livrer une langue relativement proche de la langue parlée. Les grands domaines scientifiques et techniques représentés sous le genre "traité ou essai" sont les suivants : administration publique, arts, astronomie, bâtiment, biologie, chimie, critique littéraire, droit, économie, énergie, ethnologie, histoire, industries, information, linguistique, loisirs, mathématiques, occultisme, philosophie, physique, politique, psychologie, religions, sciences de la terre, sports. On notera aussi que les indications de genres et de

domaines qui viennent d'être citées sont affectées comme descripteurs à chaque texte dans la Base, s'ajoutant aux descriptions bibliographiques proprement dites, les unes et les autres permettant la constitution de corpus de travail pertinents et personnalisés du type : *les essais publiés entre 1918 et 1939, et traitant de politique.* Le choix des éditions a été guidé dans la mesure du possible par le désir initial d'enregistrer le texte de la première édition de chaque œuvre. Il a fallu bien entendu faire de nombreuses exceptions.

Ce corpus original s'accroît régulièrement - par saisie traditionnelle, lecture optique et acquisition de bandes de photocomposition - d'œuvres de tous les siècles, à commencer par la littérature contemporaine (on introduit actuellement l'œuvre complète de R. CHAR et, pour l'ancienne langue, une collection d'œuvres de moyen français /14e et 15e siècles/)[6].

Une codification sommaire introduite lors de la saisie (en pré- ou post-édition) permet essentiellement le repérage des découpages du texte (tomes, parties, actes, scènes...) et des éléments significatifs de sa typographie, notamment les passages en italiques.

Les principes d'utilisation de FRANTEXT

La maîtrise du système est très vite acquise. Toutes les commandes sont en français. L'exploration est facilitée par de nombreux "menus".

Les corpus de travail sont ce qu'on veut qu'ils soient. On peut être précis et choisir d'explorer par exemple *les romans de Montherlant écrits entre 1934 et 1939, ou les œuvres de J. Verne parues chez Hetzel.* On peut faire porter son exigence sur un domaine et une période seulement sans qu'on ait en tête la moindre indication de nom d'auteur ou de titre, par exemple *un manuel de droit écrit en 1949* (peu importe lequel). On peut réunir plusieurs œuvres pour les comparer, par exemple le *Journal* de J. Green et celui de Fr. Mauriac pour l'année 1934.

Le "littéraire" choisit son corpus : l'œuvre, ou certaines œuvres d'un écrivain, la production littéraire d'une tranche chronologique, d'une époque, les textes appartenant à un genre ou à un domaine donnés. Le "linguiste", lui, est généralement plus soucieux de confirmer ses hypothèses par l'examen systématique et rapide du plus vaste corpus possible. Il fallait donc prévoir de travailler sur *le tout, ou* de grands ensembles du *tout.* Dans ce dernier cas, la machine fait alors elle-même un tri en fonction du thème de recherche, en l'occurrence, du mot ou des

mots qu'il s'agit de situer (c'est la fonction *index),* d'illustrer par ses contextes7 (c'est la fonction *cherche),* de dénombrer (c'est la fonction *fréquences).* On répond ainsi instantanément à des questions du type : *liste chronologique des emplois du mot* égalité *de 1600 à 1900.*

La vedette, la cible, est souvent une forme, ou graphie, délimitée par deux blancs[8] *(opérette, chantait, chanter, Opéra, opéra),* une séquence *(opéra-comique, aujourd'hui, merci pour, musée imaginaire, cime indéterminée des forêts),* un modèle syntaxique avec des variantes, ou une simple chaîne de caractères (eu-, *-phile, -ismel/-ique, -chron-)*[9]. On peut préciser la place que la cible doit avoir dans la citation, au moins *début ou fin de phrase, avant ou après une autre cible.* Le système donne en outre la possibilité, très appréciable dans beaucoup de cas, de créer des listes de formes qui seront alors traitées dans la même exploration. Dans le cas des formes fléchies, les listes sont constituées automatiquement autour de l'infinitif et à partir d'un dictionnaire de référence préétabli10.

Dans tous les cas, le système fournit instantanément les occurrences *(vie* chez Sartre, *bonheur* chez Racine, *jeunesse* dans le théâtre des années 60), mais aussi les cooccurrences *(bonheur* et vie dans un contexte de 30 mots dans le théâtre d'Anouilh, *fanatisme* et *jeunesse* séparés par moins de 20 mots chez Voltaire, *théorie* en cooccurrence avec *littérature, ou peinture, ou sculpture* dans les textes traitant d'art à la fin du 19e s.).

L'exploitation de FRANTEXT déborde déjà largement les cadres dans lesquels le système a été mis en oeuvre. Comme on pouvait s'y attendre, la Base est essentiellement utilisée comme source documentaire pour l'étude des textes. Mais elle permet aussi

l'exploration des signifiés par le biais des signifiants, des idées et des faits par les mots.

En analyse de surface, les premières questions portent sur les variantes graphiques *(thibétain/tibétain ; caesium/césium),* sur les structures morphologiques *(concurrence de* -logue/logiste *dans les textes scientifiques ; le préfixe* ana- *chez Artaud),* sur la syntagmatique *(les séquences fréquentes comportant le mot* femme *dans* Les *Mandarins* de S. de Beauvoir : femme de ménage, femme du monde, femme de tête, femme bien, femme coiffée, etc. ; ou les locutions dans lesquelles entre le substantif *complexe* dont bon nombre, il faut le noter, ne figurent dans aucun autre recensement *complexe de retraité, de situation privilégiée, du camisard, du petit profit, du verrou,* etc.).

Dans le domaine syntaxique l'éventail de l'interrogation est large : exemples de constructions multiples *(concerto de/pour piano ; continuer à/de ; merci de/pour ; travailler pour/avec/chez) ;* tours *(verbe de perception /liste fournie/* + n *mots* + qui, au 20e s. ; ce/lui/ux/Ile/Iles/ des + n *mots* + qui, chez Proust) ; thèmes syntaxiques *(l'alternative avec ou et* soit *dans le* Journal *de Gide).*

Dans le domaine plus large de la sémantique, l'aide apportée par FRANTEXT est, comme on pouvait s'y attendre, moins directe, puisque les repérages portent sur des formes seulement. Néanmoins on peut classer sous cette rubrique les listes chronologiques d'attestations rendant compte des conditions d'emploi, de l'évolution, du cheminement du sens, ou de la banalisation du mot (par exemple celle de *paranoïaque* dans la première moitié du 20e s. littéraire (chez Bernanos, Eluard, Malraux, Green, Cendrars, Queneau...). Les inventaires de type onomasiologique font l'objet d'une demande courante (les *mots* de *la* famille de "porc" chez *Giono,* les *désignations* de "l'enfant du premier âge" ou de "la grand'mère" au *19e s.),* mais il est évident que la demande doit être accompagnée d'une liste de vocables dressée par l'utilisateur lui-même. La recherche d'exemples définitoires est également grandement facilitée par FRANTEXT : on les repère par la présence dans le contexte de tours comme *j'entends par là, je veux dire, qu'entend-on par,* autrement dit, ... *ou* la présence dans un contexte proche d'un verbe comme *définir.*

Il faut surtout noter l'apport de FRANTEXT en ce qui concerne des nuances de sens dont parfois on ne trouve aucune trace dans les dictionnaires *(émeute/insurrection, manque/insuffisance, association/participation, peinture/description,* etc.).

FRANTEXT est aussi une mine d'emplois figurés : alors *qu'absinthe* n'est défini dans la plupart des dictionnaires que comme plante ou boisson, le TLF donne trois exemples de l'acception "amertume" extraits de la Base (chez Hugo, Chénier et Aragon)[11]. Il en est de même pour *allumette* dans le sens de "qui excite" ou *amande* dans celui de "l'essentiel caché sous les apparences", et bien d'autres ...

Les listes chronologiques de contextes sont toujours éclairantes. Un simple survol de la concordance de *complexe* permet de faire sur-le-champ un certain nombre de constatations premières. On constate notamment que le terme de *complexe* n'a pas semble-t-il connu une aussi grande désaffection chez les spécialistes que le prétendent les dictionnaires ; sa banalisation dans la langue littéraire et courante n'est

pas limitée aux locutions *faire/avoir un/(des) complexe(s), être bourré de complexes/sans complexe* … D'autre part, de nombreux contextes font allusion à la conscience que les sujets ont de leur(s) "complexe(s)", ce qui tendrait à prouver que le terme est souvent mal employé (mais on le trouve chez de "bons auteurs") ou que ce trouble n'est pas aussi inconscient qu'on a pu l'écrire dans la plupart des définitions, ce que confirmeraient les attestations répétées de *complexe inconscient* … Le complexe d'Œdipe fait l'objet d'une soixantaine de citations dans le corpus, parfois appelé *complexe œdipien,* parfois *Œdipe* tout court, et *Œdipe-complexe* en traduction de FREUD. Ces attestations apparaissent surtout dans les textes scientifiques, elles sont régulièrement réparties dans le temps avec cependant des éclipses de *1923* à *1935,* de *1939* à *1946* et de *1957* à *1966* (peu d'attestations littéraires, en *1935, 1949* et *1954)...*

L'exploration de quatre siècles de littérature a permis de dater de *1665* la première attestation (du moins dans le corpus) de *comédie humaine* ("L'orgueil comme lassé de ses artifices et de ses différentes métamorphoses, après avoir joué tout seul tous les personnages de la *comédie humaine,* se montre avec un visage naturel." La Rochefoucauld, *Maximes).*

Enfin, la Base est l'outil privilégié des études quantitatives (nombre des adverbes en *-ment* chez Balzac, fréquence des formes fléchies de voir chez Gracq, des mots de la famille de *chat* chez Colette, ...). Il suffit de consulter la bibliographie d'Etienne BRUNET pour en avoir la démonstration la plus riche et la plus convaincante.

On pouvait prévoir cette forme d'exploitation tant les préoccupations des linguistes et des littéraires sont proches de celles des lexicographes du *TLF,* mais très vite le public s'est élargi aux historiens dans le sens le plus large (historiens des faits, des civilisations, des sciences), sachant bien qu'à travers le style des écrivains, FRANTEXT illustre, outre les faits de langue, les faits illustrés par la langue. Et toute une série de questions qu'on pourrait qualifier d'encyclopédiques ont déjà trouvé réponse par FANTEXT : les représentations artistiques d'Orphée d'après les critiques d'art, les impressions ou les observations comparées des auteurs de mémoire pour une même journée (14 *juillet, 11 novembre …),* l'image de l'alcoolisme dans les romans du *19e s.,* les citation de Marx de son vivant, les allusions à l'Algérie dans les années *50.*

Les contextes fournis par la Base peuvent aussi permettre accessoirement de développer la bibliographie du terme, et de la notion

qu'il recouvre, par le relevé des citations d'auteurs qui y sont faites dans un environnement plus ou moins proche, par exemple pour *complexe,* ADLER, BAUDOIN, FREUD, FROMM, GURVITCH, P.-J. JOUVE, JUNG, KARDINER, MALINOWSKI ou MÜNSTERBERG.

Base d'étude, d'analyse, d'observation, d'exploration des textes, FRANTEXT est utilisée aussi comme fournisseur d'un matériau riche et différencié, propre à de nombreuses applications pédagogiques (exercices de style, de syntaxe, de sémantique, d'onomasiologie), et source de nouveaux textes, du simple recueil de morceaux choisis sur un thème donné à la génération d'autres textes à partir d'éléments existants.

Des expériences sont menées dans ce sens dans le cadre de l'enseignement secondaire notamment, par exemple au sein du groupe C L E O [12] : recherche des thèmes dominants de *L'Argent* de Zola, *d'Antigone* d'Anouilh, de L'île mystérieuse de J. Verne..., examen des champs dérivés, des conditions d'emploi, etc. Noter aussi qu'une bonne liste d'exemples appelée par FRANTEXT se prête aisément à une étude conceptuelle, proche des travaux sur les centres d'intérêt, mais qui suggère souvent des associations plus riches et plus nuancées. On peut ainsi demander à l'élève de relever les mots-pleins utilisés en cooccurrence avec ami dans un contexte un peu long, d'en extraire les termes quasi-synonymes et de sens proche *(compagnon,* camarade, complice, confrère, allié, *confident,* conseiller, admirateur, disciple), puis d'attester ces derniers chez le même auteur ou dans un autre corpus, de les employer dans des exemples créés, etc. Ce développement lexical, consacré par l'usage, mais ici présenté dans un contexte situationnel, incite à réfléchir sur la nature des différents sèmes ainsi identifiés, illustrés, et des liens qui les unissent, en même temps qu'il favorise la recherche et la découverte du mot juste. Le locuteur ou scripteur, débutant ou non, peut faire un choix éclairé dans le catalogue de formes et d'emplois qui sont proposés, catalogue de messages aussi qui impressionnent l'élève ponctuellement, un peu au hasard, qui, par leur diversité même invitent l'interlocuteur à les relire et à en rédiger d'autres implicitement ou explicitement.

Songeons aussi aux auteurs d'anthologies centrées autour de l'"écureuil", de la "tauromachie", de la "guerre", du "divorce" ou de l'"Académie" et qui, grâce à la production automatique des

répartitions de fréquences disposent au départ d'une documentation brute, certes à utiliser avec précaution et discernement, mais abondante souvent, et rapidement constituée toujours. Notons enfin que, bien que ce

ne soit pas sa destination première, FRANTEXT pourrait offrir de nouvelles perspectives aux écrivains de l'OULIP0[13] de l'ALAM0[13], et plus largement à tous ceux qui travaillent à la génération de texte.

On voit que, dans l'état actuel de la Base, par le choix des corpus et les combinaisons de commandes, l'utilisateur peut multiplier selon ses besoins le nombre des applications : donner une illustration nouvelle des sens connus, confirmer, attester des sens nouveaux, rares ou seulement présumés, mais aussi relever des citations, des témoignages, constituer un fichier de séquences fonctionnant sur un modèle donné, dater une forme ou un syntagme, sélectionner des exemples définitoires, lister des hapax d'auteurs, dégager des types de discours, des thèmes syntaxiques, lister et dénombrer des graphies, etc.

On pourra toujours déplorer l'absence de tris sémantiques à proprement parler, notamment dans les cas souvent cités d'homographie et de polysémie. Mais il faut se rappeler qu'on peut dès maintenant lever certaines ambiguïtés notamment par le choix des corpus, le jeu des cooccurrences et la place des mots.

Il faut surtout savoir que les perspectives de développement sont nombreuses et que l'INaLF s'emploie activement à accroître le corpus, à optimiser les méthodes de stockage et les procédures d'exploration et donc à diversifier et à rendre plus précis et plus spécifiques les produits, le principe restant toujours celui d'une introduction représentative et fidèle des données et d'une exploration de plus en plus fine, sélective et conviviale.

L'accroissement du corpus se fait chaque jour, par saisie traditionnelle et peu à peu, pour certaines collections et moyennant certaines précautions, par saisie optique. Outre la complémentation chronologique qui s'impose tout naturellement, des apports sont prévus siècle par siècle pour combler des lacunes et renouveler le cas échéant les éditions. On introduit en outre de la littérature régionale et régionaliste. Des idées sont lancées qui seront prises en considération pour un programme de saisie ultérieur : littérature populaire, francophone, biographies, prix littéraires, dialogues de film. On tentera en outre de constituer des "œuvres complètes", notamment pour les écrivains déjà bien représentés (Claudel, Montherlant, Malraux ...). Il sera également tenu compte de suggestions d'utilisateurs en fonction de thèmes de recherche et de programmes officiels, en linguistique et littérature, mais aussi en histoire des civilisations, philosophie, et plus généralement en histoire des sciences.

Le corpus s'accroît également peu à peu de bandes de photocomposition confiées par les éditeurs eux-mêmes[4] et de textes français saisis ailleurs, à Stockholm, à Chicago, à Louvain, à Liège, à Gênes, Montréal, Toronto, et bien entendu dans les équipes de recherche françaises, l'idée directrice de cet accroissement restant toujours le souci de la meilleure adéquation possible à la recherche en cours en sciences humaines, et aux programmes et orientations de l'enseignement. En ce qui concerne les conditions de stockage, l'utilisation d'un support disque-compact est envisagé ; il aura pour conséquence de permettre l'utilisation de FRANTEXT (version 2 ou 3) non plus seulement sur réseau, mais sur micro-ordinateur.

L'optimisation du système est possible à court terme, en premier lieu, par l'utilisation de bases de connaissances déjà réalisées à l'INaLF à d'autres fins. Elle devra permettre notamment :

- d'extraire des contextes terminologiques (les termes les plus courants du vocabulaire de la psychiatrie, avec leur environnement dans les romans et mémoires de la fin du 19e s.)

- de relever et de dénombrer les régionalismes d'une oeuvre (les mots du parler lorrain chez Barrès)

- de regrouper toutes les graphies d'un même vocable (phantasme avec fantasme, yoghourt avec yaourt)

- d'explorer le corpus à partir de champs synonymiques, antonymiques ou thématiques (les mots désignant la mort chez Zola, ou la folie chez Stendhal)15.

Enfin l'informatisation du TLF, envisagée à moyen terme constituera sans nul doute le dictionnaire de référence par excellence par le fait que les formes, morphèmes, lexies, syntagmes y seront stockés avec leurs propriétés et relations.

Par ailleurs, des logiciels d'analyse automatique du discours, en cours d'élaboration à l'INALF, pourront sans doute être au moins partiellement utilisés, permettant notamment les levées d'ambiguïté par exploration automatique des contextes, et une plus grande finesse de l'interrogation par le repérage des classes morphosyntaxiques et des catégories fonctionnelles.

La base de données FRANTEXT est appelée, on le voit à des développements considérables qui, accroissant, ses performances, accroîtront son champ d'action et d'application, et la gamme de ses

utilisateurs. Jamais en concurrence avec le livre qu'elle sert en même temps que le lecteur, elle constitue un outil documentaire simple, sûr et rapide d'exploration du texte français.

FRANTEXT est interrogeable dans la plupart des bibliothèques universitaires et la Bibliothèque publique de Beaubourg.

Évelyne MARTIN
CNRS-INaLF mai 1988

NOTES

1 Trésor de la langue française, Dictionnaire de la langue du 19e et du 20e s., (abrév. *TLF),* éd. par le CNRS (Diffusion Gallimard-Sodis), 12 vol. parus (t. 1 à 7 sous la dir. de P. IMBS ; t. 8 à 12 sous la dir. de B. QUEMADA, Directeur de l'INaLF).

2 Pour *droit* par exemple : 75 exemples cités dans le *TLF* pour 40 000 existants dans la Base.

3 Soit environ 600 000 pages imprimées, 300 000 formes différentes, plus de cent millions de mots, plus d'un milliard de caractères. À ce corpus s'est ajouté un ensemble de textes saisis au Laboratoire d'analyse lexicologique qui fut dirigé à Besançon par B. QUEMADA de 1956 à 1969.

4 La Société des auteurs multimédia et la Société civile de l'édition littéraire française ont étudié avec l'INaLF une formule d'accord qui permet l'exploration, à des fins de recherche, des textes protégés. Il faut voir là une preuve, s'il en est besoin, de la prise en considération par les écrivains et les éditeurs de l'outil informatique dans le traitement des textes, et plus encore de la conviction, qui gagne du terrain, que cet outil, loin de détourner de la lecture proprement dite, y engage et la facilite d'une certaine façon.

5 Le choix s'est fait essentiellement en fonction "de la sûreté de la langue, de la richesse du vocabulaire, et à cause d'une influence possible sur l'usage" (P. IMBS, *Préface* du *TLF,* t. 1, p. XXIII).

6 Ce sous-ensemble sera notamment utilisé par les rédacteurs du *Dictionnaire de moyen français* qui fait partie du programme de l'INaLF.

7 La longueur du contexte donné est modulable jusqu'à concurrence de 300 mots par citation.

8 Les tirets et apostrophes des mots composés sont considérés comme des blancs.

9 Le nombre de lettres exigé dans la partie variable peut être précisé (pré... *préavis, préface, prélegs,* etc.)

10 Ce dictionnaire a été constitué à partir d'une table des radicaux verbaux, progressivement enrichie puis combinée grâce à un programme de conjugaison distinguant plus de 40 classes avec les systèmes désinentiels appropriés.

11 Noter que la Base fournit aussi, pour la même famille morphologique, un exemple rare de *absintheuse :* /"Et dans ce garni, des déclassés de tous les sexes, étranges : une vieille femme de la société, une *absintheuse,* se mettant sous la peau dans un jour vingt-deux absinthes"/ (GONCOURT E. et J., *Journal,* t. 4, 1986).

12 CLEO : Centre Lorrain d'Enseignement assisté par Ordinateur.

13 OULIPO : OUvroir de la Llttérature POtentielle. ALAMO : Atelier de Littérature Assistée par Mathématique et Ordinateurs.

14 À commencer par les Presses et les Éditions du CNRS, et les Éditions Gallimard (diffuseur du *TLF).*

15 La plupart des demandes de ce type supposeraient en réalité, pour être pleinement satisfaites, l'existence de listes thématiques ou notionnelles diversifiées qui prennent en compte, outre le niveau de langue, les conditions ou les domaines d'emploi. Un tel répertoire, dont un projet a déjà été exposé, serait consultable à toutes les étapes de la recherche, constituerait un dictionnaire de référence utilisable non seulement pour l'exploration thématique pure et simple de FRANTEXT, mais pour l'étude comparative de la répartition des mots d'un thème et leur collocations dans les textes d'un corpus donné, voire pour l'indexation thématique assistée par ordinateur des textes, notamment en vue de la constitution quasi automatique de corpus et sous-corpus de travail. Un travail d'indexation artisanale des textes (affectation à chaque titre d'œuvre de descripteurs de thèmes, milieux socioprofessionnels, périodes, etc.) entrepris à l'INaLF gagnerait à être poursuivi à la faveur de ce nouveau dictionnaire.