Voyage au cœur du langage : le *Trésor de la langue française* et *Frantext*

Fruit d'un énorme travail lexicographique qui dura 30 ans, le *Trésor de la langue française*, aujourd'hui informatisé, a été le premier dictionnaire fondé sur une analyse des usages effectifs des mots, réalisée à travers l'exploitation d'un vaste ensemble de textes français devenu la base *Frantext*.

PASCALE BERNARD et VÉRONIQUE MONTÉMONT
ATILF-CNRS

Le TLFI : **http://atilf.atilf.fr/tlf.htm**
Frantext : **www.frantext.fr**

La question de la production d'un grand dictionnaire de la langue française du XX^e siècle s'est posée avec force à la fin des années cinquante. Elle a été formalisée en novembre 1957, lors d'une conférence de Paul Imbs, professeur à la Faculté des Lettres et directeur du Centre de philologie romane de Strasbourg, à l'occasion du colloque « Lexicologie et lexicographie françaises et romanes ». L'alternative qui s'offrait était la suivante : fallait-il republier le Littré, tombé dans le domaine public, ou au contraire repartir sur de nouveaux frais pour offrir un « exemple-type de lexicographie scientifique moderne[1] » ? C'est la deuxième solution qui a été retenue, et qui a constitué le point de départ de l'aventure du *Trésor de la langue française*. Il faut rappeler que le contexte de l'époque invitait à ce type de renouvèlement épistémologique : la cybernétique, qui essaimait aux États-Unis depuis la fin des années cinquante, avait familiarisé avec l'idée que les machines pouvaient se révéler de puissants relais de l'intelligence humaine, et avait jeté les bases de ce qui allait devenir l'informatique. La révolution structuraliste, dont la linguistique serait le fer de lance, s'apprêtait quant à elle à rénover la discipline : son vœu était d'offrir une pensée panoramique et systémique de la langue. Le projet du futur *Trésor de la langue française (TLF)* présentait une parfaite convergence avec ces manières ambitieuses d'envisager le savoir. Ses objectifs, définis par Paul Imbs, étaient multiples : bâtir un dictionnaire de référence du français, le doter d'une dimension historique (chaque mot bénéficierait de sa rubrique étymologique) et linguistique (avec une description du mot en contexte). Un corpus d'exemples particulièrement riche viendrait étayer les définitions, et permettrait d'offrir une analyse concrète des usages du mot dans la langue[2].

Dans les faits, la réalisation du dictionnaire s'est traduite par la création d'un centre de recherche, installé à Nancy en 1960, qui deviendra rapidement l'INaLF (Institut national de la langue française) – aujourd'hui ATILF-CNRS. Des moyens humains et matériels considérables ont été mis en œuvre : une équipe de cent personnes a ainsi œuvré pendant trente ans à la confection de ce dictionnaire, laissant un fonds de centaines de milliers d'archives désormais conservé à Nancy. Un noyau de mille textes, littéraires (80 % du corpus) et scientifiques et techniques (les 20 % restants) a été « mécanographié », de manière à pouvoir faire l'objet d'une consultation informatisée. Chaque mot a ensuite été traité dans un « dossier » : compilation systématique de toutes les définitions de dictionnaires existants, compléments d'information éventuels puisés dans des dictionnaires étrangers ou dans la littérature spécialisée (jusqu'à des copies de schémas si nécessaire !), auxquels on ajoutait les informations offertes par le traitement automatisé. On pouvait ainsi calculer la fréquence de chaque mot, sa distribution, éditer une fiche le présentant dans son contexte d'apparition, et l'assortir d'un ou plusieurs exemples – au total, le dictionnaire en offrira 430 000. Le premier tome du *TLF* parait en 1971, le dernier en 1994 : au total, 100 000 mots, 270 000 définitions, 16 volumes de référence, salués pour leur qualité et leur aboutissement, mais dont le cout reste un obstacle pour une diffusion véritablement large.

C'est encore la bonne fée informatique, en la personne de Jacques Dendien, qui viendra se pencher une seconde fois sur le berceau du *TLF* pour lui donner une nouvelle jeunesse, et surtout, élargir son public. En effet, le dictionnaire, peu maniable, reste d'abord utilisé par des spécialistes, et se consulte plus volontiers dans les bibliothèques que chez soi. Le développement des réseaux, couplé à la mise au point de systèmes de balisage fin des données informatisées, a permis une formidable démocratisation de l'outil. En premier lieu, le dictionnaire a été « rétroconverti », c'est-à-dire encodé au format numérique, après ressaisie des huit premiers volumes et nettoyage des bandes de photocomposition de tous les suivants. Le processus s'effectuait en deux temps : d'abord une rétroconversion de premier niveau, qui découpait les articles en objets (définition, indicateurs grammaticaux, auteurs, synonymes, antonymes, etc.). Un second niveau analysait quant à lui la structure hiérarchique de chaque article, liant les définitions à leurs différents indicateurs (de nom de domaine, sémantique, stylistique), de manière à proposer l'éventail de requêtes le plus large possible.

Un moteur de recherche, Stella, qui est en fait un véritable logiciel, a été élaboré au laboratoire et greffé sur l'ensemble. Au lieu de simplement feuilleter le dictionnaire en quête d'un mot, l'utilisateur peut pro-

1. Paul Imbs dir., *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles (12-16 novembre 1957)*, Paris, Éditions du CNRS, 1961.

2. Sur la genèse du dictionnaire, voir : Jean-Marie Pierrel et Éva Buchi, « Research and Resource Enhancement in French Lexicography : the ATILF Laboratory Computerised Resources », *in* : Silvia Bruti, Roberta Cella et Marina Foschi Albert dir., *Perspectives on Lexicography in Italy and Europe*, Newcastle-upon-Tyne, Cambridge Scholars Publishing, 2009, p. 79-117.

« Pour le père de ces Chenouville on disait notre oncle, car on n'était pas assez gratin à Féterne pour prononcer notre "onk", comme eussent fait les Guermantes, dont le baragouin voulu, supprimant les consonnes et nationalisant les noms étrangers, était aussi difficile à comprendre que le vieux français ou un moderne patois. »

Marcel Proust, *Sodome et Gomorrhe*, 1922. Citation extraite du TLFi à la définition du mot « baragouin »

fiter de toutes les ressources croisées du *TLF* : recherche d'un mot en orthographe approximative, tri des différentes acceptions, recherche d'un mot à l'intérieur d'un champ disciplinaire donné, recherche de séquence comportant telle ou telle catégorie grammaticale, affichage et mise en surbrillance de tous les exemples contenant un mot donné, etc. Ce *Trésor de la langue française informatisé (TLFi)* a été diffusé sous forme de cédérom, version Mac et PC, en 1998, par CNRS Éditions. En parallèle, le dictionnaire a été mis en ligne en 2001 et il est rapidement devenu la ressource lexicographique de référence de l'Internet, cité jusque dans certains sites de discussion ou blogs ! Actuellement, le *TFLi* sert environ 300 000 requêtes par jour, et a pleinement rempli son contrat initial : être un grand dictionnaire de langue française, fondé sur une exploration minutieuse des textes, offrant et une information linguistique exhaustive, et des outils intelligents pour y accéder.

En parallèle, le réservoir de textes à partir duquel a été élaboré le corpus d'exemples n'a cessé d'être abondé. Au départ simple outil et « sous-produit de la recherche », cet ensemble de ressources numérisées est vite apparu comme une richesse intrinsèque, méritant une plus large mise à disposition. Baptisée *Frantext* (pour « Textes français »), équipée d'un logiciel de recherche, Stella, celui-là même qui accompagne le *TLFi*, la base de données a été mise en service dès 1984. Ses fonctionnalités, très étendues, autorisent les requêtes simples comme les plus complexes. Dans un premier temps, chaque texte est décrit selon un ensemble de métadonnées, ce qui permet à l'utilisateur de composer un corpus trié en fonction de paramètres précis : par exemple, tous les romans écrits entre 1950 et 1970, l'ensemble des volumes de *À la recherche du temps perdu*, ou encore tous les ouvrages comportant le mot *révolution* dans leur titre.

Les textes eux-mêmes ont fait l'objet d'un balisage (d'abord propriétaire, puis normalisé selon les standards XML-TEI). Un quart d'entre eux a été catégorisé grâce à un analyseur qui a identifié la nature grammaticale de chaque élément. Le logiciel de recherche recourt à des outils linguistiques classiques, comme les expressions régulières ; mais il y ajoute une série de liens logiques, permettant toutes sortes de combinaisons. C'est pourquoi l'utilisateur peut formuler des demandes variées : un mot, un lemme (c'est-à-dire l'ensemble des formes fléchies d'une vedette de dictionnaire), un syntagme, une catégorie grammaticale, une liste, une expression de séquence (combinaison de formes…). Ainsi, il est possible de rechercher dans un texte toutes les occurrences de l'adjectif *blanc*, mais aussi des combinaisons du type *chat (noir|blanc|gris)*, le mot *chat* non suivi de *blanc*, les co-occurrences de *chat* et *moustaches*, etc. À un niveau supérieur, l'utilisateur peut composer des séquences appelées *grammaires*, qui modélisent des séries d'informations complexes : expression de la date, occurrences de noms de nombre, motifs syntaxiques. Une fois les occurrences trouvées, elles sont affichées, en surbrillance, dans un contexte de 350 signes ; l'affichage localise aussi la référence, à la page près, de la citation. Enfin, *Frantext* peut trier, en un temps record, le matériau lexical d'un corpus, classer la liste des fréquences, analyser les co-occurrences de deux termes, ou extraire à partir d'une expression régulière tous les mots d'un corpus contenant ou ne contenant pas une série de lettres. Les combinaisons sont infinies, et il n'est guère de requête qui ne puisse être exaucée par Stella, moyennant un petit apprentissage de ses fonctionnalités.

D'abord mise partiellement à disposition sous forme d'un cédérom, « Discotext » (1984), puis par Minitel, Frantext a suivi les développements de la micro-informatique et a trouvé en 1998 un nouveau débouché sur Internet. Proposée sur abonnement – contrainte liée à la présence de textes sous droits –, la base compte aujourd'hui près de 250 bibliothèques et institutions abonnées à travers le monde, ainsi qu'un grand nombre d'utilisateurs individuels. Son enrichissement se poursuit, puisqu'elle propose désormais plus de 4 000 titres et 243 millions de mots ; l'offre se spécialise, aussi, avec une base proposant chaque année les textes au programme de l'agrégation et du concours d'entrée à Normale Sup. Bien que, désormais, de grandes institutions ou entreprises, comme Gallica ou Google Livres, aient considérablement élargi l'offre numérique, *Frantext* a gardé toute sa pertinence au sein de la communauté scientifique : à la fois base de données et logiciel perfectionné, elle permet de mener à bien des enquêtes lexicales ou lexicographiques poussées, sur un corpus échantillonné et segmentable en fonction des besoins. De plus, les passerelles entre la base et le dictionnaire permettent d'hypernaviguer de l'un à l'autre d'un simple clic et de visualiser les mots dans leur contexte d'emploi. *Frantext* et son grand frère le *TLFi* représentent donc deux outils privilégiés d'exploration de la langue au cœur du monde francophone. ■