

Communication de Monsieur le Professeur Jean-Marie Pierrel



Analyse et traitement informatique de la langue française

À ses débuts, l'informatique avait pour principal objet d'automatiser le calcul sur les nombres dans des domaines aussi divers que la balistique ou la comptabilité. C'est d'ailleurs là qu'il convient de rechercher l'origine du terme « calculateur ». Mais, progressivement, il est apparu que l'informatique pouvait également nous aider à gérer et structurer des flux d'informations et ne pas se limiter au seul calcul : l'informatique est alors définie comme le domaine d'activité scientifique, technique et industriel concernant le traitement automatique de l'information par des machines. Un peu plus tard, dans les années 1990, le mariage entre les télécommunications et l'informatique fait émerger un nouveau domaine, au sein duquel la gestion de l'information dans l'entreprise (Intranet) ou au niveau mondial (Internet, Web, autoroutes de l'information) devient centrale : les TIC (technologies de l'information et de la communication). Une véritable révolution se met en place : la « société de l'information », caractéristique de notre XXI^e siècle prend la suite de la « société industrielle » qui caractérisa le début du XX^e siècle. Dans ce cadre, les aspects d'acquisition, de gestion, de structuration, d'analyse et d'interprétation, de modélisation et d'exploitation informatiques des informations et connaissances, pour la plupart sous formes langagières et textuelles, apparaissent au centre des grands débats du monde de la recherche ou de l'économie.

Il convient pourtant de ne pas limiter les relations entre la langue et l'informatique à ces seuls aspects mis récemment sous le feu des projecteurs. Dès les années 1950, alors que l'informatique n'en était qu'à ses débuts, est apparue l'idée d'utiliser les ordinateurs pour traduire des textes d'une langue à l'autre. Puis, assez rapidement, ce champ d'exploitation des langues s'est diversifié et peu à peu étendu à la vérification et correction orthographiques, à l'indexation et au résumé de textes, etc. Ainsi, deux disciplines *a priori* très différentes,

l'informatique et la linguistique, se sont confrontées et rapprochées pour donner naissance à une informatique linguistique ou une linguistique informatique.

Le traitement automatique des langues (TAL) et la linguistique de corpus sont ainsi devenus 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 [Pierrel 2000]. 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 » [Habert 1995]. Un nouveau champ de recherche se structura : 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.

De nécessité de ressources de références pour le traitement automatique de la langue

Ces études et recherches en traitement automatique des langues 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.

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 ou le linguiste, 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 :

- 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.
- 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 impossibles 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, dès 2002, le projet « technolangue », 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*. Plus récemment, la création d'une Très Grande Infrastructure de Recherche (TGIR) « Corpus » (<http://www.corpus-ir.fr/>) se propose de poursuivre cet effort. 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), et la nécessité de vastes corpus normalisés, annotés et validés s'impose.

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. 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.atilf.fr/tlfi) 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/>).

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.

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 !

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

En conclusion de cette première partie, nous souhaitons faire partager notre conviction de la nécessité de mutualiser 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. À 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 le traitement automatique de la langue 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 automatique), qui nécessite, pour des besoins de comparaison, la disponibilité de ressources de référence (corpus textuels, corpus d'exemples sur un phénomène de langue, ressources dictionnaires) 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 leurrions 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 Centre National de Ressources Textuelles et Lexicales (CNRTL, www.cnrtl.fr) mis en place à Nancy au sein de notre laboratoire.

Quelles contraintes pour de telles ressources linguistiques informatisées ?

Une nécessité de normalisation

L'une des caractéristiques essentielles de la recherche, à laquelle ce domaine n'échappe pas, est aussi de permettre à la communauté scientifique d'échanger, d'évaluer, de confronter et de 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 SC4 dédié aux ressources linguistiques et à leur normalisation) ou dans le cadre de la Text Encoding Initiative (TEI : <http://www.tei-c.org>), consortium international qui définit des recommandations de codage de ressources textuelles. Dans ce cadre, à travers un partenariat entre trois laboratoires (l'ATILE, l'INIST et le LORIA), Nancy est devenu centre européen support de cette initiative.

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 Sciences Humaines et Sociales 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).

Exemples de contributions nancéiennes

Le Trésor de la Langue Française informatisé www.atilf.fr/tlfi

Nous avons présenté dans une contribution antérieure [Pierrel 2011] ce Trésor de la Langue Française informatisé. Nous nous contenterons ici d'en rappeler les principales caractéristiques et ses différentes modes d'interrogation.

Caractéristique du TLFi

Reflète fidèle de la version papier [CNRS 1976-1994], 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 métatextuels 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 (Dendien & 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 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.

Quels accès au TLFi ?

Le TLFi correspond à une rétroconversion 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 permettent 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'hypernavigation 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.

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 [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



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 correspondants 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 telle « 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 les définitions :

Objets de la recherche : 1. ¶Paragraphe 1

① 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 sonnante 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. trompette (infra dér.). Le trompette noir du dancing* (BEAUVOIR, *Mandarins*, 1954, p. 306).

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, HOCHÉ-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 : *&mvoile^[1]*]. Voici un extrait des 61 réponses que l'on obtient :

| |
|--|
| ABRIER, ABREYER, verbe trans. |
| 3. Empêcher 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. |
| 3. Fixer 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 : ɢliberté] ; on obtient ainsi 306 réponses dont :

| Objets de la recherche : 1 Définition 1 |
|---|
| ABUSER, verbe trans. 1 Exagérer 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. 1 Grande liberté. 1 |
| ALIÉNANT, ANTE, part. prés. et adj. 1 Qui prive l'homme de son humanité, de sa liberté : 1 |
| Etc..... |

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 :

Objets de la recherche : 1 Entrées 1 3 Domaine technique 3 4 Définition 4 5 Source 5 6
Langue empruntées 6

| | |
|---|---------------------|
| MORTIFICATION, subst. fém. | |
| 1 MORTIFICATION, subst. fém. 1 | 3 ART CULIN. 3 |
| 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 |
| 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. | |

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 hypernavigation avec le dictionnaire de l'Académie, FRANTEXT^[2] 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.

Le CNRTL : un outil de mutualisation des ressources linguistiques

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 ainsi :

- contrôler et valider 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 ;
- sauvegarder, stocker et maintenir 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 ;

- partager et enrichir de ressources, afin d'assurer une réelle mutualisation entre équipes de recherche ;
- administrer des ressources et aider les utilisateurs afin de permettre 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 ;
- enfin former des producteurs et les utilisateurs de telles ressources aux méthodologies d'annotation, de codage et de normalisation. Sur ce point fort du fait que Nancy est centre support de la TEL, on s'appuie autant que faire se peut sur les recommandations de la TEL.

Le CNRTL nœud de réseaux européens ou internationaux plus vastes

Au-delà de sa seule mission nationale, le CNRTL participe au réseau européen CLARIN des centres de gestion de ressources linguistiques (Commun Language Resources and Technology Infrastructure : www.clarin.eu) et à DARIAH (Digital Research Infrastructure for the Arts and Humanities : <http://www.dariah.eu/>), deux infrastructures européennes de recherche récemment mises en place. Ces projets sont également l'occasion d'une réflexion commune sur la gestion de plates-formes ouvertes de gestion et d'archivage de documents numériques.

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 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 droit. Parmi les ressources déjà intégrées au CNRTL, outre le portail lexical sur lequel nous allons revenir dans le paragraphe suivant, il convient de noter :

- Le corpus journalistique « Est Républicain » : dans le cadre d'un accord de collaboration avec L'Est Républicain, le CNRTL offre après en avoir assuré le traitement informatique, l'accès à l'un des plus vastes corpus journalistiques. Ce corpus est constitué des données textuelles correspondant à ce jour à trois années de toutes les éditions intégrales du quotidien régional.
- Les corpus de textes libres de droits 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 Définies (DEDE : issu d'une coopération entre les laboratoires 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 : 8e et 9e éditions) qu'anciens (du XVI^e au XIX^e siècle).

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 morphosémantique 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.

- FastKwic : un outil d'indexation automatique permettant de produire un concordancier. Il fonctionne sur le français et l'anglais encodés en ISO-latin-1 (ISO-8859-1) ou UTF-8. Il s'appuie sur une ressource lexicale/terminologique qui doit être compilée avec FastKwic dans une première étape. Dans une seconde étape, FastKwic permet d'indexer un texte et de produire un concordancier à partir du résultat de cette indexation.

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 et TLF-Etym (www.atilf.fr/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 *riaït* par : <http://www.cnrtl.fr/morphologie/riaït>

| Morphologie | | | | | |
|--|-----------|-------------------|-----------|--|-------|
| Lexicographie | | Etymologie | | Synonymie | |
| Concordance | | Aide | | | |
| Entrez une forme: <input type="text" value="riaït"/> | | Catégorie: toutes | | <input type="button" value="Lancer la recherche"/> | |
| rire : <i>verbe</i> | | | | | |
| Orthographe | Mode | Temps | Nombre | Personne | Genre |
| rire | infinitif | | | | |
| ris | indicatif | présent | singulier | 1 ^{ère} personne | |
| ris | indicatif | présent | singulier | 2 ^{ème} personne | |
| rit | indicatif | présent | singulier | 3 ^{ème} personne | |
| riens | indicatif | présent | pluriel | 1 ^{ère} personne | |
| riez | indicatif | présent | pluriel | 2 ^{ème} personne | |
| rient | indicatif | présent | pluriel | 3 ^{ème} personne | |
| riaïs | indicatif | imparfait | singulier | 1 ^{ère} personne | |
| riaïs | indicatif | imparfait | singulier | 2 ^{ème} personne | |
| riaït | indicatif | imparfait | singulier | 3 ^{ème} personne | |
| riens | indicatif | imparfait | pluriel | 1 ^{ère} personne | |
| riez | indicatif | imparfait | pluriel | 2 ^{ème} personne | |
| riaient | indicatif | imparfait | pluriel | 3 ^{ème} personne | |
| ris | indicatif | passé simple | singulier | 1 ^{ère} personne | |
| ris | indicatif | passé simple | singulier | 2 ^{ème} personne | |

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

Morphologie **Lexicographie** Etymologie Synonymie Concordance Aide

Entrez une forme:

Catégorie: toutes

RIRE¹, verbe

I. – *Empl. intrans.*

A. – [Le suj. désigne une pers.]

I. 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, rigoler¹, 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 lèvres pour ne pas rire; crever, crouler, éclater, étouffer, s'étouffer, s'étrangler, mourir, pisser, pouffer, se*

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

Morphologie Lexicographie **Etymologie** Synonymie Concordance Aide

Entrez une forme:

Voir la liste des notices étymologiques corrigées Catégorie: toutes

RIRE¹, verbe

Étymol. et Hist. A. I. a) *Ca* 1100 *rire* « marquer un sentiment de gaieté... » (*Roland*, éd. J. Bédier, 302); **b)** *ca* 1350 *rire dessous son chaperon* (*Tristan de Nanteuil*, éd. K. V. Sinclair, 13090); 1643 *rire sous cappe* (SAINT-AMANT, *Rome ridicule*, 840, éd. J. Lagny: la piété ... **rit** sous cappe); 1690 *rire sous barbe* (FUR., s.v. *barbe*); 1694 *rire dans sa barbe* (Ac.); **c)** *ca* 1440 *rire aux anges* (*L'Amant rendu cordelier*, éd. A. de Montaiglon, 456); 1441-47 (P. DE HAUTEVILLE, *La Confession et Testament de l'amant trespasé de deuil*, éd. R. M. Bidler, 894); 1867 (DELVAU, p. 427: **Rire** aux anges. Sourire doucement en dormant); **d)** 1640 (OUDIN *Curiositez*: il **rit** jaune comme farine); 1740-55 *rire jaune* (cf. *jaune*); **2.** *ca* 1165 « se moquer » (CHRÉTIEN DE TROYES, *G. d'Angleterre*, éd. M. Wilmette, 1294: s'i ot moult **ris** et moult gabé); *ca* 1165 *rire de* « se moquer de »

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

Morphologie Lexicographie Etymologie **Synonymie** Concordance Aide

Entrez une forme:

Synonymes de "rire"

| | |
|------------|------------|
| se tordre | ██████████ |
| se moquer | ██████████ |
| railler | ██████████ |
| s'amuser | ██████████ |
| plaisanter | ██████████ |
| se gausser | ██████████ |
| blaquier | ██████████ |

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

Morphologie | Lexicographie | Etymologie | Synonymie | **Concordance** | Aide

Entrez une forme:

Résultat: 1 à 30 sur 249

« 1 2 3 4 5 6 7 8 9 »

ame... *Pomaré satisfaite de sa facétie riait sous cape. Elle avait mis à profit le t
répondre ; elle détournait la tête, et riait sous les plis de la mousseline... j'éca
amies ; elle avait essuyé ses pleurs et riait aux éclats. Elle ne parlait point franç
ut à fait comme il faut, dit *Yves, qui riait assez rarement. Mais je vais vous dire,
ace de dents, plus rien, et, quand elle riait, on voyait à la place ses gencives rond

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 :

http://www.cnrtl.fr - Concordance - Mozilla Firefox

■ Bibliographie

Titre Le Mariage de Loti : Rarahu
Auteur Pierre LOTI
Année 1882
Edition Paris : Calmann-Levy, 1891.

■ Concordance

profondeurs de la montagne, comme un orchestre formidable soulignant la situation tendue d'un mélodrame... *Pomaré satisfaite de sa facétie riait sous cape. Elle avait mis à profit le trouble qu'elle venait d'occasionner pour marquer deux fois té téné (l'homme), c'est-à-dire le roi...

Le portail lexical permet également, à partir d'un simple double-clic sur un mot, une hypernavigation 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'hypernaviguer vers les informations lexicales de ce mot :

Morphologie | Lexicographie | Etymologie | Synonymie | **Concordance** | Aide

Entrez une forme:

Résultat: 1 à 30 sur 249

« 1 2 3 4 5 6 7 8 9 »

ame... *Pomaré satisfaite de sa **facétie** riait sous cape. Elle avait mis à profit le t
répondre ; elle détournait la tête, et riait sous les plis de la mousseline... j'éca
amies ; elle avait essuyé ses pleurs et riait aux éclats. Elle ne parlait point franç
ut à fait comme il faut, dit *Yves, qui riait assez rarement. Mais je vais vous dire,
ace de dents, plus rien, et, quand elle riait, on voyait à la place ses gencives rond
eux fiancés, qui dansaient à l'air très bon, en les voyant tous
ueneau, sanglotant. ah ! Monsieur, on riait ! On riait ! *Cyrano oui, ma vie

Chercher 'facétie' en: Fermer

- morphologie
- lexicographie
- etymologie
- synonymie
- concordance

Conclusion

Le traitement informatique de la langue française s'inscrit dans un contexte plus large du traitement automatique des langues, un des domaines incontournables dans notre société de l'information. Il nécessite le partage et la mutualisation des ressources textuelles et lexicales et d'outils de traitement pour permettre un outillage de notre langue comparable à celui existant pour l'anglais. C'est sur cette base que le CNRS a créé 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/article-5133166.html>). La généralisation de telles exploitations et valorisations de ressources linguistiques modifie les modes de travail et d'échanges scientifiques au sein des communautés de recherche en traitement automatique des langues. En ce domaine Nancy a su prendre une place de choix conforté récemment par la décision d'y structurer un équipement d'excellence ORTOLANG (Open Resources and TOols for LANGuage : www.ortolang.fr) dont le but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui :

- permette, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international ;
- facilite l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels, en particulier vers les PME qui souvent ne peuvent pas se permettre de développer de telles ressources et outils de traitement de la langue compte tenu de leurs coûts de réalisation ;
- valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics.

Notes

- [1] &msubs permet de tester toutes les formes d'un substantif, de même que &cverbe toutes les formes d'un verbe
- [2] FRANTEXT : base de données textuelles de l'ATILF riche de plus de 4 000 œuvres, accessible à l'adresse www.atilf.fr/frantext.



Bibliographie

- ATILF *Trésor de la Langue Française informatisé*, CNRS Editions, 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, 2004, Version Mac OS 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. et Pierrel J.M. « Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *Traitement Automatique des Langues*, Vol 44 – n° 2/2003, Hermes Sciences Edition, p. 11-37.
- 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
- Pierrel J.M. *Ingénierie des Langues*, *Traité Information - Commande - Communication*, Hermès, octobre 2000.
- Pierrel J.M. *À la découverte d'un trésor : le mariage de l'informatique et de la lexicographie au service de la valorisation de la langue française Mémoires de l'Académie de Stanislas 8e série, Tome XXIV*