Ressources linguistiques informatisées de l'ATILF

Pascale Bernard, Charles Bernet, Jacques Dendien, Jean-Marie Pierrel,
Gilles Souvay, Zina Tucsnak

ATILF " Analyses et Traitements Informatiques du Lexique Français "
UMR CNRS Université Nancy 2
44, avenue de la Libération, BP 30687
54063 Nancy cedex
contact@inalf.fr http://www.inalf.fr/atilf

Résumé

Cette contribution présente les ressources linguistiques informatisées du laboratoire ATILF (Analyses et Traitements Informatiques du Lexique Français) disponibles sur la toile et sert de support aux démonstrations prévues dans le cadre de TALN 2001. L'ATILF est la nouvelle UMR créée en association entre le CNRS et l'Université Nancy 2 qui, depuis le 2 janvier 2001, a succédé à la composante nancéienne de l'INaLF. Ces importantes ressources sur la langue française regroupent un ensemble de plus de 3500 textes réunis dans Frantext et divers dictionnaires, lexiques et autres bases de données. Ces ressources exploitent les fonctionnalités du logiciel Stella, qui correspond à un véritable moteur de recherche dédié aux bases textuelles s'appuyant sur une nouvelle théorie des objets textuels. La politique du laboratoire consiste à ouvrir très largement ses ressources en particulier au monde de la recherche et de l'enseignement.

Abstract

This paper presents the computerized linguistic resources of the Research Laboratory ATILF (Analyses et Traitements Informatiques du Lexique Français) available via the Web, and will serve as a helping document for demonstrations planned within the framework of TALN 2001. The Research Laboratory ATILF is the new UMR (Unité Mixte de Recherche) created in association between the CNRS and the University of Nancy 2 since January 2nd, and succeeds to the local component of the INaLF situated in Nancy. This considerable amount of resources concerning French language consists in a set of more than 3500 literary works grouped together in Frantext, plus a number of dictionaries, lexis and other databases. These web available resources are operated and run through the potentialities and powerful capacities of a software called Stella, a search engine specially dedicated to textual databases and relying on a new theory of textual objects. The general policy of our laboratory is to welcome and give the research and teaching world the widest access to all our resources.

Mots Clés : Ressources linguistiques, Bases de données textuelles, Dictionnaires, Web

1 Introduction

L'évolution des études en linguistique tout comme en traitement automatique ou en ingénierie des langues (Pierrel, 2001) donne une place de plus en plus grande à l'exploitation systématique de ressources linguistiques informatisées, corpus, dictionnaires, lexiques, etc. L'ATILF, Laboratoire d'Analyses et Traitements Informatiques du Lexique Français, qui a pris la succession de la composante nancéienne de l'INaLF depuis le 2 janvier 2001, dispose d'un fonds très important de ressources linguistiques informatisées. Ces ressources sur la langue française regroupent un vaste ensemble de textes réunis dans Frantext et divers dictionnaires et lexiques. Accessibles sur la toile, ces ressources exploitent les fonctionnalités d'un logiciel de consultation, Stella, qui correspond à un véritable moteur de recherche dédié aux bases textuelles s'appuyant sur une nouvelle théorie des objets textuels. La politique du laboratoire consiste à ouvrir très largement ses ressources en particulier au monde de la recherche et de l'enseignement. Cet article court se donne comme objectif de présenter ces ressources linguistiques informatisées disponibles et sert de support aux démonstrations prévues dans le cadre de TALN 2001.

Nous y présenterons successivement :
- Les principales fonctions utilisateur du logiciel Stella
- Les produits accessibles librement, de loin les plus nombreux,
- Les produits accessibles par abonnement, en particulier Frantext et l'encyclopédie de Diderot et d'Alembert, et qui sont soumises à des droits d'usage spécifiques,
- Les produits actuellement uniquement à usage interne, base de synonymes, base de Moyen Français, etc., pour lesquels nous sommes ouverts à des utilisations externes dans le cadre de conventions de partenariat spécifiques
- Enfin, des produits à usage réservé, même au personnel du laboratoire, et pour lesquels il convient de faire une demande spécifique au service concerné, Marge, Région ……

2 Principales fonctionnalités d'accès aux ressources linguistiques du laboratoire via le logiciel Stella

Le logiciel Stella est une "boîte à outils" (bibliothèque de logiciels) offrant aux développeurs d'applications trois axes de services dont nous ne citerons que les points essentiels :

1) **Interfaces Web :** gestion des formulaires, de la notion de "session" utilisateur, intégration de menus déroulants à base de technologie Java, hypernavigation dynamique (c'est-à-dire sans hyper-liens prédéfinis) entre applications résidant ou non sur le même serveur.
2) **Services d'utilité générale :** tris optimaux par arbres binaires, "expressions régulières" au plus haut niveau du standard, base de connaissances lexicales permettant la flexion ou la lemmatisation des verbes, substantifs et adjectifs.
3) **Gestion des bases textuelles :** fabrication/maintenance de bases textuelles, système d'indexation mathématiquement optimal, architecture totalement ouverte grâce à la notion d'objet textuel abstrait, élaboration de requêtes de très haut niveau grâce à un compilateur intégré prenant en compte des schémas d'objets textuels arbitrairement complexes, décrits par l'utilisateur à l'aide de grammaires formelles.

Du point de vue de l'utilisateur, l'ensemble des services offerts par Stella se traduit par :
- **Le confort d'utilisation** : la notion de session lui permet de se créer un véritable environnement de travail, grâce à la possibilité de créer à titre temporaire des fichiers sur

le serveur (listes de mots, grammaires, fichiers résultats des requêtes) et de leur appliquer successivement les différents traitements informatiques offerts par l'interface.
- **Le temps de réponse optimal**, garanti par les mécanismes d'optimisation mis en œuvre.
- **La qualité de service** : Stella contient un "savoir linguistique" (flexion, existence de bases textuelles catégorisées) permettant à l'utilisateur de rechercher, bien plus que de simples chaînes de caractères, des entités complexes (verbes fléchis, catégories grammaticales…).
- **La puissance de l'interrogation** : la possibilité d'écrire des grammaires paramétrables (et donc réutilisables dans différents contextes) permet de réaliser des recherches textuelles impensables avec tout autre système de base textuelle.
- **La souplesse de l'hypernavigation :** elle permet de faire communiquer entre elles toutes les applications gérées par Stella : un simple clic de souris sur tout mot de n'importe quelle page Web produite par une application Stella permet de déclencher sa recherche dans les différentes bases de données gérées par Stella. Ainsi, les ressources textuelles de l'ATILF sont puissamment interconnectées les unes aux autres, et l'utilisateur peut naviguer, par exemple, en tous sens entre Frantext, le TLFI et les dictionnaires de l'Académie.

3 Produits accessibles librement :

3.1 Les dictionnaires de la langue dont le Trésor de la Langue Française

Au cours de la période 1960-1990, l'un des objectifs majeurs de l'INaLF fut de produire, à partir d'un vaste corpus incluant 80% de textes littéraires et 20% de textes techniques un dictionnaire de référence sur la langue du 19^{e} et du 20^{e} siècle : Le *Trésor de la Langue Française*, (*TLF*), paru en 16 tomes (CNRS 1971-1994). Au cours des dernières années, ce dictionnaire fut informatisé. Le TLFi (Trésor de la Langue Française Informatisé) se distingue des autres dictionnaires électroniques existants par la finesse de la structuration des données. Accessible à l'adresse http://www.inalf.fr/tlfi/ ce dictionnaire est doté d'une interface simple et conviviale offrant, essentiellement, trois niveaux de consultation via le logiciel Stella :
- lecture du dictionnaire, article par article, avec possibilité de mettre en évidence tel ou tel type d'information, par exemple, repérer les définitions, les syntagmes...
- consultation transversale. Il est possible de visualiser les mots d'origine espagnole, les mots utilisés dans la marine, les régionalismes canadiens, ...
- requêtes les plus arbitrairement complexes, par exemple, extraire les noms d'arbres, visualiser les termes de marine en rapport avec la manœuvre des voiles…

Notons par ailleurs que sur le site du laboratoire sont aussi accessibles, libres de droits, les 8ème édition http://www.inalf.fr/academie8/ et 9ème édition http://www.inalf.fr/academie9/ du dictionnaire de l'*Académie française* dans le cadre d'un partenariat entre l'Académie et le laboratoire. L'accès à ces dictionnaires est réalisé, comme pour le TLFi, via le logiciel Stella. Sont aussi accessibles, à l'adresse http://www.inalf.fr/dictionnaires/, un ensemble de dictionnaires du 16^{e} au 19^{e} siècle, dont les 1ère (1694), 5^{ième} (1798), et 6^{ième} (1835) éditions du dictionnaire de l'Académie française, le *dictionarium latinogallicum* de Robert Estienne, le *Thresor de la langue françoyse* de Jean Nicot, et le *dictionnaire historique et critique* de Pierre Bayle ; ainsi que diverses données sur le français quotidien, en particulier la présentation des travaux de l'équipe sur les expressions familières à l'adresse : http://www.inalf.fr/richlex/ et le guide de la féminisation http://www.inalf.fr/feminisation/ . Concernant les dictionnaires anciens, on peut y effectuer des recherches simples, par mot-

vedette ainsi que des recherches plein texte. Une consultation transversale, uniquement par mot vedette est proposé pour les trois dictionnaires de l'Académie et le dictionnaire de J. Nicot. Ainsi on peut remarquer que l'orthographe du mot "beste" a changé au fil du temps : on retrouve cette graphie uniquement dans le TLF de 1606 et dans l'édition de 1694 du Dictionnaire de l'Académie. Dans les éditions de 1798 et de 1835 du dictionnaire de l'Académie la graphie retenue est "bête".

3.2 La base historique du français

Cette base de données est constituée de datations du vocabulaire français. Les 30 premiers volumes des *DDL* (Datations et Documents Lexicographiques, Matériaux pour l'histoire du français) sont informatisés et consultables à l'adresse : http://www.inalf.fr/ddl/, via Stella.

3.3 Le catégoriseur Winbrill

Le catégoriseur Winbrill est une interface Windows du catégoriseur d'Éric Brill [Brill 1994] développé à l'Université de Pennsylvanie sous Unix, et paramétré pour le français à l'INaLF. Le module d'étiquetage a été porté sous Windows. Il est en téléchargement libre sur le site de l'ATILF. Les paramètres pour le français sont distribués librement après signature d'une convention, à des fins exclusives de recherches.
FLEMM [Namer 2000] est un programme Perl qui effectue l'analyse morphologique flexionnelle de textes français préalablement étiquetés par Brill ou TreeTagger. FLEMM calcule le lemme de chaque mot fléchi (en fonction de son étiquette) et fournit également ses principaux traits morphologiques. Ces deux outils (Winbrill et FLEMM v1.1) sont accessibles à l'adresse http://www.inalf.fr/winbrill/ ainsi que FLEMM v 2.0 à l'adresse http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm

3.4 Texte Balisés

Le service des bases textuelles a commencé à mettre à la disposition des chercheurs des textes intégraux, libres de droits, à consulter en ligne : *Cromwell, Gambara, Séphaphîta* de Balzac, *Le Barbier de Séville* de Beaumarchais, *les Complaintes* de Laforgue, et *Britanicus* de Racine. Pour certains d'entre eux, il est possible de télécharger une version au format XML accompagnée d'une feuille de Style à l'adresse http://www.inalf.fr/sbt/accueil.htm.

4 Produits accessibles par abonnement :

4.1 La base textuelle Frantext [Frantext 1992]

FRANTEXT peut se définir comme un vaste corpus, à dominante littéraire, constitué de plus de 3500 textes français qui s'échelonnent du 16^{e} au 20^{e} siècle. Cette base s'enrichit régulièrement d'œuvres nouvelles et, pour les textes déjà présents dans la base, d'éditions plus récentes. Il est prévu de l'étendre à la période 1330-1500, correspondant au moyen français, dont les données sont actuellement consultables dans une base autonome.

Le logiciel d'interrogation est conçu pour répondre à des questions lexicales, syntaxiques ou littéraires. Il permet d'explorer soit l'ensemble de la base, soit un sous-ensemble défini par l'utilisateur par tranches chronologiques, par auteurs, par titres, par genres littéraires ou par diverses combinaisons de ces critères. Les requêtes portent sur des chaînes de caractères, des

mots-formes isolés ou groupés en listes, des vocables lemmatisés, sur des catégories grammaticales (dans la base de textes catégorisés) ou encore sur des expressions complexes combinant ces éléments. L'exploration permet d'extraire des listes de formes attestées, par exemple celles qui sont construites sur un même radical ou comportent un même suffixe, de rechercher tous les cooccurrents d'un mot donné, d'obtenir des informations quantitatives sur la répartition des mots dans l'ensemble traité et surtout de consulter tous les contextes dans lesquels apparaissent des mots ou des suites de mots, éventuellement discontinues, au moyen d'un outil qui offre de riches possibilités d'investigation.

L'ensemble de ces données, dont la plupart sont aujourd'hui encore sous droits d'auteur ou d'éditeur, ne sont accessibles que sous conditions particulières d'abonnement limitées au monde de la recherche et de l'enseignement (coût actuel de l'abonnement : 2000 francs pour un sous-domaine de réseau) à l'adresse http://www.inalf.fr/frantext/ via le logiciel Stella.

4.2 L'encyclopédie de Diderot et d'Alembert

Une version de la première édition de cette encyclopédie est disponible à l'adresse http://www.inalf.fr/encyclopedie/. Réalisée en collaboration avec ARTFL de l'Université de Chicago elle comporte dix-sept volumes de texte et onze volumes de planches au sein desquels on peut naviguer par recherche d'articles de plusieurs façons, notamment par vedette ou auteur ou par recherche plein texte sous Philologic [Olsen 1996 ; Andreev 1999]. Philologic est un moteur de recherche pour les bases textuelles. C'est un outil qui englobe des scripts Perl qui communiquent avec les butineurs www via la traditionnelle interface CGI et des programmes écrits en langage C.

Cette base de données contient plusieurs modules coordonnées : une base des objets textuels, une base des mots, un index de concordance des mots (qui reconnaît les objets textuels) et un gestionnaire d'objets. Le mécanisme des interrogations plein texte est identique à celui d'interrogation par mot vedette à l'exception du traitement des meta-données. La recherche plein texte exploite les 15 éléments de la spécification *Dublin Core* qui est ultilisé pour les méta données. L'extraction du contenu des entêtes se fait conformément au spécifications ATE (ARTFL Text Encoding). Ce codage permet quelques extensions : identificateurs de page, balises spécifiques pour les fins de phrase et pour les noms propres, par exemple.

Pour l'Encyclopédie Diderot, le système Philologic permet des recherches par mot-vedette, par auteur, etc., ainsi qu'une recherche par mot adjacent (pattern matching). La recherche plein texte accepte les expressions régulières et les opérateurs logiques. Pour l'Encyclopédie, les textes sont traités comme une hiérarchie d'objets textuels : volumes, articles (72 000), sous articles, paragraphes, phrases et mots (environ 21 millions) , tout cela en parallèle avec d'autres structures (images, titres de pages, planches (3000), etc.). La recherche d'information comporte quatre parties : définition du corpus (sélection d'objets ou de mots par co-occurrence), recherche dans les index, extraction du contexte et formatage final du format de sortie (conversion des balises SGML dans du HTML, résolution des références croisées, etc.).

5 Produits à accès réservés

Outre les ressources linguistiques informatisées, présentées ci-dessus et qui, dès aujourd'hui peuvent être accessibles, avec ou sans abonnement suivant les cas, via le site Web du laboratoire http://www.inalf.fr/atilf, diverses autres ressources sont disponibles au laboratoire

et peuvent être ouvertes dans le cadre de partenariats spécifiques. En particulier nous disposons d'une Base d'Index Techniques Cumulatifs qui permet de localiser rapidement dans quels dictionnaires techniques sont attestés un terme, une locution ou un syntagme. La base est interrogeable à l'adresse http://www.inalf.fr/itc/ à partir de deux index :

- ITC1 : comprend les dictionnaires monolingues de spécialités pour la période 1800-1980 (530 dictionnaires indexés à partir de 183 domaines techniques).
- ITC2 : nouvelle version de 230 dictionnaires techniques indexés à partir d'une liste de domaines mise à jour et dans lesquels ont été relevées des vedettes d'articles, des syntagmes, des synonymes, des équivalents étrangers pour la période 1980-1998.

6 Conclusion

L'ensemble de ces ressources sera présenté sous forme de démonstration. Notre objectif à travers une telle présentation est tout à la fois de faire connaître ces ressources, fruits de nombreuses années de travail au sein de l'INaLF et aujourd'hui disponibles à l'ATILF. Mais notre effort de recherche et développement se poursuit aujourd'hui et, au terme de cette rapide présentation, nous tenons à affirmer notre volonté de créer des partenariats avec d'autres pour contribuer à rendre plus largement disponibles et exploitables dans le monde de la recherche et de l'enseignement des ressources sur notre langue indispensables au développement d'études et de recherches sur sa connaissance et son informatisation.

Références

Brill E. Some Advances in Transformation-Based Part-of-Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94*), 1994

CNRS (1976-1994) *TLF, Dictionnaire de la langue du 19e et 20e siècle*, CNRS, Gallimard, Paris, 1971 - 1994

Dendien J. Access to information in a textual database : access functions and optimal indexes, Oxford : Clarendon press, 1991.

FRANTEXT. *Autour d'une base de données textuelles ; témoignages d'utilisateurs et voies nouvelles*, Paris, Didier Érudition, 1992, 359 p.

Namer F., FLEMM : un analyseur flexionnel du français à base de règles", in *Traitement automatique des langues pour la recherche d'information*, Ch. Jacquemin (éd.) T.A.L., volume 41, n°2/2000, HERMES, Paris.

Pierrel J.M., *Ingénierie des langues*, Paris, Hermès Science Publication, 340 p.

Olsen M., Text Theory and Coding Practice: Assessing the TEI, in Joint Annual Conference of the Association for Comuters and the Humanities and Association for Literary and Linguistic Computing, Bergen, Norway, June 1996.

Andreev L., Olsen M., Conception de systèmes Hypermedia à grande échelle pour les sciences humaines: présentation de Philologic : le logiciel d'ARTFL, in 67th Congrès de l' ACFAS (l'Association canadienne-française pour l'avancement des sciences), May 11, 1999 ,University of Ottawa.