Proposition d'Action de Recherche Concertée [ARC]

RLT
Acquisition et représentation de Ressources Linguistiques pour les TAG

Eric de la Clergerie

20 décembre 2000

Note : Ce document est accessible en ligne à `http://atoll.inria.fr/~clerger/ARC00/index.html` en accès protégé (utilisateur `arc`, mot de passe `rlt`).

1 Présentation Scientifique

1.1 Thème et motivations

Dans le cadre du traitement de la langue, l'acquisition de grammaires à large couverture linguistique et des lexiques associés à ces grammaires est une tâche de longue haleine et d'une importance capitale. En pratique, le développement d'applications linguistiques est souvent freiné par le manque de ressources facilement accessibles et exploitables. De plus, la qualité de la représentation des ressources linguistiques est essentielle pour faciliter leur distribution et assurer leur pérennité. L'émergence actuelle de XML (*eXtended Markup Language*) apporte une solution adéquate pour assurer cette qualité de représentation, en particulier en s'appuyant sur les spécifications précises qu'offrent les DTD (*Document Type Definition*).

Nous nous proposons dans cette action d'examiner ces problèmes d'acquisition et de représentation des ressources dans le contexte des grammaires d'arbres adjoints (**TAG** – *Tree Adjoining Grammars*). Ce formalisme grammatical est apprécié des linguistes informaticiens car il permet des représentations élégantes de nombreux phénomènes linguistiques. Sur le plan informatique, il possède aussi de bonnes propriétés, en particulier celle de pouvoir être analysable en temps polynomial (pour le modèle de base des TAG). Divers projets de grammaires TAG à large couverture linguistique sont actuellement en cours de développement pour l'anglais (groupe XTAG [1] [7]) et pour le français (équipe TALaNa [2]). Ces grammaires couvrent un large éventail de phénomènes linguistiques mais sont aussi couplées à des lexiques importants.

Les équipes participant à cette proposition sont déjà toutes impliquées, à des degrés divers, dans l'étude et le traitement des TAG ainsi que dans les représentations XML de ressources linguistiques. L'action envisagée est l'occasion de mettre en commun leurs efforts pour définir, réaliser et valider un scénario d'acquisition de ressources qui nécessite de regrouper une forte expertise linguistique et de nombreux outils informatiques.

1.2 Scénario

Le scénario d'acquisition que nous souhaitons définir au cours de cette action s'articule autour des points présentés dans cette partie.

[1] `http://www.cis.upenn.edu/~xtag/`

1.2.1 Grammaires et méta-grammaires

Suivant l'architecture XTAG [7], la grammaire **FTAG** du français développée par TALaNa lie un ensemble d'environ 5000 arbres, organisés en familles, avec un lexique syntaxique de *lemmes* (par exemple le verbe `/DONNER/`) et un lexique de *formes fléchies* (par exemple la forme `donne`), comme illustré par la figure 1. Les liaisons entre formes fléchies, lemmes et arbres sont actuellement construites manuellement (ou presque), ce qui représente un effort considérable et freine le développement de **FTAG**.

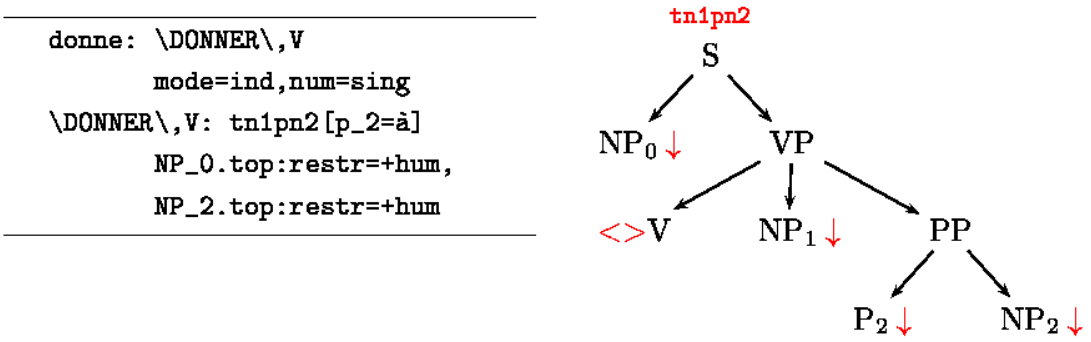

FIG. 1: Liens entre forme fléchie, lemme et arbre

Les problèmes posés par cette approche directe sont selon nous de deux ordres. D'une part, les descriptions linguistiques pré-existantes doivent être entièrement retraduites dans le formalisme des TAG, c'est-à-dire, en pratique, entièrement refaites ; d'autre part, étant donnée la liaison forte entre la grammaire (les arbres TAG) et le lexique, toute modification de la grammaire, telle qu'un ajout ou une suppression d'arbre TAG entraîne une modification du lexique sous peine d'incohérence.

Pour répondre à ce problème, Marie-Hélène Candito et Anne Abeillé ont proposé que les arbres et familles de la grammaire **FTAG** soient en fait dérivés à partir d'une *méta-grammaire* [6]. Concrètement, la méta-grammaire est constituée d'un ensemble de classes organisées en hiérarchies ; chaque classe spécifie des contraintes (de domination, d'ordre linéaire et d'étiquetage entre « quasi-nœuds »). Les arbres de la grammaire sont engendrés sont engendré par croisement de plusieurs classes correspondant à plusieurs dimensions linguistiques (valence, redistribution de fonctions et réalisation des fonctions). Par ce mécanisme, pour autant que les classes de la méta-grammaire correspondent à des phénomènes linguistiques pertinents, les arbres obtenus sont étiquetés par les classes dont ils dérivent et donc par les phénomènes linguistiques qu'ils encodent.

De ce fait, le lien entre le lexique et les arbres de la grammaire peut se faire indirectement en associant aux lemmes les phénomènes qu'ils admettent. (ex : *donner* : di-transitif [N0VN1àN2], passivable, ...). Cette fois, une modification de la grammaire (c-à-d des classes) aboutissant à la disparition ou à l'ajout d'un ou plusieurs arbres ne donne pas lieu à une modification du lexique.

Le processus de dérivation des arbres est cependant sous-spécifié : certains arbres peuvent être trop généraux, certains sont probablement inutiles et certains phénomènes ne sont pas nécessairement entièrement couverts. Il en est de même pour les liens avec le lexique où il est nécessaire d'identifier pour chaque lemme les constructions qu'il admet et les contraintes supplémentaires qu'il apporte. Des processus d'acquisition sont cependant envisageables pour filtrer les arbres possibles et lier les lemmes aux arbres en examinant les constructions syntaxiques valides obtenues sur un corpus de textes.

Il est à noter que l'utilisation d'une méta-grammaire pour dériver une grammaire est encore un domaine très neuf. De nombreuses questions subsistent sur les aspects linguistiques (comme la caractérisation des classes) mais aussi sur les propriétés formelles d'une telle méta-grammaire. Sur le plan linguistique, on peut ainsi étudier les rapports précis entre méta-grammaire et TAG. Il est possible que la méta-grammaire soit relativement neutre, ce qui permettrait (a) d'incorporer plus facilement des ressources existantes non prévues pour les TAG et (b) des produire des ressources (lexiques) non spécifiques aux TAG. Sur le plan formel, il est intéressant de voir les liens des méta-grammaire avec les langages de spécifications partielles

d'arbre, par exemple les *grammaires d'interactions* [13] développées par Calligramme. Sur un plan informatique, une meilleure caractérisation formelle permet la conception d'une représentation adéquate et le développement d'outils informatiques efficaces. Cette action est aussi l'occasion d'examiner ces diverses questions.

1.2.2 Constitution de corpus annotés

Traditionnellement, l'acquisition de ressources se fait à partir de corpus manuellement annotés, qui sont bien évidemment rares à trouver et fastidieux à construire. Nous voulons plutôt privilégier la constitution automatique de corpus annotés par analyse syntaxique de corpus de textes. Cela nécessite des analyseurs syntaxiques efficaces pour les TAG, ce qui est le cœur des activités actuelles de l'équipe Atoll [5, 12, 4] et, dans une moindre mesure, de « Langue et Dialogue » [11].

Comme, par hypothèse, les lexiques et les grammaires ne sont pas connus, nous serons obligés d'être sur-générateurs. Par exemple, lorsque nous ne saurons pas déterminer à quelle catégorie linguistique précise appartient un verbe, nous devrons envisager toutes les catégories verbales possibles. D'autre part, certains mots et constructions syntaxiques seront inconnus. Cela nécessite des analyseurs syntaxiques robustes (n'échouant pas complètement) et capables de gérer un nombre important d'analyses. Nous nous appuierons donc sur les forêts de dérivation produites par les analyseurs d'Atoll pour garder factorisées les ambiguïtés localement non pertinentes pour un arbre élémentaire TAG donné.

Il faut également étudier et normaliser la nature des annotations, qui peuvent décrire des structures syntaxiques (catégories syntaxiques, arbres ancrés par un mots, frontières entre constituants, relations entre constituants, ...) ou plus généralement des phénomènes linguistiques (extraction, construction ergative, ...).

Enfin, si l'analyse syntaxique est au coeur de ce que nous proposons pour annoter les corpus, il faut noter que d'autres outils de traitement linguistiques sont nécessaire en amont, tels des ségmenteurs (découpe des documents en phrase et mots) et surtout des étiqueteurs (*taggers*). Le rôle des ces étiqueteurs est ainsi de restreindre les catégories syntaxiques (nom, verbe, ...) possibles pour un mot. Il y aura donc un travail de recensement des outils nécessaires.

1.2.3 Création de ressources

Une fois disponibles les corpus annotés, il faut faire émerger les ressources linguistiques candidates. Par rapport à des corpus manuellement annotés et donc à priori sans erreur, des corpus automatiquement annotés vont être incomplets, présenter des ambiguïtés et comporter des erreurs. L'émergence des ressources va donc nécessiter l'emploi de méthodes statistiques pour passer outre ces limitations des corpus. Enfin, comme le processus d'émergence ne sera pas non plus parfait, les ressources candidates devront ensuite être soumises à validation par un linguiste. Lors de cette phase de validation, le linguiste devra pouvoir accéder aux exemples ayant conduit à proposer la ressource. Pour lier un lemme à des arbres ou des classes linguistiques, cet examen peut inclure les phrases où apparaît les formes de ce lemme mais aussi les arbres d'analyse des phrases en question et les annotations du corpus. Nous conjecturons que les choix proposés seront généralement corrects, évitant au linguiste d'écrire manuellement la description de la ressource (et éventuellement d'introduire des coquilles). Pour les cas litigieux restants, l'examen des exemples permettra d'accélérer la prise de décision.

Après validation d'un nouveau jeu de ressources, il est intéressant de réitérer le processus d'acquisition, jusqu'à convergence éventuelle. En effet, au fur et à mesure des itérations, les grammaires et lexiques deviennent de moins en moins sur-générateurs, introduisent moins de bruit et permettent ainsi l'émergence de nouvelles ressources.

1.2.4 Représentation

Au cours des différentes étapes, il est important de pouvoir représenter et échanger les grammaires (arbres et lexiques) ainsi que les corpus annotés. Différents outils et interfaces sont également nécessaires pour la visualisation (des corpus, des grammaires, des analyses, des lexiques, ...) et de maintenance (cohérence des grammaires et de leur lexique).

Ces sujets vont dans le sens de travaux antérieurs réalisés par les équipes participantes (TALaNa, Atoll et « Langue et Dialogue ») dans le groupe de travail TAGML [2], qui cherche à promouvoir des représentations XML pour les TAG ainsi que les outils associés. Elles complètent également l'approche de l'équipe « Langue et Dialogue », également fondée sur l'emploi de XML, dans les projets de gestion de corpus (Silfide [3] [14] et Elan) et de lexiques-dictionnaires (DHYDRO [8]).

1.3 Programme

En l'état actuel, il est envisagé de consacrer la première année à préciser et à définir les différentes étapes du scénario présenté ci-dessus. Une liste non limitative inclut les tâches suivants :

- Constitution ou sélection d'un corpus de textes
- Recensement des outils nécessaires et de ceux disponibles
- Etude du rôle de la méta-grammaire
- Définition des représentations
- Définition du processus de constitution des corpus annotés
- Définition du processus d'émergence des ressources candidates
- Définition de l'interface de validation des ressources

Ces différents points donneront lieu à des expériences ponctuelles (non incluses dans une chaîne complète de traitement) servant à valider certains choix.

L'objectif à la fin de cette première année est disposer (a) d'un scénario précis pour la mise en place d'un prototype et (b) de certains des composants de ce prototype.

La deuxième année sera consacrée à la réalisation de ce prototype et à sa validation par la constitution d'un jeu de ressources.

1.4 Résultats attendus

Les résultats attendus à la fin de cette action sont :

1. des représentations normalisées pour les différents types de ressources manipulées (méta-grammaire, grammaires, corpus, lexiques) ;
2. l'amélioration des techniques d'analyse syntaxique pour les TAG ;
3. une meilleure compréhension des méta-grammaires pour les TAG ;
4. une méthodologie pour l'acquisition supervisée de ressources linguistiques, si possible fondée sur l'emploi d'une méta-grammaire ;
5. un prototype de chaîne d'acquisition ;
6. l'acquisition d'un premier jeu de ressources validant notre approche.

Une telle action est pluridisciplinaire, requérant des compétences linguistiques importantes, fournies par TALaNa, et des outils et techniques informatiques variés (traitement linguistique, gestion de corpus, gestion de bases de données, interfaces de traitement pour les linguistes), fournis par les équipes INRIA impliquées. Il est important de noter que le but essentiel de cette action n'est pas de construire un jeu particulier de ressources pour le français mais de dégager une méthodologie et des prototypes d'outils (de construction et de diffusion) pouvant être adaptés pour d'autres langues et d'autres formalismes linguistiques.

2 Participants

Equipe coordinatrice : Atoll

[2] `http://www.loria.fr/~lopez/TAG_XML/`
[3] `http://www.loria.fr/projets/Silfide/Index.html`

2.1 Atoll (INRIA Rocquencourt)

Page Web : `http://www.inria.fr/Equipes/ATOLL-fra.html`
Membres de l'équipe impliqués :

- Éric de la Clergerie (Chargé de Recherche INRIA)
  coordinateur (`Eric.De_La_Clergerie@inria.fr`)
- Pierre Boullier (Directeur de Recherche INRIA)
- Philippe Deschamp (Chargé de Recherche INRIA)
- François Barthélemy (Maître de conférence au CNAM)

L'équipe Atoll apporte son expérience dans la construction d'analyseurs syntaxiques pour les TAG (au travers de l'approche RCG [5] et du système DyALog [4]). Elle travaille également sur des formats XML de représentation des grammaires TAG [4] et des forêts de dérivations [5].

2.2 Calligramme (LORIA)

Page Web : `http://www.inria.fr/Equipes/CALLIGRAMME-fra.html`
Membres de l'équipe impliqués :

- Guy Perrier (Maître de conférence à l'université de Nancy 2)
  coordinateur (`perrier@loria.fr`)
- Philippe de Groote (Chargé de Recherche INRIA)
- François Lamarche (Directeur de Recherche INRIA)

L'équipe Calligramme travaille sur les liens entre logique linéaire et analyse syntaxique de la langue. Dans cette perspective, Guy Perrier est à l'origine du formalisme des grammaires d'interaction [13]. Dans ce formalisme, le processus d'analyse syntaxique est le calcul de modèles de descriptions partielles d'arbres. Comme pour les TAG, le lexique peut être dérivé d'une méta-grammaire structurée sous forme de descriptions d'arbres. L'équipe Calligramme apportera son expérience à ce sujet et s'intéressera particulièrement à la formalisation précise de la méta-grammaire.

2.3 Langue et Dialogue (LORIA)

Page Web : `http://www.inria.fr/Equipes/LANGUEETDIALOGUE-fra.html`
Membres de l'équipe impliqués :

- Bertrand Gaiffe (Chargé de Recherche CNRS)
  coordinateur (`Bertrand.Gaiffe@loria.fr`)
- Laurent Romary (Chargé de Recherche CNRS)
- Azim Roussanaly (Maître de conférences à l'université de Nancy 2)
- Samuel Cruz-Lara (Maître de conférences à l'université de Nancy 2)
- Jean-Luc Husson (Maître de conférences à l'université Henri Poincarré Nancy 1)
- Djamé Seddah (Doctorant)

L'équipe « Langue et Dialogue » compte s'impliquer dans l'étude de l'implémentation de la méta-grammaire et sa représentation normalisée. Elle dispose également d'analyseurs syntaxiques TAG qu'elle mettra à disposition du projet. Par ailleurs, elle apportera son expérience et ses outils en ce qui concerne la gestion de ressources de type corpus (projets Silfide et Elan) et lexiques-dictionnaires (projet DHYDRO).

[4] `http://atoll.inria.fr/~clerger/tag.dtd,xml`
[5] `http://atoll.inria.fr/~clerger/forest.dtd,xml`

2.4 TALaNa (Université Paris 7)

Page Web : `http://talana.linguist.jussieu.fr/`
Membres de l'équipe impliqués :

- Anne Abeillé (Professeur)
  coordinateur (`abeille@linguist.jussieu.fr`)
- Laura Kallmeyer (poste ATER)
- Kim Gerdes (Doctorant)
- Nicolas Barrier (Doctorant)
- Sébastien Barrier (Doctorant)
- Alexandra Kinyon (Doctorant)
- Lionel Clement (Doctorant)

L'équipe TALaNa est impliquée depuis longtemps dans la promotion du formalisme TAG en Linguistique Informatique. Elle développe la seule grammaire TAG à large couverture du français et s'intéresse aux problèmes de représentation des grammaires et à leur traitement (analyse, génération, « tagging »). Elle est à l'origine de la notion de méta-grammaire pour les TAG. Elle est déjà impliquée dans les problèmes d'acquisition avec en particulier la constitution du premier corpus pour le français annoté par des informations syntaxiques (indiquant les catégories syntaxiques et les frontières de constituants) [3]. TALaNa a de plus organisé la dernière édition de **TAG+5**, la conférence de la communauté internationale sur les TAG.

Références

[1] Anne Abeillé. Une grammaire électronique du français. à paraître chez CNRS Edition, 2000.

[2] Anne Abeillé, Marie-Hélène Candito, and Alexandra Kinyon. The current state of FTAG. In S. Winter, editor, *ESSLI Workshop*, 2000.

[3] Anne Abeillé, Lionel Clément, and Alexandra Kinyon. Building a treebank for french. In *LREC*, Athenes, 2000.

[4] Miguel Alonso Pardo, Djamé Seddah, and Éric Villemonte de la Clergerie. Practical aspects in compiling tabular TAG parsers. In *Proceedings of the 5^{th} International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5)*, pages 27–32, Université Paris 7, Jussieu, Paris, France, May 2000.

[5] Pierre Boullier. On tag parsing. In *$6^{ème}$ conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN'99)*, pages 75–84, Cargèse, Corse, France, July 1999.

[6] Marie-Hélène Candito. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. PhD thesis, Université Paris 7, January 1999.

[7] Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. XTAG system — a wide coverage grammar for English. In *Proc. of the 15th International Conference on Computational Linguistics (COLING'94)*, pages 922–928, Kyoto, Japan, August 1994.

[8] Jean-Luc Husson, Nadia Viscogliosi, Laurent Romary, Sylviane Descotte, and Marc Van-Campenhoudt. Dhydro : a generic environment developed to edit and access multilingual terminological data on the Internet. In *Second Conference on Maritime Termonology, Turku, Finlande*, 2000.

[9] Aravind K. Joshi. An introduction to tree adjoining grammars. In Alexis Manaster-Ramer, editor, *Mathematics of Language*, pages 87–115. John Benjamins Publishing Co., Amsterdam/Philadelphia, 1987.

[10] Laura Kallmeyer. *Tree description grammar*. PhD thesis, Unniversité de Tubingen, 1999.

[11] Patrice Lopez. Repairing Strategies for Lexicalized Tree Grammars. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, June 1999.

[12] Mark-Jan Nederhof, Miguel Alonso Pardo, and Eric Villemonte de la Clergerie. Tabulation of automata for tree-adjoining languages. *Grammars*, 2000. à paraître.

[13] G. Perrier. Interaction grammars. In *CoLing '2000, Sarrebrücken*, 2000.

[14] Laurent Romary, Patrice Bonhomme, Florence Bruneseaux, and Jean-Marie Pierrel. Silfide : A System for Open Access and Distributed Delivery of TEI Encoded Documents. *Computers and the Humanities*, 33(1-2) :31–38, 1999.