Traduire

Revue française de la traduction

222 | 2010

Traduire pour le théâtre

Le Trésor de la langue française informatisé

Pascale Bernard

Édition électronique

URL : <http://journals.openedition.org/traduire/458>

DOI : 10.4000/traduire.458

ISSN : 2272-9992

Éditeur

Société française des traducteurs

Édition imprimée

Date de publication : 15 juin 2010

Pagination : 125-136

ISSN : 0395-773X

Référence électronique

Pascale Bernard, « Le Trésor de la langue française informatisé », *Traduire* [En ligne], 222 | 2010, mis en ligne le 12 novembre 2013, consulté le 30 avril 2019. URL : <http://journals.openedition.org/traduire/458> ; DOI : 10.4000/traduire.458

Le Trésor de la langue française informatisé

Pascale Bernard

Cet article fait suite à l'intervention présentée par l'auteur lors de la Journée de la traduction professionnelle 2009 à Lyon.

Le dictionnaire *le Trésor de la langue française, dictionnaire de la langue du XIX^e et du XX^e siècle (TLF)*, (CNRS, 1976-1994), est le plus grand et le plus complet des dictionnaires de langue. Il est paru en 16 volumes de plus de 1 500 pages chacun, le premier volume en 1971 et le dernier en 1994.

Comme dans la plupart des dictionnaires, la majorité des mots est présentée selon l'ordre alphabétique, à l'exception des participes passés qui, dans le *Trésor de la langue française*, se trouvent à la suite du verbe dont ils dépendent. Ainsi, le participe passé **pris** se trouve après le verbe **prendre** indépendamment de son ordre alphabétique et comme dans la plupart des dictionnaires imprimés en plusieurs volumes, il n'est pas aisé d'y faire des recherches.

Grâce au TLFi (*Trésor de la langue française informatisé*), suite logique du TLF (Dendien, 1996), cette importante ressource sur la langue française est disponible sur la toile. Associé à un moteur de recherche puissant (Stella), le TLFi permet de satisfaire les besoins de tous les linguistes, aussi bien sémanticiens que stylisticiens, qui peuvent exploiter toutes les possibilités offertes par Stella, mais il satisfait également les besoins de tous les utilisateurs, linguistes ou non. Ces recherches peuvent s'articuler autour des axes suivants : études en vue de repérer des cooccurrences et collocations, extraction de sous-lexiques, études morphologiques, études de syntaxe locale, de sémantique, de stylistiques mais on peut aussi faire la recherche d'un mot dont on ne connaît pas exactement l'orthographe ou lancer une recherche par mots clefs pour retrouver un mot oublié mais dont on connaît le sens, etc.

Le TLFi peut être considéré comme une base de données lexicales en même temps qu'une base de connaissances finement structurée.

Son originalité est fondée sur des spécificités liées à son **contenu** :

- Sa nomenclature est riche d'environ 100 000 entrées (vedettes ou sous-vedettes, c'est-à-dire présentées en remarques ou dérivés). Original aussi le traitement de morphèmes grammaticaux au même titre que les autres mots, ainsi on compte 62 mots traités sous la vedette –O, finale et finale par apocope.
- Sa liste d'objets métatextuels sur lesquels repose la structuration de l'article du dictionnaire : vedettes, codes grammaticaux, définitions, indicateurs stylistiques et/ou sémantiques, exemples, etc. soit, en tout, environ 40 objets métatextuels.
- Sa grande quantité d'exemples référencés : environ 430 000.
- La grande diversité des rubriques proposées, concernant l'étymologie, la prononciation, la bibliographie, etc.

1. Sa structure

- Le dictionnaire a été balisé en XML et des balises spéciales ont été injectées pour pouvoir délimiter et atteindre chaque type d'objet textuel. Ainsi, si la recherche plein texte sur la totalité du contenu du dictionnaire reste toujours permise, il est possible aussi d'éliminer le « bruit » inhérent à un tel type de requêtes. Il est en effet possible de restreindre cette recherche plein texte à une recherche plus « pointue » sur un type particulier d'objet textuel.
- De plus, une seconde dimension a été introduite : ces objets textuels ont été hiérarchisés à l'aide d'une série de balises spécifiques et d'une grammaire de contrôle.

De par le traitement hautement élaboré qui a permis de décomposer le texte du TLF en objets élémentaires (définitions, exemples, indicateurs de domaine technique, indicateurs sémantiques, grammaticaux, stylistiques, etc.) et d'analyser les relations hiérarchiques liant ces objets, il résulte des possibilités d'interrogation virtuellement illimitées.

2. Les recherches

L'accès aux possibilités les plus raffinées exige la compréhension des concepts permettant de mettre en œuvre des recherches très puissantes.

Trois niveaux d'utilisation de puissance croissante sont proposés :

La consultation simple des articles, qui apporte d'emblée des facilités de recherche grâce au correcteur d'erreurs.

Les recherches assistées, qui fournissent des possibilités étonnantes de consultation du dictionnaire, et dont la mise en œuvre ne nécessite que quelques exemples types très simples.

Les recherches complexes, qui offrent les possibilités maximales d'investigation. Elles demandent la consultation d'une documentation présentant les différents objets rencontrés dans le TLF et les concepts permettant d'effectuer les recherches les plus raffinées.

L'utilisateur pressé pourra se contenter des deux premiers niveaux d'utilisation. Qu'il sache qu'il a en réserve un potentiel considérable. Il pourra tester l'**hypernavigation** à travers les pages gérées par le même logiciel, ce qui apporte un plus à sa connaissance.

2.1. Consultation simple des articles

On dispose de **trois possibilités de consultation** pour visualiser un article.

2.1.1. Avec le correcteur d'erreurs

De nombreuses requêtes échouent dans la consultation d'un dictionnaire, indépendamment de son support, papier ou version électronique, car la recherche d'un mot dans un dictionnaire suppose que le lecteur connaisse l'orthographe exacte du mot. Ne plus savoir écrire **chrysanthème** ou **ornithorynque**, se tromper dans les accents (**régulièrement** ou **reguliere-ment**) ou les doubles consonnes (**ramonnage**) n'est pas un problème puisqu'il suffit de taper phonétiquement le mot recherché en un ou plusieurs mots (**règue uli ère man, crise en taime, corps nichon**).

Deux modes sont proposés, mode automatique ou mode forcé.

En mode automatique, le correcteur intervient si vous avez tapé une forme fautive comme celles citées précédemment, il n'intervient pas si vous avez tapé une forme existante, le logiciel affichera seulement la solution trouvée.

Ainsi si vous cherchez le mot **étique**, il vous affichera uniquement ce mot puisqu'il l'a trouvé orthographié ainsi.

En mode forcé, le correcteur intervient et vous propose toutes les solutions trouvées. Ainsi pour le mot **étique**, que vous pouvez taper **etik**, il vous proposera **étique** et **éthique**.

2.1.2. Avec l'utilisation de listes défilantes

Les listes défilantes permettent de parcourir le TLF, comme on pourrait le faire dans la version papier, mais elles affichent uniquement les mots traités en vedette d'articles, c'est-à-dire en entrées principales. Si un mot ne figure pas dans cette liste, il ne faut pas en conclure qu'il n'est pas dans le TLF, on pourra le chercher grâce aux autres possibilités de recherche des autres formulaires.

La recherche se fait en deux étapes :

- Choix de la tranche alphabétique dans laquelle se situe le mot cherché
 - Exemple : VENTOUSE à VETEMENT
- Choix du mot dans la tranche retenue
 - Exemple : VERNAL

2.1.3. Avec la saisie phonétique

Cette consultation permet de rechercher un mot par sons et non par caractères alphabétiques. Certains sons sont regroupés. Lorsque le pointeur de souris est positionné sur un son il donne, dans le cadre « Explication », un ou des exemples du son ou des sons représentés.

2.2. La recherche assistée

Cette recherche se fait par l'intermédiaire d'un formulaire de recherche sur lequel nous allons pouvoir faire cinq types de recherche que nous pourrons combiner entre elles.

Illustration : Marlène Junius, <http://alotoftralala.over-blog.com>

Ce formulaire permet alors de considérer le *Trésor de la langue française* comme une véritable base de données.

Les cinq possibilités de base permettent de rechercher par :

- Vedette contenant un mot donné
 - Exemple : **tête**
 - Le logiciel affiche toutes les entrées contenant ce mot, et permet de découvrir tous les mots composés avec cette entrée,
 - Jadis et autrefois et leurs fréquences.
- Code grammatical
 - Un menu déroulant propose 12 codes grammaticaux. On pourra obtenir les autres codes dans le formulaire de recherche complexe (élément formant, préfixe, suffixe, interjection, ...).
- Domaine technique particulier
 - On peut choisir dans la liste proposée soit par ordre alphabétique, soit par thème, exemple : « mythologie ».
- Indicateur d'emploi (grammatical, stylistique ou sémantique)
 - Exemple : « péjoratif ».
- Objet textuel à choisir à partir d'un menu déroulant. Si l'on se sert de cette proposition, il faut impérativement choisir un contenu.
 - Exemple : « auteur d'exemple » contenant **zola** permet de trouver tous les exemples de Zola cités dans le dictionnaire.

La recherche peut être restreinte à l'une des possibilités ou croiser plusieurs critères :

- Chercher les animaux du domaine de la mythologie
 - Domaine : « mythologie ».
 - Définition contenant le mot **animal**.
- Chercher tous les verbes employés par Victor Hugo dans un contexte populaire
 - Code grammatical : verbe.
 - Indicateur : populaire.
 - Auteur d'exemple : **hugo**.
- Chercher tous les substantifs du domaine de la gastronomie
 - Code grammatical : substantif.

- Domaine : gastronomie

(on notera que sous ce domaine nous avons regroupé arts culinaires, cuisine et gastronomie qui sont assez proches et dont l'utilisateur peut ne pas saisir la nuance). La recherche détaillée de chacun des domaines est toujours possible soit dans le cadre 5 de la recherche assistée, soit dans la recherche complexe.

Ce formulaire présente de nombreuses possibilités de recherches.

On peut ainsi extraire de ce grand dictionnaire, *le Trésor de la langue française*, des sous-ensembles de dictionnaires. On peut se construire, comme on l'a montré précédemment, un dictionnaire des animaux de la mythologie, mais aussi tous les sous-dictionnaires possibles, tel un dictionnaire des mammifères rongeurs, un dictionnaire des arbres, etc. Toutefois, pour ce dernier, il faudra placer l'objet textuel recherché « arbre » dans la définition car on obtient des réponses justes, certes, correspondant à notre requête, mais trop vagues pour se faire un sous-ensemble des arbres.

On peut chercher par mots clefs pour retrouver un mot oublié dont on connaît au moins un des éléments de la définition. Ainsi lorsqu'on recherche un mot qui ne nous revient pas, il suffit de taper un ou plusieurs mots qui pourraient se trouver dans l'objet définition pour retrouver ainsi le mot oublié.

Par exemple :

- Comment se nomment les pierres projetées par les volcans ?
- Comment retrouver les noms de ces grosses bouteilles de champagne dont on a oublié tous les noms ?
- Quel est le mot précis qui décrit cette tendance qu'on a à toujours remettre au lendemain ?
- Comment retrouver l'adjectif ou les adjectifs qui sont relatifs au printemps et à l'hiver ? On se souvient bien de printanier et d'hivernal plus communément employés mais on a un vague souvenir d'avoir rencontré d'autres mots à connotation plus littéraire. Comment les retrouver ?
- Qui a dit « ce siècle avait deux ans » ?

Pour trouver les pierres projetées par les volcans :

Dans le cadre 5 de la recherche assistée, on sélectionne dans le menu déroulant, on choisit l'objet « Définition » et dans « Contenu 1 », il faut taper **pierre** et dans « Contenu 2 », **volcan**.

La solution trouvée est LAPILLI.

Pour retrouver les noms des grosses bouteilles de champagne :

Au même endroit, on sélectionne « Définition » et on tape dans « Contenu 1 », **bouteille** et dans « Contenu 2 », **champagne**

Les solutions affichées donnent plusieurs résultats dans lesquels on retrouve BALTHAZAR, JÉROBOAM, MATHUSALEM et NABUCHODONOSOR.

Pour retrouver cette tendance à remettre toujours ce qu'on peut au lendemain :

Toujours le même principe, dans le champ « Définition », taper dans « Contenu 1 », **remettre** et dans « Contenu 2 » **au lendemain**.

Les solutions affichées sont PROCRASTINATION et PROCRASTINER.

Pour retrouver les mots littéraires qui se rapportent au printemps et à l'hiver :

Dans le cadre 4, « Le passage est consacré à un indicateur d'emploi », il suffit de sélectionner dans le menu déroulant « Littéraire » et dans le cadre 5, de sélectionner l'objet « Définition » et de taper dans « Contenu 1 » **printemps**. Dans les trois solutions, on retrouve aussitôt le mot recherché VERNAL, et on procède de même avec le mot **hiver** qui donnera parmi les solutions proposées le mot recherché HIÉMAL.

Et enfin pour retrouver une citation d'auteur, il suffit de sélectionner dans le menu déroulant du cadre 5 « Paragraphe quelconque ». Cela indique que le logiciel va effectuer une recherche plein texte, tant dans la partie synchronique que diachronique, qui n'a pas été balisée. Cet exemple de Victor Hugo a été cité trois fois dans le TLF sous les mots PERCER, SIÈCLE et SOUS.

Cette possibilité de recherche plein texte dans le dictionnaire est très utile pour retrouver les mots cachés sous d'autres mots. En effet, par gain de place, nous avons souvent regroupé les mots d'une même famille sous un nom de base, ce qui ne facilite pas la consultation et la recherche de mots dans le TLFi et ce qui la rend quasiment impossible dans la version imprimée.

Ainsi, le mot **myxomatose** est traité en dérivé sous un mot moins connu **myxome** ; grâce à cette possibilité de recherche plein texte, l'utilisateur du TLFi peut le retrouver en un clic de souris.

Lorsqu'un mot absent des volumes précédemment traités était nécessaire dans une définition ou qu'il était présent dans un exemple, le rédacteur avait alors la possibilité de gloser ce mot, c'est-à-dire de le définir à l'endroit où il en avait besoin. Ces mots cachés sont impossibles à trouver dans la version papier du dictionnaire, alors qu'ils sont repérables facilement dans la version électronique grâce à cette recherche plein texte.

Par exemple, le mot régional lorrain **migaine** n'a pas été traité lors de la rédaction de la lettre M, or le rédacteur du mot **quiche** citait un auteur lorrain qui employait ce mot dans sa définition de quiche. Le rédacteur a alors défini le mot **migaine** sous le mot quiche.

La recherche plein texte permet aussi de chercher des expressions sur lesquelles on aimerait avoir des précisions quant à leur bon usage.

Ainsi pour des expressions de plus en plus répandues comme **au jour d'aujourd'hui**, ou **voire même** on peut obtenir rapidement le maximum de renseignements sur leur bon emploi et bon usage.

2.3. La recherche complexe

Les requêtes peuvent être élémentaires, un seul objet recherché, ou multi-objets. Le formulaire de Recherche assistée permet déjà de nombreuses recherches mais il a ses limites. Dès qu'on veut travailler par listes de mots ou imposer la place du mot recherché dans un objet textuel il faut utiliser le formulaire de recherche complexe.

2.3.1. Requêtes élémentaires

Les requêtes les plus simples portent sur un seul objet.

Vous devez impérativement spécifier le type de l'objet que vous recherchez en le sélectionnant dans le menu déroulant et éventuellement lui imposer un contenu textuel.

Ainsi, pour extraire le dictionnaire des arbres, il suffit de rechercher les définitions qui contiennent le mot arbre. Les nombreux résultats affichés indiquent clairement qu'il y a beaucoup de bruit dans la requête et que tous les résultats obtenus ne sont pas tous des arbres. L'aide proposée et associée à la case « Contenu », nous indique comment procéder pour imposer la place d'un mot dans un objet textuel, à savoir **&d** suivi d'un nombre indiquant le nombre de mots qu'on désire avant le mot recherché.

Dans notre exemple, on peut tolérer 0 mot (si arbre est en début de définition), 1 mot (pour des définitions du genre « Petit arbre... », ou 2 mots (pour des définitions du genre « Sorte d'arbre... »).

2.3.2. Requêtes multi-objets

Il est possible de faire des requêtes sur plusieurs objets, par exemple, on désire rechercher les définitions qui contiennent le mot **instrument** et qui sont empruntées au dictionnaire de l'Académie.

Une telle requête met en jeu deux objets :

- 1) un objet de type « Définition » ayant le contenu textuel instrument,
- 2) un objet de type « Source » ayant le contenu textuel académie.

La notion de liens

Pour exprimer une telle requête complexe, il ne suffit pas d'énumérer deux objets, définition et source, il faut définir dans ce formulaire un lien logique entre les objets. Si ce lien était calculé automatiquement dans la recherche assistée, l'utilisateur doit le définir lui-même à ce niveau. Il existe deux types de liens de nature distincte, inclusion ou dépendance.

Pour exprimer la nature du lien, il ne faut pas oublier de se référer à l'aide fournie dans le cadre « Liens ».

Lien « objet X inclus dans objet Y »

Si on reprend l'exemple précédemment cité et que l'on observe la structure d'une telle définition, on voit bien que l'objet Définition contient à la fois l'énoncé de la définition et la source. Autrement dit, l'objet « Source » est inclus dans la « Définition ».

Grâce au cadre intitulé « Peindre les objets », on voit également très clairement quels objets sont inclus les uns dans les autres :

- les sources des définitions sont incluses dans les définitions,
- les sources des exemples (auteur, titre, date) sont incluses dans les exemples,
- les sources des syntagmes sont incluses dans les syntagmes,
- les codes grammaticaux dans les entrées,
- les mots vedettes dans les entrées.

Lien « objet X dépendant de l'objet Y »

Un objet X est dépendant d'un objet Y si X est hiérarchiquement inférieur.

La vedette d'un article, ainsi que toute information qui apparaît au début de l'article avant que ne commence la numérotation des niveaux hiérarchiques, sont de toute évidence hiérarchiquement supérieures à tout autre objet de l'article. Par convention, tout objet de l'article est considéré comme dépendant de la vedette, ou des informations du début.

Voyons, à travers quelques exemples, la puissance que confère aux recherches la notion de dépendance hiérarchique des objets :

Quelles sont les définitions contenant le mot « mets », utilisées dans le domaine de l'art culinaire, et illustrées par des exemples de Zola ?

Objet 1 : type « Définition », contenu **mets**, lien « Dépendant de l'objet 2 »

Objet 2 : type « Domaine technique », contenu **art culinaire**

Objet 3 : type « Auteur d'exemple », contenu **Zola**, lien « Dépendant de l'objet 1 » (on estime raisonnablement qu'un exemple dépendant d'une définition illustre cette définition)

NOTE : Plusieurs liens peuvent être attachés à un même objet, comme le montre l'exemple ci-dessous :

Quelles sont les locutions populaires ayant trait à la religion ?

Objet 1 : type « Indicateur », contenu **populaire**

Objet 2 : type « Domaine », contenu **religion**

Objet 3 : type « Indicateur », contenu **locution**, liens « Dépendant de l'objet 1 » **et** « Dépendant de l'objet 2 »

Le double lien « Dépendant de l'objet 1 » et « Dépendant de l'objet 2 » indique que la locution est à la fois sous l'influence de **populaire** et de **religion**.

Le remplissage du formulaire ne vous impose pas de penser dans tel ou tel ordre : vous ne faites que décrire une collection d'objets, dans l'ordre où ils vous viennent à l'esprit, en énonçant des liens logiques simples qui les unissent.

2.4. Les listes de mots

Les listes de mots peuvent être créées manuellement, ou créées par flexion de lemmes, ou créées à partir de l'ensemble des graphies du TLF. Elles peuvent être modifiées, elles peuvent être détruites.

Pour la création à partir des graphies du TLF, le critère de sélection permet de définir quelles graphies seront sélectionnées.

Grâce à ce formulaire de « Listes de mots », nous allons pouvoir établir des recherches par début de mots, par radical ou par finales.

Prenons, par exemple la recherche des outils qui se terminent en **—oir**.

Pour obtenir les entrées du dictionnaire qui se terminent en **—oir**, on ne peut pas demander directement ces entrées, il faut procéder en deux temps :

- d'abord établir la liste des mots qui se terminent en **—oir** en tapant dans le critère de sélection du cadre « Création d'une liste extraite du TLF » : **.*oir** (lire l'aide associée à cette recherche). Ne pas oublier de donner un nom à la liste, **oir** par exemple. La liste des mots obtenus s'affiche dans le cadre 1.
- et chercher dans cette liste les mots qui contiennent **outil** dans leur définition.
 - Dans Recherche complexe :
- Type de l'objet : « Entrée », contenu **&loir**
- Type de l'objet : « Définition », lien : « Dépendant de l'objet 1 », contenu **outil** (56 réponses)

2.5. L'hypernavigation

Le TLFi fournit un ensemble de plus d'environ 100 000 entrées et un *corpus* de 430 000 exemples.

Grâce à l'hypernavigation il est possible de faire des liens, en un clic, vers

- d'autres dictionnaires,
 - le TLFi,

- 3 dictionnaires de l'Académie :
 - 4^e édition : 1762,
 - 8^e édition : 1932-1935,
 - 9^e édition : 1994, en cours de rédaction, de A à Plébéien.
- d'autres exemples de la base Frantext,
- une base de connaissance lexicale,
- la base historique du vocabulaire français.

Conclusion

Certaines de ces ressources, financées essentiellement par le CNRS, constituent un patrimoine important et nous avons le souhait de les faire connaître et de les rendre accessibles à tous. Le dictionnaire *Le Trésor de la langue française*, conçu au départ sous sa version papier comme un dictionnaire à l'usage des spécialistes de la langue, est devenu grâce à son informatisation et sa mise à la disposition de tous, une ressource incontournable sur la langue française.

pascale.bernard@atilf.fr

Pascale BERNARD est ingénieure de recherche au laboratoire ATILF (Analyse et traitement informatique de la langue française) UMR 7118, Nancy. Sa formation d'angliciste (licence d'anglais) et de linguiste (licence, maîtrise et DEA de linguistique et phonétique) lui ont permis de pratiquer ses expériences professionnelles comme lexicographe à l'Institut national de la langue française (ancien nom du laboratoire ATILF), en tant que rédactrice du dictionnaire du TLF et comme linguiste en linguistique computationnelle dans un projet européen de traduction automatique pour la CEE (EUROTRA).

Elle a travaillé à l'informatisation du Trésor de la langue française (TLF), et outre ses travaux de recherche lexicographiques actuels, notamment une mise à jour d'un volume de Supplément du TLF, elle participe à l'évolution de la base Frantext et à la valorisation des bases du laboratoire au travers de cours et de conférences.

Références

- CNRS (1976-1994) *TLF, Dictionnaire de la langue du XIX^e et XX^e siècle*, CNRS, Gallimard, Paris.
- DENDIEN J. (1996). Le projet d'informatisation du TLF, in *Lexicographie et informatique*, p. 25-34.
- BERNARD P., BERNET C., DENDIEN J., PIERREL J.-M., SOUVAY G. et TUCSNAK Z. (2001) « Un serveur de ressources linguistiques informatisées via le Web », *Actes de TALN 2001*, Tours, Juillet 2001, p. 333-338.
- BERNARD P., DENDIEN J., LECOMTE J., PIERREL J.-M. (2002) Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella, *8^e Journées Internationales d'Analyse Statistique des Données Textuelles JADT 2002*, p. 137-149.
- BERNARD P., LECOMTE J., DENDIEN J., PIERREL J.-M. (2002) Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: TLFi, Frantext and the software Stella, *LREC-2002*.
- BERNARD P., DENDIEN J., LECOMTE J., PIERREL J.-M. (2002) *Un ensemble de ressources informatisées et intégrées pour l'étude du français : Frantext, TLFi, Dictionnaires de l'Académie et logiciel Stella, présentation et apprentissage de leurs exploitations*, TALN, 2002, Nancy.