Informatisation et valorisation sur le Net : un deuxième vie pour le TLF

Jean-Marie Pierrel

Informatisation et valorisation sur le Net : une deuxième vie pour le TLF

LEXICOGRAPHIE ET INFORMATIQUE : BILAN ET PERSPECTIVES

Colloque international à l'occasion du 50e anniversaire du lancement du projet du *Trésor de la Langue Française*

Dates : 23-25 janvier 2008

Lieu : Nancy, Campus Lettres et Sciences Humaines

Organisation : UMR ATILF (CNRS/Nancy-Université)

Pierrel Jean-Marie (1)
Jean-Marie.Pierrel@atilf.fr

(1) ATILF Nancy Université & CNRS

Mots-clés : dictionnaire électronique, informatisation, valorisation, ressources, mutualisation, accès structuré, , interface, hyper-navigation.

Keywords: electronic dictionary, computerization, valorization, resources, mutualisation, structured access, interface, hyper navigation, XML encoding.

Résumé : Le *Trésor de la Langue Française* (TLF) est un grand dictionnaire de langue française en 16 volumes réalisé par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS) entre le début des années 60 et le milieu des années 90. Ce dictionnaire était initialement conçu pour être édité uniquement sous forme papier. La décision de le transformer en dictionnaire électronique a été prise alors que le dictionnaire était pratiquement achevé. Après avoir rappelé l'importance d'une meilleure valorisation de nos productions de recherche, cet article présente les principales caractéristiques du *Trésor de la Langue Française informatisé* (TLFi), son insertion au sein du portail lexical du Centre National de Ressources Lexicales et Textuelles (CNRTL) et les impacts de ces versions informatisées du TLF sur sa diffusion internationale.

Abstract : The *Trésor de la Langue Française* (TLF) is a large 16 volumes of French language dictionary, released by the Institut National de la Langue Française (INaLF, former laboratory of the CNRS) between the beginning of the 60's and the middle of the 90's. This dictionary was intended to be published in a printed form. The decision to make an electronic version came as the dictionary was almost completed. Having to remind the importance of a better valorization of our research productions, this article presents the main characteristic of the *Trésor de la Langue Française informatisé* (TLFi), its insertion within the lexical common network of the National Center of Lexical and Textual Resources (CNRTL) and the impact of these TLF computerized versions on its international distribution.

Introduction

Dès que l'on s'intéresse à la langue, que cela soit pour un usage strictement humain ou pour une intégration dans une chaîne de traitement automatique, les informations lexicales liées aux mots de la langue occupent une importance primordiale (Laporte, 1997 ; Pierrel 2000). Pourtant, pour le français, il n'existe pas à ce jour de lexique ou de dictionnaire informatique optimal adapté tout à la fois à l'homme et à la machine. Pour les dictionnaires électroniques commerciaux et les lexiques spécifiques développés par telle ou telle équipe, une des questions essentielles porte sur la qualité et la couverture linguistique de tels outils : nombre d'entrées, richesse des informations disponibles, validité linguistique, facilité d'accès, etc. Dans ce contexte, le *Trésor de la Langue Française* (TLF) et sa version informatisée occupent une place de choix

Le *Trésor de la Langue Française* (TLF) est le dictionnaire de langue de référence réalisé entre le début des années 60 et le milieu des années 90 (CNRS, 1976-1994) par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS) dont notre laboratoire ATILF est aujourd'hui le successeur nancéien. Dans son ouvrage sur les dictionnaires de la langue française, Jean Pruvost présente ainsi cet ouvrage : *« Ce projet, qui correspond à une entreprise publique ayant requis une centaine de chercheurs pendant 30 ans, avec un dépouillement de plus de 3 000 textes littéraires, scientifiques et techniques, a bénéficié des compétences nationales et internationales les plus éminentes […] Il en résulte, au-delà de la très grande qualité scientifique des articles, une description du fonctionnement de la langue qui ne manque pas d'être impressionnante : 23 000 pages, 100 000 mots, 450 000 entrées, 500 000 citations précisément identifiées. Le TLF relève pleinement d'une lexicographie philologique et historique, recourant aux citations-attestations qui permettent de fonder toutes les analyses morphologiques et sémantiques »* (Pruvost, 2002, page 78). Robert Martin, quant à lui, écrit « *De nombreux dictionnaires sont aujourd'hui disponibles sous un format électronique – des ouvrages encyclopédiques mais aussi des dictionnaires de langue. Un apport déterminant, en lexicographie française, est l'informatisation du Trésor de la Langue Française (TLF)* » (Martin, 2001 page 61).

L'objectif du présent article est de montrer comment, à travers ses versions CDROM (www.tlfi.fr) et web (www.atilf.fr/tlfi) d'une part, son intégration au sein du portail lexical du Centre National de Ressources Textuelles et Lexicales (www.cnrtl.fr) et ses produits dérivés tel le lexique morpho-syntaxique MORPHALOU (www.atilf.fr/morphalou) d'autre part, le *Trésor de la Langue Française informatisé* (TLFi) a permis de mieux valoriser le TLF et de lui offrir une seconde vie en terme d'usage effectif.

Après un premier paragraphe montrant l'importance d'une meilleure valorisation et mutualisation de nos résultats de recherche, nous rappelerons les principales caractéristiques du TLFi et de ses produits dérivés puis présenterons sa valorisation dans le cadre du portail lexical du CNRTL dont l'objectif est de regrouper le maximum d'informations sur le lexique français. Nous terminerons enfin par une rapide analyse des usages actuels du TLFi qui permit en quelques années de faire du TLF, dictionnaire de référence longtemps considéré comme un dictionnaire d'une élite pour une élite, un des dictionnaires français les plus utilisés sur le Net.

1 De la nécessité d'une meilleure valorisation des productions de recherche sur notre langue

Au-delà des besoins sociétaux de diffusion de connaissances sur notre langue sur lesquels nous reviendrons en fin de cet article, nous nous focalisons dans ce paragraphe sur les impératifs de recherche, primordiaux pour tous nos laboratoires, en particulier en sciences du langage.

1.1 Un enjeu pour la linguistique, la linguistique de corpus et le traitement automatique des langues

Une analyse de l'évolution de la linguistique au cours du dernier demi-siècle montre que sa confrontation avec l'informatique et les mathématiques lui a permis de se définir de nouvelles approches. C'est ainsi qu'au-delà d'une simple linguistique descriptive s'est développée une linguistique formelle, couvrant aussi bien les aspects lexicaux que syntaxiques ou sémantiques, qui tend à proposer des modèles s'appuyant sur une double validation, explicative d'un point de vue linguistique, opératoire d'un point de vue informatique. Par ailleurs, la disponibilité de ressources textuelles électroniques de grande taille (corpus, bases de données textuelles, dictionnaires et lexiques) et les progrès de l'informatique, tant en matière de stockage que de puissance de calcul, ont créé, au cours des années 1990, un véritable engouement pour les approches statistiques et probabilistes sur « corpus » (Habert, 1995). Ainsi se structura petit à petit un nouveau champ de recherche : la linguistique de

corpus (Habert et col., 1997) permettant au linguiste d'aller au-delà de l'accumulation de faits de langue et de confronter ses théories à l'usage effectif de la langue.

Ces études et recherches en TAL et en linguistique de corpus nécessitent de plus en plus l'usage de vastes ressources linguistiques : textes et corpus, si possible annotés, dictionnaires, outils de gestion et d'analyse de ces ressources. Le coût de réalisation de telles ressources justifie pleinement des efforts de mutualisation pour permettre à la communauté de recherche de bénéficier, pour le français, de ressources comparables à celles existant pour d'autres grandes langues tel l'anglais.

1.2 Quelles ressources pour l'étude des langues aujourd'hui ?

1.2.1 Des corpus textuels

Le premier type de ressources, indispensable pour le développement de nombreuses études sur la langue, son analyse et son traitement, concerne les corpus textuels. Leur rôle est en effet central pour permettre la construction de modèles représentatifs de l'usage effectif de la langue. Il s'agit le plus souvent de faire émerger des invariants ou, au contraire, des comportements particuliers d'entités linguistiques. Si, pendant longtemps, ce type d'activité a pu se satisfaire des connaissances intrinsèques sur la langue qu'a le chercheur, les besoins de validation objective du monde scientifique nécessitent de plus en plus le maniement de vastes ensembles d'exemples attestés. La question fondamentale est alors de savoir comment recueillir des données fiables sur l'usage effectif de la langue. Le Web est aujourd'hui une source importante d'extraction de corpus, mais deux travers de taille caractérisent les textes qui y sont disponibles (Pierrel, 2005) :

- leur qualité est souvent très discutable.
- la pérennité de leur disponibilité n'est pas toujours assurée.

La question de la qualité et de la disponibilité de corpus de référence reste donc importante et, pour s'en convaincre, il suffit d'analyser certains projets nationaux ou internationaux. Ainsi, en France, le projet « technolangue »[1] lancé par le Ministère de la Recherche et des Nouvelles Technologies indiquait parmi ses quatre thèmes d'appel à proposition un volet sur les ressources linguistiques dont l'objectif était « *de stimuler la production, la validation et la diffusion de ressources linguistiques pour répondre aux besoins minimaux pour l'étude de la langue française, favoriser la réutilisabilité de ces ressources et diminuer le coût du « ticket d'entrée » dans le secteur* ». Les besoins sont en effet très diversifiés : que ce soit en terme de types de textes (littéraires, scientifiques ou techniques, mono et multilingues) ou en termes d'usages (professionnels ou grand public), la nécessité de vastes corpus normalisés, annotés et validés s'impose.

1.2.2 Des dictionnaires et des lexiques

Le second type de ressources concerne les dictionnaires et les lexiques. Bon nombre des arguments développés ci-dessus peuvent aussi s'appliquer à ce domaine. Or aucun traitement de la langue ne peut se passer du niveau lexical, et la disponibilité de ressources lexicales est absolument indispensable. Là encore les besoins sont très divers dans un contexte mono ou multilingue : dictionnaires spécialisés et dictionnaires généraux de langue, lexiques techniques ou bases terminologiques, par exemple.

Si, une fois de plus, la toile offre des réponses diversifiées à ce besoin, nombre de questions demeurent, concernant tout à la fois la qualité, la richesse, la couverture et la disponibilité de telles ressources. Nous sommes pour notre part convaincu qu'il importe de développer et partager des ressources de ce type et c'est cette conviction qui nous amena à proposer sur le Net une version informatisée du *Trésor de la Langue Française* et d'en dériver un lexique ouvert des formes fléchies du français (540 000 formes issues de 68 000 lemmes : http://www.cnrtl.fr/lexiques/morphalou/).

1.2.3 Des outils d'accès et de traitements

Un troisième type de ressources, complément des deux précédents, concerne les outils d'accès et de traitement de ces ressources. Deux types d'outils méritent une attention toute particulière :

- Les outils de gestion et d'exploitation des ressources textuelles, lexicales ou dictionnairiques. Que seraient en effet des ressources textuelles ou dictionnairiques du type de celles envisagées ci-dessus sans les logiciels d'exploration de ces ressources ?
- Les outils de base indispensables pour permettre à une équipe de recherche de proposer des avancées sur tel ou tel point : lemmatisation, conjugaison ou étiquetage morphosyntaxique.

[1] http://www.recherche.gouv.fr/appel/2002/technolangue.htm.

Une fois de plus on ne peut que noter, tout en le regrettant, le manque de disponibilité d'outils fiables et généraux de ce type. Faute de cette disponibilité, la première tâche d'une équipe de recherche ou de développement travaillant sur des ressources linguistiques et plus généralement sur la langue consiste souvent, aujourd'hui, à redévelopper de tels outils !

1.3 Une nécessité : mutualiser les ressources et mieux prendre en compte leur production dans l'évaluation des chercheurs

En conclusion de ce paragraphe introductif, nous souhaitons faire partager notre conviction de la nécessité de mutualiser, au sein de la communauté francophone des sciences du langage, des ressources de références (corpus textuels, dictionnaires et lexiques, outils d'exploitation de ces ressources) pour la construction de modèles ou outils linguistiques, leur validation et leur comparaison.

Le coût de définition et de production de vastes ressources linguistiques de qualité (corpus, dictionnaires et lexiques) est important et c'est un gâchis énorme de vouloir, pour chaque projet, redéfinir l'ensemble des ressources dont on a besoin. A titre d'exemple, la construction d'un dictionnaire de langue tel le *Trésor de la Langue française* a nécessité près de cent personnes durant trente ans, et l'établissement d'une base de données textuelle tel FRANTEXT (www.atilf.fr/frantext) s'est chiffré aussi en dizaines d'hommes-an. Sans vouloir plaider ici pour une rentabilisation extrême de la recherche à travers une taylorisation de notre domaine, il convient néanmoins de prendre conscience que, sans une véritable mutualisation de telles ressources dans un domaine aussi vaste que les sciences du langage qui nécessite d'aborder des aspects aussi divers que le lexique, la syntaxe, la sémantique, la pragmatique, chaque équipe de recherche ou chaque chercheur se verrait dans l'obligation de tout réinventer, alors même que nul ne peut être spécialiste de chacun de ces sous-domaines.

Un second point plaidant pour la mutualisation des ressources concerne l'évaluation, de plus en plus indispensable, de nos productions de recherche (analyseurs, systèmes de traitement) qui nécessite, pour des besoins de comparaison, la disponibilité de ressources de référence (corpus textuels, corpus d'exemples sur un phénomène de langue, ressources dictionnairiques) accessibles, partagées et clairement identifiables.

Enfin, il convient de noter qu'en termes de valorisation de la recherche et de partage de connaissances avec nos concitoyens, une disponibilité accrue, en particulier sur le Web, de nos productions de recherche est indispensable. Outre le fait que cela peut permettre un meilleur partage entre le monde de la recherche et la société civile, cela répond aussi à un besoin de plus en plus grand de connaissances chez nos concitoyens.

Mais ne nous leurrons pas, la constitution et la valorisation de telles ressources de qualité nécessitent des investissements en temps importants. Si l'on souhaite que des chercheurs puissent consacrer une partie de leur temps à de telles tâches au service de l'ensemble de la communauté scientifique, il convient de mieux prendre en compte cette activité de production de ressources numériques dans leur évaluation et de mettre en place une structure servant à la fois de validation et de diffusion de ces productions. C'est en partie du moins le rôle que le CNRS a confié aux Centres Nationaux de Ressources, dont le CNRTL.

2 Le TLF, une ressource inestimable valorisée à travers le TLFi

2.1 Caractéristiques du TLFi

Reflet fidèle de la version papier, jusque dans sa présentation typographique à l'écran, le TLFi (www.atilf.fr/tlfi) se caractérise, comme le TLF, par la richesse de son matériau et la complexité de sa structure :

- Importance de sa nomenclature : 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 plus de deux siècles de production littéraire française.
- 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 (Dendien et Pierrel, 2003) 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, 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.

Nous ne reviendrons pas ici sur les étapes d'informatisation du TLF traitées par ailleurs (Dendien et Pierrel 2003), mais nous nous contenterons ici de rappeler, essentiellement par l'exemple, les accès offerts par la version informatisée du TLF.

2.2 Quels accès au TLFi ?

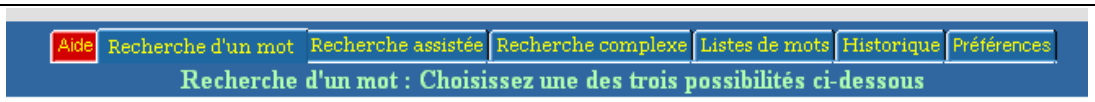
Le TLFi correspond à une rétro-conversion de la version papier du TLF pour laquelle, par des procédures de repérage semi-automatique des objets textuels composant les articles du dictionnaire original, nous avons introduit un balisage fin, tant typographique (de manière à conserver une image 100 % fidèle 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 92 997 entrées et 64 346 locutions faisant l'objet de 271 166 définitions et illustrées par 427 493 exemples.

C'est ce balisage fin du TLF et l'exploitation du document XML correspondant qui nous permet de proposer des accès 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.

2.3 Exemples de recherches dans le TLFi

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations sur les modes de recherche offerts dans le TLFi, mais la meilleure façon de se rendre compte de l'intérêt d'une telle transformation du TLF en document numérique consiste soit à accéder au Cédérom du TLFi (ATILF, 2004), soit à se connecter directement à l'adresse : www.atilf.fr/tlfi. Trois principaux types d'accès sont alors proposés : la recherche d'un mot, la recherche assistée et la recherche complexe.



2.3.1 Recherche d'un mot ou d'une expression

Cette recherche permet un 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 ;* de même un accès à partir de la forme *sussiez* permet d'obtenir automatiquement l'article *savoir*. Elle donne aussi la possibilité d'obtention directe des définitions et conditions d'usage d'expression tel « le trompette » en focalisant la réponse sur l'élément pertinent demandé et en offrant la possibilité, à l'aide d'une sorte de « stabilo boss » électronique, de surligner tel ou tel objet textuel. Ici, par exemple, la définition :

Objets de la recherche : 1 **Paragraphe** 1

H **TROMPETTE, subst.**

II. — *Subst. masc.* Personne qui joue de la trompette.

1 **A.** — Soldat chargé d'exécuter les sonneries. ***Le trompette*** *de l'escadron, d'un régiment de cavalerie. Tu seras capitaine, avec une nuée de trompettes courant et sonnant devant toi* (HUGO, *Légende*, t. 3, 1877, p. 390). 1

— *Loc. fam., vieilli. Il est bon cheval de trompette.* Il ne se laisse ni effrayer, ni intimider. *Son air, un air de bon cheval de trompette qui ne craignait pas le bruit* (A. DAUDET, *Tartarin de T.*, 1872, p. 13).

B. — Musicien jouant dans une fanfare, un orchestre. Synon. *trompettiste* (*infra* dér.). *Le trompette noir du dancing* (BEAUVOIR, *Mandarins*, 1954, p. 306).

noir du dancing (BEAUVOIR, *Mandarins*, 1954, p. 306).

2.3.2 Recherche assistée

Ce second type d'accès permet par exemple de rechercher des expressions composées d'une forme : ainsi, en demandant les mots contenant la forme *queue* on obtient 35 réponses dont :

COURTE-QUEUE, adj. et subst.
DEMI-QUEUE, subst. fém.
HOCHEQUEUE, HOCHE-QUEUE, subst. masc.
PAILLE-EN-CUL, PAILLE-EN-QUEUE, subst. masc.
PORTE-QUEUE, subst. masc.
QUEUE(-)D'ARONDE, voir ARONDE.
Etc.

ou de 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 : [code grammatical : *verbe* ; domaine : *marine* ; type d'objet : *définition*, contenu : &m*voile*[2]]. Voici un extrait des 61 réponses que l'on obtient :

ABRIER, ABREYER, verbe trans.
3Empêcher 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. : `` (WILL. 1831) :3
AMURER, verbe.
3Fixer l'amure d'une voile pour l'orienter selon le vent :3
ETC……

Autre exemple : pour l'ensemble des mots dont la définition utilise le mot *liberté* [type d'objet : définition, contenu : &mliberté], on obtient 306 réponses dont :

Objets de la recherche : 1**Définition**1

ABUSER, verbe trans.
1Exagérer 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.
1Grande liberté.1
ALIÉNANT, ANTE, part. prés. et adj.
1Qui prive l'homme de son humanité, de sa liberté :1
Etc…..

2.3.3 Recherche complexe

Les interrogations possibles au sein de ce dictionnaire peuvent prendre des formes encore plus complexes. Ainsi, il est possible de répondre à la requête suivante : « *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 pour cela d'utiliser l'onglet « recherche complexe » et de 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

[2] &ms*ubs* permet de tester toutes les formes d'un *substantif*, de même que &c*verbe* toutes les formes d'un *verbe.*

"*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, 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, et le lien "dépendant de l'objet 1" de l'objet 6 exprime que l'objet est dans l'article dont l'entrée est l'objet 1.
Une telle interrogation nous fournit quatre résultats dont :

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

MORTIFICATION, subst. fém.
1 MORTIFICATION, subst. fém. 1 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 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
ETC.

3 Le portail lexical du CNRTL : un outil de mutualisation de résultats en lexicographie

3.1 Objectifs du CNRTL

Créé par le CNRS en 2005, le Centre National de Ressources Textuelles et Lexicales (CNRTL : www.cnrtl.fr) est adossé au laboratoire Analyse et Traitement Informatique de la Langue Française (ATILF / CNRS - Nancy Université). Son objectif est de réunir au sein d'un portail unique le maximum de ressources informatisées et d'outils de consultation pour l'étude, la connaissance et la diffusion de la langue française.

Grâce à une mutualisation de connaissances issues des travaux de différents laboratoires, le CNRTL se propose d'optimiser la production, la validation, l'harmonisation, la diffusion et le partage de ressources, qu'il s'agisse de données textuelles et lexicales informatisées ou d'outils permettant un accès intelligent à leur contenu.

La décision de création du CNRTL s'inscrit dans la politique du CNRS visant à la création de nouvelles infrastructures indispensables aux travaux de recherche menés par l'ensemble de la communauté scientifique et résulte d'une action commune à la Direction de l'Information Scientifique et au Département Homme et Société du CNRS. Il est aujourd'hui l'une des composantes du très grand équipement d'accès unique aux documents numériques en sciences humaines et sociales ADONIS.

L'expertise scientifique reconnue ainsi que les nombreux projets coopératifs nationaux et internationaux des laboratoires auxquels il est adossé ont permis en outre au CNRTL de se positionner au niveau européen à travers :

- Des collaborations directes avec des centres partenaires, en Grande-Bretagne (Université d'Oxford), en Allemagne (Centres de compétence de Trèves et de Würzburg, DFKI à Sarrebruck, MPI) et aux Pays-Bas (Université de Nimègue).
- La participation au réseau européen CLARIN (http://www.mpi.nl/clarin/) des centres de gestion de ressources et de technologies linguistiques.

3.2 Les ressources gérées au sein du CNRTL

Le CNRTL se structure autour de cinq pôles de compétence : un portail lexical sur le français ; des corpus et données textuelles, annotés ou non ; des dictionnaires encyclopédiques et linguistiques (anciens et modernes) ; des lexiques phonétiques, morphologiques, syntaxiques, sémantiques ; des outils linguistiques (étiqueteurs, analyseurs, aligneurs, concordanciers, outils d'annotation). Parmi les ressources déjà intégrées au CNRTL, outre le portail lexical sur lequel nous allons revenir dans le paragraphe suivant, il convient de noter, entre autres :

- Les corpus de textes libres de droit d'auteur et d'éditeur (dans un premier temps 500 textes issus de Frantext) : à travers une sélection par auteurs, titres, dates ou genres, nous offrons la possibilité de

télécharger les textes sélectionnés au format XML dans une DTD respectant les recommandations de la TEI (www.tei-c.org)[3], l'utilisateur récupèrant une archive contenant la DTD et le codage XML/TEI des textes. A notre connaissance, le CNRTL est le premier site offrant un ensemble de corpus français normalisés XML/TEI d'environ 150 millions de caractères.

- Le lexique Morphalou, dérivé de la nomenclature du TLF en accès libre tant en consultation qu'en téléchargement : lexique ouvert des formes fléchies du français qui fournit 524 725 formes fléchies, appartenant à 95 810 lemmes, linguistiquement valides (responsabilité d'un comité éditorial) et respectant les propositions de normalisation pour les ressources lexicales de l'ISO (TC37/SC4).
- Des versions informatisées de dictionnaires tant modernes (TLFi ; Dictionnaire de l'Académie française : 8ème et 9ème éditions) qu'anciens (Dictionnaires de R. Estienne (1552), de Jean Nicot (1606), de Bayle (1740), de Ferraud (1787-1788), de l'Académie (1ère édition, 1694 ; 4ème édition, 1762 ; 5ème édition, 1798 ; 6ème édition, 1835)), ainsi que de l'Encyclopédie de Diderot et d'Alembert[4].

Le CNRTL se propose également de mettre à disposition de la communauté des outils linguistiques utilisables directement sur le site Web à partir d'un simple navigateur Internet. Parmi les différents projets en cours ou à venir, nous comptons offrir aux utilisateurs un accès simple et convivial à des outils comme :

- FLEMM : outil d'analyse flexionnelle de textes en français au préalable étiquetés, au moyen de l'un des deux catégorisateurs : Brill ou TreeTagger.
- POMPAMO : outil de détection de candidats à la néologie formelle et catégorielle basé sur l'utilisation de lexiques d'exclusion. Ce projet exploite des ressources lexicales comme Morphalou et permet d'en constituer de nouvelles.

3.3 Un exemple d'intégration de données lexicographiques : le portail lexical

Le portail lexical a pour vocation de valoriser et de partager, en priorité avec la communauté scientifique, un ensemble de données issues des travaux de recherche sur le lexique français. Projet évolutif, cette base de connaissances lexicales propose, à partir d'une forme lexicale, d'intégrer un maximum de connaissances disponibles.

Informations lexicographiques

Au premier rang de ces connaissances se placent les informations lexicographiques. A ce jour nous avons intégré dans ce portail les informations issues du TLF (www.atilf.fr/tlfi) qui apparaissent par défaut lorsqu'on demande des informations lexicographiques. Elles sont complétées par des informations facilement accessibles via un menu et issues :

- du dictionnaire de l'Académie Française (4ème, 8ème et 9ème éditions) (www.atilf.fr/academie) dont l'informatisation a été réalisée au sein du laboratoire dans le cadre d'un partenariat avec l'Académie.
- de la Base de Données Lexicographiques Panfrancophone (BDLP : http://www.tlfq.ulaval.ca/bdlp/), projet d'envergure internationale qui s'inscrit dans l'entreprise du Trésor des vocabulaires français lancée par le professeur Bernard Quemada dans les années 1980. L'objectif de la BDLP est de constituer et de regrouper des bases représentatives du français de chacun des pays et de chacune des régions de la francophonie. Les bases de données sont conçues de façon à pouvoir être interrogées de façon séparée ou comme un seul corpus et à servir de complément au *Trésor de la Langue Française informatisé*. Dans sa dimension internationale, le projet de la BDLP est patronné par l'Agence Universitaire de la Francophonie qui l'appuie à travers son réseau d'étude du français en francophonie (http://www.eff.auf.org/).
- De la Base Historique du Vocabulaire Français (Datations et Documents Lexicographiques : www.atilf.fr/ddl) constituée de datations du vocabulaire français, s'appuyant sur des données des 48 volumes de la collection *Datations et Documents Lexicographiques*.

L'ensemble de ces informations lexicographiques sont également accessibles directement, pour une forme donnée, par *http://www.cnrtl.fr/lexicographie/suivi de la forme que l'on souhaite interroge.* Ainsi : http://www.cnrtl.fr/lexicographie/aguerrit permet d'accéder aux informations lexicographiques du verbe *aguerrir*.

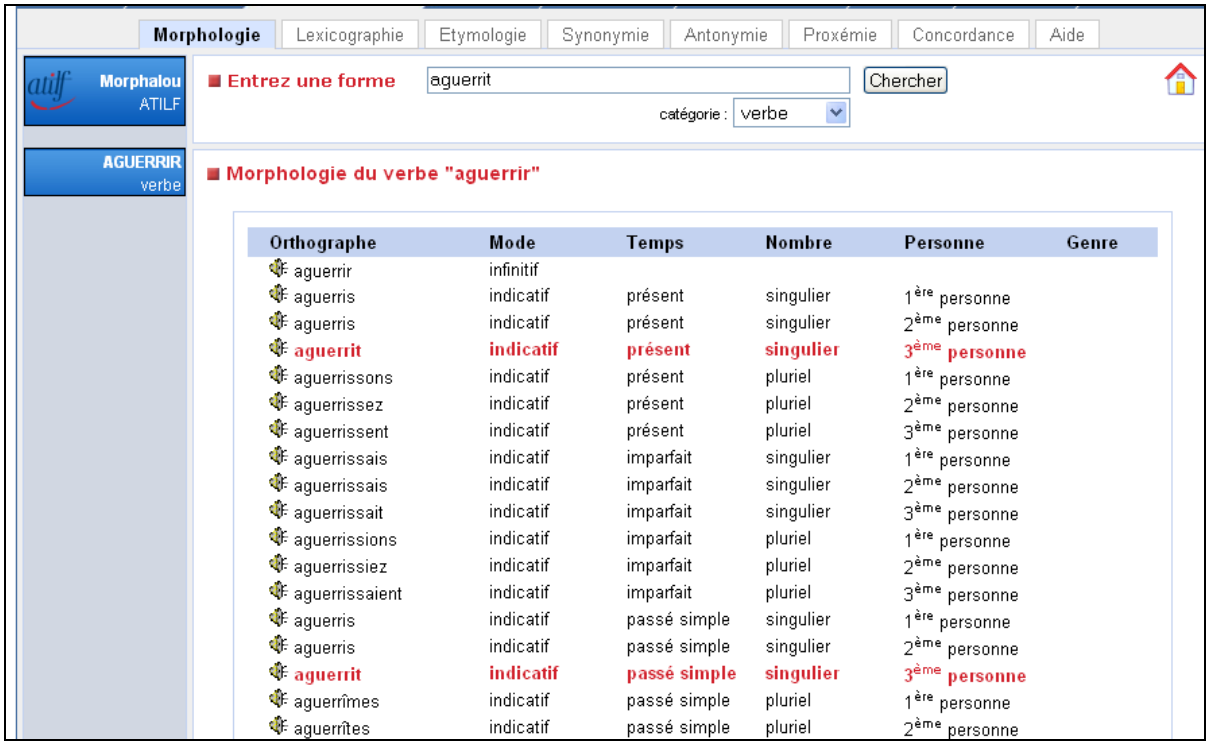
---

[3] Notons que Nancy, à travers une association entre l'ATILF, l'INIST et le LORIA, est aujourd'hui centre support européen de la TEI.

[4] La plupart des versions informatisées de dictionnaires anciens, tout comme celle de l'Encyclopédie de Diderot et d'Alembert, sont le fruit d'un partenariat avec l'ARTLF (http://humanities.uchicago.edu/orgs/ARTFL/).

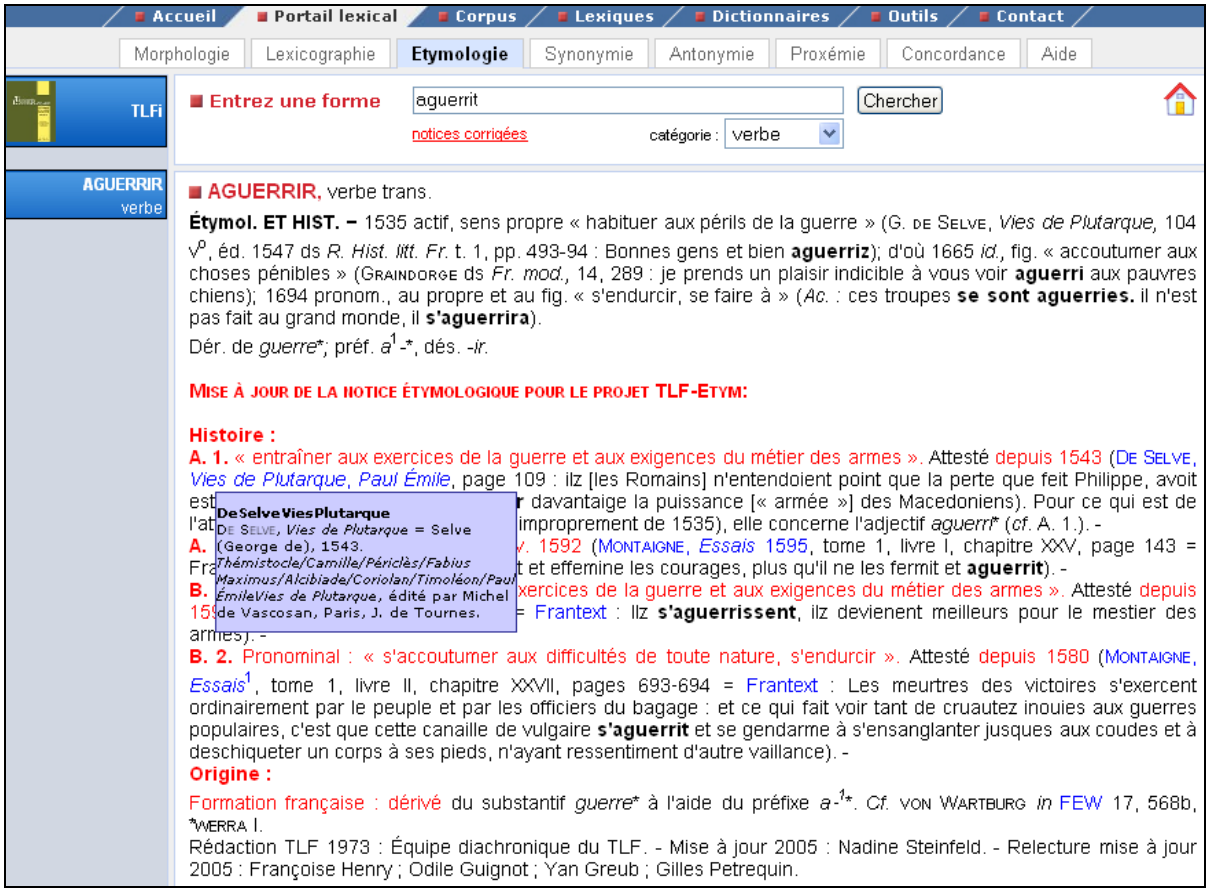
Informations morphosyntaxiques

Ces informations morphologiques sont issues de la base Morphalou (www.atilf.fr/morphalou), construite au départ à partir de la nomenclature du TLFi. Elles sont aussi accessibles directement pour la forme, telle *aguerrit*, par : http://www.cnrtl.fr/morphologie/*aguerrit*



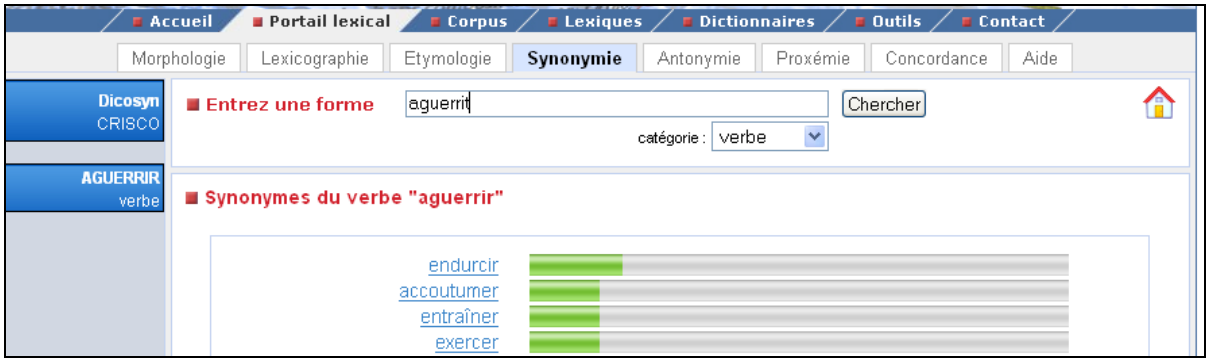
Informations étymologiques

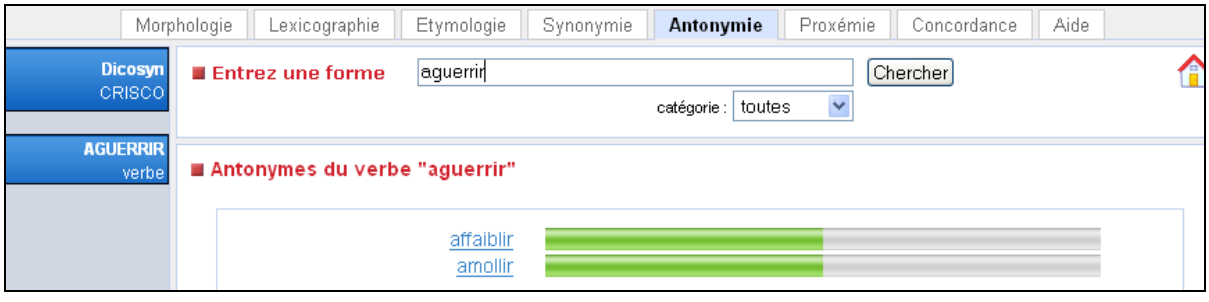
Ces informations étymologiques sont issues du TLF (www.atilf.fr/tlfi) et du projet TLF-Etym de mise à jour des rubriques étymologiques du TLF (www.atilf.fr/tlf-etym). Elles sont accessibles directement, pour une forme, par : http://www.cnrtl.fr/etymologie/*aguerrit*



Synonymies et antonymies

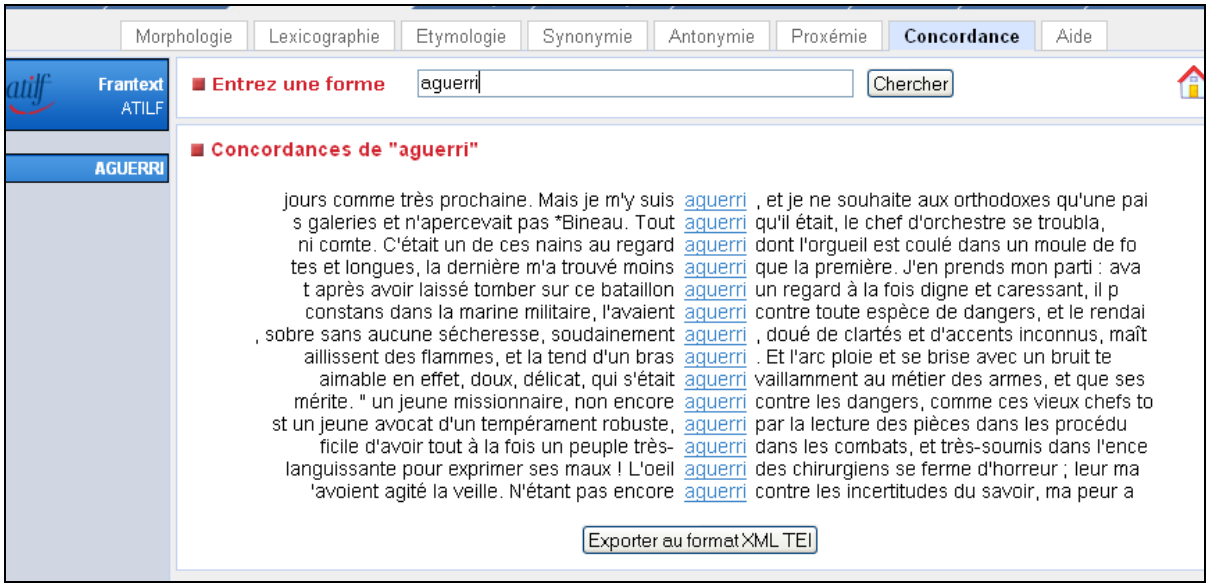
Ces informations de synonymie et d'antonymie proviennent du dictionnaire de synonymes de Caen (http://www.crisco.unicaen.fr/), construit à partir de données issues de l'INaLF. Ces informations sont aussi directement accessibles par : http://www.cnrtl.fr/synonymie/*aguerrit* ou http://www.cnrtl.fr/antonymie/*aguerrit*





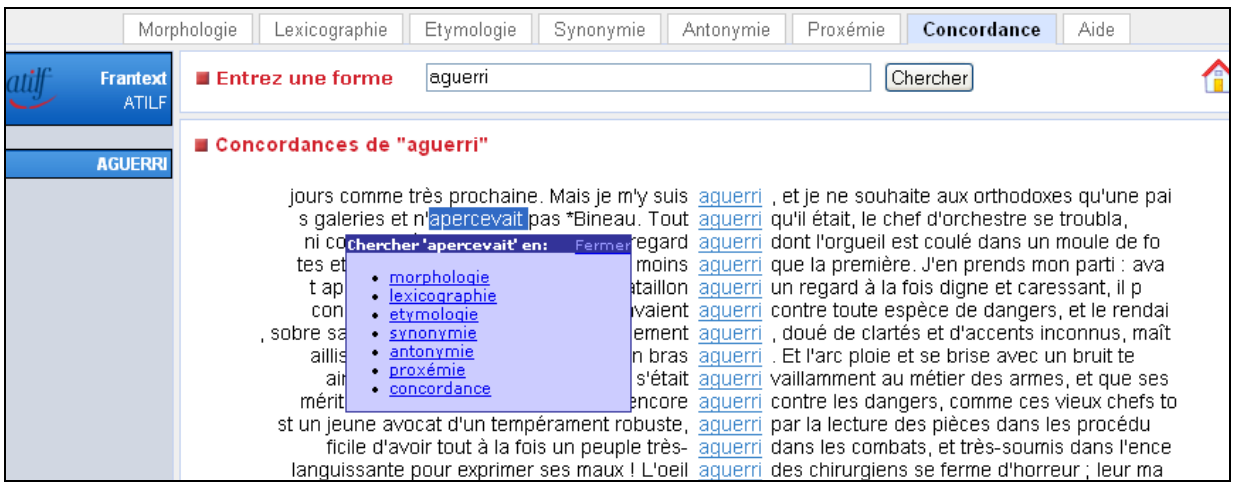
Concordance

Cette concordance utilise le corpus des textes de la base Frantext (www.atilf.fr/frantext) qui offre aussi la possibilité d'exporter les résultats du concordancier au format XML/TEI. C'est à notre connaissance le seul site permettant à un utilisateur d'exporter dans un format normalisé un concordancier français d'une telle importance. Ces concordances sont aussi directement accessibles par : http://www.cnrtl.fr/concordance/*aguerri*



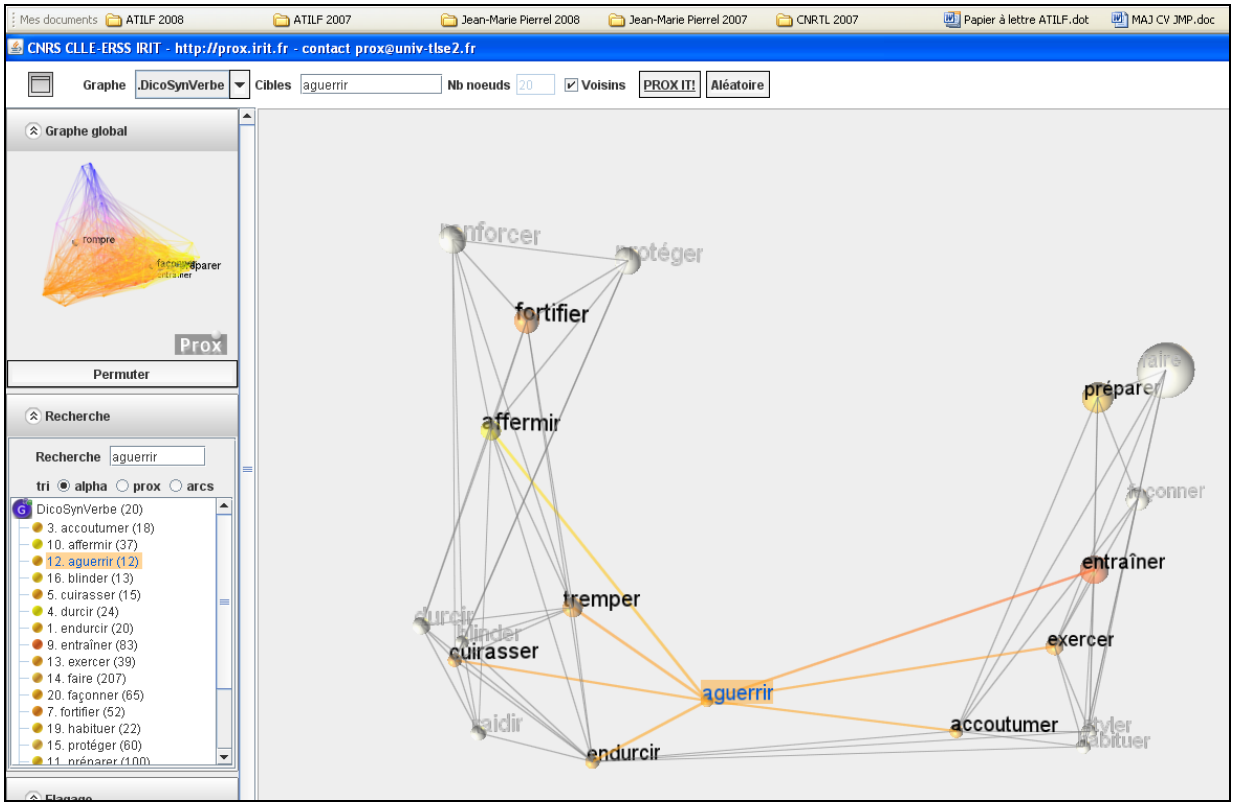
De plus, un simple clic droit sur un des exemples permet d'obtenir la référence complète de l'exemple sélectionné. Ainsi pour le premier exemple :

Le portail lexical permet également, à partir d'un simple double clic sur un mot, une hyper-navigation vers toutes les informations lexicales disponibles pour ce mot. Par exemple, si l'on veut obtenir des informations sur la forme « apercevait » du deuxième exemple de concordance, un double clic sur le mot affiche un menu qui permet d'hyper-naviguer vers l'ensemble des informations disponibles sur cette forme :



Proxémie

Une représentation en trois dimensions de la proxémie des mots de langues réalisée en coopération entre l'IRIT et l'ERSS (http://Prox.irit.fr) (Gaume 2006) est accessible elle aussi directement par : http://www.cnrtl.fr/proxemie/aguerrir.



4 Impact de la valorisation de ressources lexicales sur le web : une seconde vie pour le TLF

Sans aucun doute le plus grand dictionnaire informatisé consacré à la langue française, le TLFi, grâce à la richesse de son contenu entièrement encodé en XML, a ouvert des perspectives intéressantes. Le TLF a eu pendant longtemps la réputation tenace d'être un dictionnaire réservé à une élite. Cette perception du TLF pouvait s'expliquer par au moins trois caractéristiques de sa version papier :

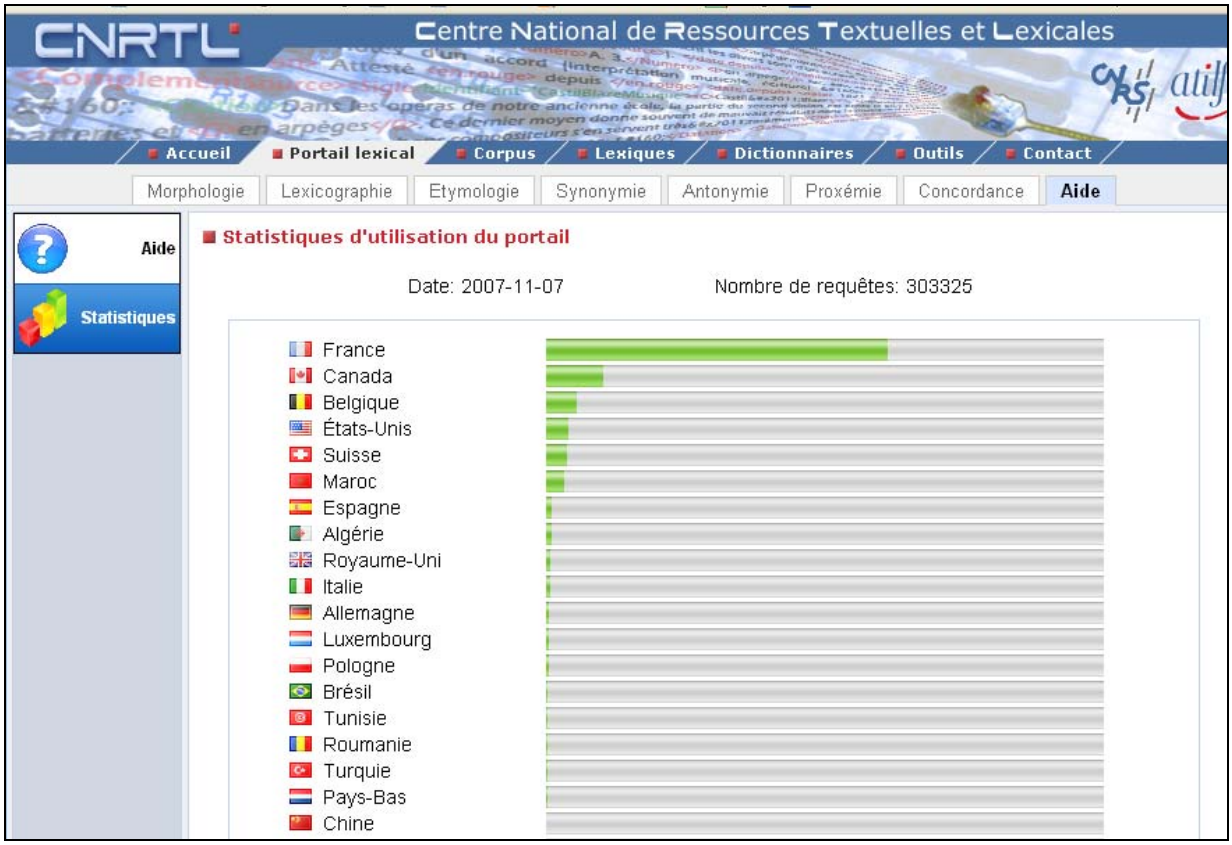
- Sa taille : 16 volumes de plus de 1 000 pages chacun.
- Sa richesse de description qui parfois nuisait à sa lecture, au moins pour les articles les plus lourds : l'article « aimer » se développe ainsi sur 12 pages soit 24 colonnes, et il n'est pas toujours aisé pour un non-spécialiste d'appréhender cette information très riche.
- Son coût, environ 1 500 euros, qui ne le rendait pas facilement accessible à tous.

S'il a su se positionner comme une référence en lexicographie française, la diffusion de sa version papier s'est néanmoins limitée à quelques millers d'exemplaires au sein d'une intelligentsia somme toute limitée.

Sa version informatique sous forme de Cédérom (environ 15 000 exemplaires vendus en moins de 4 ans) ou de ressources librement accessibles sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Sa version web fait l'objet d'environ 300 000 connexions quotidiennes en provenance de tous les continents, et il est référencé par d'innombrables sources. La notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française.

Son intégration plus récente encore au sein du portail lexical du CNRTL et ses interconnexions avec d'autres types de ressources sur le vocabulaire français le positionnent 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 devenue largement injustifiée et sa diffusion au sein du portail lexical du CNRTL, qui fait l'objet d'environ

300 000 requêtes par jour provenant d'horizons très divers, en fait aujourd'hui l'un des dictionnaires les plus exploités sur le Web



Notons pour conclure que le partage d'une telle version informatisée d'une production scientifique de référence offre aujourd'hui des modes nouveaux de valorisation de ressources ou de résultats de recherche. Au-delà du seul monde universitaire, ces techniques permettent de mettre à disposition de l'ensemble de la société nos résultats de recherche. On peut, pour s'en convaincre, analyser les commentaires apparaissant sur le web dans des sites institutionnels (http://www.terminometro.info/article.php?ln=fr&lng=fr&id=4546 par exemple) ou professionnels (http://www.entreprisesaletranger.org/archive-01-05-2007.html). La généralisation de telles exploitations et valorisations de versions électroniques est ainsi en train de modifier notablement les modes de travail et d'échanges scientifiques au sein des communautés de recherche SHS.

Bibliographie

ATILF *Trésor de la Langue Française informatisé*, CNRS Editions, Livre d'accompagnement 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, 2004, Version Mac OX X, ISBN 2-271- 06365-5, 2005.

CNRS, *TLF, Dictionnaire de la langue du 19e et 20e siècle*, CNRS, Gallimard, Paris, 1976-1994.

Dendien J., Pierrel J.M. Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, *TAL* Vol 44 – n° 2/2003, Hermès Sciences Edition, p. 11-37.

Gaume B., Cartographier la forme du sens dans les petits mondes lexicaux, in *JADT* 2006, p 541-465.

Habert B., Traitements probabilistes et Corpus, *TAL* Vol 36- N°1-2, Paris, Hermès, 1995.

Habert B., Nazarenko A. et Salem A. *Les linguistiques de corpus*, Paris, Armand Colin, 1997.

Laporte E., Mots et niveau lexical, TAL, vol. 38, n°2, 1997.

Martin R., *Sémantique et automate*, PUF, Paris 2001.

Pierrel J.M., *Ingénierie des langues*, Hermès Editions, 2000, 354 p.

Pierrel J.M., Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella , *Revue Québécoise de Linguistique*, volume 32/1, TAL Web et Corpus, p. 155-176, 2005.

Pruvost J., *Les dictionnaires de la langue française*, collection Que Sais-je ?, PUF, Paris 2002.