Linguistique de corpus et Traitement Automatique de la Langue.

Jean-Marie Pierrel

Communication et connaissance:

Supports & médiations à l'âge de l'information

Coordinateur : Jean-Gabriel Ganascia, Professeur à Pierre et Marie Curie (Paris VI)

Ont participé à cet ouvrage Nicolas Balacheff, Michel Beaudoin-Lafon, Pierre Beauvillain, Danièle Bourcier, Didier Bourigault, Philippe Breton, Jean Caelen, Francesco Cara, Pierre Chavel, Claude Chappert, Michel de Rougemont, Gilbert de Terssac, Didier Decoster, Rose Dieng, Philippe Dolffus, Serge Fdida, Patrick Flandrin, Patrick Gallinari, Christine Gaspin, Line Garnero, Marie-Claude Gaudel, Jean-Michel Hoc, Christian Jacob, Philippe Jorrand,Olivier Joubert, Gérard Loiseau, Antonio Munoz-Yague, Jean-Marie Pierrel, Brigitte Plateau, Raymond Quéré, François Rastier, Paul-Allain Rolland, Joseph Saillard, Jean-Michel Salaün, Jean Pierre Sanchez, Michel Scholl et Denis Trystram.

Linguistique de corpus et Traitement Automatique de la Langue

Jean-Marie Pierrel, professeur à l'Université Henri Poincaré
ATILF (UMR 7118 / CNRS - Université Nancy 2 - Université Henri Poincaré.)
Analyse et **T**raitement **I**nformatique de la **L**angue **F**rançaise
44, avenue de la Libération
BP 30687
54063 Nancy cedex
Jean-Marie.Pierrel@atilf.fr

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 (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 et col. 1995). Ainsi se structura petit à petit un nouveau champ de recherche : la *linguistique de corpus* (Habert et col. 1997) permettant au linguiste d'aller au-delà de l'accumulation de faits de langue et de confronter ses théories à l'usage effectif de la langue. Parallèlement, les besoins applicatifs ont conduit à de nombreux travaux en *TAL*.

Aujourd'hui, traitement automatique des langues et linguistique de corpus structurent un nouveau champ disciplinaire aux finalités multiples, en particulier :

- modélisation de la langue, de sa structure et de son usage conduisant la linguistique à des exigences d'opérationnalité effective sur les formes d'usage de la langue, par opposition aux exemples construits, encore trop souvent utilisés en linguistique ;
- mise en place d'applications concrètes : indexation et accès à l'information, résumé de textes, extraction de connaissances, dialogue homme-machine, par exemple.

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 telle 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.

1. Quels ressources et corpus pour l'étude des langues aujourd'hui ?

1.1. Corpus textuels

Le premier type de ressources, indispensables pour le développement de nombreuses études sur la langue, son analyse et son traitement, concerne les corpus textuels et les corpus d'exemples. Leur rôle est en effet central pour permettre la construction de modèles. Cette activité de modélisation a comme objectif premier de proposer, parfaire et évaluer des modèles opératoires ou des théories linguistiques 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'activités 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 on peut à juste titre s'interroger sur la fiabilité des ressources textuelles qu'on y trouve ! Deux travers de taille caractérisent les textes disponibles sur le Web :

(i) 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.

(ii) 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 pour notre domaine de recherche 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*

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

secteur[2]. Le nombre de projets soumis sur ce volet, en association entre des chercheurs et des industriels, montre l'importance de ce thème. 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 (industriels, professionnels ou grand public), la nécessité de vastes corpus normalisés, annotés et validés s'impose.

1.2. 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 convaincu 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 (ATILF 2004 ; www.tlfi.fr) et d'en dériver un lexique ouvert des formes fléchies du français (540 000 formes issues de 68 000 lemmes : www.atilf.fr/morphalou).

[2] Le coût de développement des ressources textuelles est important et demeure souvent un frein pour de nombreuses études sur notre langue.

1.3. Des outils de traitements partagés

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

(i) 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 ?

(i)(ii) Les outils de base de traitement de la langue. Indispensables pour permettre à une équipe de recherche ou de développement de proposer des avancées sur tel ou tel point spécifique, ils doivent devenir disponibles pour notre langue, ce qui permettrait d'éviter de réinventer sans cesse la roue dans des domaines tels que la lemmatisation, la conjugaison ou l'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 re-développer de tels outils !

2. Les enjeux actuels de la linguistique de corpus

2.1. Œuvrer à une meilleure connaissance et modélisation de notre langue, de son lexique, de sa structure et de son fonctionnement.

La modélisation, qui est au cœur de toute activité de recherche, tant en informatique qu'en linguistique, a pour but premier, dans notre domaine, de proposer ou parfaire des théories linguistiques et des modèles informatiques qui soient tout à la fois opératoires et valides d'un point de vue de l'usage de la langue. Pendant longtemps, pour atteindre ces objectifs de modélisation de la langue (informatique et/ou linguistique) et faire émerger des invariants ou, au contraire, des comportements particuliers d'entités linguistiques, la recherche s'est

appuyée sur les connaissances qu'a le chercheur de cet objet multiforme qu'est la langue (aux niveaux phonétique, phonologique, morphologique, syntaxique ou sémantique). Aujourd'hui, que cela soit en informatique ou en linguistique, l'accent est mis de façon plus forte sur l'usage effectif de la langue tel qu'il peut être analysé à travers l'exploitation de vastes corpus textuels représentatifs d'un domaine applicatif ou plus généralement d'un usage nouveau tel qu'il apparaît sur le Web. Ce courant fut initié dès la fin des années 60 en lexicographie à travers le projet de dictionnaire du Trésor de la Langue Française, aujourd'hui disponible sous forme électronique (www.tlfi.fr), premier dictionnaire de langue se fondant sur une méthodologie systématique d'analyse des usages effectifs des mots de notre langue à travers l'exploitation d'une vaste base de données textuelles dont le but premier était de fournir, à travers des concordances, des données organisées aux rédacteurs du dictionnaire. Cette base de données textuelles, enrichie et mise à jour, a donné naissance à FRANTEXT (www.atilf.fr/frantext), sans aucun doute le plus grand corpus diachronique sur la langue française (220 millions d'occurrences de mots, soit plus de 1,5 milliards de caractères), support aujourd'hui de nombreuses recherches en linguistique de corpus.

Au cours des dernières années, parmi les résultats les plus remarquables, il convient de noter ceux obtenus aux niveaux lexical et terminologique (création de lexiques multilingues), morphologique et morphosyntaxique (en particulier étiquetage morphosyntaxique de textes pour lever des ambiguïtés de formes telle *portes*, substantif ou verbe), ou syntaxique (analyse syntaxique robuste d'énoncés) ainsi que leur exploitation soit comme prétraitement pour un accès au contenu de données textuelles, soit comme aide à l'enseignement des langues assistée par ordinateur.

2.2. Vers une véritable exploitation du contenu informationnel le plus souvent sous formes langagières et textuelles

Les aspects d'acquisition, de gestion, de structuration, d'analyse et d'interprétation, de modélisation et d'exploitation des informations et connaissances, pour la plupart sous formes langagières et textuelles, sont au centre des grands débats du monde de la recherche ou de l'économie. La linguistique de corpus et le TAL (ou *ingénierie des langues)* sont ainsi devenus des domaines-clés répondant aux besoins actuels de notre société de l'information.

Les principales activités visées sont les industries de la langue, l'édition numérique, la veille technologique, mais aussi la gestion de patrimoines scientifiques et techniques (IST), industriels (mémoire d'entreprise), linguistiques ou culturels et leurs exploitations à travers les vastes réseaux aujourd'hui disponibles et de plus en plus accessibles par chacun de nous. A ce niveau, les résultats les plus marquants concernent l'usage de méthodes statistiques pour la détermination ou la sélection de thèmes (pour faciliter l'accès au contenu) ou de genres textuels (pour une classification automatique de textes). Pour l'avenir, l'un des domaines les plus importants concerne sans aucun doute les modélisations sémantiques et les procédures d'accès au contenu telles qu'elles émergent aujourd'hui dans ce que l'on qualifie de *web sémantique* et qui contribuent à l'amélioration des accès aux informations scientifiques, techniques et culturelles le plus souvent sous formes textuelles.

2.3. La normalisation incontournable pour échanger et confronter des données et résultats sur le fonctionnement de la langue

L'analyse et le traitement informatique de la langue est en effet aussi un outil indispensable aux recherches en linguistique pour échanger des données structurées sur les langues (données brutes, annotées, alignées – dans le cadre de recherche multilingue – ou des résultats d'analyses codés dans un format normalisé, comme par exemple XML). L'une des caractéristiques essentielles de la recherche, à laquelle la linguistique et le TAL n'échappent pas, est en effet cette nécessité de permettre à la communauté scientifique de pouvoir échanger, évaluer, confronter et reproduire des résultats de recherche ou d'analyse. Cela nécessite le développement et l'usage de normes pour les ressources linguistiques et textuelles. La communauté française est fort bien placée en ce domaine : dans le cadre du comité technique 37 de l'ISO, le sous-comité dédié aux ressources linguistiques et à leur normalisation est présidé par un français, et dans le cadre de la Text Encoding Initiative, consortium international qui définit des recommandations de codage de ressources textuelles

(www.tei-c.org), la France, à travers une coopération entre plusieurs laboratoires, est devenue le centre européen support de cette initiative[3].

3. Un devoir vis-à-vis du traitement de notre langue

Nous disposons en France d'équipes de recherche de qualité en linguistique de corpus et en TAL et, s'il est vrai qu'un des objectifs premier de la linguistique reste de rechercher des invariants sur le fonctionnement des diverses langues, il est nécessaire, pour permettre au français de rester présent dans ce champ fortement internationalisé, de poursuivre le développement de recherches, outils et ressources liés plus spécifiquement à notre langue. Il apparaît donc naturel et nécessaire de développer en France des études plus spécifiques sur la langue française, pour elle-même (la langue étant fortement liée aux notions de culture et de communication d'une société) et en interaction avec des langues partenaires (pour faciliter les échanges entre sociétés différentes), mais aussi pour permettre au français de rester présent, à côté de la langue dominante qu'est aujourd'hui l'anglais, dans le plus grand nombre de champs applicatifs de notre société de l'information (gestion d'informations et accès par le contenu, traduction assistée par ordinateur, industries de la langue, etc.). C'est, nous semble-t-il, l'un des enjeux incontournable pour les années à venir si l'on souhaite maintenir et renforcer l'usage du français dans un contexte de mondialisation de plus en plus marqué.

Références générales

ATILF (ouvrage collectif publié sous le nom du laboratoire) *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, novembre 2004, Version Mac OX X, ISBN 2-271- 06365-5, septembre 2005

Coandamines Anne (Ed.) *Sémantique et corpus*, Traité IC2, Paris, Hermès Lavoisier, 2005

Daille B. et Romary L. (Eds.) Linguistique de Corpus, *Traitement Automatique des Langues* Vol 42- N°2, Paris, Hermès, 2001

Habert B. Traitements probabilistes et Corpus, *Traitement Automatique des Langues* Vol 36- N°1-2, Paris, Hermès, 1995

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

[3] Ce centre support européen de la TEI, officialisé en novembre 2004, s'est constitué sur Nancy entre l'ATILF (laboratoire de linguistique), l'INIST (Unité spécialisée en IST) et le LORIA (laboratoire d'informatique).

Pierrel J.M. (Ed.) *Ingénierie des langues*, Traité IC2, Paris, Hermès Lavoisier, 2000