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
Bernard Lang [DR]
Responsable permanent
Pierre Boullier [DR]
Assistante de projet
Josy Baron [AJT]
Personnel Inria
Philippe Deschamp [CR]
Éric Villemonte de la Clergerie [CR]
Collaborateur extérieur
François Barthélemy [Maître de conférence, CNAM]
Chercheur invité
Alexandre Agustini [Novembre 2001, Université Nouvelle de Lisbonne]
Chercheur post-doctorant
Lionel Clément [A partir du 1er novembre]
Doctorants
Vitor Rocio [Thèse en co-tutelle avec l'Université Nouvelle de Lisbonne]
François Role [Fonctionnaire au DISTNB-MESR, Université d'Orléans, départ au 1er Septembre]
Stagiaires
Vartika Bhandari [stage ingénieur IIT Kanpur (Inde), été 2001]
Abdelaziz Khajour [stage ingénieur ENSIAS (Maroc), printemps 2001]
Julien Collet [stage ingénieur École Polytechnique, printemps 2001]
Stéphanie Werli [stage DEA Université Paris 7, printemps été 2001]

2 Présentation et objectifs généraux

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 de la Langue Naturelle** [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, mettre en place un système complet de traitement pour l'analyse de documents ou la traduction automatique dépasse nos moyens et compétences actuels.

Nous cherchons donc à développer progressivement des aspects plus appliqués du traitement de la langue en nous appuyant sur nos autres points forts liés à nos compétences informatiques et en nous associant à d'autres acteurs plus directement impliqués dans les problèmes de traitement de documents électroniques et de linguistique appliquée.

L'usage en plein essor des documents électroniques et structurés, dû en grande partie au développement de la « toile » WWW (le « World Wide Web »), nous paraît une opportunité à exploiter, notamment en raison de notre expérience concernant les environnements de programmation. En conséquence, nous cherchons à nous diversifier vers des secteurs plus appliqués, à l'occasion 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 implantent.

Le développement de nos activités présente donc actuellement deux aspects, que nous ferons converger à terme :

1. Poursuite de nos travaux sur les techniques fondamentales en analyse syntaxique et évaluation tabulaire de programmes et grammaires logiques, avec développements de prototypes distribuables.
2. Recherche, traitement et gestion des documents électroniques, en particulier dans leur dimension linguistique.

Nos travaux étant nécessairement limités à un champ étroit de la linguistique informatique, il nous faut pouvoir travailler dans le contexte de ressources et d'outils développés par d'autres équipes. Malheureusement, dans ce domaine comme dans d'autres, le libre accès aux ressources scientifiques et techniques se fait de plus en plus difficile et coûteux. Cela nous a amené à nous pencher sur la possibilité du développement de ressources libres. Ce thème est devenu un sujet à part entière, 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.

Résumé : *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.*

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*
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.

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 [Car92] ou éventuellement sur des contraintes. Nous avons ainsi les *Lexical Functional Grammars* (LFG) [KB82] et *Head-Driven Phrasal Structure Grammars* (HPSG) [PS94].

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.

[Car92] B. Carpenter, *The Logic of Typed Feature Structures with Applications to Unification Grammars, Logic Programs and Constraint Resolution*, ISBN 0-521-41932, Cambridge University Press, 1992.

[KB82] R. M. Kaplan, J. Bresnan, « Lexical-Functional Grammar: A formal system for grammatical representation », *in : The Mental Representation of Grammatical Relations*, J. Bresnan (éditeur), The MIT Press, Cambridge, MA, 1982, p. 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.

[PS94] C. Pollard, I. A. Sag, *Head-Driven Phrase Structure Grammar*, University of Chicago Press, Chicago, 1994.

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 CF). 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 [5].

Cette approche trouve ses origines dans les analyseurs à chartes initialement développés par Earley [Ear70]. Elle permet de généraliser différentes méthodes proposées en analyse syntaxique mais aussi en programmation en logique.

Le système DyALog (cf. 5.3) implante 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 (c.f. Guidage dans la Section 6.2), 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

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.

Cette réflexion se traduit par le développement de systèmes de construction d'analyseurs syntaxiques comme SYNTAX (cf. 5.1), DyALog (cf. 5.3) et RCG (cf. 5.2). Plus récemment, nous avons également examiné les problèmes de normalisation de grammaires d'arbre adjoints en utilisant XML ainsi que la normalisation des forêts de dérivation produites par les analyseurs. Un environnement pour les grammaires d'arbres adjoints est issu de ce travail de normalisation (cf. 6.1).

3.3 Acquisition de ressources

Participant : Éric Villemonte de la Clergerie.

Ce nouvel axe concerne l'exploration des relations existantes entre analyse syntaxique et res-

[Ear70] S. Earley, « An Efficient Context-Free Parsing Algorithm », *in : Communications ACM 13(2)*, ACM, 1970, p. 94–102.

sources 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.

Cet axe de recherche démarre 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. 6.4).

3.4 Le Poste de Travail Informationnel

Participants : Bernard Lang, François Role.

La recherche de débouchés applicatifs à nos travaux, de pair avec un certain intérêt de l'équipe, nous pousse vers les nouveaux média (principalement cédérom et internet) dont le rôle économique, social et culturel va croissant. Cela nous amène naturellement à nous impliquer dans diverses actions dont nous espérons à terme des synergies avec nos compétences en analyse syntaxique et déduction, ainsi qu'avec celles plus anciennes en génie logiciel et traitement de documents structurés.

Plus applicatif, cet axe présente deux volets complémentaires, à savoir d'une part la conception et le développement d'outils pour des supports matériels des documents qui sont en pleine évolution, et d'autre part le développement de techniques d'analyse et de gestion des contenus des documents eux-mêmes. Ces deux aspects sont parfois difficilement dissociables. Par exemple, la réalisation d'un outil de recherche sur le Web requiert à la fois une maîtrise des techniques strictement informatiques de l'accès à l'information, mais aussi des outils sophistiqués d'extraction du contenu des documents (par exemple la lemmatisation des mots pour un indexeur sophistiqué).

Il est également clair que ces problèmes font appel à une grande variété de techniques liées au traitement des documents, à l'analyse de la langue naturelle et à la recherche documentaire. Bien entendu, il ne saurait être question d'acquérir une expertise universelle avec les moyens dont nous disposons, et nous cherchons au maximum à réutiliser des outils existants pour nos travaux, tout en nous efforçant d'identifier et d'explorer des problèmes originaux.

Le thème unificateur que nous fixons à ces activités est le développement d'un *Poste de Travail Informationnel*, permettant à un travailleur intellectuel de gérer facilement son capital d'informations et de documents, tant en ce qui concerne la recherche de nouveaux documents, qu'en ce qui concerne leur mémorisation et leur organisation (indexation) pour une réutilisation ultérieure.

4 Domaines d'applications

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 thesaurus 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.

5 Logiciels

5.1 Logiciel SYNTAX

Participants : Pierre Boullier [correspondant], Philippe Deschamp.

Une version de SYNTAX est en cours de portage sous Linux. Cette version étend la version 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), et des analyseurs non déterministes (à la GLR et à la Earley) qui reposent sur des automates LR, RLR ou Left-Corner. Cette tâche, non prioritaire pour l'instant, pourrait le devenir si nous décidons de participer à la « compétition d'analyseurs » organisée par John Carroll (`http://www.cogs.susx.ac.uk/lab/nlp/carroll/elsps.html`).

Pour rappel, la version 3.9 de SYNTAX existe pour les environnements SunOs, Solaris, DOS et Windows. Elle comporte notamment :

- un traitement amélioré des caractères 8 bits, fonctionnant dans tous les environnements ;
- un constructeur de dictionnaires utilisant des techniques de représentation de matrices creuses.

5.2 Logiciel RCG

Participants : Pierre Boullier [correspondant], Philippe Deschamp.

Le système RCG permet de produire des analyseurs syntaxiques pour les RCG. Ce prototype, utilisé pour l'instant en interne, réalise d'excellentes performances.

5.3 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 (1.9.0) sont disponibles pour les plates-formes Linux (Pentium) et SunOS (Sparc) 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, Vartika Bhandari, Abdelaziz Khajour.

Mots clés : Grammaire d'Arbres Adjoints, XML.

Glossaire :
TAG *Tree Adjoining Grammars*
XML *eXtensible Markup Language*
DTD *Document Type Definition*
servlet scripts Java exécutés au sein d'un serveur HTTP comme Apache
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 ».

Résumé : *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.*

Le développement de l'atelier TAG du projet ATOLL s'est poursuivi cette année en complétant les composants déjà disponibles et en l'enrichissant.

Ainsi, une nouvelle DTD pour la représentation des grammaires d'arbres adjoints est actuellement disponible, permettant la prise en compte de grammaires plus complexes. Le travail de conversion des grammaires existantes et des outils actuels vers cette nouvelle DTD n'est cependant pas encore achevé.

Les différents analyseurs TAG que nous construisons sont accessibles de manière uniforme par l'intermédiaire d'un serveur de parseurs. Nous avons ajouté à ce serveur un accès au parseur TAG de référence développé par le groupe XTAG à l'université de Pennsylvanie. Nous avons construit un module de conversion du format de sortie de ce parseur XTAG vers notre format de représentation canonique des forêts de dérivation. En plus des vues brutes et sous forme de

grammaires des forêts de dérivation, nous avons ajouté deux vues graphiques supplémentaires sous forme de forêts d'arbres ou de forêts de dépendance.

Enfin, nous avons mis en place deux nouveaux serveurs. Le premier, conçu par Abdelaziz Khajour, permet de consulter le contenu d'une grammaire (lexiques, familles, arbres, ...) tandis que le second, conçu par Vartika Bhandari, permet de consulter le contenu d'une banque de forêts de dérivation (produite par analyse d'un corpus de document). Quoique légèrement différents dans leur réalisation, ces deux serveurs ont une conception similaire :

- les données XML (grammaires ou forêts) sont stockées dans des DB relationnelles (Postgres et MySQL) en respectant leur structure XML.
- un langage de requêtes spécifique permettant une conversion en SQL. Ce langage doit faciliter les requêtes posées par des linguistes.
- la consultation se fait par l'intermédiaire de servlets Java sous Cocoon au sein d'un serveur Apache.

L'ensemble de cette infrastructure (en particulier les serveurs) doit nous servir dans le cadre de l'ARC RLT (cf. 7.2). Notre atelier TAG a été présenté à [10, 11] ainsi qu'au groupe XTAG [17].

6.2 Analyse contextuelle

Participant : Pierre Boullier.

Mots clés : formalismes grammaticaux contextuels, guidage, temps d'analyse polynomial, modularité grammaticale.

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

Résumé : *Nos recherches sur les grammaires à concaténation d'intervalles se sont poursuivies selon deux axes : l'étude de leur propriétés formelles et leur utilisation pour implanter les analyseurs pour des grammaires à large couverture de l'anglais et du français.*

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, englobant les grammaires non-contextuelles (CFG) et les formalismes faiblement dépendant du contexte (*mildly context-sensitive*—MCS). Elles permettent même la description de phénomènes linguistiques qui nécessitaient auparavant des grammaires indexées[1], voire de phénomènes 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, 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

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

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 la grammaire du français, définie à l'université Paris 7 à l'aide d'une grammaire d'arbres adjoints, contient plus de 5000 arbres élémentaires. Non seulement cette taille est supérieure d'au moins un ordre de grandeur à la taille des grammaires décrivant les langages de programmation, mais aussi l'information contenue dans un arbre élémentaire est bien plus grande que celle contenue dans une production non-contextuelle. De plus, les temps d'analyse passent de linéaire en n pour le sous-ensemble des CFG qui décrit les langages de programmation à $\mathcal{O}(n^6)$ pour les TAG, si n désigne la longueur du texte source.

Nos recherches se sont donc poursuivies essentiellement selon deux axes. D'une part, nous avons continué l'étude des propriétés formelles des RCG et d'autre part nous avons poursuivi la réalisation d'analyseurs RCG pour des grammaires à large couverture de l'anglais et du français. Le but de cette expérimentation est de montrer que cela est possible avec la technologie RCG, et aussi que les analyseurs obtenus ont des performances égales ou même supérieures aux analyseurs dédiés originaux.

6.2.1 Propriétés formelles des RCG

Ce premier axe, consacré à l'étude des propriétés formelles des RCG et des RCL, a donné lieu cette année à deux publications internationales.

La première [8] montre que, contrairement aux CFG, les RCG permettent de compter. Il est bien connu que les CFG ne savent compter que jusqu'à deux : elles savent reconnaître les parenthèses ouvrantes et fermantes de structures bien parenthésées. Nous avons montré que les RCG peuvent décrire des propriétés que même les grammaires indexées sont incapables de définir.

Une deuxième publication [13] présente comment traduire une grammaire contextuelle[2] (CG) en une RCG équivalente. Bien que le traitement des langues naturelles soit l'un des buts principaux qui a conduit à la définition des CG, très peu d'études ont été consacrées à ce domaine. Une des raisons est certainement l'absence d'analyseur syntaxique efficace pour les langages contextuels. Les CG sont aussi appelées grammaires *pures* car elles définissent des langages sans utiliser de symboles auxiliaires (de non-terminaux). À partir de phrases de base appelées *axiomes*, on fabrique de nouvelles phrases en insérant des couples de séquences de mots appelés *contextes*, autour de sous-chaînes appartenant à des langages appelés *sélecteurs*. Dans ce processus de dérivation, chaque proto-phrase est en fait une phrase qui contient tous les mots de l'étape précédente, et les mots du contexte ajoutés à l'étape courante. Nous avons montré que parmi les nombreuses variantes des CG, certaines pouvaient se traduire directement en RCG et pouvaient donc s'analyser en temps polynomial, alors que pour d'autres variantes, cette traduction était impossible. Pour pouvoir traiter ces secondes variantes, nous avons étendu le

[2] Ce type de grammaire ne correspond pas au type 2 de la classification de Chomsky.

formalisme des RCG et défini les RCG *dynamiques* (DRCG). Cette extension permet, au cours d'une analyse, de fabriquer, à partir du texte source courant, un nouveau texte qui peut être lui-même utilisé comme nouveau texte source courant. Bien sûr, la puissance supplémentaire des DRCG sur les RCG se traduit par un accroissement du temps de l'analyse qui passe de polynomial à exponentiel.[3]

6.2.2 L'analyse des langages à concaténation d'intervalles

L'exécution en temps polynomial est, pour le traitement de la langue naturelle, l'un des atouts des analyseurs RCG. De plus, l'implantation prototype que nous avons réalisée s'est révélée déjà très efficace. Nous avons encore essayé d'accentuer cette particularité en concevant et réalisant une technique d'analyse *guidée*.

Nous avons montré que pour toute RCG positive G définissant le langage L, il était possible de construire une RCG positive à prédicats unaires (une 1-PRCG) G_1 qui définit un sur-langage L_1 de L. D'autre part, dans [1], nous avons montré qu'une très large sous-classe des 1-RCG définissait des langages qui pouvaient s'analyser en temps cubique. Si l'on suppose que la grammaire G_1 vérifie les conditions de cette sous-classe, on sait donc analyser L_1 en temps cubique. Dans sa version la plus simple, une analyse guidée est une analyse dans laquelle le texte source w est analysé deux fois : une première phase analyse w (en temps cubique) selon G_1 et construit une structure, le *guide*, qui sera utilisée par une deuxième phase qui réanalyse w, mais cette fois selon G, en s'aidant des renseignements contenus dans le guide. En fait, les renseignements contenus dans le guide permettent à l'analyseur selon G de ne s'engager dans une analyse (partielle) que si la ou les analyses correspondantes ont réussi selon G_1.

Cette technique d'analyse, qui s'applique à (presque) toutes les RCG, a été expérimentée sur la TAG à large couverture de l'anglais définie dans le projet XTAG, à l'Université de Pennsylvanie à Philadelphie.

6.2.3 Analyse de l'anglais

Les expérimentations que nous avons menées sur des analyseurs RCG guidés ont été présentées dans l'article [12]. Ces expérimentations ont porté sur la grammaire de l'anglais la plus récente issue des travaux menés à l'Université de Pennsylvanie à Philadelphie dans le cadre du projet XTAG. Le formalisme utilisé dans cette description est celui des TAG. Cette TAG décrivant l'anglais contient plus de 1 200 arbres élémentaires et est associée à un lexique contenant plus de 300 000 formes fléchies. Conformément à [2], cette TAG est tout d'abord transformée en une RCG équivalente.

Dans l'expérimentation la plus simple, cette RCG G est transformée en une autre RCG G_1 qui décrit un sur-langage de la première et dont l'analyseur sert à produire un guide. L'analyseur pour G utilise ce guide pour diriger ses actions d'analyse. L'expérimentation a été menée sur 32 phrases anglaises assez courtes (la longueur ne dépasse pas 17 mots). Sur ce petit ensemble l'analyse guidée va de 2 à 3 fois plus vite que l'analyse non guidée. Plus les phrases sont longues, plus le gain est important ; il est de l'ordre de 60 pour une phrase de 35 mots.

[3] Le temps d'une analyse peut même ne pas être fini car la longueur des textes source créés dynamiquement peut ne pas être bornée.

Nous avons aussi comparé les temps mis par une analyse RCG avec l'analyseur du groupe XTAG qui est une référence dans la communauté TAG. Sur les phrases du petit ensemble le gain des analyseurs RCG par rapport à l'analyseur natif XTAG dépasse 3 ordres de grandeur alors qu'il dépasse 5 ordres de grandeur sur la longue phrase !

6.3 DyALog : Automates à piles et Programmation dynamique

Participants : Éric Villemonte de la Clergerie, Julien Collet.

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

Glossaire :
TAG *Tree Adjoining Grammars*
Feature TAG TAG avec attributs
2SA *2-stack Automata*
TA *Thread Automata*

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

Automates et Programmation Dynamique Nous avons proposé en 1998 un formalisme d'automates à 2 piles, associé à une interprétation en programmation dynamique [5], permettant l'analyse tabulaire des grammaires d'arbres adjoints [TAG].

L'interprétation en Programmation Dynamique proposée assure une complexité optimale en place et en temps, à savoir respectivement $O(n^5)$ et $O(n^6)$ où n est la longueur de la chaîne analysée. Néanmoins, son implantation dans DyALog a montré sa complexité et surtout n'a pas fourni les résultats escomptés. L'analyse des particularités des TAG nous ont conduits à concevoir et à implanter une nouvelle interprétation en Programmation Dynamique pour les 2SA. Cette nouvelle interprétation est qualifiée de « faible » car n'assurant pas une complexité optimale (mais néanmoins polynomiale). Elle est bien plus simple à mettre en œuvre et donne effectivement de bien meilleurs résultats sur les grammaires linguistiques que nous avons testées [14].

Cette nouvelle interprétation met en avant la notion de *continuation* : la reconnaissance d'un constituant A est suspendue pour reconnaître un segment d'un constituant B et reprendre par la suite. Cette suspension se traduit par la création d'une continuation qui est tabulée (temporairement ou définitivement).

Généralisant ce schéma, nous avons alors introduit un nouveau formalisme d'automates appelés *Thread Automata* [TA] et une interprétation en Programmation Dynamique pour ces automates assurant une complexité polynomiale. Ces automates peuvent s'utiliser sur un large spectre de formalismes linguistiques, que nous conjecturons être équivalent à celui couvert par les RCG simples et les LCFRS (Linear Context-Free Rewriting Systems). D'autres algorithmes tabulaires en complexité polynomiale ont déjà été présentés pour ces formalismes (en particulier par l'intermédiaire des RCG) mais sans assurer la validité des préfixes reconnus pendant l'analyse (*prefix-valid property*), ce que permettent les TA. Nous sommes en train d'explorer

ces TA ainsi qu'une classe de grammaires bien adaptées pour les TA et que nous avons dénommées *Grammaires à Constituants Entrelacés* (ICG – Interleaved Constituent Grammars). Un article est en cours de rédaction [19].

Plus généralement, ces travaux permettent d'explorer les liens existant entre tabulation et continuations que ce soit dans le domaine de l'analyse syntaxique ou dans celui de la programmation en logique.

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. 5.3)

Nous avons amélioré la robustesse de DyALog, concernant la récupération de la mémoire (GC), la gestion des alternatives et la taille des exécutables produits par DyALog. Le mini-assembleur utilisé par DyALog a été enrichi. Dans une version prototype, nous avons également cherché à réduire les temps de démarrage, dus à la construction de termes, et qui sont importants pour des grammaires de grande taille.

Formalismes Nous avons cherché à enrichir l'ensemble des formalismes grammaticaux couvert par DyALog.

Un stage d'ingénieur a ainsi été effectué par Julien Collet sur l'implantation des HPSG pour exploiter les structures de traits typées [TFS] disponibles dans DyALog. Ce travail, quoique très incomplet, a permis de vérifier que cette implantation est possible. Pour aller dans cette direction, nous avons ainsi étendu DyALog pour permettre la gestion de termes infinis (création, affichage, unification).

À titre expérimental, nous avons également réalisé une implantation des RCG dans DyALog en les étendant par des arguments logiques.

Dans le cadre de sa visite, Vitor Rocio a complètement réécrit en DyALog les divers composants de son analyseur syntaxique pour le Portugais, certains de ces composants étant initialement écrits en SICSTUS. Cette réécriture a permis de confirmer l'intérêt des nouvelles fonctionnalités offertes par DyALog.

Il est prévu à court terme de réaliser une implantation des Thread Automata.

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.

Nous avons ainsi testé l'utilisation de stratégies d'analyse par coin-gauche (left-corner relation) dans le cadre de DCG. Les résultats n'ont pas été satisfaisants et une analyse plus poussée semble suggérer quelques modifications à apporter dans DyALog.

L'autre optimisation en cours d'étude concerne la factorisation des préfixes communs rencontrés lors du parcours d'une clause DCG ou d'un arbre élémentaire TAG. Une version prototype a été réalisée qui doit être testée et surtout étendue pour gérer les arguments des non-terminaux.

6.4 Projet Botanique

Participant : Éric Villemonte de la Clergerie.

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 commençons à nous intéresser au traitement de corpus botaniques décrivant des espèces végétales.

Cette collaboration a donné lieu cette année à un stage DESS qui s'est déroulé à Orléans sur la structuration en XML d'un corpus botanique à partir d'une version électronique « plate » [15]. En plus de la structuration, une interface de navigation WEB a été conçue et implantée pour un meilleur accès à ce corpus.

À terme, ce travail doit se poursuivre par des phases de traitements linguistiques (extraction de vocabulaire, acquisition d'ontologie et fouille de texte).

En collaboration avec l'IRD, le projet IMEDIA et le LIFO (Université d'Orléans), il est envisagé de déposer un dossier pour une action RNTL ou PRIAM.

6.5 Bibliothèques électroniques

Participant : François Role.

Mots clés : métadonnée, bibliothèque électronique.

F. Role a soutenu cette année sa thèse [7] concernant les apports de la documentation structurée et la prise en compte des métadonnées dans la conception d'un environnement de travail pour l'étude de textes numériques.

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.[4] 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 quatre ans.

[4] Notre attention fut initialement attirée sur ce sujet par la mise en œuvre du système d'exploitation libre Linux. Des discussions avec plusieurs collègues nous ont amené à voir ce problème sous l'angle de la disponibilité des ressources scientifiques.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Projet RNTL e-COTS

Participant : Bernard Lang.

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).

Ce projet a été accepté et devait démarrer en 2001. Il est cependant actuellement retardé.

7.2 ARC RLT

Participants : Éric Villemonte de la Clergerie, Pierre Boullier, Philippe Deschamp, François Barthélemy, Lionel Clément, Stéphanie Werli.

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 permettant de préciser le scénario à déployer en 2002. Elles ont également été l'occasion 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 accueille actuellement Lionel Clément sur bourse Post-Doctorale afin de coordonner la mise en place d'une chaîne de traitement linguistique pour les TAG. Cette chaîne doit intégrer un ensemble d'outils et de ressources linguistiques existant ou en cours de développement. Nous disposons déjà d'outils pour la segmentation en mots et mots composés, l'étiquetage morphosyntaxique, l'analyse syntaxique superficielle (*shallow-parsing*) et l'analyse syntaxique.

L'ARC a aussi permis le déroulement du stage DEA de Stéphanie Werli [20], qui s'est déroulé à TALaNa (sous la direction d'Anne Abeillé).

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 depuis peu membre de la Commission spécialisée de terminologie et de néologie des télécommunications.

B. Lang est secrétaire 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 (`http://www.isoc.asso.fr`), branche française de l'Internet Society.

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 ORLINSROC

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 2001 dans le cadre du programme d'action ICTII entre le Portugal et la France. Cette coopération nommée **ORLINSROC** 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 (Eric de la Clergerie, Vitor Rocio et Alexandre Agustini).

Coté portugais, V. Rocio ainsi pu compléter son analyseur portugais partiellement écrit en DyALog en en exploitant les nouvelles fonctionnalités. Cet analyseur comprend plusieurs couches, dont deux exploitent DyALog, et s'appuie en particulier sur le formalisme des *grammaires à mouvements restreints* (BMG). Cette expérience est décrite dans [9].

De notre côté, nous avons pu découvrir plus en détail cet analyseur et les applications linguistiques en cours à Lisbonne qui s'appuient sur cet analyseur.

Il est envisagé de convertir cet analyseur et certaines des applications associées pour le français.

8.2.3 Collaboration XTAG

Suite à une visite de É. de la Clergerie effectuée au sein du groupe XTAG à l'université de Pennsylvanie (Philadelphie), nous sommes en train de mettre en place une collaboration dans le cadre du programme NSF-INRIA. Cette collaboration doit porter sur l'analyse syntaxique des TAG et sur l'infrastructure pour les TAG.

9 Diffusion de résultats

9.1 Encadrement

É. de la Clergerie a encadré les stages d'ingénieurs de A. Khajour [16], V. Bhandari, et J. Collet. Il a également été impliqué dans le suivi du stade DESS de Stéphanie Balva à Orléans [15], dans le cadre du projet Botanique (cf. 6.4).

9.2 Jury

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

B. Lang a été membre du jury de thèse de F. Role.

P. Boullier est un des rapporteurs de l'Habilitation à diriger des recherches de Jacques Farré (Université de Nice).

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

9.3 Enseignement

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

É. de la Clergerie et F. Role sont intervenus dans la filière transversale « Ingénierie des Industries Culturelles » (2IC) de l'Université de Technologie de Compiègne.

9.4 Comités de programme

É. de la Clergerie a été membre du comité de programme de IWPT'01 (Pékin). Il a également présidé une session à IWPT'01, à ACL'01 et à ICLP'01.

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 membre du comité éditorial de la revue FUTUR(e)S `http://www.futur-e-s.com`.

B. Lang est ou était membre du comité de programme de diverses manifestations professionnelles, dont la conférence Net 2001 et l'atelier « Logiciel Libre » Tunis 2001.

F. Role a été membre du comité d'organisation de la conférence ISKO 2001.

P. Boullier a jugé des articles proposés à ACL'01 et É. de la Clergerie des articles proposés à ACL'01, ICLP'01 et IWPT'01.

9.5 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 a participé aux réunions de l'action de recherche RLT et y a effectué plusieurs présentations.

Participation de É. de la Clergerie à NAACL'01 (Pittsburgh) (présentation [14]), TALN'01 (présentation [11]), ACL'01 (présentation dans un atelier satellite [10]), IWPT'01 (Pékin) avec présentation de dernière minute.

Présentation par É. de la Clergerie d'un tutoriel à ICLP'01 (Chypre) [18].

Présentation par É. de la Clergerie de l'atelier TAG [17] au groupe XTAG lors d'une visite à l'université de Pennsylvanie (Philadelphie). Visite en Septembre de É. de la Clergerie au groupe CENTRIA de l'université nouvelle de Lisbonne dans le cadre de l'action ORLINSROC et présentation de ses travaux.

Participation de P. Boullier à ACL'01 (présentation [12]) et à FGMOL'01 (présentation [13]).

Participation de Ph. Deschamp à ACL'01.

P. Boullier a présenté les RCG au cours d'un séminaire organisé par le projet Contraintes de l'INRIA-Rocquencourt.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

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

[2] P. Boullier, « On TAG Parsing », *Traitement Automatique des Langues (T.A.L.) 41*, 3, 2000, p. 111–131, issued June 2001.

[3] B. Lang, « Complete Evaluation of Horn Clauses : an Automata Theoretic Approach », *rapport de recherche n°913*, INRIA, Rocquencourt, France, novembre 1988.

[4] B. Lang, « Towards a Uniform Formal Framework for Parsing », *in : Current issues in Parsing Technology*, M. Tomita (éditeur), Kluwer Academic Publishers, 1991, ch. 11, also appear in the Proc. of Int. Workshop on Parsing Technologies – IWPT89.

[5] E. 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`.

[6] E. 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.

Thèses et habilitations à diriger des recherches

[7] F. Role, *Vers un formalisme abstrait implémentable pour l'étude savante des textes numérisés*, thèse de doctorat, LIFO – Université d'Orléans, mai 2001, `http://www.inria.fr/rrrt/tu-0678.html`.

Articles et chapitres de livre

[8] P. Boullier, « Counting with Range Concatenation Grammars », *Theoretical Computer Science 5317*, 2001, Elsevier Science.

[9] V. J. Rocio, G. P. Lopes, E. Villemonte de la Clergerie, « Tabulation for multi-purpose parsing », *Grammars 4*, 1, 2001, p. 41–65.

Communications à des congrès, colloques, etc.

[10] F. Barthélemy, P. Boullier, P. Deschamp, L. Kaouane, A. Khajour, E. Villemonte de la Clergerie, « Tools and resources for Tree Adjoining Grammars », *in : Proceedings of ACL'01 workshop on Sharing Tools and Resources*, p. 63–70, Toulouse, France, juillet 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/wACL01.ps.gz`.

[11] F. Barthélemy, P. Boullier, P. Deschamp, L. Kaouane, E. Villemonte de la Clergerie, « Atelier ATOLL pour les grammaires d'arbres adjoints », *in : Proceedings of TALN'01*, p. 63–72, Tours, France, juillet 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/TALN01.ps.gz`.

[12] F. Barthélemy, P. Boullier, P. Deschamp, E. Villemonte de la Clergerie, « Guided Parsing of Range Concatenation Languages », *in : Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, p. 42–49, Université de Toulouse, France, juillet 2001.

[13] P. Boullier, « From Contextual Grammars to Range Concatenation Grammars », *in : Sixth Conference on Formal Grammar and Seventh Conference of Mathematics of Language (FG/MOL'01)*, University of Helsinki, Helsinki, Finlande, août 2001. Electronic Notes in Theoretical Computer Science 53 (`http://www.elsevier.nl/locate/entcs/volume53.html`), 12 pages.

[14] E. Villemonte de la Clergerie, « Refining Tabular Parsers for TAGs », *in : Proceedings of NAACL'01*, p. 167–174, CMU, Pittsburgh, PA, USA, juin 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/NAACL01-clerger.ps.gz`.

Divers

[15] S. Balva, *Analyse automatique d'un document botanique*, Mémoire DESS, Université d'Orléans, 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/RapportBalva01.doc`.

[16] A. Khajour, *Serveur de grammaire d'arbres adjoints*, Mémoire ingénieur, ENSIAS (Maroc), juin 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/RapportKhajour01.doc`.

[17] E. Villemonte de la Clergerie, « ATOLL TAG Workbench », mai 2001, Slides presented at IRCS, University of Pennsylvania.

[18] E. Villemonte de la Clergerie, « Natural Language Tabular Parsing », Slides for a tutorial delivered at ICLP'01, novembre 2001, `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/ICLP01-small.ps.gz`.

[19] E. Villemonte de la Clergerie, « Prefix-valid tabular parsers for Mildly-Context Sensitive Grammars », To be submitted, 2001.

[20] S. Werli, *Développement et évaluation d'une grammaire TAG pour le français*, Mémoire de DEA, Université Paris 7, septembre 2001.