INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet Atoll

Atelier d'Outils Logiciels pour le Langage naturel

Rocquencourt

THÈME 3A

Table des matières

1. Composition de l'équipe

Responsable scientifique
Éric Villemonte de la Clergerie [CR, responsable scientifique à compter de juin 2002]

Responsable permanent
Pierre Boullier [DR]

Assistante de projet
Josy Baron [AJT]

Personnel Inria
Bernard Lang [DR, responsable scientifique jusqu'en juin 2002]
Philippe Deschamp [CR]

Collaborateur extérieur
François Barthélemy [Maître de conférence, CNAM]

Chercheurs invités
Areski Nait Abdallah [invitation partagée avec le projet COQ]
Alexandre Agustini [novembre 2002, Université Nouvelle de Lisbonne]
Vitor Rocio [novembre 2002, Université Nouvelle de Lisbonne]

Chercheur post-doctorant
Lionel Clément

Doctorant
Benoît Sagot [Détachement du corps des Télécoms à partir de septembre 2002]

Ingénieur
Stéphane Laurière

Stagiaires
Edwige Fangseu Badjio [DEA CHM, Université du Mans, février - septembre 2002]
Fayçal Chami [DEA LIAFA, Université Paris 6, mars - septembre 2002]
José-Manuel Aguirre-Ruiz [DESS INALCO, Université Paris 6, juin - décembre 2002]

2. Présentation et objectifs généraux

2.1. Des outils pour le traitement linguistique

L'équipe Atoll s'est constituée autour d'une compétence dans les techniques d'analyse syntaxique et d'évaluation tabulaire des programmes logiques. Cette compétence, essentiellement acquise dans le cadre de la compilation des langages de programmation, est maintenant appliquée pour le **Traitement Automatique des Langues** [TAL], dans ses aspects syntaxiques, voire sémantiques. Ce domaine de recherche est riche de problèmes sur le plan scientifique, peut bénéficier d'une approche formelle et algorithmique solide et est prometteur quant aux applications industrielles.

Cependant, notre équipe ne peut couvrir qu'un champ restreint des nombreux problèmes liés au traitement de la langue. Ainsi, développer l'ensemble des outils et ressources nécessaires dans une chaîne de traitement pour l'analyse de documents ou la traduction automatique dépasse nos moyens et compétences actuels.

Notre axe principal et historique de recherche porte donc sur les techniques fondamentales en analyse syntaxique, en parallèle avec le développement de prototypes distribuables (SYNTAX et DyALog). Nous nous appuyons en particulier sur des techniques tabulaires, quasi indispensables pour gérer les ambiguïtés inhérentes au langage.

La mise en place de chaînes de traitement linguistique pour permettre la validation de nos outils se fait par intégration d'autres outils et au travers de collaborations avec des partenaires pouvant nous apporter des ressources linguistiques. Outre les questions de génie logiciel liées à la mise en place de telles chaînes, les

problèmes d'accès et de réutilisabilité de celles-ci nous amènent à nous intéresser aux questions de production, de normalisation, de diffusion et d'exploitation de ces ressources. En particulier, pour remédier aux coûts de production des ressources, nous explorons les possibilités d'acquisition automatique ou semi-automatique. Cet axe de recherche offre également l'opportunité de tester nos outils à grande échelle.

Plus généralement, le thème de l'acquisition se relie à celui de l'extraction d'information à partir de corpora ou de documents, domaine en plein essor avec le développement de la « toile » WWW (le « World Wide Web ») et du récent concept de WEB sémantique. Ce dernier concept met en effet en avant l'accès à l'information plus que l'accès aux documents bruts, nécessitant en conséquence des outils permettant ce transfert des documents vers l'information. Outre l'intérêt de le test d'outils linguistiques, ce domaine applicatif se trouve être également dans la lignée de travaux antérieurs sur la gestion de documents au sein d'environnements de traitement (*poste de travail informationel*).

Cette diversification vers des secteurs plus appliqués n'est possible qu'au travers de thèses, mémoires et coopérations. Cependant nous souhaitons aussi, au travers de coopérations, établir des liens nous permettant de faire valoir nos résultats algorithmiques et les systèmes qui les implémentent.

En marge des travaux linguistiques mais néanmoins reliée aux questions d'accès aux ressources et outils linguistiques, le projet ATOLL, au travers de Bernard Lang, mène une réflexion sur les questions de libre accès aux ressources scientifiques et techniques, réflexion dont l'intérêt scientifique, économique et politique ne cesse de croître.

3. Fondements scientifiques

3.1. Formalismes grammaticaux

Mots clés : *TAL*, *analyse syntaxique*, *linguistique*, *programmation dynamique*, *programmation logique*.

Participants : Pierre Boullier, Éric Villemonte de la Clergerie.

Glossaire

CFG *Context-Free Grammars*
DCG *Definite Clause Grammars*
TAG *Tree Adjoining Grammars*
LIG *Linear Indexed Grammars*
LFG *Lexical Functional Grammars*
HPSG *Head-driven Phrasal Structure Grammars*
RCG *Range Concatenation Grammars*
MCG *Mildly Context-sensitive Grammars*
LPDA *Logical Push-Down Automata*
2SA *2-Stack Automata*
TA *Thread Automata*
Programmation Dynamique technique de construction d'algorithmes consistant à diviser un problème en sous-problèmes élémentaires dont les solutions sont tabulées pour pouvoir être réutilisées plusieurs fois si nécessaire.

Ce thème concerne l'analyse syntaxique appliquée à différents formalismes grammaticaux servant au traitement de la langue naturelle. L'ensemble de ces formalismes forme un continuum très large pour lequel sont étudiées des techniques génériques d'analyse qui permettent de traiter au mieux l'ambiguïté inhérente à toute langue.

3.1.1. Des langages de programmation aux grammaires linguistiques

Le passage des grammaires pour les langages de programmation vers des grammaires pour les traitements linguistiques se traduit avant tout par un saut en complexité et l'obligation de gérer les ambiguïtés du langage. Il est bien connu que les problèmes d'ambiguïté en linguistique sont, entre autres problèmes, source d'explosions combinatoires mal maîtrisées.

De plus, alors que la syntaxe des langages de programmation se définit souvent par une (sous-classe d'une) grammaire non contextuelle (CFG), aucun formalisme de description de la syntaxe des langues naturelles n'a fait l'unanimité des linguistes. On assiste au contraire à l'éclosion régulière de nouveaux formalismes grammaticaux, avec en particulier les grandes catégories suivantes :

Formalismes dépendant faiblement du contexte : Ils regroupent entre autres les grammaires d'arbres adjoints (TAG) et linéaires indexées (LIG) et possèdent une base structurelle qui assure l'existence d'évaluateurs travaillant en temps polynomial.

Grammaires d'unification : Elles combinent un squelette non contextuel et une décoration donnée par des attributs logiques. Les représentants les plus connus sont les Grammaires de Clauses Définies (DCG) où l'unification à la PROLOG est utilisée pour calculer et propager ces attributs. Les formalismes plus récents s'appuient sur des structures typées de traits [18] ou éventuellement sur des contraintes. Nous avons ainsi les *Lexical Functional Grammars* (LFG) [20] et *Head-Driven Phrasal Structure Grammars* (HPSG) [21].

Les spécificités évoquées précédemment peuvent se combiner, par exemple en ajoutant des contraintes et des attributs logiques sur une grammaire d'arbres adjoints. Ajoutons que nous participons à ce foisonnement de formalismes grammaticaux avec les RCG (Section 6.2).

Cependant, malgré cette diversité, la plupart des formalismes grammaticaux linguistiques trouvent place dans ce qu'on peut appeler le **« continuum de Horn »**, c'est-à-dire un ensemble de formalismes de complexité croissante, allant des clauses de Horn propositionnelles aux clauses de Horn du premier ordre (grosso-modo PROLOG), et même au-delà.

Ce constat motive notre travail de développement de techniques générales d'analyse permettant de couvrir ce continuum, ceci au travers de deux approches complémentaires qui utilisent, toutes les deux, les techniques de la programmation dynamique afin de réduire l'explosion combinatoire due au traitement des ambiguïtés :

Approche multi-passe. Elle consiste, lorsque c'est possible, à découper un traitement en une séquence dont les composants ont une complexité (pratique ou théorique) croissante ;

Approche globale. Elle repose essentiellement sur la description du formalisme grammatical et des stratégies d'analyse à l'aide d'automates à piles.

Ces deux approches ne s'opposent pas. Au contraire, chacune enrichit l'autre. L'examen de particularités mises en évidence par l'approche multi-passe permet des avancées théoriques ; réciproquement, des concepts théoriques bien compris et identifiés se traduisent par un élargissement du champ d'action de l'approche multi-passe.

3.1.2. Approche multi-passe

Le traitement des langages de programmation est traditionnellement découpé en phases successives de complexité croissante : analyse lexicale, analyse syntaxique, traitement de la sémantique statique, ... Ce découpage se justifie par des raisons théoriques et pratiques. Les automates finis qui modélisent l'analyse lexicale n'ont pas la puissance formelle nécessaire pour décrire la partie syntaxique qui nécessite une description par une (sous-classe des) CFG. Les CFG elles-mêmes ne permettent pas de décrire les phénomènes contextuels de la sémantique statique. Outre une efficacité potentielle accrue (chaque phase est traitée avec le bon niveau de formalisme), ce découpage augmente la modularité du processus.

L'approche multi-passe du traitement des langues naturelles résulte d'une vision similaire. On essaie d'isoler dans les formalismes grammaticaux des parties de complexité moindre sur lesquelles le reste du traitement va

pouvoir s'appuyer. En fait, on constate que la plupart des formalismes du continuum de Horn sont structurés par une base non-contextuelle forte. Ces grammaires peuvent donc être vues comme une CFG décorée par un système de contraintes. L'approche multi-passe consiste pour tous ces formalismes à utiliser un analyseur non-contextuel général (très performant) sur lequel est greffé le système de contraintes, particulier à chaque formalisme traité. Le traitement du squelette non-contextuel est confié au système SYNTAX(cf. 5.1).

3.1.3. Approche globale

L'approche multi-passe s'applique moins bien lorsque la structure CF du formalisme est faible (par exemple dans le cas de PROLOG) ou lorsque les phases sont interdépendantes (par exemple lorsque le traitement des contraintes conditionne fortement l'analyse CFG). Il est alors préférable d'utiliser une approche globale où les contraintes (d'unification ou autres) sont gérées en même temps que l'analyse.

Cette approche, très générale, repose sur des formalismes abstraits d'automates à piles permettant de décrire diverses stratégies d'analyse pour divers formalismes grammaticaux à base logique ou non [6]. Ces automates sont ensuite évalués à l'aide de techniques de programmation dynamique. La notion de pile se prête en effet bien à la division des calculs en sous-calculs élémentaires et réutilisables dans différents contextes : il suffit essentiellement d'oublier provisoirement l'information disponible dans le bas des piles. Ces sous-calculs élémentaires sont représentables sous forme compacte par des *items*. L'utilisation d'automates à 2 piles [2SA] nous a ainsi permis de traiter les formalismes grammaticaux TAG et LIG [7]. L'introduction récente des Automates à fils [*Thread Automata* - TA] doit permettre de traiter une gamme encore plus large de formalismes.

Cette approche trouve ses origines dans les analyseurs à chartes initialement développés par Earley [19]. Elle permet de généraliser différentes méthodes proposées en analyse syntaxique mais aussi en programmation en logique.

Le système DYALOG (cf. 5.2) implémente cette approche pour la programmation en logique et pour différents formalismes grammaticaux.

3.1.4. Forêts partagées d'analyse et de dérivation

Les deux approches précédentes partagent de nombreuses caractéristiques, par exemple l'utilisation des techniques de programmation dynamique. Nous pouvons également citer la notion de forêt partagée d'analyse ou de dérivation. De telles forêts regroupent sous forme compacte l'ensemble des analyses ou dérivations possibles pour une phrase et sont en général assimilables à des grammaires ou à des programmes logiques [4]. Ainsi, alors que l'analyse par une CFG peut conduire à un nombre exponentiel (ou même non borné) d'analyses, la forêt d'analyse reste cubique en la longueur de la phrase analysée. Les forêts d'analyse ou de dérivation, qui sont les structures intermédiaires de l'approche multi-passe, constituent de surcroît un point de départ pour des traitements linguistiques ultérieurs (prise en compte de contraintes syntaxiques ou sémantiques complémentaires, traduction, ...).

3.2. Infrastructure linguistique et normalisation

Participants : Éric Villemonte de la Clergerie, Pierre Boullier, Philippe Deschamp, François Barthélemy.

Nous nous intéressons aux problèmes liés à la mise en place d'une chaîne de traitement linguistique ainsi qu'aux problèmes d'accès et de représentation de ressources linguistiques (cf. 6.1).

Cette réflexion se traduit par le développement de systèmes de construction d'analyseurs syntaxiques comme SYNTAX (cf. 5.1) et DyALog (cf. 5.2). Plus récemment, nous avons également examiné les problèmes de normalisation de grammaires d'arbres adjoints en utilisant XML ainsi que la normalisation des forêts de dérivation produites par les analyseurs. Cet effort se poursuit maintenant au sein d'organismes et d'actions pour la normalisation de ressources linguistiques (cf. 7.4).

Par ailleurs, un environnement pour les grammaires d'arbres adjoints est issu de ce travail de normalisation (cf. 6.1), qui a été largement complété dans le cadre de l'action de recherche RLT.

3.3. Acquisition de ressources

Participant : Éric Villemonte de la Clergerie.

Cet axe concerne l'exploration des relations existantes entre analyse syntaxique et ressources linguistiques de type lexiques. Nous comptons regarder comment l'analyse syntaxique peut servir à l'acquisition de lexiques et dans un second temps, comment des lexiques avec des informations riches peuvent améliorer l'analyse syntaxique.

Des expériences sont en cours dans le cadre de l'ARC « Ressources Linguistiques pour les TAG » (cf. 7.2) et d'un projet encore informel de traitement de corpora botaniques (cf. 6.5).

4. Domaines d'application

4.1. Applications

Le projet ATOLL se situe dans le domaine de la Linguistique informatique dont le champ d'application est très vaste et au cœur des besoins actuels des systèmes d'information. Pour cibler plus spécifiquement les domaines d'applications pour ATOLL, nous pouvons citer :

Correction grammaticale Utilisation de l'analyse syntaxique pour identifier les erreurs grammaticales dans un document et la proposition de correction.

Acquisition de connaissance Les techniques linguistiques (et statistiques) peuvent être utilisées pour extraire de la connaissance à partir de corpora. Ces connaissances peuvent aller d'une simple liste terminologique à un réseau sémantique identifiant des relations entre concepts. Entre ces extrêmes, nous avons l'acquisition de lexiques, de thésaurus et d'ontologies. Nous pensons que ce domaine d'application peut bénéficier de l'utilisation de techniques d'analyse syntaxique plus sophistiquées que celles actuellement utilisées.

Fouille de textes et Questions/Réponses Une analyse syntaxique éventuellement complétée par une analyse sémantique et pragmatique peut permettre l'extraction d'informations précises dans un document, en vue d'alimenter, par exemple, une base de données (ou de connaissances) ou de répondre à une question formulée par un utilisateur.

Parmi ces domaines applicatifs, ATOLL porte en priorité ses efforts vers l'acquisition de connaissances (cf. 6.1 et 7.2) et la fouille de textes (cf. 6.5).

5. Logiciels

5.1. Logiciel Syntax

Participants : Pierre Boullier [correspondant], Philippe Deschamp.

Notre version de travail (non encore distribuée) de SYNTAX a été baptisée 6.0 et est actuellement développée sous Linux. Rappelons que la version 3.9 traite essentiellement les grammaires non-contextuelles déterministes de la catégorie LALR(1). La version 6.0 étend la 3.9 en lui ajoutant à la fois le RLR (extension du LR qui permet l'utilisation si nécessaire d'un nombre non borné de symboles de prévision), des analyseurs non déterministes (à la GLR et à la Earley) qui reposent sur des automates LR, RLR ou Left-Corner et le constructeur pour les grammaires contextuelles de type RCG. L'architecture de développement retenue permet d'obtenir facilement cette version pour des plateformes diverses, notamment Solaris, HP, Linux et Windows.

5.2. Logiciel DyALog

Participant : Éric Villemonte de la Clergerie [correspondant].

DyALog : http://atoll.inria.fr Rubrique « Logiciels »

Le logiciel DyALog est un compilateur de grammaires et de programmes logiques produisant des exécutables tabulaires. Il est principalement dédié à la construction d'analyseurs syntaxiques pour le traitement de la langue naturelle mais est également utile pour remplacer des systèmes PROLOG traditionnels dans le cadre d'applications très ambiguës avec potentiellement du partage de calculs.

Les sources de la version courante de DyALog (1.10.1) sont disponibles pour les plates-formes Linux (Pentium) sous FTP.

La version actuelle permet le traitement des programmes logiques, des DCG (*Definite Clause Grammars*), des FTAG (*Feature Tree Adjoining Grammars*) et des RCG (*Range Concatenation Grammars*). DyALog permet l'utilisation des structures typées de traits et des domaines finis pour des écritures plus compactes des grammaires. Les termes infinis sont maintenant disponibles.

Il est également possible d'interfacer DyALog avec du code C.

Outre un usage interne au projet ATOLL, DyALog est largement utilisé dans le cadre d'un analyseur robuste du Portugais développé à l'Université Nouvelle de Lisbonne.

6. Résultats nouveaux

6.1. Atelier TAG

Participants : Pierre Boullier, Philippe Deschamp, Éric Villemonte de la Clergerie, François Barthélemy, Lionel Clément, Fayçal Chami, José-Manuel Aguirre-Ruiz.

Mots clés : *Grammaire d'Arbres Adjoints*, *XML*.

Glossaire

TAG *Tree Adjoining Grammars*
MG Méta-grammaire, formalisme permettant la génération de grammaires TAG
XML *eXtensible Markup Language*
DTD *Document Type Definition*
Cocoon Environnement tournant au sein d'un serveur HTTP et permettant l'accès et la transformation de documents XML http://xml.apache.org/cocoon1/index.html

Atelier TAG : http://atoll.inria.fr Rubriques « Logiciels » et « Démo ».

Nos travaux sur l'analyse syntaxique des TAG ont conduit au développement d'un atelier de travail pour les TAG, comprenant divers outils et ressources et s'appuyant sur une représentation XML des grammaires. Cet atelier est utilisé dans le cadre de l'ARC RLT.

Dans le cadre de l'ARC RLT (cf. 7.2), nous avons poursuivi notre mise en place d'un atelier de travail pour les TAG.

F. Barthélemy a ainsi continué son travail de conversion des formats XTAG vers le format TAGML basé sur XML que nous mettons en avant pour la représentation de grammaires TAG. Il a en particulier travaillé sur deux grosses grammaires, l'une de la langue anglaise (environ 1100 arbres), développée à l'université de Pennsylvanie, l'autre de la langue française (environ 6000 arbres), développée à Paris 7. Lors de ces conversions, nous nous sommes heurté à différents problèmes, notamment la variabilité des formats lexicaux et syntaxiques, l'ambiguïté sémantique de certaines constructions (par exemple le co-ancrage) et les incohérences des grammaires. À cet égard, ce travail de conversion a ouvert la voie à une technique de validation partielle des grammaires et a permis un retour d'information vers leurs concepteurs.

Des évolutions de la DTD pour TAGML ont été proposées, donnant naissance à une nouvelle version de cette DTD nommée TAGML2. Une mise à jour des outils et ressources est en cours, mais pas encore achevée.

L. Clément a complété l'atelier par une chaîne de traitement couvrant les étapes antérieures à l'analyse syntaxique. Cette chaîne comprend des outils développés par L. Clément comme un segmenteur et un gestionnaire de lexiques (`lexed`). D'autres outils ont été récupérés (l'analyseur morphologique `FLEMM`), éventuellement modifiés (comme `SPAK`, un analyseur syntaxique superficiel développé par Alexandra Kinyon)

ou réentrainé (comme `TreeTagger`, un étiqueteur réentrainé pour le français). Un modèle de *pipeline XML* a été défini et mis en place pour la communication entre ces divers outils. Un document XML est décodé pour passer les informations à un outil de la chaîne dans le format attendu ; les résultats produits sont mis sous forme XML et réinsérés dans le document XML initial. Cet enrichissement du document initial a posé quelques problèmes de synchronisation qui ont été résolus.

L. Clément a également collecté plusieurs corpora pour entraîner les outils et servir à la conduite de l'acquisition de lexique prévue dans le cadre de l'ARC RLT.

Les derniers modules de la chaîne permettent l'appel d'analyseurs syntaxiques construits (dans notre cas) avec SYNTAX ou avec DyALog.

En sortie d'analyse syntaxique, les forêts de dérivation produites sont représentées sous forme XML. Durant son stage DEA, F. Chami a complété un prototype existant permettant le stockage de ces forêts dans des *banques de dérivations*, implantées dans des bases de données relationnelles (en l'occurrence `Postgres`). Les banques de dérivations sont accessibles comme un service WEB sous `Cocoon2` à l'aide d'un langage de requêtes adapté. F. Chami a étendu l'accès à ces banques pour calculer des informations de nature distributionnelle sur l'usage des mots dans un corpus. Pour des mots présents dans le corpus mais non présents dans le lexique de la grammaire d'analyse, il devient possible de compter combien de fois telle construction syntaxique a été utilisée avec succès sur l'ensemble du corpus ou s'il existe des exemples dans un corpus où une seule construction est possible (notion de goulot d'étranglement).

Le regroupement et recoupement de ces informations sur l'ensemble des arbres d'une famille d'une grammaire TAG permet en théorie de caractériser le comportement syntaxique d'un mot et traduit une partie de sa nature sémantique. José-Manuel Aguirre-Ruiz a ainsi réalisé une ébauche d'interface de validation d'entrées lexicales. À partir des informations distributionnelles issues d'une banque de dérivations, cette interface doit proposer la classification des mots absents du lexique. Le linguiste en charge de la validation accepte ou amende la proposition en se basant sur des informations linguistiques et statistiques ainsi que sur des exemples extraits de la banque de dérivation.

Il est prévu que les informations linguistiques soient en relation avec une méta-grammaire (MG) ayant servi à engendrer la grammaire TAG. Cette idée de méta-grammaire [17] est reprise dans le cadre de l'ARC RLT et a donnée lieu au développement d'outils d'édition et de compilation dans l'équipe Langue & Dialogue au LORIA. L. Clément a utilisé ces outils pour réaliser une petite méta-grammaire du français. Il les a modifiés pour une meilleure ergonomie et pour pouvoir engendrer aussi des grammaires LFG (*Lexical Functional Grammars*). J.-M. Aguirre-Ruiz a également utilisé ces outils pour décrire quelques phénomènes linguistiques pour l'espagnol avec une méta-grammaire.

6.2. Analyse contextuelle

Participant : Pierre Boullier.

Mots clés : *formalismes grammaticaux contextuels*, *forêts partagées*, *grammaires à concaténation d'intervalles*, *temps d'analyse polynomial*, *modularité grammaticale*.

Glossaire

MCS *Mildly Context-sensitive Grammars*
RCG *Range Concatenation Grammars*
TAG *Tree Adjoining Grammars*

Nos recherches sur les grammaires à concaténation d'intervalles se sont essentiellement concentrées sur deux axes. D'une part, nous avons regardé comment il était possible d'ajouter de la robustesse aux RCG et d'autre part nous avons regardé s'il était possible de définir des sur-langages des RCL qui puissent s'analyser en temps linéaire. Le but de cette dernière étude est double : d'une part augmenter la vitesse de nos analyseurs en utilisant la technique du guidage et d'autre part fabriquer (automatiquement) à partir d'une RCG à large couverture d'une langue un certain nombre d'outils tels que des analyseurs superficiels ou des étiqueteurs.

Nous avons introduit en 1998 un nouveau formalisme syntaxique, la grammaire à concaténation d'intervalles (RCG), qui définit une classe de langages appelée RCL. Les RCG sont puissantes ; elles englobent les grammaires non-contextuelles (CFG) et les formalismes faiblement dépendant du contexte (*mildly context-sensitive*-MCS). Elles permettent de plus de décrire des phénomènes linguistiques dont certains ne pouvaient se décrire que par des grammaires indexées[1] et d'autres phénomènes qui sont au-delà de la puissance formelle de ces grammaires indexées. Cette puissance n'est pas atteinte au détriment du temps d'analyse qui, comme nous l'avons montré, pour toute grammaire donnée, reste polynomial en la taille du texte source et linéaire en la taille de la grammaire. Ce formalisme grammatical possède en outre un certain nombre de propriétés théoriques (citons par exemple sa clôture par intersection et par complémentation) qui lui permettent de briguer la place occupée actuellement par les CFG au cœur des systèmes définissant les langues naturelles.

Cependant, les propriétés théoriques d'un formalisme grammatical permettent de le distinguer mais ne suffisent pas à le faire adopter et utiliser : il doit non seulement permettre la description des grammaires à large couverture des langues naturelles mais aussi permettre la réalisation des analyseurs syntaxiques correspondants. La difficulté du passage à la pratique provient ici du gigantisme de ces descriptions. Rappelons que le français, défini par l'équipe TALaNa/LATTICE à Paris 7 à l'aide d'une grammaire d'arbres adjoints, contient plus de 5000 arbres élémentaires. Il faut remarquer que cette taille est non seulement supérieure d'au moins un ordre de grandeur à la taille des grammaires décrivant les langages de programmation, que l'information contenue dans un arbre élémentaire est généralement bien plus grande que celle contenue dans une production non-contextuelle, mais également que, si n désigne la longueur du texte source, le temps d'analyse du sous-ensemble des CFG qui décrit les langages de programmation et le temps d'analyse des TAG passe de linéaire en n à $\mathcal{O}(n^6)$.

Cette année, nos recherches se sont essentiellement concentrées sur deux axes. D'une part, nous avons regardé comment il était possible d'ajouter de la robustesse aux RCG et d'autre part nous avons regardé s'il était possible de définir des sur-langages des RCL qui puissent s'analyser en temps linéaire. Le but de cette dernière étude est double : d'une part augmenter la vitesse de nos analyseurs en utilisant la technique du guidage et d'autre part fabriquer (automatiquement) à partir d'une RCG à large couverture d'une langue un certain nombre d'outils tels que des analyseurs superficiels ou des étiqueteurs.

6.2.1. Analyse robuste des RCL

L'idée de robustesse provient de considérations pratiques. Le but d'un analyseur est double ; il doit être un reconnaisseur, il décide si un texte source est grammatical (c'est-à-dire s'il appartient au langage défini par la grammaire), mais il doit également, à toute phrase (texte source grammatical), associer sa structure grammaticale (ou ses structures grammaticales si la phrase est ambiguë). Sur un texte source erroné (agrammatical), le comportement minimal d'un analyseur est de rapporter l'échec de l'analyse. La robustesse d'un analyseur va caractériser sa capacité à analyser et à fournir le maximum de renseignements sur les parties correctes du texte erroné. Traditionnellement, la robustesse est approchée par des modifications *ad hoc* de l'analyseur, modifications qui peuvent être plus ou moins sophistiquées mais qui dépendent toujours très fortement de la méthode d'analyse et qui, de toute façon, ne permettent jamais de traiter tous les cas.

Nous proposons une approche radicalement différente de la robustesse en déplaçant le problème de l'analyseur vers la grammaire : pour un formalisme grammatical donné, est-il possible de définir une notion de grammaire *robuste* G_R associée à une grammaire donnée G ? Dans ce cas, d'une part la notion de robustesse devient beaucoup plus formelle et d'autre part un analyseur robuste n'est rien d'autre qu'un analyseur usuel fondé sur une grammaire robuste.

La grammaire G_R doit posséder deux propriétés. D'une part elle doit pouvoir accepter n'importe quel texte source (formellement, pour un certain vocabulaire terminal T, on doit avoir $\mathcal{L}(G_R) = T^*$). D'autre part, sur toute portion de texte correcte selon G, les structures syntaxiques (partielles) produites par G_R et G sont identiques.

[1] Les langages indexés forment une classe de langages pour laquelle aucun algorithme d'analyse en temps polynomial n'est connu.

Dans le cadre des RCG positives (PRCG), il est très facile de fabriquer des grammaires robustes. Rappelons tout d'abord que les PRCG forment une classe très importante des RCG car elle définit une classe de langages (les PRCL) qui coïncide avec la classe PTIME des langages qui peuvent s'analyser en temps polynomial. Si G est une PRCG, on construit la grammaire robuste associée G_R en ajoutant aux clauses de G, pour chaque prédicat A de G la clause $A(X_1, ..., X_p) \longrightarrow \varepsilon$, où p est l'arité de A. Ces nouvelles clauses peuvent s'instancier quels que soient les arguments de leur partie gauche (on a donc $\mathcal{L}(G_R) = T^*$), mais ces clauses assurent également que chaque appel de prédicat va réussir. Un simple élagage dans la forêt partagée produite par une analyse selon G_R des clauses instanciées propres à G_R suffit à produire toutes les analyses partielles selon G d'un texte source quelconque.

6.2.2. Sur-langages réguliers des RCL

Beaucoup d'applications en traitement de la langue mettent en œuvre non une description complète de cette langue (grammaire à large couverture), mais plutôt des descriptions approchées et partielles, chaque description étant conçue spécifiquement pour le besoin de l'application. Pour une même langue, la cohérence entre ces descriptions pléthoriques est souvent difficile à maîtriser. Là encore, nous proposons une approche originale, consistant à partir d'une grammaire à large couverture de la langue et à lui appliquer un certain nombre de transformations grammaticales, chaque transformation produisant une grammaire qui définit un sur-langage du niveau précédent. L'idée sous-jacente est que la classe du langage d'un niveau donné est plus simple (et donc plus facile à traiter) que la classe du niveau précédent.

Le deuxième axe de recherche concerne donc la définition de sur-langages de certains types de RCL. Nous avons déjà montré comment toute k-PRCG simple qui définit le langage L peut se transformer en une 1-PRCG simple qui décrit un sur-langage L' de L. Comme cette 1-PRCG simple est équivalente à une CFG, les phrases de L' peuvent s'analyser en temps cubique. Cette analyse peut bien sûr être effectuée par un analyseur RCG mais elle peut également s'effectuer par un analyseur non-contextuel général classique (GLR, CYK, Earley...). L'étape suivante consiste à transformer L' en un sur-langage L'' qui puisse s'analyser en temps linéaire. Or, les langages réguliers peuvent s'analyser en temps linéaire, ils forment donc une classe cible possible pour L''. La définition d'une couverture régulière d'un CFL n'est pas neuve et nous l'avons nous-mêmes utilisée en RLR. Cependant, tous les résultats publiés sont relativement décevants. Soit le langage régulier est trop *éloigné* de son modèle CF (rappelons qu'il n'existe pas de plus petite couverture régulière), soit la taille de l'automate fini obtenu est trop importante pour que ce dernier puisse être utilisé. On peut même parfois cumuler ces deux inconvénients. Malgré ces expériences négatives, les nouvelles méthodes que nous avons définies donnent des résultats préliminaires encourageants. Nous envisageons des retombées de cette méthode dans plusieurs domaines. Ces langages réguliers peuvent être utilisés pour guider des analyseurs CF ou des analyseurs RCG afin d'augmenter leur vitesse. Ils peuvent également permettre de faire de l'analyse superficielle, de l'étiquetage (*tagging*) ou du super-étiquetage (*supertagging*). Alors que l'étiquetage permet à un mot de lui associer sa catégorie (par exemple `nom`), le super-étiquetage lui associe un ou plusieurs arbres qui donnent tous les contextes possibles d'utilisation de ce nom.

6.3. DyALog : Automates à piles et Programmation dynamique

Participants : Éric Villemonte de la Clergerie, Areski Nait Abdallah.

Mots clés : *tabulation*, *analyse syntaxique*, *programmation en logique*, *programmation dynamique*, *automate à pile*, *TAG*.

Glossaire

TAG *Tree Adjoining Grammars*
TA *Thread Automata*

Le développement du système DyALog se poursuit et valide l'approche globale par automates (section 3.1.3).

6.3.1. Automates et Programmation Dynamique

Nous poursuivons notre recherche de formalismes d'automates permettant la description de stratégies d'analyse pour des formalismes grammaticaux complexes tout en assurant l'existence d'analyseurs tabulaires efficaces pour ces automates. Un travail antérieur sur une variante d'automates à 2 piles pour les TAG [5] nous a amenés à mettre en avant la notion de *continuation*, i.e. de suspension et reprise de calculs. Cette idée a été généralisée en introduisant un nouveau formalisme d'automates appelés *Automates à fils* [*Thread Automata*-TA]. Ces automates permettent un traitement naturel de constituants grammaticaux discontinues et entrelacés, qui sont en particulier présents dans des phénomènes de déplacement à longue distance. Un « thread » est associé à un constituant A, est suspendu lors du traitement d'un autre constituant B et est réactivé lors du retour à A. Nous avons montré comment les TA permettent la description naturelle de stratégies d'analyse pour divers formalismes faiblement dépendant du contexte dont les TAG, les « *local Multi-Component TAG* » (tree et set local MC-TAG) et la sous-classe des RCG simples ordonnées (osRCG). De plus, nous avons montrés qu'il est possible de décrire des stratégies d'analyse vérifiant la propriété de validité des préfixes reconnus lors de l'analyse (*prefix-valid property*), une propriété généralement difficile à assurer. Par ailleurs, la classe des osRCG traitée par les TA est intéressante car faisant le lien avec les travaux de l'équipe sur les RCG et recouvrant bien la notion de constituants entrelacés. Elle est de plus équivalente aux LCFRS (*Linear Context-Free Rewriting Systems*), une classe très large de formalismes faiblement dépendant du contexte. Outre l'aspect descriptif des TA, nous avons aussi exhibé une interprétation en Programmation Dynamique pour ces TA permettant des analyses en temps et espace polynomial. Les complexité ont été précisées pour les divers formalismes grammaticaux étudiés. Ces résultats sont présentés dans deux articles [13][12].

Cependant, malgré leur nombreux avantages, les TA ne couvrent pas l'ensemble des phénomènes linguistiques. En particulier, certains phénomènes de grande liberté dans l'ordre des mots (*scrambling*) rencontrés par exemple en allemand sont difficiles à modéliser avec les TA. Nous explorons donc comment certaines extensions des TA pourraient traiter ce problème. Des liens entre TA et automates de parcours d'arbres sont également explorés.

6.3.2. Développement du système DyALog

Le travail théorique autour de la notion de tabulation s'est poursuivi en parallèle avec le développement du système DyALog (cf. 5.2).

DyALog a maintenant atteint une bonne stabilité et robustesse. Il est capable de traiter des grammaires conséquentes (par exemple, plusieurs centaines d'arbres TAG) et le traitement de larges grammaires TAG (plusieurs milliers d'arbres) est en vue.

Nous avons donc cherché cette année à étendre les fonctionnalités de DyALog, pour faciliter son utilisation et pour ouvrir la voie au traitement de nouveaux formalismes.

Ainsi, les nombres flottants ont été ajoutés pour permettre à terme le traitement de grammaires probabilistes.

Par ailleurs, l'ajout de nouveaux formalismes et de nouvelles stratégies dans DyALog pose le problème de la taille et de la cohérence du compilateur. Nous réfléchissons à une architecture plus modulaire où il serait possible de construire facilement divers compilateurs à partir d'une bibliothèque de base. Pour préparer le terrain, DyALog offre maintenant un système de modules avec des espaces de noms séparés. Des points de configuration ont également été ajouté pour faciliter une personnalisation du compilateur. Ces points de configuration peuvent en particulier s'appuyer sur l'existence d'un « toplevel » accessible pendant la phase de compilation.

La partie la plus stable des fonctionnalités de DyALog a fait l'objet d'un article [11].

6.3.3. Formalismes

Nous réfléchissons à l'implantation des « Thread Automata » au sein du système DyALog, avec des extensions pour traiter les arguments logiques. Ces TA étendus pourraient alors servir de modèle unificateur pour les divers types d'automates actuellement utilisés dans DyALog.

6.3.4. Optimisations

Le traitement de grammaires linguistiques de plus en plus grandes nous oblige à mettre en œuvre de nouvelles optimisations, à la fois pour réduire la taille des analyseurs syntaxiques produits et pour améliorer leur efficacité.

Par exemple, nous avons optimisé le traitement des prédicats devant être tabulés mais qui sont non-récursifs.

Suite aux bons résultats pour les TAG, nous avons également systématisé l'emploi des conditions d'activation par défaut pour les autres formalismes grammaticaux couverts par DyALog (comme les DCG). Ces conditions d'activation permettent d'exploiter les propriétés de lexicalisation des grammaires en fonction des mots de la phrase, en ne chargeant que les parties de grammaires concernées.

Nous avons aussi implanté dans DyALog la stratégie d'analyse par coin-gauche (*left-corner relation*) pour les DCG avec également ajout d'un module de calcul de cette relation coin-gauche. La conduite de tests a confirmé l'efficacité de cette nouvelle stratégie, qui s'est montrée jusqu'à trois fois plus rapide sur certains exemples.

Enfin, nous avons réalisé des essais de guidage pour les TAG. Un premier essai a consisté à inclure (sur demande) une phase d'analyse des TAG sans arguments de traits précédant l'analyse complète avec traits. Nous travaillons actuellement sur une phase préliminaire où la grammaire TAG initiale est approximée par une grammaire CFG.

6.3.5. Modèles de tabulation

Profitant de sa visite dans l'équipe ATOLL, A. Nait Abdallah a étudié avec Éric de la Clergerie comment intégrer dans DyALog les idées de l'interpréteur PIILOG (*Partial Information Prolog*), conçu et écrit en collaboration avec J.R. Rajnovich. Le but de cette intégration est d'enrichir DyALog avec un traitement des informations hypothétiques ou incomplètes sur le lexique.

Dans une vision à plus long terme, il explore avec É. de la Clergerie l'application de la logique avec information partielle aux problèmes de correction d'erreurs pour le traitement linguistique. Au cours de ces discussions, les jalons d'une combinaison entre les techniques de tabulation de DyALog et la programmation en information partielle ont ainsi été posés. Ceci soulève des questions sur la formalisation plus fine du modèle de tabulation utilisé par DyALog et sur les possibilités non encore pleinement exploitées ouvertes par l'existence d'une table conservant une trace des calculs et des relations entre ces calculs. Il devient possible de considérer des arbres de dérivations (ou arbres de preuves) comme des objets à part entière permettant de déclencher des corrections d'erreurs ou de mieux gérer les phénomènes de coordination dans le langage.

6.4. Morphologie et transducteurs à états finis

Participant : François Barthélemy.

François Barthélemy s'est intéressé à la description morphologique au moyen de machines à états finis. La morphologie est le domaine de la linguistique qui décrit les unités de sens élémentaires de la langue, appelés lexèmes. Ce lexèmes composent des aspects lexicaux associés à une racine fixe et des aspects morphotactiques permettant diverses adjonctions d'information à la racine sous forme de suffixes, préfixes ou infixes. La conjugaison, la déclinaison, l'accord, sont des phénomènes morphotactiques.

L'utilisation de machines à états finis pour mettre en œuvre une description morphologique est une approche très largement utilisée depuis longtemps. Elle a l'avantage de l'efficacité et de l'homogénéité, tous les aspects du problème pouvant être traités de façon compatible. Le lexique aussi bien que les règles de dérivation ou de flexion peuvent être décrits par des machines pour lesquelles existent des opérateurs compatibles avec des opérations linguistiques. Ces opérateurs sont l'union, l'intersection, la composition.

Nous avons travaillé sur un exemple de morphologie très ambiguë, à savoir les verbes akkadiens notés au moyen de l'écriture cunéiforme. Nous avons essayé de décrire leur morphologie complexe (morphologie de type sémitique) et leur écriture très variable (absence de notion d'orthographe) au moyen d'une combinaison de contraintes élémentaires. Les problèmes de recherche qui émergent de notre démarche sont essentiellement ceux de la définition des transducteurs finis. Les transducteurs finis sont des machines à état finis mettant en correspondance deux langages réguliers. Il se pose à leur sujet deux problèmes. Le premier est celui de

l'intersection. Cette opération est essentielle dans l'optique de composer des contraintes élémentaires pour obtenir un système complexe. Or les transducteurs ne sont pas clos par intersection. Il faut donc identifier une sous-classe de transducteur close par intersection ou modifier les définitions de transducteur ou d'intersection. L'autre problème qui se pose est celui de la notation, de la description de transducteurs. Il n'existe pas pour eux l'équivalent des expressions régulières pour les automates finis, c'est à dire une notation strictement équivalente et agréable à manipuler. Nos recherches dans la littérature n'ont pas permis de trouver de solution pleinement satisfaisante à ces deux problèmes.

Pour nos expériences, nous avons utilisé des boîtes à outils de machines à états finis, notamment FSM de AT&T. Les transducteurs offerts par ces boîtes à outils ne conviennent pas à notre application car ils ne sont pas clos par intersection.

Nous espérons pouvoir spécifier le type de machines à états finis utiles pour l'analyse morphologique : des machines mettant en relation plusieurs langages réguliers (pas nécessairement deux), permettant de relier des chaînes de longueur variables, closes par intersection, union et composition et pour lesquelles la composition et l'intersection seraient deux variantes de la même opération.

6.5. Projet Botanique

Participants : Éric Villemonte de la Clergerie, Edwige Fangseu Badjio.

Projet Botanique : http://atoll.inria.fr Rubrique « Projets »

Dans le cadre d'une collaboration naissante avec l'unité Biodival d'Orléans (IRD, ex ORSTOM), nous nous intéressons au traitement de corpus botaniques décrivant des espèces végétales.

Cette collaboration a donné lieu cette année à un stage DEA effectué par Edwige Fangseu Badjio, portant sur la correction automatique de documents numérisés comportant de nombreuses erreurs typographiques. Des résultats encourageants ont été obtenus en s'appuyant sur la notion de distance d'édition et en faisant l'hypothèse que les mots sont en général bien reconnus. Un apprentissage endogène sur l'ensemble du corpus permet ainsi de repérer les mots corrects à fréquence plus forte que les mots erronés proches par édition. Des expérimentations ont été menées avec des automates et transducteurs à états finis pour pouvoir travailler à grande échelle.

Ce travail se situe en amont d'un travail antérieur sur la structuration automatique en XML de documents à structure répétitive. Ces travaux sont préparatoires et doivent se poursuivre par des phases de traitement linguistique (extraction de vocabulaire, acquisition d'ontologie et fouille de texte).

En collaboration avec l'IRD, le projet IMEDIA, le LIFO (Université d'Orléans), MNHN (Museum National d'Histoire Naturel) et LIS (Laboratoire Informatique et Systématique, EA 3496, UPMC), une proposition d'ARC a été déposée. La rédaction d'une proposition pour une action nationale (RIAM ou RNIL) est également en cours.

6.6. Logiciels libres

Participant : Bernard Lang.

Mots clés : *logiciel libre*, *Linux*.

L'évolution du marché et de la disponibilité des ressources logicielles et linguistiques (dictionnaires, grammaires, corpus) nous a amené à nous intéresser au développement des ressources libres. Ce nouveau modèle de production et de distribution des biens immatériels a émergé depuis comme une composante majeure de l'évolution économique et politique, autant que technique, des technologies de l'information, ce qui justifie le travail que nous lui avons consacré depuis environ cinq ans.

7. Contrats industriels

7.1. Projet RNTL e-COTS

Participants : Bernard Lang, Stéphane Laurière.

Le projet **e-COTS** a pour objectif de réaliser un portail internet coopératif et ouvert, au contenu librement réutilisable, sur les composant logiciels commerciaux ou libres et leur utilisation industrielle.

Il s'agit d'un projet financé par le RNTL auquel participent, outre l'INRIA représenté par le projet Atoll, les sociétés Thomson-CSF (gestionnaire du projet), EDF et Bull (équipe Pharos du projet Dyade).

Retardé, ce projet a finalement démarré en 2002.

7.2. ARC RLT

Participants : Éric Villemonte de la Clergerie, Pierre Boullier, Philippe Deschamp, François Barthélemy, Lionel Clément, Fayçal Chami, José-Manuel Aguirre-Ruiz.

Une Action de Recherche Concertée [ARC] intitulée « Ressources Linguistiques pour les TAG » [RLT] http://atoll.inria.fr/RLT/ a été acceptée pour 2001 et 2002, coordonnée par É. de la Clergerie. Les participants sont le projet ATOLL, le laboratoire TALaNa (Université Paris 7), le projet Langue et Dialogue (LORIA) et le projet Calligrame (LORIA). Les objectifs de cette ARC sont de dégager une méthodologie d'acquisition semi-automatique de ressources lexicales pour la grammaire française d'arbres adjoints et de contribuer au développement d'un environnement de travail pour les TAG.

Plusieurs réunions se sont tenues cette année avec de nombreuses présentations concernant l'architecture de la grammaire TAG du français, la méta-grammaire sous-tendant cette grammaire, et les infrastructures existantes parmi les partenaires telles l'atelier TAG d'ATOLL (cf. 6.1).

Dans le cadre de cette ARC, le projet ATOLL a accueilli Lionel Clément sur bourse Post-Doctorale afin de coordonner la mise en place d'une chaîne de traitement linguistique pour les TAG.

Pour poursuivre le travail entrepris dans RLT, nous avons déposé une nouvelle proposition d'ARC axée sur les méta-grammaires. Celles-ci se sont en effet révélées de grande importance au cours de l'ARC RLT.

7.3. ARC Geni

Participants : Éric Villemonte de la Clergerie, Benoît Sagot.

Nous sommes partenaires de l'ARC Geni « Génération et Inférence », coordonnée par « Langue et Dialogue » (LORIA). Cette ARC a débuté courant 2002 et cherche à améliorer la qualité de la génération de textes par utilisation de processus d'inférences à partir d'informations de sémantique lexicale. Outre ATOLL, les partenaires de cette ARC sont ILPL (IRIT), Langue et Dialogue (LORIA), Lattice (Université Paris 7) et Orpailleur (LORIA).

7.4. Action Normalangue

Participants : Éric Villemonte de la Clergerie, Lionel Clément.

ATOLL est un partenaire important de l'action Normalangue, financée par le programme national Technolangue. Cette action doit favoriser l'émergence de formats de standardisation pour les ressources et outils linguistiques, en parallèle avec la définition d'API pour certains outils linguistiques. Cette action doit coopérer avec le groupe miroir AFNOR du sous-comité ISO TC37 SC4 portant sur ces activités de normalisation au niveau international. É. de la Clergerie a en particulier été nommé animateur de ce groupe miroir.

L'action Normalangue ne démarre qu'en 2003, mais des réunions de lancement ont déjà lieu fin 2002 : réunion ISO TC37 SC4 (Pont à Mousson, novembre 2002), réunion du groupe miroir AFNOR (décembre 2002).

7.5. Action EVALDA

Participants : Éric Villemonte de la Clergerie, Pierre Boullier, Lionel Clément.

ATOLL a déposé une demande de participation dans la campagne d'évaluation d'analyseurs syntaxiques devant être organisée par l'action EVALDA dans le cadre du programme national Technolangue.

8. Actions régionales, nationales et internationales

8.1. Actions nationales

Ph. Deschamp est membre de la Commission spécialisée de terminologie de l'informatique et des composants électroniques, et diffuse sur la toile le glossaire http://www-rocq.inria.fr/who/Philippe.Deschamp/CMTI/ résultant de ses travaux (plus de 130 000 téléchargements). Ph. Deschamp est également membre de la Commission spécialisée de terminologie et de néologie des télécommunications.

B. Lang est vice-président de l'AFUL (http://www.aful.org), Association Francophone des Utilisateurs de Linux et des Logiciels Libres, et membre du conseil d'administration de l'ISoc-France, branche française de l'Internet Society. Il est également membre du comité de surveillance de l'association SOISSON Informatique Libre.

8.1.1. Logiciels Libres

B. Lang a présenté les logiciels libres dans des séminaires, tables-rondes et conférences organisés par plusieurs entreprises, collectivités locales et administrations.

8.2. Réseaux et groupes de travail internationaux

8.2.1. Logiciels Libres

B. Lang a été invité à plusieurs reprises à s'exprimer sur les logiciels libres.

B. Lang est membre du groupe d'experts sur le logiciel libre réuni par la DG Société de l'Information (ex DG 13) de la Commission Européenne (http://eu.conecta.it/).

8.2.2. Action FASTLING

Une demande de coopération entre le groupe ATOLL, le groupe CENTRIA de l'Université Nouvelle de Lisbonne et l'université d'Orléans a été acceptée pour 2002 et 2003 dans le cadre du programme d'action INRIA-ICTII avec le Portugal. Cette coopération nommée **FASTLING** prolonge une déjà longue collaboration entre nos équipes.

Cette année, ce programme d'action ICTII a permis le financement de visites dans les deux sens (É. de la Clergerie pour ATOLL, Vitor Rocio et Alexandre Agustini pour CENTRIA, Silvie Billot et Cristel Vrain pour l'université d'Orléans).

Dans le cadre de cette coopération, une grammaire du portugais est développé avec le système DyALog (Vitor Rocio) et utilisée pour diverses tâches, dont des tâches d'acquisition de ressources linguistiques (Alexandre Agustini).

Cette année, V. Rocio a utilisé DyALog pour convertir sa grammaire du portugais dans un format plus déclaratif exploitant les fonctionnalités offertes par DyALog, à savoir les structures de traits et les domaines finis. Cette conversion doit permettre une meilleure maintenance de la grammaire et la possibilité de décrire des phénomènes plus complexes. Le nouveau format devrait également faciliter des expériences de transfert vers le français.

Par ailleurs, Alexandro Agustini a réalisé des expériences d'inferfaçage de l'analyseur du portugais avec un lexique stocké dans une base de données MySQL, utilisant pour cela l'API MySQL disponible sous DyALog.

De côté français, É. de la Clergerie a assisté à la soutenance de thèse de V. Rocio et a pu découvrir les expériences d'acquisition et d'alignement actuellement en cours dans l'équipe portugaise.

8.2.3. Action intégrée PICASSO

Une demande de coopération a été déposée, sans réponse à l'heure actuelle, avec l'université de la Corogne dans le cadre de programme d'actions intégrées PICASSO avec l'Espagne. Cette demande s'inscrit dans le cadre d'une longue collaboration avec cette université.

8.2.4. Collaboration XTAG

Nous sommes actuellement en discussion avec le groupe XTAG de l'université de Pennsylvanie (Philadelphie) pour monter une coopération dans le cadre du programme NSF-INRIA. Cette coopération doit également inclure le projet Langue et Dialogue du LORIA. Les thèmes envisagés portent en autres sur l'évaluation des analyseurs TAG, sur les questions de normalisation autour des TAG et sur l'utilisation de méta-grammaires.

8.3. Visites et invitations de chercheurs

Conjointement avec le projet COQ, nous avons accueilli en 2002 A. Nait Abdallah.

Pendant 2 semaines (novembre 2002), nous avons accueilli Vitor Rocio et Alexandro Agustini dans le cadre de l'action FASTLING. De son côté, É. de la Clergerie a séjourné une semaine à Lisbonne.

9. Diffusion des résultats

9.1. Animation interne à l'INRIA

B. Lang est membre élu au Conseil Scientifique de l'INRIA.

É. de la Clergerie est membre élu suppléant au Conseil Scientifique de l'INRIA et membre du comité de pilotage Cours-Colloques de Rocquencourt. Il a participé à la section locale d'audition du concours CR2 pour 2002.

9.2. Encadrement

É. de la Clergerie a encadré le stage d'Edwige Fangseu Badjio [14]. L. Clément et É. de la Clergerie ont coencadré les stages de Fayçal Chami [15] et de José-Manuel Aguirre-Ruiz [16]. É. de la Clergerie est co-directeur de thèse de Benoît Sagot avec Laurence Danlos (TALaNa/LATTICE, Université Paris 7).

9.3. Jury

B. Lang est membre de la commission de spécialistes du CNAM pour les enseignements d'informatique.

É. de la Clergerie est membre de la commission de spécialistes de l'université d'Orléans.

É. de la Clergerie a participé aux jurys de thèse d'Isabelle Debourges (LIFO, Université d'Orléans, juillet 2002), de Vitor Roccio (Université Nouvelle de Lisbonne, Portugal, octobre 2002) et de David Cabrero Souto (Université de la Corogne, Espagne, octobre 2002).

9.4. Enseignement

9.4.1. Enseignement universitaire.

É. de la Clergerie est intervenu dans l'option « Langage Naturel » du DEA d'Informatique de l'Université d'Orléans et dans l'option TALN de l'ENST.

9.5. Comités de programme

É. de la Clergerie est membre du comité éditorial de la revue T.A.L. http://www.atala.org/tal/tal.html.

B. Lang est ou était membre du comité de programme de diverses manifestations professionnelles.

P. Boullier participe à la sélection des notes de recherche et des démonstrations de EACL'03 (*11th Conference of the European Chapter of the Association for Computational Linguistics*). Il est membre du comité de programme de MOL'8 (*Mathematics of Language*).

É. de la Clergerie a participé à la sélection d'articles proposés à ACL'02, ICLP'02 et PPDP'02. Il participe à la sélection des articles pour EACL'03. Il est membre des comités d'organisation et de programme de IWPT'03 (*International Workshop on Parsing Technologies*).

9.6. Participation à des colloques, séminaires, invitations

B. Lang a contribué à de nombreux colloques ou salons portant sur l'utilisation des logiciels libres et leur rôle économique.

B. Lang est intervenu dans plusieurs manifestations concernant la propriété intellectuelle, notamment en ce qui concerne le développement des logiciels ou l'édition scientifique.

É. de la Clergerie et L. Clément ont participé aux réunions de l'ARC RLT et y ont effectué plusieurs présentations. Ils ont participé à des réunions de l'ARC GENI. L. Clément a également presenté ses travaux au LORIA.

É. de la Clergerie a présenté des articles à TAG+6 [13], TALN02 [11] et COLING02 [12].

L. Clément a présenté un article à LFG'02 [10].

A. Nait Abdallah a participé à l'École PPS sur les types (Agay, Mars 02), à ESSLLI (Trente, août 2002), à l'École INRIA sur les types (Giens, septembre 2002), ainsi qu'au Workshop Types 2002 (Nimegue).

É. de la Clergerie et L. Clément ont participé au séminaire du groupe de travail ISO TC37 SC4 (Pont à Mousson, 21-23 novembre) sur les activité de normalisation de ressources linguistiques et à la réunion de lancement du groupe miroir français AFNOR (décembre 2002).

10. Bibliographie

Bibliographie de référence

[1] P. BOULLIER. *A Cubic Time Extension of Context-Free Grammars.* in « Grammars », numéro 23, volume 3, 2000.

[2] P. BOULLIER. *On TAG Parsing.* in « Traitement Automatique des Langues (T.A.L.) », numéro 3, volume 41, 2000, pages 111-131, issued June 2001.

[3] B. LANG. *Complete Evaluation of Horn Clauses : an Automata Theoretic Approach.* rapport technique, numéro 913, INRIA, Rocquencourt, France, novembre, 1988, http://www.inria.fr/rrrt/rr-0913.html.

[4] B. LANG. *Towards a Uniform Formal Framework for Parsing.* éditeurs M. TOMITA., in « Current issues in Parsing Technology », Kluwer Academic Publishers, 1991, chapitre 11, also appear in the Proc. of Int. Workshop on Parsing Technologies - IWPT89.

[5] É. VILLEMONTE DE LA CLERGERIE. *Refining Tabular Parsers for TAGs.* in « Proceedings of NAACL'01 », pages 167-174, CMU, Pittsburgh, PA, USA, juin, 2001, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/NAACL01-clerger.ps.gz.

[6] É. VILLEMONTE DE LA CLERGERIE. *Automates à Piles et Programmation Dynamique. DyALog : Une application à la programmation en Logique.* thèse de doctorat, Université Paris 7, 1993.

[7] É. VILLEMONTE DE LA CLERGERIE, M. A. ALONSO PARDO. *A tabular interpretation of a class of 2-Stack Automata.* in « Proc. of ACL/COLING'98 », août, 1998, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/SD2SA.ps.gz.

Articles et chapitres de livre

[8] M. A. ALONSO, É. DE LA CLERGERIE, V. J. DIAZ, M. VILARES. éditeurs G. S. JOHN CARROLL, H. BUNT., *Relating Tabular Parsing Algorithms for LIG and TAG.* Kluwer Academic Publishers, 2002, chapitre 1, à paraître, révision d'un article présenté à IWPT'00.

[9] P. BOULLIER. éditeurs G. S. JOHN CARROLL, H. BUNT., *RANGE CONCATENATION GRAMMARS.* Kluwer Academic Publishers, 2002, chapitre 12, à paraître, révision d'un article présenté à IWPT'00.

Communications à des congrès, colloques, etc.

[10] L. CLÉMENT, K. GERDES, S. KAHANE. *an LFG-type Grammar for German Based on topological Model.* in « Proceedings of the LFG02 Conference », CSLI Publications, éditeurs M. BUTT, T. H. KING., 2002, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/lfg02-cgk.ps.gz.

[11] É. VILLEMONTE DE LA CLERGERIE. *Construire des analyseurs avec DyALog.* in « Proc. of TALN'02 », juin, 2002, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TALN02.pdf.

[12] É. VILLEMONTE DE LA CLERGERIE. *Parsing Mildly Context-Sensitive Languages with Thread Automata.* in « Proc. of COLING'02 », août, 2002, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/COLING02.pdf.

[13] É. VILLEMONTE DE LA CLERGERIE. *Parsing MCS languages with Thread Automata.* in « Proc. of TAG+6 », mai, 2002, ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TAG6.pdf.

Divers

[14] E. F. BADJIO. *Traitement de corpus botaniques.* mémoire de DEA, DEA CHM, Université du Mans, septembre, 2002.

[15] F. CHAMI. *Étude de filtrage pour la lexicalisation de la grammaire FTAG.* mémoire de DEA, DEA IARFA, Université Paris 6, octobre, 2002.

[16] J.-M. AGUIRRE RUIZ. *Le passif de l'espagnol dans le contexte d'une métagrammaire pour le formalisme TAG.* mémoire de DESS, CRIM INALCO, 2002.

Bibliographie générale

[17] M.-H. CANDITO. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées.* thèse de doctorat, Université Paris 7, janvier, 1999.

[18] B. CARPENTER. *The Logic of Typed Feature Structures with Applications to Unification Grammars, Logic Programs and Constraint Resolution.* numéro ISBN 0-521-41932, Cambridge University Press, 1992.

[19] S. EARLEY. *An Efficient Context-Free Parsing Algorithm.* in « Communications ACM 13(2) », ACM, 1970, pages 94-102.

[20] R. M. KAPLAN, J. BRESNAN. *Lexical-Functional Grammar : A formal system for grammatical representa-*

tion. éditeurs J. BRESNAN., in « The Mental Representation of Grammatical Relations », The MIT Press, Cambridge, MA, 1982, pages 173-281, Reprinted in Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*, 29-130. Stanford : Center for the Study of Language and Information. 1995..

[21] C. POLLARD, I. A. SAG. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago, 1994.