

Syntex, analyseur syntaxique de corpus

Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques & Sylwia Ozdowska

ERSS – CNRS & Université Toulouse le Mirail
5, allées Antonio Machado, 31 058 Toulouse Cedex 9
{didier.bourigault,cfabre,frerot,mpjacques,ozdowska}@univ-tlse2.fr

Résumé

Cet article est un document de présentation de l'analyseur syntaxique de corpus Syntex, dans lequel nous décrivons les principes à la base du développement de l'analyseur et son architecture informatique. Une bibliographie du projet SYNTEX est donnée à la fin du document.

1 Analyseur de corpus

L'analyseur SYNTEX a été développé à l'origine (Bourigault, Fabre, 2000) pour remplacer le logiciel *LEXTER*¹, un analyseur syntaxique robuste dédié au repérage des syntagmes nominaux dans les corpus spécialisés et utilisé dans des applications de construction de terminologies ou d'ontologies spécialisées. Les diverses expérimentations réalisées avec *LEXTER* avaient mis en évidence la nécessité d'étendre la couverture du logiciel à l'extraction des syntagmes verbaux. A partir de ce constat, nous avons décidé d'entreprendre à l'ERSS la réalisation d'un nouvel analyseur, avec l'objectif d'en faire un outil opérationnel d'analyse syntaxique de corpus, utilisable dans différents contextes applicatifs, dont la construction de ressources lexicales spécialisées pour des systèmes de traitement de l'information (Bourigault *et al.*, 2004 ; Ozdowska *et al.*, 2005). SYNTEX doit traiter des corpus de phrases réelles, de taille importante (de quelques centaines de milliers à plusieurs millions de mots). Ceci impose des contraintes d'efficacité (temps de traitement), de robustesse (tolérance aux malformations syntaxiques et aux mots ou structures inconnues, possibilité de rendre des analyses partielles et incomplètes) et d'adaptabilité (prise en compte de certaines propriétés syntaxiques particulières des mots dans des corpus spécialisés). Les principes de base de l'analyseur sont les suivants : Syntex analyse des corpus préalablement étiquetés (section 2), il effectue une analyse syntaxique en dépendance (section 3), il est organisé sous la forme d'un enchaînement de modules de reconnaissance de relations

¹ BOURIGAULT D. (1994), *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

syntaxiques (section 4) et il exploite de façon combinée des procédures d'apprentissage endogène et des ressources lexico-syntaxiques de sous-catégorisation (section 5).

2 Etiquetage préalable

L'organisation du partage des tâches entre étiquetage morphosyntaxique (attribution d'une étiquette morphosyntaxique aux mots de la phrase) et analyse syntaxique (identification de constituants syntaxiques ou de relation de dépendance syntaxique) est un problème délicat. Disposer des étiquettes morphosyntaxiques des mots pour identifier les relations syntaxiques est extrêmement pratique. Mais, dans certains cas, la levée d'ambiguïtés catégorielles exige une analyse syntaxique partielle du contexte large. Le problème reste ouvert. Notre choix a été de séparer nettement les deux tâches et de confier la tâche préalable d'étiquetage à un outil extérieur. Même s'il y a interdépendance forte entre étiquetage et analyse, quantitativement l'analyse syntaxique a beaucoup plus à profiter de l'étiquetage que l'inverse. Des outils d'étiquetage de bonne qualité sont disponibles pour le français. SYNTAX prend en entrée les résultats du Treetagger², développé à l'Université de Stuttgart. Treetagger est un étiqueteur efficace et robuste. Il présente l'intérêt fondamental d'être ouvert, en ce sens qu'il est possible de faire en amont, à sa place, une partie du travail de tokénisation et d'étiquetage. Nous avons développé des procédures (lexiques, règles) de reconnaissance d'unités syntaxiques complexes qui viennent poser sur le corpus des étiquettes sur lesquelles le Treetagger s'appuie pour étiqueter les mots environnants. Nous avons aussi introduit dans la chaîne de traitement la possibilité d'intégrer un fichier de règles de tokénisation et de pré-étiquetage, données sous la forme d'expressions régulières, spécifiques au corpus à analyser. Cette fonctionnalité est essentielle quand il s'agit de traiter des corpus comportant des mots inconnus ou des structures « bizarres » (codes de produits, nomenclature d'éléments chimiques, etc.). Enfin, la frontière entre étiquetage et analyse n'est pas étanche. Dans certains contextes syntaxiques, l'analyseur effectue des retours en arrière sur l'étiquetage en venant modifier des étiquettes attribuées par le Treetagger (Jacques, 2005).

3 Analyse en dépendance

SYNTAX effectue une analyse en dépendance. Nous ne nous basons sur aucune théorie syntaxique particulière et nous n'avons pas élaboré une grammaire de dépendance spécifique pour ce projet. Notre base d'appui est la grammaire traditionnelle, tant au niveau des catégories morphosyntaxiques que des relations syntaxiques. Les principales relations syntaxiques actuellement reconnues sont les suivantes : sujet, objet direct, complément prépositionnel (de nom, de verbe et d'adjectif), antécédence relative, modification adjectivale (épithète, attribut), subordination. Les théories syntaxiques ou les descriptions linguistiques sont utiles pour définir des modes de représentation des relations (pour telle structure complexe, quel est le recteur, quel est le régi et dans quel sens s'établit la relation de dépendance, comment représenter les dépendances dans le cas des complexes verbaux et dans des structures discontinues, comme les structures comparatives, etc.). En revanche, pour faire de la syntaxe opérationnelle, c'est-à-dire pour écrire des règles de repérage de relations syntaxiques dans une chaîne étiquetée, le recours

² <http://www.ims.uni-stuttgart.de>

aux théories et descriptions syntaxiques est moins nécessaire. En particulier, le traitement des coordonnants et des virgules (apposition, incise, coordination), qui foisonnent dans les textes réels, exigent le développement de procédures d'analyse complexes, qui empruntent peu aux descriptions linguistiques classiques.

4 Architecture modulaire séquentielle

Nous décomposons le problème de l'analyse syntaxique d'une phrase en sous-problèmes élémentaires du type : soit m un mot de catégorie C dans la phrase étiquetée S , quel est le recteur syntaxique de m dans S ? De façon simplifiée, la résolution de ce problème s'effectue par un enchaînement en cascade d'une suite de modules qui prennent en charge chacun une relation syntaxique. Chaque module prend en entrée les sorties du module précédent. Cette organisation séquentielle des traitements impose de choisir un ordre dans l'analyse. On est face à un dilemme du type de celui du partage entre étiquetage et analyse. Par exemple, faut-il reconnaître les relations sujet avant de chercher à identifier les relations de coordination, ou faire l'inverse, ou répartir le traitement à deux moments de la chaîne ? Le choix de l'ordre est un choix difficile qui a un impact fort sur la programmation des différents modules, et sur lequel il est de plus en plus difficile de revenir au fur et à mesure que l'analyseur s'enrichit et se complexifie. A l'intérieur même de la chaîne d'analyse, les retours en arrière sont là aussi possibles, certains modules venant détruire et remplacer des relations syntaxiques posées par des modules antérieurs. Les modules sont constitués d'un ensemble d'heuristiques de parcours de la chaîne étiquetée et partiellement analysée, qui partent d'un régi (resp. recteur) potentiel pour aboutir à son recteur (resp. régi). Ils sont développés « à la main » par des linguistes informaticiens (dans le langage Perl), selon une méthode qui met en œuvre le recours à la connaissance grammaticale et à des tests nombreux et variés sur des corpus diversifiés.

5 Ressources lexicales

L'analyseur SYNTEX est peu (mais de plus en plus) lexicalisé. Nous avons fait le choix initial de la table rase. Contrairement aux approches qui choisissent, pour réaliser un analyseur syntaxique, de développer au préalable un lexique syntaxique très riche recensant les propriétés syntaxiques des mots de la langue, nous avons commencé sans aucune information de ce type. Cette approche est possible à partir du moment où l'on a choisi de s'appuyer sur les résultats d'un étiqueteur (on bénéficie indirectement des ressources lexicales éventuellement exploitées par celui-ci). Des informations lexicales sont intégrées dans l'analyseur au fur et à mesure des besoins : liste de locutions prépositionnelles, liste de verbes transitifs, liste de verbes se construisant avec des compléments en *que*, en *de*, etc. Pour résoudre les ambiguïtés de rattachement prépositionnel, l'analyseur exploite des informations de sous-catégorisation associées aux couples (mot, préposition). Depuis l'origine de nos travaux sur l'analyse syntaxique, ces informations sont acquises de façon endogène sur le corpus en cours de traitement. Les expériences menées sur de nombreux corpus spécialisés ont montré que ces corpus renferment des spécificités lexicales, en particulier que certains mots, fréquents dans le corpus, manifestent des comportements syntaxiques spécifiques et imprédictibles. C'est pourquoi, nous avons porté nos efforts depuis une dizaine d'années sur le développement de procédures d'apprentissage endogène sur corpus qui permettent à l'analyseur d'acquérir lui-même, par analyse du corpus à traiter, des informations de sous-catégorisation spécifiques à ce corpus. Devant les limites inhérentes à l'exploitation exclusive d'informations de sous-

catégorisation endogènes, nous travaillons à l'élaboration de ressources générales, susceptibles d'être exploitées pour tout corpus (Frérot *et al.*, 2003). Nous avons expérimenté l'utilisation d'un lexique de sous-catégorisation construit à la main à partir des tables du Lexique-Grammaire (Frérot, à paraître), puis de lexiques construits automatiquement à partir de corpus. Dans son état actuel, l'analyseur exploite un lexique de probabilités de sous-catégorisation construit à partir d'un corpus de 200 millions de mots (Bourigault, Frérot, 2005).

Bibliographie du projet Syntax

BOURIGAUT D., AUSSENAC-GILLES N., CHARLET J. (2004), Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, *Revue d'Intelligence Artificielle (RIA)* , « Techniques Informatiques et structuration de terminologies », PIERREL J.-M. et SLODZIAN M. (Ed.) , Paris : Hermès. Vol. 18, n°1/2004, pp. 87-110

BOURIGAUT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, Université Toulouse le Mirail, pp. 131-151.

BOURIGAUT D., FRÉROT C. (2005), Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France, juin 2005

FRÉROT C. (2005), *Etude en corpus variés de l'intégration de ressources linguistiques générales dans un analyseur syntaxique*, Thèse en sciences du langage de l'Université Toulouse le Mirail

FRÉROT C., BOURIGAUT D., FABRE C. (2003), Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in *TAL*, 44-3

JACQUES M.-P. (2005), Que : la valse des étiquettes. *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France,

OZDOWSKA S., BOURIGAUT D. (2004) Détection de relations bilingues entre termes à partir d' une analyse syntaxique de corpus *Actes du 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, RFIA'04*, Toulouse, France, janvier 2004

OZDOWSKA S., CLAVEAU V. (2005) Alignement de mots par apprentissage artificiel de règles de propagation syntaxique en corpus. *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France

OZDOWSKA S., NÉVÉOL A., THIRION B.. Traduction automatique compositionnelle de bitermes dans des corpus alignés anglais/français. *Actes de la Conférence Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France, avril 2005