

Le Trésor de la Langue Française Informatisé : un dictionnaire de référence accessible à tous

Jean-Marie Pierrel

▶ **To cite this version:**

Jean-Marie Pierrel. Le Trésor de la Langue Française Informatisé : un dictionnaire de référence accessible à tous. AMOPA, 2006, 174, pp.25-28. 10.1007/s00726-011-0887-1 . halshs-00398661

HAL Id: halshs-00398661
https://shs.hal.science/halshs-00398661

Submitted on 24 Jun 2009

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.

Le Trésor de la Langue Française informatisé :

Un dictionnaire de référence accessible à tous

Jean-Marie Pierrel

ATILF – CNRS / Nancy Université
44, avenue de la Libération, BP 30687, F-54063 Nancy cedex
Jean-Marie.Pierrel@atilf.fr

Le TLFi, Trésor de la Langue Française informatisé, est la version informatique du Trésor de la Langue Française (TLF)[1], dictionnaire de la langue du 19e et du 20e siècle, publié par le CNRS en 16 volumes entre 1971 et 1994. Pur produit nancéien, ce dictionnaire est en effet le fruit d'une grande aventure lexicographique qui, sous les directions successives du Recteur Imbs et du Professeur 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[2] se veut être aujourd'hui le digne successeur.

Le TLF successeur du Littré

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 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

[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

Aout 2006

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. 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. 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, ordre alphabétique des contextes gauche et droit, ou encore ordre définis selon les constructions syntaxiques propres à chaque partie du discours. Ces concordances étaient utilisées pour un premier tri de la documentation et permettaient d'obtenir, dans un second temps, des contextes élargis parmi lesquels furent sélectionnés les exemples (plus de 430 000) finalement retenus dans le dictionnaire.

Du TLF à sa version informatisée

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. 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[3] dans l'introduction de l'ouvrage *Lexicographie et Informatique*[4], « 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 intervenus 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 rétro-conversion du TLF pouvait être réalisée au sein même du laboratoire.

Un juste retour des fruits de la recherche vers nos concitoyens

Lorsqu'en janvier 2001, à la demande du CNRS, j'ai accepté de m'investir dans la direction de l'ATILF, 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 et demandait à être peaufiné, sa pertinence devant encore être prouvée. Une des premières tâches à laquelle nous

[3] Robert Martin succéda à Bernard Quémada à la direction de l'INaLF.
[4] Piotrowski D. (sous la direction de), *Lexicographie et Informatique : autour de l'informatisation du Trésor de la Langue Française*, INaLF, Didier Erudition, Paris, 1996.

nous sommes donc attelés fut d'ouvrir très vite une 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 200 000 pages par jour ouvrable aujourd'hui (www.atilf.fr/tlfi). Ces 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. En novembre 2004, nous proposions chez CNRS Editions une version cédérom pour PC, suivie un an après d'une version MacOS (www.tlfi.fr).

Comme je le notais en conclusion de mon exposé lors du colloque international de l'AMOPA à l'Institut de France « Les dictionnaires français, une pérennité prestigieuse et des métamorphoses pionnières »[5], ce n'est pas l'informatisation du TLF qui est à proprement parler intéressante[6], c'est plutôt le fait que l'informatisation d'un grand dictionnaire institutionnel permette un partage des savoirs beaucoup plus grand. Aujourd'hui, tout le monde peut accéder à ce grand dictionnaire conçu au départ pour une élite financière et universitaire.

Caractéristique du TLFi

Reflet fidèle de la version papier, le TLFi se distingue, comme le TLF, par la richesse de son matériau et la complexité de sa structure :

- Originalité de sa nomenclature (incluant préfixes, suffixes et autres éléments formants) : c'est en tout 100 000 mots avec leur étymologie et leur histoire, et 270 000 définitions.
- Richesse des objets méta-textuels inclus dans chaque article (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…).
- Richesse des 430 000 exemples, tirés de deux siècles de production littéraire française.

[5] Cf. l'article de notre collègue Jean Pruvost dans le n° 172 de la revue de l'AMOPE, avril 2006.

[6] Ceux et celles d'entre vous qui seraient intéressés par ces aspects d'informatisation pourront se référer à l'article « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », Traitement Automatique des Langues, Vol 44 – n° 2/2003, Hermes Sciences Edition, p. 11-37, Jacques Dendien, Jean-Marie Pierrel

- Diversité des rubriques : une rubrique synchronie couvrant la période 1789 à nos jours, une rubrique étymologie et histoire, et une rubrique bibliographie pour les principaux articles.

La version informatique du TLF intègre de plus des accès à très haut niveau de tolérance permettant une insensibilité aux accents, une tolérance aux fautes d'orthographe courantes, un traitement phonétique et un traitement morphologique. Ainsi, on peut offrir une correction automatique des fautes et permettre des accès à partir de formes et non plus uniquement de lemmes ou de vedettes et proposer des procédures d'accès diversifiées pour une consultation humaine.

Quels accès au TLFi ?

Le TLFi correspond à une rétro-conversion de la version publiée du TLF pour laquelle, par des procédures de repérage semi-automatique des objets textuels composant les articles du dictionnaire original, nous avons introduit un balisage fin, tant typographique (nous avons conservé la totalité des marqueurs typographiques présents dans le texte initial, de manière à conserver une image 100 % fidèle du TLF) que sémantique (repérage des principaux objets textuels au sein de chaque article). Quelques chiffres peuvent donner un aperçu de la finesse de ce balisage : après validation sur l'ensemble des seize tomes, 36 613 712 balises XML ont été positionnées : 17 364 854 balises typographiques, 1 070 224 balises décrivant la hiérarchie, 18 178 634 balises repérant les objets textuels, dont 92997 entrées et 64346 locutions faisant l'objet de 271166 définitions et illustrées par 427493 exemples.

Ce balisage fin du TLF nous a permis de proposer divers accès possibles à l'ensemble du dictionnaire, cumulant les avantages d'un dictionnaire avec ceux d'une ressource textuelle et d'une véritable base de données lexicales :

- Recherche d'un mot, d'une expression ou d'une forme lexicale plus ou moins bien orthographiés, avec possibilité, via un « panneau de réglage », de mettre en évidence divers champs dans le résultat de la recherche (définition, code grammatical, domaine spécifique, exemple, auteur d'exemple, construction, indicateur, etc.).
- Possibilité d'hyper navigation à l'intérieur du dictionnaire permettant en un clic souris de passer d'un mot à sa définition.
- Interrogations assistées ou requêtes complexes exploitant l'ensemble de la structure du dictionnaire à travers le croisement de multiples critères.

Exemples de recherches dans le TLFI

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations sur les recherches offertes dans le TLFi parmi lesquelles :

Recherche d'un mot

Accès à un mot à travers un système de correction automatique (forcée ou non) : ainsi, en introduisant la recherche de la forme *etique* (sans accent) on accède aux deux articles correspondant aux mots *étique* ou *éthique* ; et accès à partir d'une forme : en interrogeant à partir de la forme *sussiez* on obtient automatiquement l'article *savoir*.

Recherche d'une expression

Possibilité d'obtention directe des définitions et conditions d'usage d'expression telles « droit d'aînesse » ou « le trompette » et recherche des expressions composées d'une forme : ainsi, en demandant les mots contenant la forme *queue* on obtient le résultat :

COURTE-QUEUE, adj. et subst.

DEMI-QUEUE, subst. fém.

Hironde (queue d'), loc., (dans l'article HIRONDELLE, subst. fém.)

HOCHEQUEUE, HOCHE-QUEUE, subst. masc.

LEU[1] (à la queue leu leu), voir QUEUE.

PAILLE-EN-CUL, PAILLE-EN-QUEUE, subst. masc.

PORTE-QUEUE, subst. masc.

QUEUE(-)D'ARONDE, voir ARONDE.

QUEUE-DE-CARPE, subst. fém.

QUEUE(-)DE(-)CHAT, subst. fém.

Etc.

Exploitation de la structure du dictionnaire

Par exemple pour rechercher les verbes qui, en marine, concernent le maniement des voiles, il suffit de préciser que l'on recherche dans la classe des verbes ceux qui dans le domaine de la marine correspondent à une définition incluant une forme du mot voile, soit dans une structure plus compacte : [classe : *verbe* ; domaine : *marine* ; définition : *voile*]. Voici un extrait des réponses que l'on obtient :

ABORDER, verbe trans.

3 Se mettre en panne, au vent du navire qu'on veut aborder, et se laisser dériver sur lui en lui présentant toujours le côté, en manœuvrant ses voiles de manière à ne le point dépasser, et en mouillant une ancre avant de l'accrocher.`` (BESCH. 1845). 3 (Aborder à l'ancre.)

ABRIER, ABREYER, verbe trans.
3 Empêcher le vent, en l'interceptant, de passer jusqu'à (une autre voile) : 3
AGRÉER[2], verbe trans.
3 ,,Préparer ou travailler à la garniture, aux agrès d'un bâtiment, fourrer les dormans, estroper les poulies, garnir voiles, vergues, etc. : mais on dit plus souvent qu'on travaille à la garniture des mâts et des vergues, aux voiles et aux rechanges.`` (WILL. 1831) : 3
AMURER, verbe.
3 Fixer l'amure d'une voile pour l'orienter selon le vent : 3
APPAREILLER[1], verbe.
3 ,,Déployer cette voile, et la présenter au vent de telle sorte qu'elle fonctionne convenablement.`` (JAL 1848). 3 (Appareiller une voile.)
ARBORER, verbe trans.
3 Hisser, déployer la voile. 3
ARRISER, verbe trans.
3 Amener un peu les voiles en se servant des ris, afin d'en diminuer la hauteur, en cas de gros temps : 3
ARRIVER, verbe intrans.
3 Élargir l'angle du cap du navire et de la direction du vent, pour augmenter l'effet du vent sur les voiles. 3
ETC……

Recherche des syntagmes construits avec droit (174 dont 87 syntagmes définis) [syntagme : *droit*]

À, prép.
1 à bon droit (agir, penser) 1
ACCUMULATION, subst. fém.
1 Accumulation de droit. 1
AFFÉRENT, ENTE, adj.
1 Droit afférent, part afférente, portion afférente, etc. 1
AFFOUAGE, subst. masc.
1 Droit d'affouage, affouage. 1
AGRAIRE, adj.
1 Droit agraire. 1
AÎNESSE, subst. fém.
1 Droit d'aînesse. 1
Etc…..

Création de listes de mots

Recherche des mots commençant par « c » et se terminant par « able » : 159 réponses dont : *câble, cadable, calculable, calomniable, calottable, canalisable….*

Recherche sur un champ lexical

Recherche des mots dont la définition utilise le mot *liberté* : 306 réponses dont :

Objets de la recherche : 1 Définition 1
ABUSER, verbe trans.
1 Exagérer dans l'usage d'une possibilité, d'une liberté : 1
AFFRANCHI, IE, part. passé, adj. et subst.
1 (Celui) à qui on a donné la liberté. 1
AISE[1], subst. fém.
1 Grande liberté. 1
ALIÉNANT, ANTE, part. prés. et adj.
1 Qui prive l'homme de son humanité, de sa liberté : 1
ALIÉNATION, subst. fém.
1 Privation de libertés de droits humains essentiels éprouvée par une personne ou un groupe social sous la pression de facteurs permanents (Hegel) ou historiques (Marx) qui l'asservissent à la nature ou à une classe dominante. 1
ÂME, subst. fém.
1 Aliéner sa liberté, sa dignité... en échange de quelque chose : 1 (Vendre son âme (au diable).)
Etc…..

Recherche complexe

Les interrogations possibles au sein de ce dictionnaire peuvent prendre des formes encore plus complexes. Ainsi, pour trouver une réponse à une interrogation du type : *Quels sont les substantifs empruntés à une langue étrangère (non précisée) et qui, lorsqu'ils sont employés dans le domaine de l'art culinaire, sont illustrés par une définition empruntée au dictionnaire de l'Académie ?*, il convient d'utiliser l'onglet « recherche complexe » et préciser :

Objet 1 : type "*Entrée*"
Objet 2 : type "*Code grammatical*", contenu "*substantif*", lien "*inclus dans l'objet 1*"
Objet 3 : type "*Domaine technique*", contenu "*art culinaire*", lien "*dépendant de l'objet 1*"
Objet 4 : type "*Définition*", lien "*dépendant de l'objet 3*"
Objet 5 : type "*Source*", contenu "*Académie*", lien "*inclus dans l'objet 4*"
Objet 6 : type "*Langue empruntée*", lien "*dépendant de l'objet 1*"

Le lien "inclus dans l'objet 1" de l'objet 2 exprime que l'entrée est un substantif.

Le lien "dépendant de l'objet 1" de l'objet 3 exprime que l'indication de domaine technique est dans la portée de l'objet 1 (l'entrée), et se trouve donc dans l'article consacré à cette entrée.
Le lien "dépendant de l'objet 3" de l'objet 4 exprime que la définition est valable dans le domaine de l'art culinaire.
Le lien "inclus dans l'objet 4" de l'objet 5 exprime que la source de la définition est le dictionnaire de l'Académie.
Le lien "dépendant de l'objet 1" de l'objet 6 exprime que l'objet "langue empruntée" est dans l'article dont l'entrée est l'objet 1.

Une telle interrogation nous fournit comme résultat :

Objets de la recherche : 1 Entrée 1 2 Code grammatical 2 3 Domaine technique 3 4 Définition 4 5 Source 5 6 Langue empruntée 6

MORTIFICATION, subst. fém.
1 MORTIFICATION, subst. fém. 1 2 **subst**. fém. 2 3 **ART CULIN**. 3 1 4 ,,Action de garder certaines viandes pour qu'elles deviennent tendres et gagnent du fumet`` (Ac. 1878, 1935). 4 5 **Ac**. 1878, 1935 5 6 Empr. au lat. 6
NAPOLITAIN, -AINE, adj. et subst.
1 NAPOLITAIN, -AINE, adj. et subst. 1 2 adj. et **subst**. 2 3 **ART CULIN**. 3 2 4 Gros gâteau cylindrique ou hexagonal fait d'une pâte à base d'amandes et fourré de confiture d'abricots et de gelée de groseilles (d'apr. Ac. Gastr. 1962). 4 5 d'apr. **Ac**. Gastr. 1962 5 6 Empr. à l'ital. 6
PANADE, subst. fém.
1 PANADE, subst. fém. 1 2 **subst**. fém. 2 3 **ART CULIN**. 3 3 4 Préparation à base de pain, biscottes ou farine utilisée comme liaison, le plus souvent pour les farces (d'apr. Ac. Gastr. 1962). 4 5 d'apr. **Ac**. Gastr. 1962 5 6 Empr. au prov. 6
ETC.

En guise de conclusion

Le TLFi est, sans aucun doute, le plus grand dictionnaire informatisé consacré à la langue française. La richesse de son contenu, entièrement encodé dans un format XML très poussé, a d'ores et déjà ouvert des perspectives intéressantes.

Sa mise à disposition sous forme de Cédérom ou sur le Web a rencontré un vif succès tant auprès du grand public que des utilisateurs universitaires : objet de plusieurs centaines de milliers de connexions quotidiennes en provenance de tous les continents, il est référencé par d'innombrables sources (écoles, collèges, lycées, universités, associations, publications à caractère culturel ou scientifique, etc.). La notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française.

Le TLF est considéré comme une référence, ce qui lui a paradoxalement donné la réputation tenace d'être un dictionnaire réservé à une élite. Ses interconnexions par hyper navigation avec le dictionnaire de l'Académie, FRANTEXT[7] ou la base historique du vocabulaire français le positionnent aujourd'hui au cœur d'un ensemble de ressources sur la langue française au sein desquelles il joue un rôle actif et prépondérant, démontrant ainsi que sa réputation élitiste est injustifiée.

[7] FRANTEXT : base de données textuelles de l'ATILF riche de près de 4 000 œuvres accessible à l'adresse www.atilf.fr/frantext.