Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL

Etienne Petitjean, Jean-Marie Pierrel

Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL

Jean-Marie Pierrel
jean-marie.pierrel@atilf.fr

Etienne Petitjean
etienne.petitjean@atilf.fr

ATILF & CNRTL, CNRS – Nancy Université

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

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

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 les réseaux européens CLARIN et DARIAH.

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 european networks of resource centres for the humanities (CLARIN and DARIAH).

1 Introduction

Dans le cadre des travaux de recherche en Sciences Humaines et Sociales, les aspects de valorisation et exploitation scientifiques de documents sont particulièrement importants et stratégiques pour servir de support à la fois :

aux travaux de recherche : la notion de corpus d'étude est en effet très présente dans la plupart des disciplines en SHS, que cela soit en linguistique, et plus spécifiquement en linguistique de corpus, en langue, en littérature, en histoire, en droit, en didactique, etc.

à la diffusion des résultats de ces travaux qui passe le plus souvent par la production de documents textuels : articles et livres bien sûr, mais aussi documents plus spécialisés tels des dictionnaires ou des lexiques.

Aujourd'hui, un des aspects essentiels pour l'exploitation et la valorisation de tels documents est leur informatisation et leur disponibilité sur le Web sous forme de documents numériques permettant, grâce à des outils intelligents de recherche et de navigation qui ne se limitent pas à l'exploitation de simples mots clés ou d'informations décrivant leurs structures, un véritable accès par le contenu à travers soit une recherche plein texte, soit une exploitation d'annotations et de balisages représentatifs de ce contenu informationnel.

Dans cet article, après avoir discuté de l'intérêt de proposer une valorisation ou exploitation de documents numériques pour la recherche en SHS, et plus particulièrement dans le champ de la linguistique, nous nous interrogerons sur les contraintes que doivent, à nos yeux, respecter de tels documents numériques. Puis en nous limitant à des documents numériques relatifs à notre domaine de compétence, la linguistique et plus particulièrement le lexique français, nous détaillerons deux exemples de structuration et d'exploitation de documents numériques : le TLFi (www.atilf.fr/tlfi), version informatisée du Trésor de la Langue Française [1] l'un des plus grands dictionnaires de langue du français et le portail lexical du Centre National de Ressources Textuelles et Lexicales (CNRTL : www.cnrtl.fr), récemment mis en place, au sein de l'ATILF, sous l'égide du CNRS.

Avertissement

Nous illustrerons au maximum cet article par un certain nombre de figures reflétant divers usages de ces documents numériques, néanmoins nous conseillons au lecteur d'ouvrir une fenêtre sur son navigateur préféré pour mieux percevoir l'intérêt de telles exploitations de documents structurés en suivant les liens indiqués dans le texte.

2 Importance de la valorisation et de l'exploitation de documents numériques pour la recherche linguistique

2.1 Les enjeux de la linguistique de corpus et du traitement automatique des langues

Le traitement automatique des langues (TAL) et la linguistique de corpus sont devenus, au cours des dernières années, des domaines-clés pour répondre aux besoins de notre société en terme d'analyse et d'exploitation de gisements d'information, le plus souvent sous forme textuelle, et aujourd'hui largement disponibles, en particulier sur le Web [2]. 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 grandes tailles (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 » [3]. Ainsi se structura petit à petit un nouveau champ de recherche : la linguistique de corpus [4] 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 normalisation et 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.

Par ailleurs, ce champ de la linguistique de corpus et du TAL est porteur d'enjeux incontournables tant pour une meilleure connaissance et modélisation de la langue que pour nous permettre de progresser vers une véritable exploitation du contenu informationnel le plus souvent sous formes langagières ou textuelles, ou de valider, échanger et confronter nos résultats en TAL.

2.2 Quelles ressources et corpus pour l'étude des langues aujourd'hui ?

2.2.1 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 disponibles sur le Web [5] :

1. Leur qualité est souvent très discutable. Sans parler des nombreuses fautes qui demeurent dans bien des textes disponibles sur la toile, on y retrouve un mélange de textes, de formes, de genres et de niveaux de langue ou d'époques très disparates, incompatible avec la nécessité de travailler sur des corpus homogènes de référence pour pouvoir tout à la fois construire des modèles pertinents, les valider et les confronter.
2. La pérennité de leur disponibilité n'est pas toujours assurée. Le propre du Web est de fournir des informations en constante évolution et, dans le cadre de projets de recherche, leur durée de vie est souvent inférieure à la durée de vie des projets qu'elles sous-tendent, ce qui rend très souvent impossible des comparaisons objectives de résultats.

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 français 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.

2.2.2 Dictionnaires et 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 automatique de la langue ne peut se passer du niveau lexical, et la disponibilité de ressources de ce type est unanimement reconnue comme indispensable pour la plupart des traitements. 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. Il suffit pour s'en convaincre d'analyser les réponses que l'on peut obtenir après une interrogation de la toile à partir, par exemple, de « dictionnaire + langue française » ! Nous sommes pour notre part convaincus qu'il importe de développer et partager des ressources de ce type et c'est cette conviction qui nous amena à proposer une version informatisée du Trésor de la Langue Française (www.tlfi.fr) 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/).

2.2.3 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 :

1. 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 ?
2. 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.

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 !

2.3 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 de TAL ou de linguistique, 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 complète 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 de ressources concerne l'évaluation, de plus en plus indispensable, de nos productions de recherche (analyseurs, système de traitement), qui nécessite, pour des besoins de comparaison, la

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

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 recheche 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 des documents et 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 (cf. $5), qu'il a mis en place au sein des SHS au cours des derniers 18 mois.

3 Quelles contraintes pour de tels documents numériques ?

3.1 Une nécessité de normalisation

L'une des caractéristiques essentielles de la recherche, à laquelle les sciences du langage n'échappent pas, est donc la nécessité de permettre à la communauté scientifique de pouvoir échanger, évaluer, confronter et reproduire des résultats de recherche ou d'analyse à partir de données de référence. Cela nécessite le développement et l'usage de normes pour les ressources linguistiques et textuelles. Aujourd'hui, dans le cadre des documents numériques, XML s'impose. Mais au-delà d'une simple utilisation d'un langage commun, il convient aussi de partager, voire normaliser, les schémas de balisage (ou DTD). La communauté française est fort bien placée en ce domaine que ce soit dans le cadre du comité technique TC 37 de l'ISO (le sous-comité dédié aux ressources linguistiques et à leur normalisation (SC4)) ou dans le cadre de la Text Encoding Initiative (TEI) [6, 7], consortium international qui définit des recommandations de codage de ressources textuelles (www.tei-c.org). Ainsi à travers un partenariat entre trois laboratoires, l'ATILF, l'INIST et le LORIA, Nancy est devenu centre européen support de cette initiative et aujourd'hui cette tâche est plus spécifiquement confiée au CNRTL.

3.2 Une transparence nécessaire des outils informatiques pour l'exploitation de tels documents informatiques

Une autre caractéristique indispensable au partage de ressources linguistiques (corpus, dictionnaires et lexiques) sous forme de documents électroniques au sein de la communauté SHS est la transparence des outils et structures informatiques. L'immense majorité des usages potentiels de tels documents numériques viennent en effet de collègues qui ne sont pas informaticiens et pour lesquels d'ailleurs l'outil informatique, au-delà des fonctions de base de la bureautique, fait peur. La solution que nous préconisons donc est d'offrir des services et des outils qui, en aucun cas, ne nécessitent un équipement logiciel spécifique pour les exploiter. La solution technique existe, c'est celle que nous utilisons dans les deux exemples que nous présentons ci-après, elle consiste à assurer cette diffusion à travers des applications Web qui ne nécessite aucun équipement autre, chez l'utilisateur, qu'un navigateur internet quel qu'il soit (Explorer, Netscape, Mozilla ou autre).

4 Exemple de valorisation d'un document électronique : le TLFi

4.1 Caractéristique du TLFi

Reflet fidèle de la version papier [1], jusque dans sa présentation typographique à l'écran, le 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 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 [8] intègre de plus des accès à très haut niveau de tolérance permettant une insensibilité aux accents, une tolérance aux fautes d'orthographe courantes, un traitement phonétique et un traitement morphologique. Ainsi, on peut offrir une correction automatique des fautes et permettre des accès à partir de formes et non plus uniquement de lemmes ou de vedettes et proposer des procédures d'accès diversifiées pour une consultation humaine.

4.2 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 divers accès possibles à l'ensemble du dictionnaire, cumulant les avantages d'un dictionnaire avec ceux d'une ressource textuelle et d'une véritable base de données lexicales :

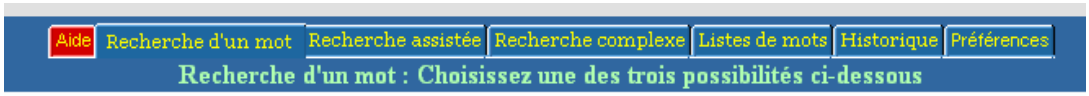
Recherche d'un mot, d'une expression ou d'une forme lexicale plus ou moins bien orthographiés, avec possibilité, via un « panneau de réglage », de mettre en évidence divers champs dans le résultat de la recherche (définition, code grammatical, domaine spécifique, exemple, auteur d'exemple, construction, indicateur, etc.).

Possibilité d'hyper-navigation à l'intérieur du dictionnaire permettant en un clic souris de passer d'un mot à sa définition.

Interrogations assistées ou requêtes complexes exploitant l'ensemble de la structure du dictionnaire à travers le croisement de multiples critères.

4.3 Exemples de recherches dans le TLFI

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations sur les recherches offertes dans le TLFi, 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 [9], soit à se connecter directement à l'adresse : http://atilf.atilf.fr/tlf.htm. Trois principaux types d'accès sont alors proposés : la recherche d'un mot, la recherche assistée et la recherche complexe



4.3.1 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).

4.3.2 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……

ou encore l'ensemble des mots dont la définition utilise le mot *liberté* [type d'objet : définition, contenu : &mliberté] ; on obtient ainsi 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…..

4.3.3 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 "*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 :

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

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 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	3 ART CULIN. 3 6 Empr. au lat. 6
NAPOLITAIN, -AINE, adj. et subst.	
1 NAPOLITAIN, -AINE, adj. et subst. 1 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	3 ART CULIN. 3 6 Empr. à l'ital. 6
ETC.	

4.4 L'impact du TLFi

Le TLFi, sans aucun doute le plus grand dictionnaire informatisé consacré à la langue française, grâce à la richesse de son contenu, entièrement encodé en XML, a ouvert des perspectives intéressantes. Sa mise à disposition sous forme de Cédérom et sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue : objet de plusieurs centaines de milliers de connexions quotidiennes en provenance de tous les continents, il est référencé par d'innombrables sources et la notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française. Alors même que le TLF avait la réputation tenace d'être un dictionnaire réservé à une élite, sa version informatique et les interconnexions par hyper-navigation avec le dictionnaire de l'Académie, FRANTEXT[3] ou la base historique du 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 injustifiée.

5 Le CNRTL : un outil de mutualisation des ressources numériques

5.1 Présentation des objectifs spécifiques du CNRTL

5.1.1 Les missions du CNRTL

Les missions du CNRTL (Centre National de Ressources Textuelles et Lexicales : www.cnrtl.fr) mis en place par le CNRS au sein de l'ATILF peuvent se résumer en sept points :

« 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.

5.1.2 Le CNRTL nœud de réseaux européens ou internationaux plus vastes

Au-delà de sa seule mission nationale, le CNRTL participe au réseau européen CLARIN (http://www.mpi.nl/clarin) des centres de gestion de ressources linguistiques qui correspond à l'une des propositions européennes d'infrastructure de

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

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. Dès à présent, le CNRS, la MPG4 (Allemagne), l'AHDS5 (Grande-Bretagne) et le DANS6 (Pays-Bas), sont directement impliqués dans la mise en place de ce projet.

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 : la possibilité pour un utilisateur de pouvoir soumettre une ou plusieurs ressources numériques (texte brut, texte 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 : permettre le téléchargement des ressources sélectionnées dans le format de sortie souhaité par les utilisateurs (XML, PDF, Word, HTML, etc.)

5.2 Les ressources gérées au sein du CNRTL

Le CNRTL se construit 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).

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 le portail lexical sur lequel nous allons revenir dans le paragraphe suivant, 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 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 (TLFi Dictionnaire de l'Académie française : 8ème et 9ème éditions) qu'anciens (du XVIe au XIXe siècle).

5.3 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 été étiqueté, au moyen de l'un des deux catégorisateurs : Brill ou TreeTagger.

[4] Max-Planck Gesellschaft : http://www.mpg.de
[5] Arts and Humanities Data Service : http://ahds.ac.uk/
[6] Data Archiving and Network Services : http://www.dans.knaw.nl

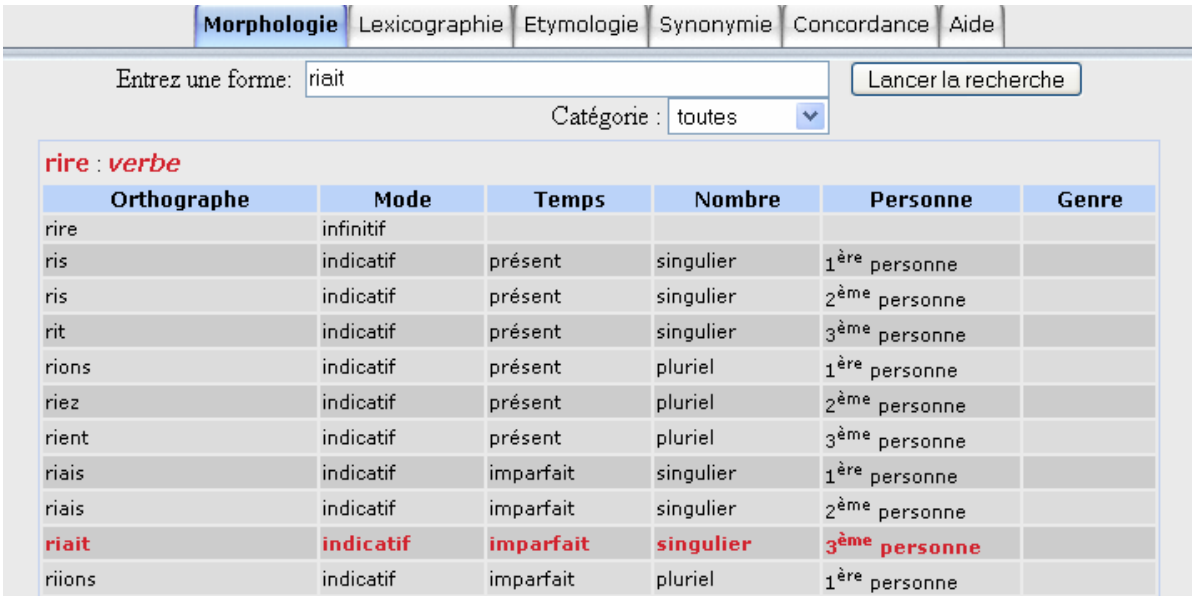
- DERIF : outil d'analyse morpho-sémantique du français qui s'applique à des entrées lexicales catégorisées issues d'un dictionnaire de la langue générale, capable de traiter des mots hors-dictionnaire et dont les résultat associent la morphologie et la sémantique
- POMPAMO : outil de détection de candidats 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.

5.4 Un exemple d'intégration de documents numériques : 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 exploite aujourd'hui divers documents numériques pour fournir, à partir d'une forme lexicale, cinq 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/) et une concordance utilisant le corpus des textes de la base Frantext (www.atilf.fr/frantext). 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.

Voici à titre d'exemple le type de résultat accessible via ce portail :

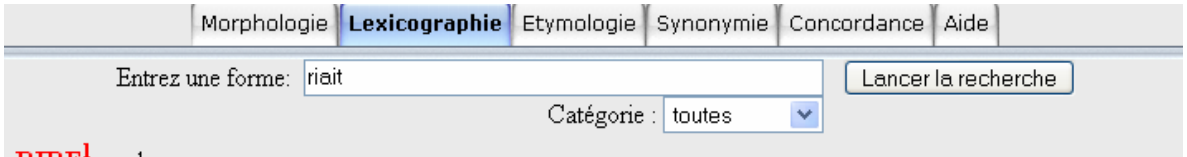
Informations morphologiques, accessibles directement pour la forme *riait* par : http://www.cnrtl.fr/morphologie/*riait*



rire : *verbe*

Orthographe	Mode	Temps	Nombre	Personne	Genre
rire	infinitif				
ris	indicatif	présent	singulier	1ère personne	
ris	indicatif	présent	singulier	2ème personne	
rit	indicatif	présent	singulier	3ème personne	
rions	indicatif	présent	pluriel	1ère personne	
riez	indicatif	présent	pluriel	2ème personne	
rient	indicatif	présent	pluriel	3ème personne	
riais	indicatif	imparfait	singulier	1ère personne	
riais	indicatif	imparfait	singulier	2ème personne	
riait	**indicatif**	**imparfait**	**singulier**	**3ème personne**	
riions	indicatif	imparfait	pluriel	1ère personne	

Informations lexicographiques, accessibles directement pour la même forme : http://www.cnrtl.fr/lexicographie/*riait*



RIRE[1], verbe

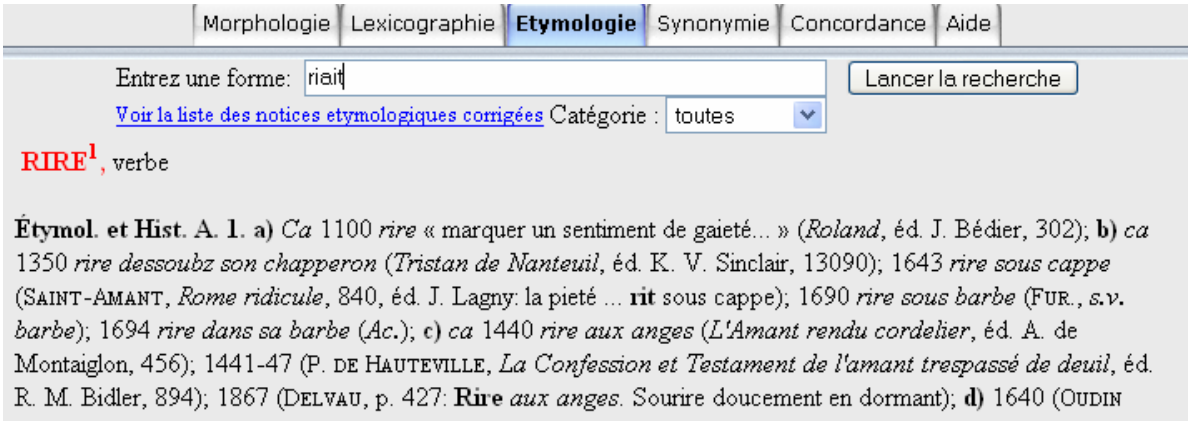
I. – *Empl. intrans.*

A. – [Le suj. désigne une pers.]

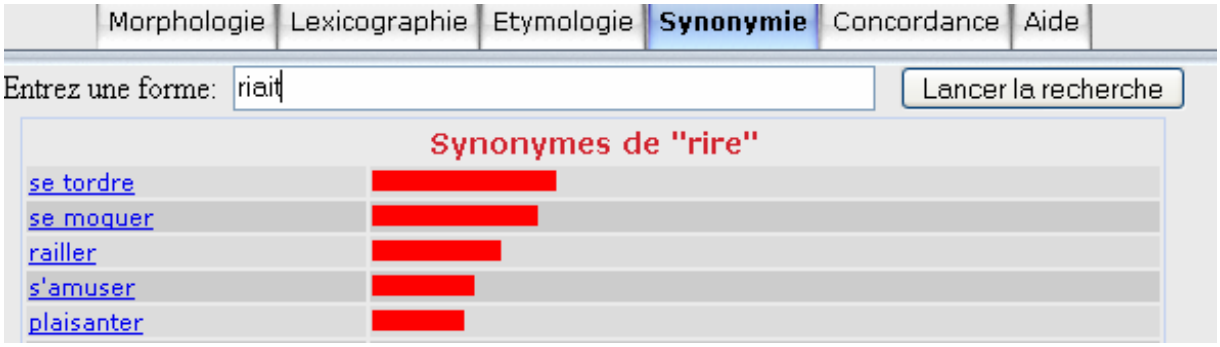
1. a) Manifester un état émotionnel, le plus souvent un sentiment de gaieté, par un élargissement de l'ouverture de la bouche accompagné d'expirations saccadées plus ou moins bruyantes et un léger plissement des yeux. Synon. fam., pop. *se bidonner, s'esclaffer, se fendre* la pêche, la pipe, se gondoler, se marrer, pouffer, rigoler*[1], *se tordre*. [*L'oncle Adolphe*] *rit de tout son cœur. Ça ne lui arrive pas souvent. Alors, je questionne, inquiète: – Je suis grotesque?... – Non (...) Tu es drôle!* (GYP, *Souv. pte fille*, 1928, p. 160).

SYNT. *Rire doucement, très fort, tout bas, tout haut; rire sans sujet, hors de propos, pour un rien; rire de surprise, d'un jeu de mots; commencer, se mettre, se prendre à rire; partir d'un éclat de rire; ne pouvoir s'empêcher de rire; s'arrêter, finir, faire semblant, avoir envie, donner envie de rire; se mordre, se pincer les*

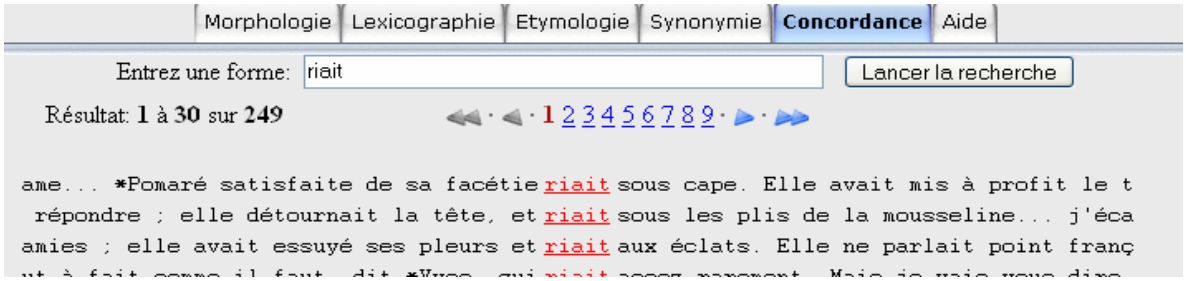
Informations étymologiques, accessibles directement par : http://www.cnrtl.fr/etymologie/*riait*



Informations de synonymie, accessibles directement par : http://www.cnrtl.fr/synonymie/*riait*

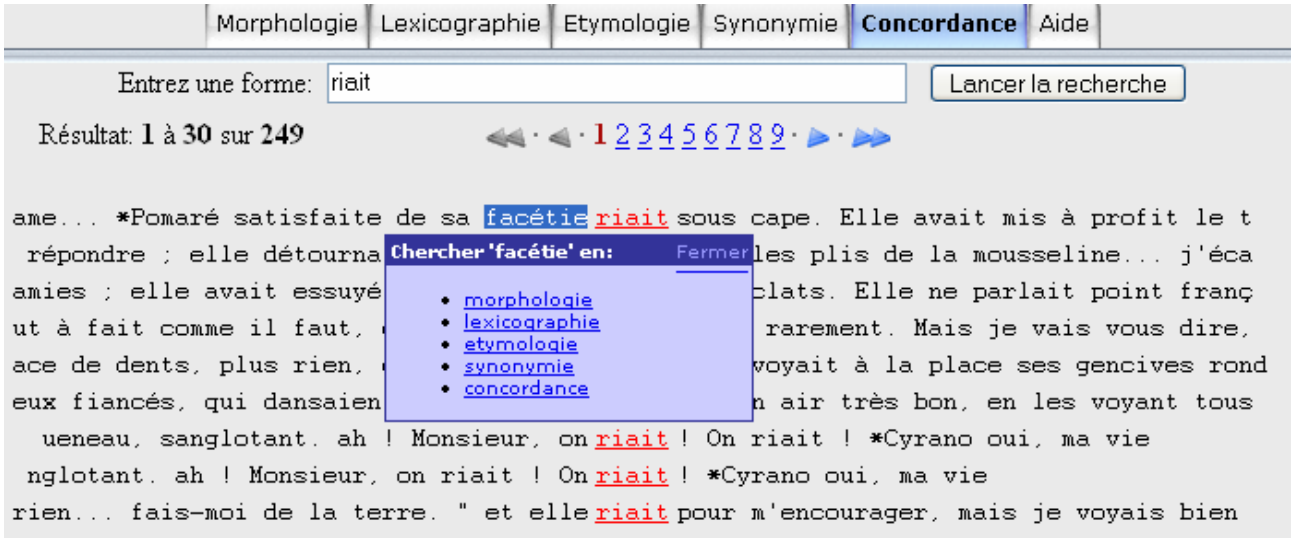


concordances, accessibles directement par : http://www.cnrtl.fr/concordance/*riait*



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 le mot « facétie » du premier exemple de concordance, un double-clic sur le mot affiche un menu qui permet d'hyper-naviguer vers les informations lexicales de ce mot :



6 Conclusion

L'exploitation et le partage de documents électroniques, version informatisée de productions scientifiques issues de nos laboratoires offrent aujourd'hui des modes nouveaux de mutualisation de ressources ou de résultats de recherche. C'est sur cette base que le CNRS a récemment décidé de créer dans le domaine des SHS des Centres Nationaux de Ressources dont le CNRTL pour l'écrit. Ce mouvement de fond sur la mutualisation de ressources sous forme de documents numériques est aussi à l'origine des efforts européens de mise en place d'infrastructures de recherche tels les projets CLARIN et DARIAH. Mais au-delà du seul monde universitaire, ces techniques permettent aussi de mieux partager avec 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 documents électroniques est ainsi en train de modifier notablement les modes de travail et d'échanges scientifiques au sein des communautés de recherche SHS.

Bibliographie

[1] CNRS 1976-1994 TLF, Dictionnaire de la langue du 19e et du20e siècle, CNRS, Gallimard, Paris.

[2] J.M. Pierrel *Ingénierie des Langues*, Traité Information - Commande - Communication, Hermès, octobre 2000.

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

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

[5] J.M. Pierrel 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

[6] Web et cOrpus

[7] Ide N., Véronis J. (eds.), "The Text Encoding Initiative: background and context", Special issue of Computers and the Humanities, vol. 29 n° 1/2/3, 1995.

[8] Ide N., Romary L., "International standard for a linguistic annotation framework", International Journal of Natural Language Engineering, vol. 10 n° 3-4, 2004 (b), p. 211-225.

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

[10] 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