

## ■ La préface du TLFi par Jean-Marie Pierrel

Le TLFi, Trésor de la Langue Française informatisé, est la version informatique du Trésor de la Langue Française (TLF)<sup>[1]</sup>, Dictionnaire de la langue du 19e et 20e siècle, publié par le CNRS en 16 volumes entre 1971 et 1994. Ce dictionnaire est le fruit d'une grande aventure lexicographique qui, sous les directions successives du Recteur Paul Imbs et du Professeur Bernard Quémada, regroupa à Nancy plus de cent collaborateurs durant 30 années au sein du Centre de Recherche pour un Trésor de la Langue Française (CRTLF) puis de l'Institut National de la Langue Française (INaLF) dont le laboratoire ATILF Analyse et Traitement Informatique de la Langue Française <sup>[2]</sup> se veut être aujourd'hui le digne successeur.

C'est en effet par décision du 20 décembre 1960 qu'est créé à Nancy un centre de recherche avec mission de mettre en œuvre la documentation, la rédaction et la publication d'un Trésor de la Langue Française. Ce choix avait été préparé et souhaité trois ans plus tôt dans les conclusions d'un colloque international de lexicologie et lexicographie françaises et romanes, organisé au Centre de Philologie Romane de Strasbourg par son directeur le Professeur Paul Imbs.

Le Littré (1863 - 1873) étant tombé dans le domaine public, une intéressante controverse démontra que s'il était sage de réimprimer le fameux dictionnaire tel quel en raison de son caractère de " monument " de la science de son temps, le moment était venu de mettre en chantier quelque chose d'entièrement nouveau, qui tiendrait compte des acquis de la lexicologie et de la lexicographie du 20e siècle, des possibilités nouvelles en matière de documentation, et bien sûr des changements survenus dans la langue française depuis le milieu du 19e siècle. Les conclusions du colloque de Strasbourg étaient très nettes sur ce point : " Instrument de travail, le *Trésor* poursuivrait donc un double but : être le témoin objectif et impartial du vocabulaire français, mieux connu parce que mieux inventorié ; être ce qu'avait été le *Littré* pour son temps : un exemple-type de lexicographie scientifique moderne ".

Si l'on tente aujourd'hui de resituer le *Trésor* par rapport aux exigences de la lexicographie, rien de tel que de reprendre les objectifs initiaux définis par Paul Imbs. Le *Trésor* sera donc :

- *Un dictionnaire du monde francophone*. La France avait en effet sur ce point à rattraper un retard, à un moment où l'Angleterre avait terminé, vingt-cinq ans plus tôt, son *New English Dictionary* (*Dictionnaire d'Oxford*) et où plusieurs pays, latins, germaniques ou slaves étaient à l'œuvre depuis plusieurs années pour publier un dictionnaire national.

- *Un dictionnaire historique*. Le *Trésor* ne se bornera pas à donner pour les mots l'usage du moment mais il inclura, pour chaque mot, une rubrique " étymologie et histoire ", riche des connaissances actuelles en ce domaine.

- *Un dictionnaire linguistique ou dictionnaire de langue*. Par opposition à une visée encyclopédique, le *Trésor* s'attachera à définir chaque mot par ses caractéristiques linguistiques : sa forme, son sens, ses emplois stylistiques et syntaxiques.

- *Un dictionnaire, œuvre d'une génération*. La création du centre de recherche pour un trésor de la langue française coïncida avec les premières utilisations en sciences humaines de moyens mécanographiques et informatiques de documentation. Ainsi, dès 1964 et grâce à un puissant matériel informatique<sup>[3]</sup>, des dépouillements systématiques et exhaustifs de plus de 1 000 œuvres littéraires<sup>[4]</sup> permirent aux rédacteurs de s'appuyer sur une collection très riche d'exemples d'usage des mots (430 000

exemples) qui donna naissance à ce qui est sans aucun doute la plus grande base de données textuelles sur une langue donnée, FRANTEXT, qui, mise à jour régulièrement, regroupe aujourd'hui, à l'ATILF, près de 4 000 œuvres littéraires françaises<sup>[5]</sup>.

Le Trésor de la Langue Française est le premier dictionnaire de langue se fondant sur une méthodologie systématique d'analyse des usages effectifs des mots de notre langue à travers l'exploitation d'une vaste base de données textuelles dont la saisie a débuté dès les années 60 et dont le but premier était de fournir des données organisées aux rédacteurs du dictionnaire *TLF*. Ainsi, un rédacteur ayant à écrire un article se trouvait doté de concordances systématiques de ce mot, triées suivant différents critères : ordre chronologique des sources dans un document nommé " bloc Gamma "<sup>[6]</sup>, ordre alphabétique des contextes gauche et droit dans un document consacré aux cooccurrences (les " groupes binaires "<sup>[7]</sup>), ou encore ordre définis selon les constructions syntaxiques propres à chaque partie du discours dans des documents mis au point, pour tenter de dominer la masse des attestations des mots les plus fréquents, et ce, alors que la rédaction du dictionnaire était déjà bien avancée. Ces concordances étaient utilisées pour un premier tri de la documentation et permettaient d'obtenir, dans un second temps, des contextes élargis parmi lesquels seraient sélectionnés les exemples finalement retenus dans le dictionnaire. Le traitement informatique était assuré par de lourds logiciels non interactifs (de type traitement par lot) procédant par traitement séquentiel du corpus. Au début des années 80, le laboratoire a réalisé une plate-forme de base de données textuelles qui a permis un gain de productivité spectaculaire grâce à la possibilité d'accès direct aux mots du corpus, et, surtout, d'envisager la réalisation d'une première interface utilisateur (1985), avec une exploitation télématique par les moyens de l'époque (terminaux Transpac ou Minitel) : ainsi est née la base textuelle Frantext.

Pour aborder au mieux le Trésor de la Langue Française, une lecture préalable de sa préface s'impose : écrite par le recteur Paul Imbs en août 1971 pour l'édition du premier tome, elle permet de bien percevoir les objectifs, originalités et limites de cet énorme travail lexicographique. C'est pourquoi nous avons fait le choix de la rééditer ci-après, tout comme la postface due à Bernard Quémada lors de la publication du tome 16 en 1994.

Ce dictionnaire est l'œuvre collective de nombreux collaborateurs qui travaillèrent entre 1960 et 1994 à son élaboration. Il n'était que justice de leur rendre hommage à l'occasion de cette publication sous forme de cédérom : nous avons choisi de publier la liste des contributeurs et rédacteurs de ce dictionnaire, de même que la nomenclature des entrées du TFL qui reflète un certain état de la langue à l'époque de sa rédaction.

L'évolution des techniques informatiques permet aujourd'hui, à travers l'informatisation d'un dictionnaire, de découvrir des usages nouveaux et des parcours véritablement novateurs qui s'affranchissent des aspects essentiellement séquentiels de la lecture et de la recherche dans les textes imprimés. Ces enjeux sont très vite apparus et, avant même la fin de sa publication papier, l'informatisation du TLF est envisagée<sup>[8]</sup>. A cette époque, hélas, les techniques informatiques disponibles et les coûts afférents, tant en moyens humains qu'en moyens financiers, ne permettent pas de faire avancer un tel projet. Comme le note Robert Martin<sup>[9]</sup>, dans l'introduction de l'ouvrage *Lexicographie et Informatique*<sup>[10]</sup> " des chiffrages exorbitants et conséquemment une réponse dilatoire, extrêmement réservée, en mai 1991, de notre partenaire éditorial, les Editions Gallimard, l'impossibilité quasi-absolue de susciter, au départ, l'investissement et la collaboration d'industriels, tout cela faisait apparaître le TLF informatisé comme une chimère ".

Pourtant, en mai 1995 à Nancy, lors d'un colloque international, dont les actes sont publiés dans l'ouvrage cité ci-dessus, Robert Martin peut affirmer : " Nous avons la certitude que le TLF sera informatisé ; il est même en bonne voie de l'être ". Que s'est donc-t-il passé ? Trois éléments sont

intervenues dans cette évolution :

- Une collaboration scientifique avec la Bibliothèque Nationale de France qui permit la saisie des huit premiers volumes pour lesquels il n'existait pas d'archives électroniques ;

- Un soutien sans défaut de la direction des Sciences Humaines et Sociales du CNRS et, plus globalement, de la communauté scientifique qui attendait beaucoup d'une telle informatisation, ressource de base pour des recherches futures ;

- Mais surtout, et ce fut là le point déterminant, l'énorme travail réalisé par le service informatique du laboratoire et tout particulièrement par Jacques Dendien, Ingénieur de Recherche CNRS qui sut démontrer, à travers un prototype construit à partir du volume XIV, que la faisabilité de la rétroconversion du *TLF* pouvait être réalisée au sein même du laboratoire. Nul mieux que Jacques Dendien ne pouvait retracer cette aventure de l'informatisation du *TLF*. Et c'est grâce à lui, et avec le soutien de mes deux prédécesseurs à la direction de notre laboratoire, nos collègues Robert Martin et Bernard Cerquiglini, que vous allez pouvoir découvrir les richesses de ce dictionnaire de référence sur notre langue.

Et lorsqu'en janvier 2001, à la demande du CNRS, j'ai accepté de m'investir dans la direction de ce laboratoire rebaptisé *ATILF* (Analyse et Traitement Informatique de la Langue Française) après fusion entre la composante nancéienne de l'INaLF et le laboratoire universitaire LanDisCo (Langue Discours et Cognition), j'ai trouvé une équipe motivée, à la tête d'un prototype déjà très élaboré de cette version informatique du *TLF*.

Cet outil demeurait, pourtant, encore inaccessible, demandait à être peaufiné, sa pertinence devant encore être prouvée. Une des premières tâches à laquelle nous nous sommes donc attelés fut d'ouvrir très vite cette première version sur le Web en vue d'un test grandeur nature. Le 5 mars 2002, avec l'accord et le soutien de la direction générale du CNRS, une présentation publique du *TFLi* était organisée. Alors que la moyenne des pages du *TLFi* consultée en février 2002 était inférieure à 50 par jour, nous sommes passés à plusieurs milliers fin mars 2002 et à plus de 160 000 pages par jour ouvrable aujourd'hui. Ces deux années d'expérimentation sous le Web nous ont permis de procéder à de nombreuses mises à jour et améliorations du *TLFi* tant en ce qui concerne l'interface qu'en ce qui concerne les fonctionnalités offertes ; elles ont aussi démontré qu'un tel produit correspondait à une véritable demande des utilisateurs, professionnels de la langue, enseignants et chercheurs, élèves et étudiants, et, plus globalement, usagers de notre langue.

Nous sommes, aujourd'hui, particulièrement heureux de cette publication du *TLFi* sous forme cédérom, convaincus qu'il est du devoir de la recherche publique de valoriser auprès du plus grand nombre cette production-phare du laboratoire.

Mais nos recherches en lexicologie et lexicographie sur la langue française ne vont pas s'arrêter là. Si cette version informatisée correspond à un reflet fidèle de la version du *TLF* précédemment publiée, notre objectif est de la faire vivre et d'y intégrer, dans des versions ultérieures, des améliorations de contenu liées à l'évolution de nos connaissances sur la langue, son histoire et son usage, mais aussi des ajouts prenant en compte les évolutions de la langue au cours des périodes les plus récentes afin de permettre au *TLFi* d'évoluer et de demeurer dictionnaire de référence de la langue française.

Nancy, le 13 juillet 2004

1. TLF, Dictionnaire de la langue du 19e et 20e siècle, CNRS Gallimard, Paris, 1971 - 1994.
2. ATILF " Analyse et Traitement Informatique de la Langue Française " UMR 7118 CNRS Université Nancy 2, 44, avenue de la Libération BP 30687 54063 Nancy cedex ; site Web : <http://www.atilf.fr> ; courriel : [contact@atilf.fr](mailto:contact@atilf.fr)
3. GAMMA 60 au début, CII 10070 ensuite, puis Multics : les besoins en informatique de ce vaste projet furent ainsi très intimement mêlés à l'évolution de l'IUCA de Nancy (Institut Universitaire de Calcul Automatique), ancêtre du CIRIL d'aujourd'hui (Centre Interuniversitaire de Ressource Informatique de Lorraine).
4. Pour la sélection des textes, P. Imbs invoque le " principe d'autorité " qui consiste à se référer d'abord à des avis externes reconnus comme dignes de créance, recueillis dans des manuels d'histoire littéraire (Ceux-ci sont mentionnés dans la Préface du *Trésor de la Langue Française*) ou directement auprès de spécialistes, sans écarter pour autant l'expression des compétences internes du laboratoire.
5. Bernet C. et Pierrel J.M, (2004) " Histoire de Frantext : constitution d'une base textuelle (1964-2002) et perspectives ", in *L'édition électronique en littérature et lexicographique : évaluation et bilan*, Presses Universitaires de Rouen, Editions Champion.  
Cette base de données textuelles, largement utilisée aujourd'hui pour l'enseignement ou la recherche est accessible par abonnement à l'adresse [www.atilf.fr/frantext](http://www.atilf.fr/frantext).
6. Par référence à l'ordinateur Gamma 60 de Bull dont le laboratoire était doté à cette époque.
7. Gorcy G., Martin R., Maucourt J., et Vienney R., (1970) " Le traitement des groupes binaires ", Cahiers de Lexicologie, n° 2, p. 15-46.
8. Dans sa postface au dictionnaire, publiée en introduction du tome 16, Bernard Quémada situe en 1981 la proposition d'informatisation.
9. Robert Martin succéda à Bernard Quémada à la direction de l'INaLF.
10. Piotrowski D. (sous la direction de), *Lexicographie et Informatique : autour de l'informatisation du Trésor de la Langue Française* , INaLF, Didier Erudition, Paris, 1996.