La base de données textuelles du français sur Internet (WWW)

Éveline MARTIN

Abstract. Development on INTERNET (WWW) of the FRANTEXT database, the world's largest repository of French quotations: around 3,000 works from the 16th to the 20th century, or over 160 million contexts available in university libraries throughout France and in other countries.

Keywords: Textual data bases, literature and computing, linguistics and computing, computer-assisted processing of textual corpora.

Mots-clés : Bases de données textuelles, littérature et informatique, linguistique et informatique, traitement informatisé de corpus textuels.

Il y a quelque dix ans, l'Institut national de la langue française rendait disponible par le réseau Transpac, et Minitel en France, la base de données textuelles FRANTEXT résultant du traitement informatique de plus de cent soixante millions de mots-occurrences représentant une collection de plus de deux mille cinq cents textes complets, à dominante littéraire, du XVI^{e} au XX^{e} siècle. Ce corpus s'est accru considérablement et s'accroît chaque jour d'œuvres (littéraires ou non) de tous les siècles, à commencer par la production contemporaine; pour l'ancienne langue, une série de textes de moyen français sera accessible après son exploitation pour le *Dictionnaire du moyen français*.

L'utilisateur constitue lui-même son propre corpus de travail : les auteurs, les titres ou mots du titre, les genres, les dates d'écriture des œuvres sont autant de critères combinables qui peuvent permettre de le définir.

✉ Centre National de la Recherche Scientifique; Institut National de la langue française; 27, rue Damesme; F–75013 Paris (France).
Fax : +33 1 45 80 79 26 E-mail : martin@ciril.fr

Le corpus à explorer peut donc être l'œuvre, ou certaines œuvres, d'un ou plusieurs écrivains, la production littéraire correspondant à une tranche chronologique, les textes appartenant à un genre ou à un domaine donnés. Le système permet aussi d'explorer l'ensemble des textes enregistrés. Dans tous les cas, on obtient en interactif des repérages de mots, séquences, ou chaînes de caractères, des contextes (questions du type *liste chronologique d'emplois du mot* tolérance *de 1600 à 1900*), des statistiques (calcul et distribution des fréquences). L'utilisateur circule librement dans la liste des résultats, qu'il peut consulter dans l'ordre de son choix, stocker et traiter sur un support personnel.

Le système donne d'autre part la possibilité de créer des listes de formes, qui seront traitées dans la même exploration, notamment autour de l'infinitif des verbes et du masculin singulier des substantifs et adjectifs. Le système fournit donc instantanément des occurrences, par exemple celles de *le plus ... possible* dans *Candide*, de *lac* en poésie, de *ville* chez Claudel et Mauriac et des cooccurrences (*bonheur* et *vie* dans un contexte de 30 mots chez Anouilh. Une commande permet en outre d'obtenir, dans un ordre alphabétique ou hiérarchique, la liste des cooccurrents du mot-vedette, par exemple ceux de *femme, jeune fille, demoiselle*, ... chez Montherlant. Si la requête la plus fréquente est du type *désert* dans *Courrier sud* d'Antoine de Saint-Exupéry, il ne faut pas oublier que FRANTEXT est capable de répondre sans sourciller à une question comme celle-ci que lui proposa un jour en guise de défi et d'hommage tout à la fois un utilisateur passionné :

> La forme *fou*, ses variantes et les formes comportant la chaîne *-fou-* à l'exclusion de la séquence *faire le fou* (et formes fléchies) et du mot *garde-fou*, dans des contextes postérieurs à 1945, où ne figure pas le tour *c'est fou* et extraits de traités de médecine et droit, et de romans. Fréquences de ces formes, et liste de leurs cooccurrents dans l'ordre de leurs fréquences décroissantes.

Par le choix des corpus et la combinaison des commandes, l'utilisateur peut multiplier selon ses besoins le nombre des applications et donner ainsi une illustration nouvelle de sens connus, confirmer, attester des sens nouveaux, rares ou seulement présumés, mais aussi pour identifier des citations, relever des plagiats, des clichés, des redondances, des tics d'auteurs, recueillir 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'écrivains, dégager des types de discours, etc.

D'autres services s'ajouteront prochainement facilités par l'utilisation de nouvelles bases de connaissances, notamment un dictionnaire de référence thématique qui reliera automatiquement les vocables à une ou plusieurs étiquettes de champs, et permettra ainsi de répondre instantanément à des questions du type *les mots du thème du « temps » chez Proust*; *les désignations de la « mort » en poésie*; *le champ sémantique de l'« ivresse » chez Baudelaire* ou *de l'« intuition » chez Merleau-Ponty*, la possibilité étant donnée en outre à tout moment à l'utilisateur de modifier, réduire ou compléter les listes proposées. Cette nouvelle possibilité de sélection mettra en œuvre la Base de lexicologie thématique en cours de réalisation. La BLT de l'INaLF est un répertoire évolutif du lexique général par champs. Comme tel, il présente un ensemble de thèmes ou concepts-clés choisis parmi les plus fréquemment étudiés et accompagnés chacun de leur représentation non ambiguë dans le discours, c'est-à-dire de leurs corrélats spécifiques. En même temps que le paysage lexicologique des grands thèmes, tel qu'il apparaît dans les répertoires traditionnels et dans un ensemble de sources nouvelles (vastes corpus textuels), elle constitue essentiellement un outil d'aide au repérage thématique. Couplée ou non à la mise en œuvre de critères catégoriels, morpho-syntaxiques, ou à des outils de pondération, elle doit contribuer avec le plus petit taux possible d'erreurs

— à la sélection, dans un corpus textuel donné, de passages illustrant un concept donné, par exemple,

> [...] Il arma son Manhurin 7,65 et visa avec soin. Une abeille chahutée par le vent se posa en catastrophe sur la mire, juste comme il *faisait feu*. L'arme tonna. Léon fit un saut prodigieux comme si la terre s'était creusée sous lui. Il disparut dans les yuccas. La *détonation* se répercuta de muret en muret.
>
> (J. VAUTRIN, *Billy-ze-kick*, 1974, p. 32.)

pour le concept de « violence »;

— au relevé des thèmes traités dans un texte ou un ensemble de textes donnés (pré-résumé, pré-indexation automatiques). Le texte suivant par exemple

> Nul n'a jamais su [...] du moins dans ces premiers mois de la *captivité*, s'il se retrouverait dans une usine, dans une entreprise agricole ou sur un chantier de travaux publics. Les goûts personnels, les aptitudes individuelles, la force physique de chacun, le droit de l'être humain à connaître la façon dont on va disposer de lui, autant de considérations dont la *Kommandantur* de Trèves ne daignait pas soupçonner l'existence. Nous n'étions rien de plus que des Stücke, terme *militaire* par quoi le règlement

> de la *Wehrmacht* définit ses propres *soldats*, littéralement « des morceaux », entendez des robots interchangeables, mécaniquement répartis selon les hasards de l'*immatriculation*.
>
> (F. AMBRIRE, *Les grandes vacances*, 1946, p. 49.)

sera automatiquement indexé à « emprisonnement », « armée » et « Allemand » ;

— aux levées d'ambiguïté, par l'exploration de l'environnement des mots polysémiques et des homographes de même catégorie grammaticale.

Par exemple, *balle* peut être automatiquement défini sans grand risque par « projectile dont on charge une arme » dans un contexte court (deux phrases avant, deux phrases après, par exemple) où il y a encore des mots comme *revolver*, *pistolet* (« armes »), *rosser* (« violence »).

> Un dimanche soir, André revint de Paris armé d'un *revolver*. C'était un vrai *pistolet* d'ordonnance, avec ses six ***balles*** tranquillement logées dans le barillet. Il avait fauché l'arme à l'amant de sa mère, et paraissait pressé de s'en servir. Toute la journée du lundi j'ai tremblé qu'il ne le fasse et, m'étant fait *rosser* en cinq secs dans les W.-C. par un grand salaud de fils de boulanger qui en voulait autant aux rouquins qu'aux juifs
>
> (J. LANZMANN, *Le têtard*, 1976, p. 49–50.)

de même *descendre* a probablement le sens d'« abattre » ou *tirer* le sens d'« envoyer un projectile avec une arme » dans un contexte où il y a encore un mot comme *arme* (« armes »).

> Bastienne se demande pourquoi elle accepte l'invitation de China au lieu d'aller gagner son fric. C'est que lorsqu'elle aura l'argent, elle aura l'***arme***. L'***arme*** dans la poche, faudra aller *descendre* Montauquier. Plus le moment approche, plus l'angoisse grandit... Si elle n'avait pas le courage de *tirer*?... Abattre une vie, rendre mort ce qui est vivant... même un monstre...
> — Fous-toi ça sur la tête! dit China en lui tendant l'autre casque.
>
> (V. THERAME, *Bastienne*, 1986, p. 159.)

Pour que ces objectifs soient atteints avec le plus petit taux possible de résultats parasites, il a fallu définir très précisément les concepts-clés étudiés, réduire la polysémie des corrélats et éliminer les corrélats homographes : ainsi *huissier* ne figure pas sous « justice », *chagrin* ne figure ni sous « livre » ni sous « tristesse » mais *bémol* figure sous « musique », *rafale* sous « violence », *dandy* sous « mode » *et* « jeunesse ». On dispose ainsi pour chaque concept-clé d'une liste de marqueurs, monosémiques ou dont la polysémie reste à l'intérieur du même concept-clé, susceptibles de repérer les textes traitant d'un thème donné et les thèmes d'un texte ou contexte donné. Les

concepts-clés ont été choisis parmi ceux qui ont fait le plus fréquemment l'objet d'études thématiques en littérature, philosophie ou histoire des civilisations, ou figurent régulièrement dans les programmes d'enseignement. Ils sont traités sous leur « étiquette d'origine ». Mais sous la rubrique LISTE ORGANISÉE DES CONCEPTS, on trouvera des regroupements et des renvois qui permettent d'autres modes de consultation.

Ex. : VIE DE SOCIÉTÉ
accueil
activité sociale
ascension sociale → évolution sociale
assistance → aide
bons rapports humains
bourgeois → bourgeoisie
bourgeoisie
concept-clé associé : capitalisme, classe sociale, ...

L'observation et le traitement statistique des appartenances multiples (d'un même corrélat à plusieurs concepts-clés) signalent tout naturellement l'existence de macro-thèmes (ou concepts-clés de plus grande extension comme COMPORTEMENT HOSTILE regroupant « violence » et « agressivité ») ou de simples affinités comme celle qui relie « mode » et « jeunesse ». La pratique de l'interrogation conduit de même à établir une liste ouverte de renvois fonctionnels. Pour chaque concept-clé traité, la BLT propose en accès interactifs multiples

— *une définition*,
— *une liste de corrélats spécifiques du concept et appartenant à l'environnement* :
  • morphologique
    ex. : dans le thème de la « guerre » : *guerroyer* (mais non *aguerrir*),
  • sémantique
    Dé = définition
    E = étude de thématique, répertoire de synonymie, d'analogie, etc.
    Te = texte du corpus de travail
— *l'indication*
  • des concepts-clés associés,
  • des concepts-clés considérés comme « opposés »,
— *la référence*
  • à un choix d'œuvres, en français, tous genres et époques confondus, illustrant dans son ensemble ou sa majeure partie le concept (bibliographie chronologique),

- à un choix d'études portant sur le concept en tant que tel (bibliographie chronologique),

— une *anthologie* extraite du corpus de travail et dans laquelle une attention toute particulière a été portée à l'étude des *nuances contextuelles*.

Le corpus de travail rassemble environ 700 textes complets de romans de 1800 à 1995. Les données de la base en ligne sont régulièrement revues et augmentées. La consultation de la Base doit notamment permettre d'obtenir :

— la liste des corrélats spécifiques d'un ou plusieurs concepts-clés donnés;
— le ou les concepts-clés auxquels peut être rattachée une lexie donnée;
— l'environnement complet d'un concept-clé donné;
— la liste alphabétique générale des concepts-clés traités avec ou sans mention des concepts associés ou opposés;
— l'index alphabétique, général ou par type d'environnement et/ou de source, des lexies traitées :
— la fréquence absolue et relative, dans le corpus textuel de base, d'une ou plusieurs lexies données;
— la bibliographie du corpus textuel de base;
— la liste générale alphabétique, chronologique des œuvres et études citées.

Elle doit permettre également

— une exploration du texte des définitions de concepts;
— l'ajout, à tous les stades et dans chaque champ, de commentaires ou ajouts personnels.

De même, une version plus élaborée encore de FRANTEXT proposera une catégorisation grammaticale des formes, probablement patielle et aléatoire dans un premier temps mais permettant de répondre à des questions du type *contextes d'adverbes en -ment précédant un adjectif dans le dernier quart du XX^e siècle* (ex.: *politiquement correct*).

La longueur des contextes obtenus dans la base FRANTEXT est de trois pages de contexte pour les textes libres de droits; en ce qui concerne les textes *protégés*, la citation est limitée à trois cents caractères.

La Base est accessible moyennant un abonnement annuel de 2 000 FF, hors taxes, à tout organisme de recherche ou universitaires qui en fait la demande au Service FRANTEXT de l'INaLF[1]. Plusieurs milliers d'utilisateurs

[1] Institut National de la Langue Française, 27, rue Damesme, 75013 Paris. Tél. +33 1 45 80 36 00. Fax : +33 1 45 80 79 26.

interrogent la base regroupés autour de quelque soixante stations de travail en France et à l'étranger.

La nouvelle version sur Internet (WWW) désormais accessible, parallèlement à l'ancienne, devrait faire augmenter très vite le nombre des interrogations. Elle repose sur une logique d'interrogation nettement améliorée. Elle associe les services de base de FRANTEXT et ceux des logiciels de consultation du *World Wide Web* ou WWW ou W3 (mode client-serveur). L'URL (*Uniform Ressource Locator*) de FRANTEXT est :

http://www.ciril.fr/~mastina/FRANTEXT

Par le jeu des couleurs, des soulignements, des boutons, le chercheur est invité à cliquer pour obtenir l'information ou le service qu'il attend ou qui lui est conseillé. Chaque entrée du menu principal comporte, outre la désignation du service qui sert à l'activation de ce service, un bouton d'aide sur lequel on peut cliquer pour obtenir une définition du service, une documentation et des recommandations relatives à ce service. Le système est auto-documenté de façon précise et conviviale, notamment ce qui concerne la recherche de mots ou séquences dans un corpus de travail donné. Cette requête, de loin la plus fréquente, peut être faite à partir d'une forme simple ou composée d'une suite de formes entièrement ou partiellement connue, le doute et l'exclusion étant admis, d'une liste rassemblant automatiquement les graphies et les flexions possibles d'une forme ou les formes comportant une chaîne de caractères commune, ou encore à partir d'une série de formes constituée par le chercheur. Toutes ces hypothèses sont envisagées dans la documentation, accompagnées chacune d'un exemple, et d'un modèle de requête. Les *retours* au début de l'opération et/ou au menu général donnent le droit à l'erreur ou au « remords » et permettent de dénouer toutes les situations.

Le système propose à l'utilisateur des *formulaires* simples correspondant chacun à un type de services, qu'il suffit d'activer par un bouton après avoir rempli les *fenêtres* qui apparaissent sur l'écran. Cette dernière démarche, simple au demeurant, est encore facilitée par les recommadations et les menus déroulants qui apparaissent à l'écran. Elles peuvent l'être davantage encore par l'utilisation des boutons d'aide.

Le système permet enfin d'utiliser les fonctions de n'importe quel logiciel de traitement de texte notamment — en ce qui concerne les résultats — les opérations de sauvegarde ou de *copier/coller*. On peut ainsi réaliser une impression totale ou partielle des documents obtenus dans la typographie et

avec la présentation souhaitées, ou un stockage sur support informatique en vue d'un nouveau traitement.

Le dépôt, en 1981, d'un double du corpus de FRANTEXT au département de français de l'Université de Chicago, plus précisément dans l'équipe baptisée à cette occasion «*American and French Research on the Treasury of the French Language*» (ARTFL), a permis à cette dernière de développer sur place l'étude des textes français : le programme d'interrogation *Philologic* a été conçu à cette intention pour les chercheurs locaux.

La nouvelle version de FRANTEXT est, comme on le voit, fonctionnelle, facile à utiliser; elle s'adapte à n'importe quel mode de restitution et de traitement.

Plus encore, mieux encore que dans ses versions précédentes, FRANTEXT assiste, informe, documente et guide le chercheur exigeant ou le simple curieux, qu'ils soient ou non enclins à utiliser l'informatique, que sa préoccupation soit la linguistique, la littérature, la textologie ou l'histoire des faits.