

TALN 2007, Toulouse, 12-15 juin 2007

Le CNRTL, Centre National de Ressources Textuelles et Lexicales, un outil de mutualisation de ressources linguistiques

Jean-Marie Pierrel, Etienne Petitjean

CNRTL/ATILF CNRS – Nancy Université

44 avenue de la libération

BP 30687

54063 Nancy Cedex

Jean-Marie.Pierrel@atilf.fr ; Etienne.Petitjean@atilf.fr

Résumé Créé en 2005 à l'initiative du Centre National de la Recherche Scientifique, le CNRTL propose une plate-forme unifiée pour l'accès aux ressources et documents électroniques destinés à l'étude et l'analyse de la langue française. Les services du CNRTL comprennent le recensement, la documentation (métadonnées), la normalisation, l'archivage, l'enrichissement et la diffusion des ressources. La pérennité du service et des données est garantie par le soutien institutionnel du CNRS, l'adossement à un laboratoire de recherche en linguistique et informatique du CNRS et de Nancy Université (ATILF – Analyse et Traitement Informatique de la Langue Française), ainsi que l'intégration dans le réseau européen CLARIN (common language resources and technology infrastructure european).

Abstract Founded in 2005 under the auspices of the French National Centre for Scientific Research (CNRS), the CNRTL offers a unified platform to access electronic resources and documents for linguistic research on the French language. Provided services include identification, documentation (metadata), standardisation, archiving, enrichment and distribution of resources. The sustainability of services and data is ensured through the CNRS institutional support, the hosting by a public research institute in linguistics and NLP of CNRS and Nancy University (ATILF – Analyse et Traitement Informatique de la Langue Française), and integration into the common language resources and technology infrastructure european project (CLARIN).

Mots-clés : Centre de ressources, lexiques, dictionnaires, corpus, Tools

Keywords: Resource Centre, Lexicons, Dictionary, Corpora, Tools

1 Les missions du CNRTL

Les missions du CNRTL (www.cnrtl.fr) mis en place par le CNRS, Département Sciences de l'Homme et de la Société et Direction de l'Information Scientifique, au sein de l'ATILF peuvent se résumer en sept points, repris de la lettre de mission transmise lors de la création du centre :

- « Entrées » : acceptation, contrôle et validation des ressources, tant d'un point de vue scientifique que technique, afin d'assurer la qualité des ressources (corpus dictionnaires, lexiques et outils de traitement) offertes par le centre ;
- « Stockage » : stockage, maintenance et récupération des ressources. Beaucoup de chercheurs et d'équipes en SHS qui développent pour leurs recherches propres des ressources informatisées ne disposent en effet pas des moyens nécessaires pour assurer cette fonction ;
- « Gestion des ressources » : partage, conservation et enrichissement de ressources, afin d'assurer une réelle mutualisation entre équipes de recherche ;
- « Administration » : administration des ressources et aide aux utilisateurs ;
- « Pérennisation et documentation » : mise à jour et évolution des supports informatiques. L'évolution des matériels et logiciels informatiques nécessite une maintenance régulière de telles ressources informatisées pour éviter des gâchis que nous avons pu connaître dans le passé où certains corpus ont été perdus par manque de maintenance et de pérennisation ;
- « Accès » : aide et réponse aux utilisateurs permettant aux non spécialistes de l'informatique que sont les chercheurs en SHS d'accéder et d'exploiter au mieux de telles ressources informatisées à travers des outils adaptés à leurs besoins ;
- « Formation » : formation des producteurs et utilisateurs aux méthodologies d'annotation, de codage et de normalisation. Sur ce point fort du fait que Nancy est centre support de la TEI, on s'appuie autant que faire se peut sur les recommandations de la TEI.

2 Le CNRTL nœud d'un réseau international européen

Au-delà de sa seule mission nationale, le CNRTL participe au réseau européen CLARIN (Common Language Resource and Technologie Infrastructure : <http://www.mpi.nl/clarin>) des centres de gestion de ressources linguistiques qui correspond à l'une des propositions européennes d'infrastructure de recherche en SHS. Menée en étroite interaction avec la proposition d'une infrastructure européenne de gestion de données numériques en SHS (DARIAH - Digital Research Infrastructure for the Arts and Humanities), cette proposition, est incluse dans la feuille de route ESFRI qui définit les infrastructures de recherches à soutenir dans le cadre du 7ème programme-cadre. Elle vise à définir une infrastructure européenne partagée par les grands centres de recherche européens et s'appuyant sur des centres régionaux « certifiés » dans leurs domaines respectifs.

Ce projet est également l'occasion d'organiser une réflexion commune sur la gestion d'une plate-forme ouverte de gestion et d'archivage de documents numériques avec nos collègues du Max Planck Institute qui travaillent actuellement sur le même sujet. Dans l'idéal, cette collaboration permettra de converger vers une plate-forme logicielle unique utilisable par le

CNRTL comme par le MPI. Cette plate-forme logicielle pourrait s'articuler autour de Fedora (<http://www.fedora.info/>) qui est un projet open-source offrant une architecture flexible pour la gestion et la distribution de documents numériques. Développé conjointement par l'université de Virginie et l'université de Cornell, ce système semble offrir les bases dont nous avons besoin pour développer cette plate-forme, à savoir :

- Le dépôt de ressources : permettre à un utilisateur de pouvoir soumettre une ou plusieurs ressources numériques (texte brut ou étiqueté morpho-syntaxiquement, etc.)
- La consultation des ressources : offrir aux utilisateurs une interface de consultation permettant la navigation et la sélection des différents corpus et ressources disponibles sur la plateforme.
- Le téléchargement des ressources : faciliter le téléchargement des ressources sélectionnées dans le format de sortie souhaité par les utilisateurs (XML, PDF, Word, HTML, etc.)

3 Les ressources accessibles au sein du CNRTL

Le CNRTL s'est structuré 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, concordances, outils d'annotation).

Afin de proposer une première offre de ressources au sein du CNRTL, nous avons travaillé dans un premier temps sur la base des ressources linguistiques informatisées actuellement disponibles à Nancy, ressources qui, suivant les cas, sont des ressources libres et téléchargeables après acceptation d'une licence de type ressources libres, des ressources sous droits accessibles uniquement via une interface web spécifique, ressources sous droits accessibles uniquement dans le cadre d'une convention de partenariat avec les ayants droits. Parmi les ressources déjà intégrées au CNRTL, outre les outils et le portail lexical sur lesquels nous allons revenir dans les paragraphes suivants, il convient de noter :

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 : l'utilisateur récupère une archive contenant la DTD et le codage XML/TEI des textes (à notre connaissance, le CNRTL est le premier site offrant un ensemble de corpus français téléchargeables et normalisés XML/TEI d'environ 150 millions de caractères) ; et un corpus annoté pour le traitement des DEscriptions DEfinies (DEDE : coopération LORIA, Metadif et ATILF).

Le lexique Morphalou 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 dictionnaires tant modernes qu'anciens : outre l'accès à la version électronique du Trésor de la Langue Française, dictionnaire de référence des 19e et 20e siècles, produit par le laboratoire ATILF, un ensemble de liens permet également la consultation des dictionnaires suivants :

- le Dictionarium latinogallicum (troisième édition - 1552) de Robert Estienne
- le Thresor de la langue françoise, tant ancienne que moderne de Jean Nicot (Paris, David Douceur, 1606)
- le Dictionnaire historique et critique de Bayle (fac-similé de la version de 1740)
- le Dictionnaire critique de la langue française de Jean-François Féraud (1787-1788)
- le Dictionnaire de l'Académie française (1^e édition 1694 - 4^e édition 1762 - 5^e édition 1798 - 6^e édition 1835 - 8^e édition 1932/1935 – 9^e édition en cours)

4 Des outils à disposition de la communauté

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 qui ont été au préalable étiquetés, au moyen de l'un des deux catégorisateurs : Brill ou TreeTagger.
- DERIF : outil d'analyse morpho-sémantique du français qui s'applique à des entrées lexicales catégorisées issues d'un dictionnaire de la langue générale, capable de traiter des mots hors-dictionnaire et dont les résultats associent la morphologie et la sémantique
- 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.

5 Un exemple d'intégration de ressources : le portail lexical

Le portail lexical, quant à lui, 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 exploite aujourd'hui divers documents numériques pour fournir, à partir d'une forme lexicale, six types d'informations importantes : des informations morphologiques issues de Morphalou (www.atilf.fr/morphalou), des informations lexicographiques et étymologiques issues des projets TLF (www.atilf.fr/tlfi) et TLF-Etym, des informations de synonymies à travers l'intégration du dictionnaire de synonymes de Caen (<http://www.crisco.unicaen.fr/>), une concordance utilisant le corpus des textes de la base Frantext (www.atilf.fr/frantext) et une présentation des résultats de proxémie du projet Prox de l'ERSS (<http://w3.univ-tlse2.fr/erss/>). Il 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 informations sont directement intégrables dans d'autres applications Web à travers des liens spécifiques à chacune des formes type d'informations tels que : www.cnrtl.fr/concordance/mot. De plus, un simple clic sur un des exemples permet d'obtenir la référence complète de l'exemple sélectionné. 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 un mot d'un exemple de concordance, un double-clic sur le mot affiche un menu qui permet d'hyper-naviguer vers les informations lexicales de ce mot.